Part II: In vitro measurement of gene expression

Alfred O. Hero III

University of Michigan, Ann Arbor, MI

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

ISBI Tutorial

April 2004

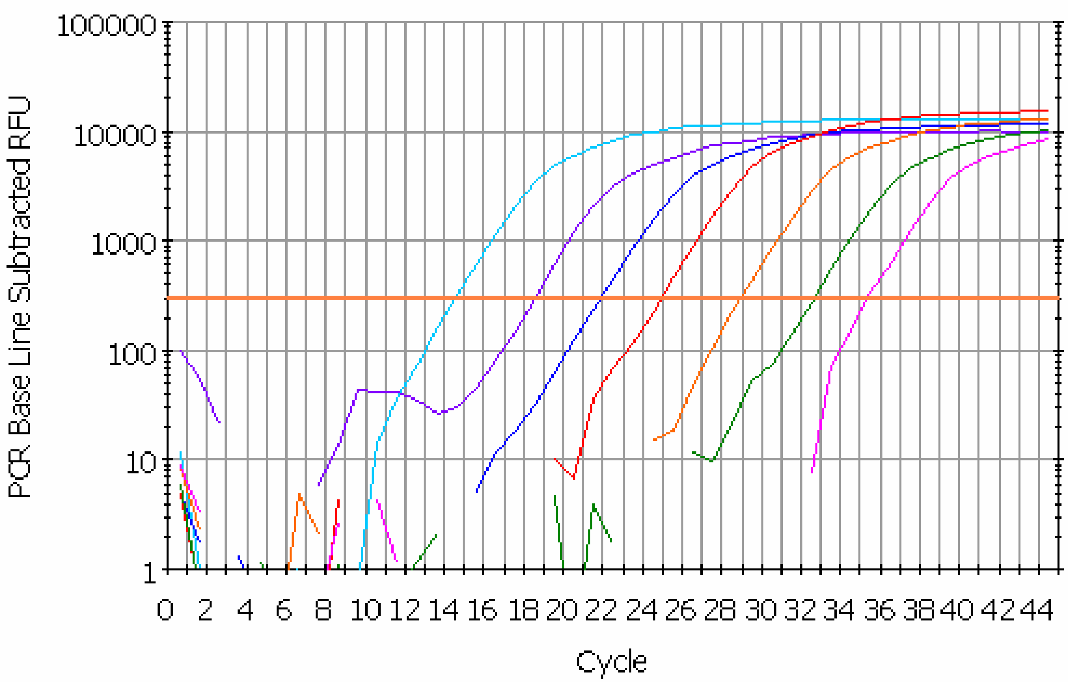
1. Hierarchy of biological questions
2. Gene Microarrays
3. Low Level Summaries of Microarray Data
4. Time/Treatment Course Studies
5. Gene Filtering, Ranking and Clustering
6. Wrap up and References

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?

Standard *in vitro* Method: Real-Time RT-PCR

- Highly accurate quantification of mRNA abundance in a sample



SERIES OF 10-FOLD DILUTIONS (y axis represents fluorescent intensity)

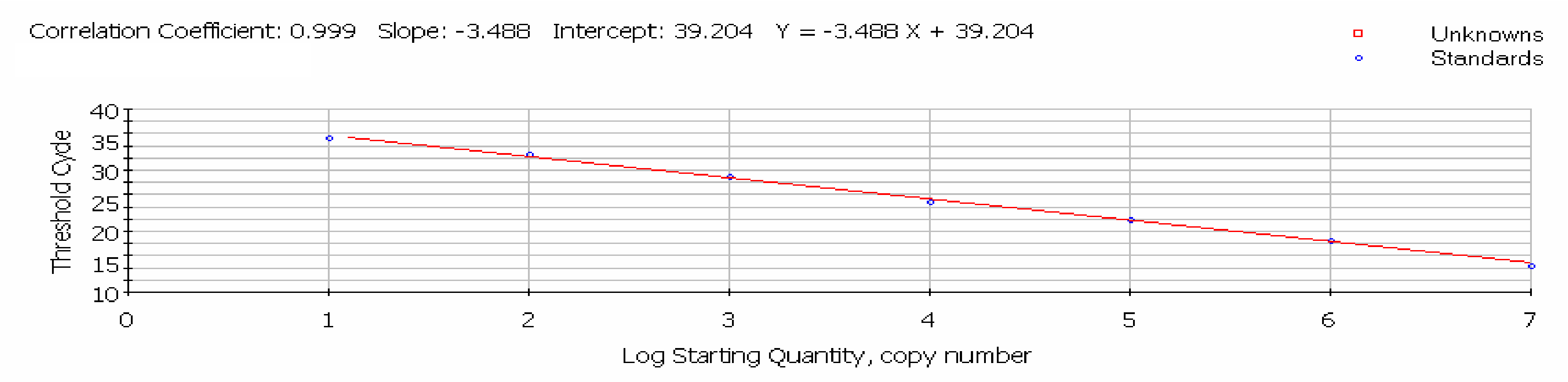
- **<u>C</u>**YCLE NUMBER IS POINT AT WHICH CURVE CROSSES Ct **<u>T</u>**HRESHOLD (Shown in Orange). THIS CROSSING POINT IS KNOWN AS THE Ct VALUE. MORE DILUTE SAMPLES WILL CROSS AT LATER Ct VALUES

Source: Sybr Green (2004)
http://www.med.sc.edu:85/pcr/realtime-home.htm

Quantification and variation of RT-PCR

- **Threshold is usually set between noise floor and saturation plateau**
- **Quality control – efficiency issues: Ct non-linear in log_2(copy number)**
- **Validation methods**:
  - Standard curve method (figure below)
  - Pfaffl Method M.W. Pfaffl, *Nucleic Acids Research* (2001) **29**:2002-2007

$$\text{ratio} = \frac{(E_{\text{target}})^{\Delta\text{Ct target (control-treated)}}}{(E_{\text{ref}})^{\Delta\text{Ct ref (control-treated)}}}$$



PCR Standard Curve: Data 27-Jan-03 1233ileff.opd

- **Low throughput method: <100 genes can be measured simultaneously**

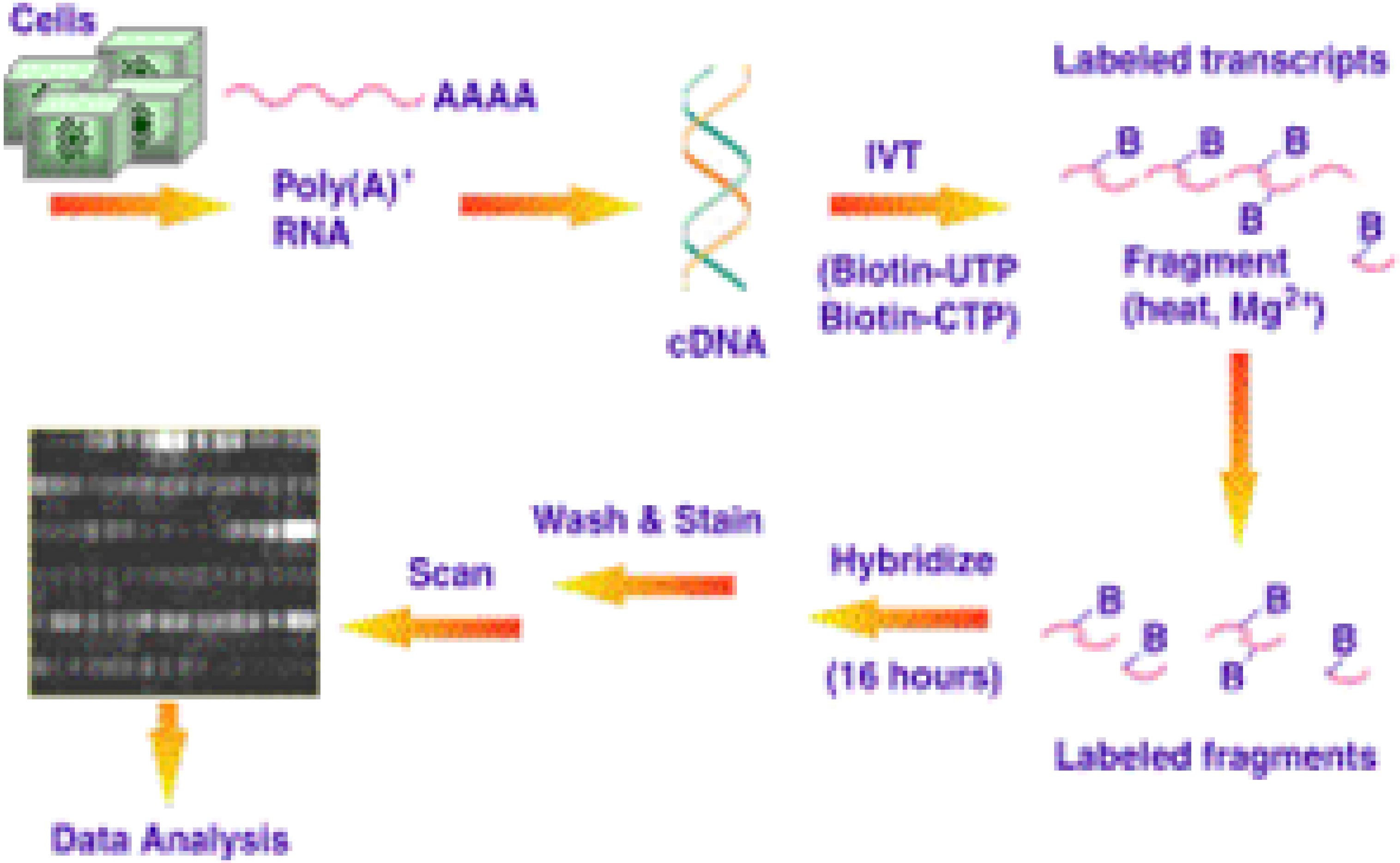
Source: Sybr Green (2004)
http://www.med.sc.edu:85/pcr/realtime-home.htm

2. Gene Microarrays

- Two principal gene microarray technologies:
  - Oligonucleotide arrays: (Affymetrix GeneChips)
    - Matched and mismatched oligonucleotide probe sequences photetched 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

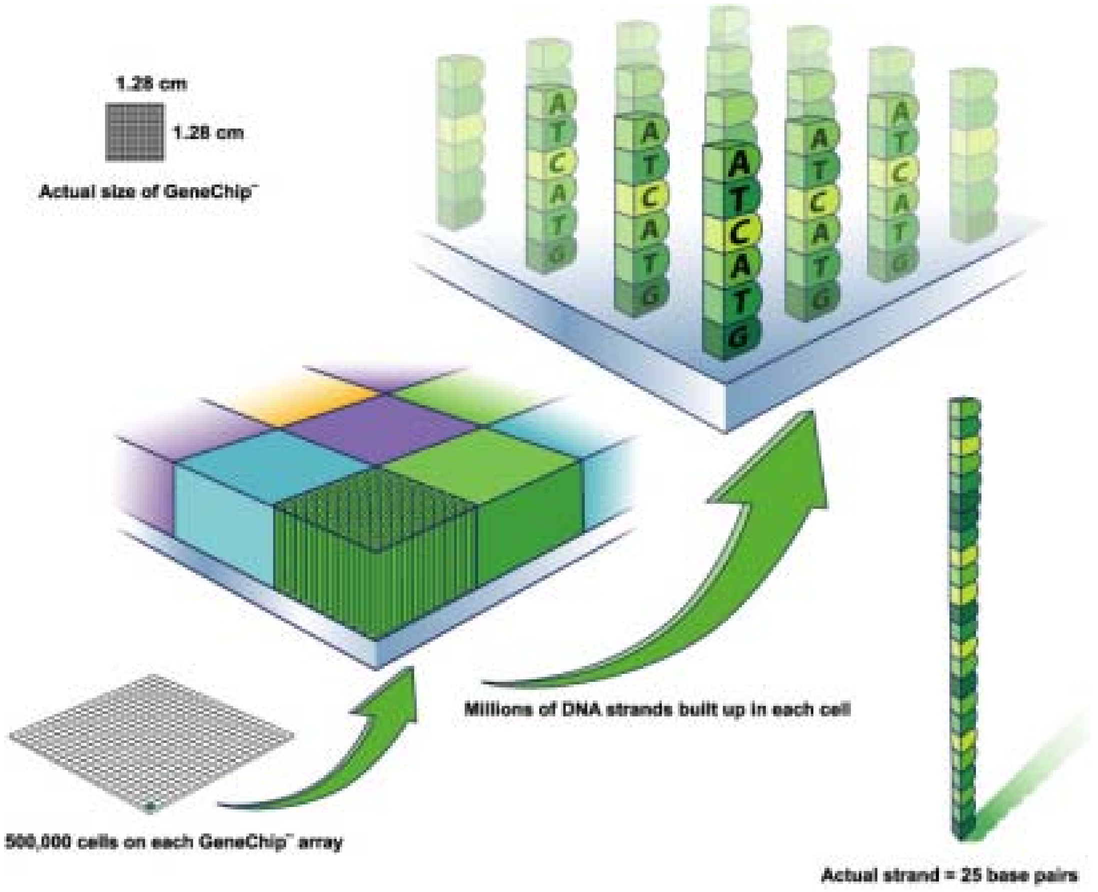
- **10,000-50,000 genes can be probed simultaneously**

Oligonucleotide Chips:



Source: http://www.tmri.org/gene_exp_web/oligoarray.htm

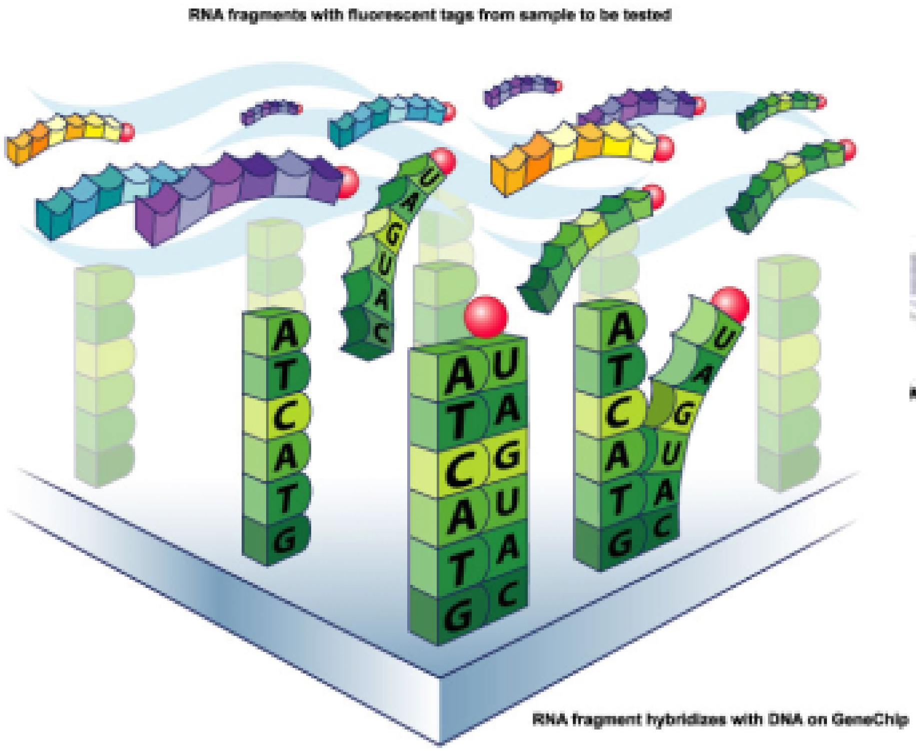
Oligonucleotide Chips


1.28 cm
1.28 cm
Actual size of GeneChip™
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip™ array
Actual strand = 25 base pairs
A
T
C
A
T
G

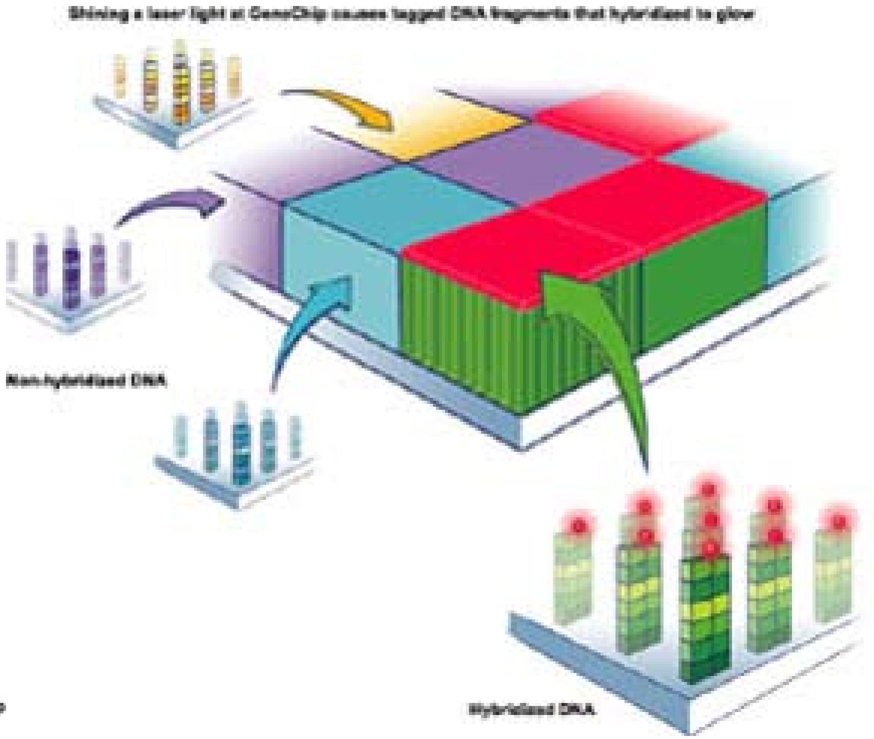
Single feature on an Affymetrix GeneChip microarray

Source: Affymetrix website

Oligonucleotide Chips


RNA fragments with fluorescent tags from sample to be tested
A T C A T G
U A G U A C
RNA fragment hybridizes with DNA on GeneChip

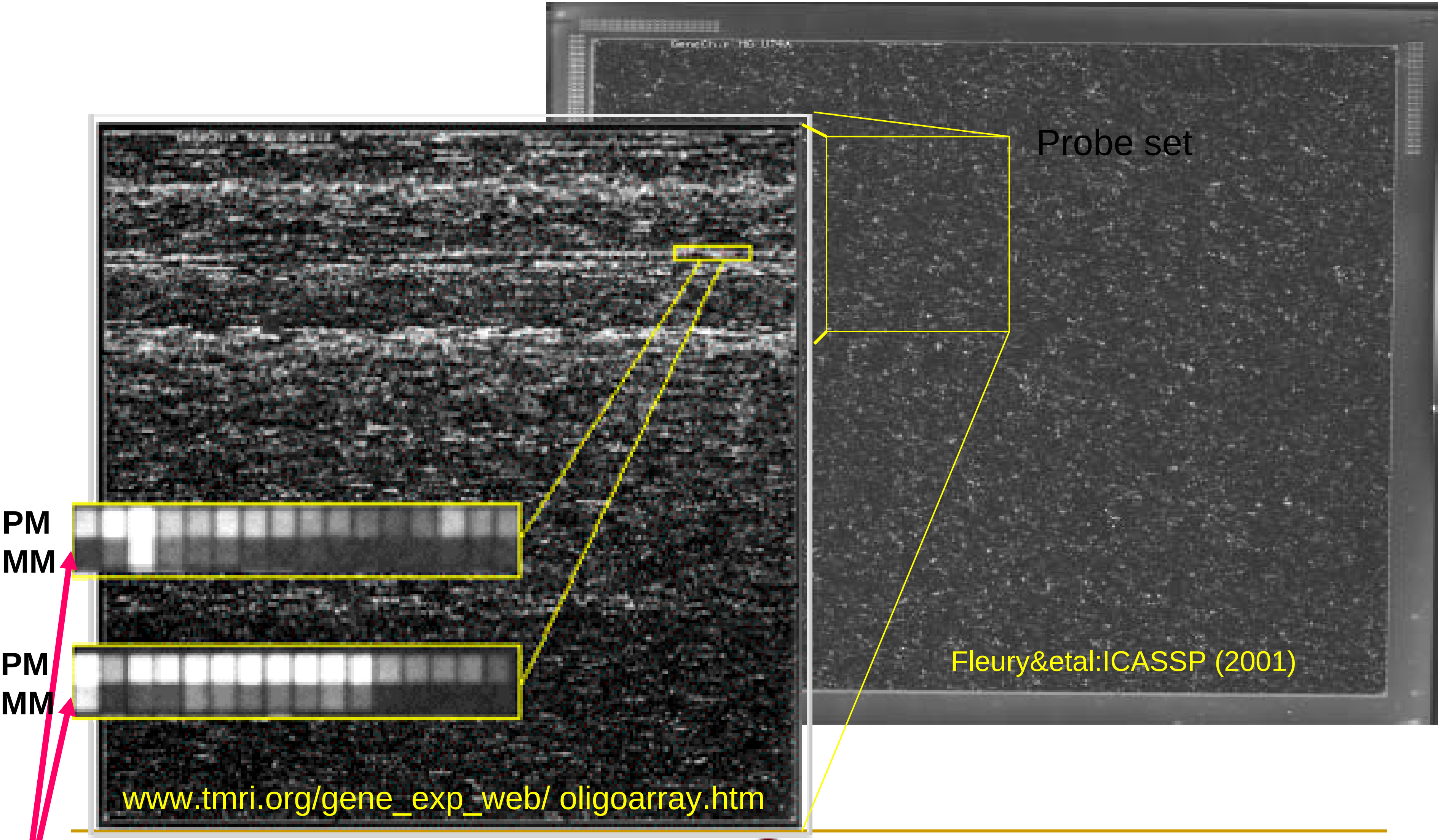
Hybridization to sample


Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA

Scanning and Readout

Source: Affymetrix website

Oligonucleotide GeneChip (Affymetrix)

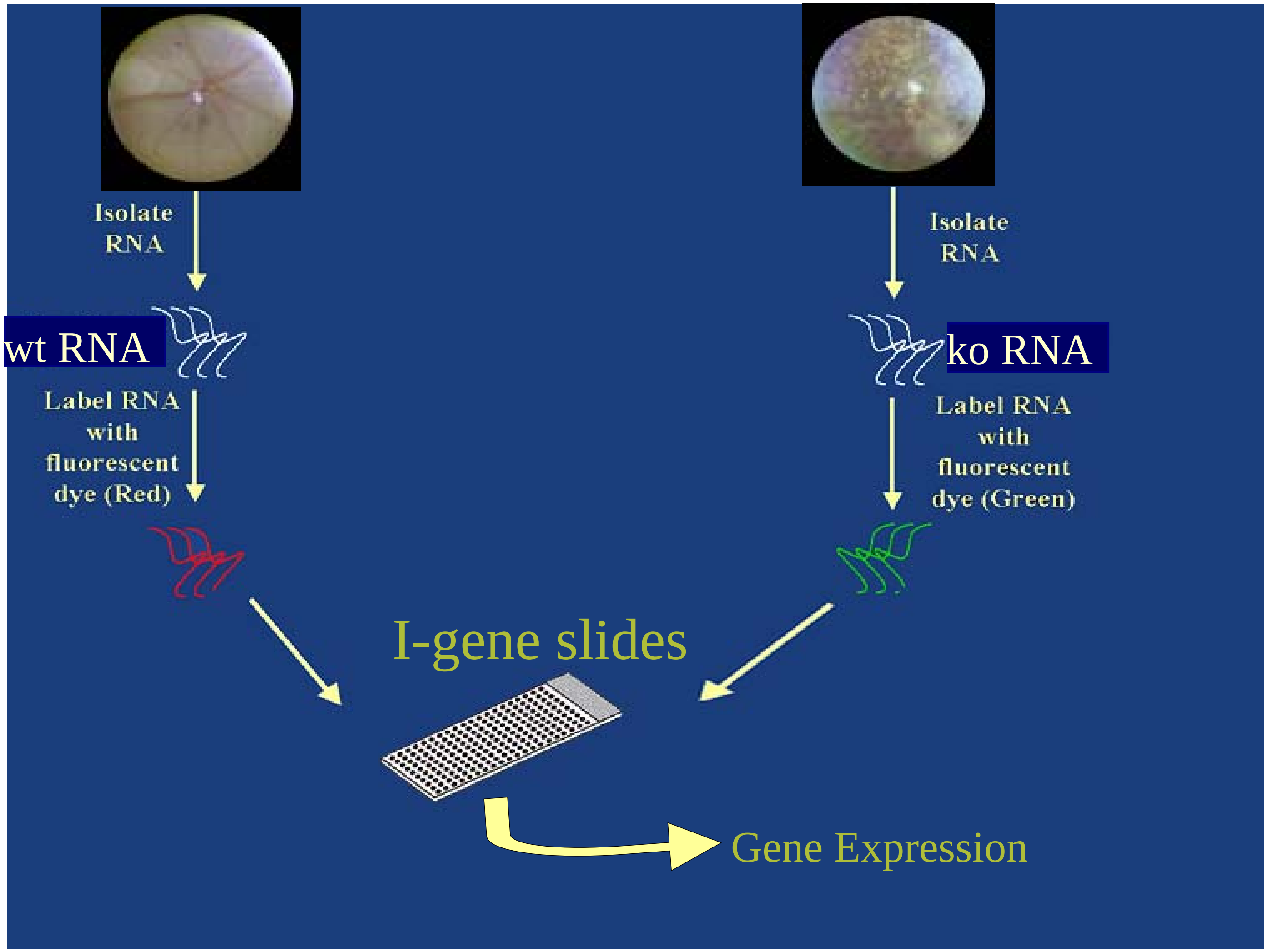


Fleury&etal:ICASSP (2001)

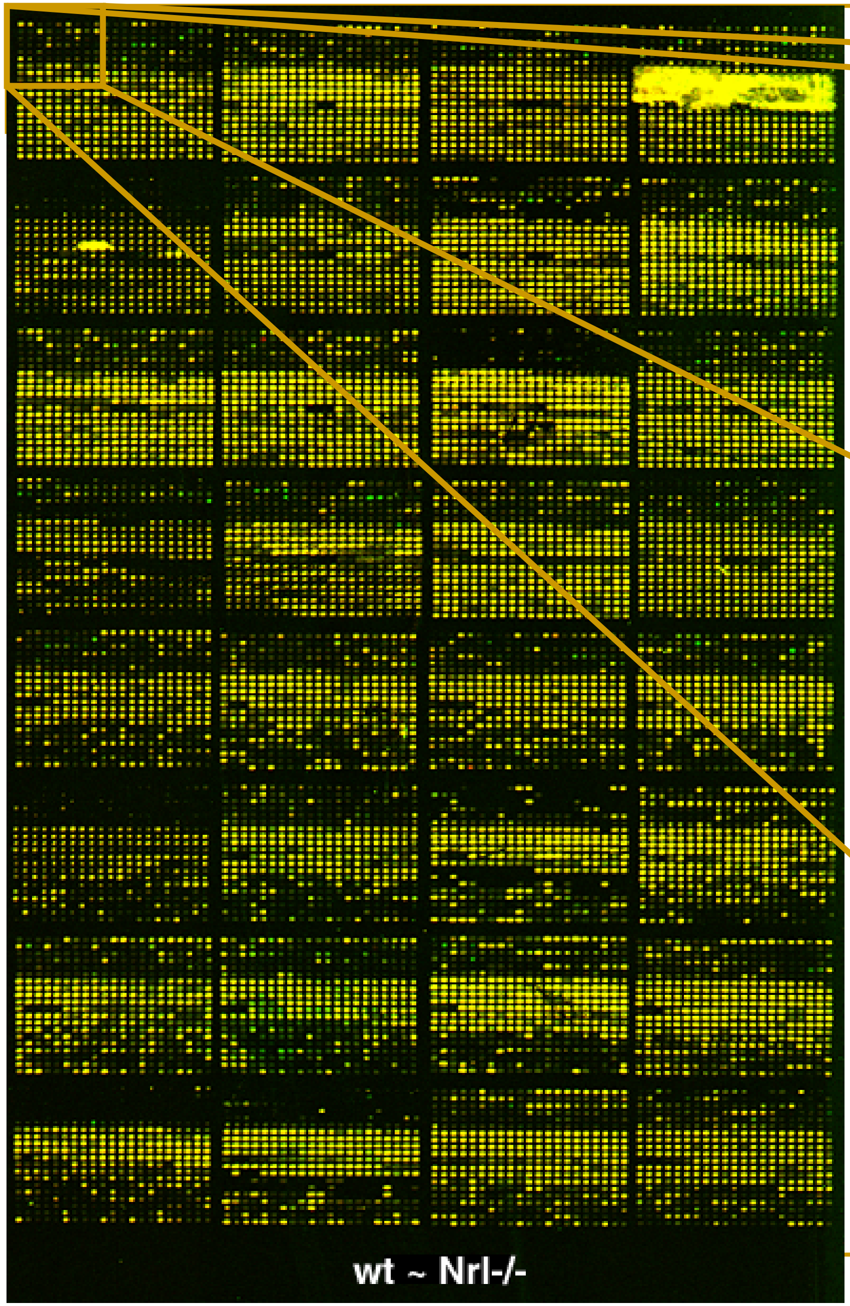
www.tmri.org/gene_exp_web/ oligoarray.htm

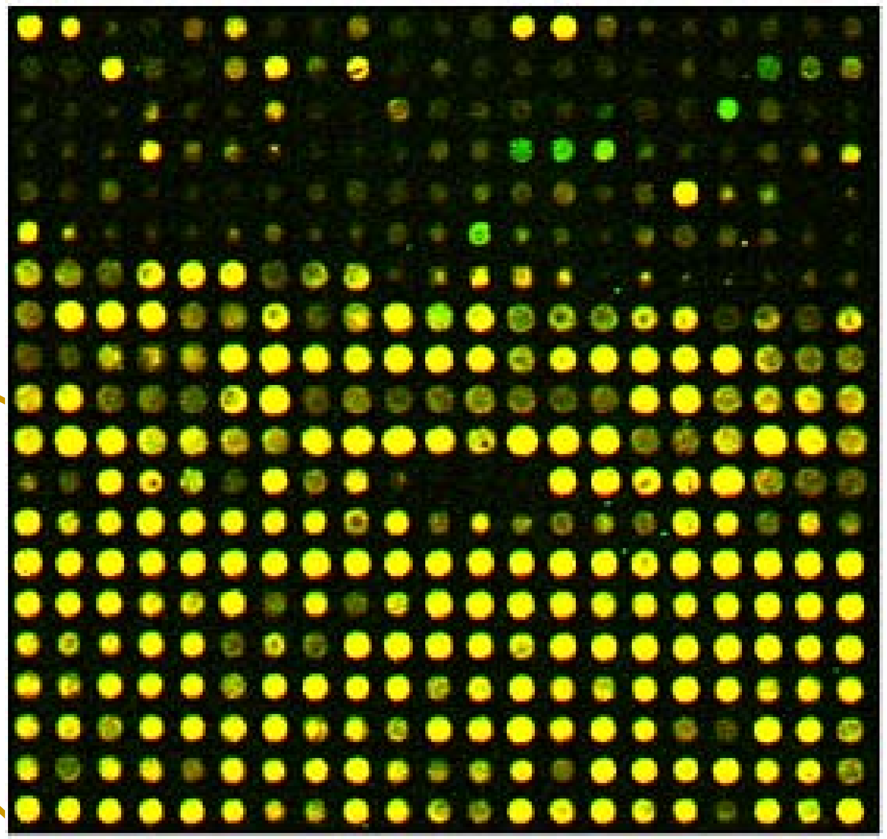
Two PM/MM Probe sets

I-Gene Microarray ko/wt Experiment


Isolate RNA
wt RNA
Label RNA with fluorescent dye (Red)
Isolate RNA
ko RNA
Label RNA with fluorescent dye (Green)
I-gene slides
Gene Expression

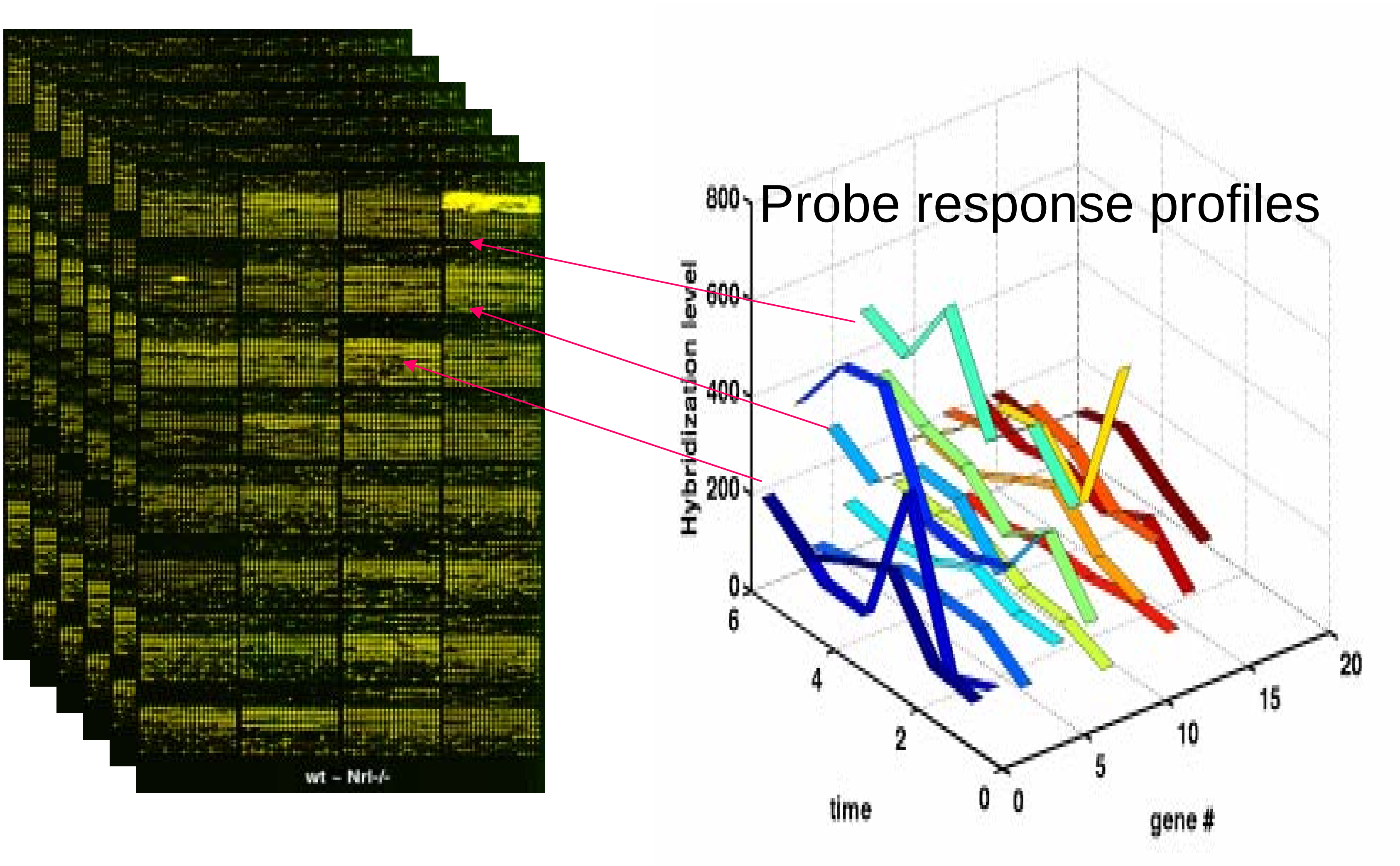
Source: J. Yu, UM BioMedEng Thesis (2004)





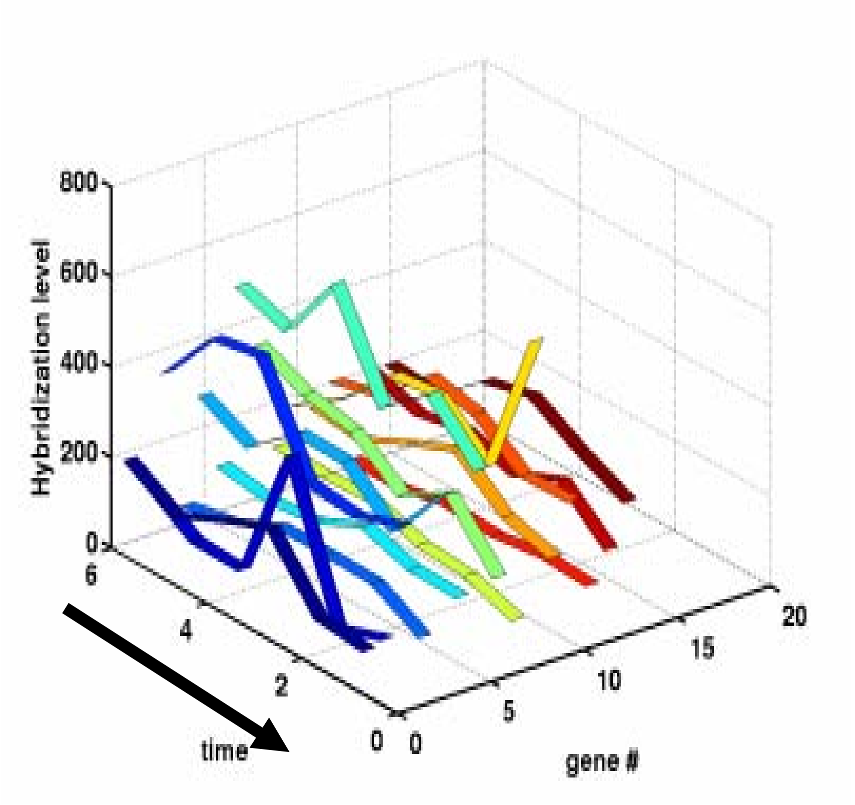
- Treated sample (ko) labeled red (Cy5)
- Control (wt) labeled green (Cy3)

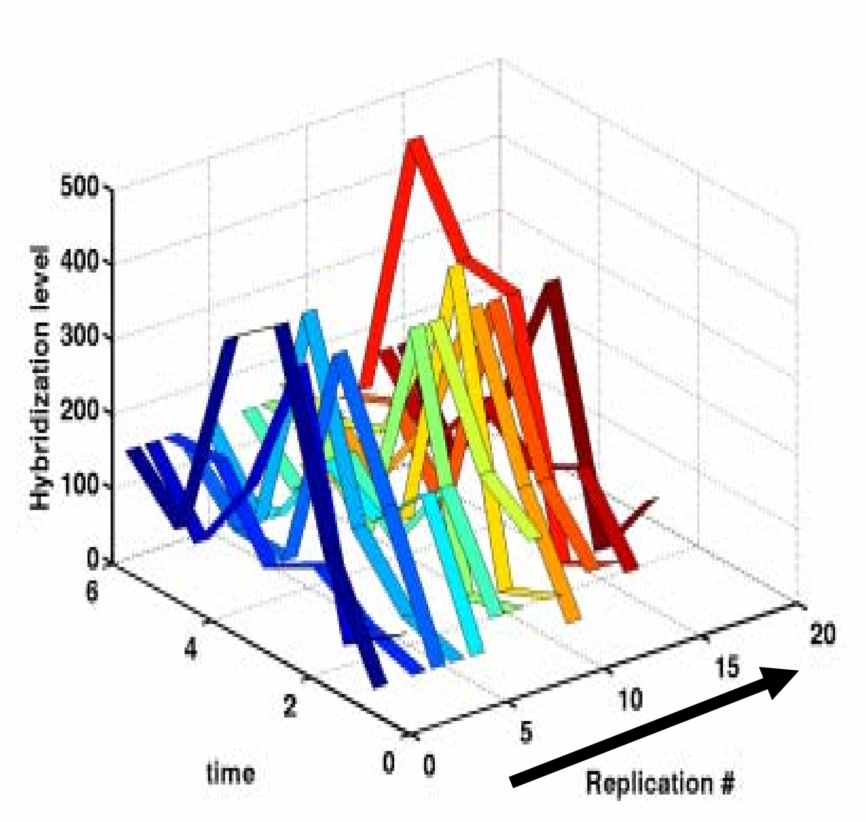
Add Treatment Dimension: Expression Profiles



Probe response profiles

Problem of Sample Variability





Across treatment variability

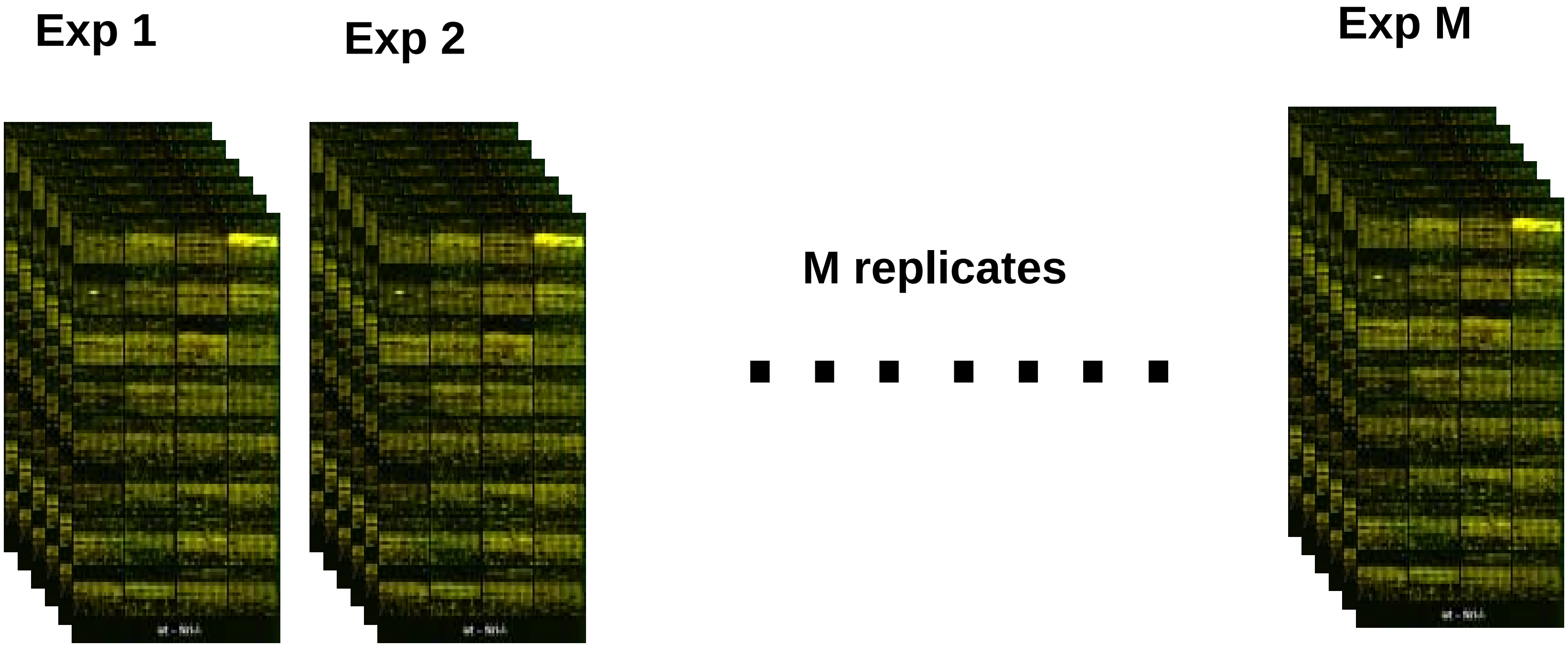
Across sample variability

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.

Solution: Experimental Replication



Exp 1

Exp 2

Exp M

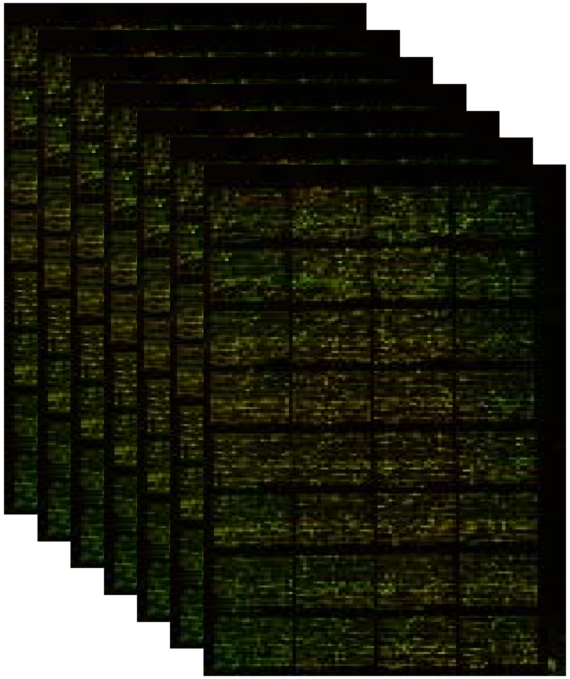
M replicates

. . . .

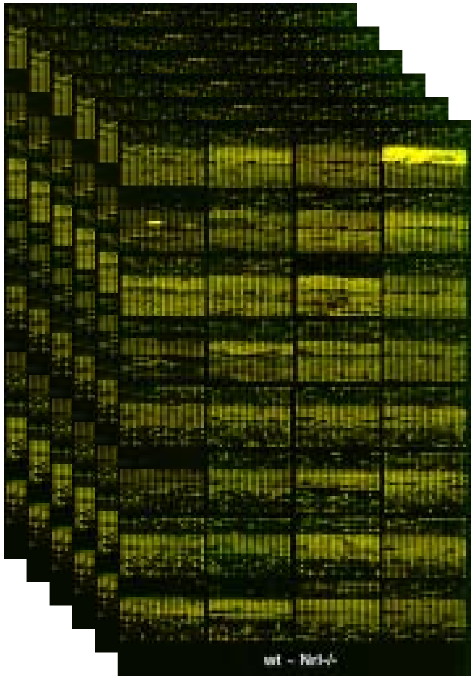
Issues:

- Control by experimental replication is expensive
- Surplus real estate allows replication in layout
- Batch and spatial correlations may be a problem

Comparing Across Microarray Experiments



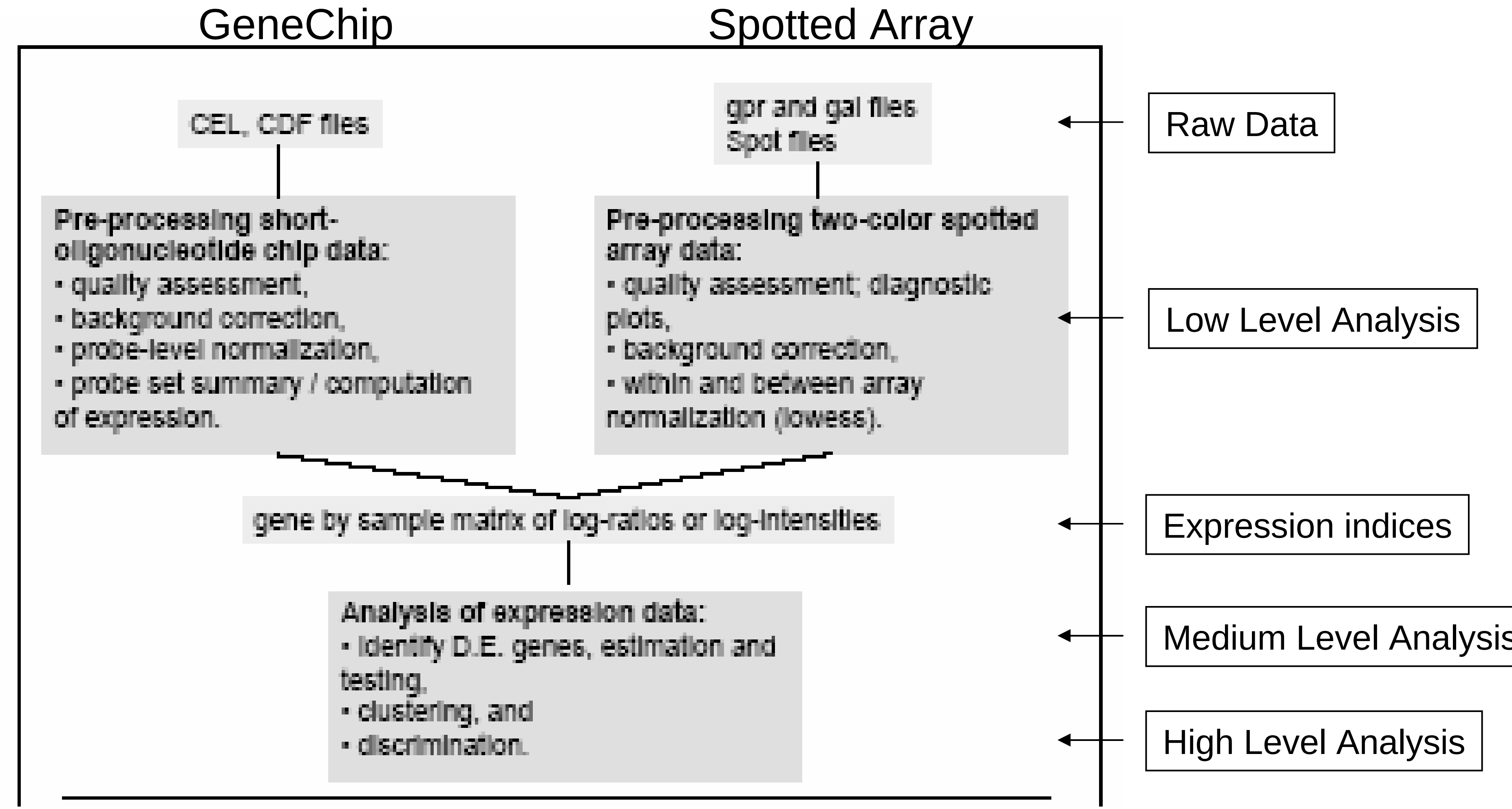
Experiment A



Experiment B

Question: How to combine or compare experiments A and B?

3. Low Level Summaries of Microarray Data



GeneChip

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.

Raw Data

Low Level Analysis

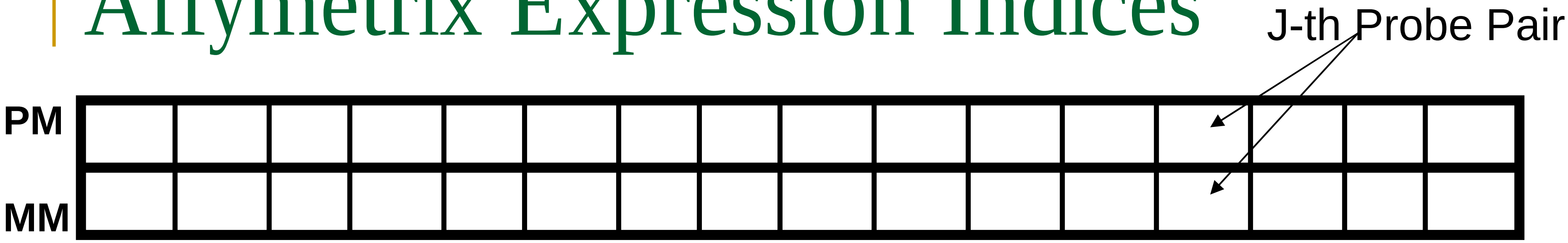
Expression indices

Medium Level Analysis

High Level Analysis

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

Affymetrix Expression Indices



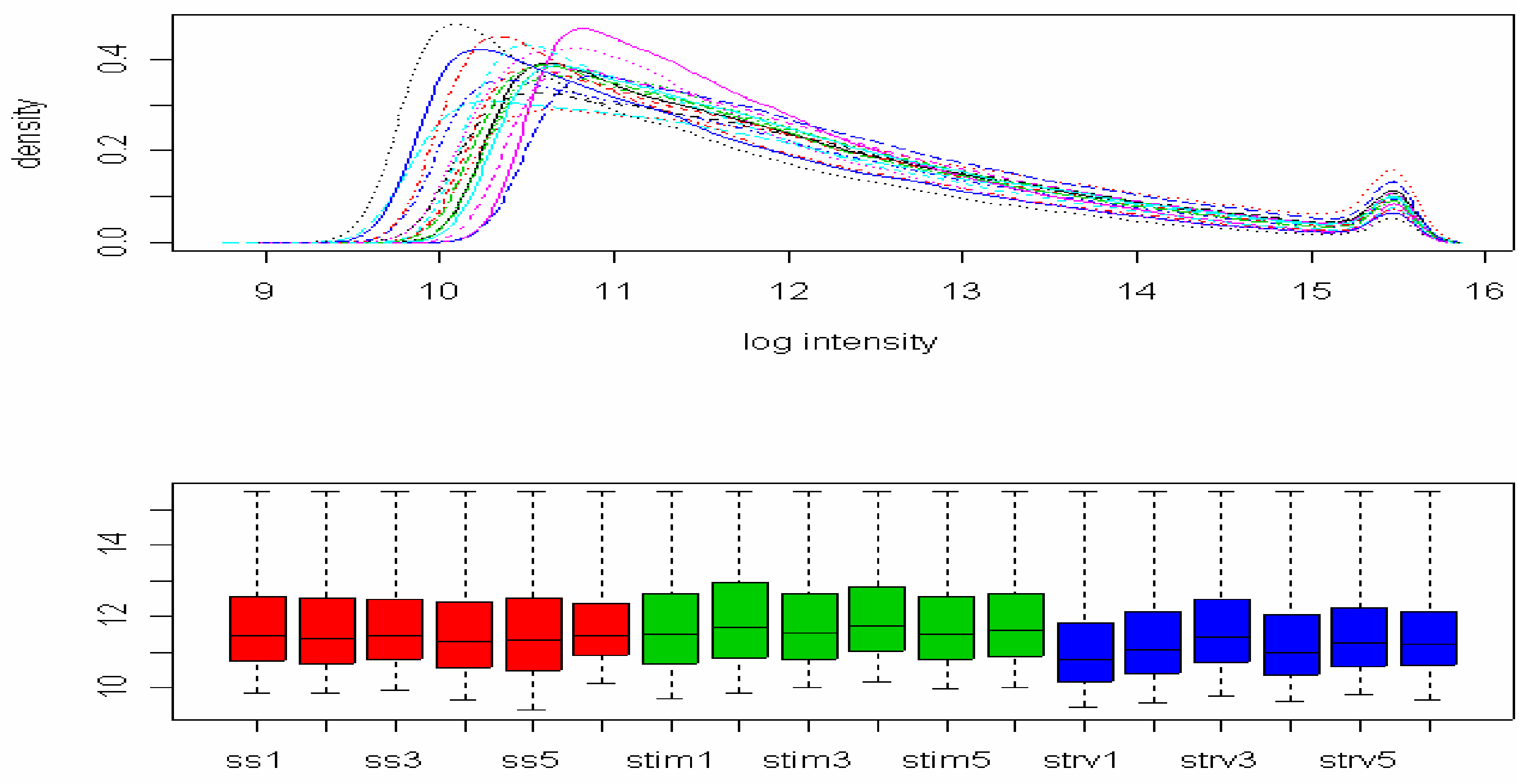
- PM_{ijg} , MM_{ijg} = Intensity for perfect match and mismatch probe in cell j for gene g in chip i
  - $i = 1,\ldots, M$
  - $j = 1,\ldots, J$
  - $g = 1,\ldots, G$
- Task: summarize for each probe set the probe level data, i.e. 20 25 PM and MM pairs, into a single index
- Expression index may then be compared within and between chips for detecting differentially expressed genes

Background Mitigation and Normalization

- Background Mitigation:
  - Subtraction: average of the lowest 2% probe cells
  - Model based methods
- Normalization: The process of identifying and removing systematic variation NOT due to real differences between treatments i.e. differential gene expression
  - Moment normalization: match all means/variances
  - Quantile normalization
    - Makes the distribution of probe intensities the same for every chip
    - Normalized distribution is obtained by averaging each quantile across chips

Refs: Irizzary&etal:2002, Bolstead&etal:2003

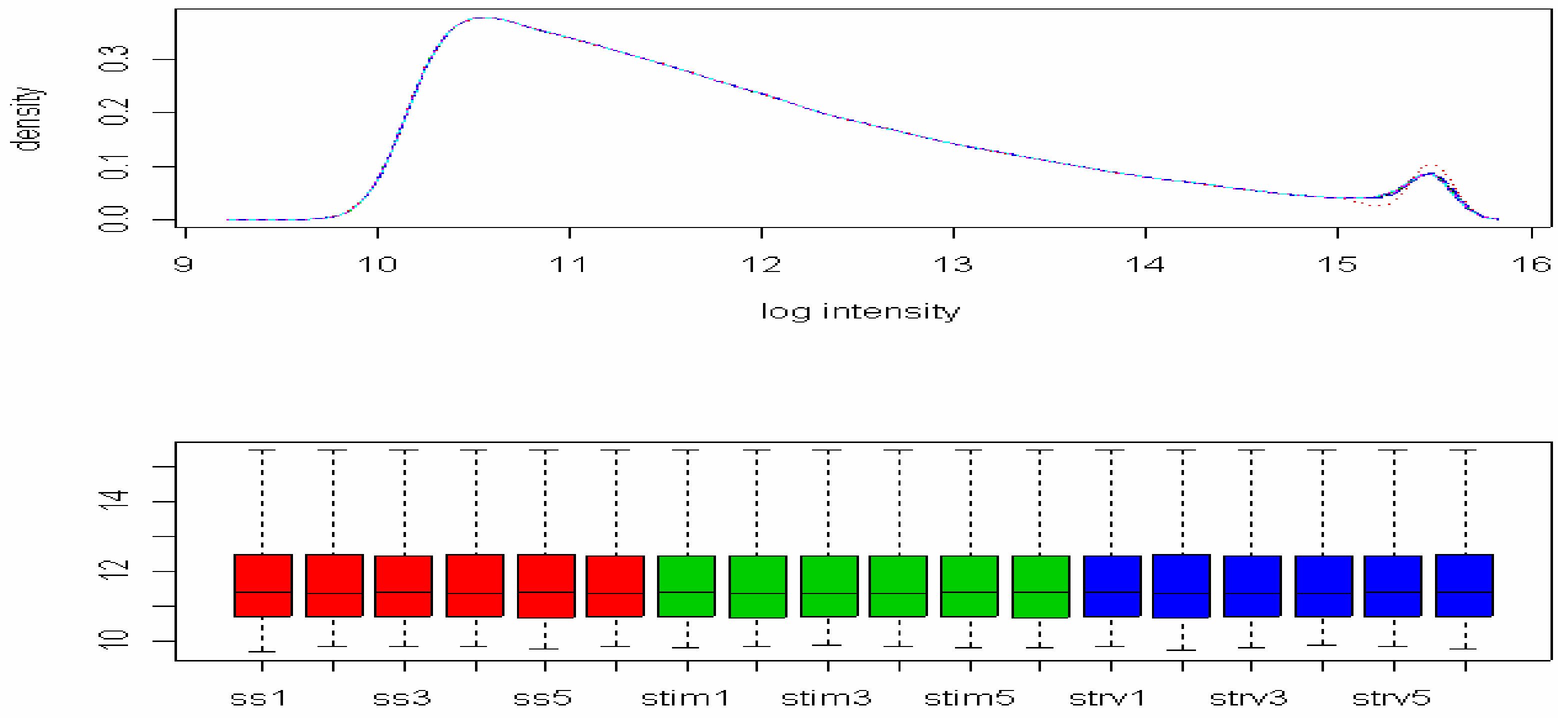
Multi-Slide Histogram: pre-Normalization



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

Data: Lemon WJ et al. 2002

Multi-Slide Histogram: post-Normalization



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

Data: Lemon JL et al. 2002

Affymetrix Expression Indices: MAS4

- GeneChip® MAS4 software made calls based on

$$Avg.diff = \frac{1}{|O|} \sum_{j \in O} (PM_j - MM_j)$$

$$LogAvg = \frac{1}{|O|} \sum_{j \in O} (\log PM_j - \log MM_j)$$

O = set of “suitable” oligo pairs chosen by software.

- Log ratio version was viewed as more reliable
- In differential studies these scores are compared between treatments/times.

Affymetrix Expression Indices: MAS5

- MAS5 uses a more sophisticated technique
  - Signal = TukeyBiweight{log(PMj-MMj*)}
  - MM* is a version of MM that is never bigger than PM
  - *Ad-hoc* background subtraction procedure and scale normalization are used

see Hubbell (2001)

http://www.stat.berkeley.edu/users/terry/zarray/Affy/GL_Workshop/genelogic2001.html

Model-Based Expression Indices: Li&Wong

Expression level

Li-Wong Full (LWF)

$$PM_{ij} = \nu_j + \alpha_j \theta_i + \phi_j \theta_i + e$$

*i*th array

*j*th probe pair

$$MM_{ij} = \nu_j + \alpha_j \theta_i + e,$$

sensitivities

$$e \sim N(0, \xi^2)$$

Identifiability constraint $\sum_j \phi_j^{\,2} = J$ Total no. probe pairs

Li-Wong Reduced (LWR) $y_{ij} = PM_{ij} - MM_{ij} = \phi_j \theta_i + \varepsilon,$

$$\varepsilon \sim N(0, \sigma^2), \sigma^2 = 2\xi^2$$

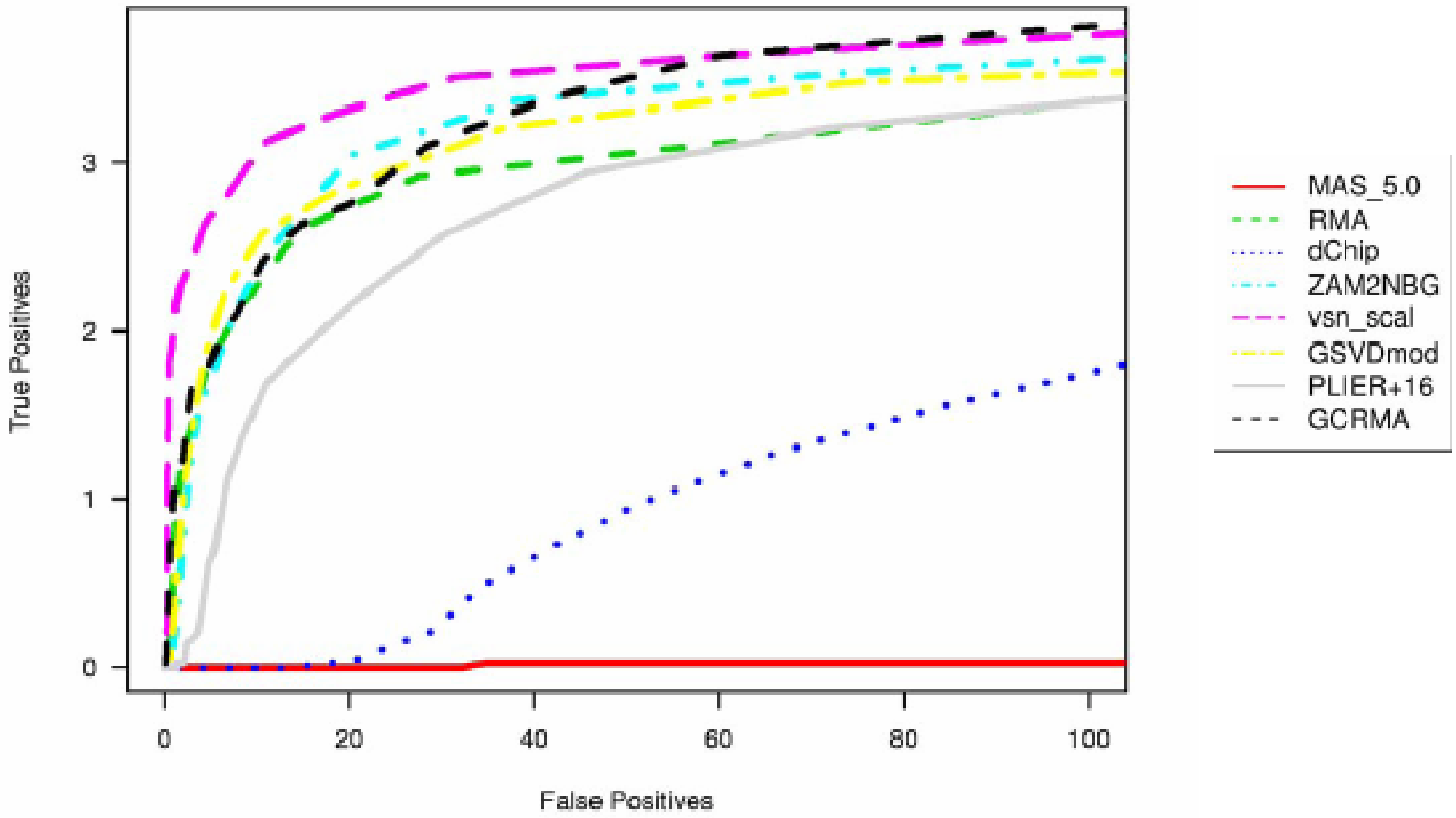
Li, C and Wong, WH, *Proc. Natl. Acad. Sci. USA*, 98:31-36, 2001.

Public code: dChip (available on web)

Model Based Expression Indices: RMA

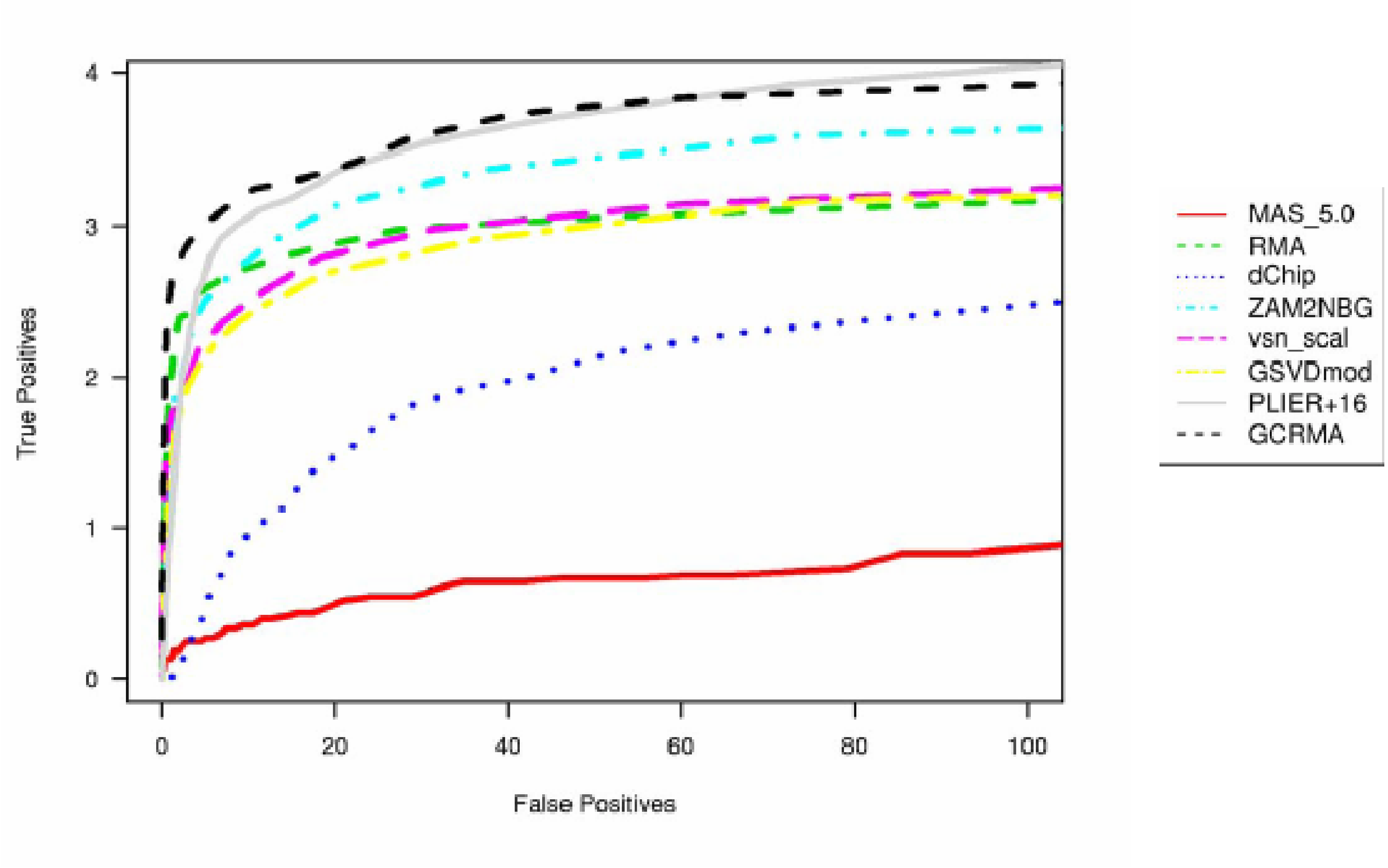
- RMA extracts expression levels from PM only
  - Background adjustment based on normal + exponential model
    - PM = Background+Signal
    - S estimated as Y=E[S|B+S], with B normal and S exponential
  - Perform quantile normalization on estimated S's
  - Post-normalization model: $\log Y_{ij} = \theta_i + \alpha_j + \varepsilon_{ij}$
  - Perform robust linear regression to find expression levels θ_i

Discovery Rate ROC: Spike-in Exp 1



Source: Irizarray R et al. The 2003 Affymetrix GeneChip Microarray Low-Level Workshop (2003)

Discovery Rate ROC: Spike-in Exp. 2 (low concentration)

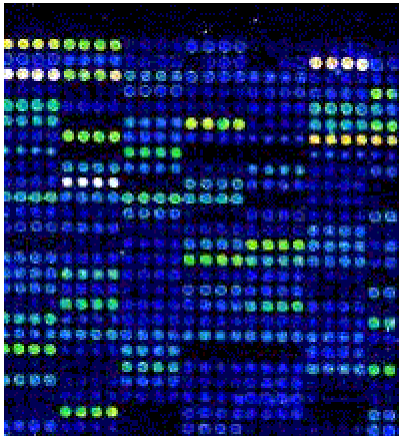


Source: Irizarray R et al. The 2003 Affymetrix GeneChip Microarray Low-Level Workshop (2003)

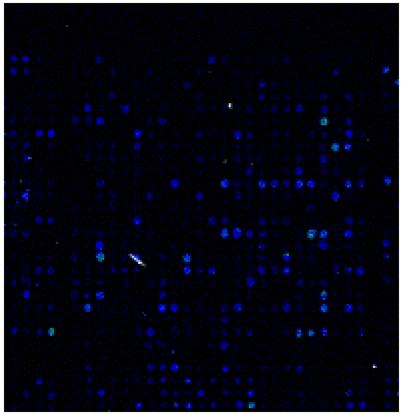
Low Level Processing of Spotted Arrays

- Image Analysis: Spot extraction
  - Addressing, estimation of spot centers
  - Segmentation, classify pixels as foreground or background
  - Signal extraction
- Quality filtering: spot quality and slide quality
- Normalization
  - Single channel normalization of log-intensities
  - Two channel normalization of log-ratios to remove systematic color bias
  - Between slide normalization to align replicates

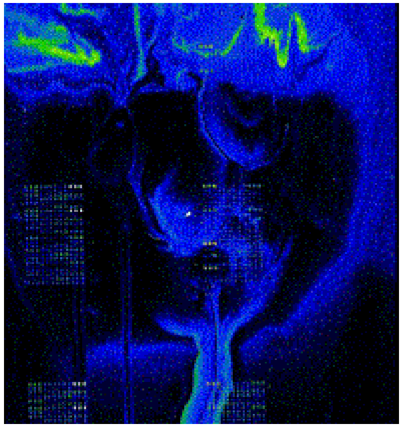
Image Analysis: Spot Extraction



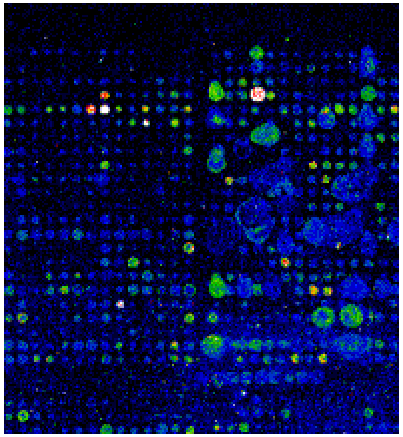
Good Signal



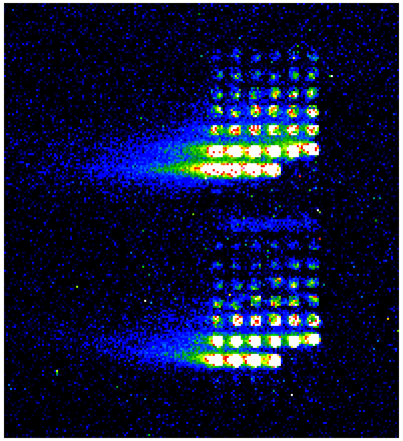
Weak Signal



Streaks



Irregular Spots

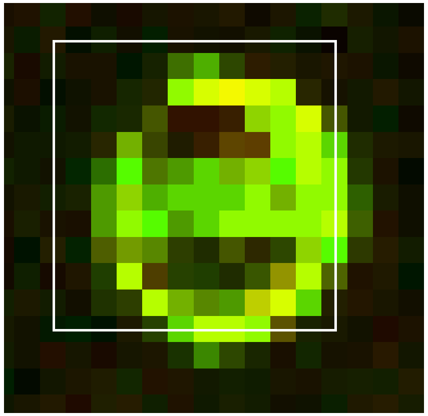


Comet Tails

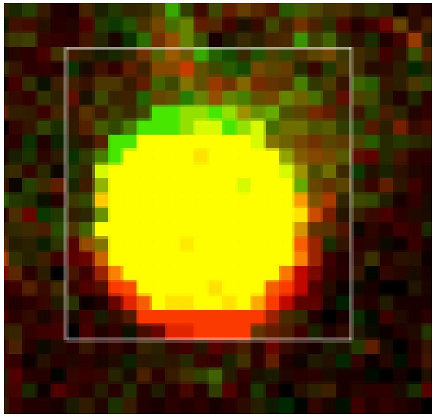
Source: http://stress-genomics.org/

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

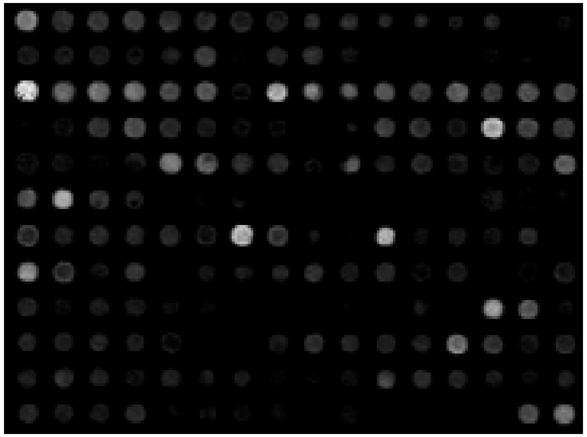
Source: C. Ball, Stanford Microarray Database

Refs: Spotfire, ScanAnalyze, GenePix, Quantarray, Spot

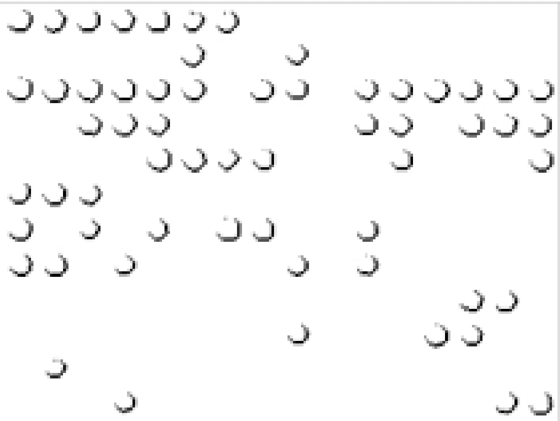
Spot Segmentation Methods

- 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

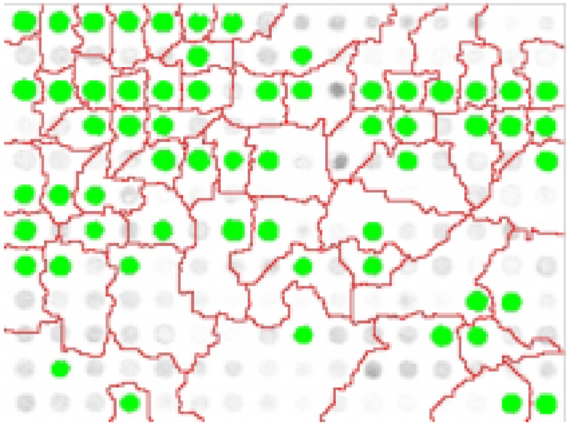
Segmentation via Morphological Operators



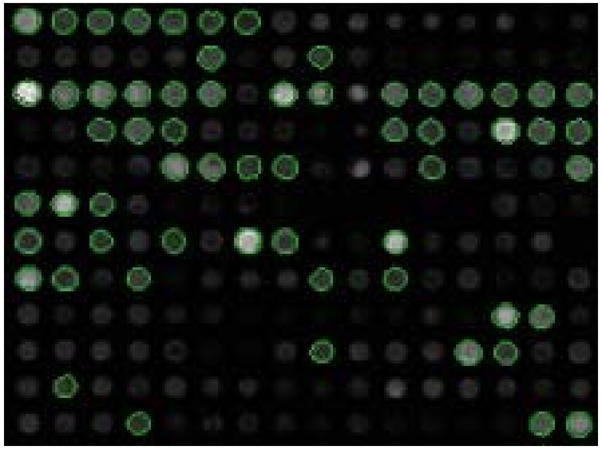
Original Image



Alternate-Sequential Filtered



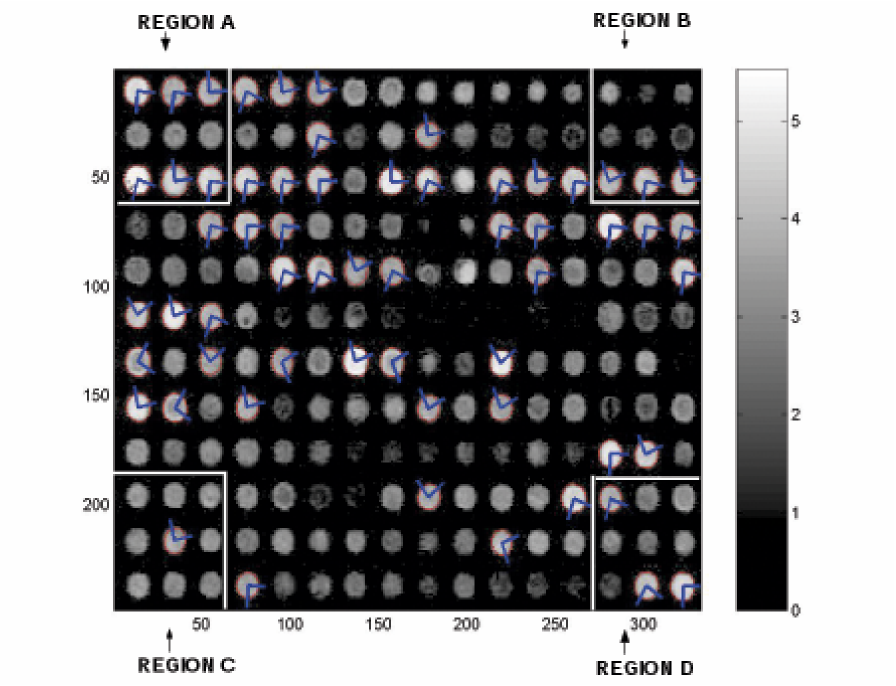
Watershed Transformed



Final Segmented Image

Ref: Siddiqui&etal:Asilomar-02

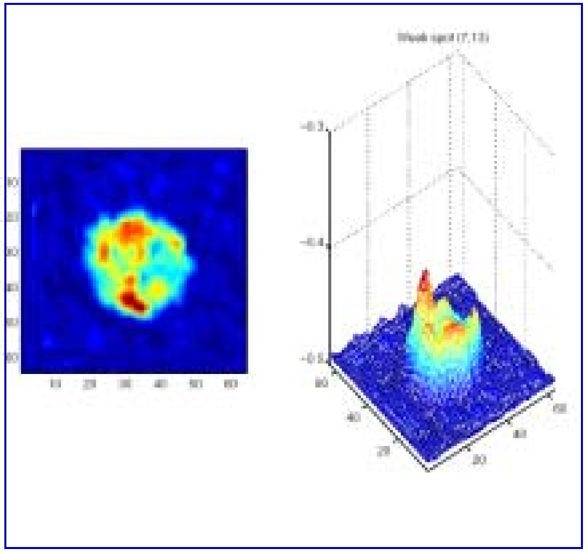
Spot EigenAnalysis



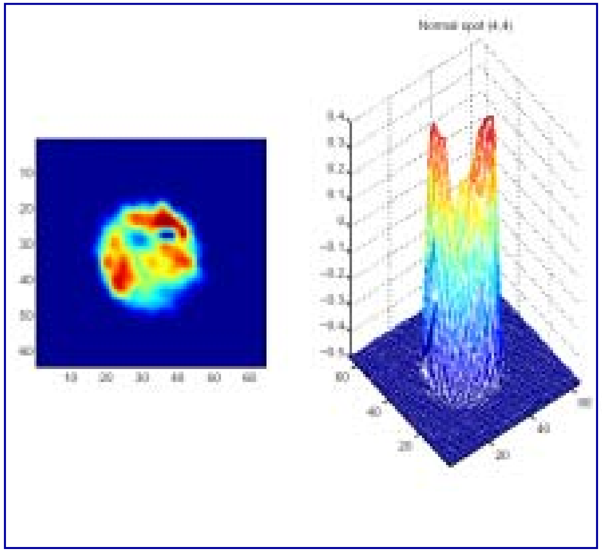
- Gray level covariance matrix over each spot boundary is calculated
- Eigen analysis of each covariance matrix is performed
- Trends in direction of eigenvectors indicate systematic bias in spot printing

Siddiqui, Hero and Siddiqui, Asilomar-02

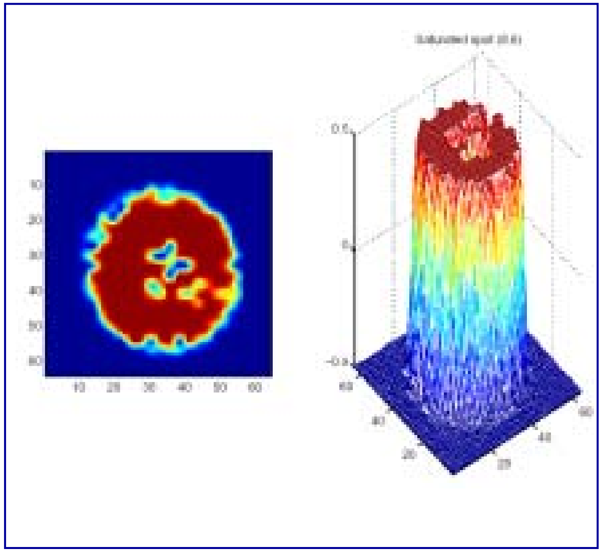
Readout Gain Effects



Weak



Normal



Saturated

- **Weak gain masks weak signals**
- **Saturated gain masks strong signals**
- **Is there a practical way to set optimal gain?**

Background and Normalization methods

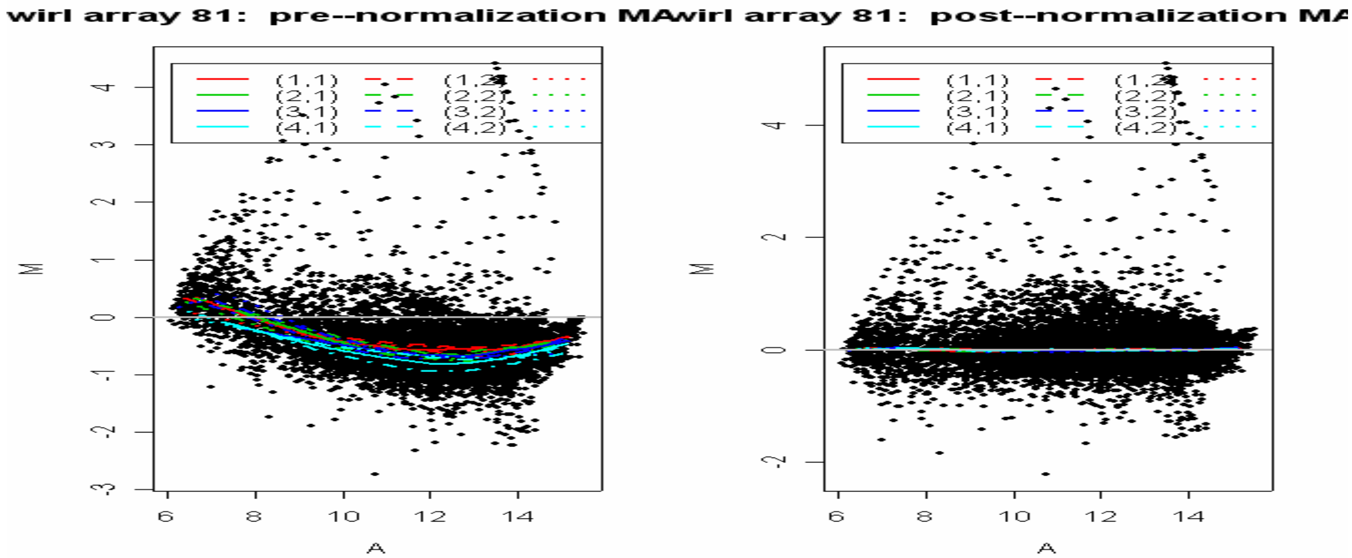
- Background correction: R = (Rf – Rb), G = (Gf - Gb)
    - Rf, Gf are Red, Green foreground and Rb, Gb are Red, Green background
- Normalization of log-ratios (M) i.e. M= $\log_2 R/G$ (two channel normalization)
    - Remove systematic color bias in ratios
    - Primary a within-slide adjustment
    - Normalization algorithms (next slide)
- Normalization of log-intensities i.e. $\log_2 R$ or $\log_2 G$ (single channel normalization)
    - Remove systematic color bias in intensities
    - Normalization algorithms
        - ANOVA
        - Quantile normalization
        - VSN

M= $(\log_2 R/G)$ *vs.* A = $\log_2(RG)^{1/2}$

- M - intensity log ratio, A - average log intensity
- For array i gene g one can assume non-linear model

$$M_{ig} = c_i(A_{ig}) + \overline{M}_{ig} + \sigma_i Z_{ig}$$

- After normalization, no visible M dependency on A in any print-tips



Graphs are generated using Bioconductor "marray" packages written by Sandrine Dudoit

Data source:http://fgl.lsa.berkeley.edu/Swirl/index.html

Within-slide normalization: Location

- Global location normalization
  - Assume that the red and green intensities are related by a constant factor, i.e., $R= kG$
  - The center of distribution of log-ratios is shifted to zero

$$\log_2 R/G \text{ -> } \log_2 R/G - c = \log_2 R/(kG)$$

  - A common choice for $c=log_2k$ is the median or mean of the intensity log ratios for a particular gene set
- Intensity $A = \log_2(RG)^{-1/2}$ dependent normalization
  - $\log_2 R/G \text{ -> } \log_2 R/G - c(A) = \log_2 R/(k(A)G)$,
  - where $c(A)$ is the LOESS fit to the M *vs.* A plot
- Within print-tip-group normalization
  - $\log_2 R/G \text{ -> } \log_2 R/G - c_i(A) = \log_2 R/(k_i(A)G)$
  - where $c_i(A)$ is the LOESS fit to the M *vs.* A plot for the ith grid only, i = 1, …..I, and I is the number of print-tips

Yee Hwa Yang et al. Normalization for cDNA Microarray Data (2001)

Within-slide normalization: Scale

- Scale normalization via maximum likelihood
- After location normalization, assume each print-tip group follows a normal distribution N(0, $a^2_i\sigma^2$)
- MLE:

$$\hat{a}_i = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{\sqrt[I]{\prod_{k=1}^{I} \sum_{j=1}^{n_i} M_{kj}^2}}$$

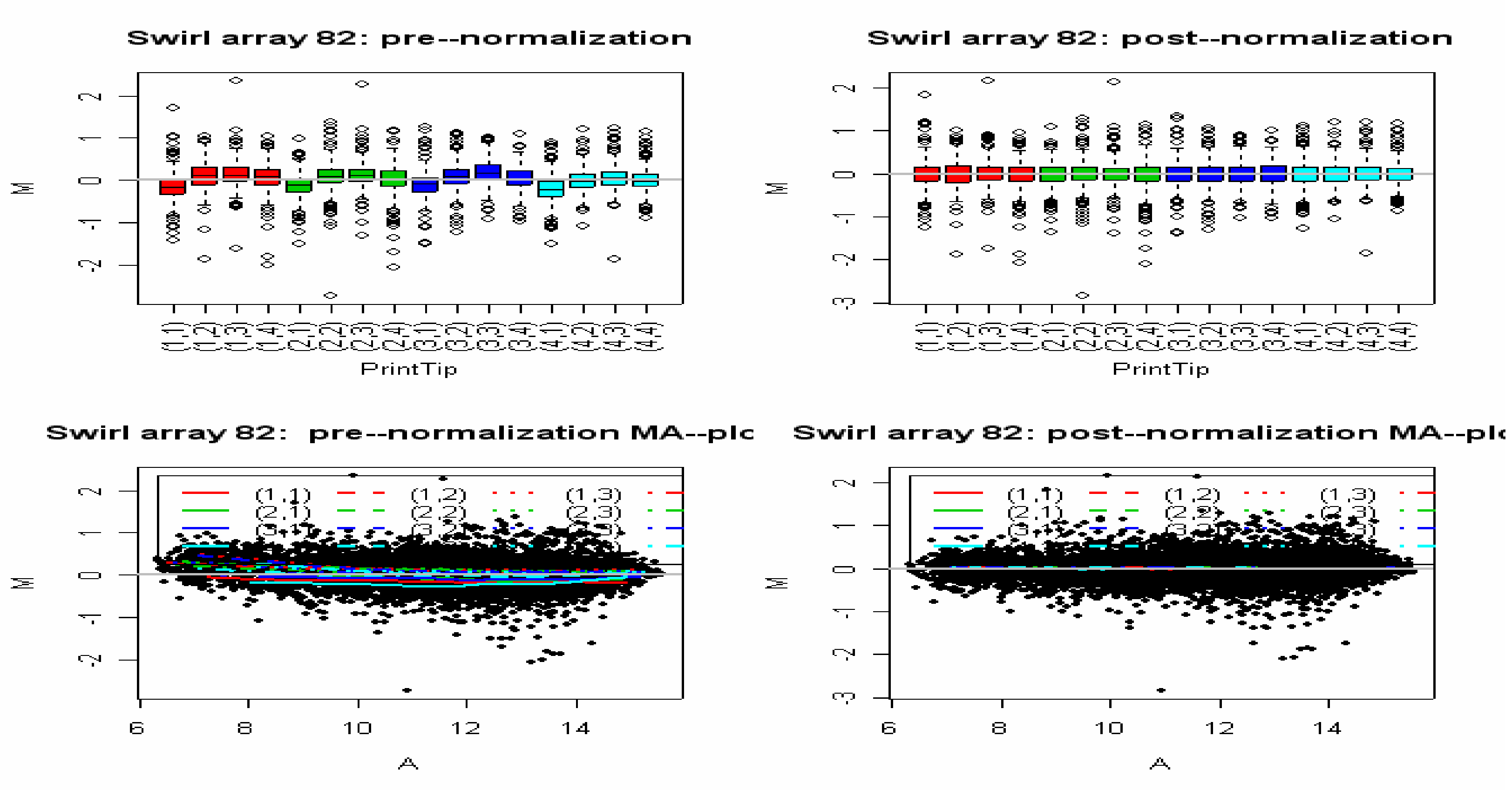
- Robust estimate

$$\hat{a}_i = \frac{MAD_i}{\sqrt[I]{\prod_{i=1}^{I} MAD_i}},$$

$$MAD_i = \text{median}_j\{\mid M_{ij} - \text{median}_j(M_{ij}) \mid\}$$

Yee Hwa Yang et al. Normalization for cDNA Microarray Data (2001)

Within-slide (print-tip-group) pre and post (both Location and Scale) normalization

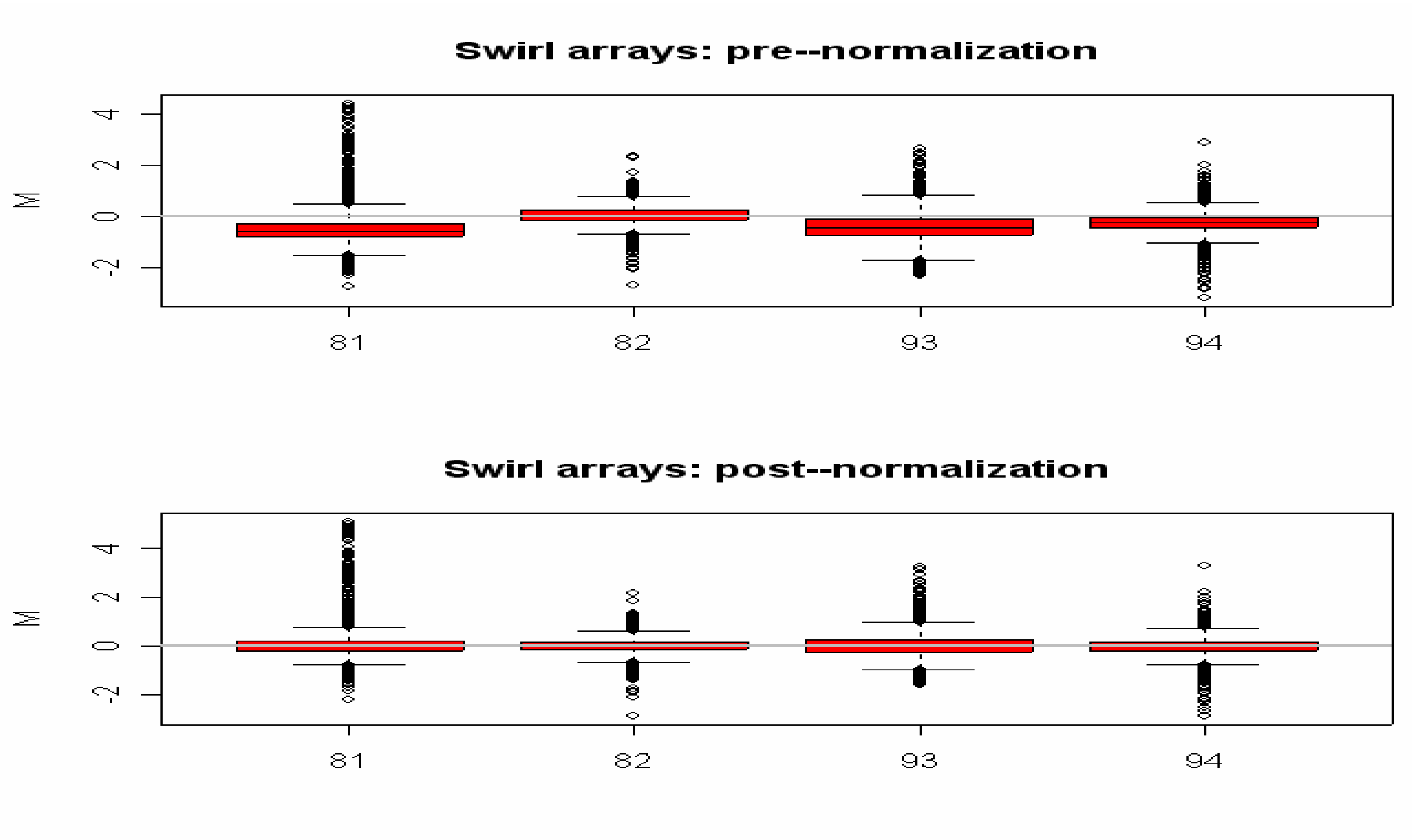


Graphs are generated using Bioconductor “marray” packages written by Sandrine Dudoit

Between-slide normalization

- Here, we are concerned with making the single-channels between slides comparable
- Quantile normalization is based on the idea of normalizing for equivalent medians or quartiles, requiring that every quantile across channels be equal and forcing the channels to have the same distribution
- This distribution is estimated by the average of each quantile across all channels
- Ref: Natalie Thorne and Gordon Smith have implemented this method in the Bioconductor package “limma”
- Use Bioconductor “marrayNorm” package written by Sandrine Dudoit, normalization is performed simultaneously for each array in the batch using the location and scale normalization procedures (next slide)

Between-slide pre and post normalization



Graphs are generated using Bioconductor "marrayNorm" packages written by Sandrine Dudoit

4. Time/Treatment Course Studies

- Objective: find all genes having significant foldchanges wrt multiple criteria $\xi_1(g), \ldots, \xi_p(g)$

$$\mathrm{fc}(g) = \overline{K}_t(g) - \overline{W}_t(g), \quad g = 1, \ldots, G$$

$\overline{K}_t, \overline{W}_t$ =log2 of the mean ko,wt expression levels

- Issues
  - Selection criteria (ratios, profiles, patterns)
  - Controlling statistical significance
  - Controlling biological significance

Possible Selection Criteria

- Some multicriteria $\xi_1(g), \ldots, \xi_p(g)$
  - Variance-normalized paired comparisons for two treatments at a single time point

$$\boxed{\xi_1(g) = (\overline{K}(g) - \overline{W}(g))/\mathsf{s}(g)}$$

  - Paired comparisons for two treatments at a single time point

$$\boxed{\xi_1(g) = \mathsf{s}(g), \quad \xi_2(g) = \overline{K}(g) - \overline{W}(g)}$$

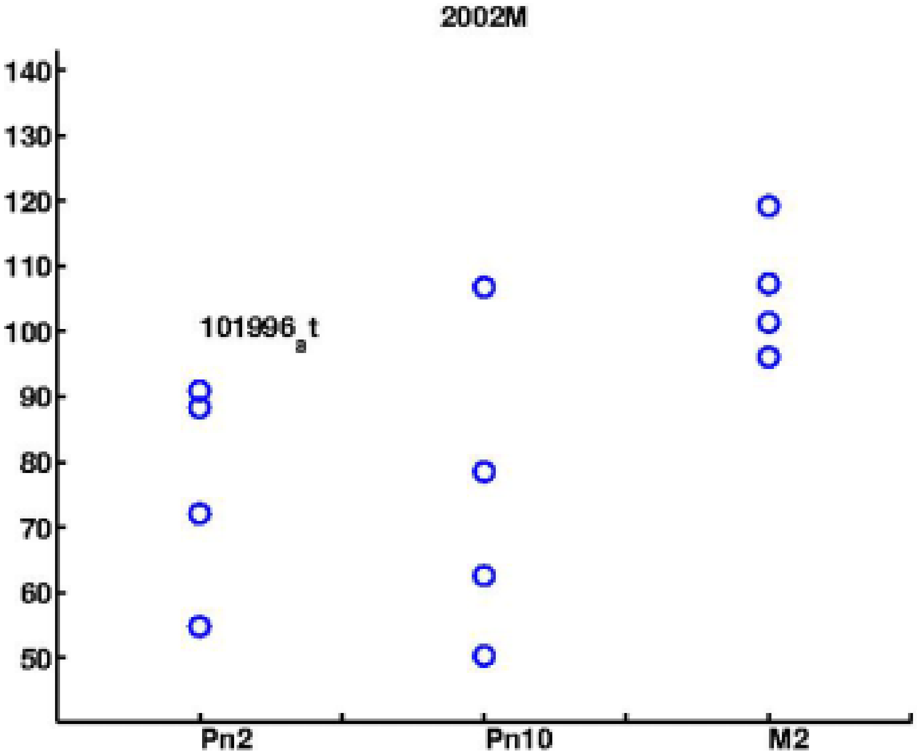
  - Paired comparisons for two treatments over T time points

$$\boxed{\xi_1(g) = \overline{K}_1(g) - \overline{W}_1(g), \quad \xi_T(g) = \overline{K}_T(g) - \overline{W}_T(g)}$$

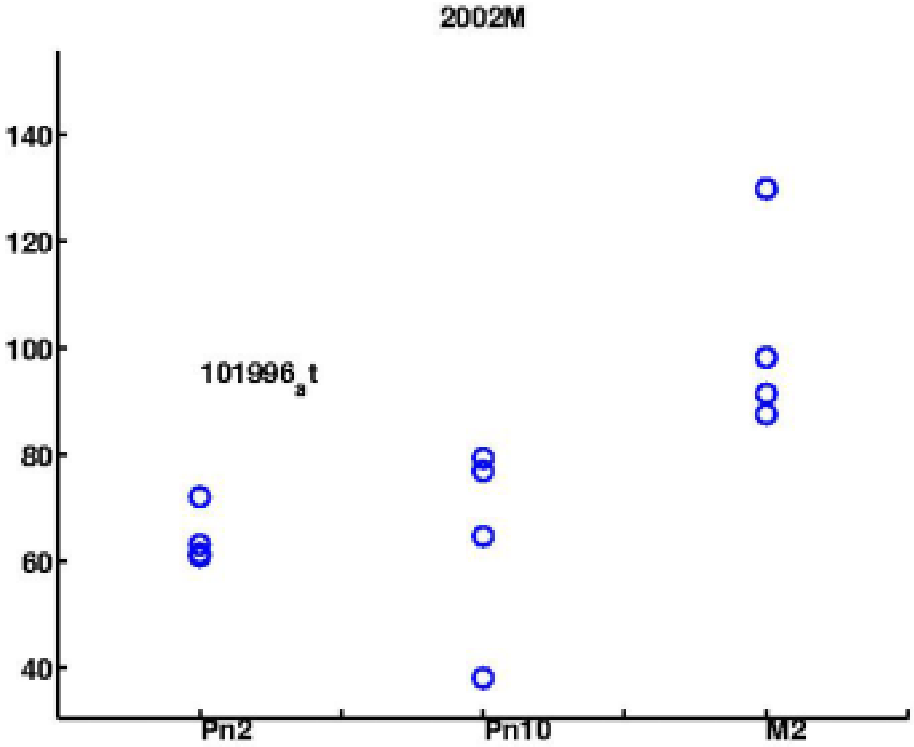
Knockout vs Wildtype Retina Study

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

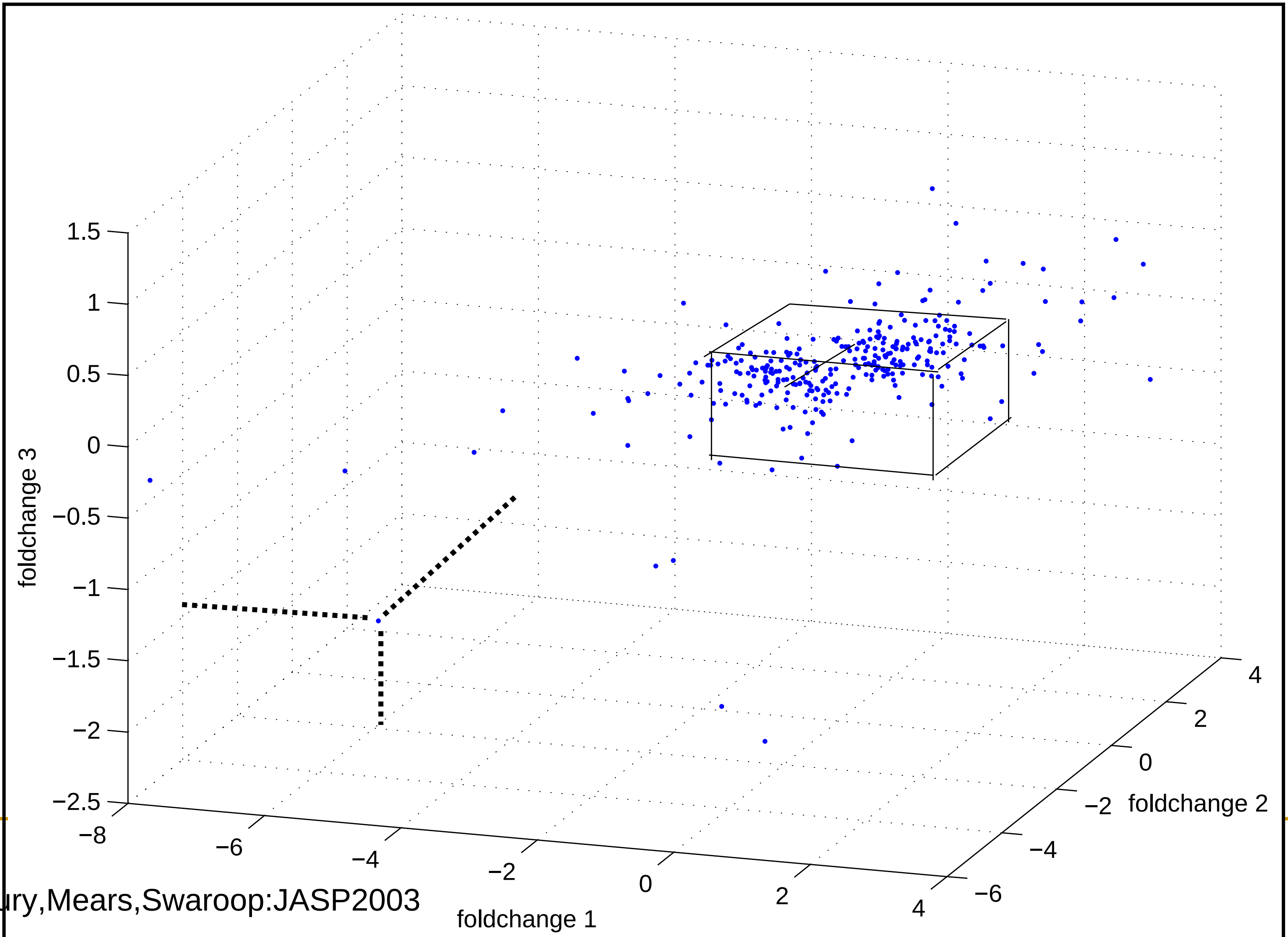
$$\mathsf{fc}(g) \neq 0$$

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

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

Hero,Fleury,Mears,Swaroop:JASP2003

Biological and Statistical Significance: Minimum Foldchange Cube



Hero,Fleury,Mears,Swaroop:JASP2003

5. 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:

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

- 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

5.1 Single-Comparison: Paired t statistic

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

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

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

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

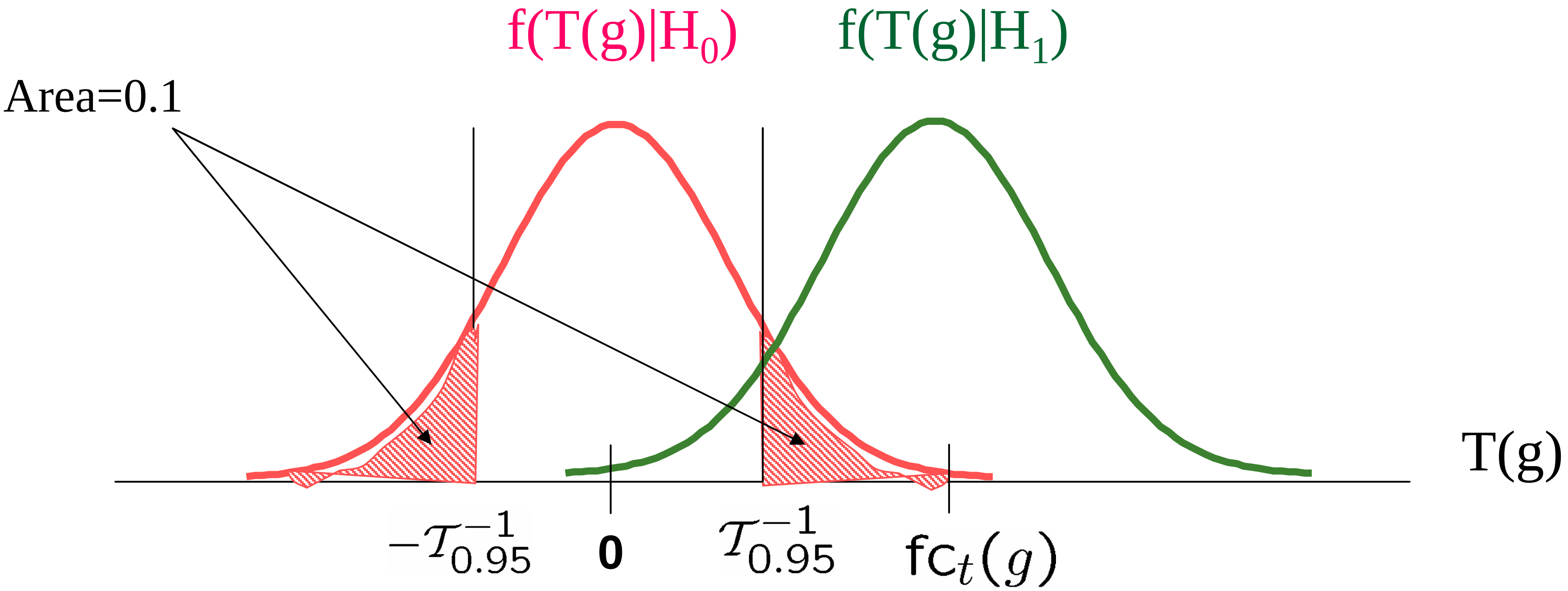
- Level 1-α onfidence interval (CI) on fc:

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

- 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

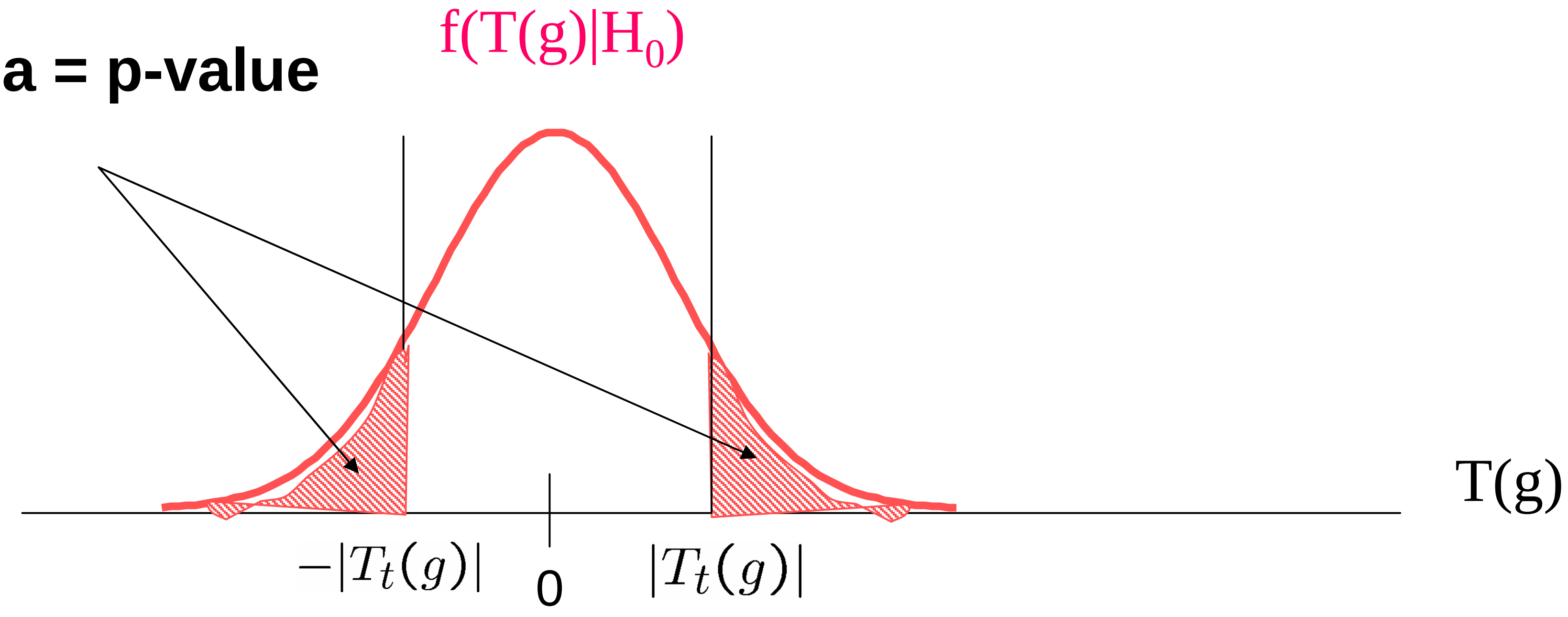
$$
\begin{aligned}
H_0 &: \mathsf{fc}_t(g) = 0 \\
H_1 &: \mathsf{fc}_t(g) \neq 0
\end{aligned}
$$



For single comparison: a false positive occurs with probability α=0.1

Stage 1: p-value of paired T test

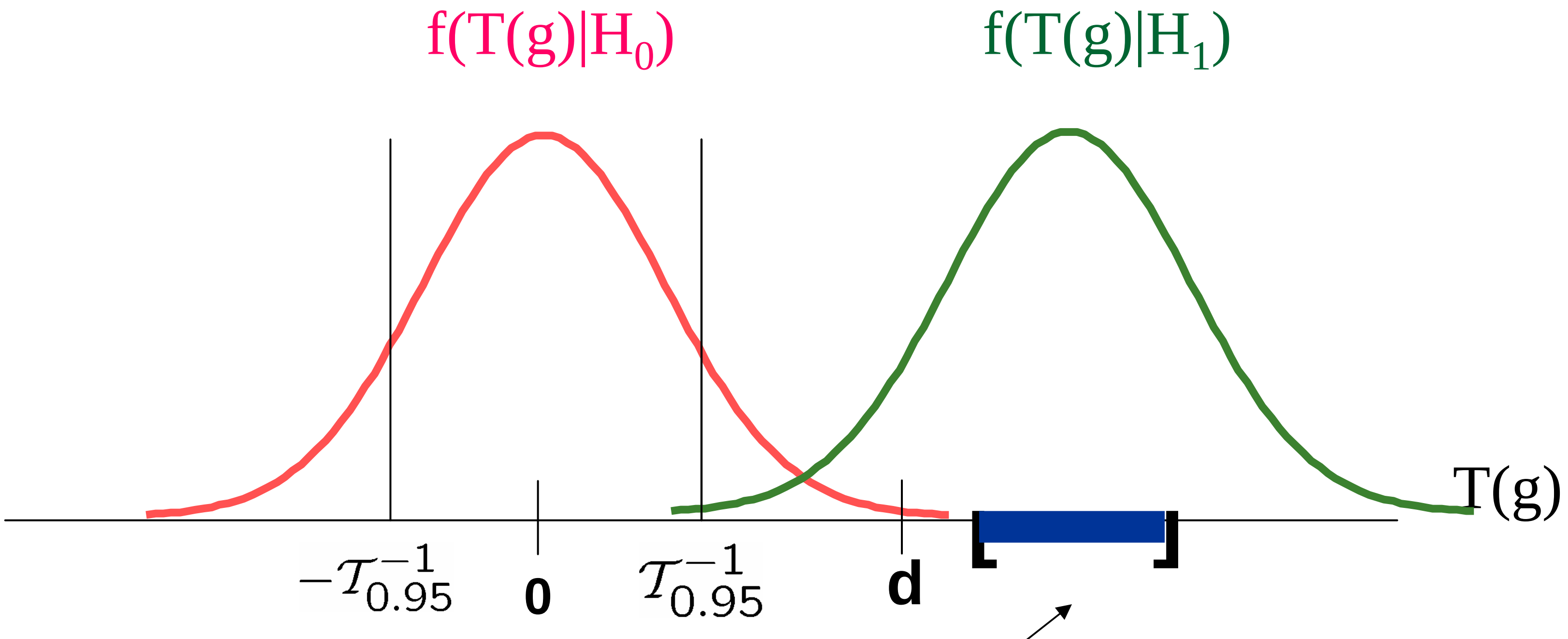
Area = p-value



In gene screening would like
p-value to be as low as possible!

Stage 2: Confidence Intervals

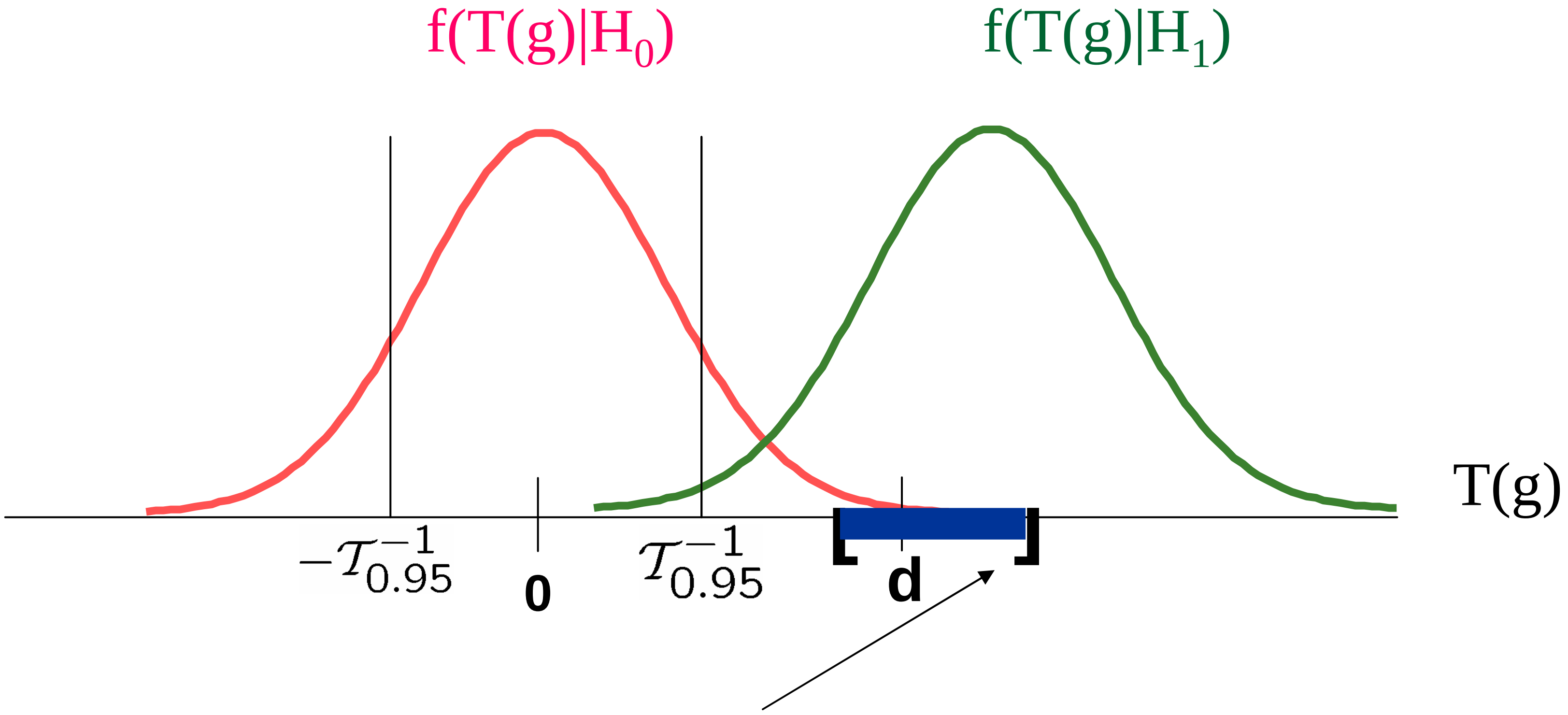
- Biologically&statistically **significant** differential response



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

Stage 2: Confidence Intervals

- Biologically&statistically **insignificant** differential response



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

P-value, FWER, FDR and FDRCI

- **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)

$$\boxed{\mathrm{FWER}(\mathcal{G}_0) = 1 - E\left[\prod_{g=1}^{G} [1 - \phi(g)]\psi_{\mathcal{G}_0}(g)\right]}$$

  - FDR: false discovery rate (Benjamini&Hochburg:1996)

$$\boxed{\mathrm{FDR}(\mathcal{G}_0) = E\left[\frac{\sum_{g=1}^{G} \phi(g)\psi_{\mathcal{G}_0}(g)}{\sum_{g=1}^{G} \phi(g)}\right]}$$

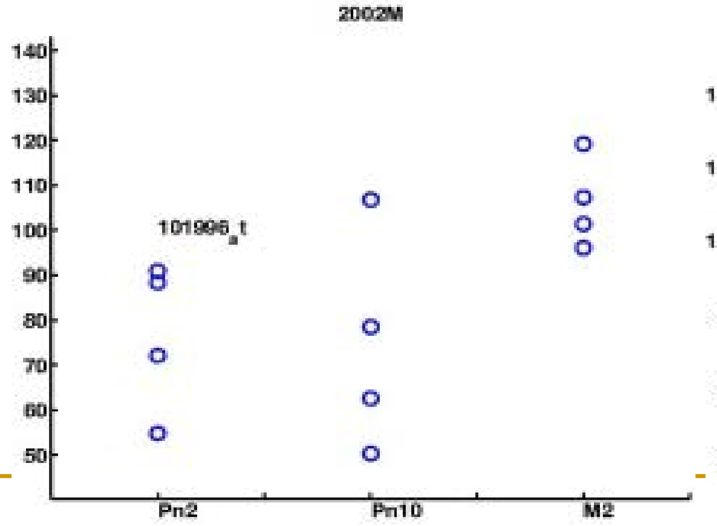
  - FDRCI: (1- α)CI on discovered fc (Benjamini&Yekutieli:2002)

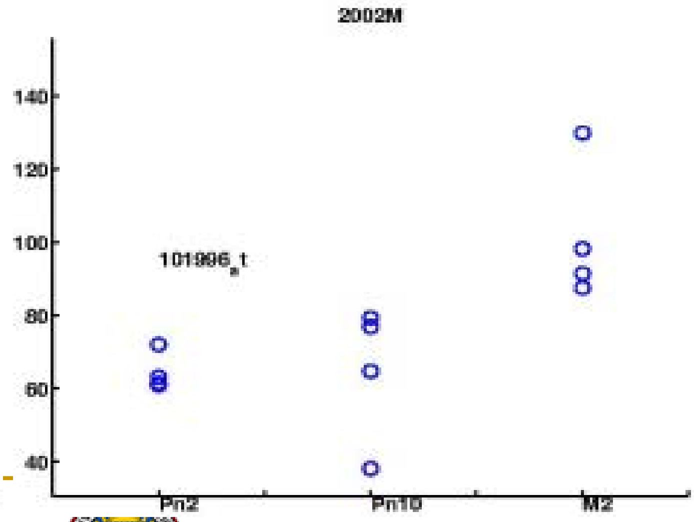
$$\boxed{\mathrm{fc}(g) \in I_g\left(\alpha\frac{P}{G}\right)}$$

  - P: number of genes discovered at FDR=α
  - $I_g(\alpha)$ standard level 1- α CI on $\mathrm{fc}(g)$

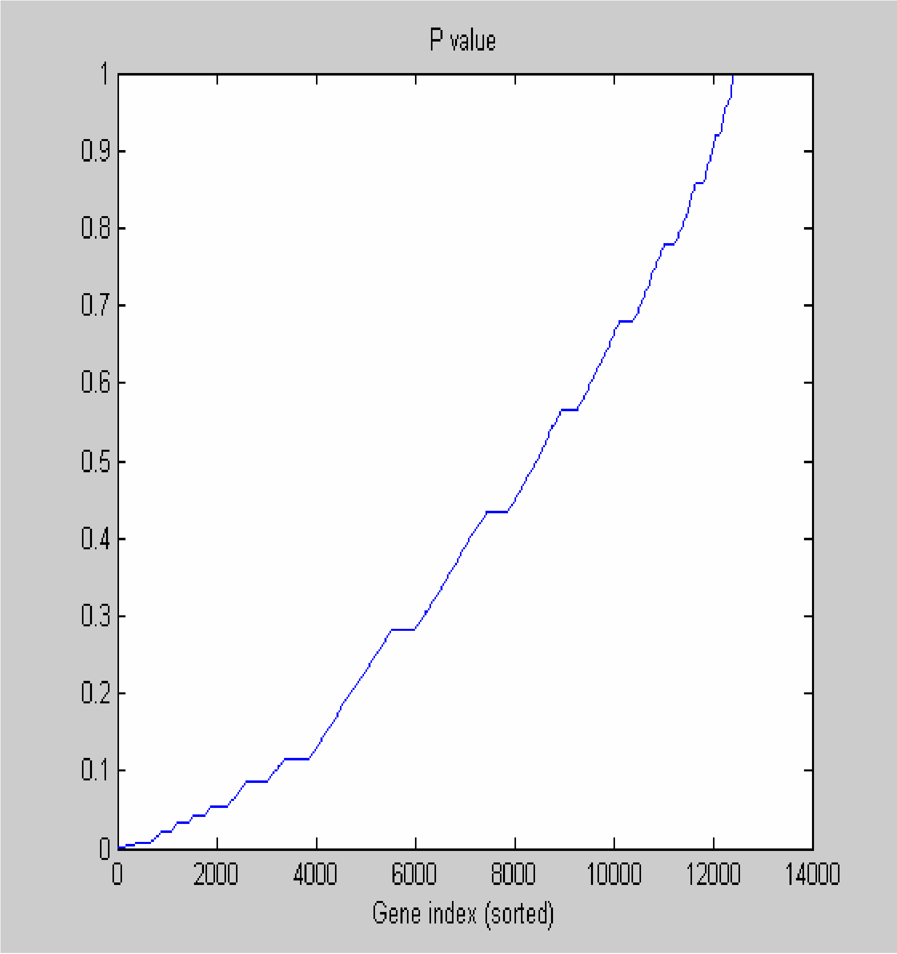
5.2 Gene Filtering: Multiple Comparisons

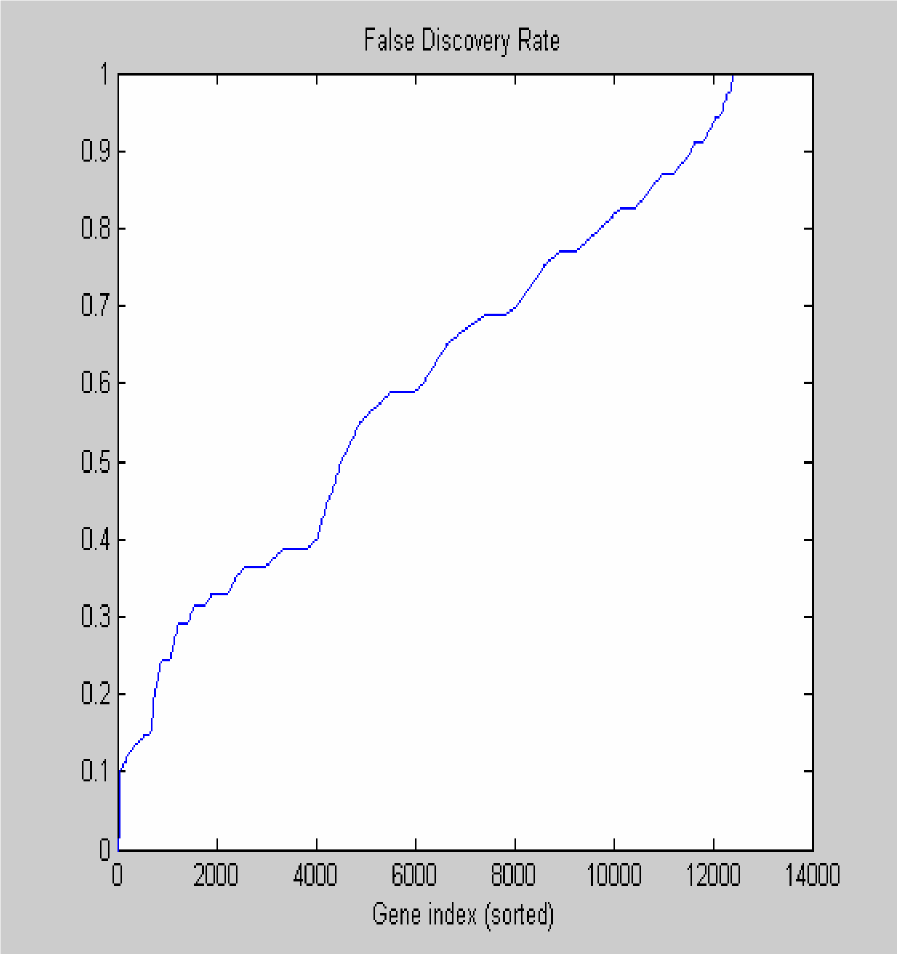
1. Find p-values of maxPT statistic over g=1…G
2. Convert p-value to FDR over g=1…G
3. Construct FDR adjusted CI’s for each t,g
4. Implement FDRCI test for MAD





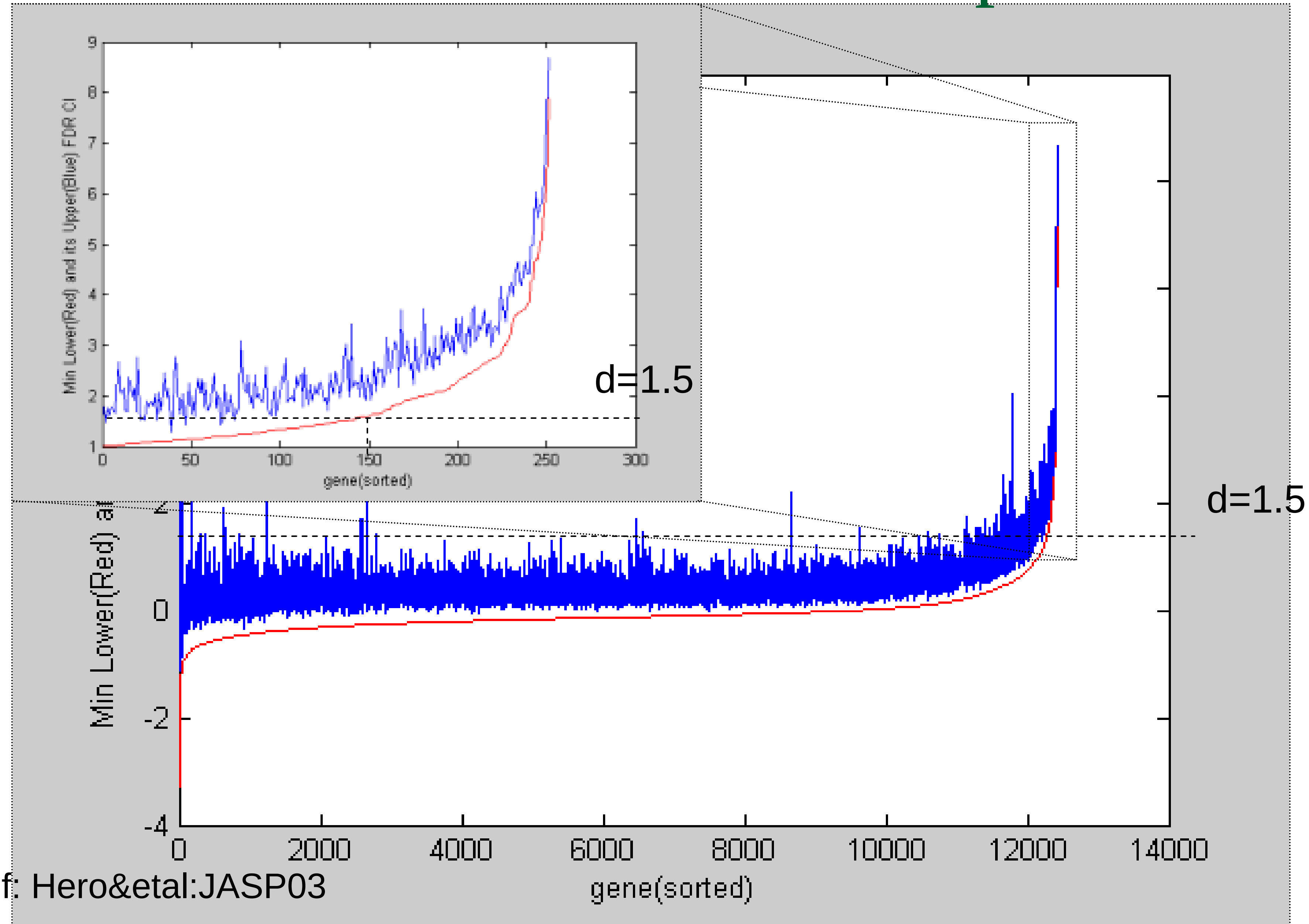
P-value vs FDR Comparison for wt/ko





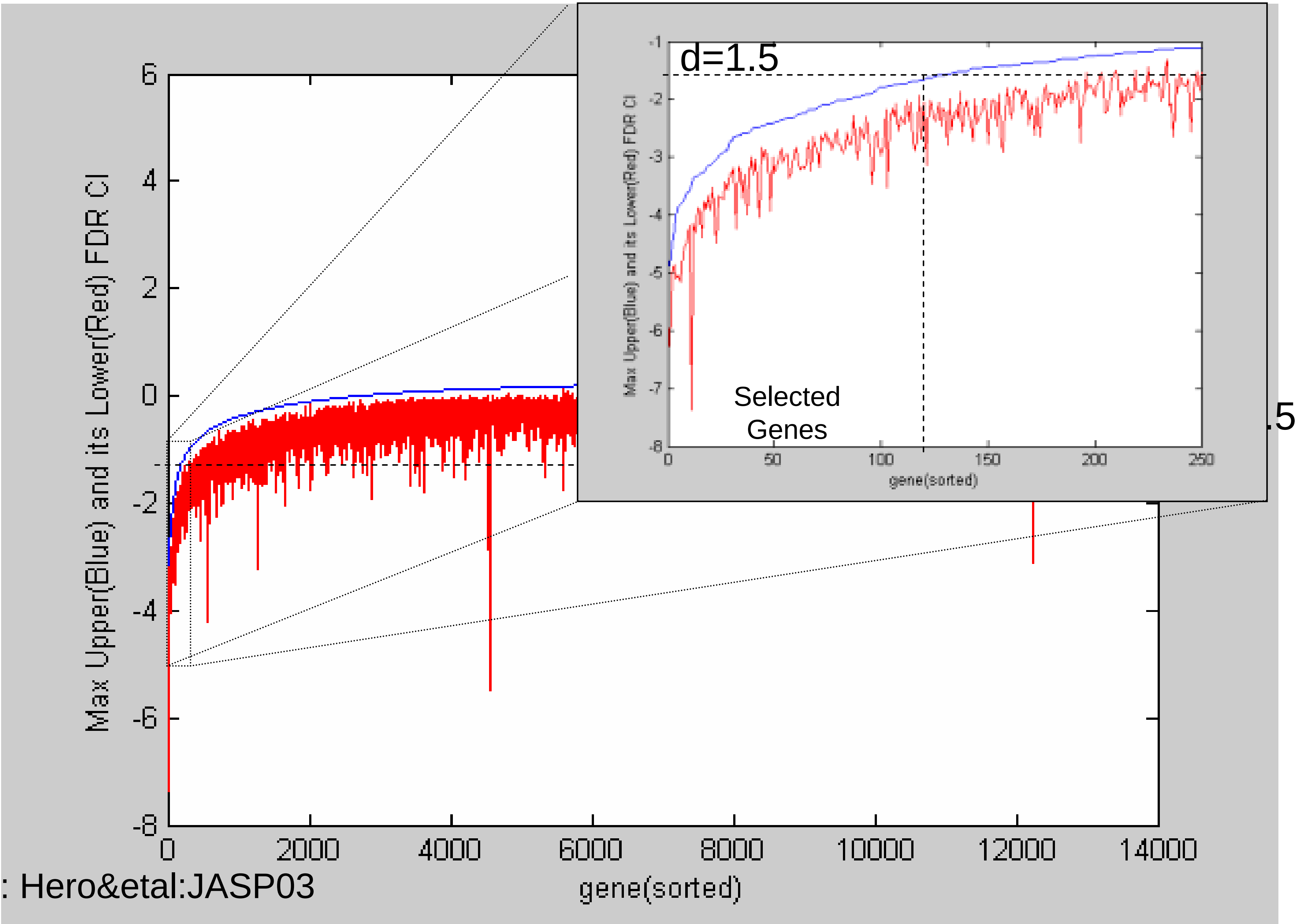
Ref: Hero&etal:JASP03

FDRCI Results for wt/ko Experiment



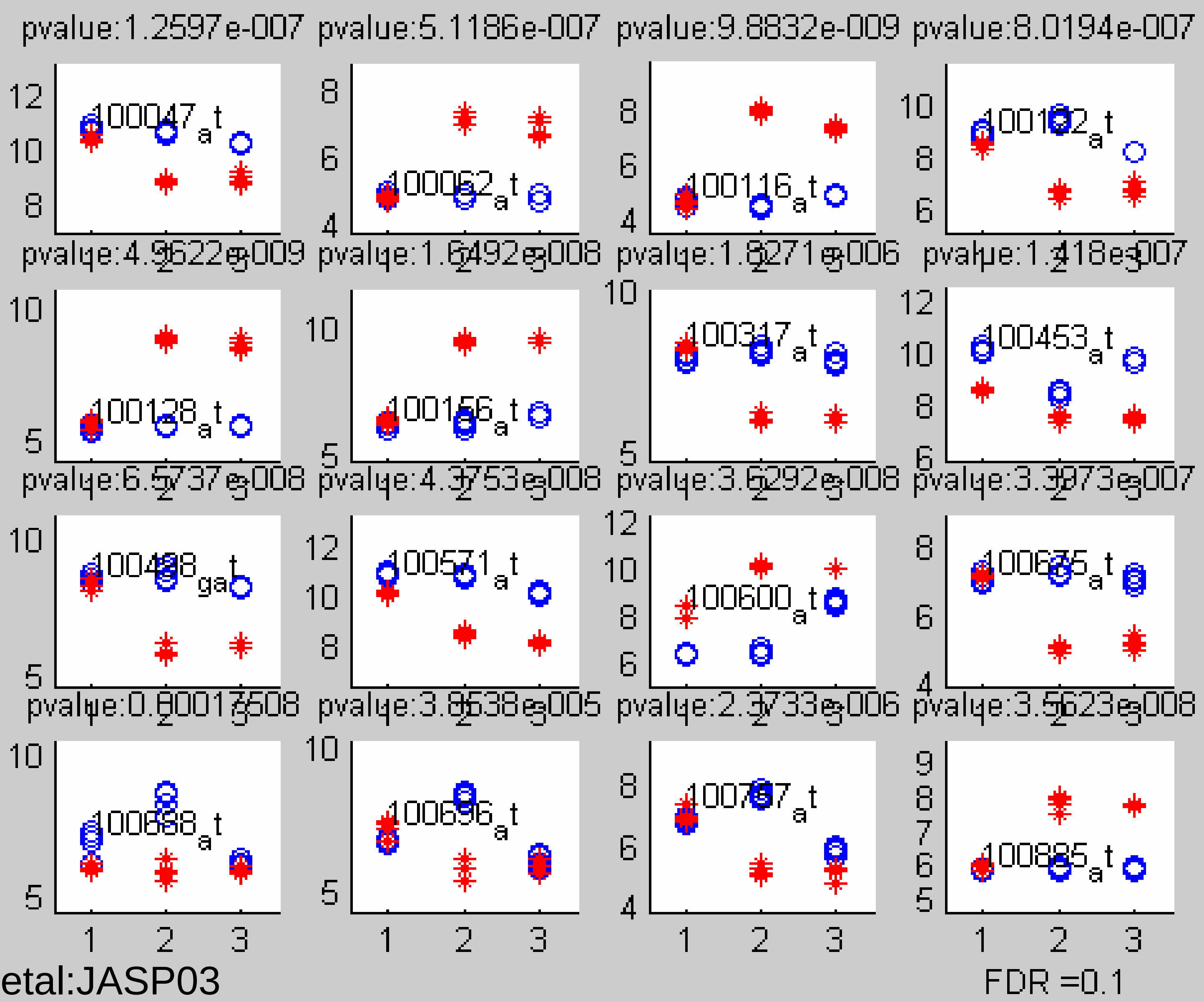
Ref: Hero&etal:JASP03

FDRCI Results for NRL Data



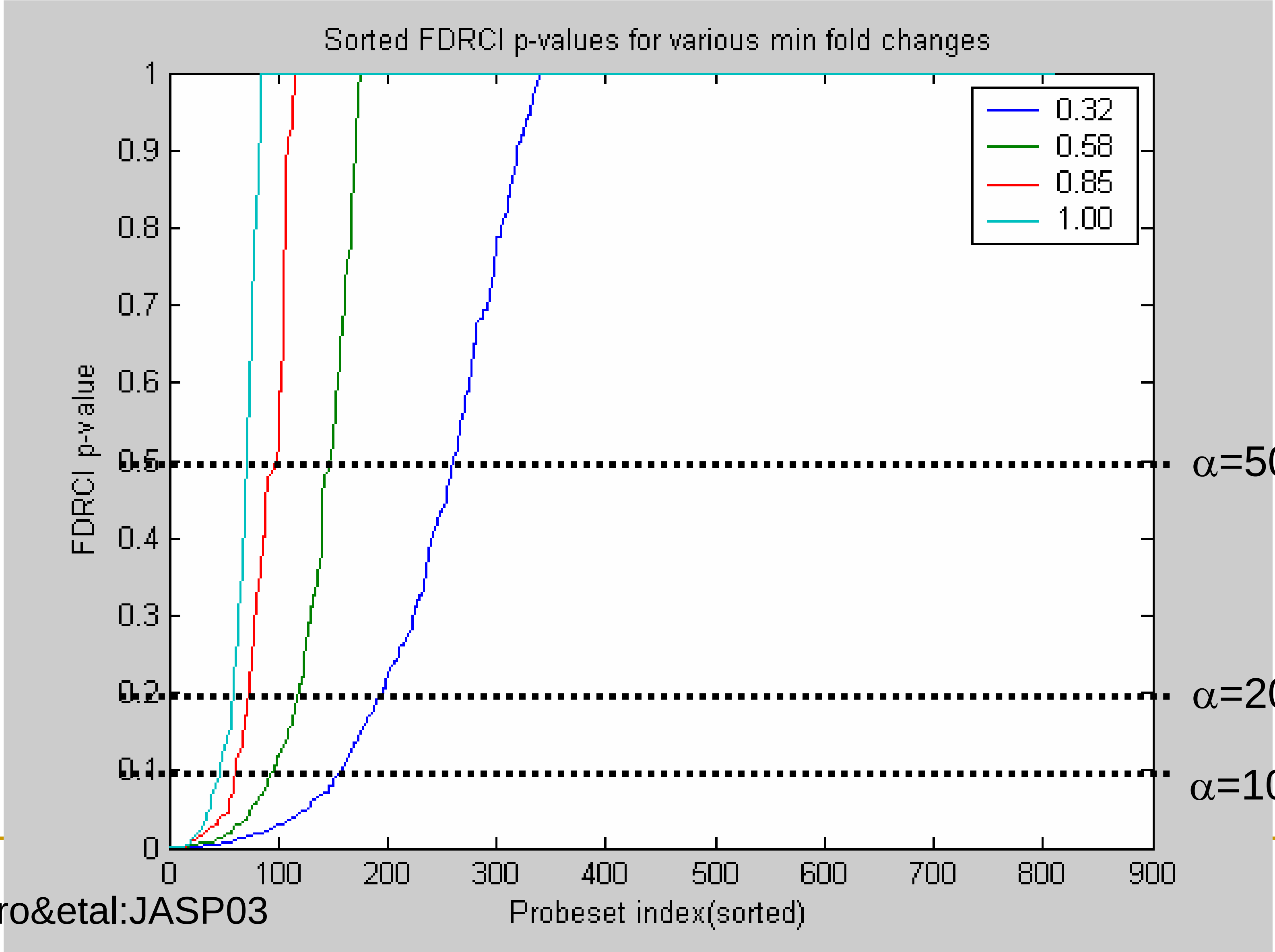
Ref: Hero&etal:JASP03

FDRCI Results for NRL Data



Ref: Hero&etal:JASP03

Sorted FDRCI pvalues for ko/wt study



α=50%

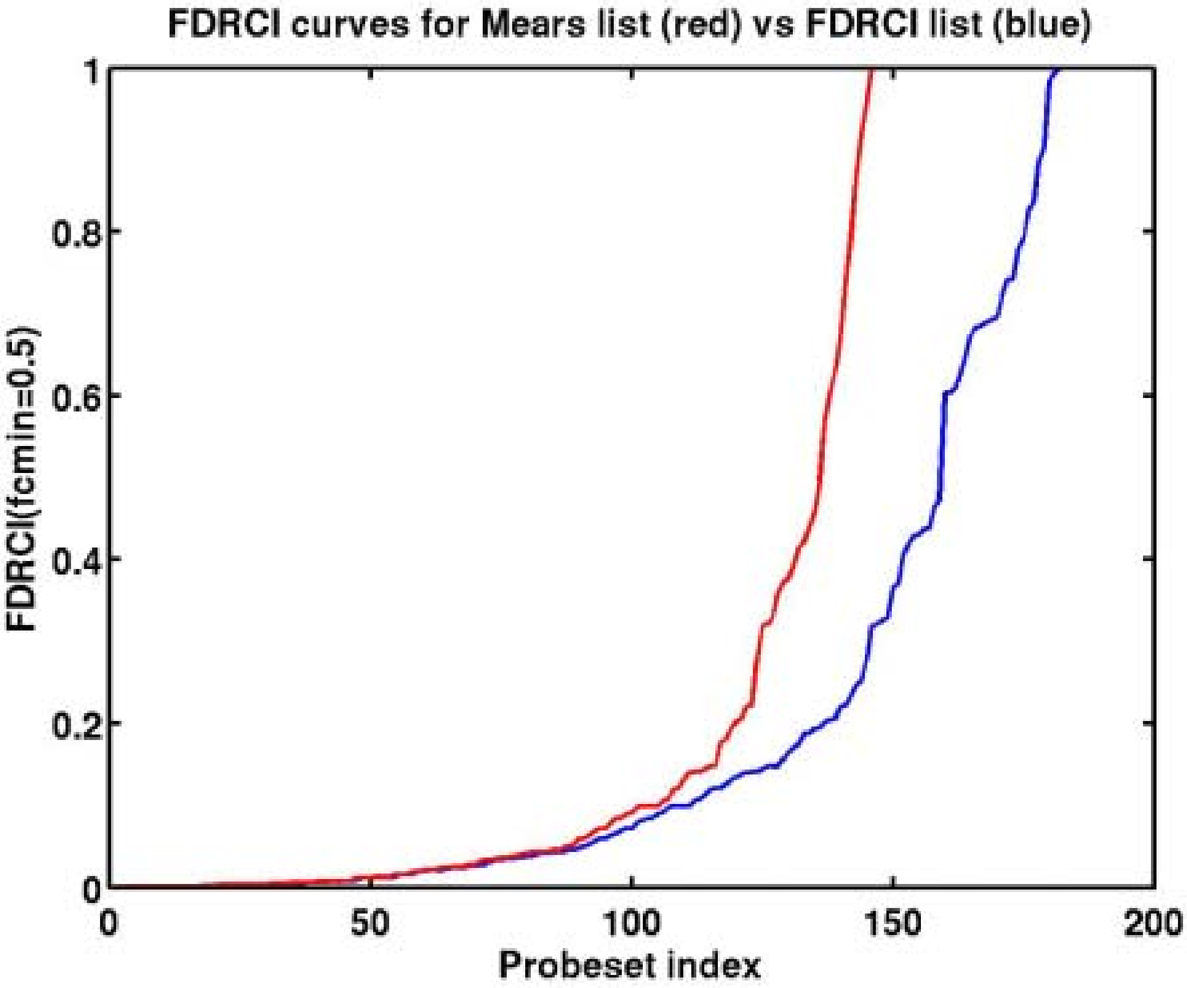
α=20%

α=10%

Ref: Hero&etal:JASP03

Mears probes	FDRCI@0.5l	FDRCI probes	FDRCI@0.5l
'92237_at'	0	'92237_at'	0
'160893_at	0	'160893_at	0
'96134_at'	0	'96134_at'	0
'96567_at'	0	'96567_at'	0
'162287_r_	0	'162287_r_	0
'94701_at'	0	'94701_at'	0
'98807_at'	0	'98807_at'	0
'95389_at'	0	'95389_at'	0
'99395_at'	0	'99395_at'	0
'94853_at'	0	'94853_at'	0
'93453_at'	0	'93453_at'	0
'102151_at	0	'102151_at	
'94139_at'	0	'94139_at'	
'98531_g_a	0	'98531_g_a	
'93330_at'	0	'93330_at'	
'96920_at'	0	'96920_at'	
'98498_at'	0	'98498_at'	
'98499_s_a	0	'98499_s_a	
'104592_i_	0	'104592_i_	
'103198_at	0	'103198_at	
'98427_s_a	0	'98427_s_a	
'104346_at	0	'104346_at	
'94150_at'	0	'94150_at'	
'161871_f_	0	'161871_f_	
'98918_at'	0	'98918_at'	
'95755_at'	0	'95755_at'	
'160754_at	0	'160754_at	
'95356_at'	0	'95356_at'	
'98957_at'	0	'98957_at'	
'99860_at'	0	'99860_at'	
'93533_at'	0	'93533_at'	
'161525_f_a	0.01	'161525_f_a	
'101855_at	0.01	'101855_at	
'162167_f_	0.01	'162167_f_	
'98967_at'	0.01	'93699_at'	
'102682_at	0.01	'98967_at'	
'160828_at	0.01	'102682_at	
'104591_g_	0.01	'160828_at	
'104643_at	0.01	'104591_g_	
'93482_at'	0.01	'104643_at	
'101923_at	0.01	'93482_at'	
'103895_at	0.01	'101923_at	
'93094_at'	0.01	'103895_at	
'103038_at	0.01	'93094_at'	

'96831_at'	0.01	'103038_at	0.01
'98852_at'	0.01	'96831_at'	0.01
'99238_at'	0.01	'98852_at'	0.01
'101344_at	0.01	'99238_at'	0.01
'92796_at'	0.01	'101344_at	0.01
'93290_at'	0.01	'92796_at'	0.01
'100696_at	0.01	'93290_at'	0.01
'100453_at	0.01	'100696_at	0.01
'98560_at'	0.01	'100453_at	0.01
'102890_at	0.01	'98560_at'	0.01
'95363_at'	0.02	'102890_at	0.01



Ref: Hero&etal:JASP03

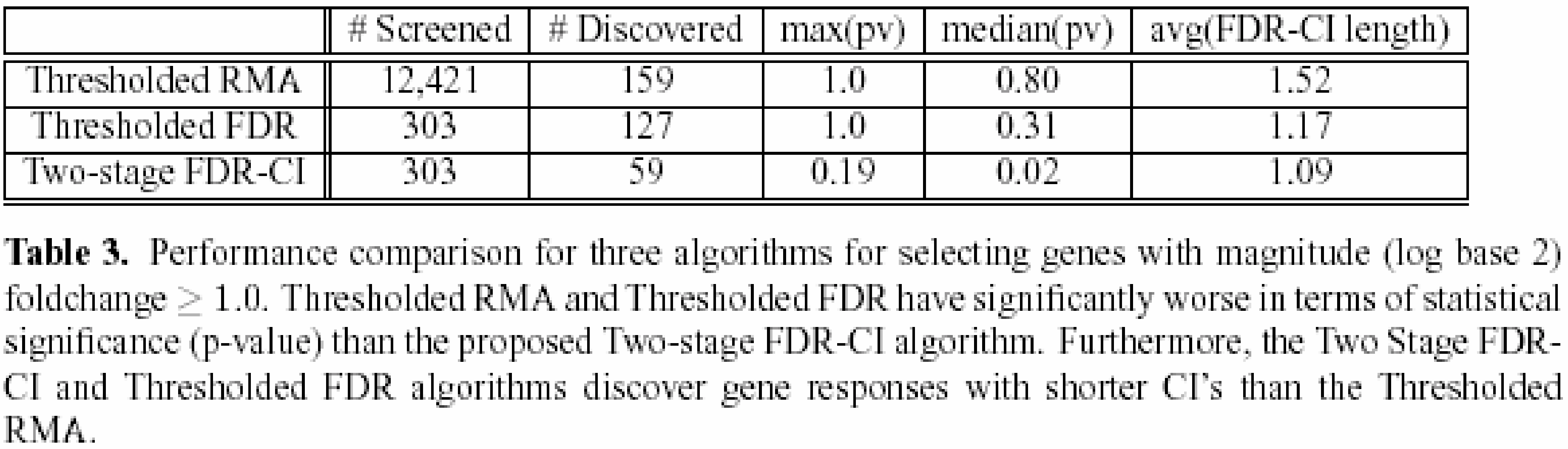
'96518_at'	0.06	'95541_at'	0.05
'93328_at'	0.06	'103033_at	0.05
'160597_at	0.06	'93269_at'	0.05
'97890_at'	0.07	'97381_s_a	0.06
'93731_at'	0.07	'96518_at'	0.06
'93887_at'	0.07	'93328_at'	0.06
'92232_at'	0.08	'160597_at	0.06
'103456_at	0.08	'103241_at	0.07
'104564_at	0.09	'97890_at'	0.07
'102292_at	0.09	'93731_at'	0.07
'104374_at	0.09	'93887_at'	0.07
'95105_at'	0.1	'92232_at'	0.08
'104206_at	0.1	'100026_at	0.08
'96596_at'	0.1	'103456_at	0.08
'97722_at'	0.1	'104564_at	0.09
'99972_at'	0.1	'102292_at	0.09
'160948_a	0.11	'104374_at	0.09
'94393_r_a	0.11	'95105_at'	0.1
'92534_at'	0.12	'104206_at	0.1
'97770_s_a	0.12	'96596_at'	0.1
'160464_s_	0.13	'97722_at'	0.1
'94739_at'	0.14	'99972_at'	0.1
'93268_at'	0.14	'160948_a	0.11
'96354_at'	0.14	'94393_r_a	0.11
'101151_at	0.14	'94872_at'	0.11
'97357_at'	0.15	'92534_at'	0.12
'97755_at'	0.15	'94733_at'	0.12
'95603_at'	0.18	'97770_s_a	0.12
'93669_f_a	0.18	'99014_at'	0.13
'97124_at'	0.19	'160464_s_	0.13
'98993_at'	0.2	'93412_at'	0.14
'104104_a	0.21	'102413_at	0.14
'99623_s_a	0.22	'94739_at'	0.14
'104761_a	0.22	'93268_at'	0.14
'93202_at'	0.28	'96354_at'	0.14
'92770_at'	0.32	'101151_at	0.14
'98111_at'	0.32	'97357_at'	0.15
'160808_a	0.33	'97755_at'	0.15
'98524_f_a	0.36	'101044_at	0.15
'101308_at	0.37	'101861_at	0.16
'104388_at	0.38	'93389_at'	0.16
'103460_at	0.39	'96766_s_a	0.17
'97579_f_a	0.42	'95603_at'	0.18
'103026_f_	0.42	'95285_at'	0.19
'100757_at	0.44	'98544_at'	0.19

'98005_at'	0.46	'97124_at'	0.19
'104469_a	0.5	'93130_at'	0.2
'103922_f_	0.57	'98993_at'	0.2
'92607_at'	0.6	'102352_at	0.2
'104171_f_	0.63	'104104_a	0.21
'96156_at'	0.67	'99623_s_a	0.22
'96586_at'	0.74	'104761_a	0.22
'101702_at	0.79	'98329_at'	0.24
'93457_at'	0.86	'99586_at'	0.25
'160894_a	0.92	'99461_at'	0.25
'104299_at	0.96	'98569_at'	0.28
'100348_at	1	'92770_at'	0.32
'100688_a	1	'102835_at	0.32
'101465_at	1	'93354_at'	0.33
'102393_at	1	'160808_a	0.33
'104518_at	1	'97732_at'	0.37
'160610_a	1	'160937_at	0.37
'160901_a	1	'95397_at'	0.41
'93391_at'	1	'94258_at'	0.42
'93606_s_a	1	'101191_a	0.43
'94255_g_a	1	'101489_a	0.43
'97142_at'	1	'100757_at	0.44
'98004_at'	1	'95453_f_a	0.44
'99126_at'	1	'93011_at'	0.46
		'160414_a	0.47
		'104743_at	0.6
		'93045_at'	0.6
		'101886_f_	0.61
		'94713_at'	0.63
		'101027_s_	0.65
		'94514_s_a	0.67
		'162237_f_	0.68
		'95555_at'	0.69
		'94270_at'	0.69
		'93191_at'	0.69
		'104217_at	0.7
		'93120_f_a	0.72
		'102317_at	0.74
		'98554_at'	0.74
		'93972_at'	0.78
		'99559_at'	0.79
		'101426_at	0.83
		'103524_at	0.84
		'103279_at	0.89
		'96762_at'	0.9

Ref: Hero&etal:JASP03

Filtering: Quantitative comparisons

- Wt vs NRL ko, Affymetrix data:



	# Screened	# Discovered	max(pv)	median(pv)	avg(FDR-CI length)
Thresholded RMA	12,421	159	1.0	0.80	1.52
Thresholded FDR	303	127	1.0	0.31	1.17
Two-stage FDR-CI	303	59	0.19	0.02	1.09

Table 3. Performance comparison for three algorithms for selecting genes with magnitude (log base 2) foldchange $\geq$ 1.0. Thresholded RMA and Thresholded FDR have significantly worse in terms of statistical significance (p-value) than the proposed Two-stage FDR-CI algorithm. Furthermore, the Two Stage FDR-CI and Thresholded FDR algorithms discover gene responses with shorter CI's than the Thresholded RMA.

Ref: Hero&etal:JASP03

5.3 Gene Ranking

- 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:

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

- Examples of mixed increasing and decreasing

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

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

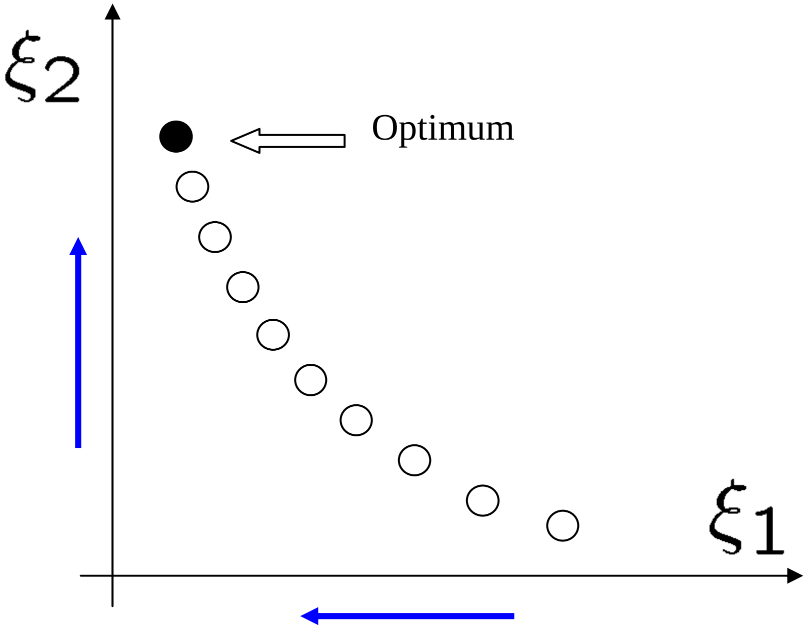
$$\{\mathsf{fc}_1(g), \mathsf{fc}_2(g), \mathsf{fc}_3(g)\}_{g \in \mathcal{G}_1} > \ldots > \{\mathsf{fc}_1(g), \mathsf{fc}_2(g), \mathsf{fc}_3(g)\}_{g \in \mathcal{G}_q}$$

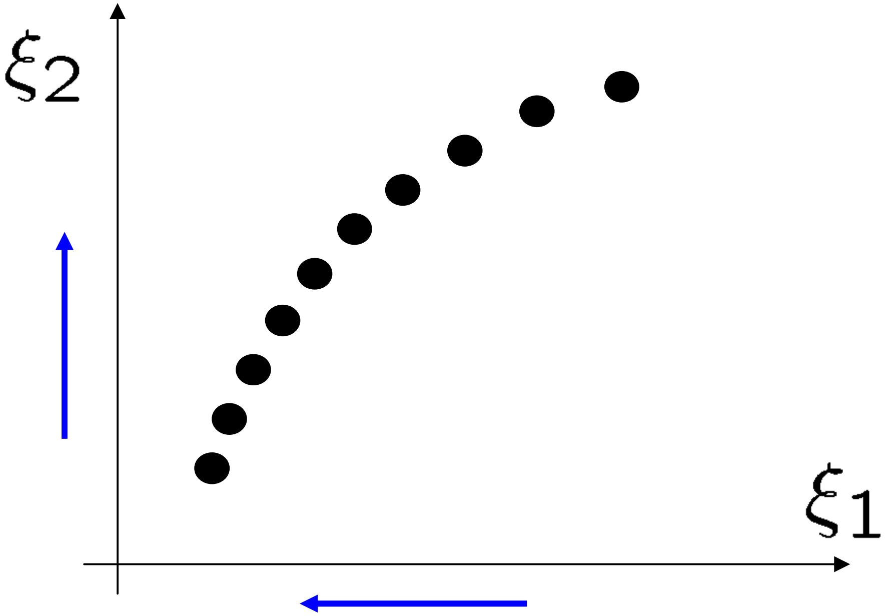
Illustration of two extreme cases

$$\xi_1 = \sqrt{(s_K^2 + s_W^2)/2} = \text{pooled sample dispersion}$$

$$\xi_2 = |\overline{K} - \overline{W}| = \text{mean treatment dispersion}$$

- A linear ordering exists
- No partial ordering exists





Comparison to Criteria Aggregation

- Assume (wolg): increasing criteria
- Linear aggregation: define preference pattern

$$\{W_p\}_{p=1}^{P}, \ \sum_{p=1}^{P} W_p = 1, \ W_p > 0$$

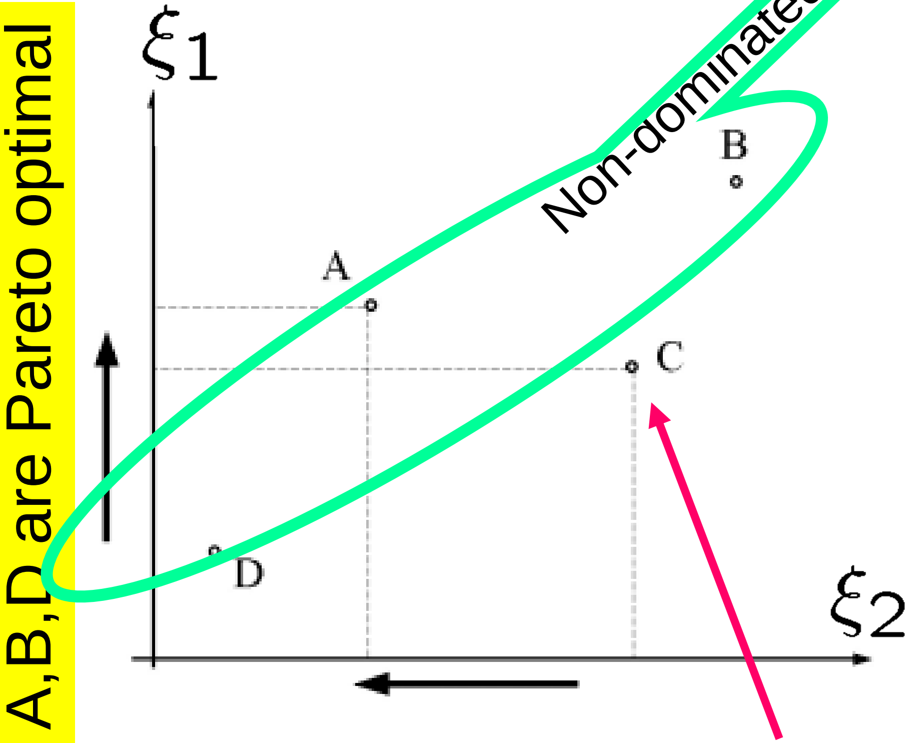
- Order genes according to ranks of

$$T(g) = \sum_{p=1}^{P} W_p \xi_p(g)$$

- Q: What are set of universally optimal genes that maximize $T(g)$ for any preference pattern?
- A: the non-dominated (Pareto optimal) genes

Multicriteria Gene Ranking

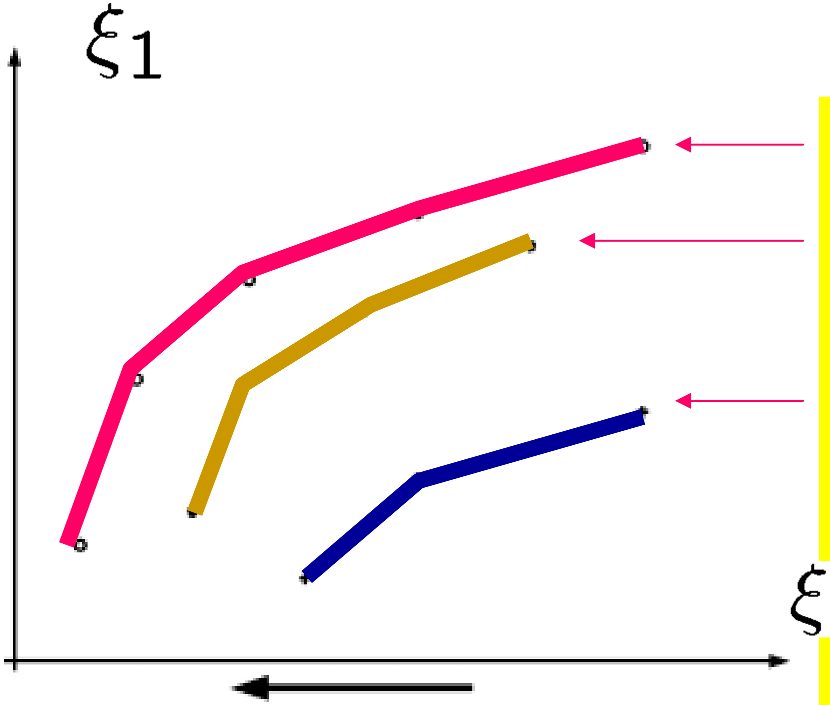
- Increasing ξ_1
- Decreasing ξ_2



Non-dominated genes=Pareto Front

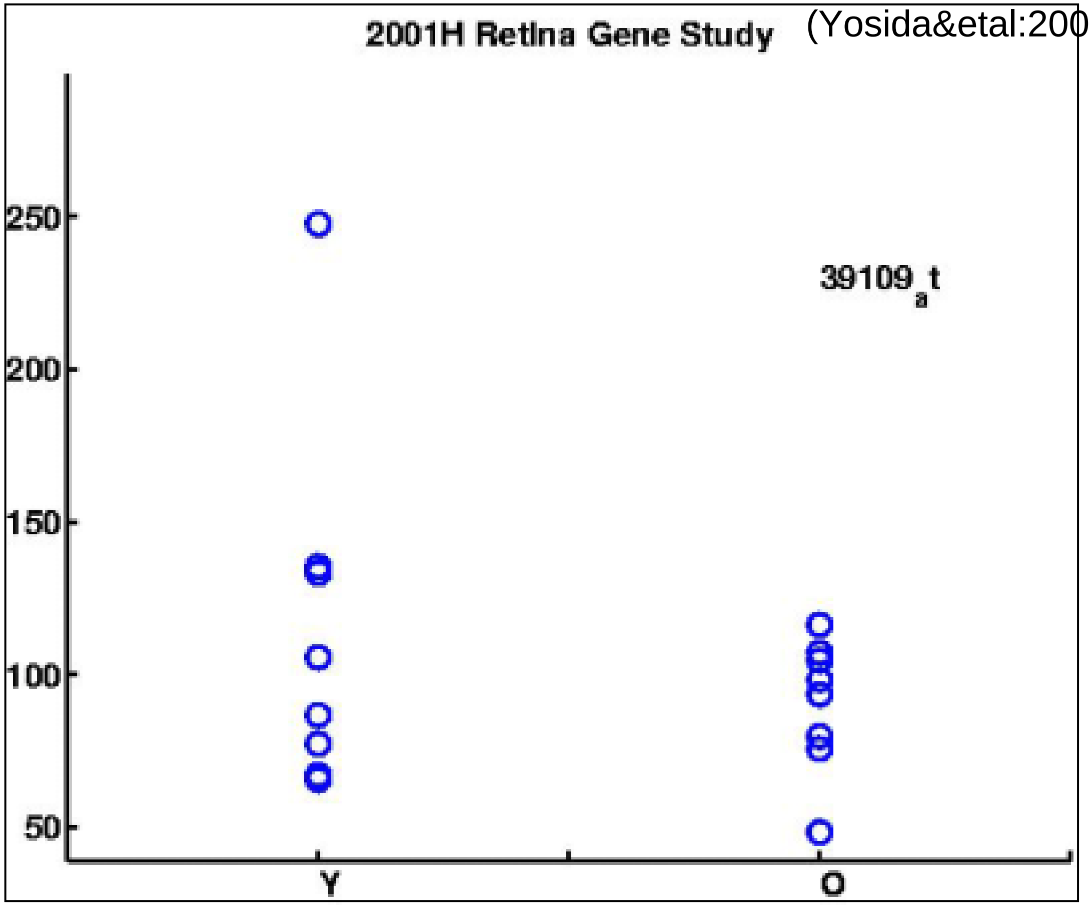
A,B,D are Pareto optimal

Dominated gene



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**

$$\xi_1(g) = \sqrt{(\sigma_O^2(g) + \sigma_Y^2(g))/2}$$

$$\xi_2(g) = |\overline{O}(g) - \overline{Y}(g)|$$

Ref: Fleury&etal ICASSP-02

Multicriteria Y/O Gene Ranking

- Paired t-test at level of significance alpha:

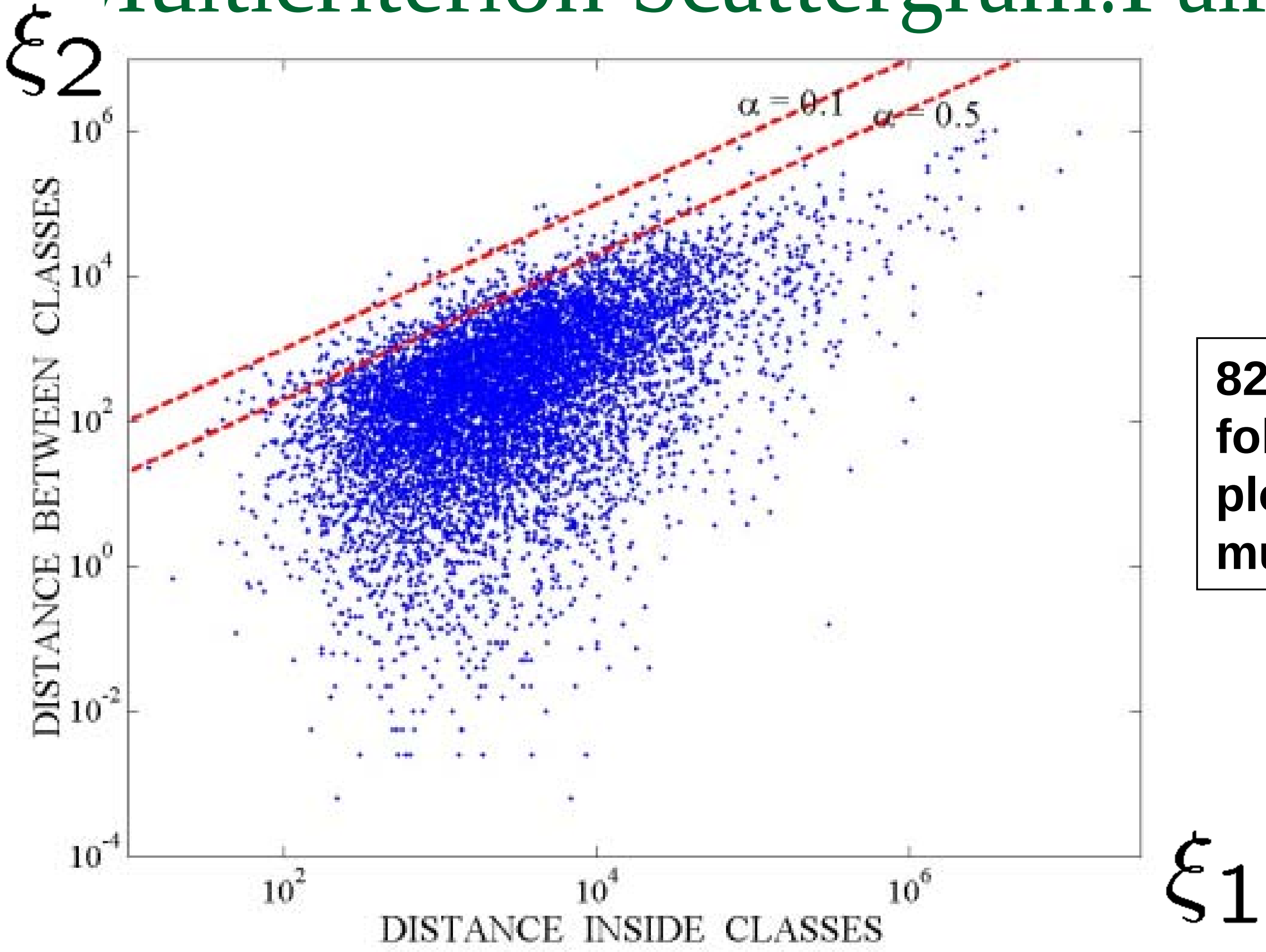
$$T(g) = \frac{\xi_2(g)}{\xi_1(g)} \quad \underset{<}{>} \quad \sqrt{2/m}\, \mathcal{T}^{-1}_{1-\alpha/2}$$

- 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

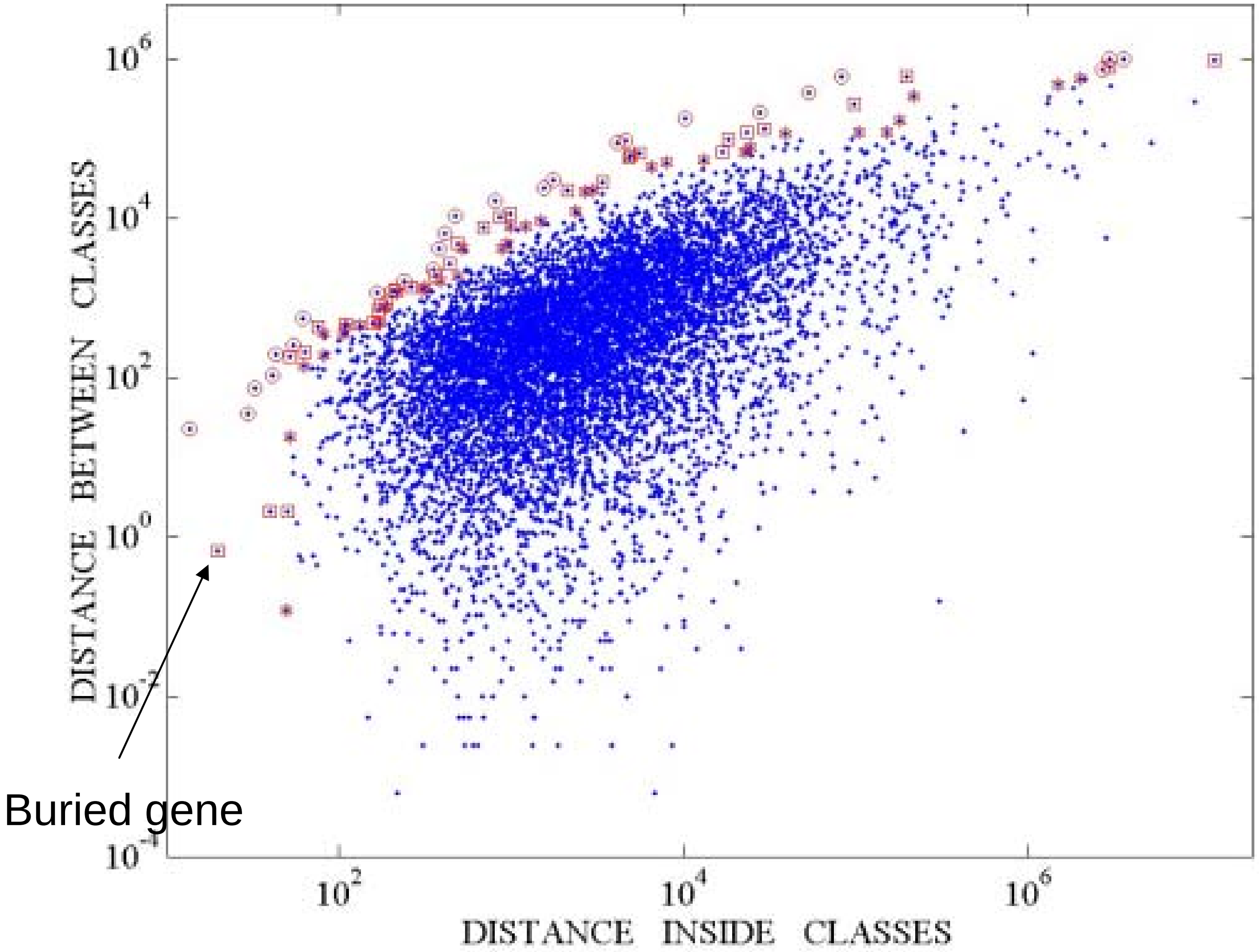
Multicriterion Scattergram:Paired t-test



8226 Y/O mean foldchanges plotted in multicriteria plane

Ref: Fleury&etal ICASSP-02

Multicriterion scattergram: Pareto Fronts



Buried gene

Pareto fronts

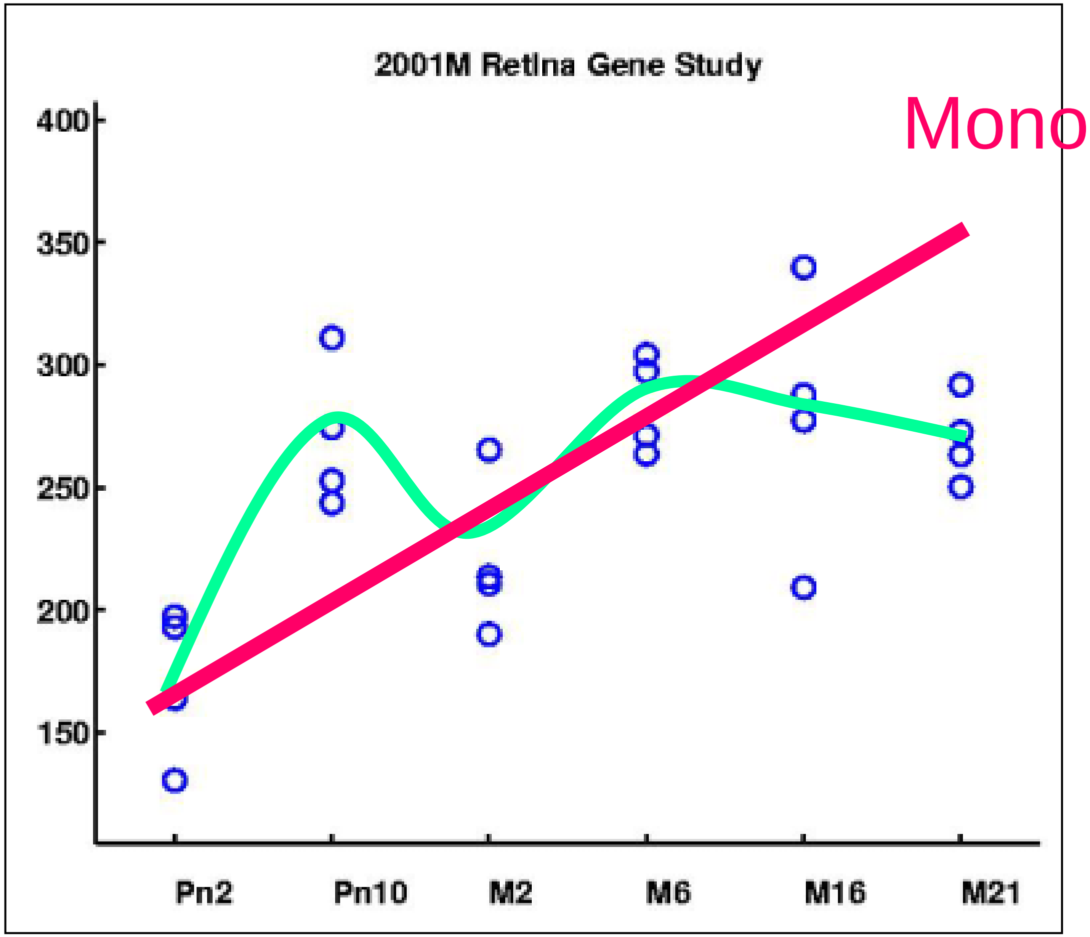
○ *first*

□ *second*

☆ *third*

Ref: Fleury&etal ICASSP-02

Ranking Based on Profile Shape



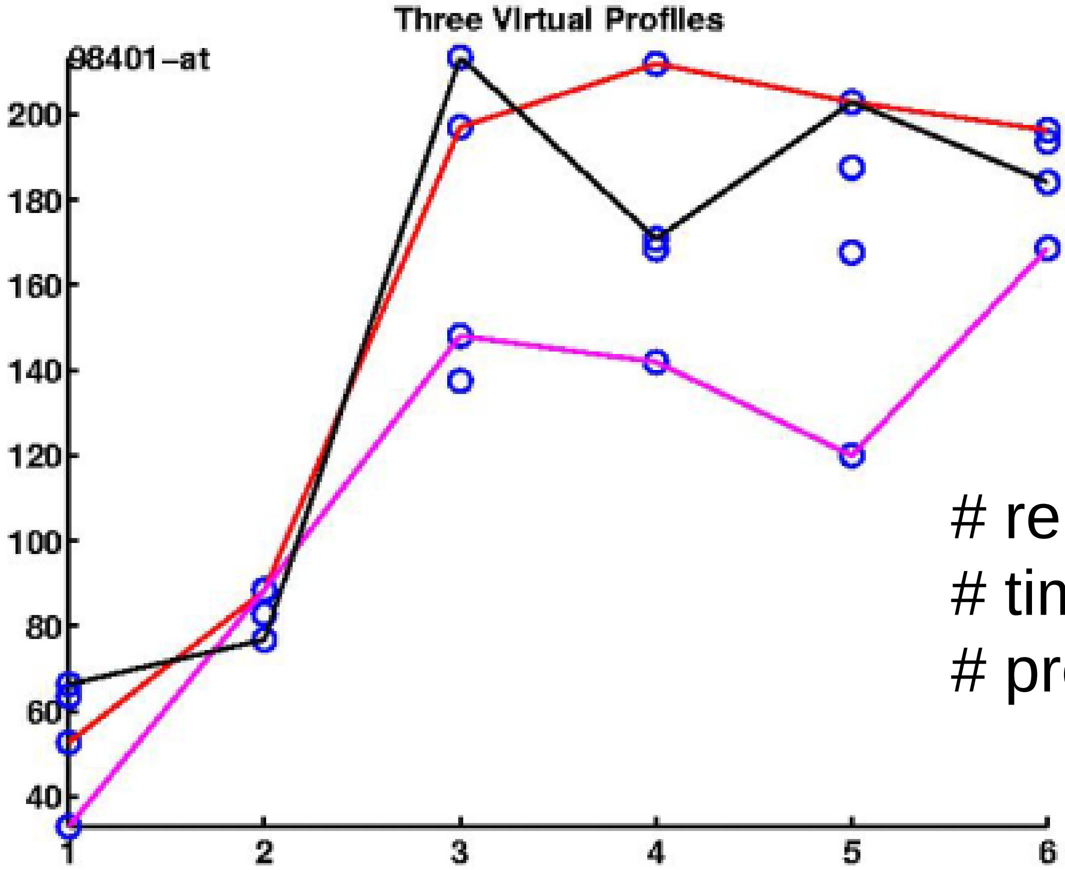
Monotonic?

Mouse Retina Aging Study

- **24 Mouse retinas**
- **6 time samples**
- **4 replicates**
- **12422 probesets**

Ref: Hero&etal:VLSI03

Jonckheere-Terpstra Statistic



# replicates=m=4
# time points=t=6
# profiles=4^6=4096

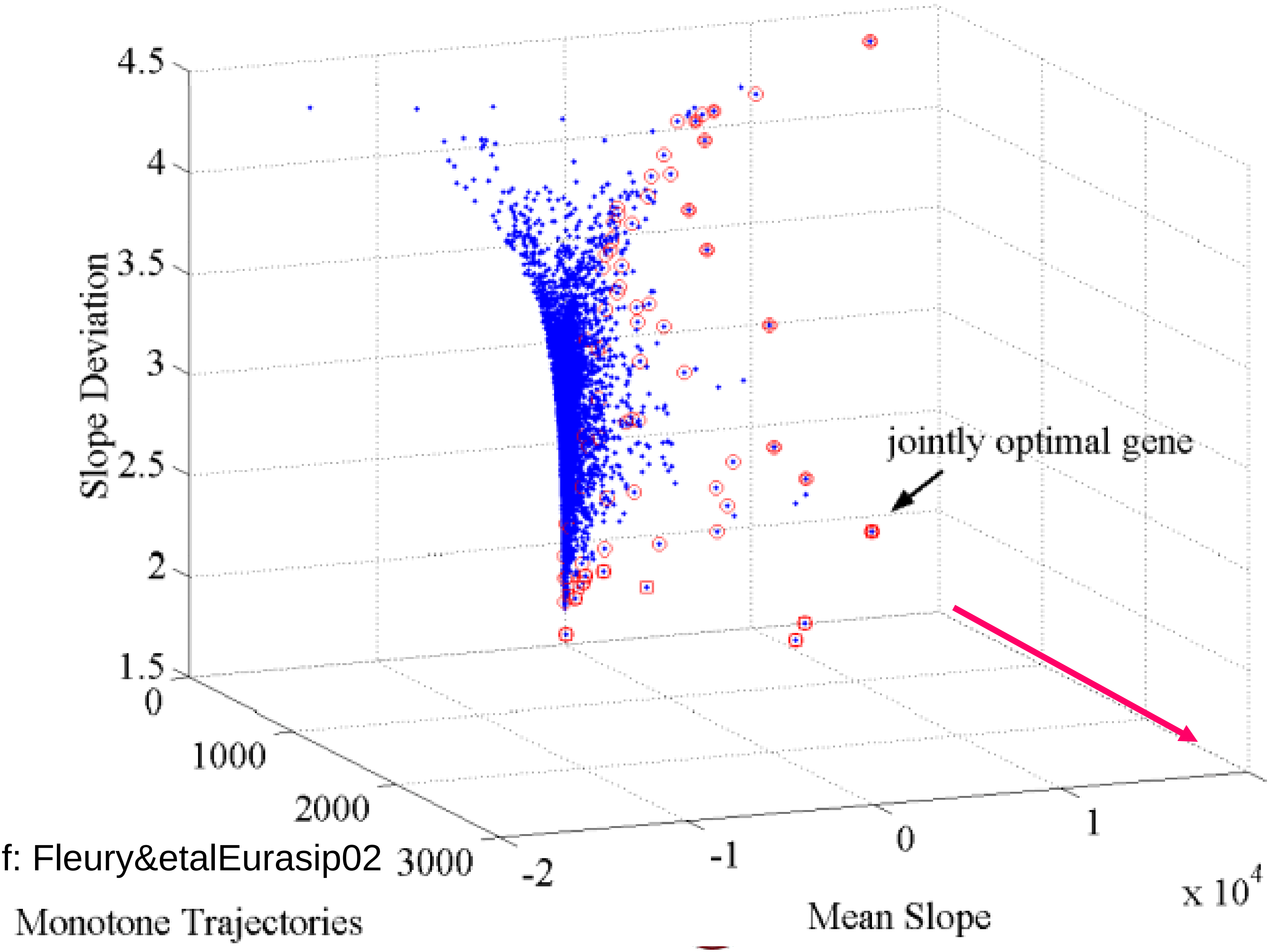
$$\xi_1(g) = \sum_{t} \sum_{t'>t} \sum_{m\neq m'} \text{sign}(y_{t',m'}(g) - y_{t,m}(g))$$

Ref: Hollander 2001

Monotonic-Profile Ranking Criteria

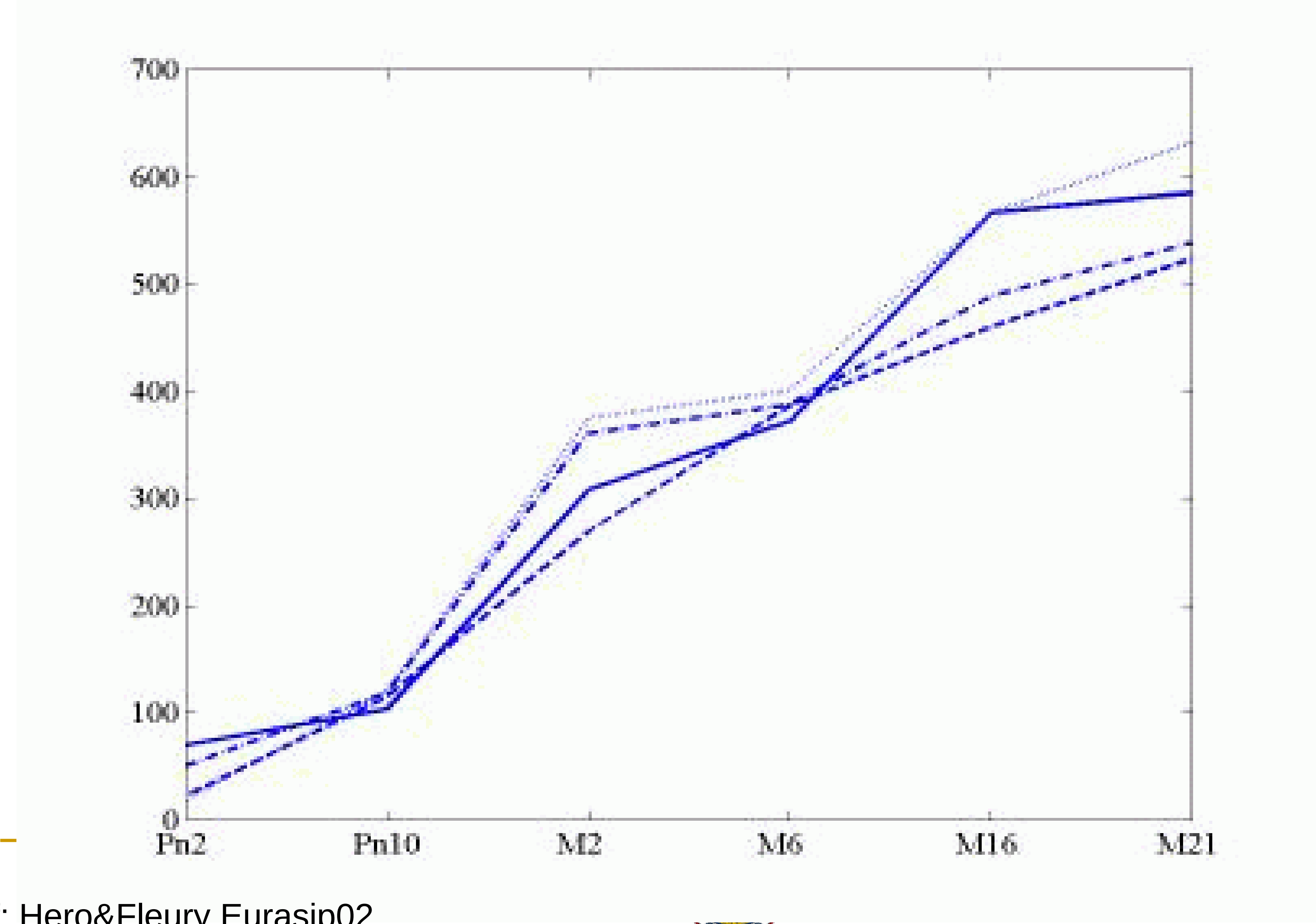
- Monotonicity: Jonckheere-Terpstra statistic
  - Large number of monotonic virtual profiles
- Curvature: Second order difference statistic
  - Small deviation from linear
- End-to-end foldchange: paired-T statistic
  - Large overall foldchange

Multicriterion Scattergram: Aging Study



Ref: Fleury&etalEurasip02

Profile of Pareto Optimal Aging Gene



Ref: Hero&Fleury Eurasip02

Accounting for Sampling Errors in PFA

- Key Concepts:
  - Pareto Depth Distribution: Fleury&etal:ISBI04, Fleury&etal:JFI03
  - Pareto Resistant Genes: Hero&Fleury:VLSI04

- 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

Pareto Depth Posterior Distribution

- Pareto front is set of non-dominated genes
- Gene i is dominated if there exists another gene g such that for some p:

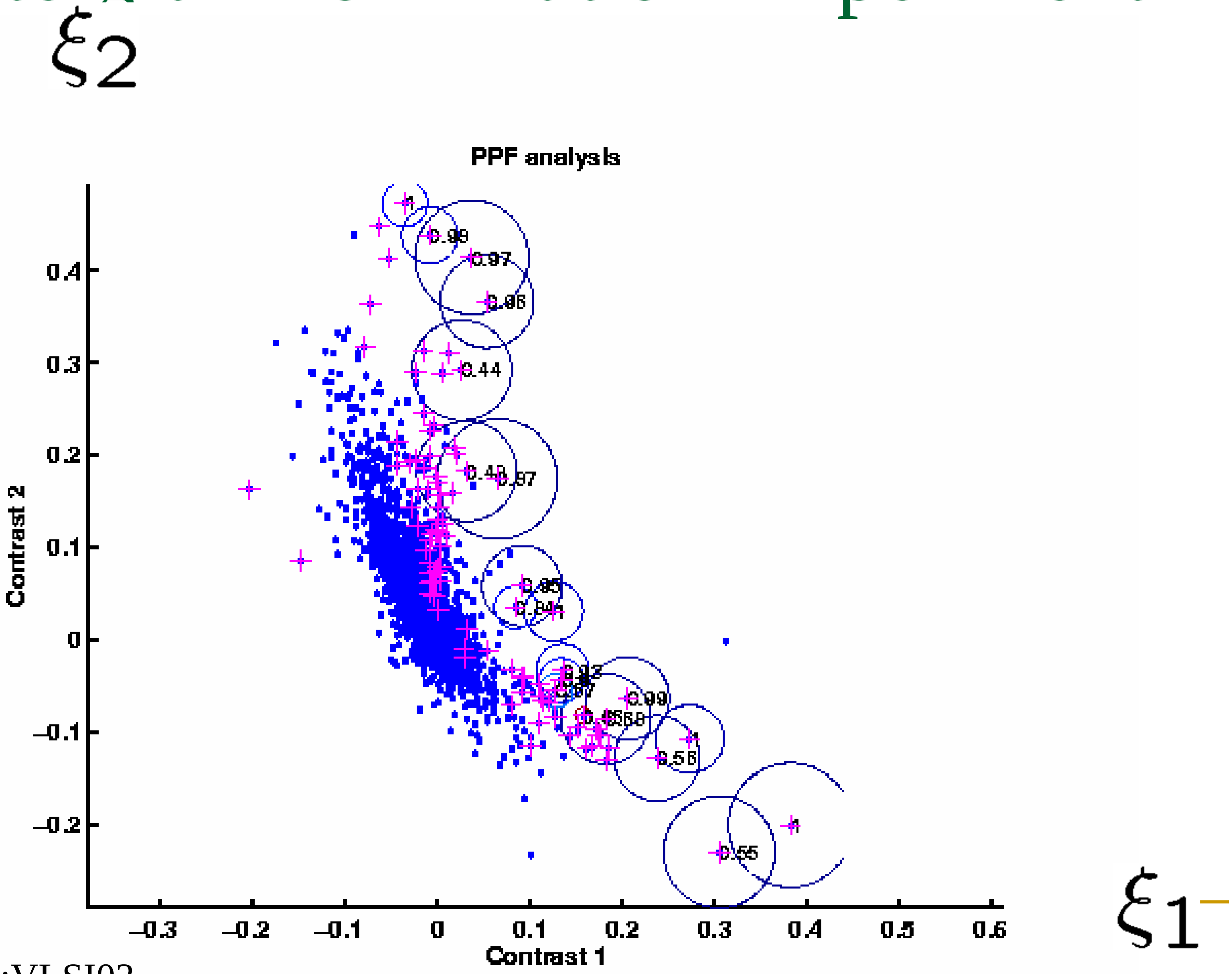
$$\boxed{\xi_q(i) < \xi_q(g) \;\; \text{and} \;\; \xi_p(i) \leq \xi_p(g),\; p \neq q.}$$

- Posterior probability: gene g is on Pareto front

$$\boxed{p(g|Y) = \int d\underline{u} f_{\underline{\xi}(g)|Y}(\underline{u}) \prod_{j\neq g} \left[1 - P\left(\underline{u} \leq \underline{\xi}(j)|Y\right)\right].}$$

- Can implement w/ non-informative prior on $\underline{\xi}(g)$

Scattergram for Dilution Experiment



Pareto Depth Sampling Distribution

- Let k be Pareto depth of gene g when leave out m-th replicate. Define

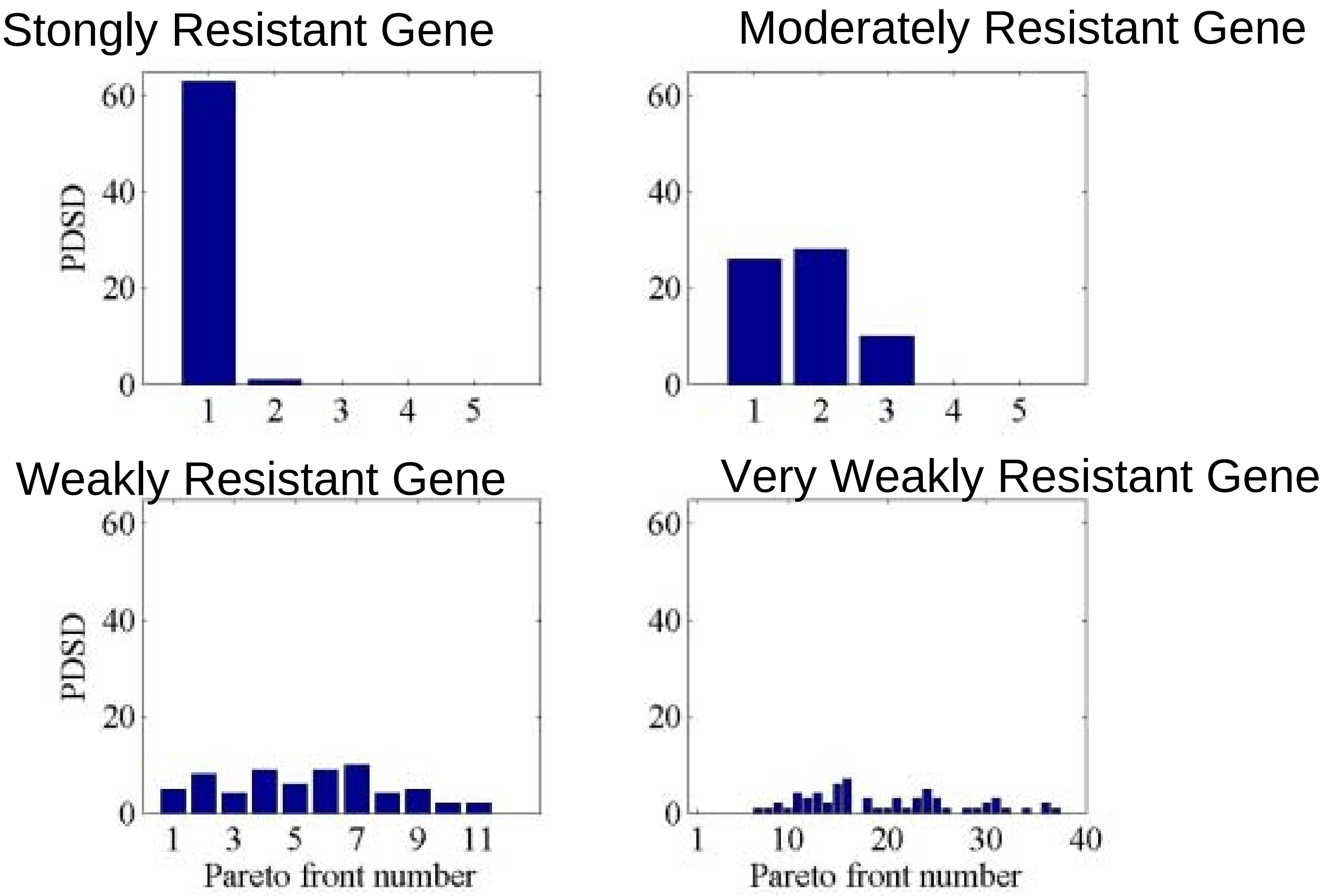
$$1_g(m,k') = \begin{cases} 1, & k' = k \\ 0, & o.w. \end{cases}$$

- (Re)sampling distribution of Pareto depth

$$\mathrm{Pdsd}_g(k) = \frac{1}{M_{\mathrm{resamp}}} \sum_{m=1}^{M_{\mathrm{resamp}}} 1_g(m,k), k = 1, \ldots, G$$

Ref: Fleury and Hero:JFI03

PDSD Examples for 4 different genes



Stongly Resistant Gene

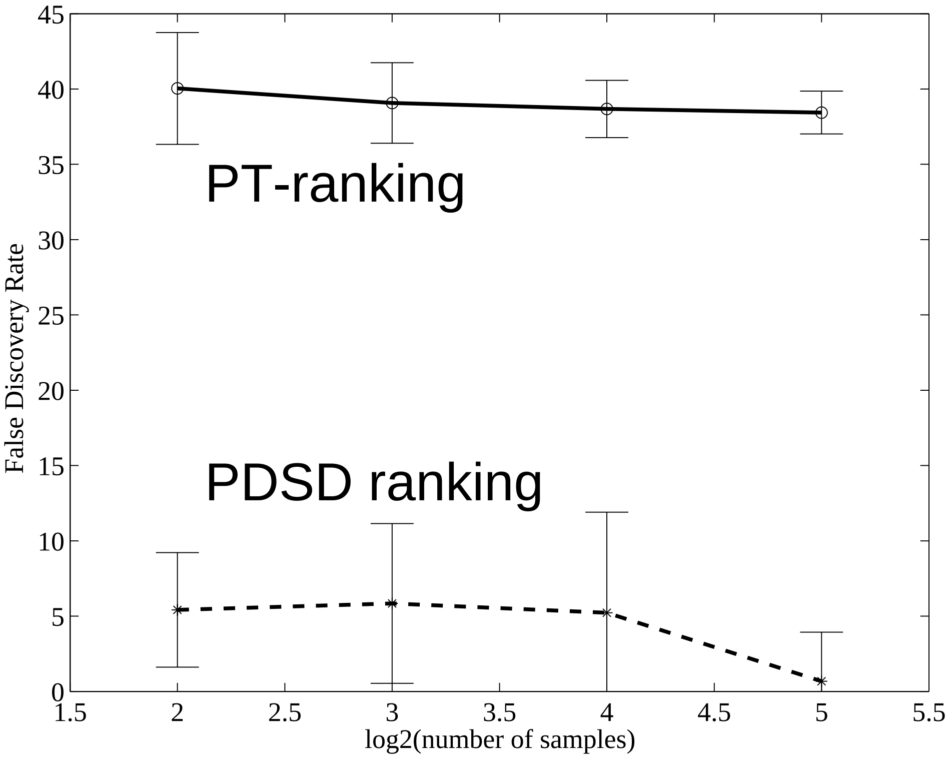
Moderately Resistant Gene

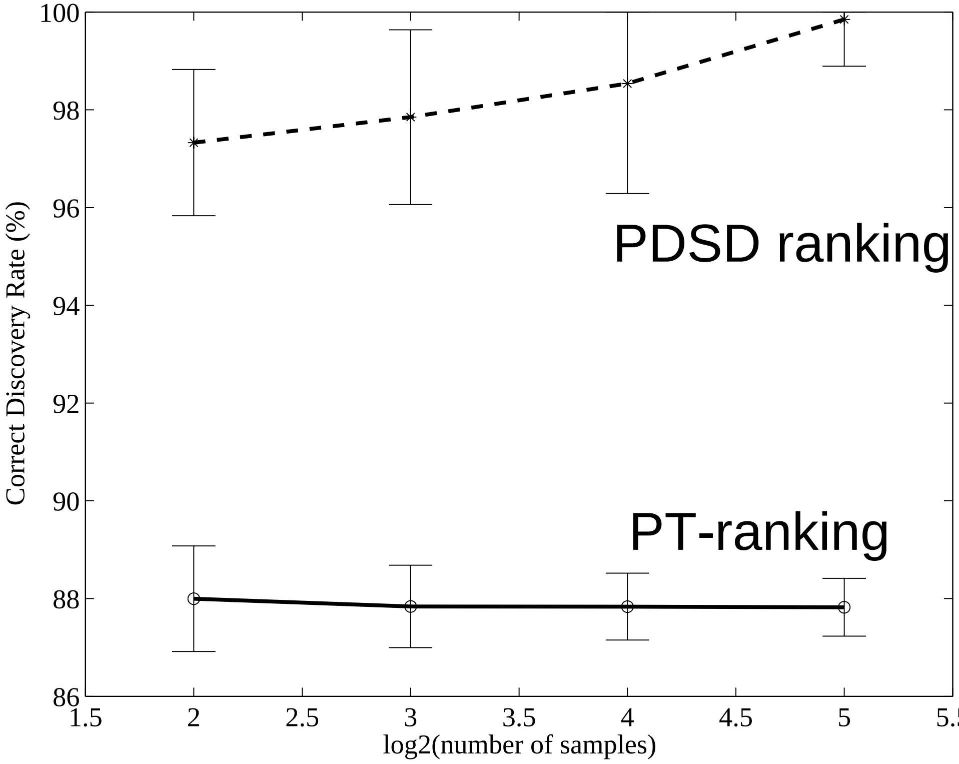
Weakly Resistant Gene

Very Weakly Resistant Gene

Ref: Fleury and Hero:JFI03

False Discovery Rate Comparisons





False Discovery Rate

Correct Discovery Rate

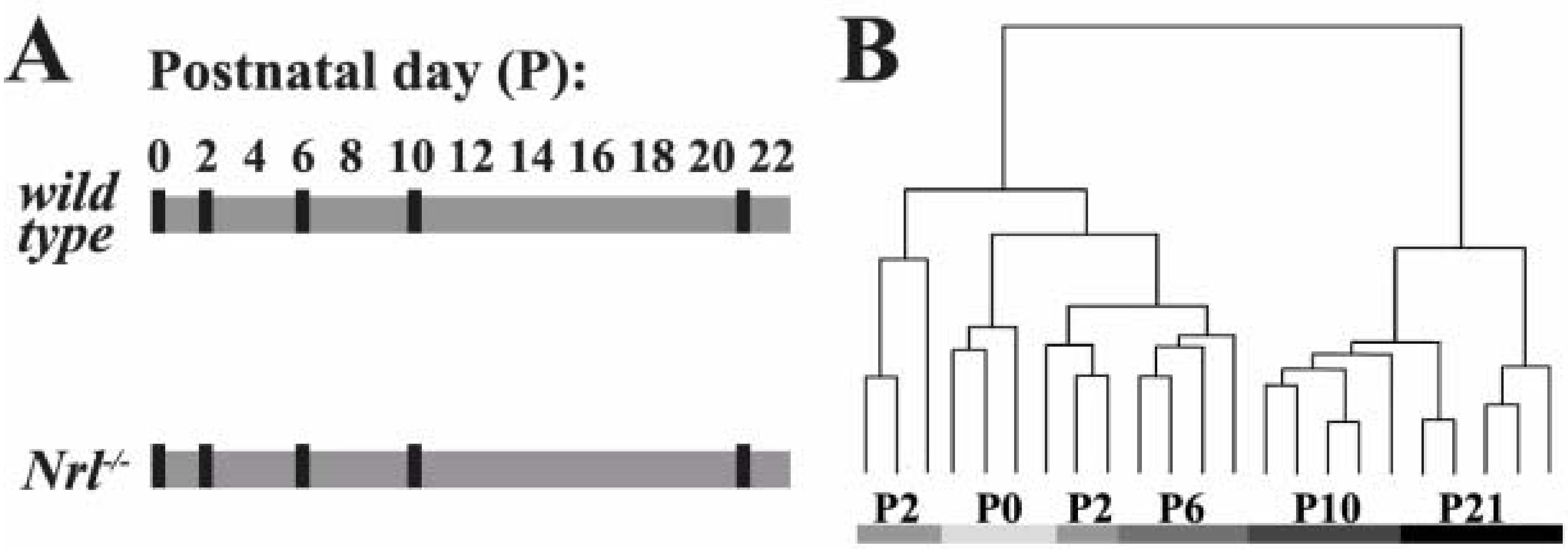
Ref: Fleury and Hero:JFI03

5.4 Clustering of Gene Expression Profiles

- Objective: find groups of genes that are similar to each other within a group and dissimilar across groups
- Clustering = classification without knowing the classes
- Common Clustering Techniques:
  - Hierarchical clustering
  - Combinatorial (partitioning): k-means, k-mediods, VQ
  - Model-based “soft”clustering
  - Spectral clustering: gene shaving, MDS, SOM, PCA
- Main issues in implementation of clustering algorithms:
  - Selecting number of genes and features to be clustered
  - Selecting number of clusters
  - Cluster validation and robustness

Clustering Case Study: cDNA wt/ko

- 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 ia 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

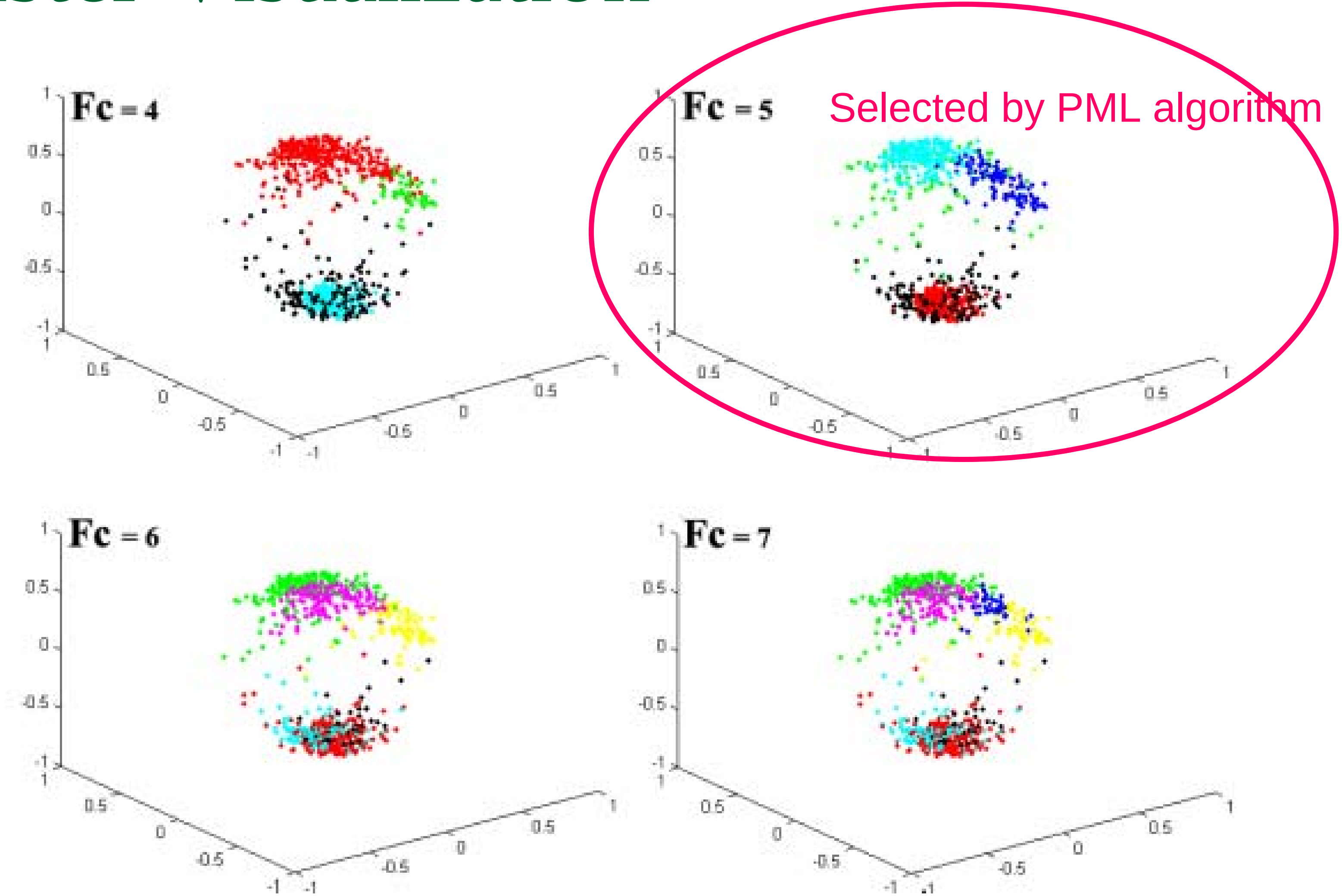
$$L(\theta, \alpha, C) = \sum_{g=1}^{G} \sum_{c=1}^{C} \alpha(c)\phi_c(X_g; \theta_c) + Q(C)$$

- 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

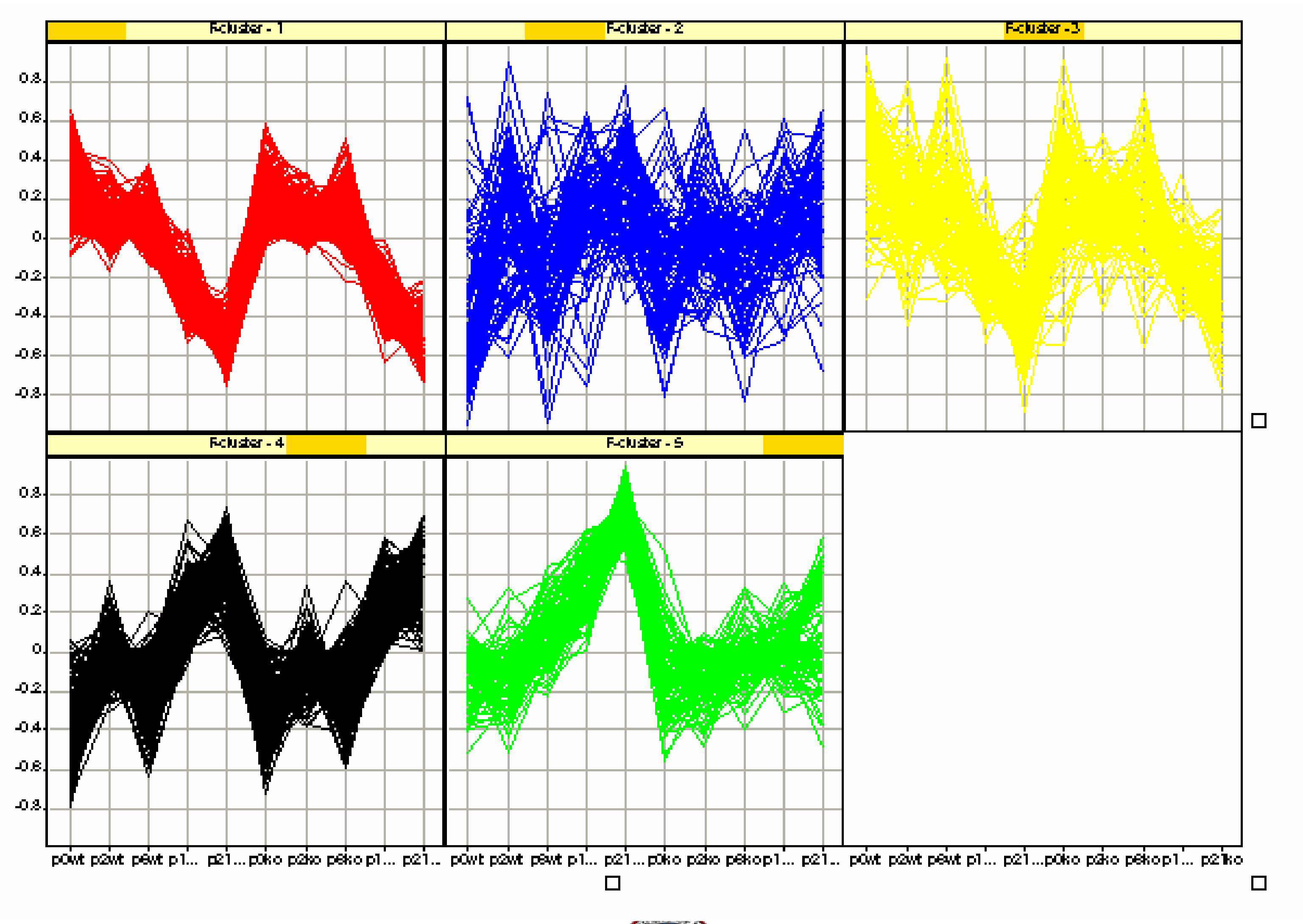


Selected by PML algorithm

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

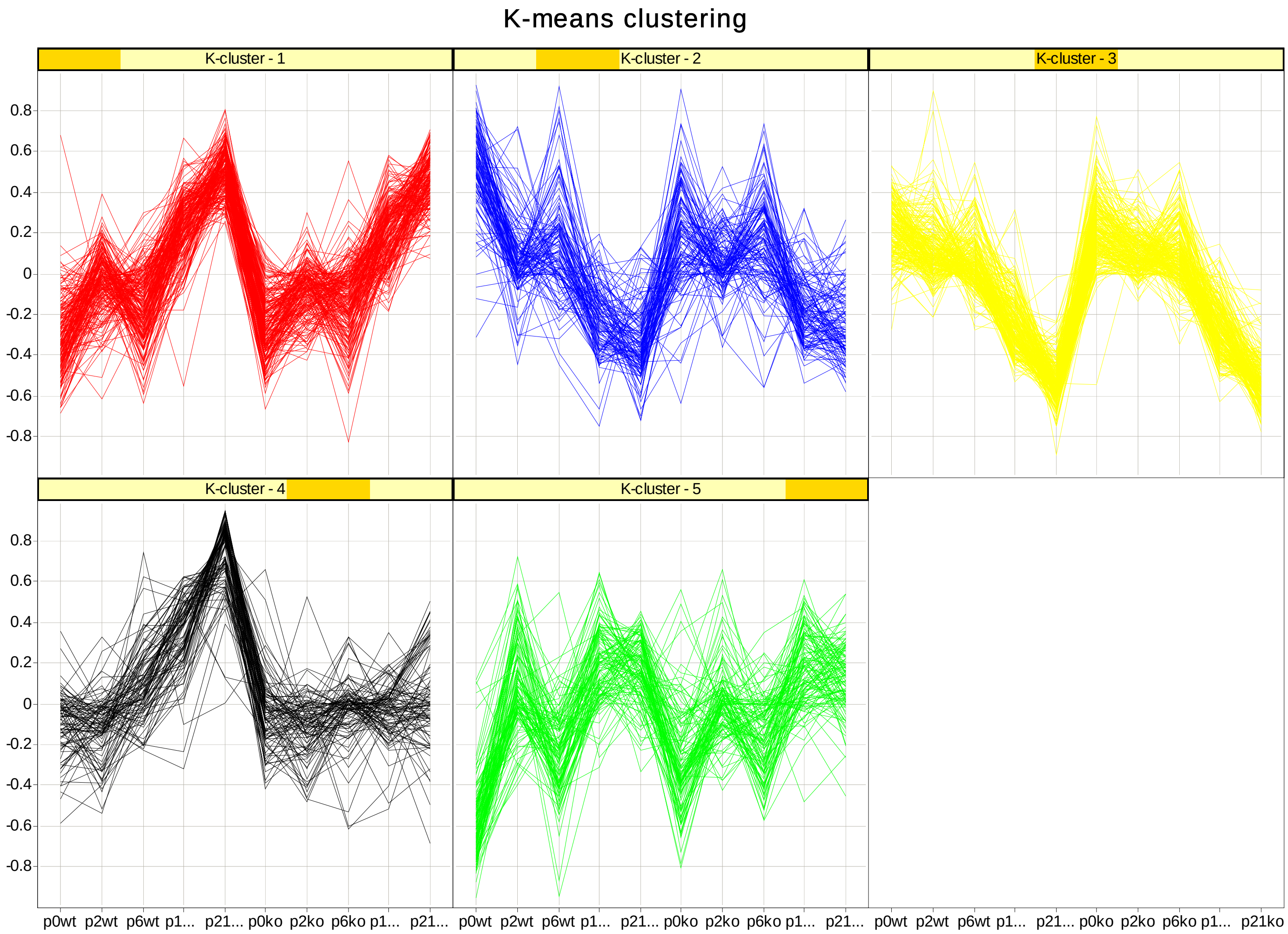
JindanYu, Stat750 Project Report, Univof Michigan, 2004.

Clustered Trajectories: PML Mixture



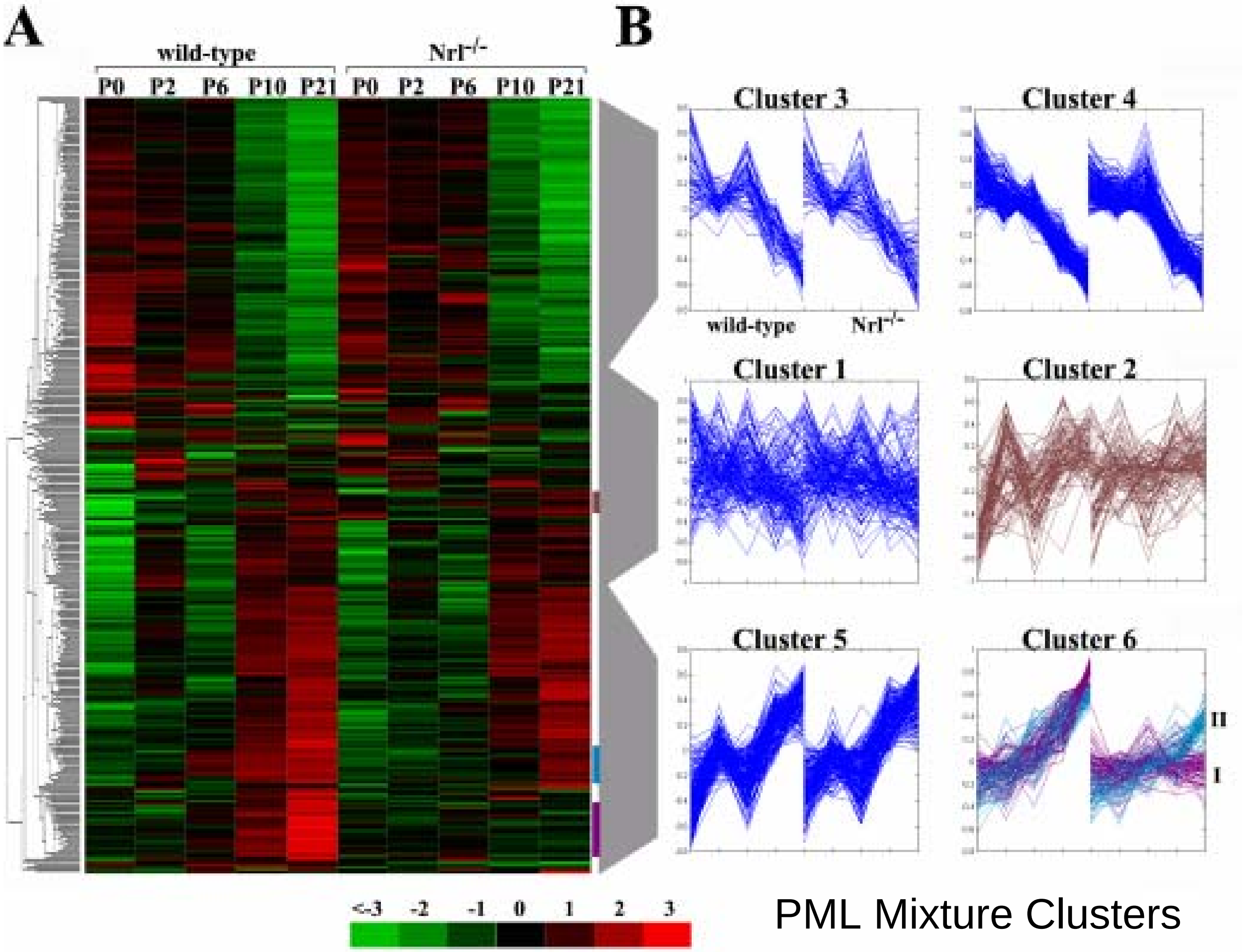
JindanYu, Stat750 Project Report, Univof Michigan, 2004.

Clustered Trajectories: k-Means



JindanYu, Stat750 Project Report, Univof Michigan, 2004.

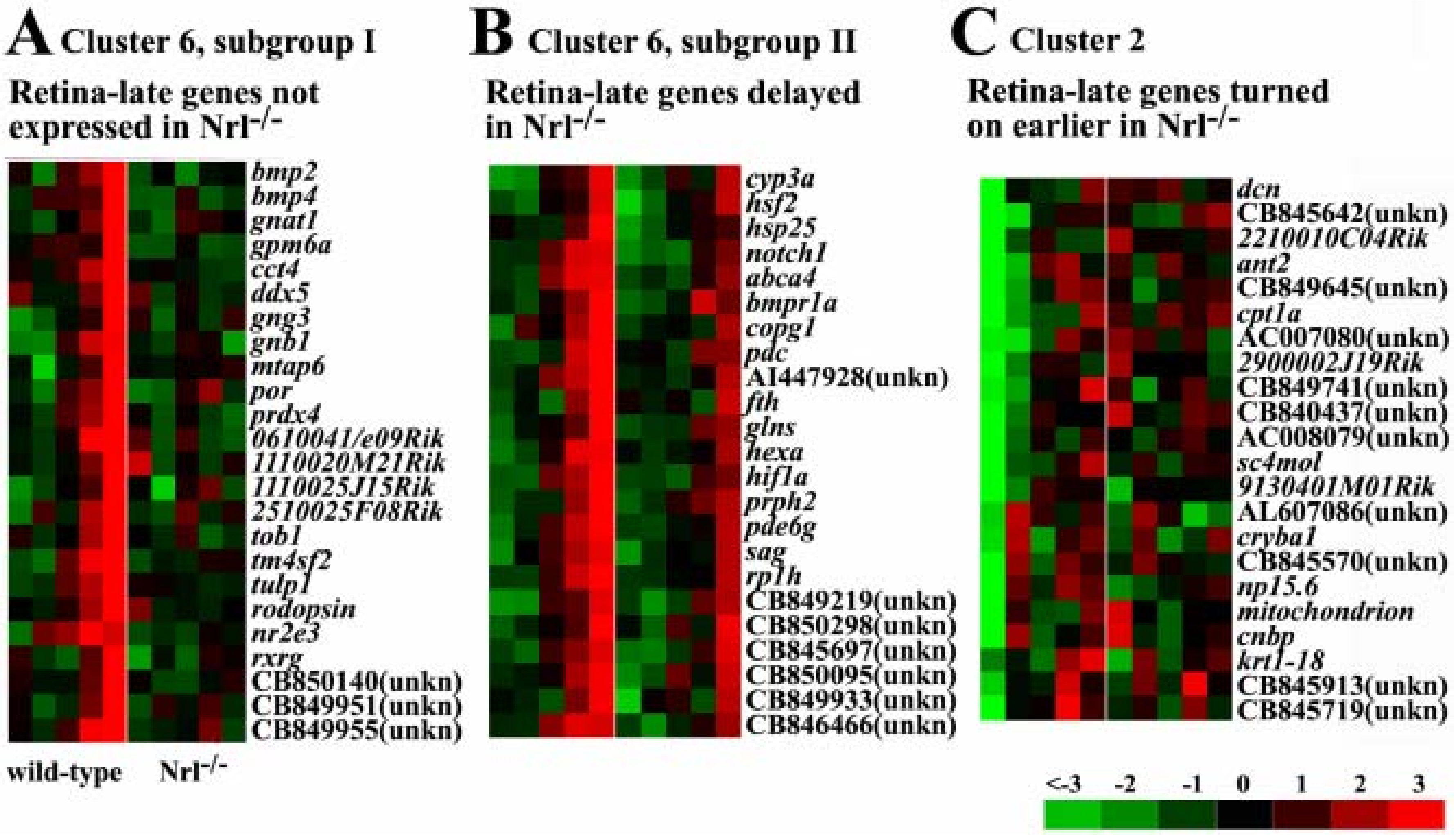
Compare to Hierarchical Clustering



PML Mixture Clusters

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

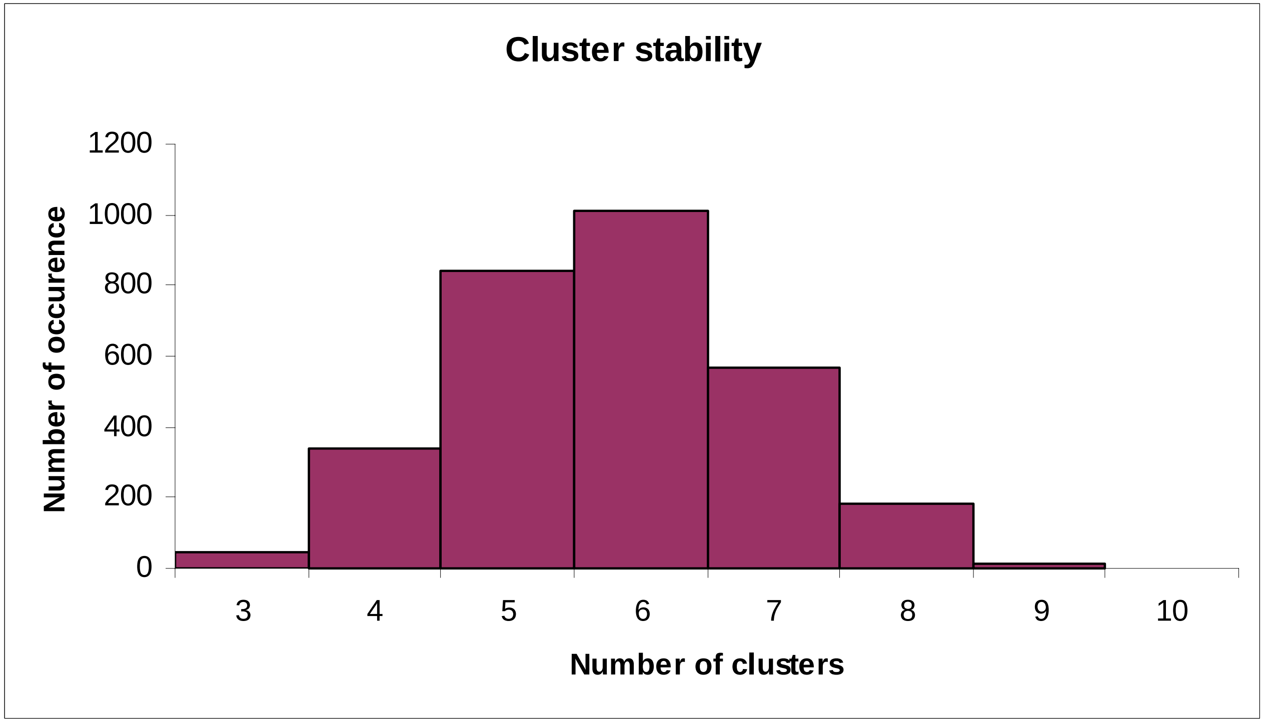
Post-Clustering Time Course Analysis



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

Cluster Validation and Robustness

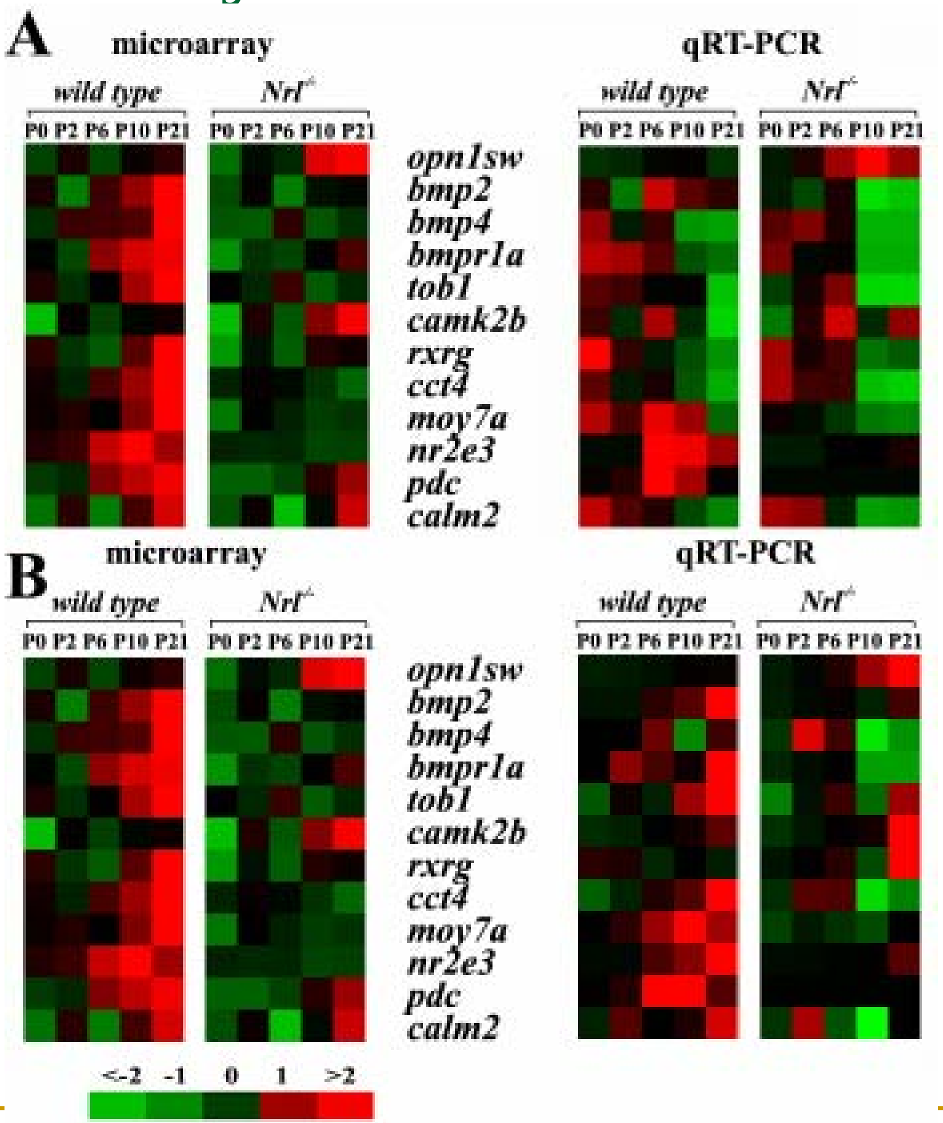
- Bootstrap resampling distribution



- Other metrics: validity indices, Sillhouettes, etc

http://www.cs.tcd.ie/Nadia.Bolshakova/validation_algorithms.html

Validation by Real Time RT-PCR



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

6. Wrap Up and References

- Low level analysis for cDNA and oligo microarray differ
- Higher level analyses on extracted expression levels are similar
- Gene filtering: accounting for biological and statistical significance
- Gene ranking: can involve optimization over multiple criteria
- Gene clustering: classify response profiles under single or multiple treatments
- Increasing importance of statistical signal and image processing approaches

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)

References on analysis of time course data

- 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

- Duda, Hart and Stork, Pattern classification (2nd Ed), Wiley, NY 2000
- 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.
- G. Fleury, A. Hero, S. Zareparsi, and A. Swaroop, "Pareto Depth Sampling Distributions for Gene Ranking,"*Proc. of IEEE Intern, Symp. On Biomedical Imaging (ISBI)*, April 2004
- 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 in 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