# Statistical Signal Processing for Gene Microarrays

**Alfred O. Hero III**

***University of Michigan, Ann Arbor, MI***

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

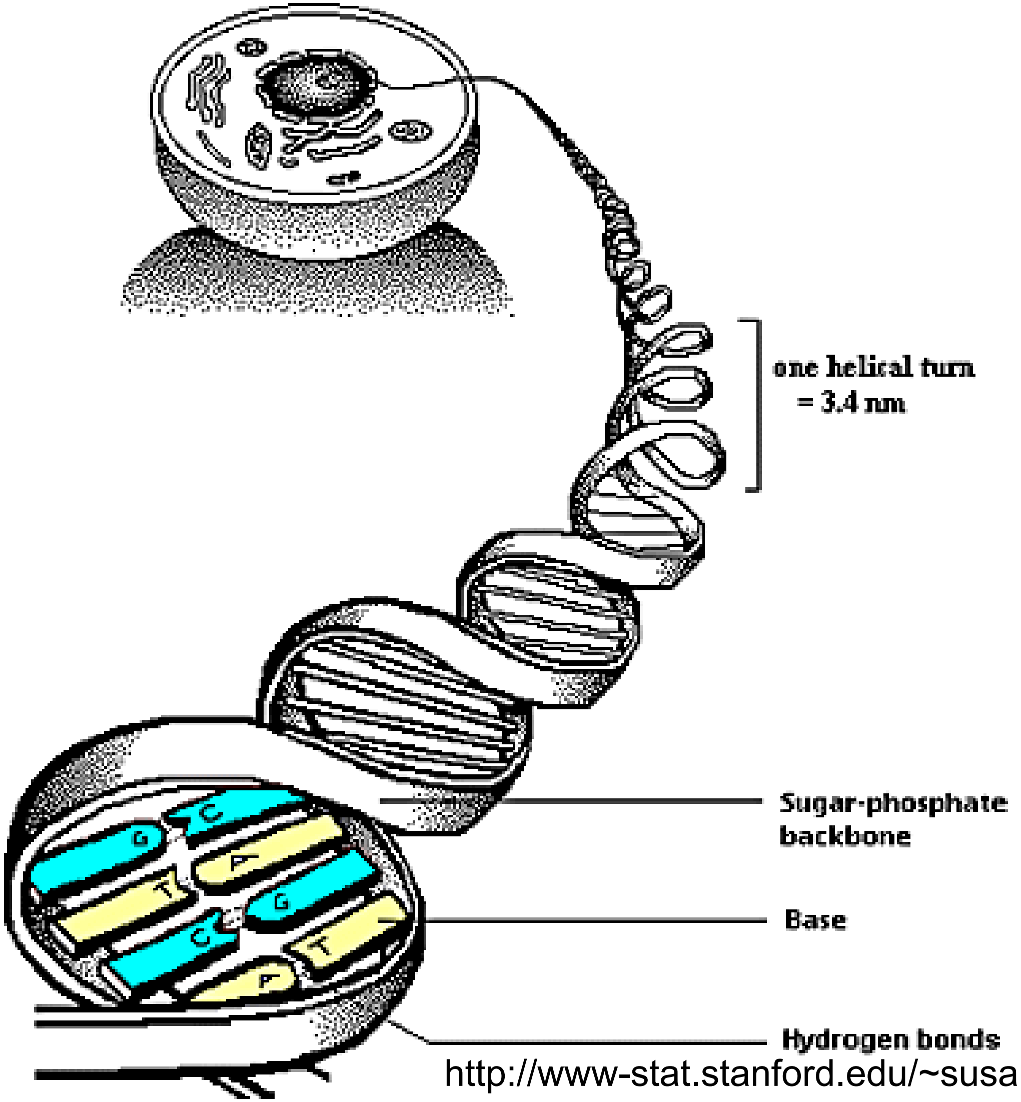
**Sept 2004**

1. Hierarchy of biological questions and gene microarrays
2. Analysis of gene microarray data
3. Gene filtering, ranking and clustering
4. Discovery or gene co-regulation networks
5. Wrap up and References

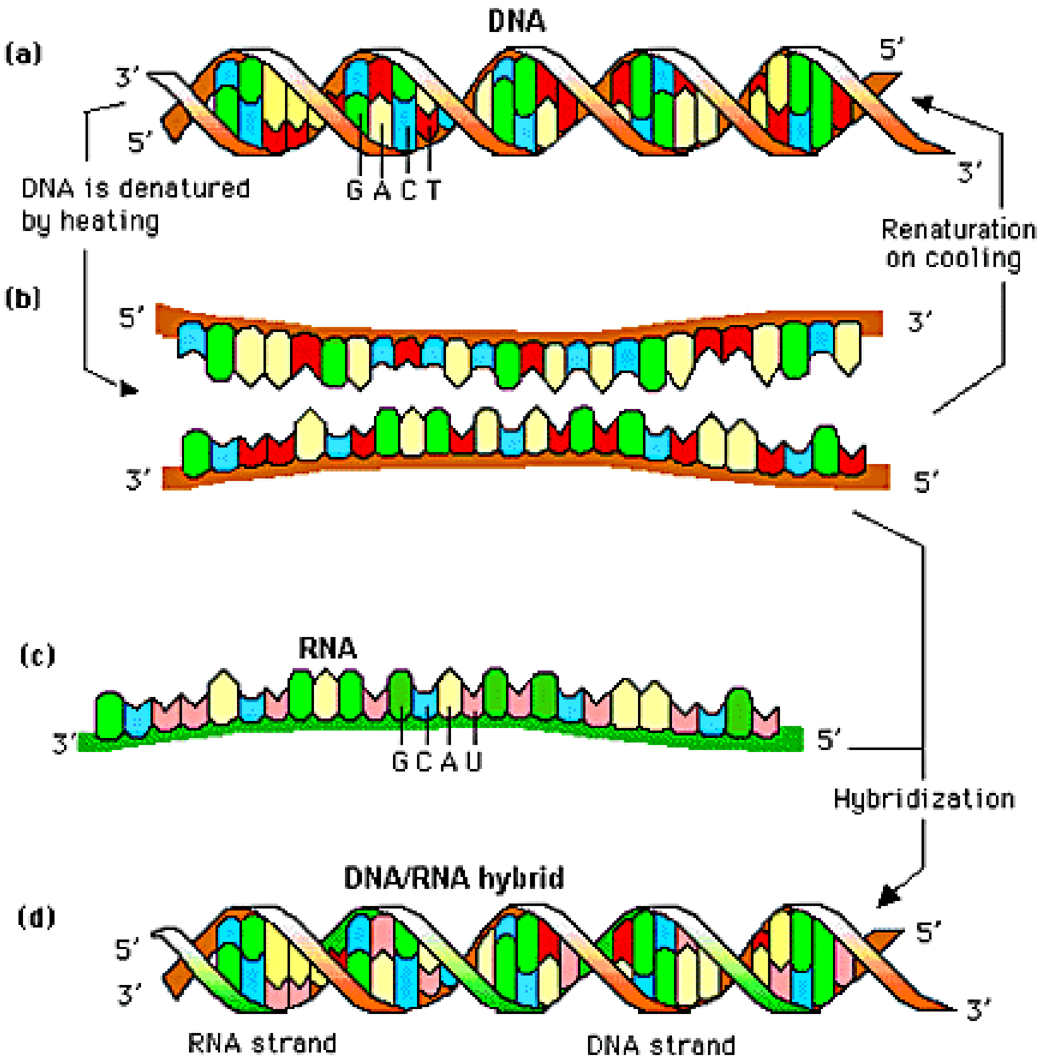
# 1. Hierarchy of biological questions

- Gene sequencing: what is the sequence of base pairs in a DNA segment, gene, or genome?
- Gene Mapping: what are positions (loci) of genes on a chromosome?
- Gene expression profiling: what is pattern gene activation/inactivation over time, tissue, therapy, etc?
- Genetic circuits: how do genes regulate (stimulate/inhibit) each other's expression levels over time?
- Genetic pathways: what sequence of gene interactions lead to a specific metabolic/structural (dys)function?

# THE STRUCTURE OF DNA



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



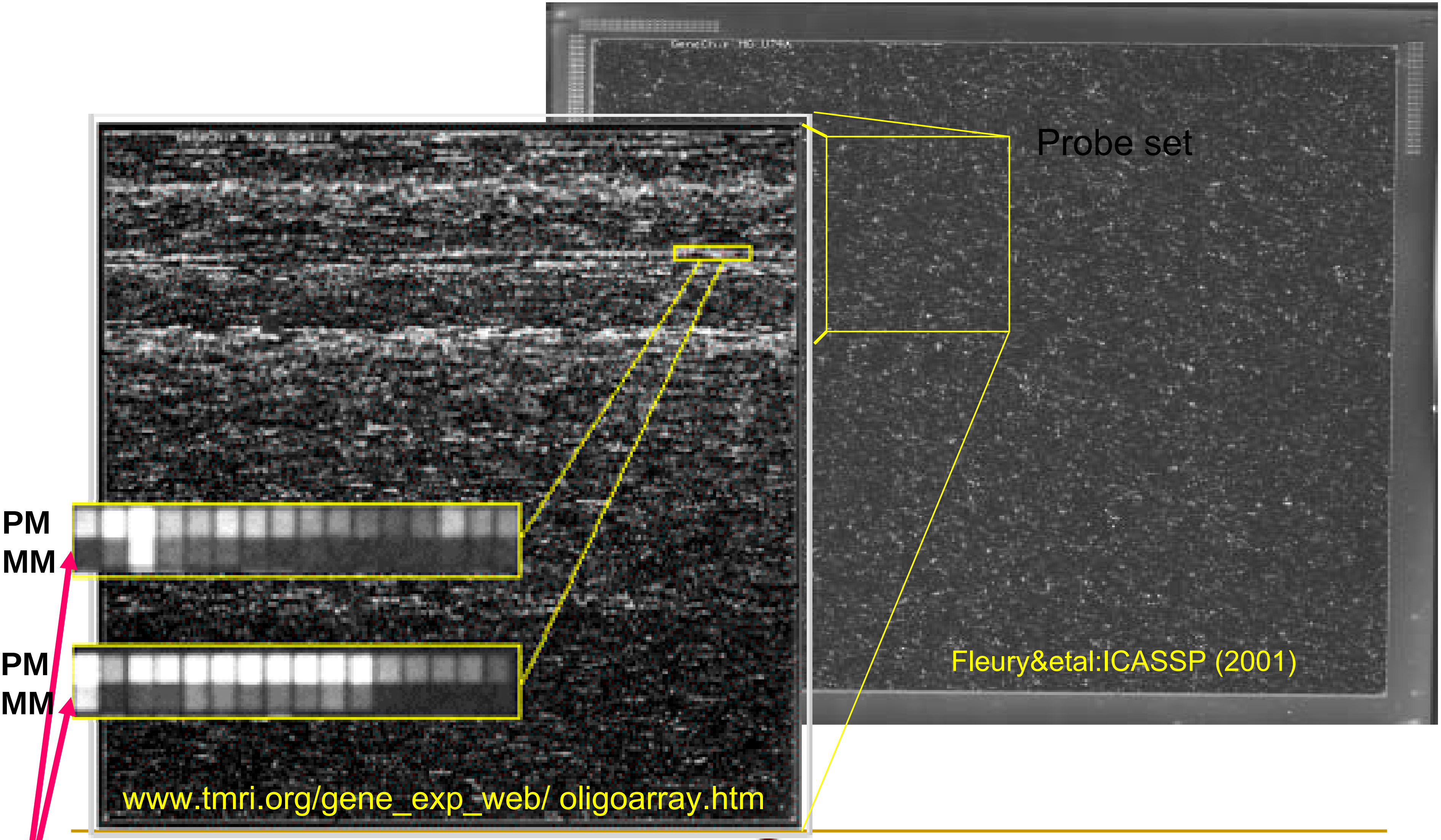
Nucleic Acid Hybridization

# 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

PM
MM

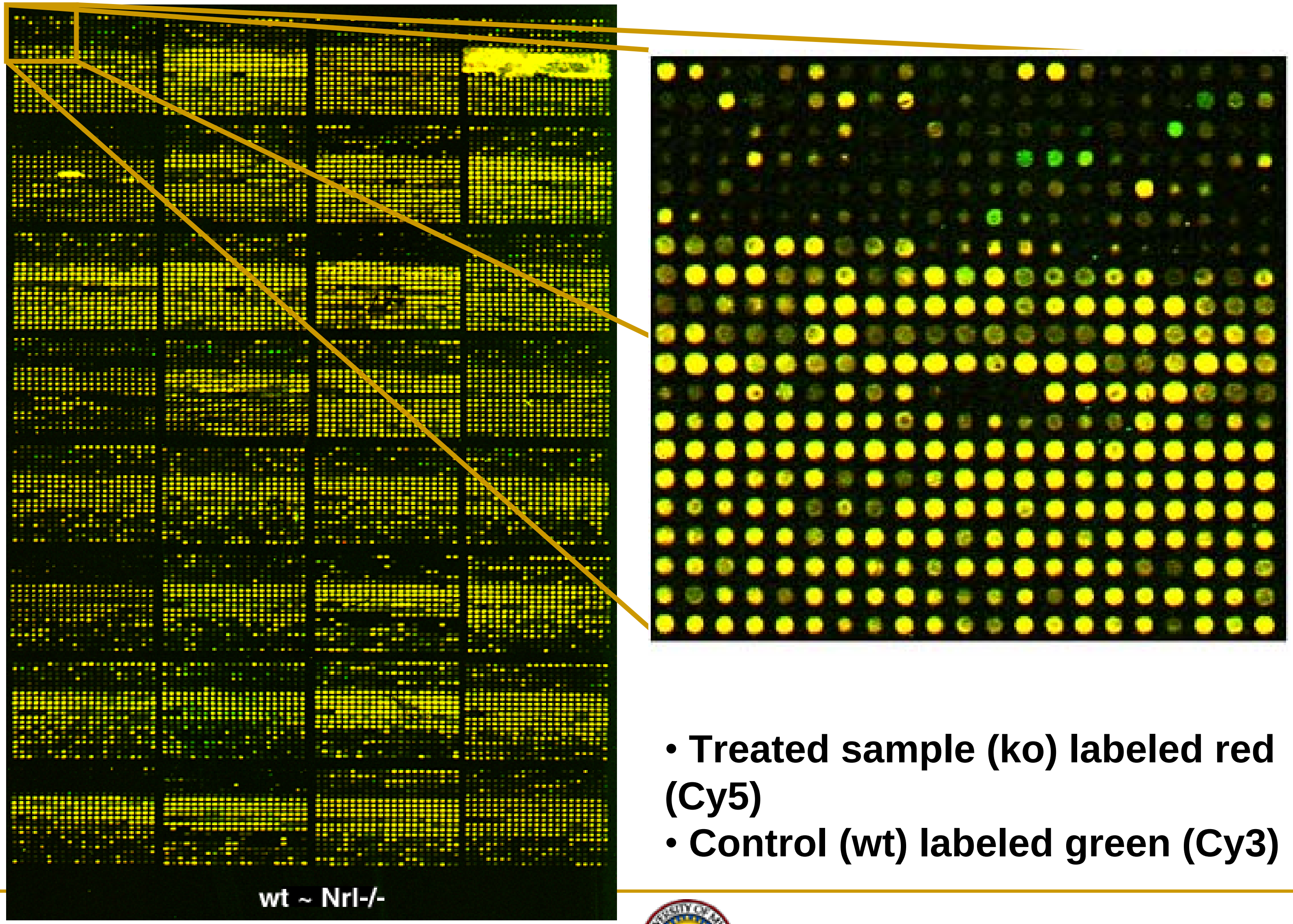
PM
MM

www.tmri.org/gene_exp_web/ oligoarray.htm

Fleury&etal:ICASSP (2001)

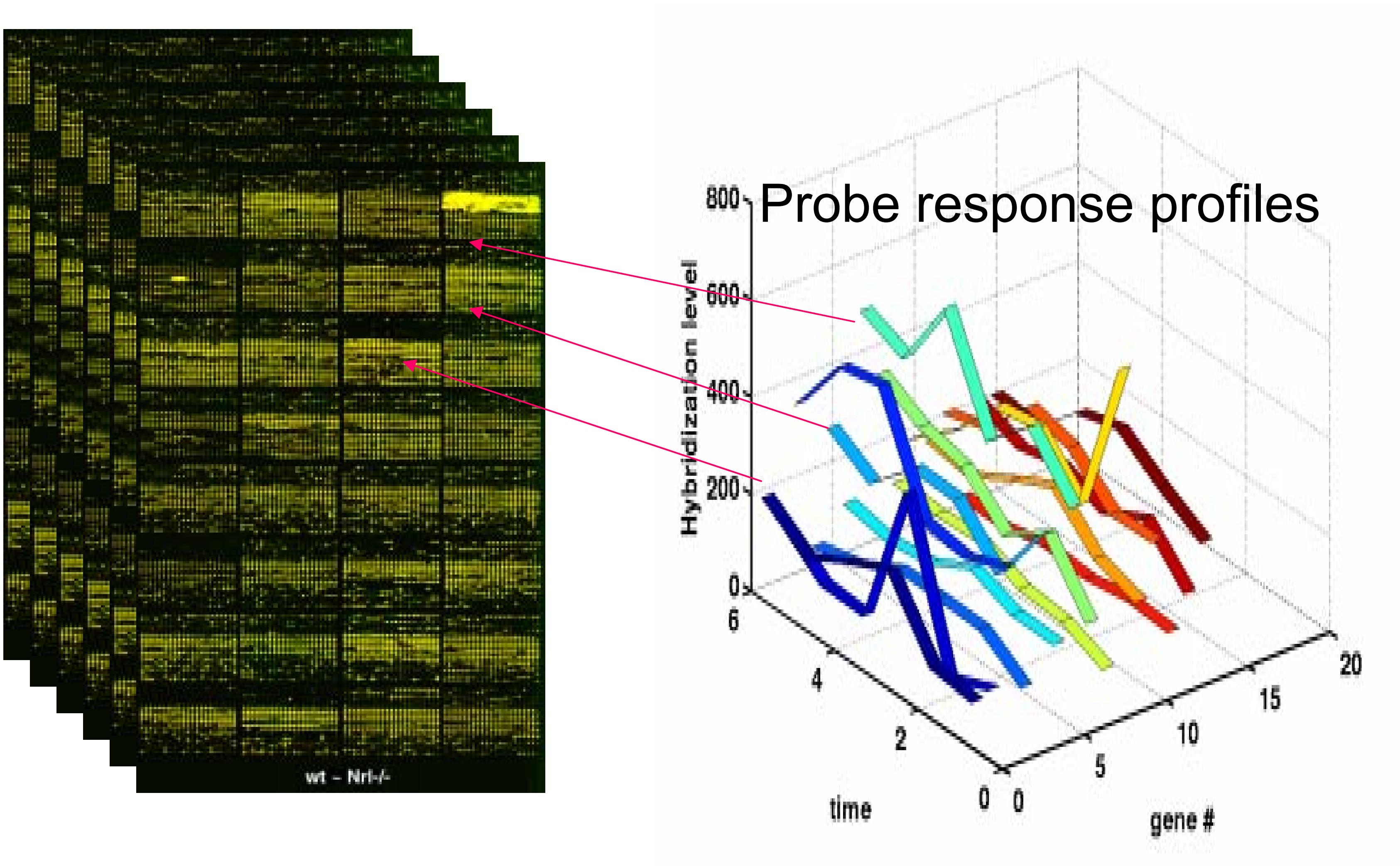
Two PM/MM Probe sets

# cDNA spotted array



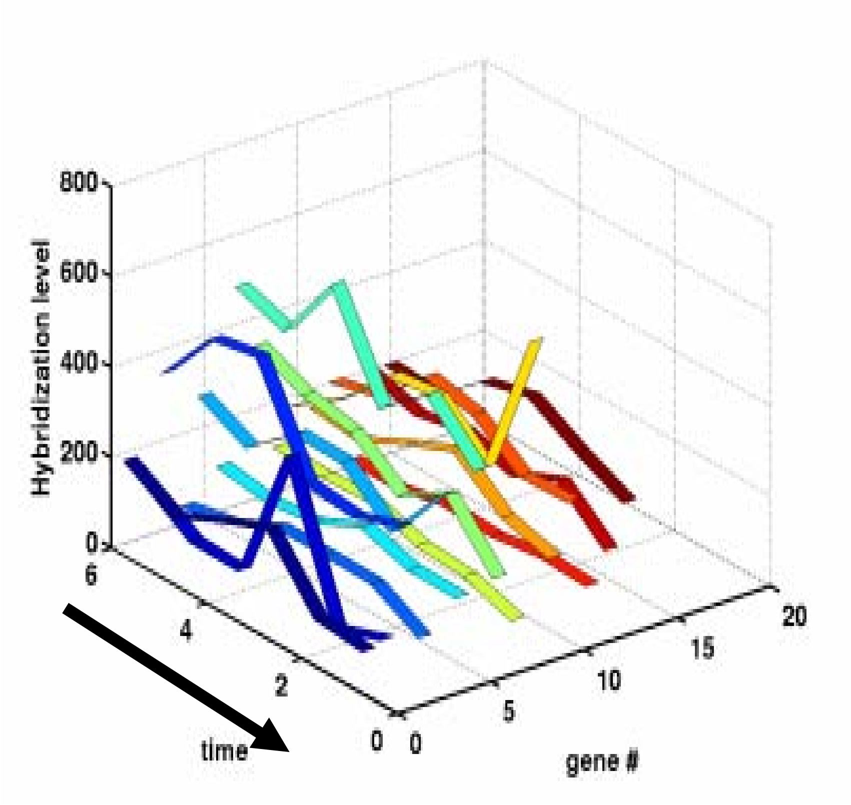
- Treated sample (ko) labeled red (Cy5)
- Control (wt) labeled green (Cy3)

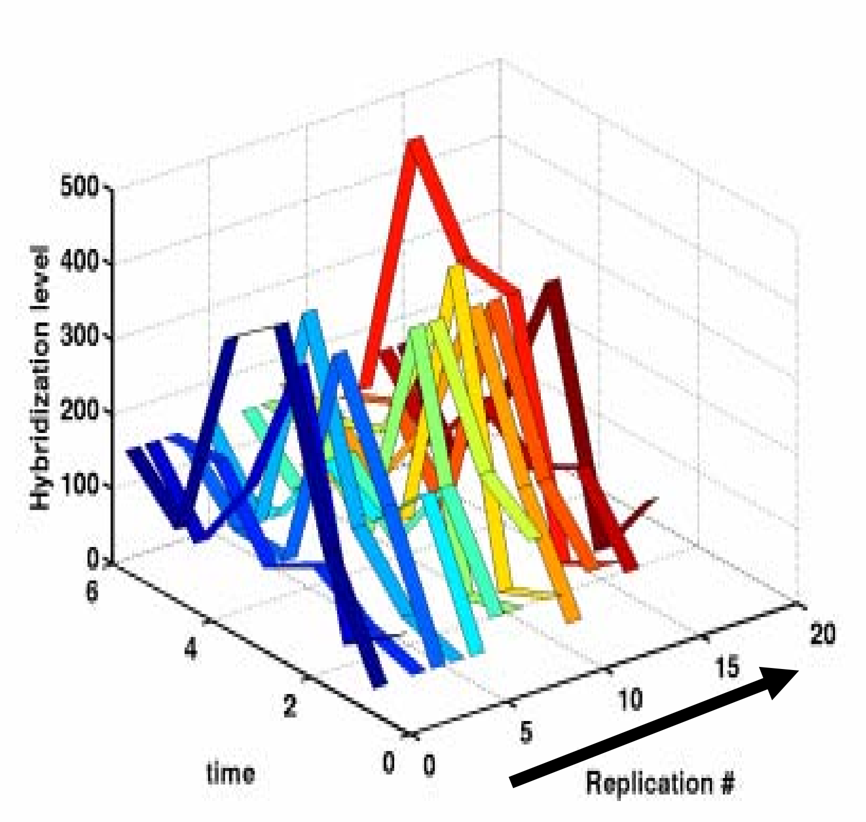
# Add Treatment Dimension: Expression Profiles



Probe response profiles

# Problem of Sample Variability





Across-treatment variability

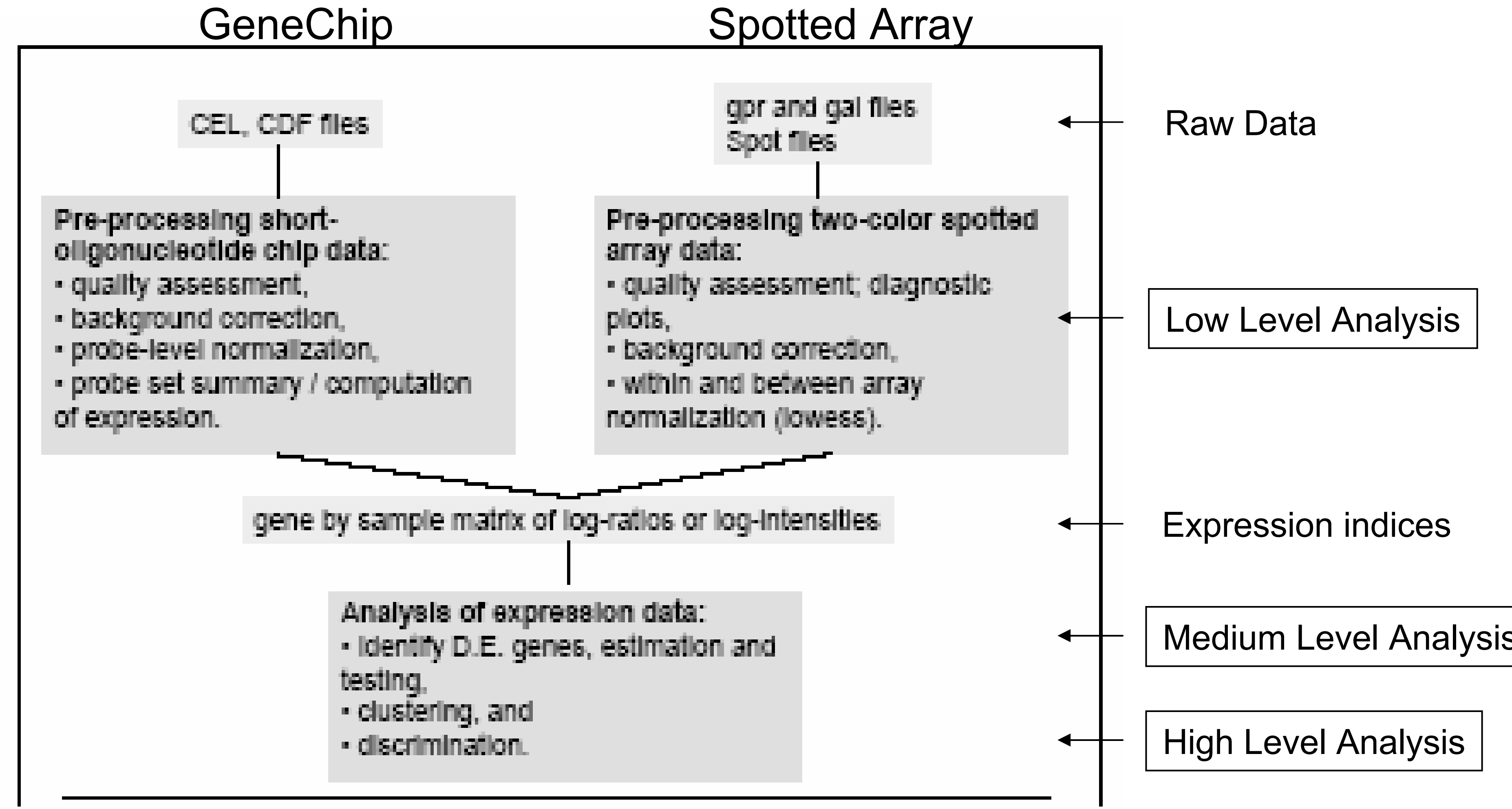
Across-sample variability

# 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

# 2. Analysis of gene microarray data



GeneChip

Spotted Array

CEL, CDF files

gpr and gal files
Spot files

Raw Data

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).

Low Level Analysis

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

Expression indices

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

Medium Level Analysis

High Level Analysis

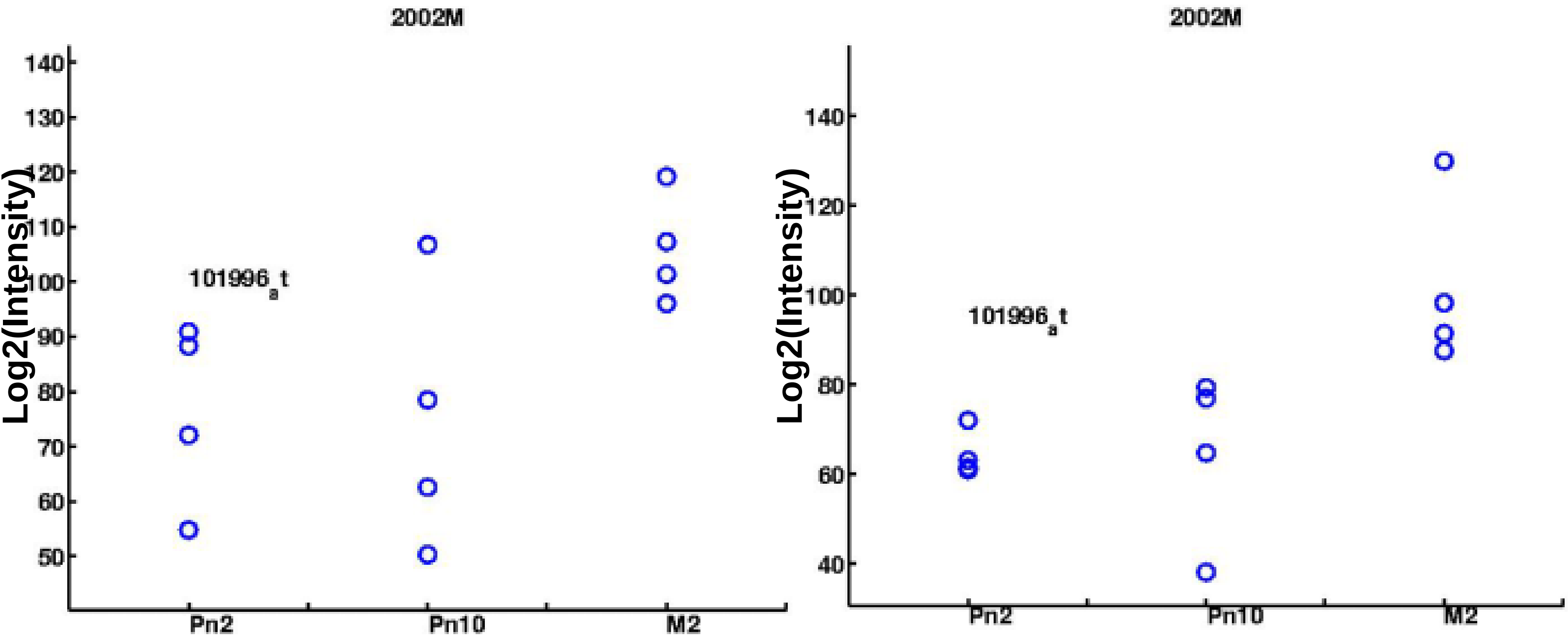
Source: Jean Yee Hwa Yang Statistical issues in design and analysis microarray experiment. (2003)

# Knockout vs Wildtype Retina Study

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:

- Statistical significance refers to foldchange being different from zero

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

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

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

Hero,Fleury,Mears,Swaroop:JASP2003

# 3. Gene Filtering, Ranking and Clustering

- Let $\mathrm{fc}_t(g)$ = foldchange of gene ‘g’ at time point ‘t’.
- We wish to simultaneously test the TG sets of 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)
- Two stage procedure:
  - **Statistical Significance**: Simultaneous Paired t-test
  - **Biological Significance**: Simultaneous Paired t confidence intervals for fc(g)’s

Hero,Fleury,Mears,Swaroop:JASP2003

# Single-Comparison: Paired t statistic

- PT statistic with ‘m’ replicates of wt&ko:

$$\boxed{T_t(g) = \sqrt{m/2}\,\frac{\overline{W}_t(g) - \overline{K}_t(g)}{s_t(g)}}$$

- Level α test: Reject H0(g,t) unless:

$$\boxed{-\mathcal{T}^{-1}_{1-\alpha/2} < T_t(g) < \mathcal{T}^{-1}_{1-\alpha/2}}$$

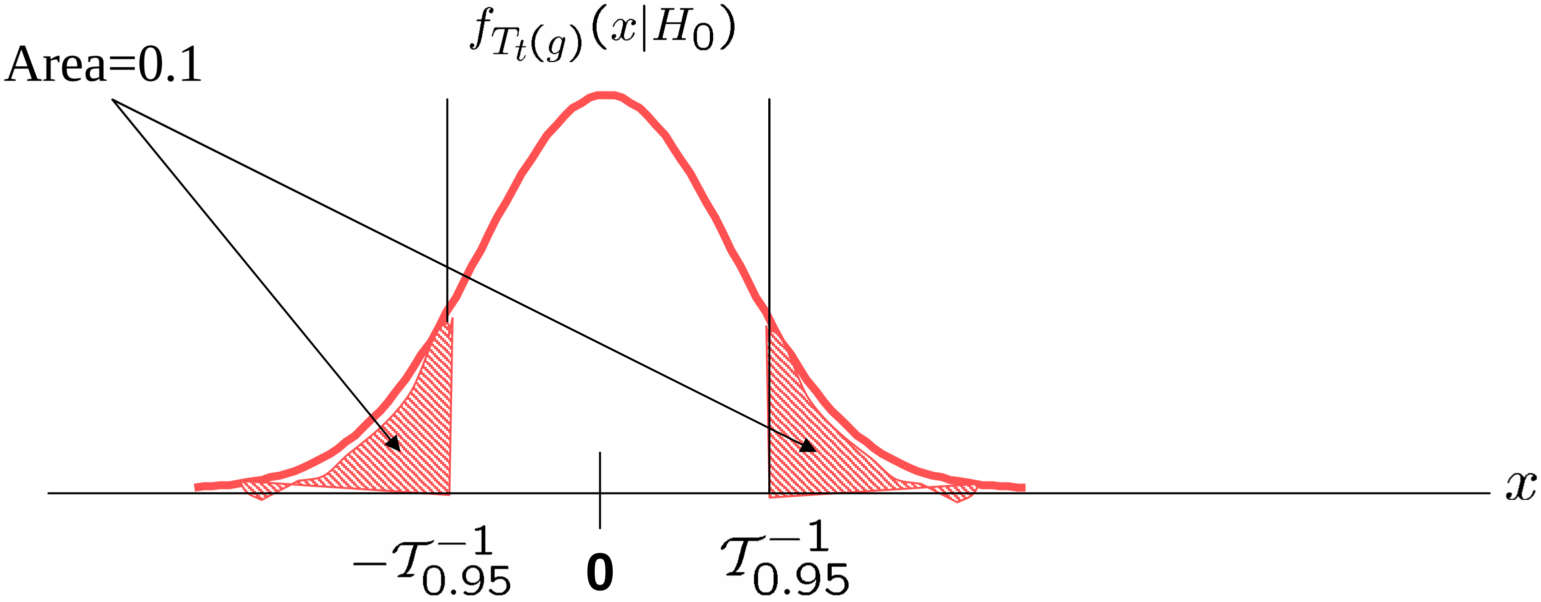
- Level 1-α confidence interval (CI) on fc:

$$\boxed{I_g(\alpha) = T_t(g) \pm \sqrt{\frac{2}{m}\mathcal{T}^{-1}_{1-\alpha/2}}}$$

- p-th quantile of student-t with 2(m-1) df: $\mathcal{T}^{-1}_p$

# Stage 1: paired T test of level alpha=0.1

$$
\begin{aligned}
H_0 &: \mathsf{fc}_t(g) = 0 \\
H_1 &: \mathsf{fc}_t(g) \neq 0
\end{aligned}
$$

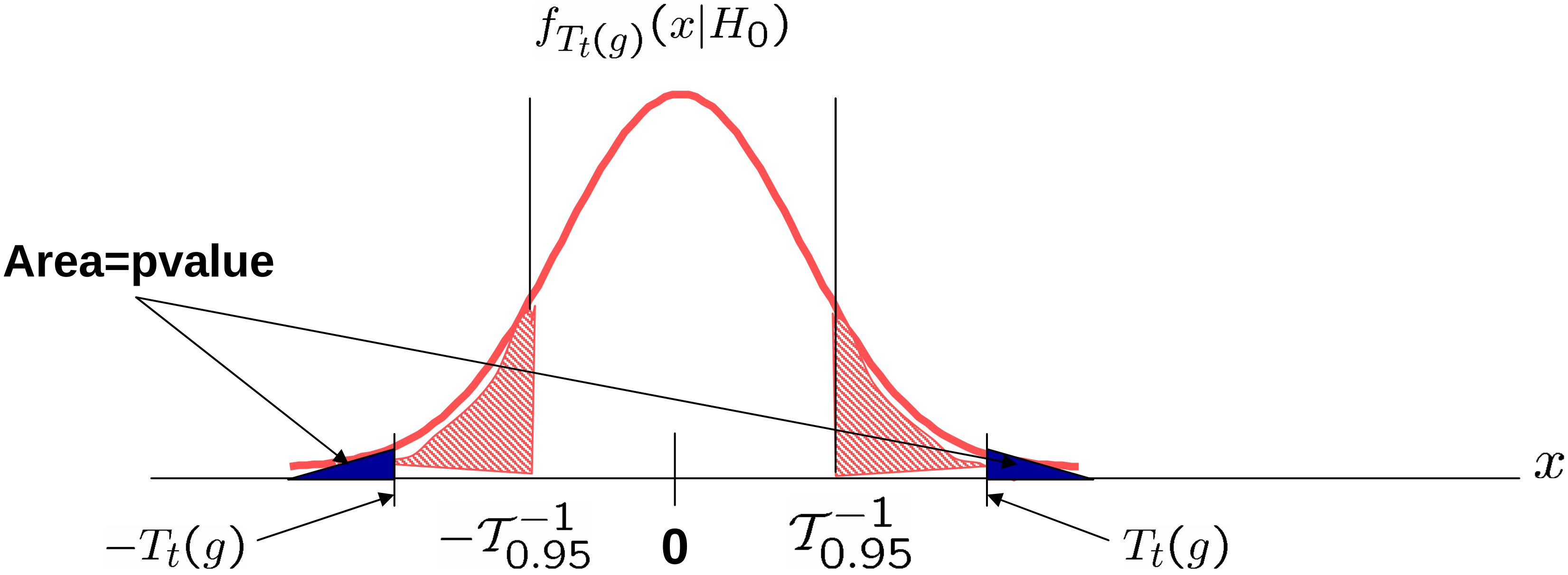


For single comparison: a false positive occurs with probability $\alpha$=0.1

# Stage 1: paired T test of level alpha=0.1

$$H_0 \;:\; \mathsf{fc}_t(g) = 0$$

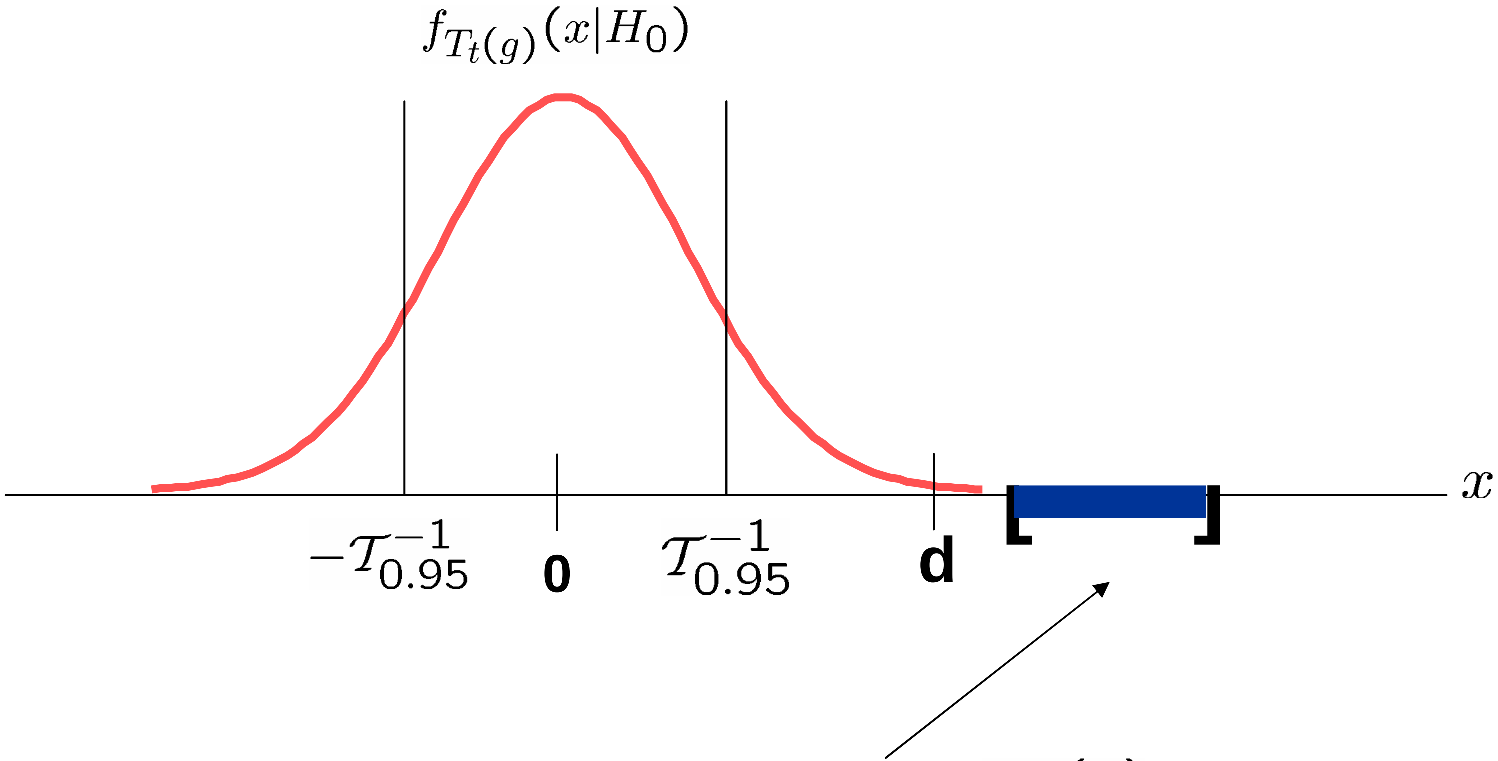
$$H_1 \;:\; \mathsf{fc}_t(g) \neq 0$$



For single comparison: a false positive occurs with probability $\alpha$=0.1

# Stage 2: Confidence Intervals

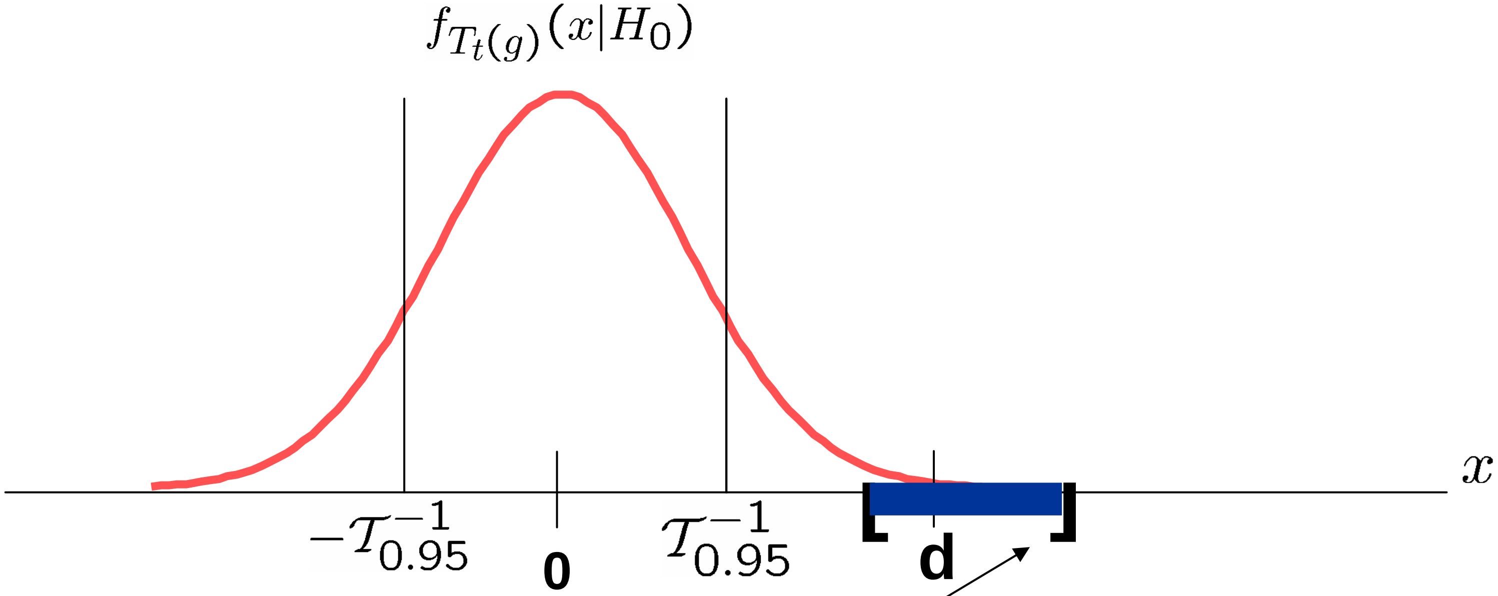
- Biologically&statistically **significant** differential response



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

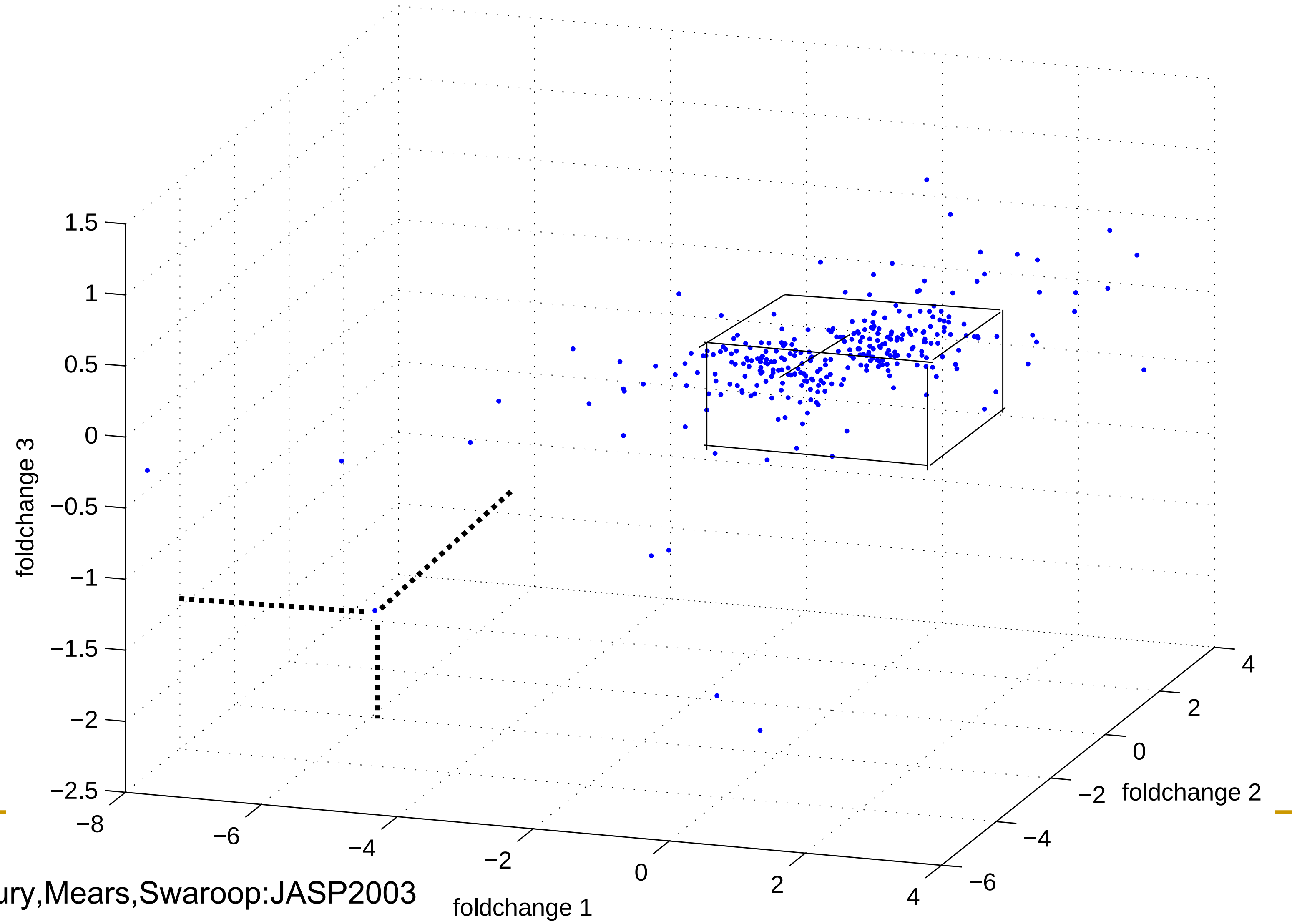
# Stage 2: Confidence Intervals

- Biologically&statistically **insignificant** differential response



**Conf. Interval on $fc_t(g)$ of level 1-alpha**

# Minimum fc cube for single gene profile



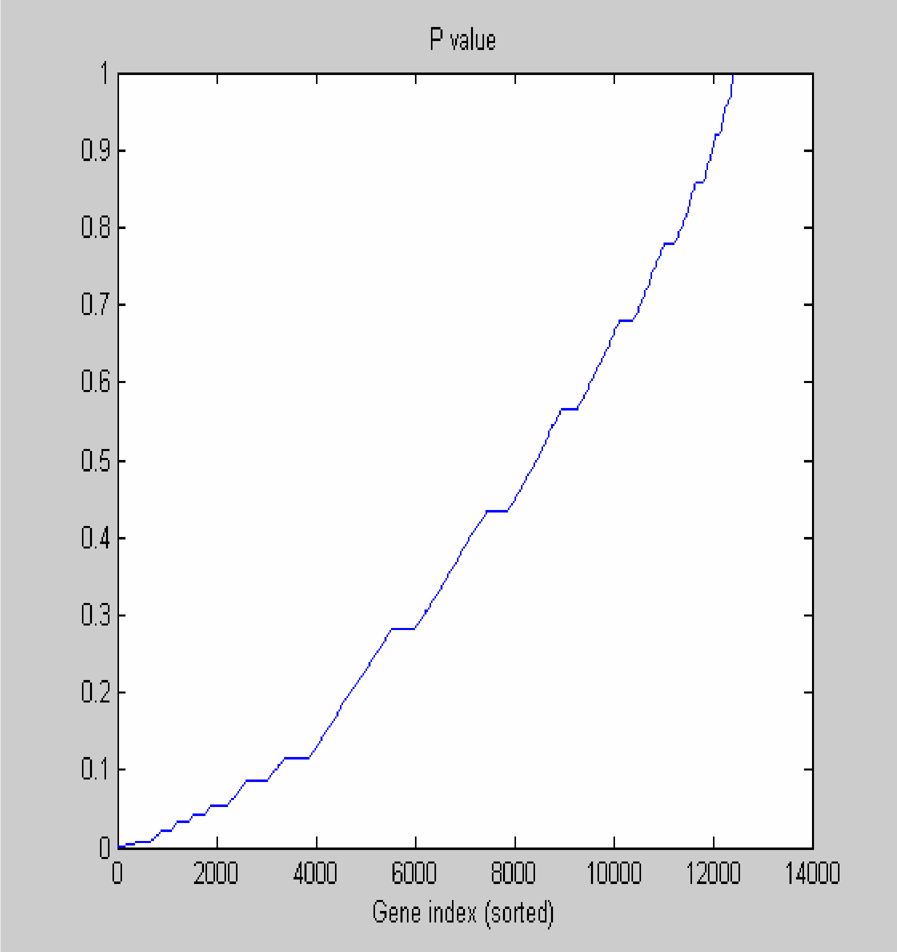
Hero,Fleury,Mears,Swaroop:JASP2003

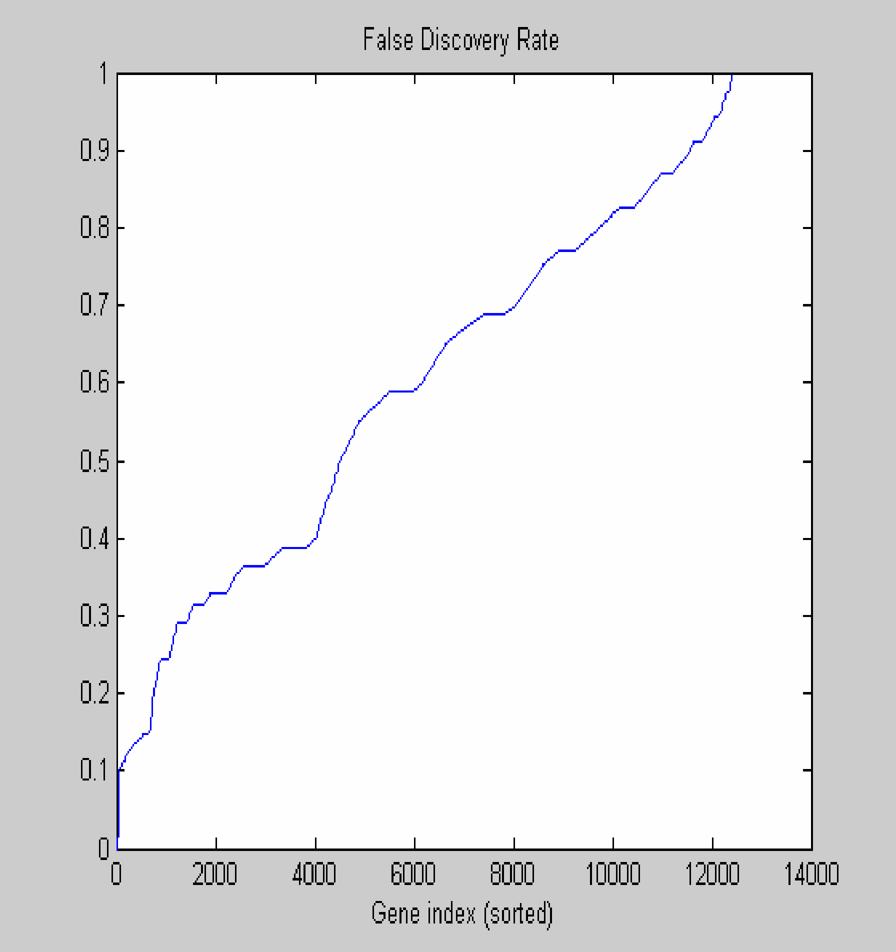
# Multiple Comparisons: FWER, FDR

- **Pvalue,CI** apply to single comparison: **T(g)** dependence.

- **FWER, FDR** and **FDRCI** depend on **{T(g), g=1, … G}**.
  - FWER: familywise error rate (Miller:1976)
    - Avg number of experiments yielding at least one false positive
  - FDR: false discovery rate (Benjamini&Hochburg:1996)
    - Avg number of false positives in a given experiment
  - FDRCI: (1-$\alpha$) CI on discovered fc (Benjamini&Yekutieli:2002)
    - Avg. number of intervals that cover true fc in a given experiment

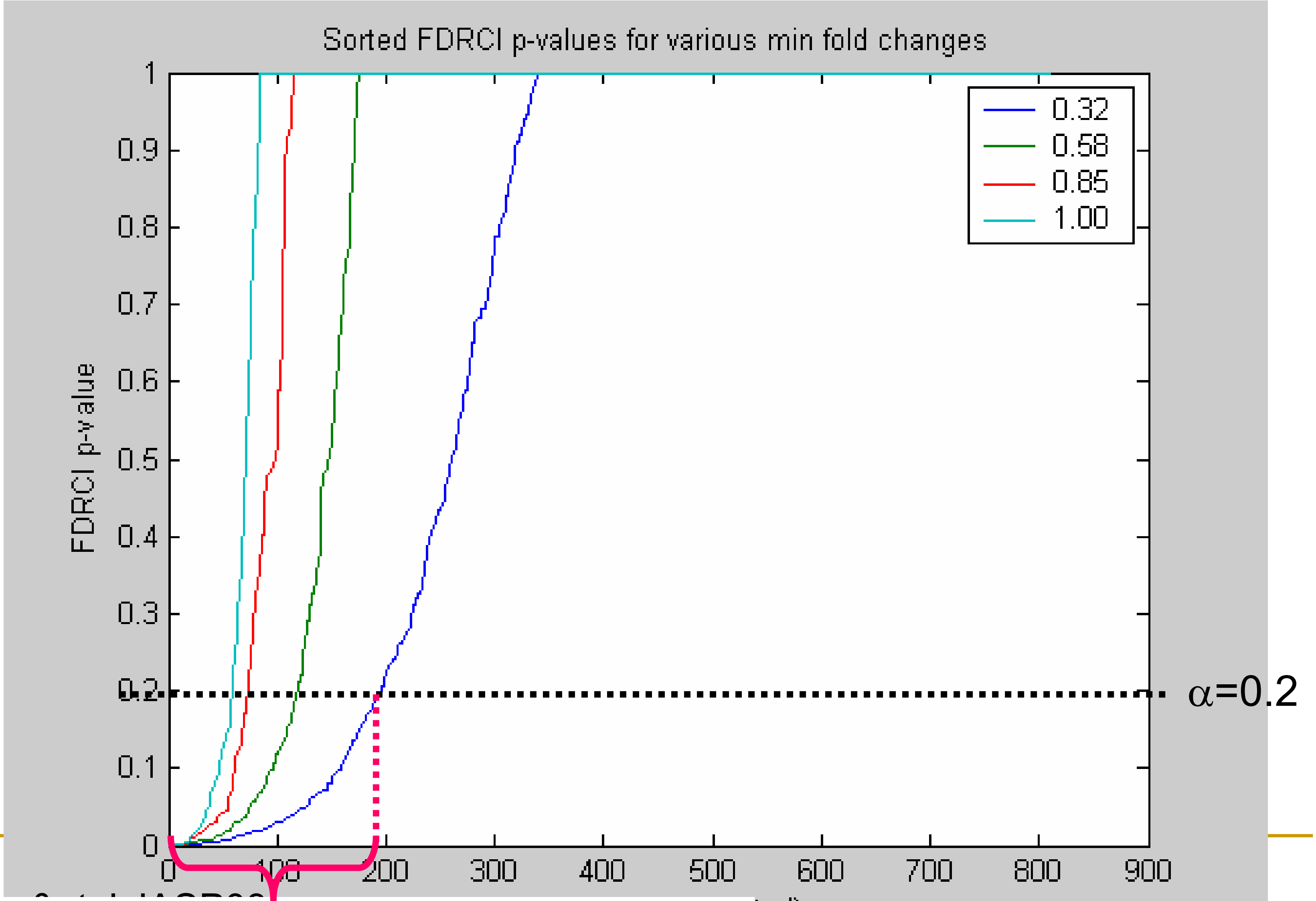
# P-value vs FDR Comparison for wt/ko





Ref: Hero&etal:JASP03

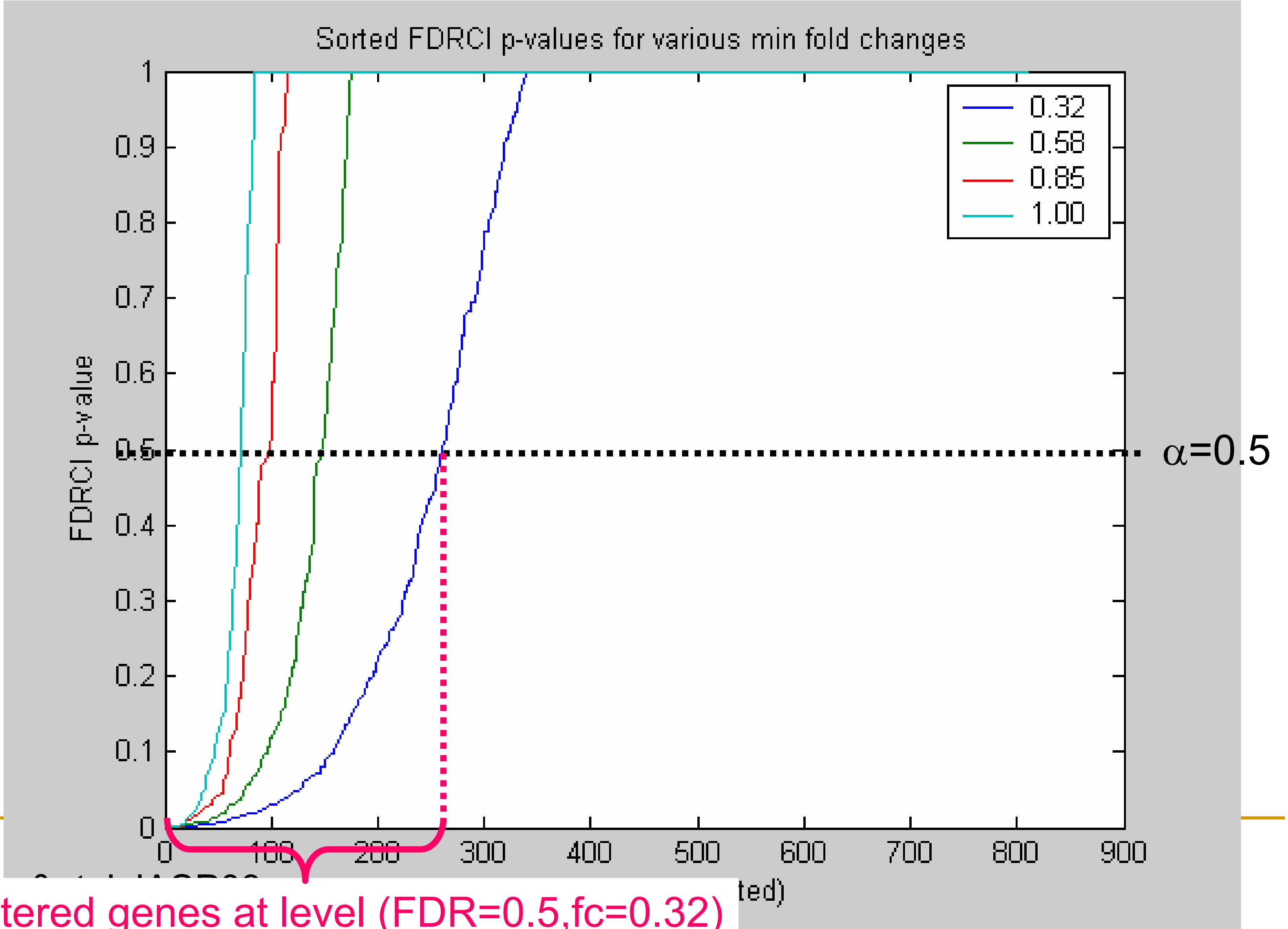
# Sorted FDRCI pvalues for ko/wt study



Ref:

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

# Sorted FDRCI pvalues for ko/wt study

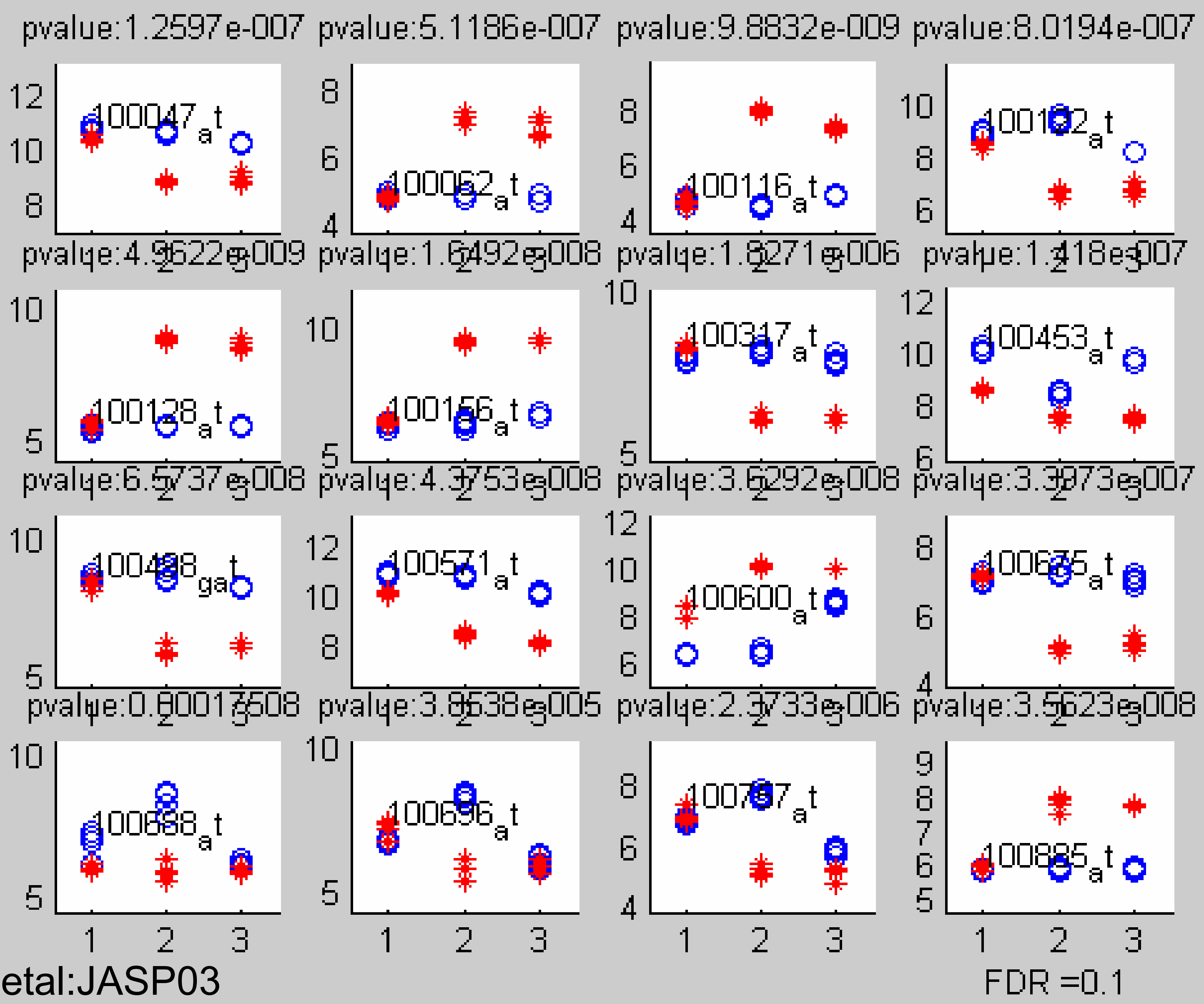


α=0.5

Filtered genes at level (FDR=0.5,fc=0.32)

Ref:

# FDRCI Results for ko/wt Data



Ref: Hero&etal:JASP03

# Ranking differential gene profiles

- Objective: find the 250-300 genes having the most significant foldchanges wrt multiple criteria

$$\xi_1(g),\ldots,\xi_P(g)$$

- Examples of increasing criteria:

$$\xi_1(g) = \overline{\mathsf{fc}}_1(g) \text{ Ko-Wt foldchange}$$
$$\xi_2(g) = \overline{\mathsf{fc}}_2(g) \text{ Ko-Wt foldchange}$$
$$\xi_3(g) = \overline{\mathsf{fc}}_3(g) \text{ Ko-Wt foldchange}$$

- Examples of mixed increasing and decreasing

$$\xi_1(g) = \mathsf{s}_K(g) = \text{Ko sample dispersion}$$
$$\xi_2(g) = \mathsf{s}^2_W(g) = \text{Wt sample dispersion}$$
$$\xi_3(g) = |\overline{K}(g) - \overline{W}(g)| = \text{Kp-Wt mean disp}$$

# Pareto Front Analysis (PFA)

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

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

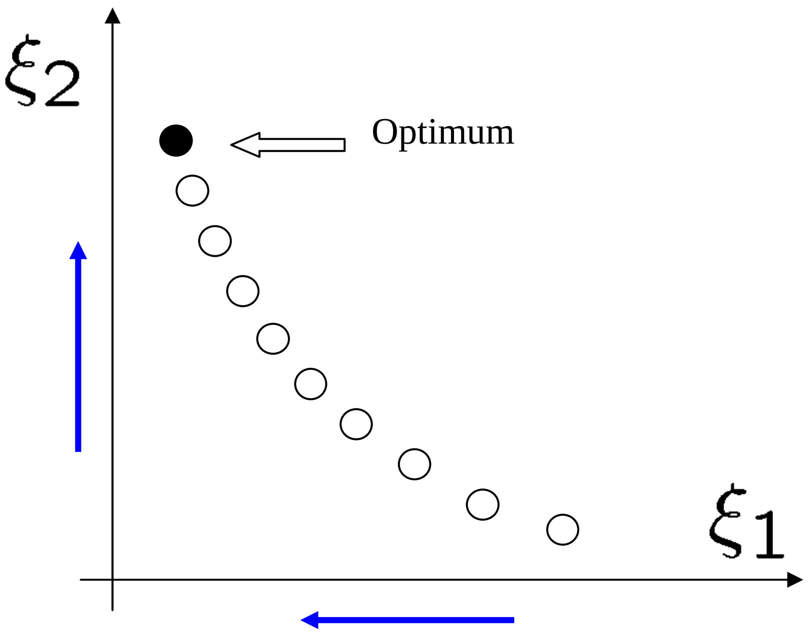
- However, a partial order is usually possible

$$\{\mathsf{fc}_1(g), \mathsf{fc}_2(g), \mathsf{fc}_3(g)\}_{g\in\mathcal{G}_1} > \ldots > \{\mathsf{fc}_1(g), \mathsf{fc}_2(g), \mathsf{fc}_3(g)\}_{g\in\mathcal{G}_q}$$

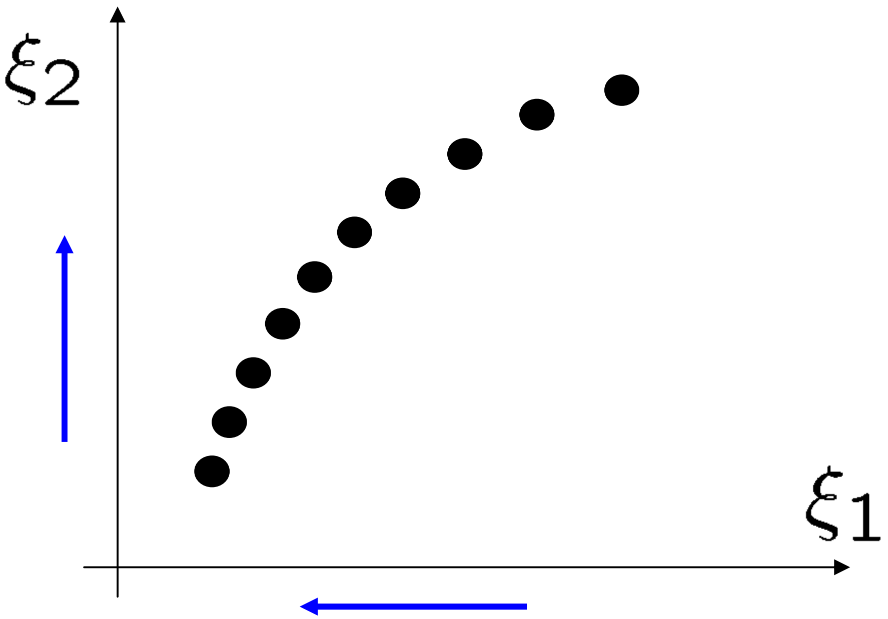
# Illustration of two extreme cases

$$\xi_1 = \sqrt{(s_K^2 + s_W^2)/2} = \text{pooled sample dispersion}$$
$$\xi_2 = |\overline{K} - \overline{W}| = \text{mean treatment dispersion}$$

- A linear ordering exists

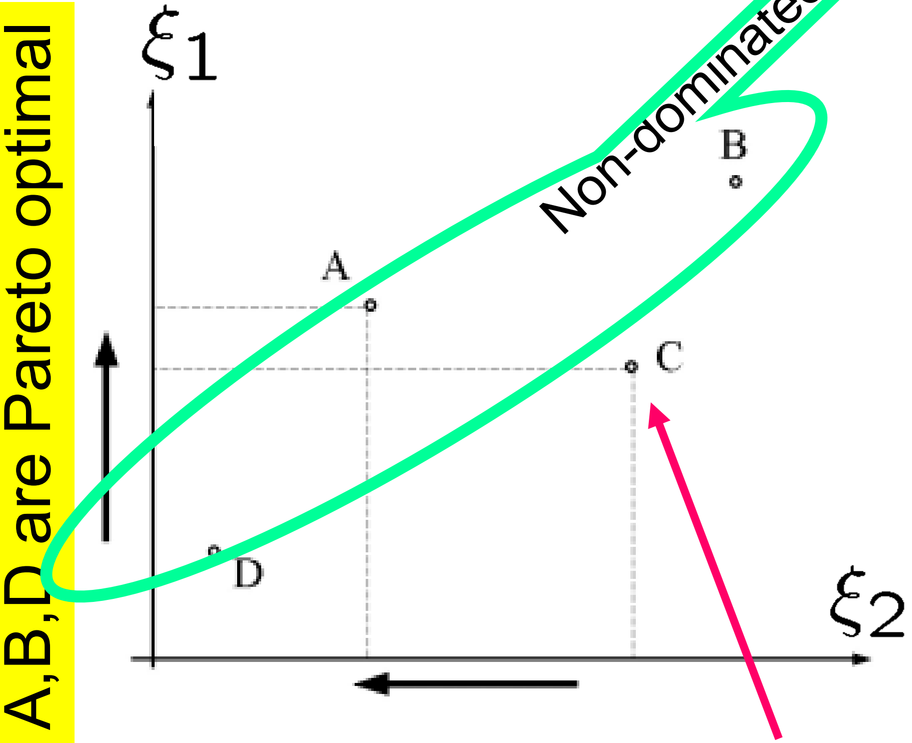


- No partial ordering exists



# Multicriteria Gene Ranking

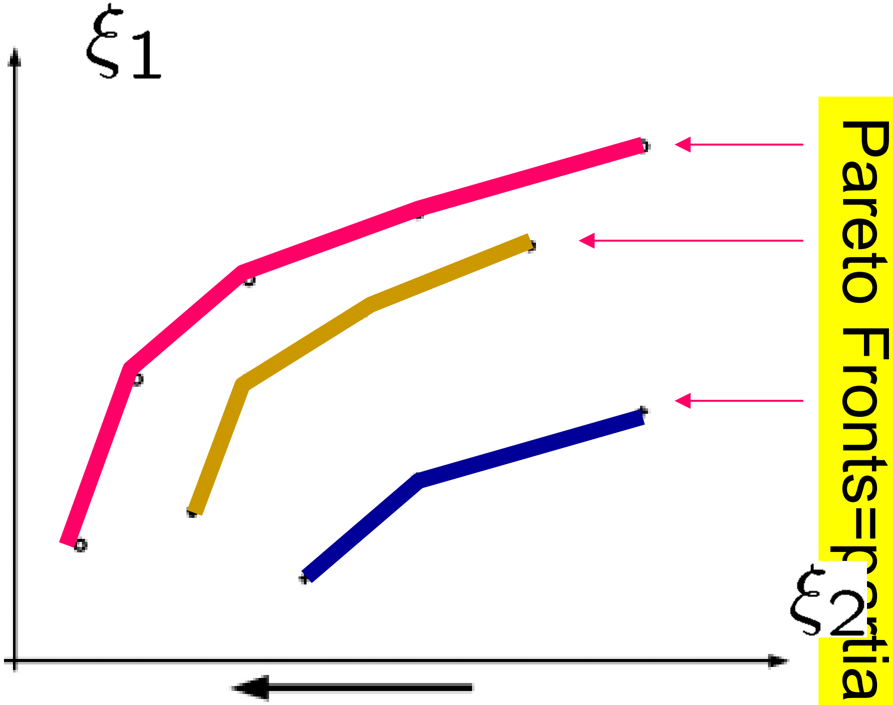
- Increasing $\xi_1$
- Decreasing $\xi_2$



Non-dominated genes=Pareto Front

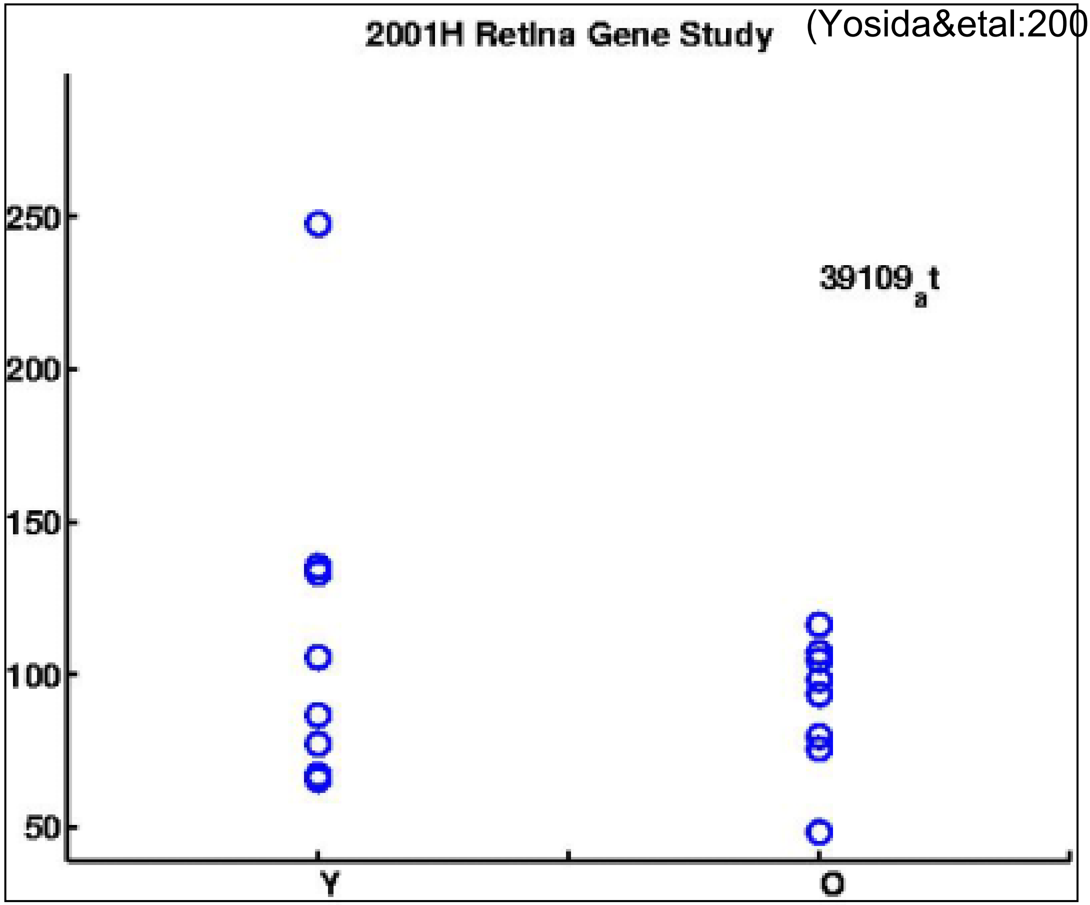
Dominated gene

A,B,D are Pareto optimal



Pareto Fronts=partial order

# Ranking Based on End-to-End Foldchange



(Yosida&etal:2002)

Y/O Human Retina Aging Data

- **16 human retinas**
- **8 young subjects**
- **8 old subjects**
- **8226 probesets**

$$\begin{aligned}\xi_1(g) &= \sqrt{(\sigma_O^2(g)+\sigma_Y^2(g))/2} \\ \xi_2(g) &= |\overline{O}(g)-\overline{Y}(g)|\end{aligned}$$

Ref: Fleury&etal ICASSP-02

# Multicriteria Y/O Gene Ranking

- Paired t-test at level of significance alpha:

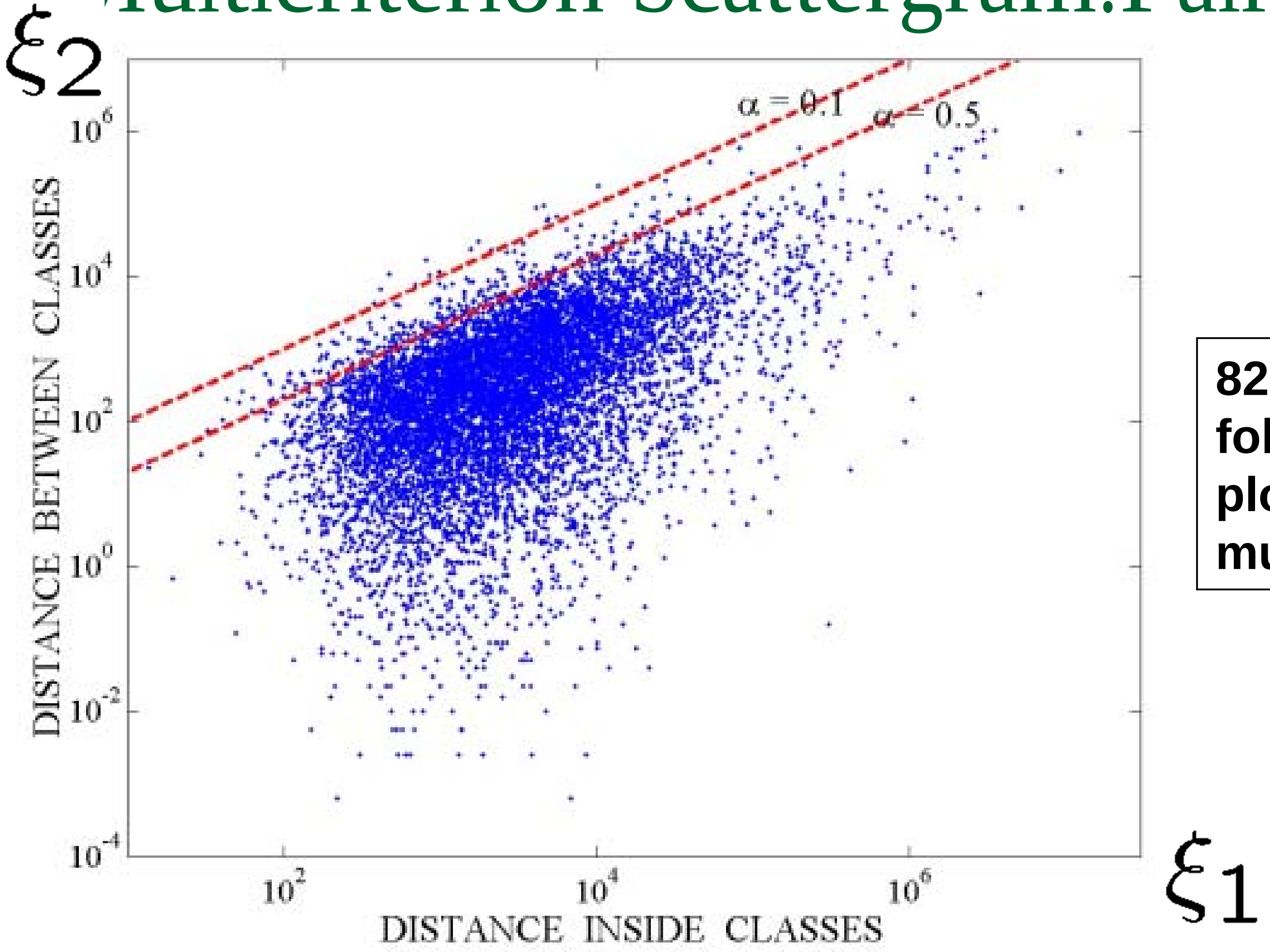
$$T(g) = \frac{\xi_2(g)}{\xi_1(g)} \begin{matrix} > \\ < \end{matrix} \sqrt{2/m}\, \mathcal{T}^{-1}_{1-\alpha/2}$$

- For Y/O Human study:

$$T(g) = \frac{|\overline{O}(g) - \overline{Y}(g)|}{\sqrt{(\sigma^2_O(g) + \sigma^2_Y(g))/2}}$$

Ref: Fleury&etal ICASSP-02

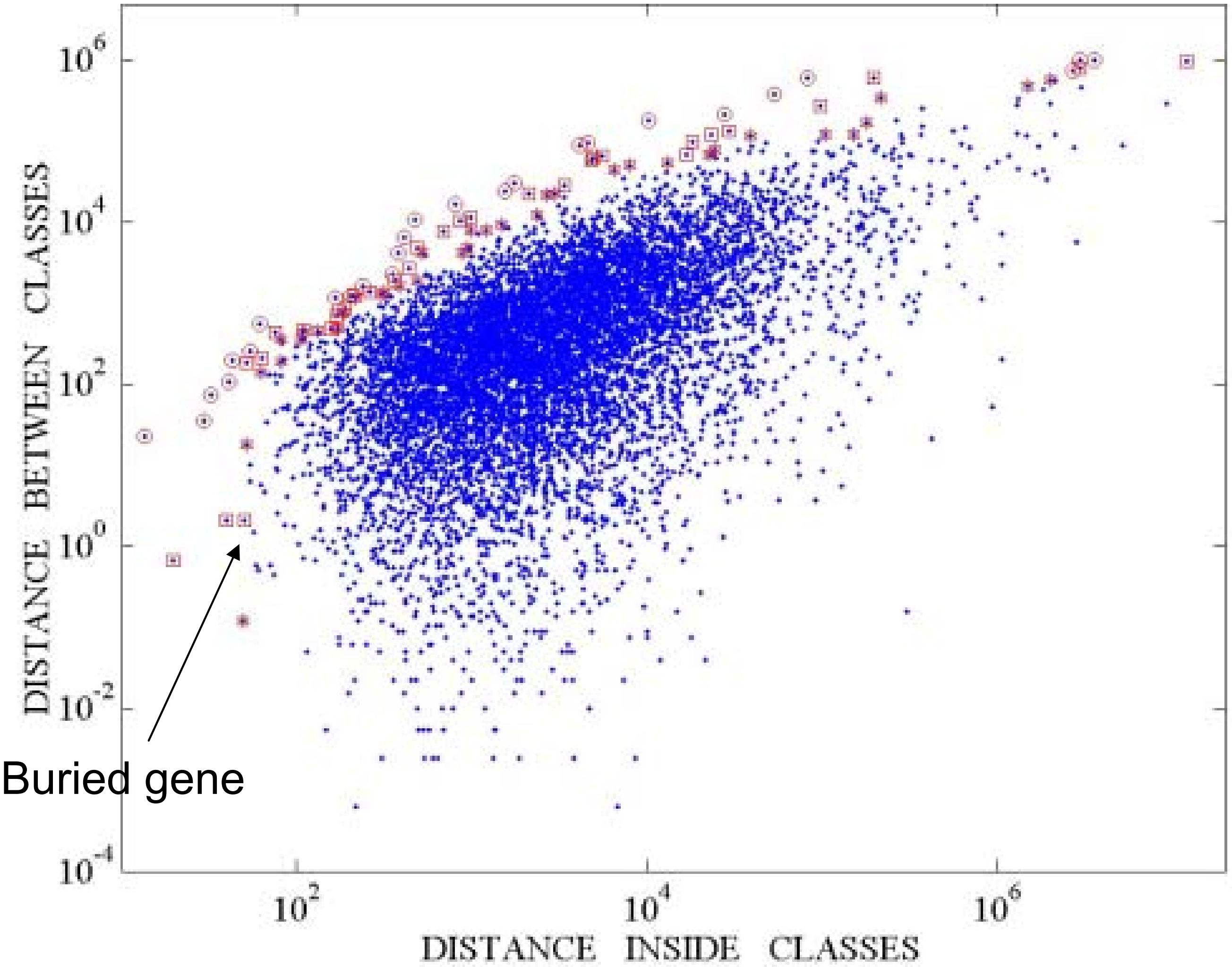
# Multicriterion Scattergram:Paired t-test



**8226 Y/O mean foldchanges plotted in multicriteria plane**

Ref: Fleury&etal ICASSP-02

# Multicriterion scattergram: Pareto Fronts



Buried gene

Pareto fronts

○ *first*

□ *second*

☆ *third*

Ref: Fleury&etal ICASSP-02

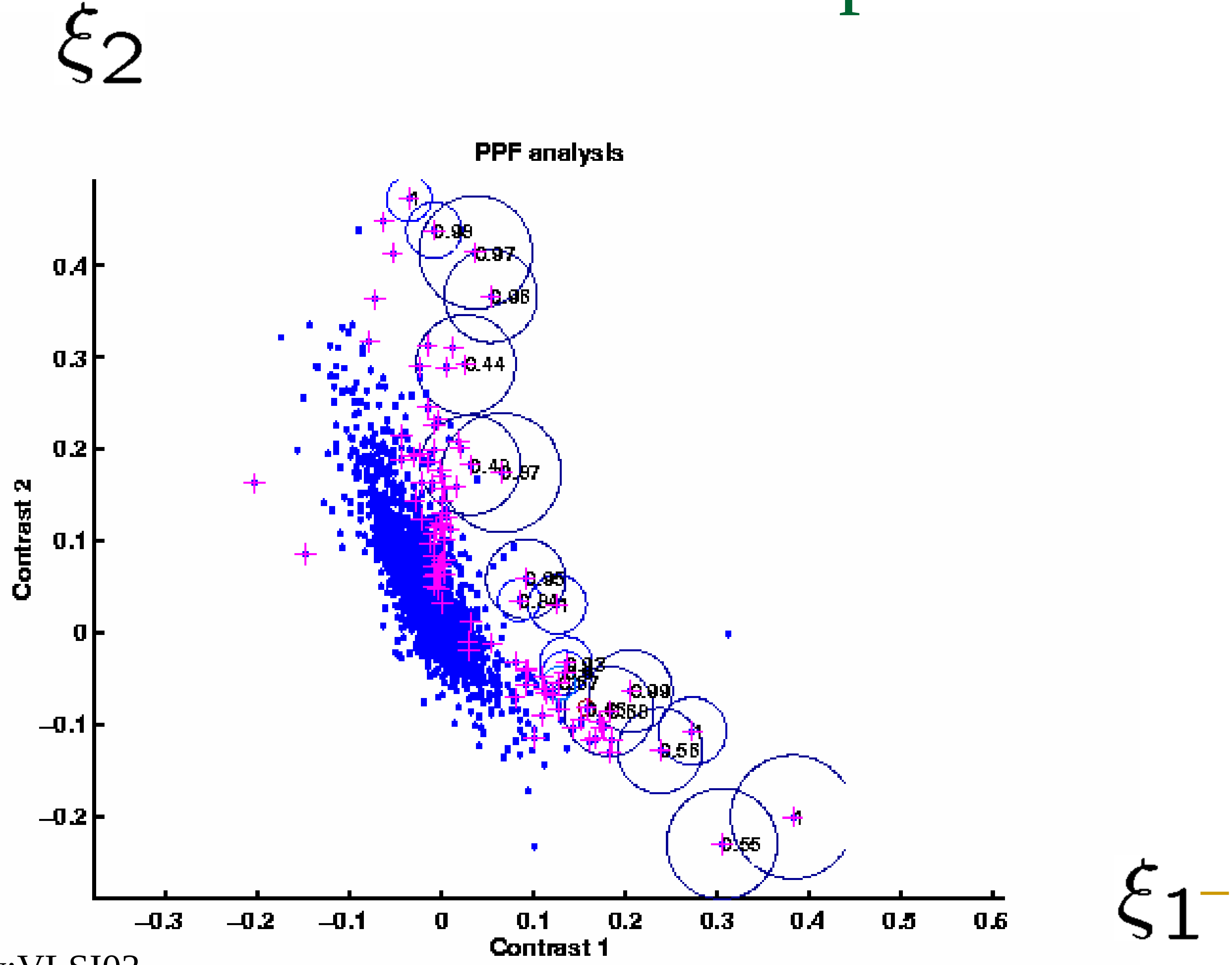
# Accounting for Sampling Errors in PFA

- Key Concepts:
  - Pareto Depth Posterior Distribution: Hero&Fleury:VLSI04
  - Pareto Depth Sampling Distribution: Fleury&etal:ISBI04, Fleury&etal:JFI03

- Bayesian perspective: Pareto Depth Posterior Distn
  - Introduce priors into multicriterion scattergram
  - Compute posterior probability that gene lies on a Pareto front
  - Rank order genes by PDPD posterior probabilities

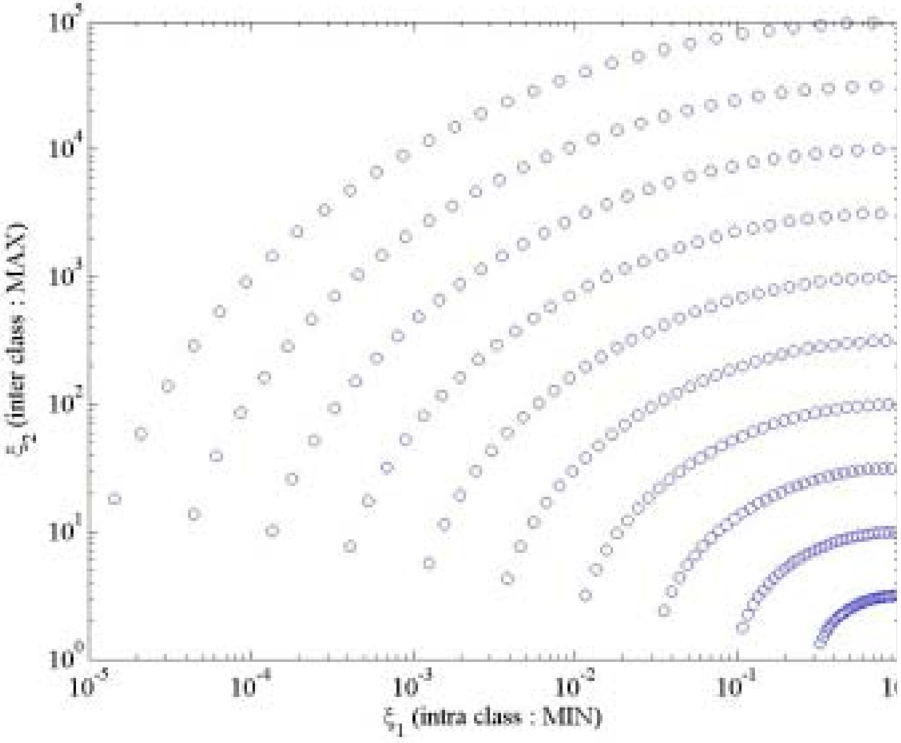
- Frequentist perspective: Pareto Depth Sampling Distn
  - Generate subsamples of replicates by resampling
  - Compute relative frequency that subsamples of a gene remain on a Pareto front
  - Rank order genes by PDSD relative frequencies

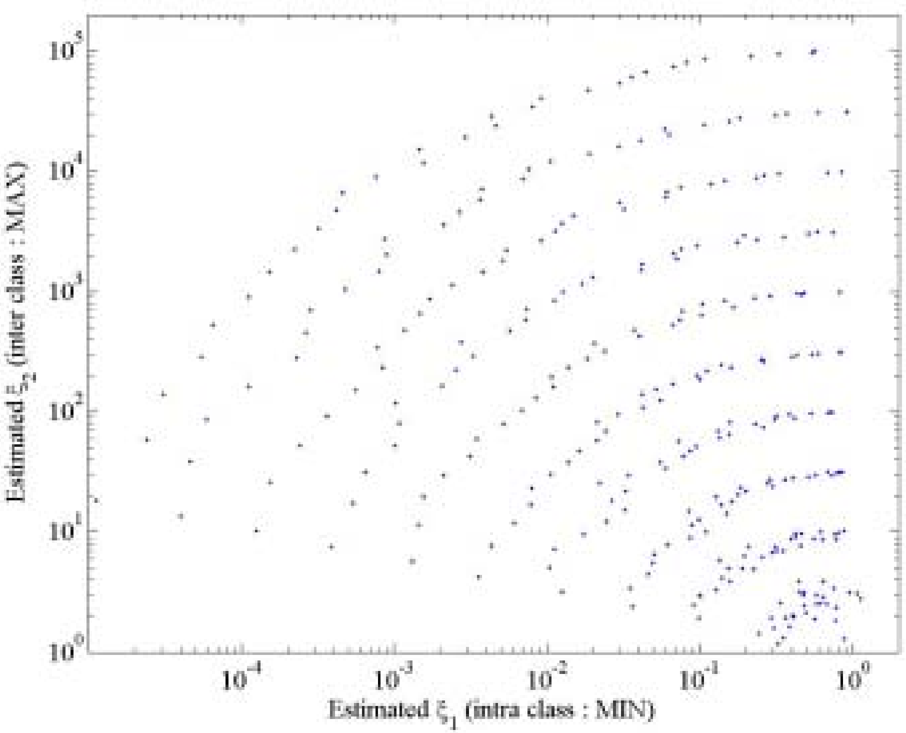
# Scattergram for Dilution Experiment



# Simulation Comparison: PT vs PDSD

## Hypothetical dual criterion planes



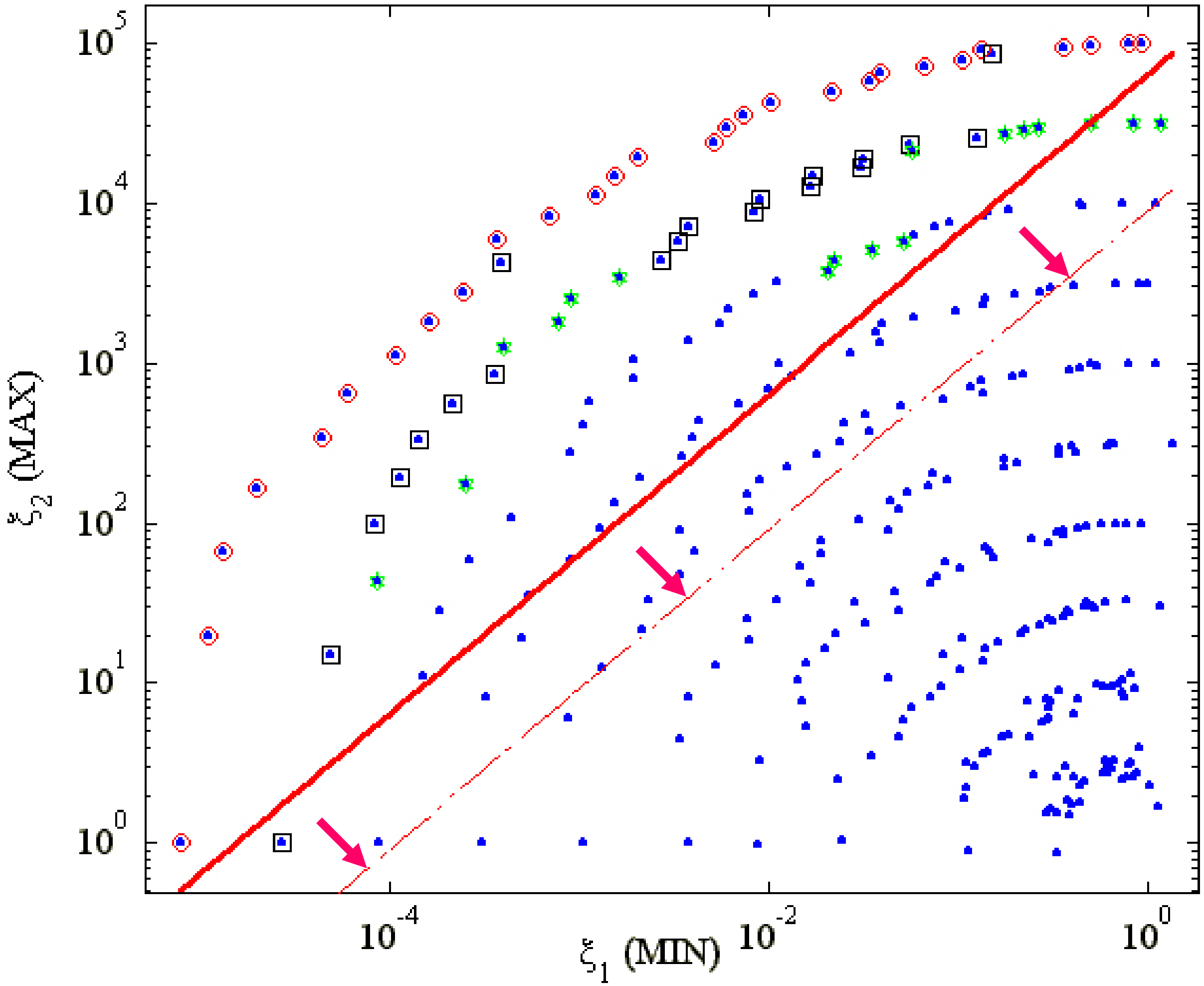


**Ensemble mean scattergram (Ground truth)**

**Sample mean scattergram (Measured)**

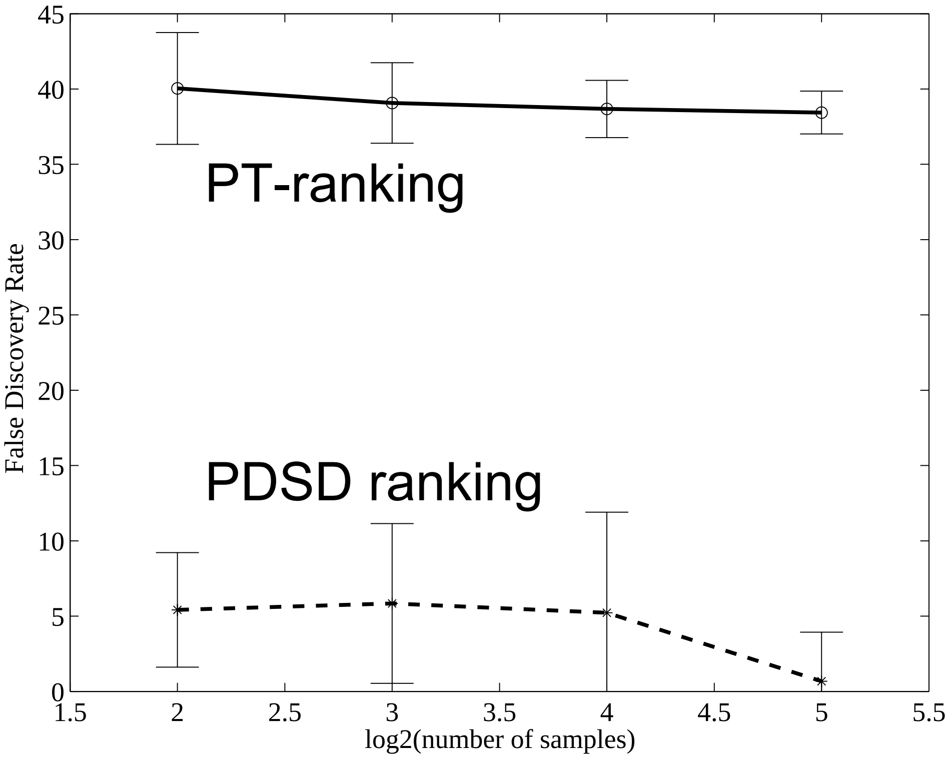
Ref: Fleury and Hero:JFI03

# Pareto Front vs. Paired T Test ranking

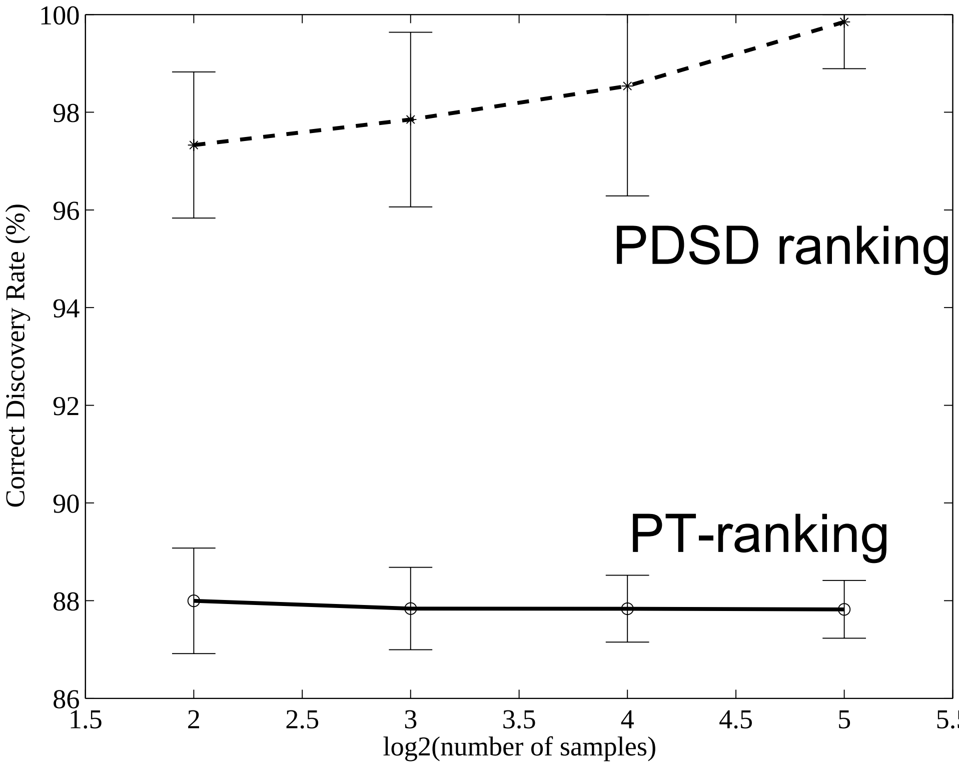


Ref: Fleury and Hero:JFI03

# False Discovery Rate Comparisons



**False Discovery Rate**

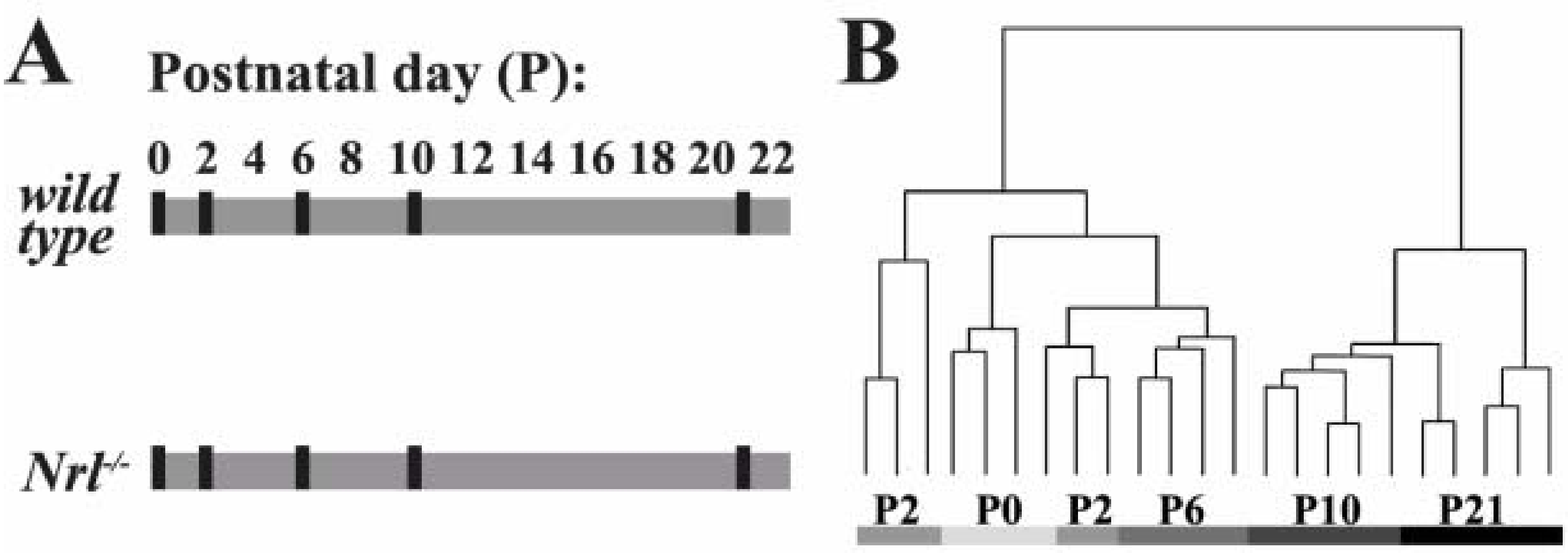


**Correct Discovery Rate**

Ref: Fleury and Hero:JFI03

# Clustering differential gene profiles

- Clustering Case Study: cDNA Microarray
  - Two treatments: Wildtype mice vs Nrl Knockout mice
  - 6 time points for each treatment
  - 4-5 replicates for each time point
  - Gene filtering via FDR produced 923 differentially expressed gene trajectories for cluster analysis



Ref: JindanYu, PhD Thesis, BME Dept, Univ of Michigan, 2004.

# Wt/ko Clustering Approach

- Objective: To find clusters of wt/ko profile differences
- Step 1: Encode each gene into a feature vector

X(g)=[wt0,wt2,wt6,wt10,wt21,ko0,ko2,ko6,ko10,ko21]

- Step 2: Cluster the rows of the 923x12 matrix

$$\mathbf{X} = [X'(1), \ldots, X'(923)]'$$

- Three clustering techniques:
  - hierarchical,
  - k-means,
  - unsupervised clustering by learning mixtures

# Clustering via PML Learning of Mixtures

- Hidden data model for class membership $Z_g(c) \in \{0,1\}$

$$X_g = \sum_{c=1}^{C} Z_g(c) S_g(c)$$

- Penalized maximum likelihood (PML) function

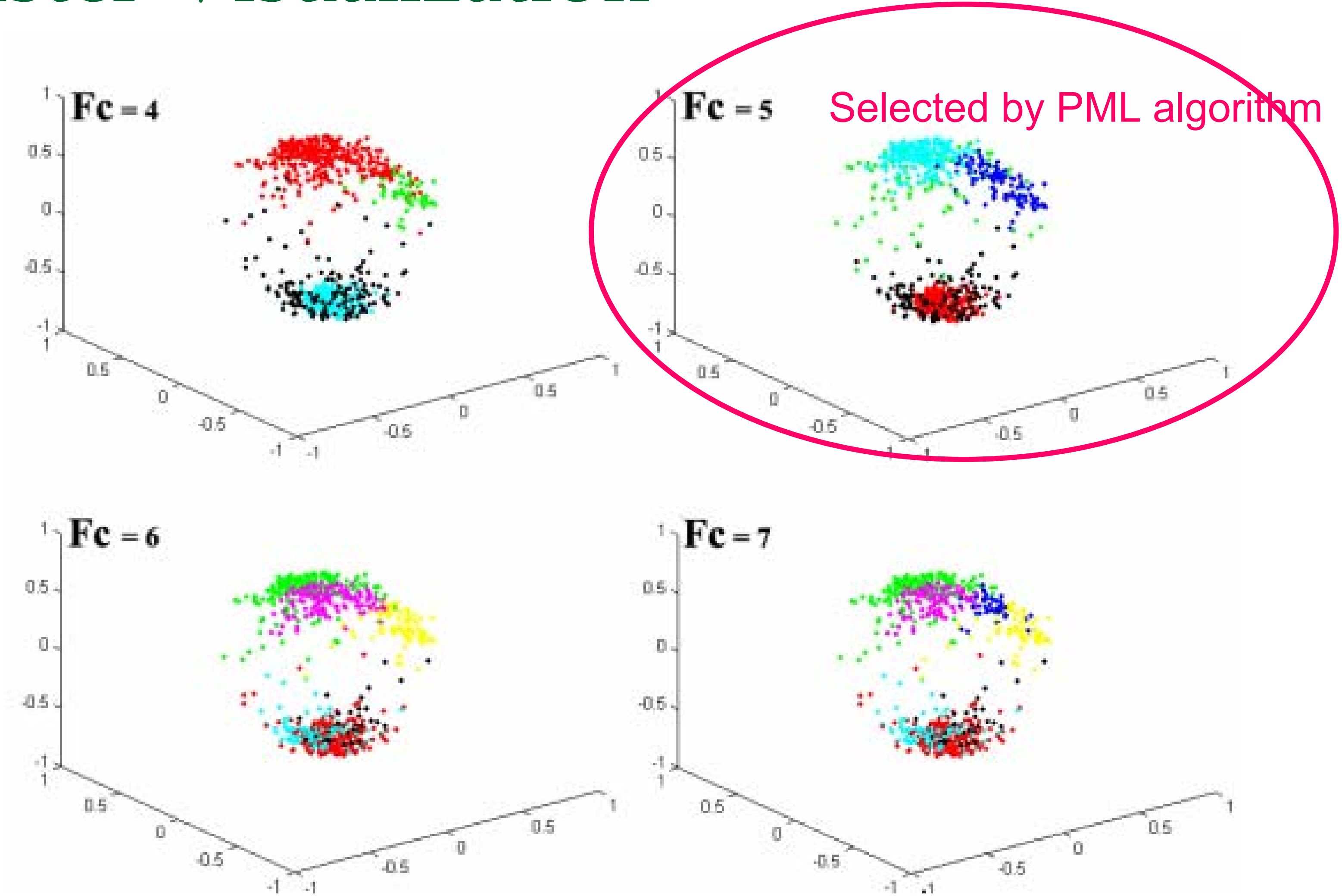
$$L(\theta, \alpha, C) = \sum_{g=1}^{G} \sum_{c=1}^{C} \alpha(c) \phi_c(X_g; \theta_c) + Q(C)$$

- Maximization of PML via EM algorithm produces
  - ❑ An estimated number C of clusters
  - ❑ A “Soft”classification to class c of each gene g

$$P(Z_g(c) = 1 | X)$$

Ref: Figuieredo&Jain:PAMI2001

# Cluster Visualization

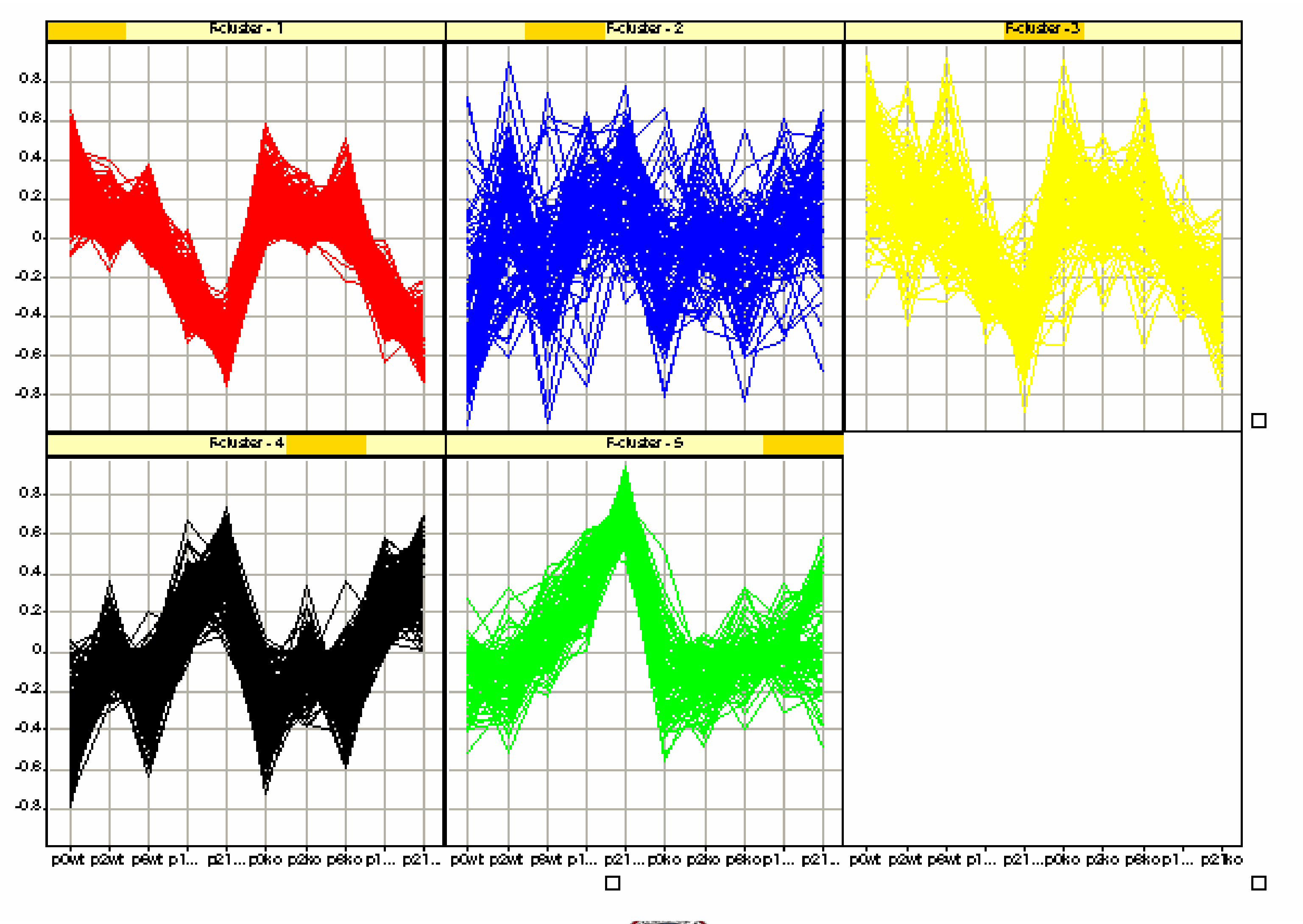


Selected by PML algorithm

**Result of PML mixture clustering of 800 genes (MDS projections onto 3D)**

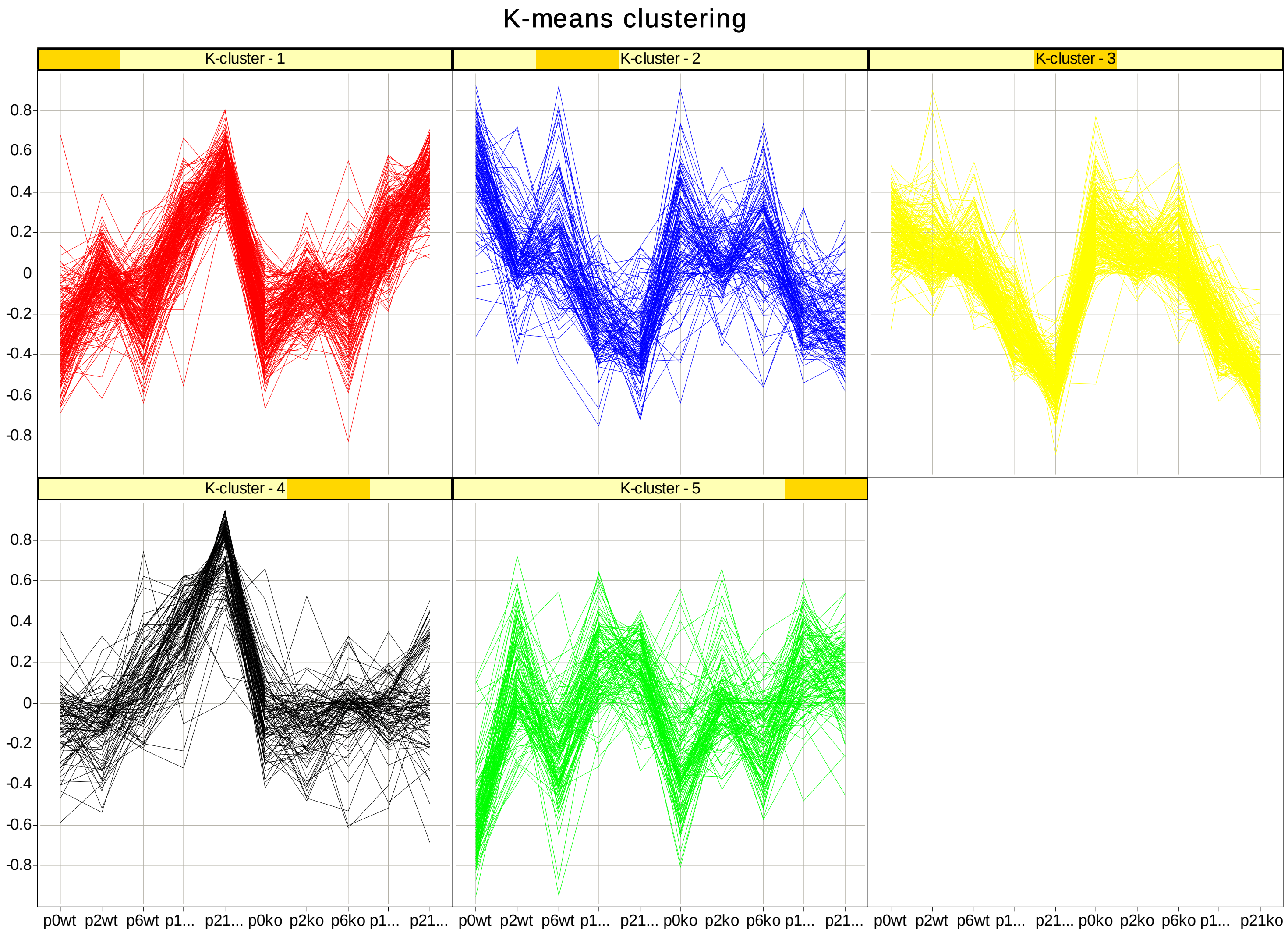
JindanYu, PhD Thesis, BME Dept, Univ of Michigan, 2004.

# Clustered Trajectories: PML Mixture



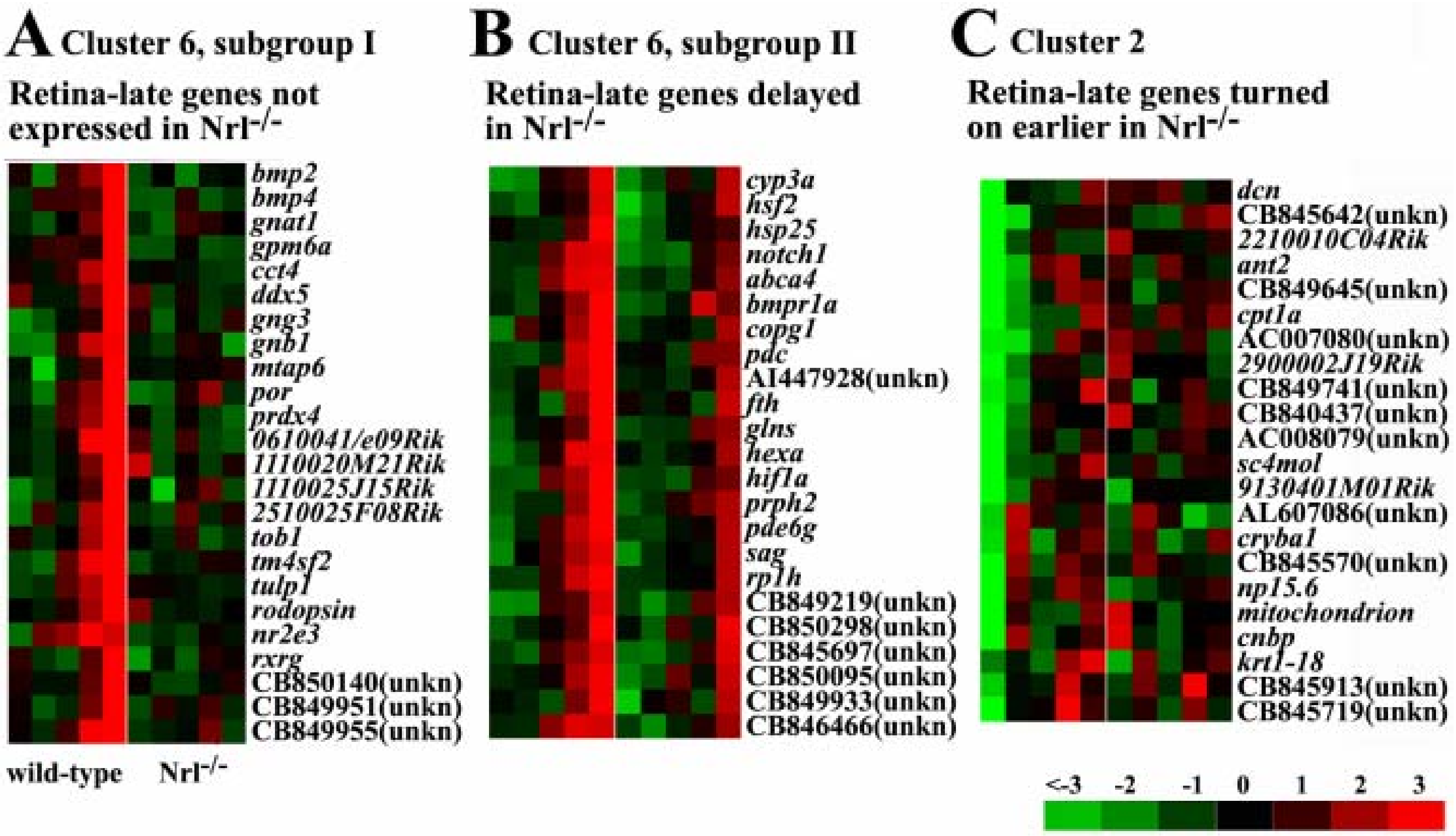
JindanYu, PhD Thesis, BME Dept, Univ of Michigan, 2004.

# Clustered Trajectories: k-Means



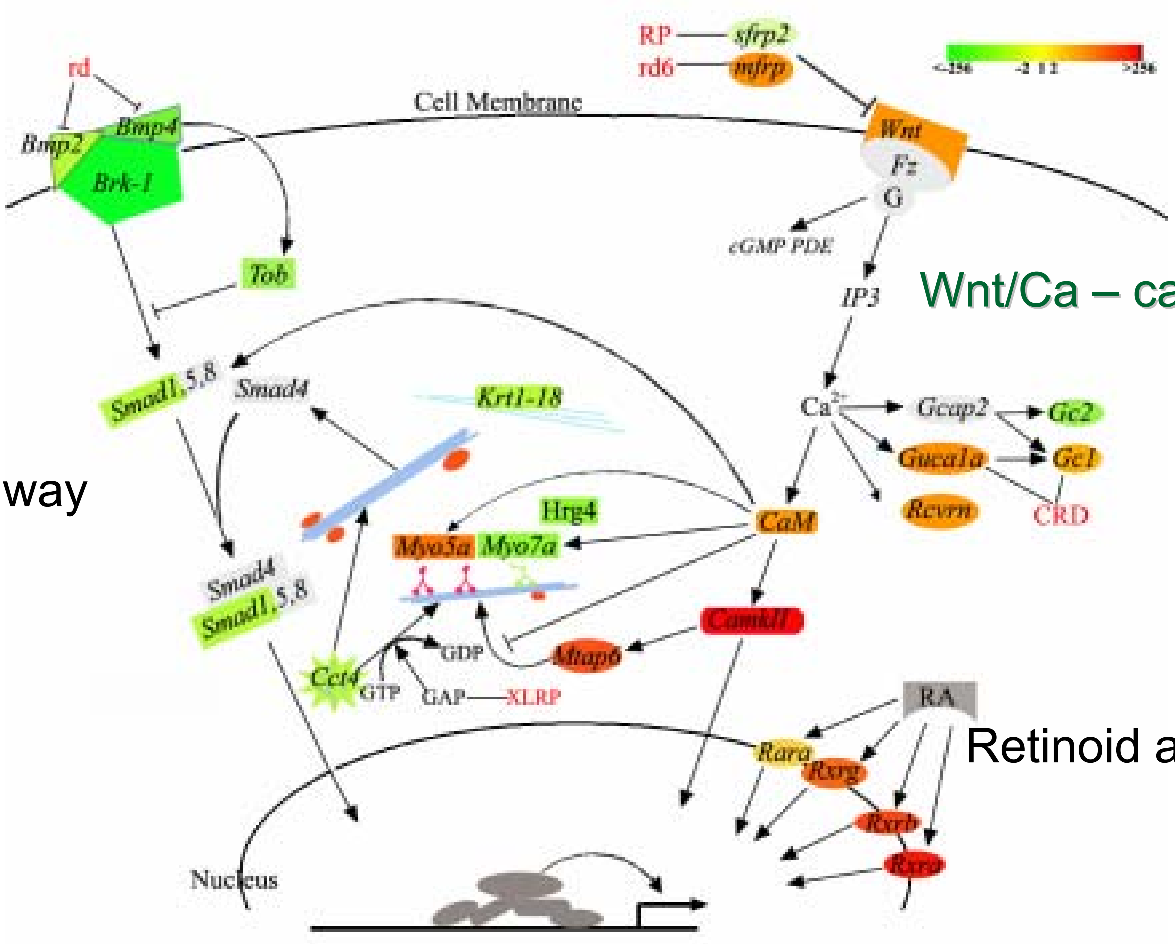
JindanYu, PhD Thesis, BME Dept, Univ of Michigan, 2004.

# Post-Clustering Time Course Analysis



JindanYu, PhD Thesis, BME Dept, Univ of Michigan, 2004.

# 4. Discovering gene regulation networks



Wnt/Ca – calmodulin pathway

Bmp pathway

Retinoid acid pathway

Draft Pathways for Photoreceptor Function

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

# Basic co-Expression Search Tools (BEST)

- Correlation measures
  - Pearson’s correlation coefficient (linear similarity)
  - Kendall’s rank correlation (non-linear similarity)
  - α-Mutual information (non-linear similarity)
- Types of correlation estimators
  - Sample covariance matrix
  - Sample partial correlation matrix
  - Resampling methods: Jackknife, Bootstrap, SIR
- Objective: Find gene dependency network from pairwise correlations between profiles
  - Relevancy network: partial ordering of correlations: $\rho(g_i, g_j)$
  - Graphical Gaussian Model: partial ordering of pairwise partial correlations $\rho(g_i, g_j | G_{-i,-j})$

Dongxiao Zhu, A. Hero, S. Qi, In preparation, Univ of Michigan, 2004.

# Two-stage pairwise correlation screening algorithm

- Statistical hypothesis for each co-expression candidate:

$$H_o : |r_0| \leq \text{cormin}$$

$$H_\alpha : |r_0| > \text{cormin}$$

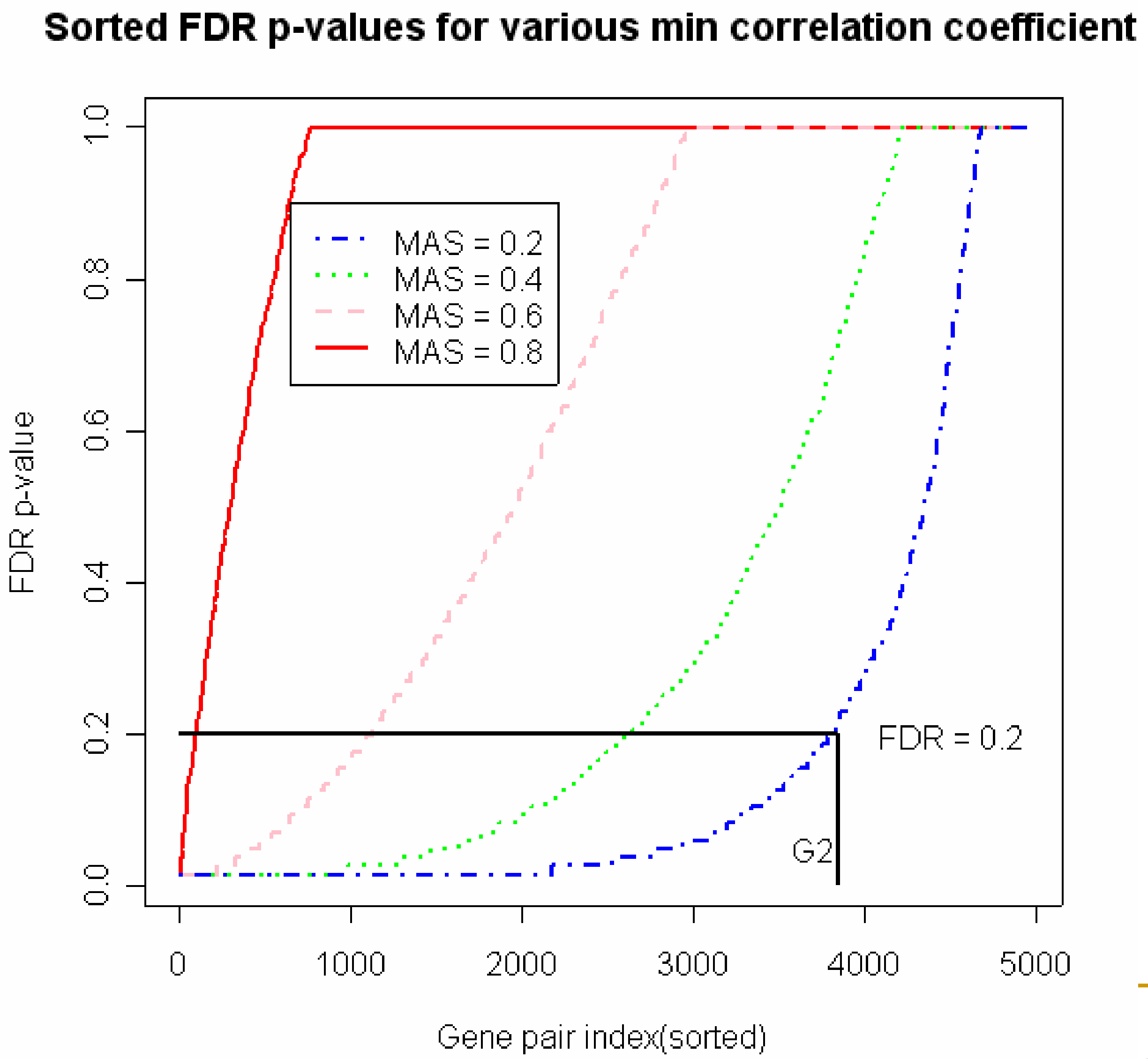
- Two-stage screen algorithm (Hero&etal:JASP 2004)
  - Stage I, controls only FDR
  - Stage II,controls both FDR and Minimum Acceptable Strength (MAS)
- Algorithm controls significance at a FDR level $\alpha$ and at a MAS level *cormin*

# Galactose metabolism experiment

- Global gene expression profiles in 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.
- Galactose metabolic pathway, “all-or-nothing”.
- Two-channel cDNA array, 5935 gene expression profiles are measured. Reference channel is dilution “wild-type + galactose”
- Missing data imputation: k-nearest neighbor (k = 12, Troyanskaya et al, 2001)
- Gene filtering eliminates expression profiles whose minimal foldchange variation <2

Dongxiao Zhu, A. Hero, S. Qi, In preparation, Univ of Michigan, 2004.

# Result of two-stage screening

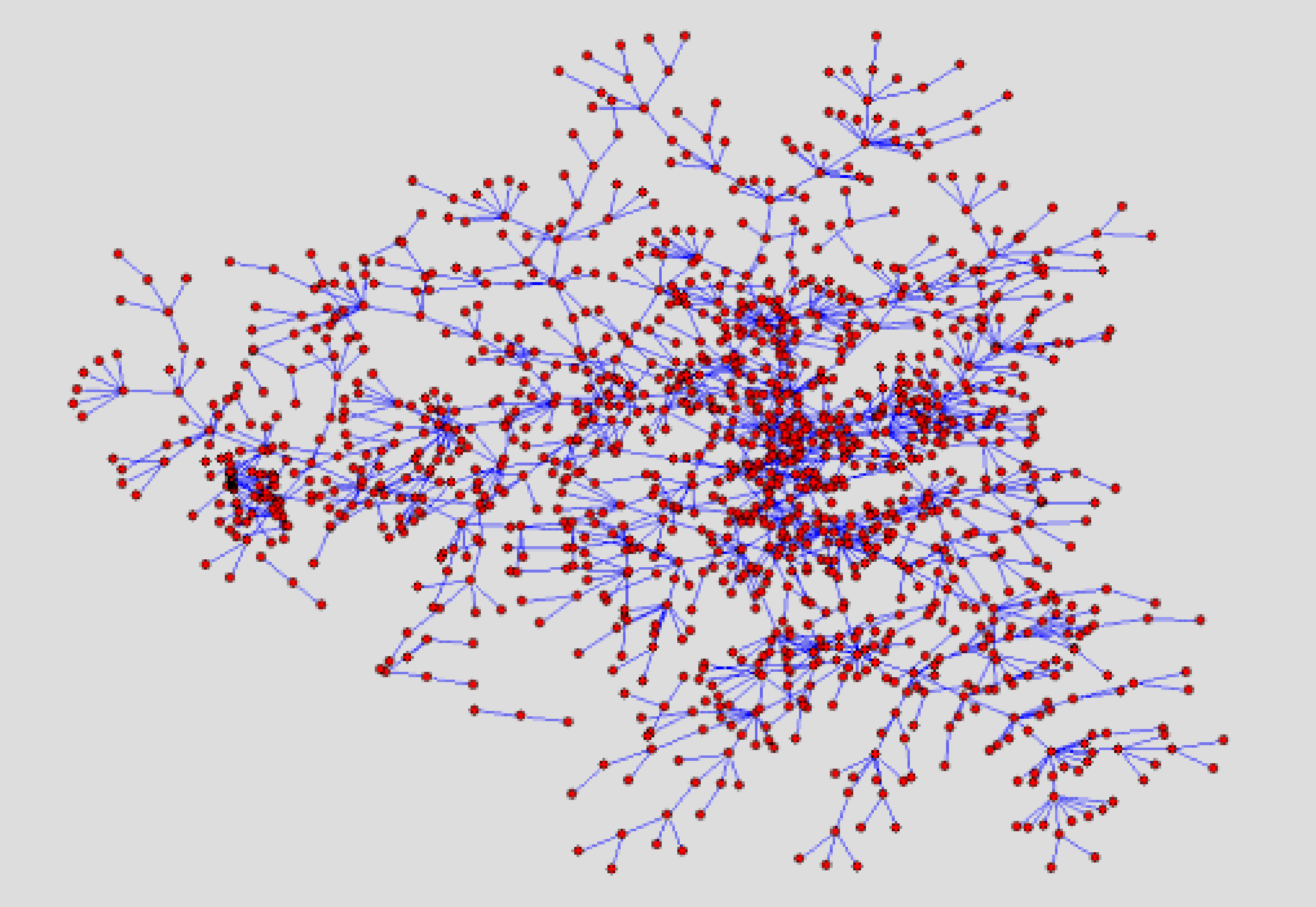


a 2004

Dongxiao Zhu, A. Hero, S. Qi, In preparation, Univ of Michigan, 2004.

# Relevance network visualization

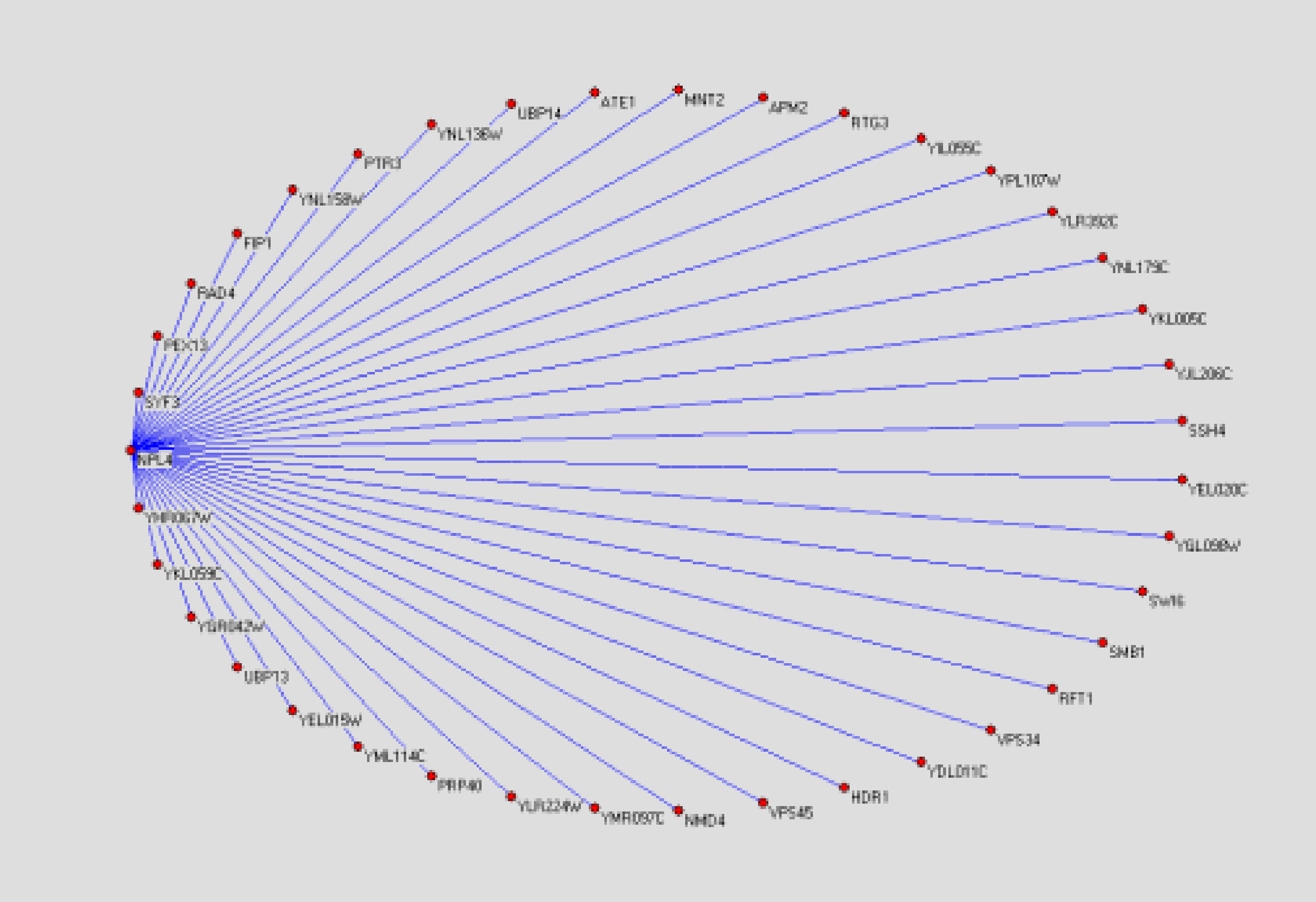
## (FDR <= 0.05, MAS = 0.7)



Dongxiao Zhu, A. Hero, S. Qi, In preparation, Univ of Michigan, 2004.

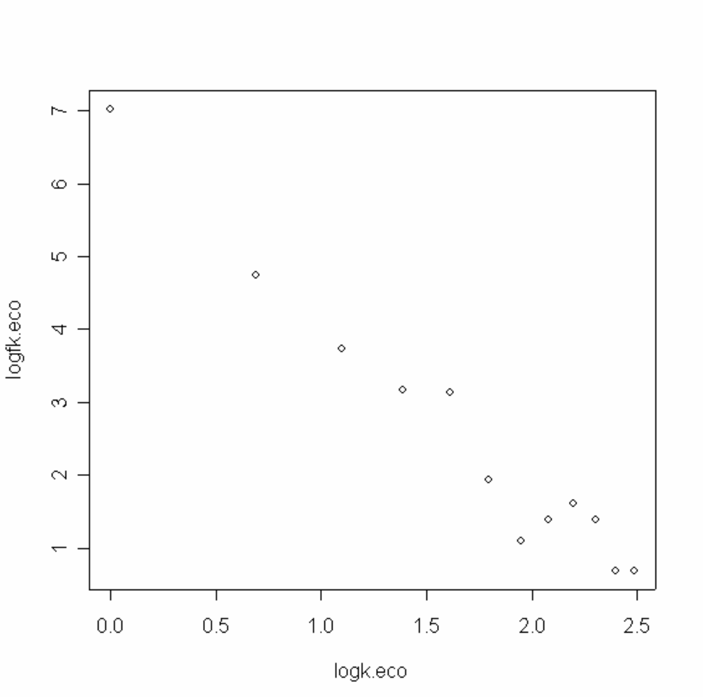
# Hub Gene “NPL4”

## (FDR <= 0.05, MAS = 0.7)

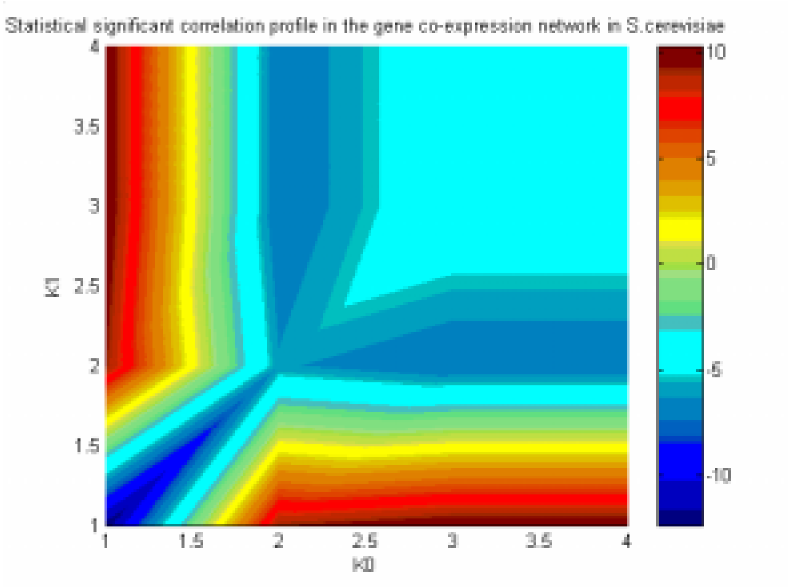


Dongxiao Zhu, A. Hero, S. Qi, In preparation, Univ of Michigan, 2004.

# Degree distribution of relevance network



Log-transformed marginal degree dsitribution



Bivariate joint degree distribution

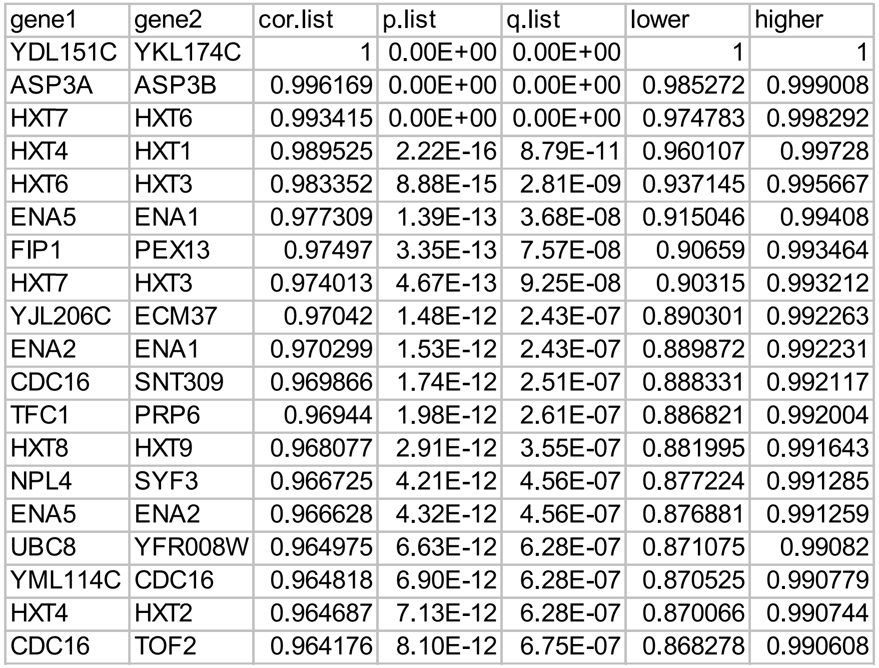
Dongxiao Zhu, A. Hero, S. Qi, In preparation, Univ of Michigan, 2004.

# Top ten "Hub Genes"

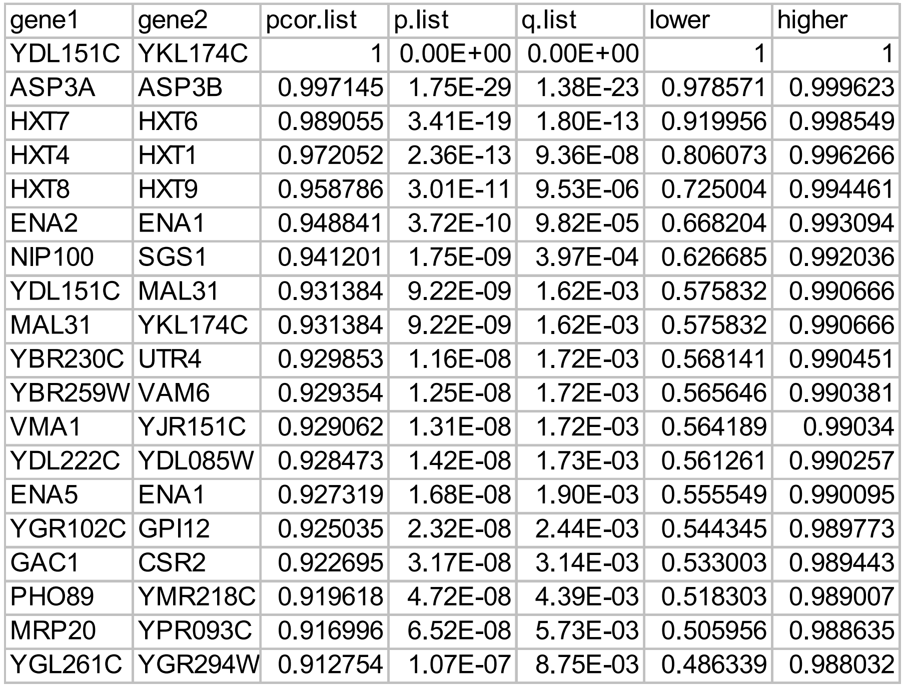
| Rank | Name | Degree | Function |
|---|---|---|---|
| 1 | NPL4 | 24 | Endoplasmic reticulum and nuclear membrane protein, forms a complex with Cdc48p and Ufd1p that recognizes ubiquitinated proteins in the endoplasmic reticulum and delivers them to the proteasome for degradation |
| 2 | **YPL107W** | 21 | Hypothetical ORF |
| 3 | CDC16 | 20 | Subunit of the anaphase-promoting complex/cyclosome (APC/C), which is a ubiquitin-protein ligase required for degradation of anaphase inhibitors, including mitotic cyclins, during the metaphase/anaphase transition; required for sporulation |
| 4 | **YEL020C** | 19 | Hypothetical ORF |
| 5 | CDC50 | 19 | Endosomal protein that regulates cell polarity; similar to Ynr048wp and Lem3p |
| 6 | SSH4 | 18 | Suppressor of SHR3; confers leflunomide resistance when overexpressed |
| 7 | **YML114C** | 17 | Hypothetical ORF |
| 8 | NBP2 | 17 | interacts with Nap1, which is involved in histone assembly |
| 9 | MTR2 | 17 | mRNA transport regulator |
| 10 | FIP1 | 15 | Subunit of cleavage polyadenylation factor (CPF), interacts directly with poly(A) polymerase (Pap1p) to regulate its activity |

Dongxiao Zhu, A. Hero, S. Qi, In preparation, Univ of Michigan, 2004.

# Comparison of co-expressed gene pairs



| gene1 | gene2 | cor.list | p.list | q.list | lower | higher |
|---|---|---|---|---|---|---|
| YDL151C | YKL174C | 1 | 0.00E+00 | 0.00E+00 | 1 | 1 |
| ASP3A | ASP3B | 0.996169 | 0.00E+00 | 0.00E+00 | 0.985272 | 0.999008 |
| HXT7 | HXT6 | 0.993415 | 0.00E+00 | 0.00E+00 | 0.974783 | 0.998292 |
| HXT4 | HXT1 | 0.989525 | 2.22E-16 | 8.79E-11 | 0.960107 | 0.99728 |
| HXT6 | HXT3 | 0.983352 | 8.88E-15 | 2.81E-09 | 0.937145 | 0.995667 |
| ENA5 | ENA1 | 0.977309 | 1.39E-13 | 3.68E-08 | 0.915046 | 0.99408 |
| FIP1 | PEX13 | 0.97497 | 3.35E-13 | 7.57E-08 | 0.90659 | 0.993464 |
| HXT7 | HXT3 | 0.974013 | 4.67E-13 | 9.25E-08 | 0.90315 | 0.993212 |
| YJL206C | ECM37 | 0.97042 | 1.48E-12 | 2.43E-07 | 0.890301 | 0.992263 |
| ENA2 | ENA1 | 0.970299 | 1.53E-12 | 2.43E-07 | 0.889872 | 0.992231 |
| CDC16 | SNT309 | 0.969866 | 1.74E-12 | 2.51E-07 | 0.888331 | 0.992117 |
| TFC1 | PRP6 | 0.96944 | 1.98E-12 | 2.61E-07 | 0.886821 | 0.992004 |
| HXT8 | HXT9 | 0.968077 | 2.91E-12 | 3.55E-07 | 0.881995 | 0.991643 |
| NPL4 | SYF3 | 0.966725 | 4.21E-12 | 4.56E-07 | 0.877224 | 0.991285 |
| ENA5 | ENA2 | 0.966628 | 4.32E-12 | 4.56E-07 | 0.876881 | 0.991259 |
| UBC8 | YFR008W | 0.964975 | 6.63E-12 | 6.28E-07 | 0.871075 | 0.99082 |
| YML114C | CDC16 | 0.964818 | 6.90E-12 | 6.28E-07 | 0.870525 | 0.990779 |
| HXT4 | HXT2 | 0.964687 | 7.13E-12 | 6.28E-07 | 0.870066 | 0.990744 |
| CDC16 | TOF2 | 0.964176 | 8.10E-12 | 6.75E-07 | 0.868278 | 0.990608 |



| gene1 | gene2 | pcor.list | p.list | q.list | lower | higher |
|---|---|---|---|---|---|---|
| YDL151C | YKL174C | 1 | 0.00E+00 | 0.00E+00 | 1 | 1 |
| ASP3A | ASP3B | 0.997145 | 1.75E-29 | 1.38E-23 | 0.978571 | 0.999623 |
| HXT7 | HXT6 | 0.989055 | 3.41E-19 | 1.80E-13 | 0.919956 | 0.998549 |
| HXT4 | HXT1 | 0.972052 | 2.36E-13 | 9.36E-08 | 0.806073 | 0.996266 |
| HXT8 | HXT9 | 0.958786 | 3.01E-11 | 9.53E-06 | 0.725004 | 0.994461 |
| ENA2 | ENA1 | 0.948841 | 3.72E-10 | 9.82E-05 | 0.668204 | 0.993094 |
| NIP100 | SGS1 | 0.941201 | 1.75E-09 | 3.97E-04 | 0.626685 | 0.992036 |
| YDL151C | MAL31 | 0.931384 | 9.22E-09 | 1.62E-03 | 0.575832 | 0.990666 |
| MAL31 | YKL174C | 0.931384 | 9.22E-09 | 1.62E-03 | 0.575832 | 0.990666 |
| YBR230C | UTR4 | 0.929853 | 1.16E-08 | 1.72E-03 | 0.568141 | 0.990451 |
| YBR259W | VAM6 | 0.929354 | 1.25E-08 | 1.72E-03 | 0.565646 | 0.990381 |
| VMA1 | YJR151C | 0.929062 | 1.31E-08 | 1.72E-03 | 0.564189 | 0.99034 |
| YDL222C | YDL085W | 0.928473 | 1.42E-08 | 1.73E-03 | 0.561261 | 0.990257 |
| ENA5 | ENA1 | 0.927319 | 1.68E-08 | 1.90E-03 | 0.555549 | 0.990095 |
| YGR102C | GPI12 | 0.925035 | 2.32E-08 | 2.44E-03 | 0.544345 | 0.989773 |
| GAC1 | CSR2 | 0.922695 | 3.17E-08 | 3.14E-03 | 0.533003 | 0.989443 |
| PHO89 | YMR218C | 0.919618 | 4.72E-08 | 4.39E-03 | 0.518303 | 0.989007 |
| MRP20 | YPR093C | 0.916996 | 6.52E-08 | 5.73E-03 | 0.505956 | 0.988635 |
| YGL261C | YGR294W | 0.912754 | 1.07E-07 | 8.75E-03 | 0.486339 | 0.988032 |

Simple correlation<br>(Relevance Network)

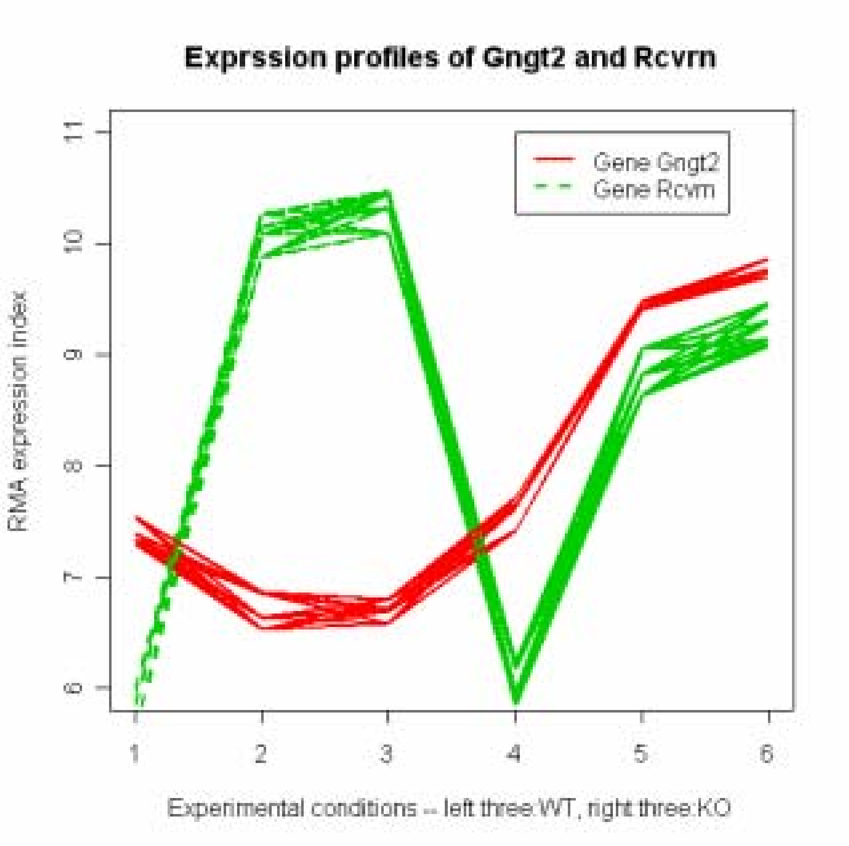
Partial correlation<br>(Graphic Gaussian Model)

Dongxiao Zhu, A. Hero, S. Qi, In preparation, Univ of Michigan, 2004.

# α-Mutual Information (Non-linearly co-expressed genes)

## (Red is up-regulated and green is down-regulated by Nrl)

| Gene1 | Gnen2 | corlist | a-MI |
|---|---|---|---|
| 160893_at | 160893_at | 1 | 1 |
| 160893_at | 100453_at | 0.771879 | 0.81483 |
| 160893_at | 160693_at | -0.42367 | 0.81088 |
| 160893_at | 102340_at | 0.738077 | 0.8036 |
| 160893_at | 160204_at | 0.12689 | 0.79348 |
| 160893_at | 94256_at | 0.242293 | 0.78675 |
| 160893_at | 93071_at | 0.049194 | 0.78327 |
| 160893_at | 97925_at | 0.02524 | 0.78173 |
| 160893_at | 96490_at | -0.53337 | 0.77259 |
| 160893_at | 101344_at | 0.82691 | 0.77083 |
| 160893_at | 98569_at | -0.28032 | 0.76988 |
| 160893_at | 98532_at | 0.093833 | 0.76495 |
| 160893_at | 160131_at | -0.28248 | 0.75963 |
| 160893_at | 98427_s_a | 0.931593 | 0.75921 |
| 160893_at | 102682_at | 0.399634 | 0.75797 |
| 160893_at | 160242_at | 0.005767 | 0.75782 |
| 160893_at | 96951_at | 0.449107 | 0.75412 |
| 160893_at | 95356_at | 0.611431 | 0.75395 |
| 160893_at | 97125_f_a | 0.445086 | 0.75371 |
| 160893_at | 97540_f_a | 0.48131 | 0.75358 |
| 160893_at | 99160_s_a | 0.301906 | 0.75236 |
| 160893_at | 98560_at | 0.978286 | 0.7493 |
| 160893_at | 93412_at | 0.743385 | 0.74621 |
| 160893_at | 102354_at | -0.09397 | 0.74365 |
| 160893_at | 93390_g_a | -0.04565 | 0.74253 |
| 160893_at | 93120_f_a | 0.480893 | 0.74087 |
| 160893_at | 104104_at | 0.705927 | 0.74051 |
| 160893_at | 96072_at | -0.10414 | 0.73879 |
| 160893_at | 104643_at | 0.981046 | 0.73793 |



MI: 0.71915   Corrcoef: -0.01989

Dongxiao Zhu, A. Hero, S. Qi, In preparation, Univ of Michigan, 2004.

# 5. Wrap Up and References

- Gene filtering: accounting for biological and statistical significance
- Gene ranking: can involve optimization over multiple criteria
- Gene clustering: group response profiles under single or multiple treatments
- Gene co-regulation networks: discover co-dependent gene profiles
- Increasing importance of statistical signal and image processing approaches
- References to UM work and software presented here: http://www.eecs.umich.edu/~hero/bioinfo.html

# Gene Microarray Software Resources

- Affymetrix software
  - http://www.affymetrix.com/products/software/index.affx
- 3rd party Affymetrix analysis software
  - http://www.affymetrix.com/support/developer/tools/genechip_compatible_software.affx
- Bioconductor, RMA, SMA software
  - http://stat-www.berkeley.edu/users/terry/Group/software.html
- R software
  - http://www.r-project.org/
- Matlab – see bioinformatics toolbox
  - http://www.mathworks.com/
- S-Plus software
  - http://www.insightful.com/products/default.asp
- dChip
  - http://www.dchip.gov

# General References

- A. Berry and J.D. Watson, DNA : The Secret of Life
  Knopf, 2003.
- C. Causton, J. Quackenbush, A. Brazma, Microarray Gene Expression Data Analysis: A Beginner's Guide, Blackwell Publishers, 2003
- S. Draghici, Data Analysis Tools for DNA Microarrays, Chapman&Hall, 2003
- ES. Garrett et al.(ed), The Analysis of Gene Expression Data: Methods and Software, Springer, New York, 2003
- Hollander&Wolfe, “Nonparametric statistical methods,” Wiley, 1999.
- Hastie, Tibshirani, Friedman, “The elements of statistical learning, Springer 2001
- T. Speed (ed), Statistical analysis of gene expression data, Chapman&Hall/CRC, 2003

# References on Microarray Image Analysis

- C. S. Brown., P. Goodwin, and P. Sorger. (2001) Image metrics in the statistical analysis of DNA microarray data. *P.N.A.S*, **98**(16):8944–8949
- Yang YH, Buckley MJ, Speed, TP (2001) Analysis of cDNA microarray images. *Brief Bioinform* **2**(4) 341-349.
- Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*,**11**: (1) 108-136
- Y. Chen, E. R. Dougherty, and M. L. Bittner.(1997) Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. *J. Biomedical Optics*, **2**(4):364–374
- M. Katzer, F. Kummert, and G. Sagerer. (2002) Robust Automatic Microarray Image Analysis. In *Proceedings of the International Conference on Bioinformatics:North-South Networking*, Bangkok.
- K.I. Siddiqui, A. Hero, and M. Siddiqui, "Mathematical Morphology applied to Spot Segmentation and Quantification of Gene Microarray Images," 2002 Asilomar Conference on Signals and Systems, Nov. 2002.
- G.C. Tseng, M.-K. Oh, L. Rohlin, J.C. Liao, and W.H. Wong. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*. **29**: 2549-2557

# References on Normalization

- Li C and Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci.*, **98**, 31-36
- Cope LM, Irizarry, RA, Jaffee HA, Wu Z, and Speed TP (2004) A benchmark for Affymetrix geneChip Expression Measures. *Bioinformatics* in press
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249-264
- Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**(4) e15.
- Bolstad BM, Irizarry, RA, Astrand A, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19** 185-193
- Y.H.Yang and N. Thorne (2003) Normalization for Two-color cDNA Microarray Data. Science and Statistics: A Festschrift for Terry Speed, D. Goldstein (eds.), IMS Lecture Notes, Monograph Series, Vol 40, pp. 403--418.

# References on Significance Analysis

- A. Hero, G. Fleury, A. Mears and A. Swaroop, "Multicriteria Gene Screening for Analysis of Differential Expression with DNA Microarrays, *JASP*, vol. 2004, No. 1, pp. 43-52, 2004.
- W. J. Lemon, J. T. Palatini, R. Krahe, and F. A. Wright, Theoretical and experimental comparison of gene expression estimators for oligonucleotide arrays," *Bioinformatics*, 2002.
- D. Reiner, A. Yekutieli and Y. Benjamini, ``Identifying differentially expressed genes using false discovery rate controlling procedures," *Bioinformatics*, vol. 19, no. 3, pp. 368-375, 2003.
- JD. Storey and R Tibshirani. Statistical significance for genomewide studies. *P.N.A.S*, 100: (16), 9440-9445
- JD. Storey et al. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc.* B (2004) **66**, *Part* 1, *pp.* 187–205
- Tusher, Tibshirani and Chu (2001): "Significance analysis of microarrays applied to the ionizing radiation response" *P.N.A.S* 2001 98: 5116-5121, (Apr 24). (SAM software source paper)
- S. Yoshida, A. Mears, J.S. Friedman, T. Carter, S. he, E. Oh, Y. Jing, R. Farjo, G. Fleury, C. Barlow, A. Hero, A. Swaroop, "Expression profiling of of the developing and mature NRL-/- mouse retina: Identification of retinal disease candidates and transcriptional regulatory targets of NRL," Human Molecular Genetics, vol/ 13, no. 14, pp. 1497-1503, 2004.

# References on analysis of time course data

- Zareparsi,S., Hero,A.O., Zack,D.J., Williams,R. and Swaroop,A. “Seeing the unseen: Microarray-based gene expression profiling in vision,” *Invest Ophthalmol Vis Sci.*, **45**, 2457-2462, 2004.
- Spellman *et al*., (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* 9, 3273-3297
- Cho RJ, Huang M, Campbell MJ, Dong H, Steinmetz L, Sapinoso L, Hampton G, Elledge SJ, Davis RW, Lockhart DJ. (2001) Transcriptional regulation and function during the human cell cycle. *Nat Genet.* **27** 48-54
- Shedden K and Cooper S (2002) Analysis of cell-cycle gene expression in Saccharomyces cerevisiae using microarrays and multiple synchronization methods. *Nucleic Acids Res.* **30** 2920-2929.
- Lu X, Zhang W, Qin ZS, Kwast KE, Liu JS. (2004) Statistical resynchronization and Bayesian detection of periodically expressed genes. *Nucleic Acids Res.* **32** 447-455.
- Wen, X. et al. Large-scale temporal gene expression mapping of central nervous system development, *P.A.N.S.*, **95**:334-339,1998
- Saban, M.R. et al. Time course of lps-induced gene expression in a mouse model of genitourinary inflammation. *Physiol. Genomics*, **5**:147-160, 2001
- Langmead, C.J. et al. Phase-independent rhythmic analysis of genome-wide expression patterns, in *Proc. Sixth Annu. Int. on Computational Molecular Biol*., Washington, D.C., 2002

# References on Pareto and clustering

- G. Fleury , A. Hero , S. Zareparsi and A. Swaroop, Gene discovery using Pareto depth sampling distributions, *Journal of the Franklin Institute*, Volume 341, Issues 1-2, pp. 55-75, 2004.
- McLachlan,G., Bean,R. and Peel,D., “A mixture model based approach to the clustering of microarray expression data,” *Bioinformatics*, **18**, 413-422, 2002.
- T. Hastie and R. Tibshirani, “Discriminant analysis by Gaussian mixtures,” J. Royal Stat. Soc. Ser. B, Volume 58, pp. 155-176, 1996.
- A. Hero and G. Fleury, "Pareto-optimal methods for gene analysis" to appear Special Issue on Genomic Signal Processing, *Journ. of VLSI Signal Processing*, 2004.
- R.E. Steuer, Multi criteria optimization: theory, computation, and application, Wiely, New York, 1986
- Tamayo, P. et al. Interpreting patterns of gene expression with self-organization maps: methods and application to hematopoietic differentiation. *P.N.A.S.*, **96**:2907-2912, 1999
- E.Zitler and L.Thiele, “An evolutionary algorithm for multi-objective optimization: the strength Pareto approach”, Technical report, Swiss Federal Insititute of Technology (ETH), May, 1998
- Duda, Hart and Stork, Pattern classification (2nd Ed), Wiley, NY 2000

# References on network discovery

- D. Zhu, A.O. Hero, Z.S. Qin, "High throughput screening of co-expressed gene pairs with controlled False Discovery Rate (FDR) and Minimum Acceptable Strength (MAS)," submitted to *Bioinformatics*, 2004.
- Barabasi,A. “Network biology: understanding the cell’s functional organization,” *Nat.Rev.Genet.*, **5**, 101-113, 2004.
- Butte,A., Tamayo,P. Slonim,D., Golub,T.R. and Kohane,I.S., “Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks,” *Proc Natl Acad Sci USA*, **97**, 12182-6, 2000.
- Dobra,A., Hans,C., Nevins,R., Yao,G. and West,M. “Sparse graphical models for exploring gene expression data,” *Journal of Multivariate Analysis*, **90**, 196-212, 2004.
- Schafer,J., and Strimmer,K., “An empirical Bayes approach to inferring large-scale gene association networks,” *Bioinformatics*, **1**, 1-13, 2004..
- Stock,M., Victoria,L. and Goudreau,P.N., “Two-component signal transduction. *Annual Review of Biochemistry*”, **69**, 183-215, 2000.
- Yeung,M., Tegner,J. and Collins,J.J., “Reverse engineering gene networks using singular value decomposition and robust regression,” *Proc Natl Acad Sci USA*, **99**, 6163-6168, 2002.
- Zareparsi,S., Hero,A.O., Zack,D.J., Williams,R. and Swaroop,A. “Seeing the unseen: Microarray-based gene expression profiling in vision,” *Invest Ophthalmol Vis Sci.*, **45**, 2457-2462, 2004.
- Zhou,X., Kao,M. and Wong,W.H, “Transitive functional annotation by shortest path analysis of gene expression data,” *Proc Natl Acad Sci USA*, **99**, 12783-12788, 2002