Multiple criteria differential expression analysis of microarray data

Alfred O. Hero III

University of Michigan, Ann Arbor, MI

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

Bioinformatics Seminar

Dec. 2003

1. Gene Microarray Data
2. Multi-criteria Screening and Ranking
3. Biological vs Statistical Significance
4. FDR-CI Gene Screening
5. Pareto Front Gene Ranking

Thanks to…

- UM Students
  - Yuezhou Jing (Stat),
  - Sebastien Cerbourg (FinEng),
  - Kashif Siddiqui (EECS),
  - Jindan Wu (BME)
- Collaborators
  - Gilles Fleury (SupElec-Paris)
  - Anand Swaroop (UM Kellog)
  - Alan Mears (U. Ottawa)
  - Sepi Zaraparsi (UM Kellog)
  - Shigeo Yosida (U. Tokyo)
- Colleagues
  - Terry Speed (UCB)
  - Peter Bickel (UCB)
  - Fred Wright (UCLA)
  - Anonymous reviewers…

Biotechnology Overview

- **Genome:** All the DNA contained in an organism**.** The operating system/program for structure/function of an organism.
- **Genomics:** investigation of structure and function of very large numbers of genes undertaken in a simultaneous fashion.
- **Bioinformatics:** Computational extraction of information from biological data.
- **Data Mining:** Algorithms for extracting information from huge datasets using user-specified criteria.

I. Gene Microarray Technologies

- High throughput method to probe gene expression in a sample
- Two principal microarray technologies:

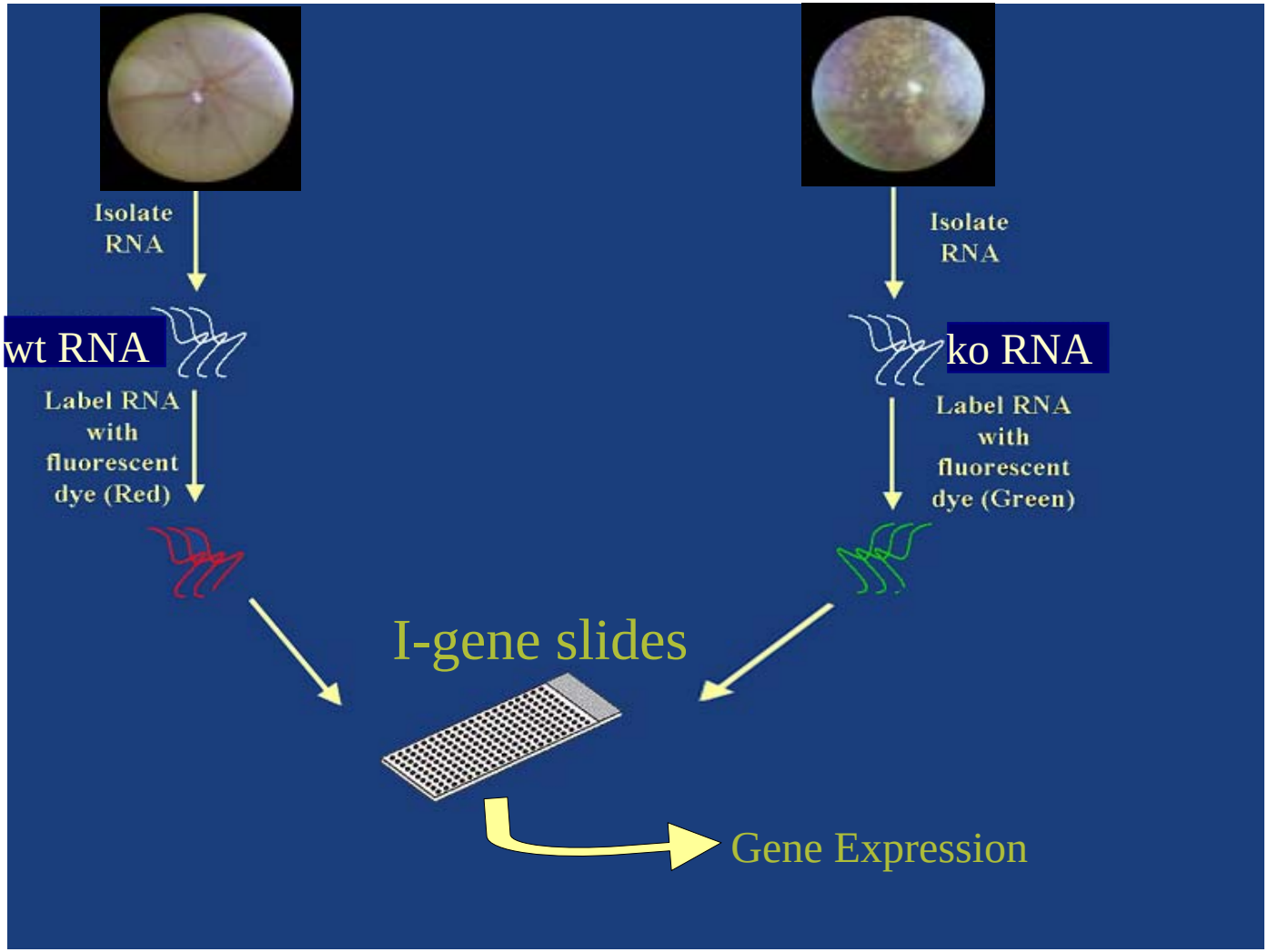
  1) oligonucleotide arrays (Affymetrix GeneChip)
  2) cDNA spotted arrays (Brown/Botstein chip)

- Main idea behind cDNA technology:

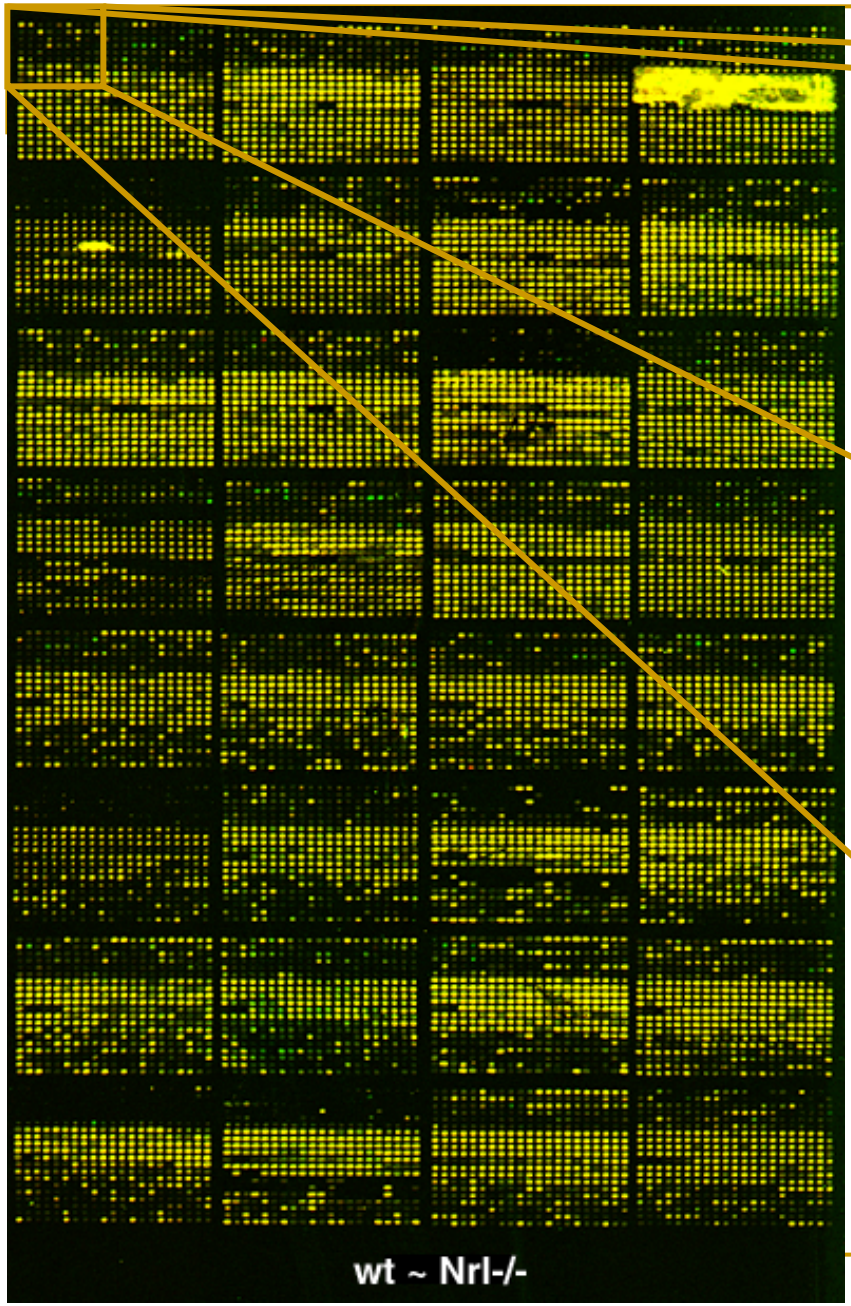
  1) Specific complementary DNA sequences arrayed on slide
  2) Dye-labeled RNA from sample is distributed over slide
  3) RNA binds to probes (hybridization)
  4) Presence of bound RNA-DNA pairs is read out by detecting spot fluorescence via laser excitation (scanning)

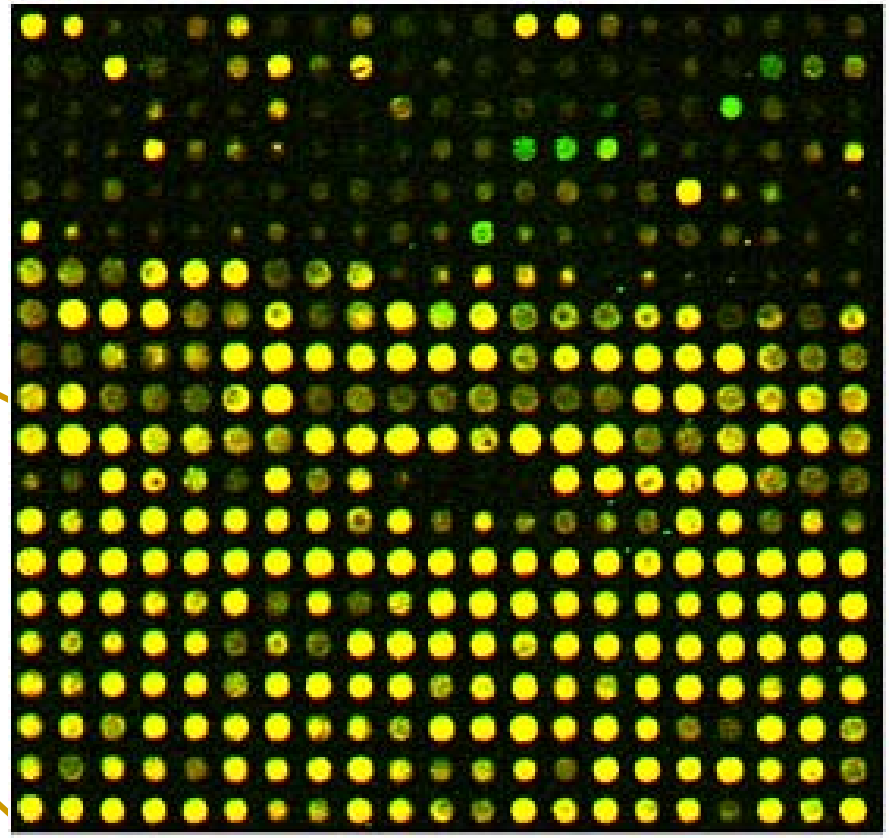
- **Result: sets of 10,000-50,000 genes can be probed**

I-Gene cDNA Microarray



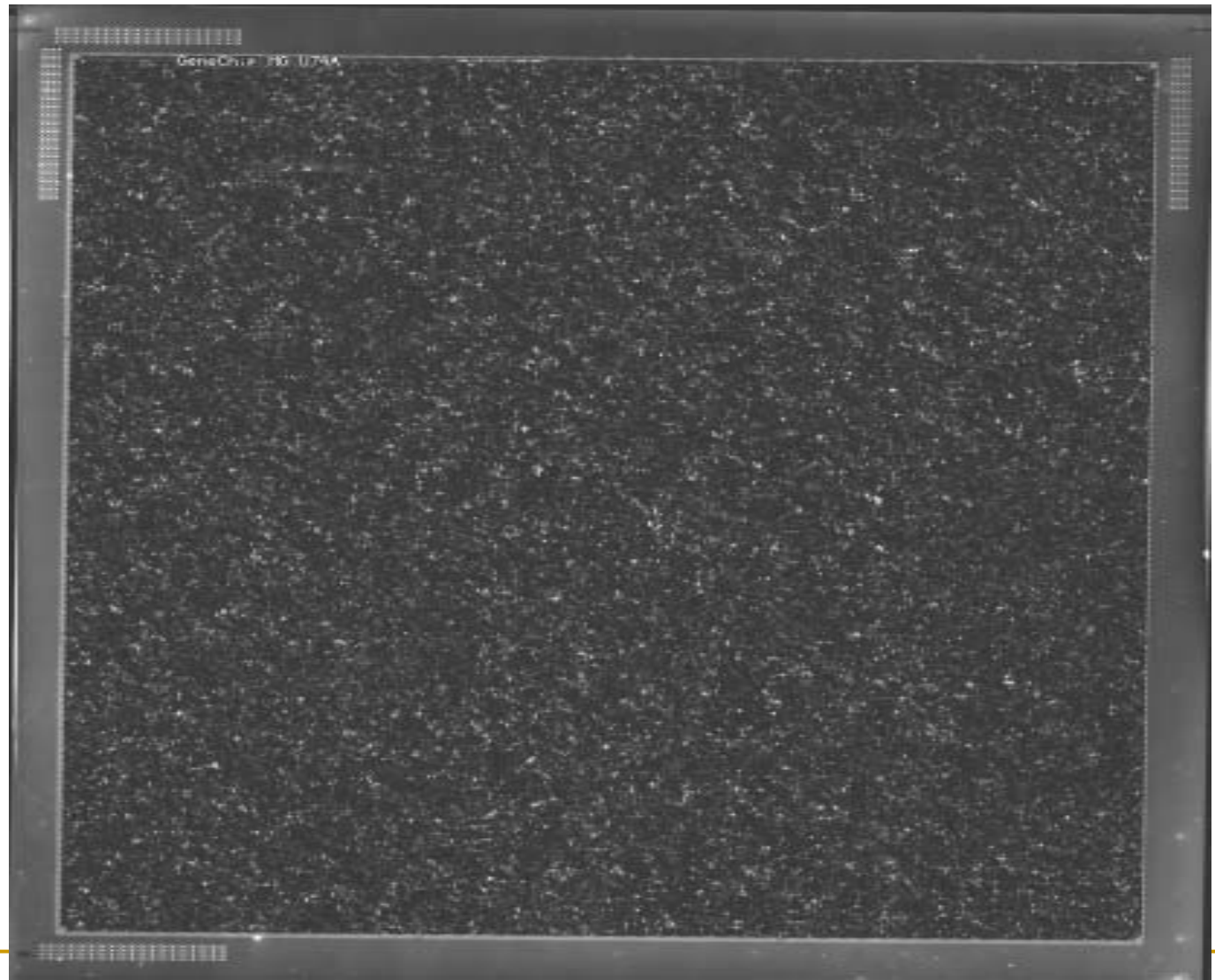
Source: J. Yu, UM BioMedEng Thesis Proposal (2002)



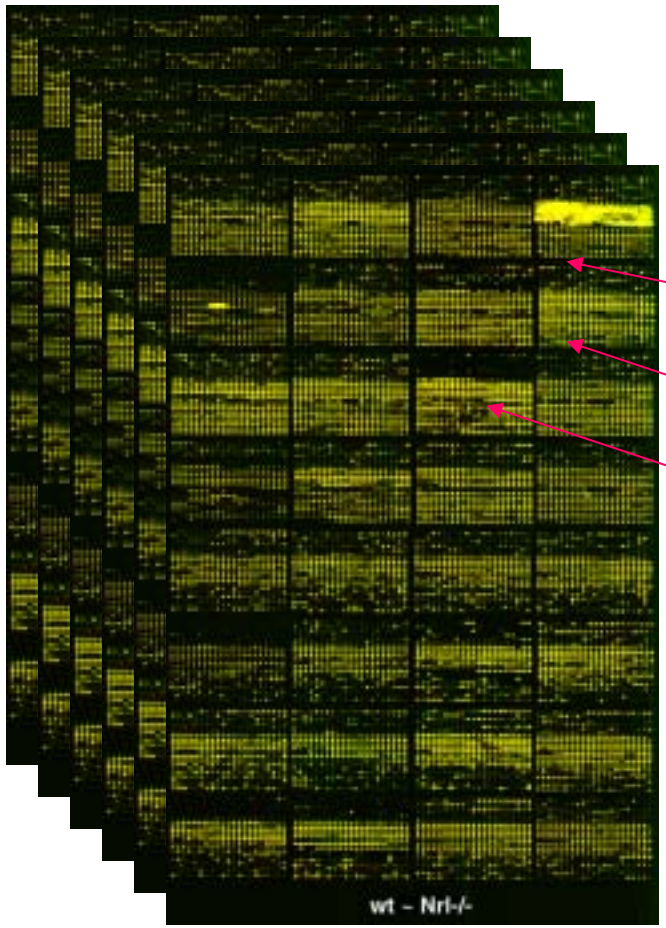


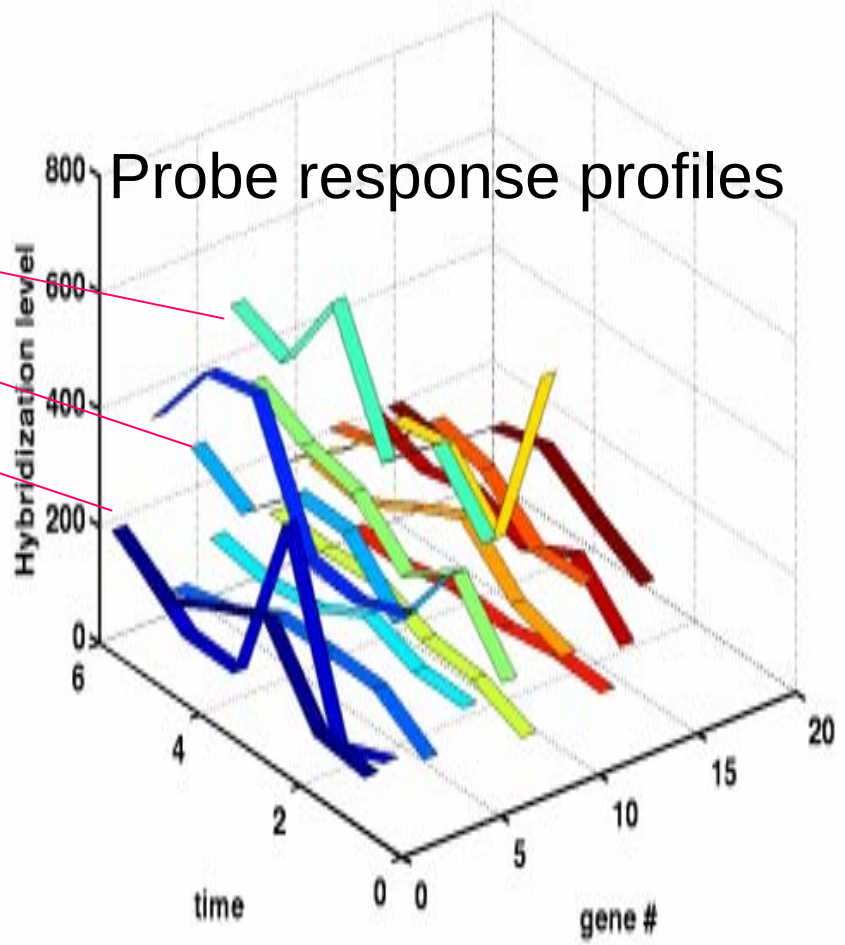
- **Treated sample (ko) labeled red (Cy5)**
- **Control (wt) labeled green (Cy3)**

Oligonucleotide GeneChip (Affymetrix)

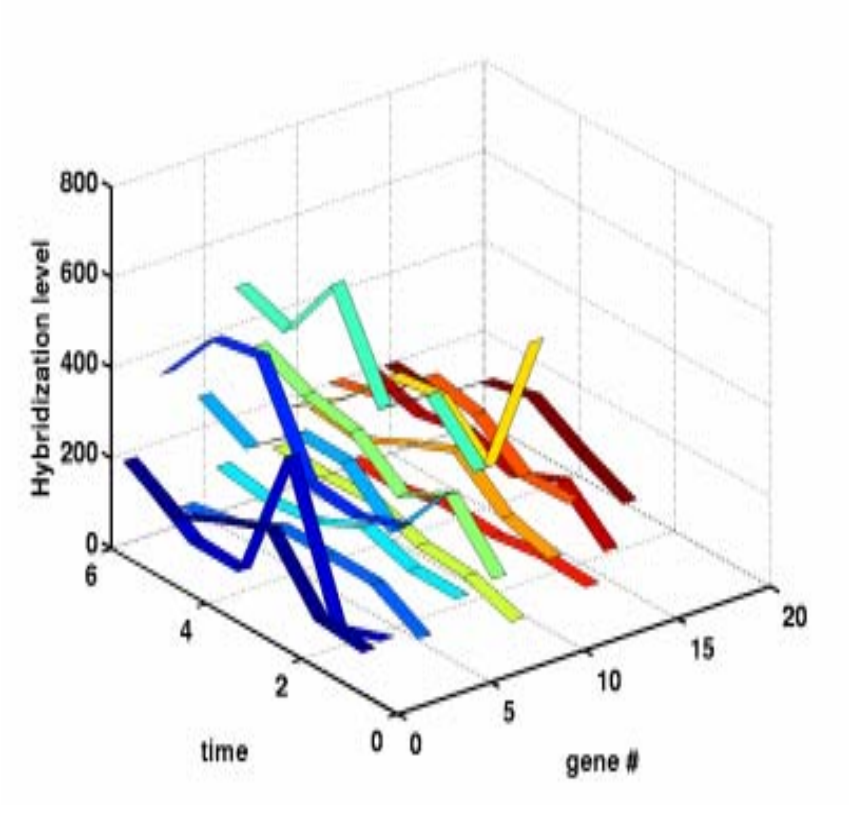


Add Treatment Dimension: Expression Profiles

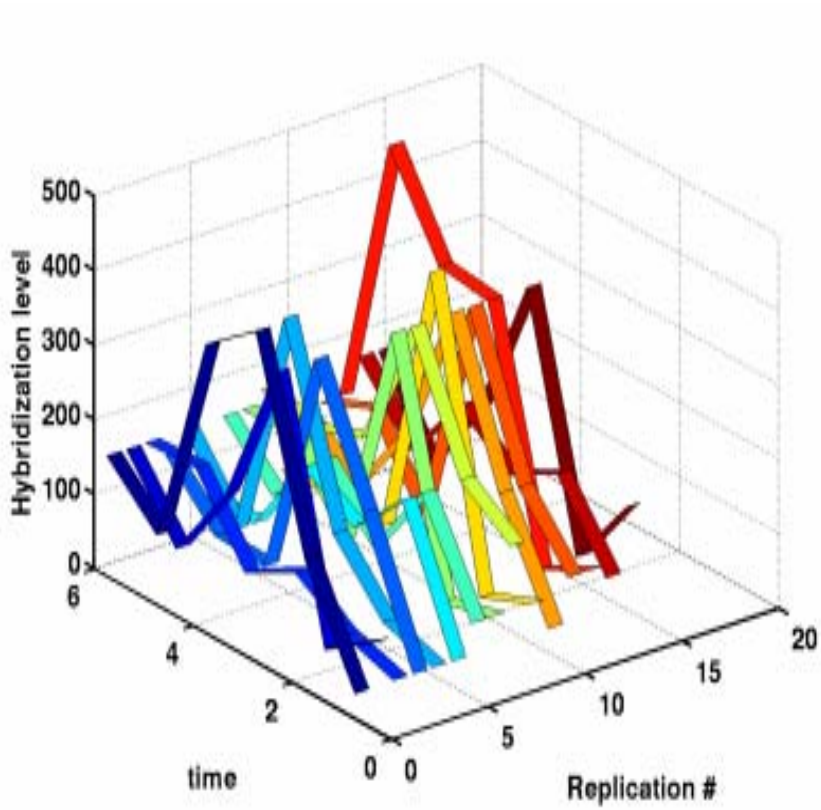




Problem of Intrinsic Profile Variability



Across-treatment variability



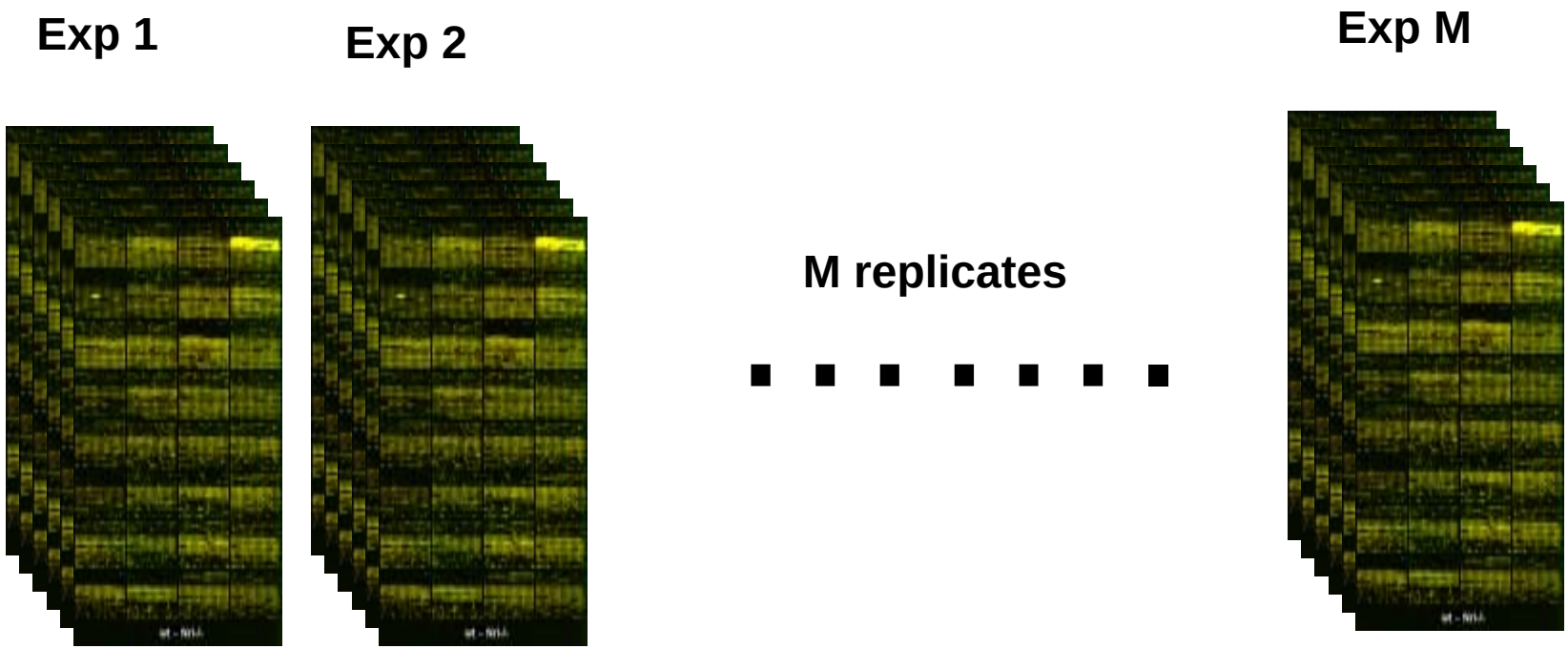
Across-sample variability

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.

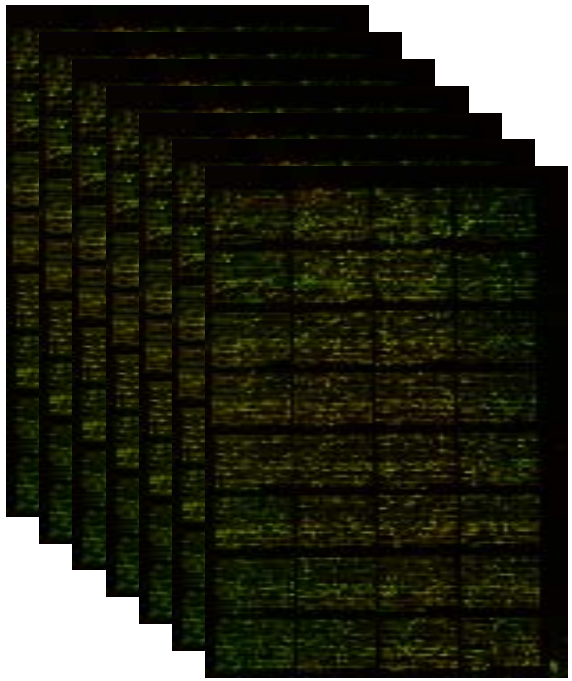
Solution: Experimental Replication



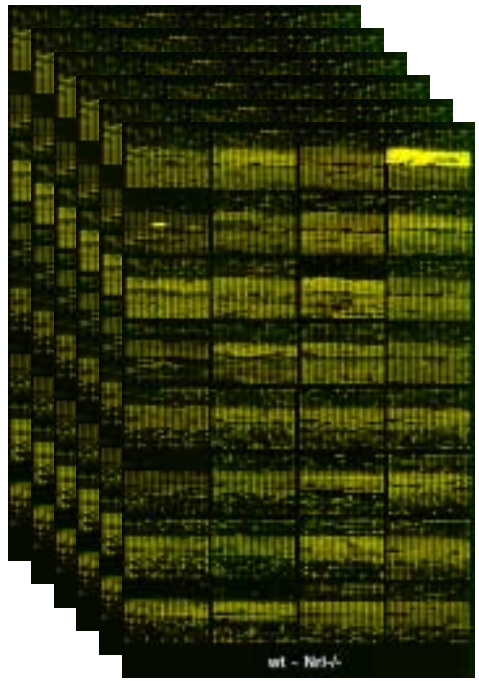
Issues:

- 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

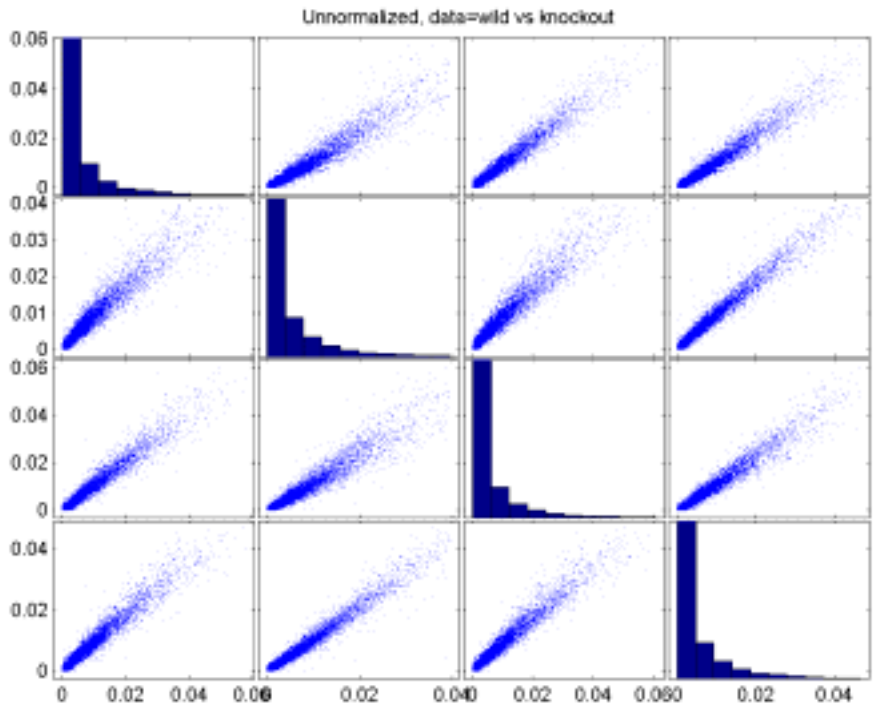


Experiment B

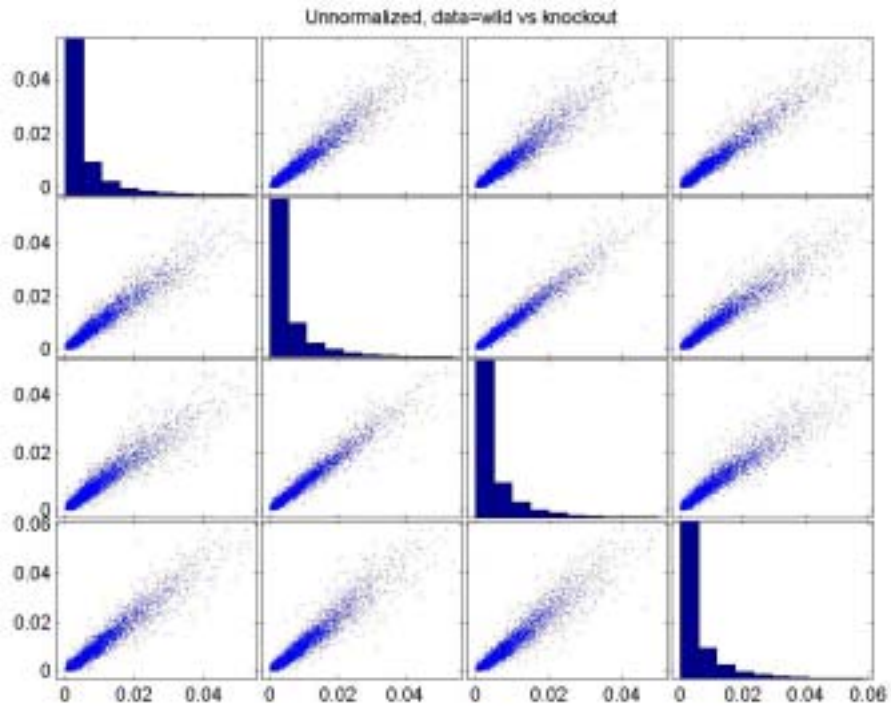
Question: How to combine or compare experiments A and B?

Un-Normalized Data Sets

Within-experiment intensity variations mask A-B differences:



Experiment A (Wildtype)



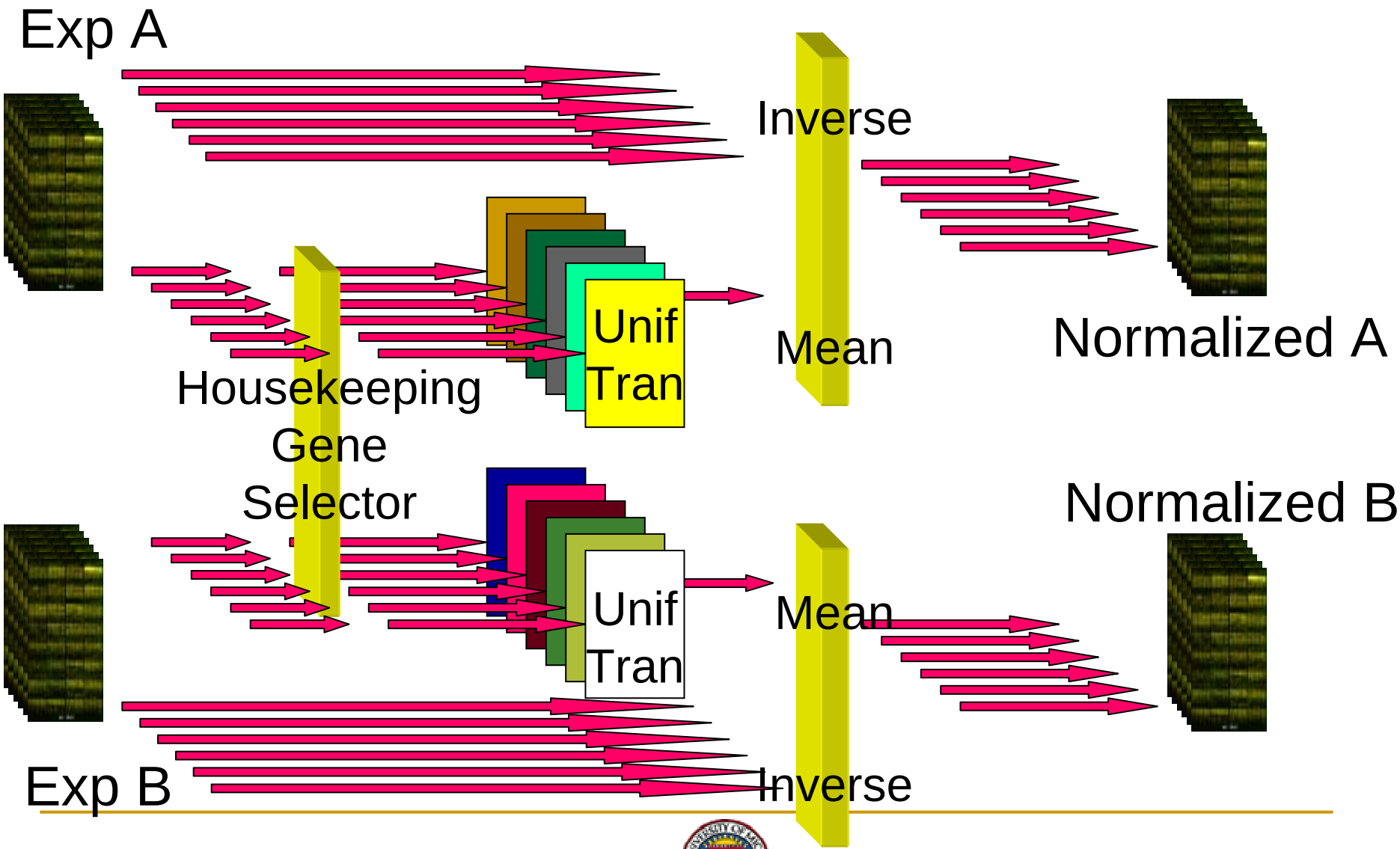
Experiment B (Knockout)

Hero&Fleury, ISSP-03

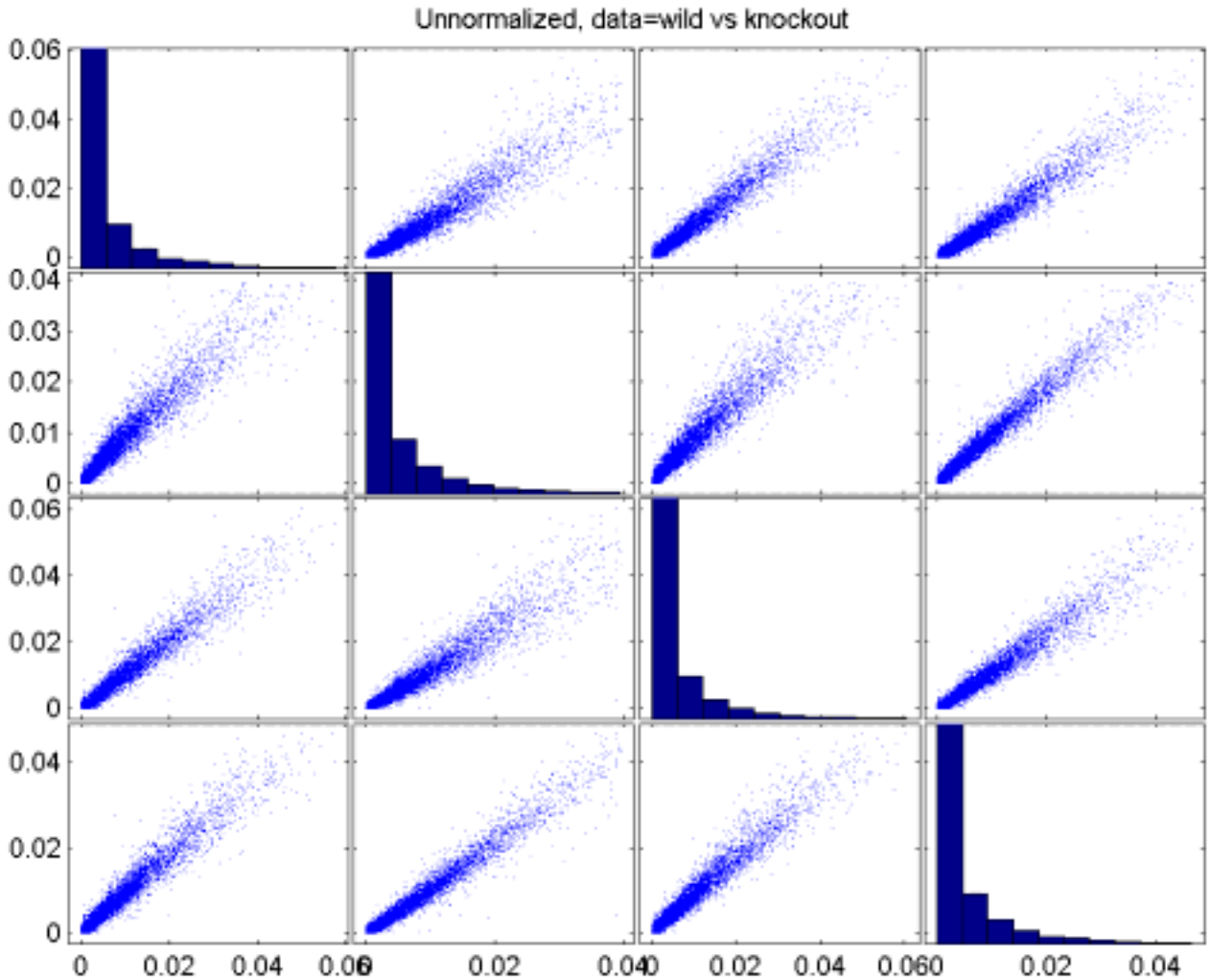
Two Possible Approaches

- If quantitative gene profile comparisons are required:
  - find normalization function to align all data sets within an experiment to a common reference. Examples: RMA, SMA, (Irizzary&etal:2002)
- If only ranking of gene profile differences is required:
  - No need to normalize: can apply rank order transformation to measured hybridization intensities (Hollander&Wolf:1999, Hero&Fleury:VLSI2003).

A vs B Microarray Normalization Method

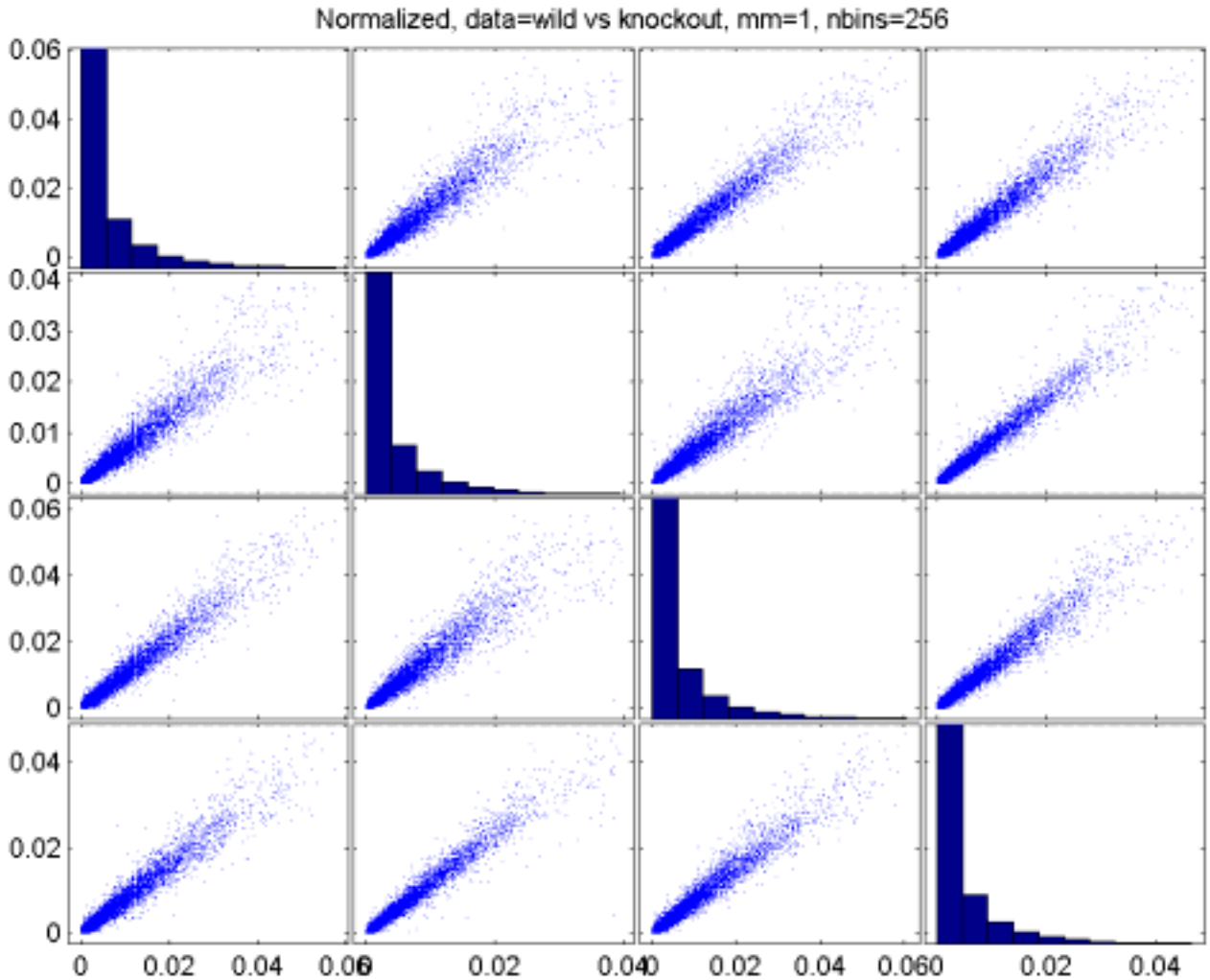


Un-Normalized Data Set (Wildtype)



Hero: ISSP-03

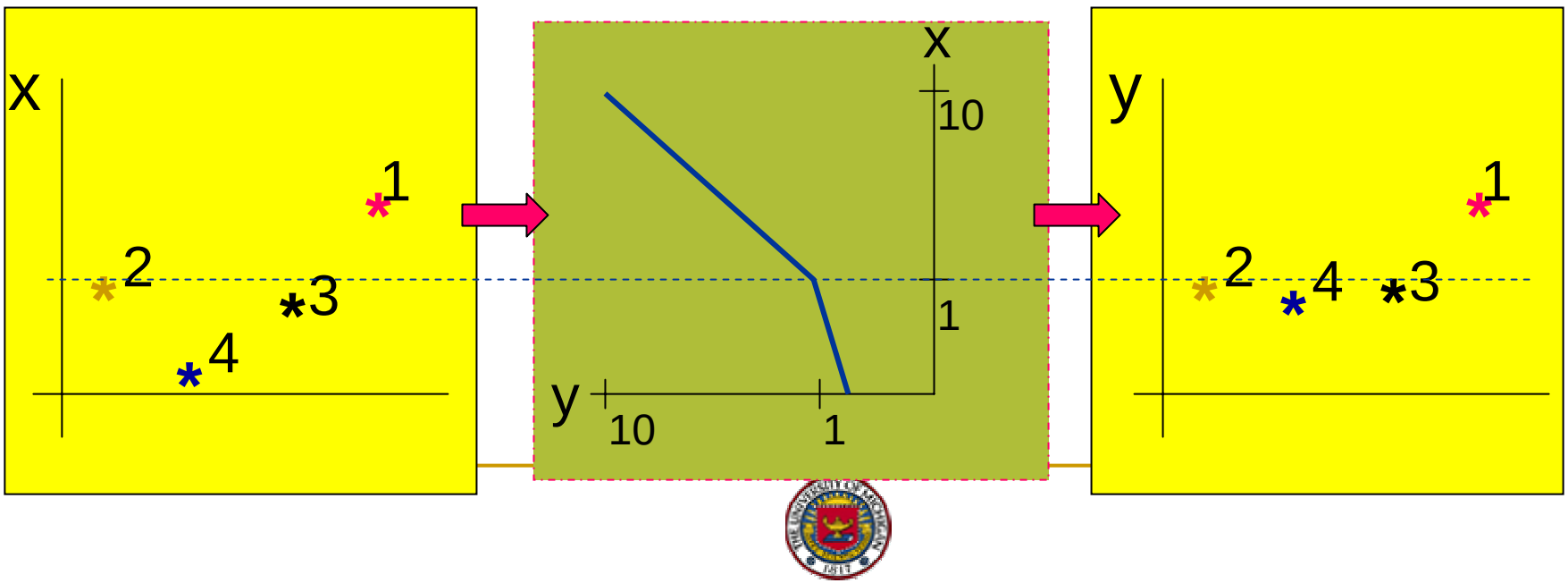
Normalized Data Set (Wildtype)



Hero:ISSP-03

Rank Order Statistical Transformation

- Rank order algorithm: at each time point replace each gene intensity with its relative rank among all genes at that time point
  - The relative ranking is preserved by (invariant to) arbitrary monotonic intensity transformations.



II. Multicriteria Gene Screening

- 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

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

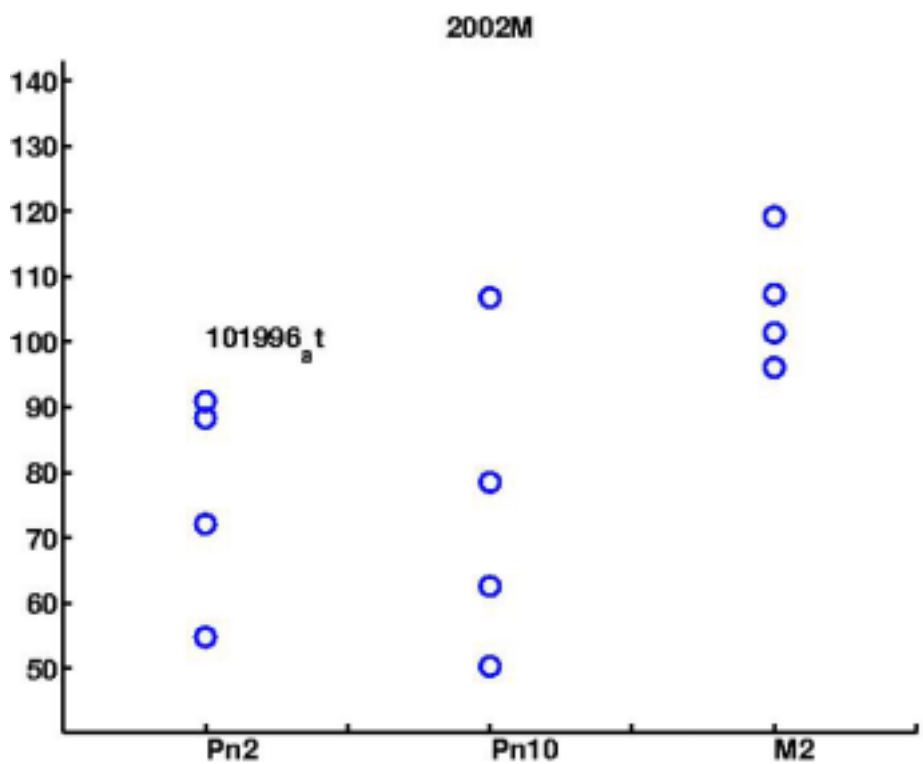
  - Paired comparisons for two treatments over T time points

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

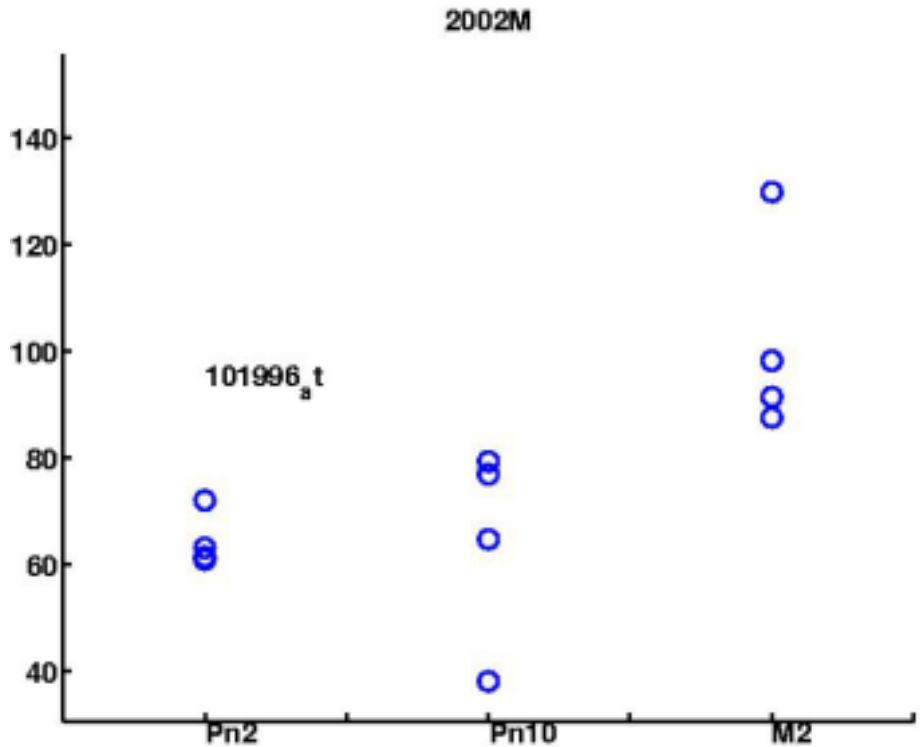
Knockout vs Wildtype Retina Study

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

Knockout



Wildtype



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

III. Biological vs Statistical Significance:

- Statistical significance refers to foldchange being different from zero

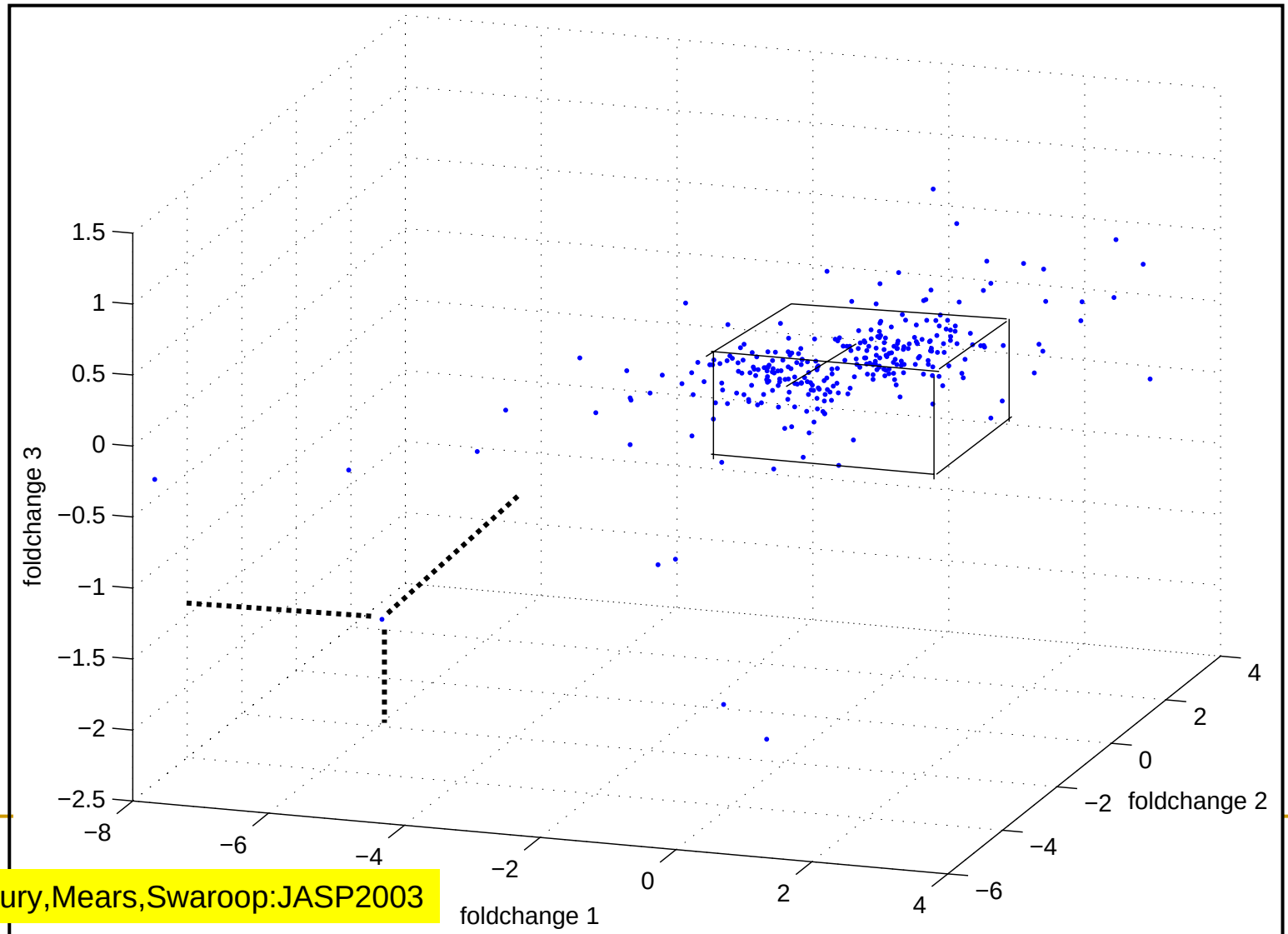
$$\mathrm{fc}(g) \neq 0$$

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

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

Hero,Fleury,Mears,Swaroop:JASP2003

Biological and Statistical Significance: Minimum Foldchange Cube



Hero,Fleury,Mears,Swaroop:JASP2003

IV. FDRCI Gene Screening

- Let $fc_t(g)$ = foldchange of gene ‘g’ at time point ‘t’.
- We wish to simultaneously test the TG sets of hypotheses:

$$H_0(g,t) \;:\; fc_t(g) \le |d|$$
$$H_1(g,t) \;:\; fc_t(g) > |d|$$

- 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 (PT) statistic

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

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

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

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

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

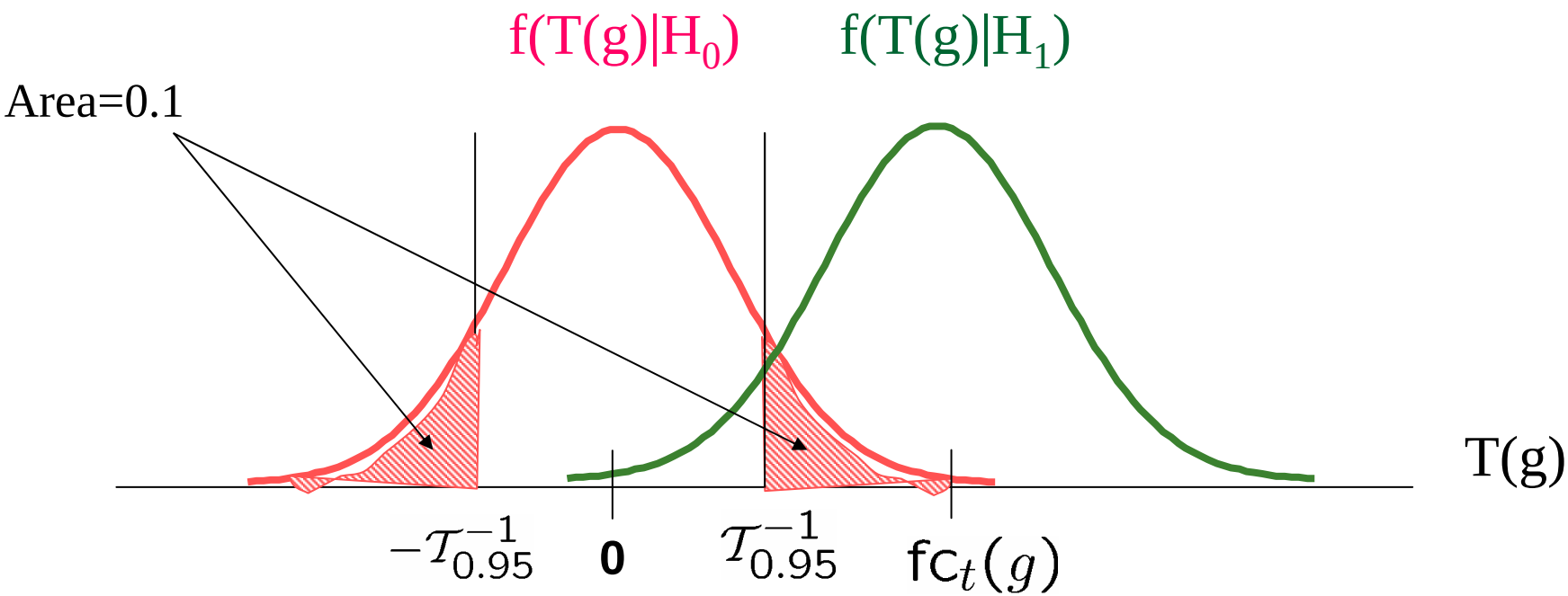
$$I_g(\alpha) = T_t(g) \pm \sqrt{\frac{2}{m}}\mathcal{T}^{-1}_{1-\alpha/2}$$

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

Stage 1: paired T test of level alpha=0.1

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

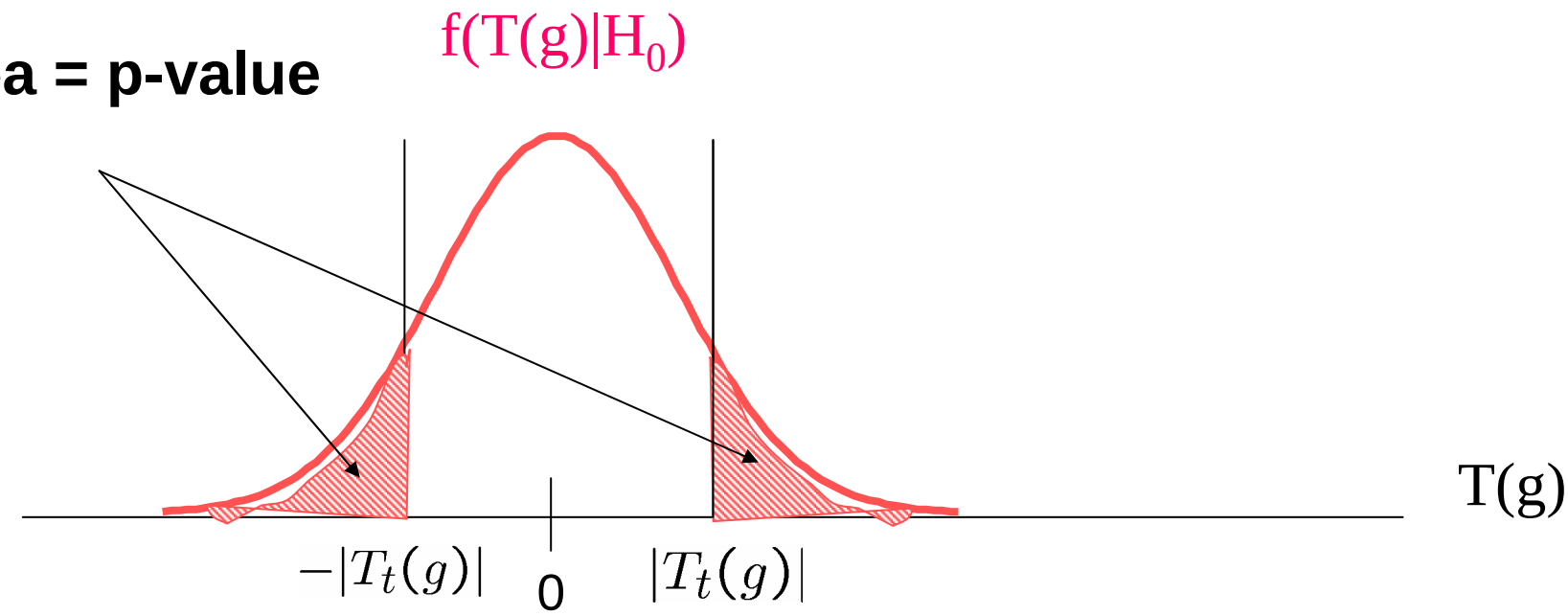
$$H_1 \;:\; \mathrm{fc}_t(g) \neq 0$$



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

Stage 1: p-value of paired T test

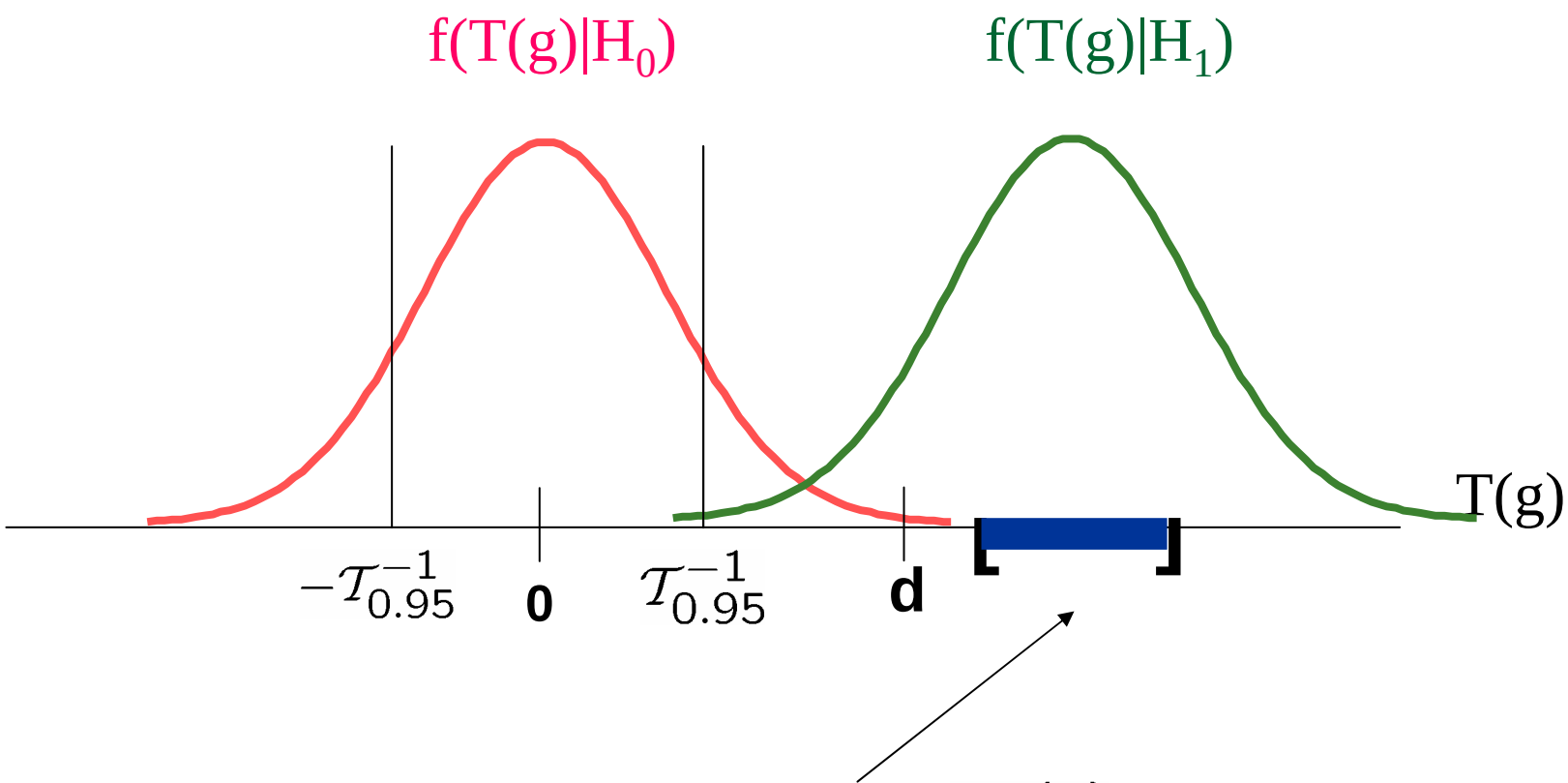
Area = p-value



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

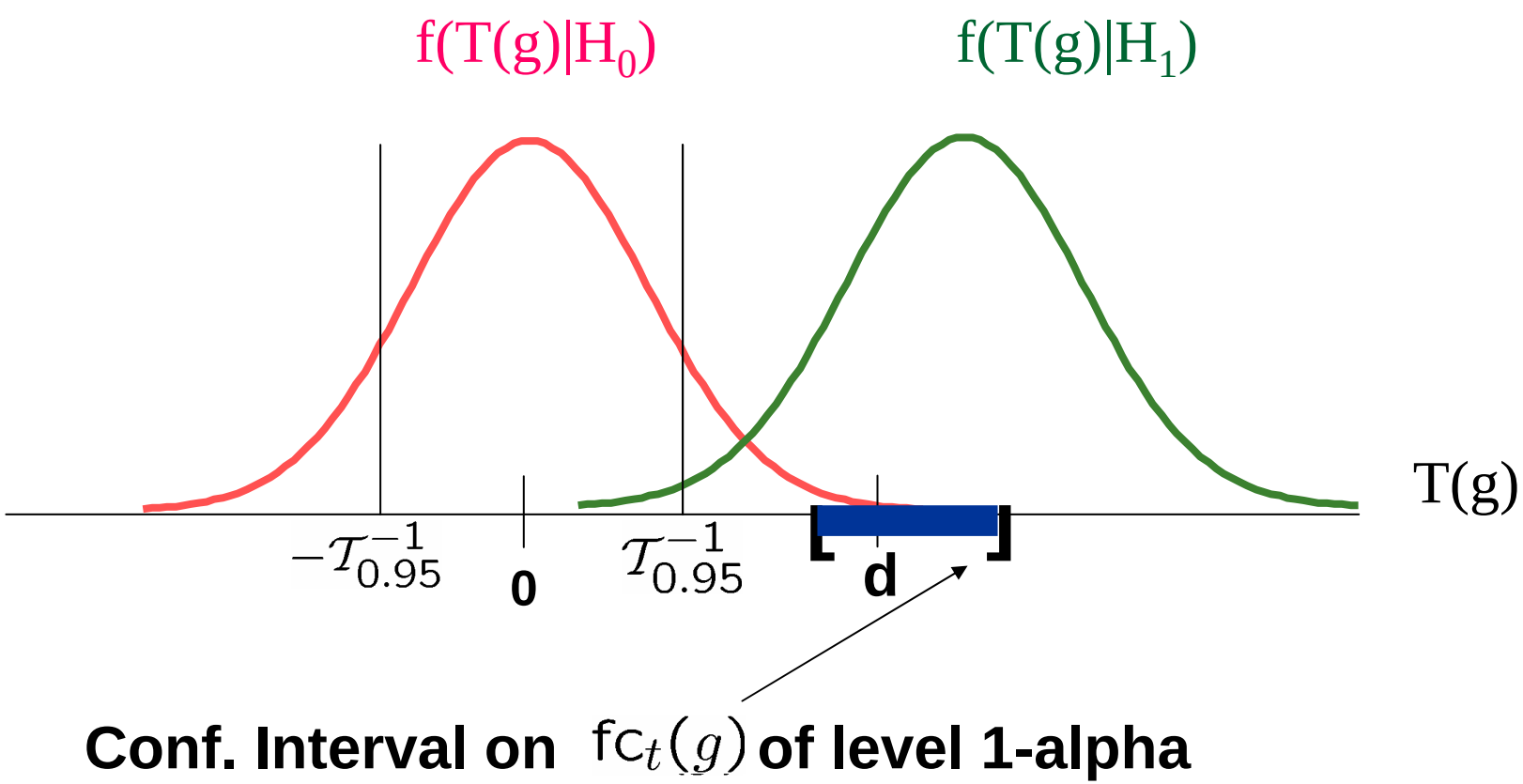
Stage 2: Confidence Intervals

- Biologically&statistically **significant** differential response



Stage 2: Confidence Intervals

- Biologically&statistically **insignificant** differential response



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)

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

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

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

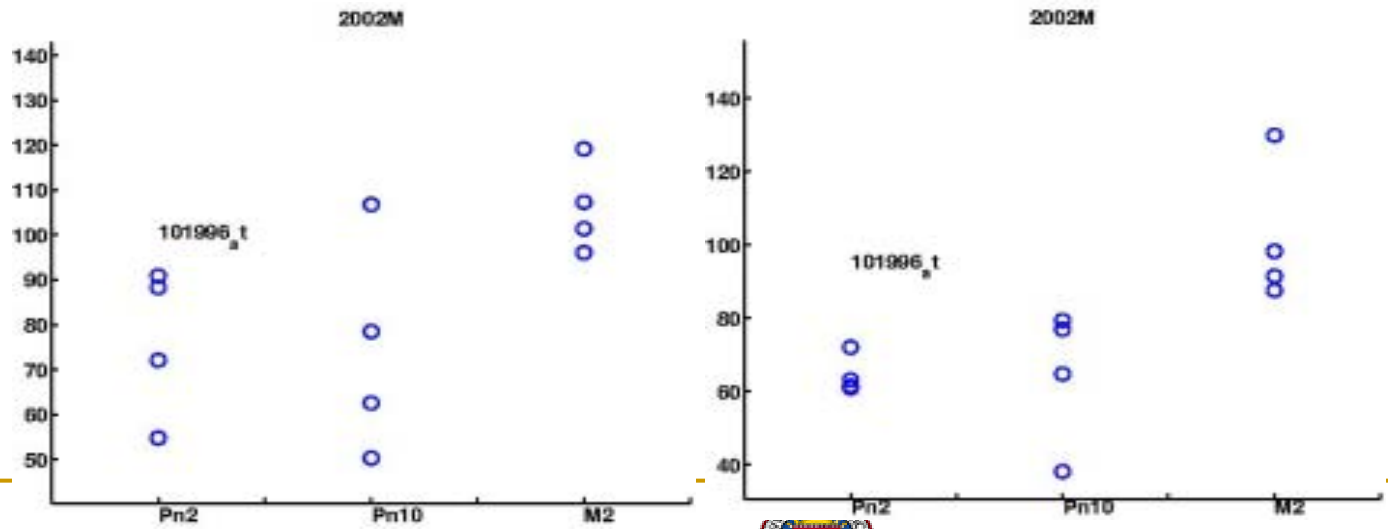
  - FDRCI: (1-α) CI on discovered fc (Benjamini&Yekutieli:2002)

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

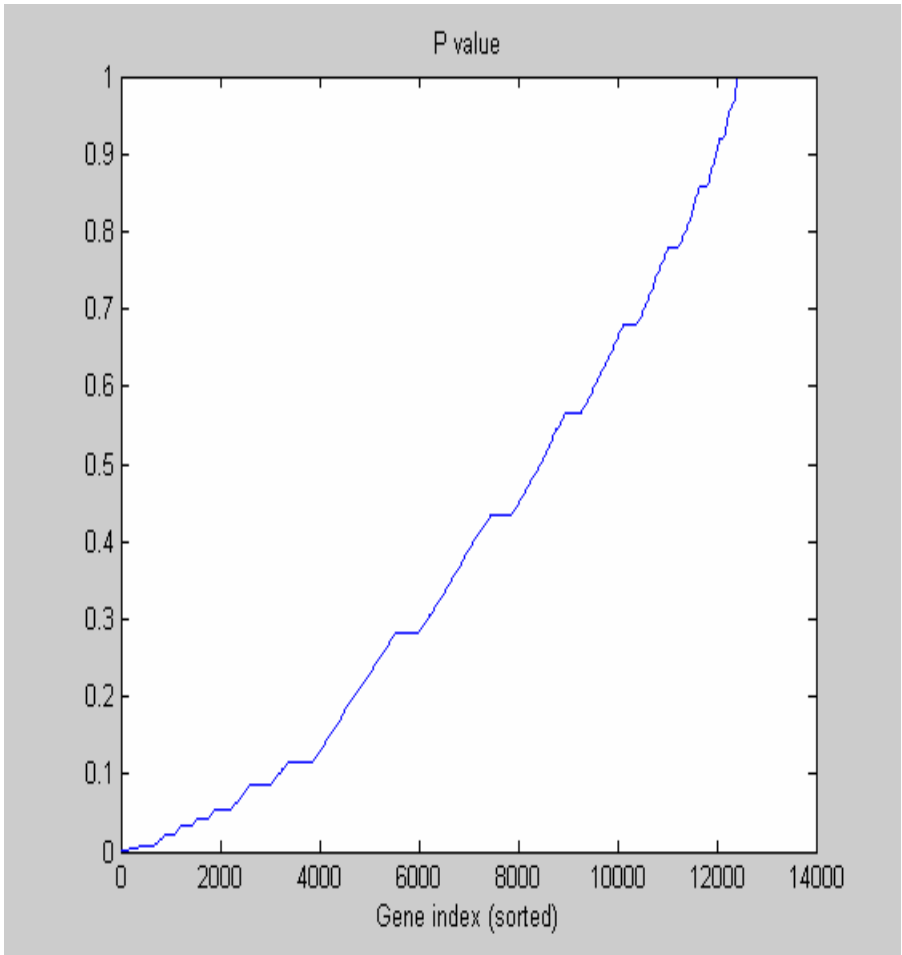
  - P: number of genes discovered at FDR=α
  - $I_g(\alpha)$ standard level 1-α CI on $\mathrm{fc}(g)$

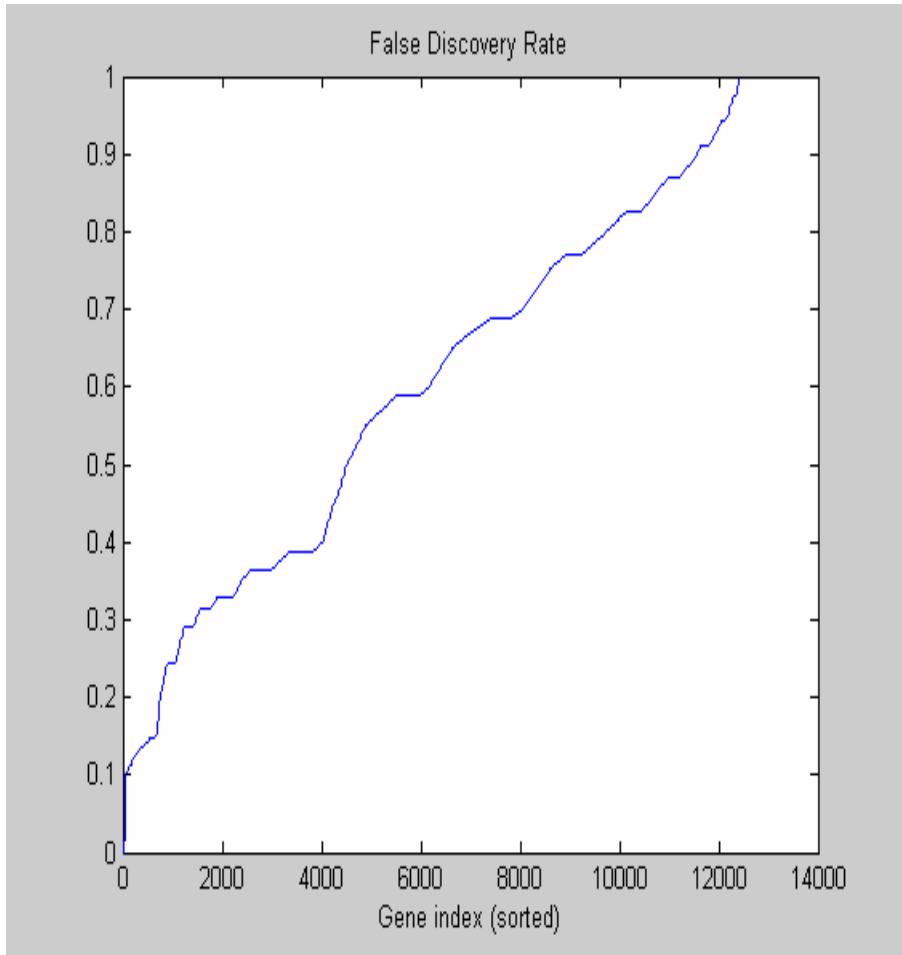
Simultaneous Testing Procedure: ko/wt Data

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



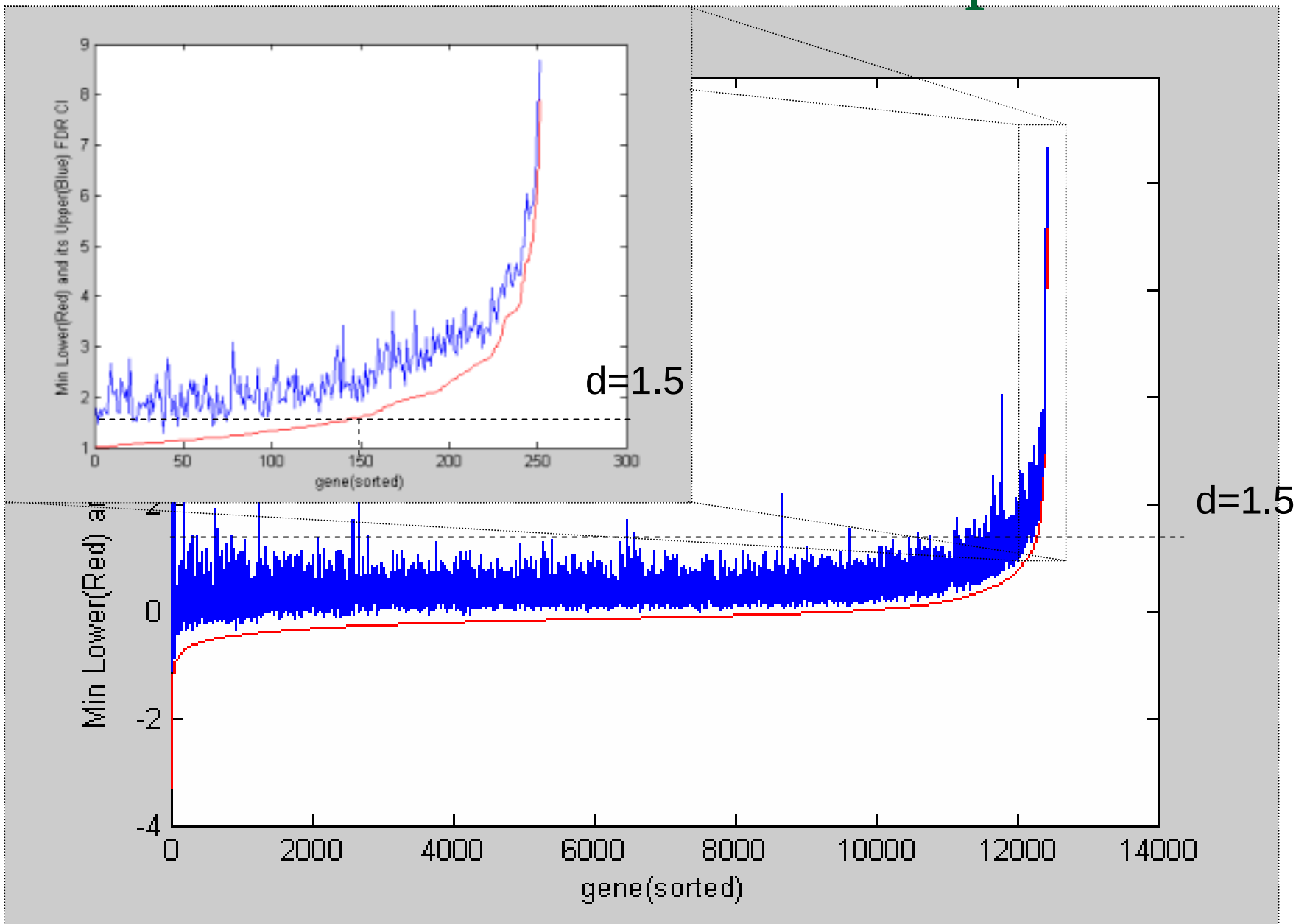
P-value vs FDR Comparison for wt/ko



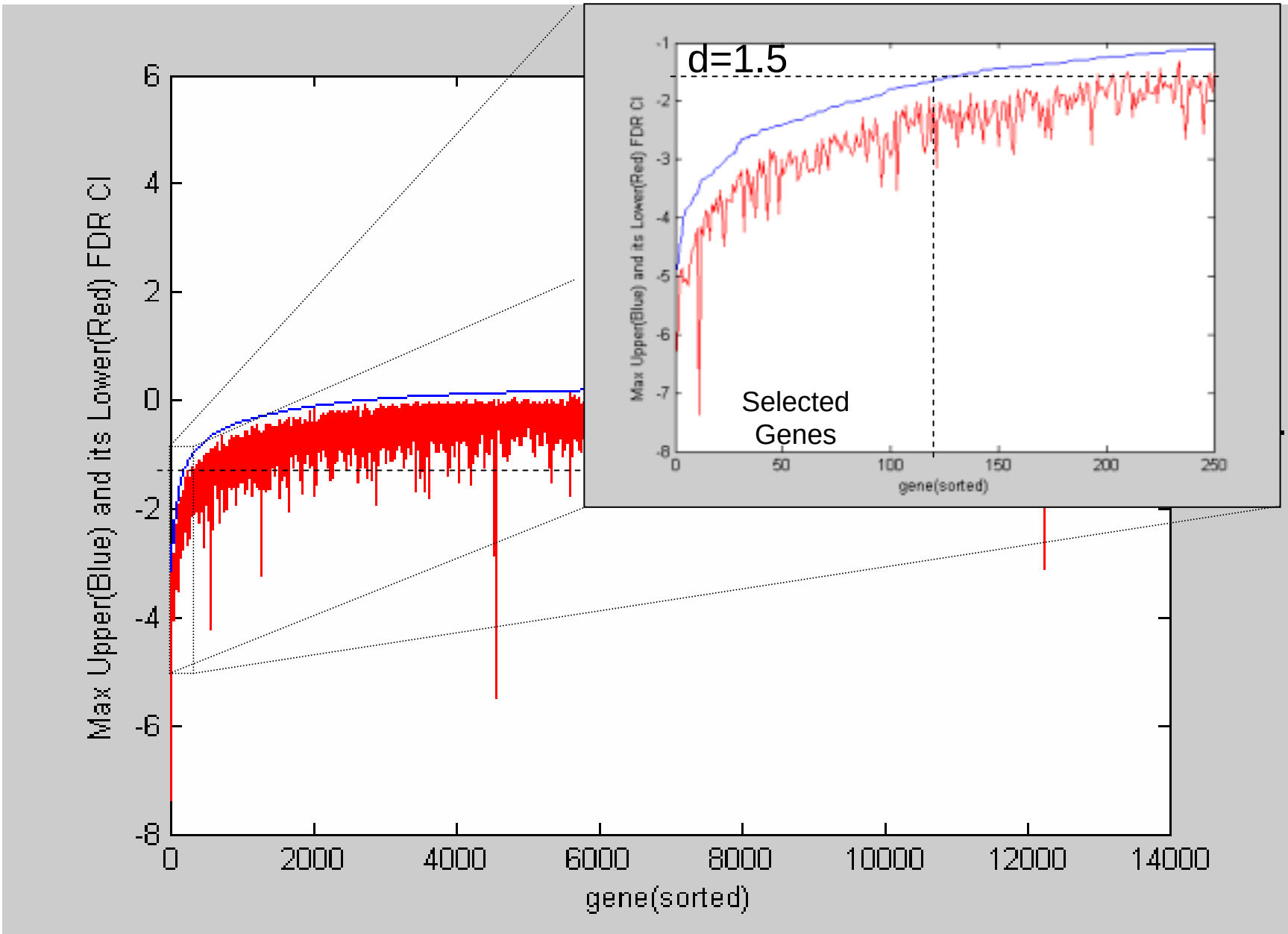


Ref: Hero:ISSPA2003

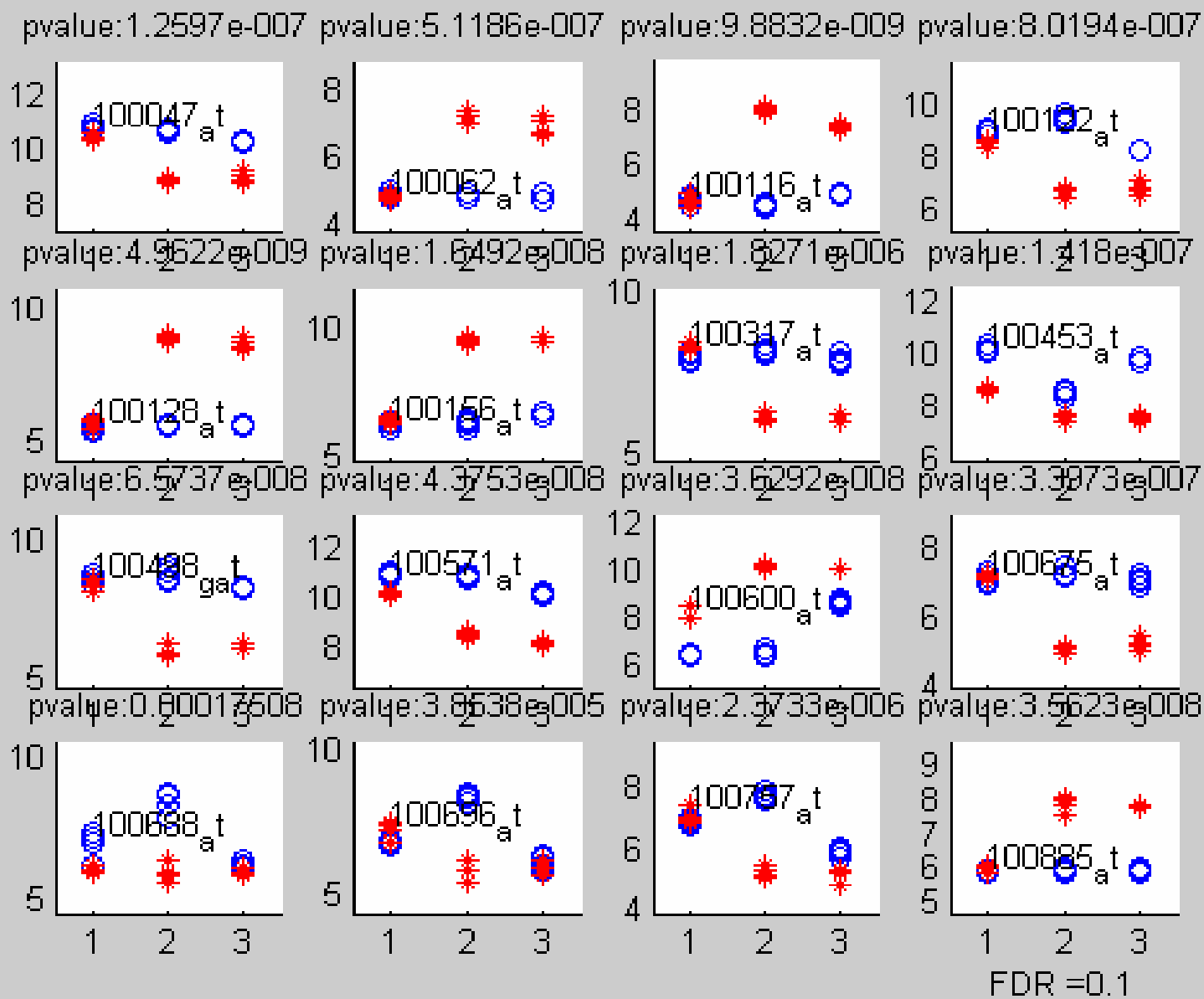
FDRCI Results for wt/ko Experiment



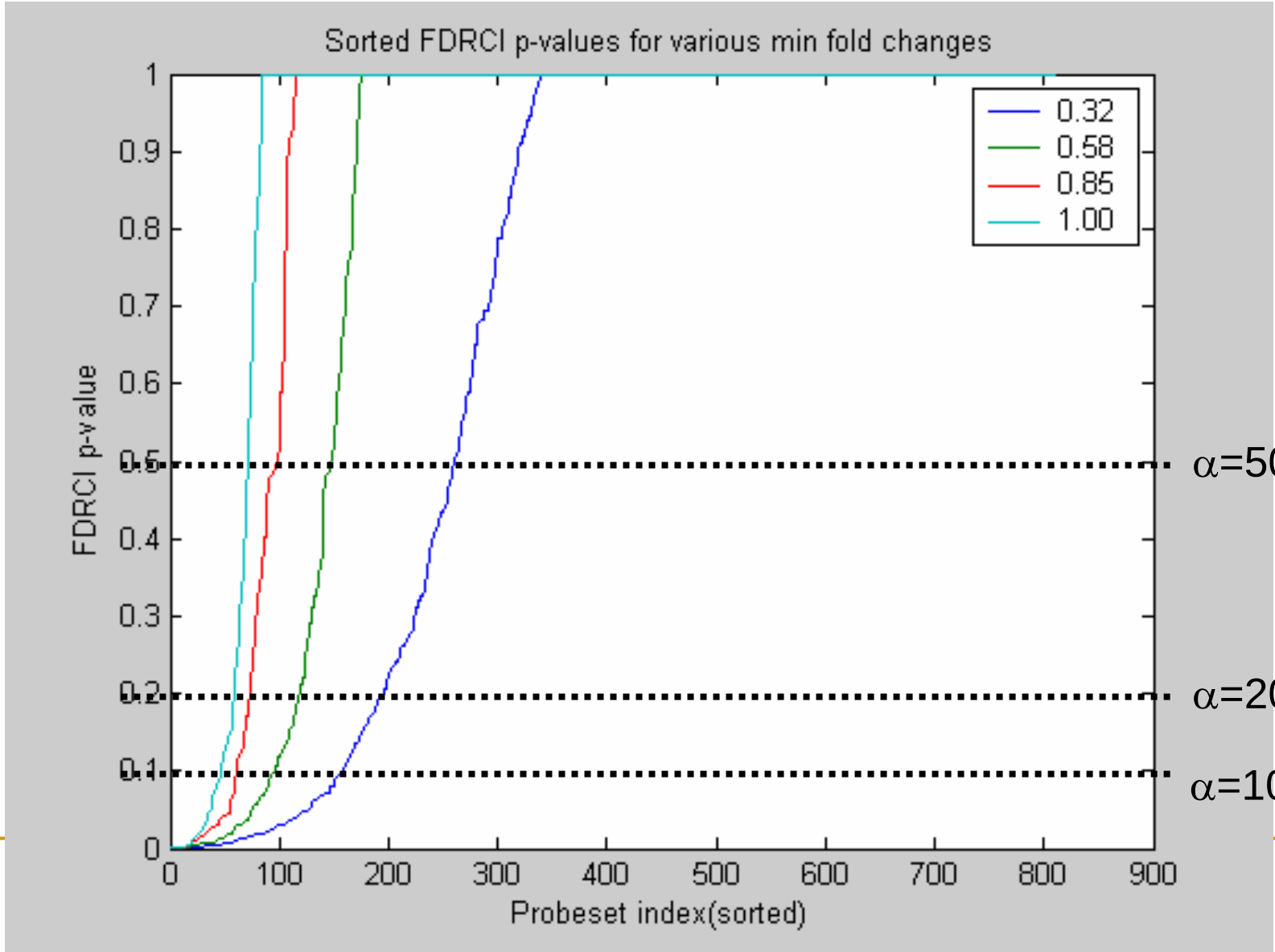
FDRCI Results for NRL Data



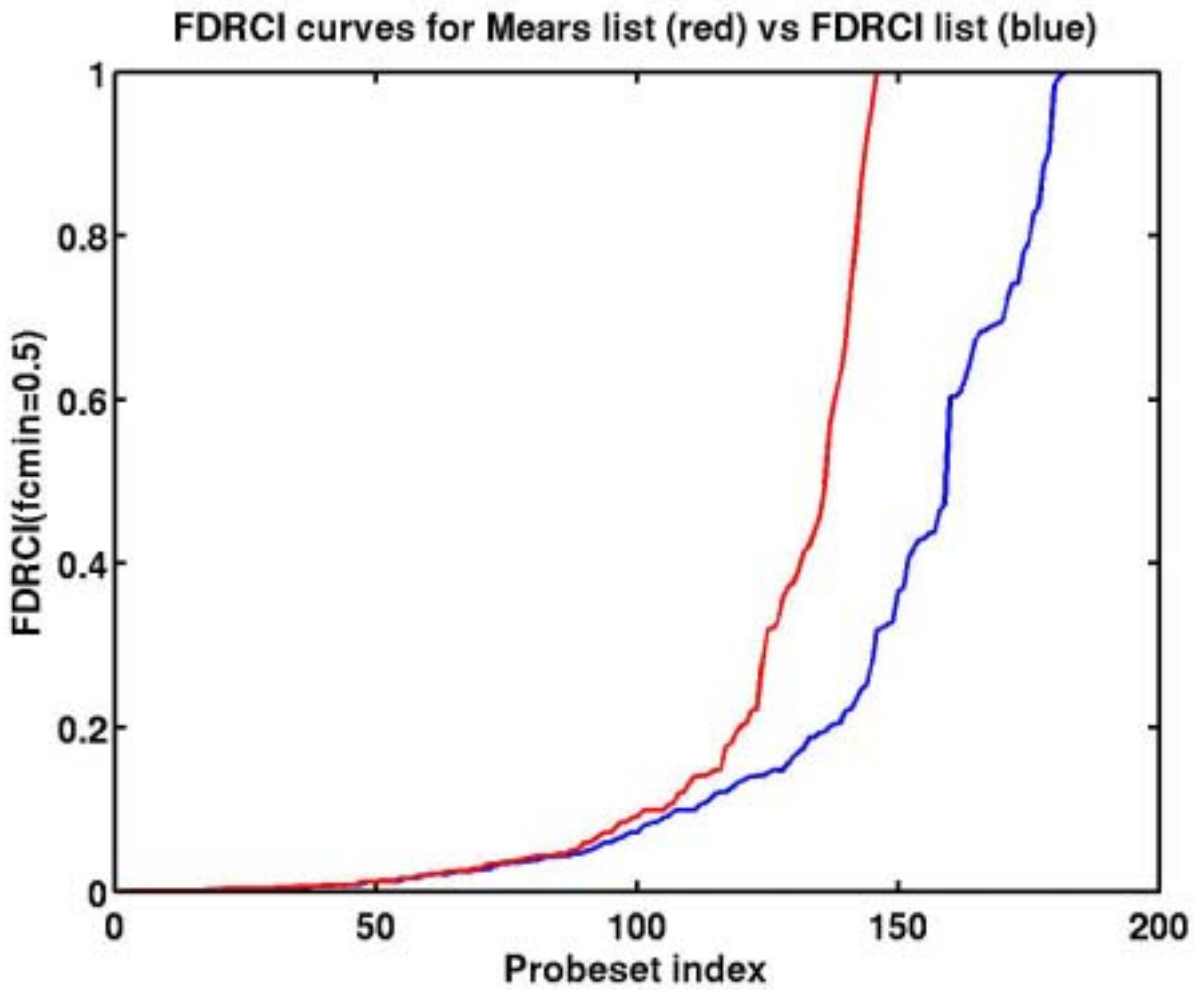
FDRCI Results for NRL Data



Sorted FDRCI pvalues for ko / wt study



Mears probes	FDRCI@0.5	FDRCI probes	FDRCI@0.5
'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_	0.01	'161525_f_	
'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



'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_at	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_at	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_at	0.21	'102413_at	0.14
'99623_s_a	0.22	'94739_at'	0.14
'104761_at	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_at	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_at	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_at	0.21
'96156_at'	0.67	'99623_s_a	0.22
'96586_at'	0.74	'104761_at	0.22
'101702_at	0.79	'98329_at'	0.24
'93457_at'	0.86	'99586_at'	0.25
'160894_at	0.92	'99461_at'	0.25
'104299_at	0.96	'98569_at'	0.28
'100348_at	1	'92770_at'	0.32
'100688_at	1	'102835_at	0.32
'101465_at	1	'93354_at'	0.33
'102393_at	1	'160808_at	0.33
'104518_at	1	'97732_at'	0.37
'160610_at	1	'160937_at	0.37
'160901_at	1	'95397_at'	0.41
'93391_at'	1	'94258_at'	0.42
'93606_s_a	1	'101191_at	0.43
'94255_g_a	1	'101489_at	0.43
'97142_at'	1	'100757_at	0.44
'98004_at'	1	'95453_f_at	0.44
'99126_at'	1	'93011_at'	0.46
		'160414_at	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

Quantitative comparisons

- Wt vs 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 ≥ 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.

V. Pareto Optimal Gene Ranking: Pareto Front Analysis (PFA)

- 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:

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

- Examples of mixed increasing and decreasing

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

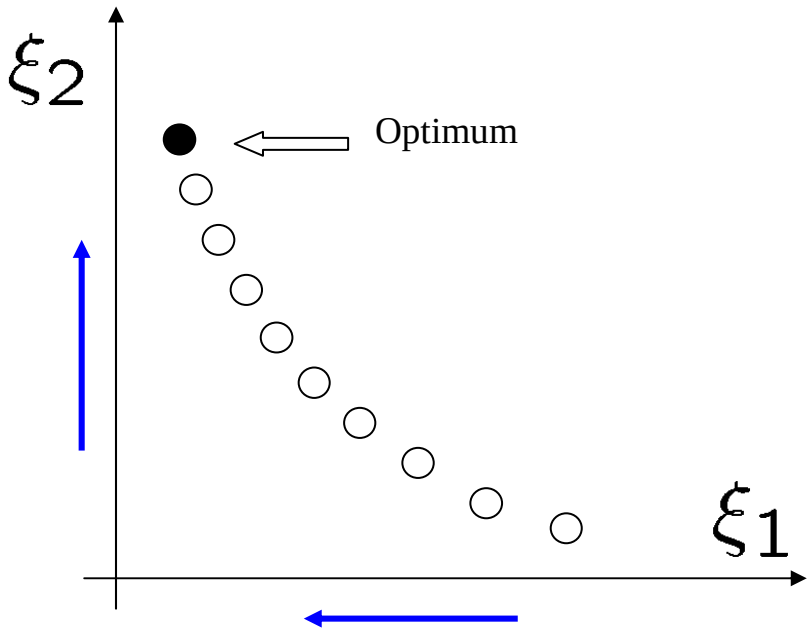
$$\{\mathsf{fc}_1(g), \mathsf{fc}_2(g), \mathsf{fc}_3(g)\}_{g \in \mathcal{G}_1} > \ldots > \{\mathsf{fc}_1(g), \mathsf{fc}_2(g), \mathsf{fc}_3(g)\}_{g \in \mathcal{G}_q}$$

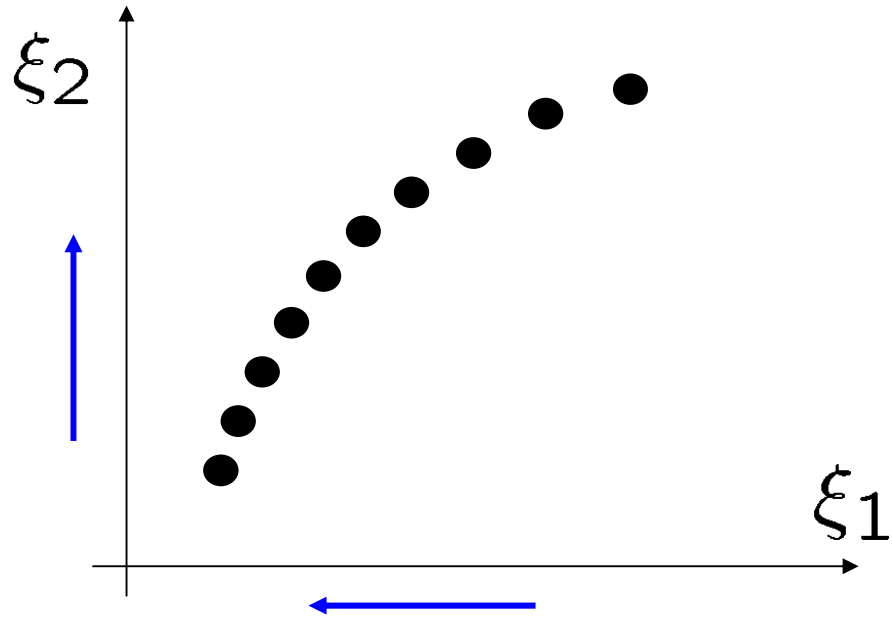
Illustration of two extreme cases

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

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

- A linear ordering exists
- No partial ordering exists





Comparison to Criteria Aggregation

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

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

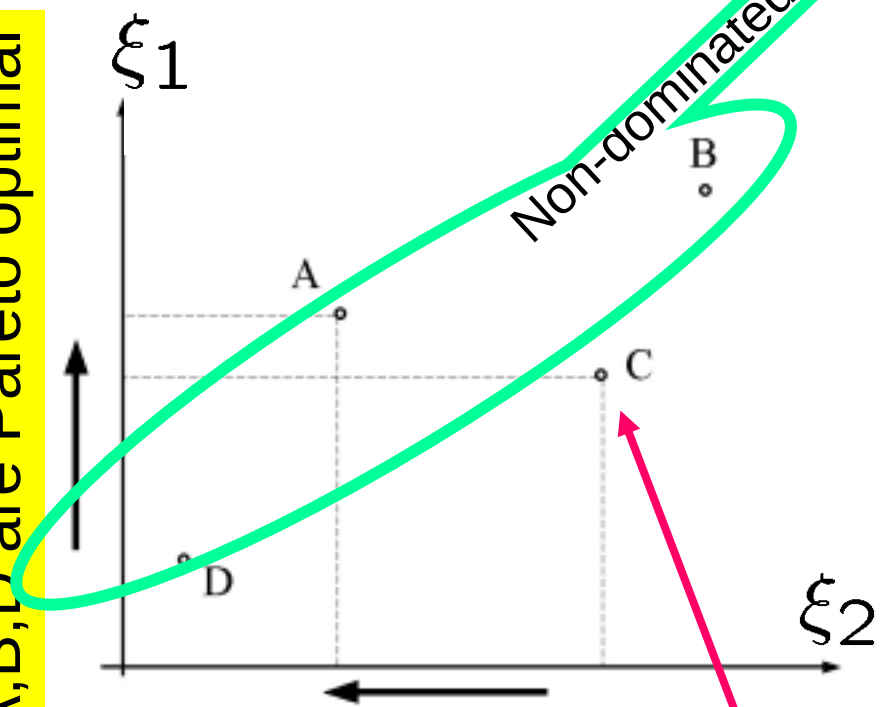
- Order genes according to ranks of

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

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

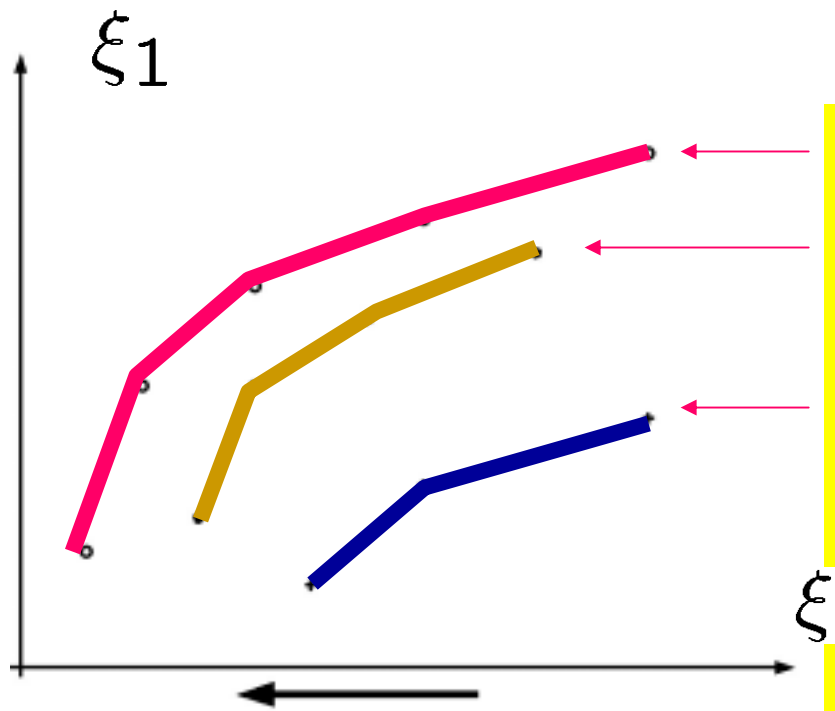
Multicriteria Gene Ranking

- Increasing ξ_1
- Decreasing ξ_2



Non-dominated genes=Pareto Front

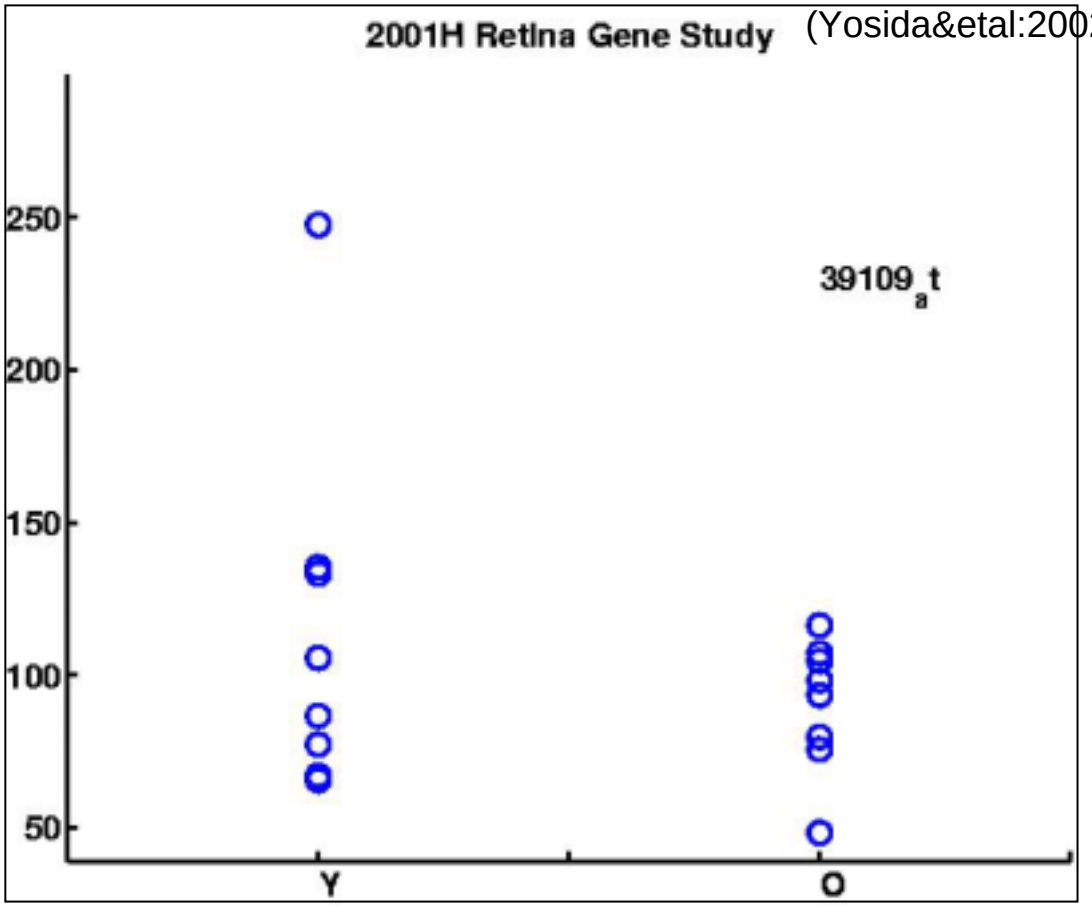
A,B,D are Pareto optimal



Pareto Fronts=partial order

Dominated gene

Ranking Based on End-to-End Foldchange



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:

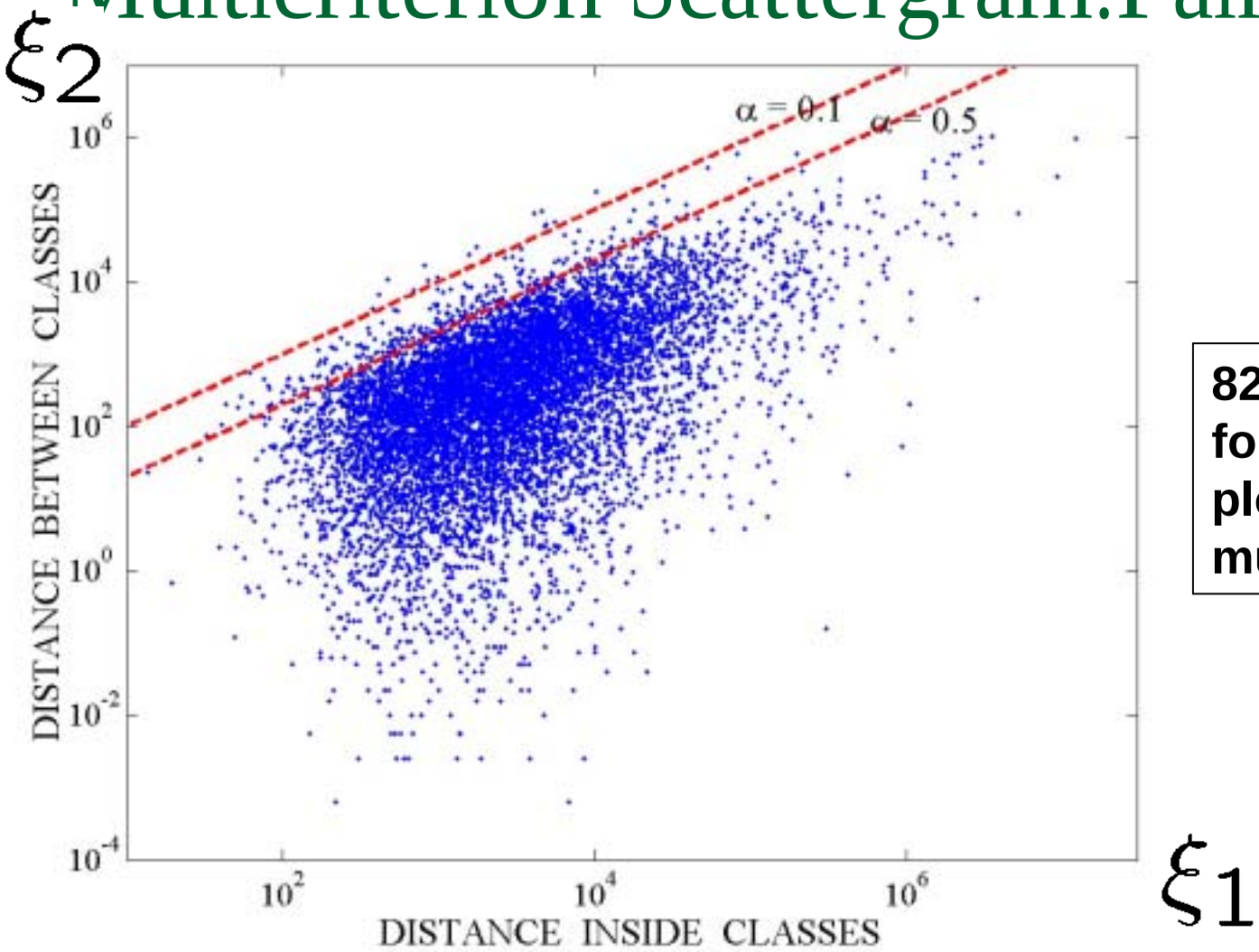
$$T(g) = \frac{\xi_2(g)}{\xi_1(g)} \begin{matrix} > \\ < \end{matrix} \sqrt{2/m}\, \mathcal{T}^{-1}_{1-\alpha/2}$$

- 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

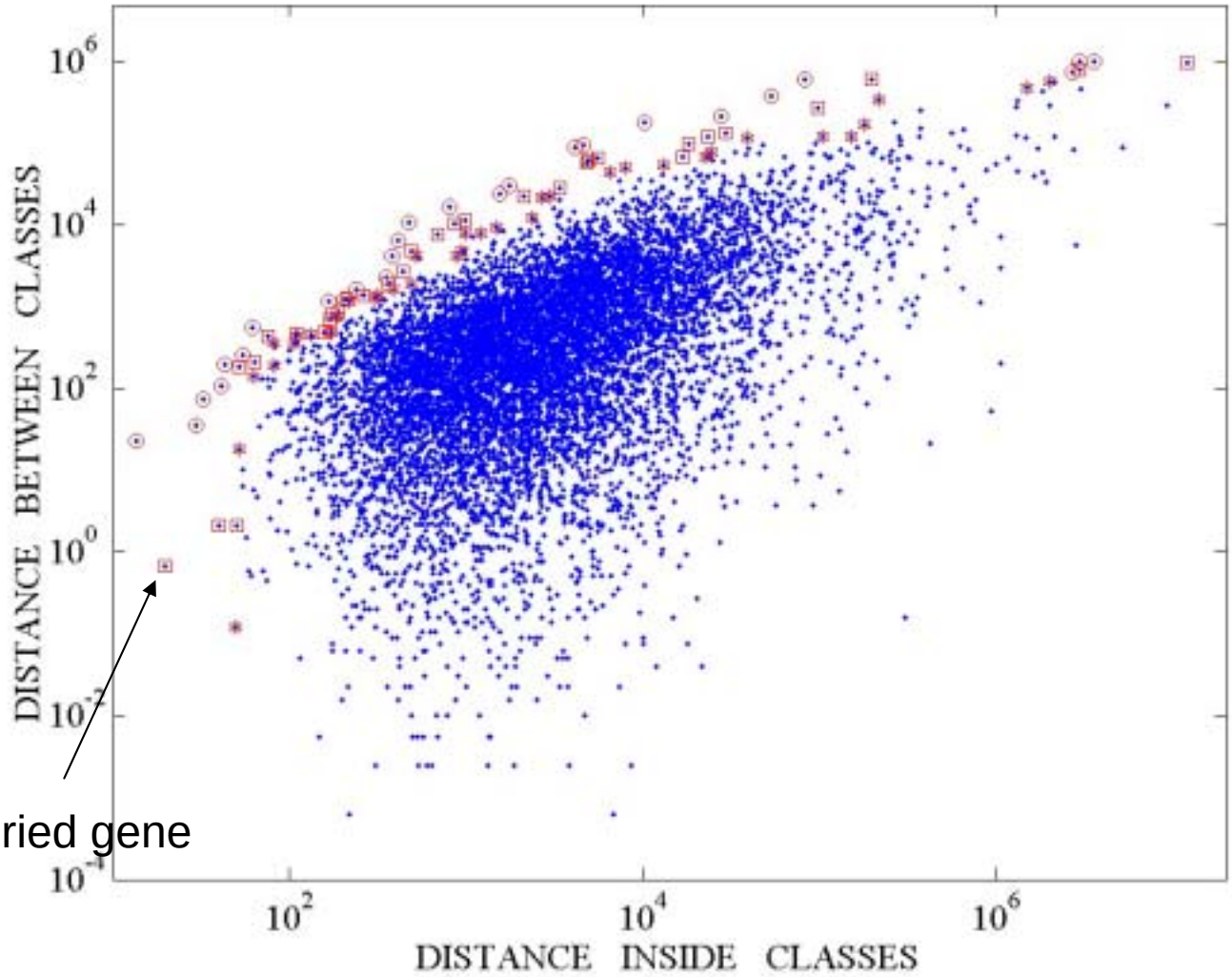
Multicriterion Scattergram:Paired t-test



8226 Y/O mean foldchanges plotted in multicriteria plane

Ref: Fleury&etal ICASSP-02

Multicriterion scattergram: Pareto Fronts



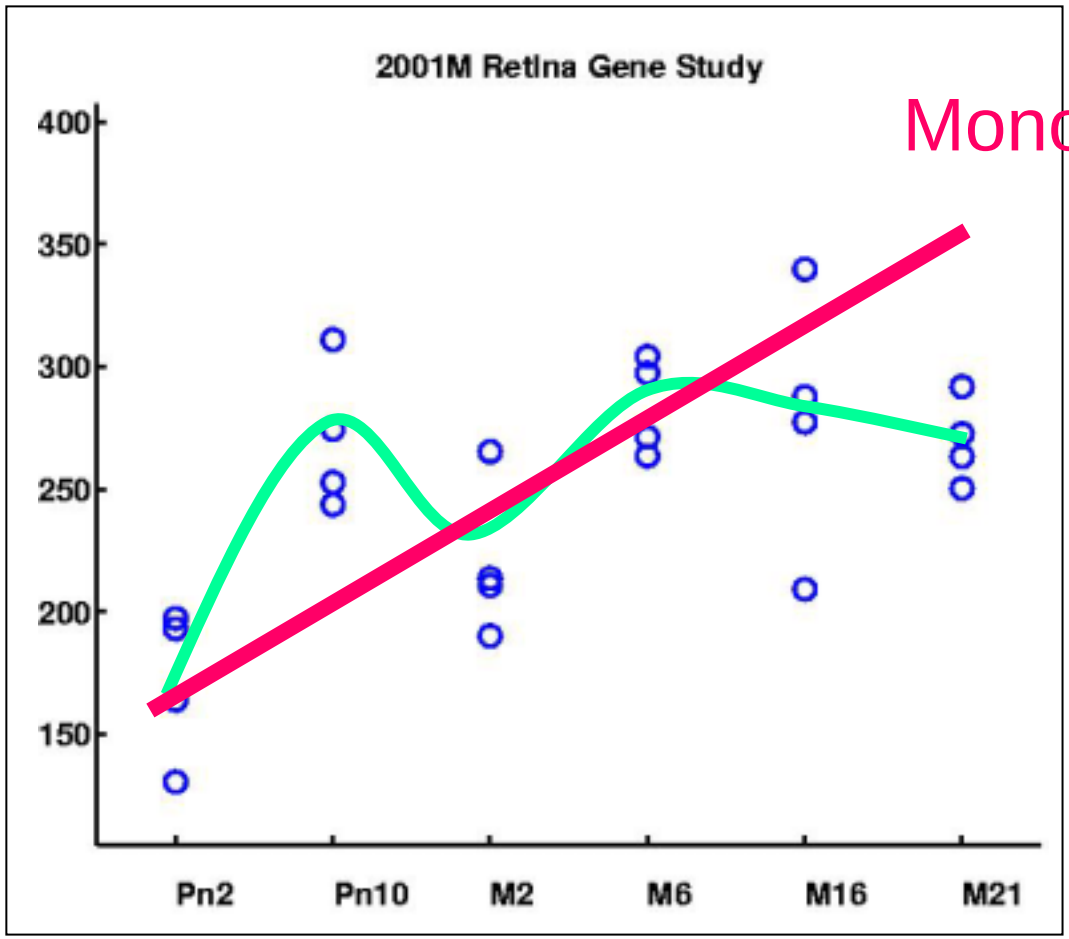
Buried gene

Pareto fronts

- ○ *first*
- □ *second*
- ☆ *third*

Ref: Fleury&etal ICASSP-02

Ranking Based on Profile Shape



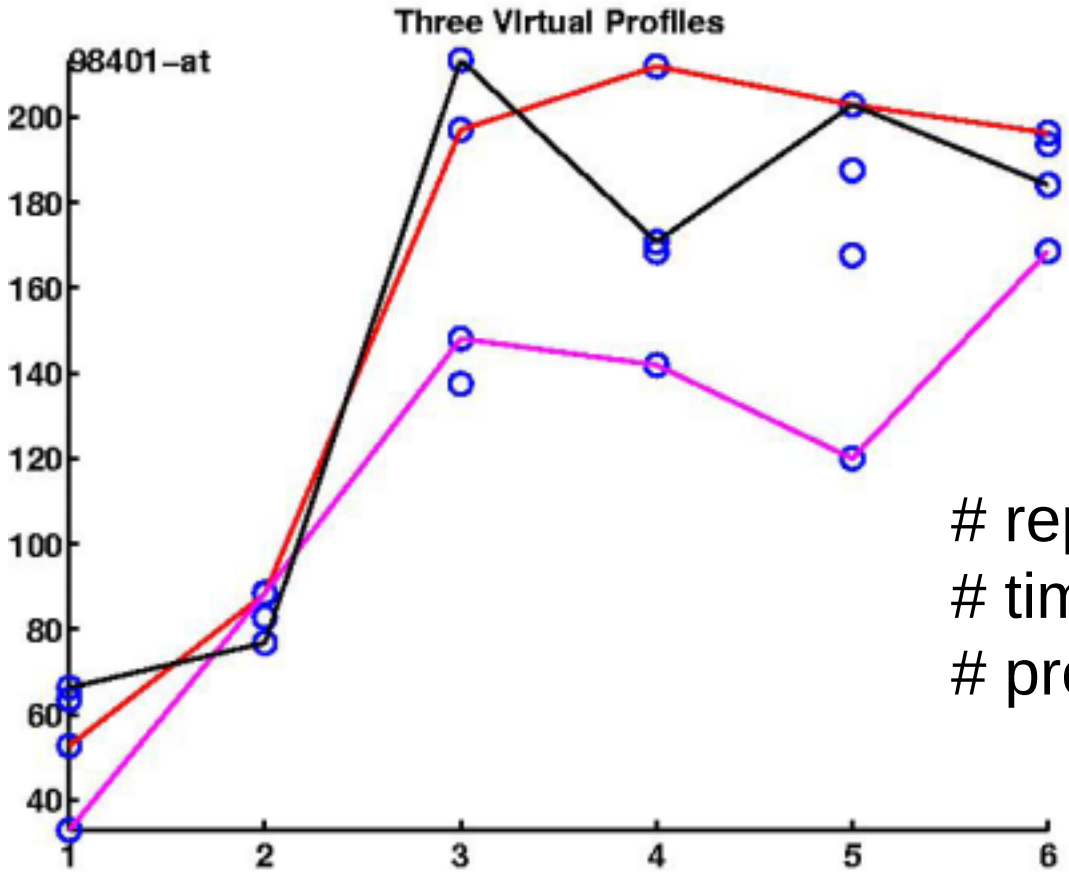
Monotonic?

Mouse Retina Aging Study

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

Ref: Hero and Fleury:VLSI03

Jonckheere-Terpstra Statistic



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

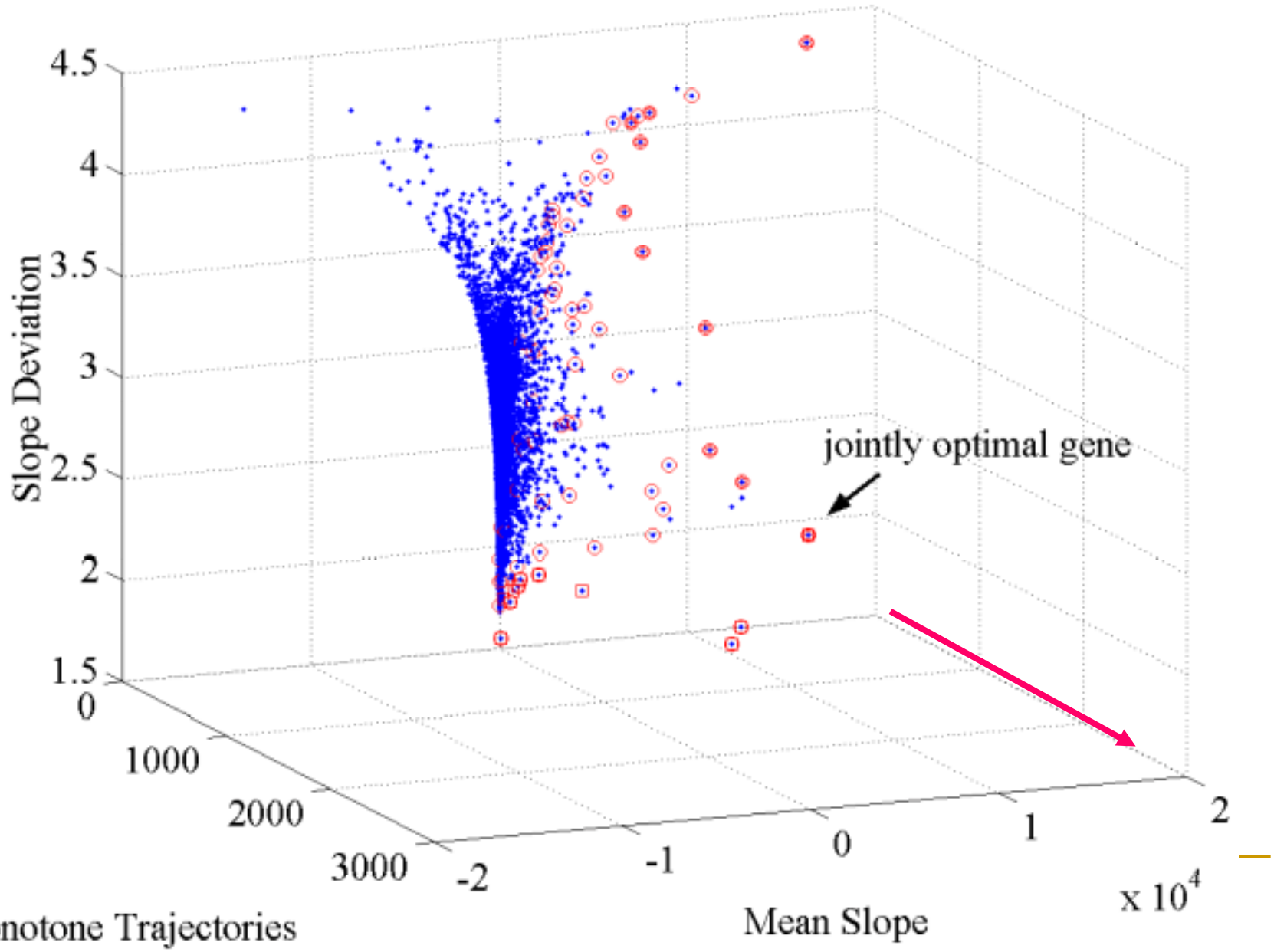
$$\xi_1(g) = \sum_{t} \sum_{t'>t} \sum_{m \neq m'} \text{sign}(y_{t',m'}(g) - y_{t,m}(g))$$

Hollander 2001

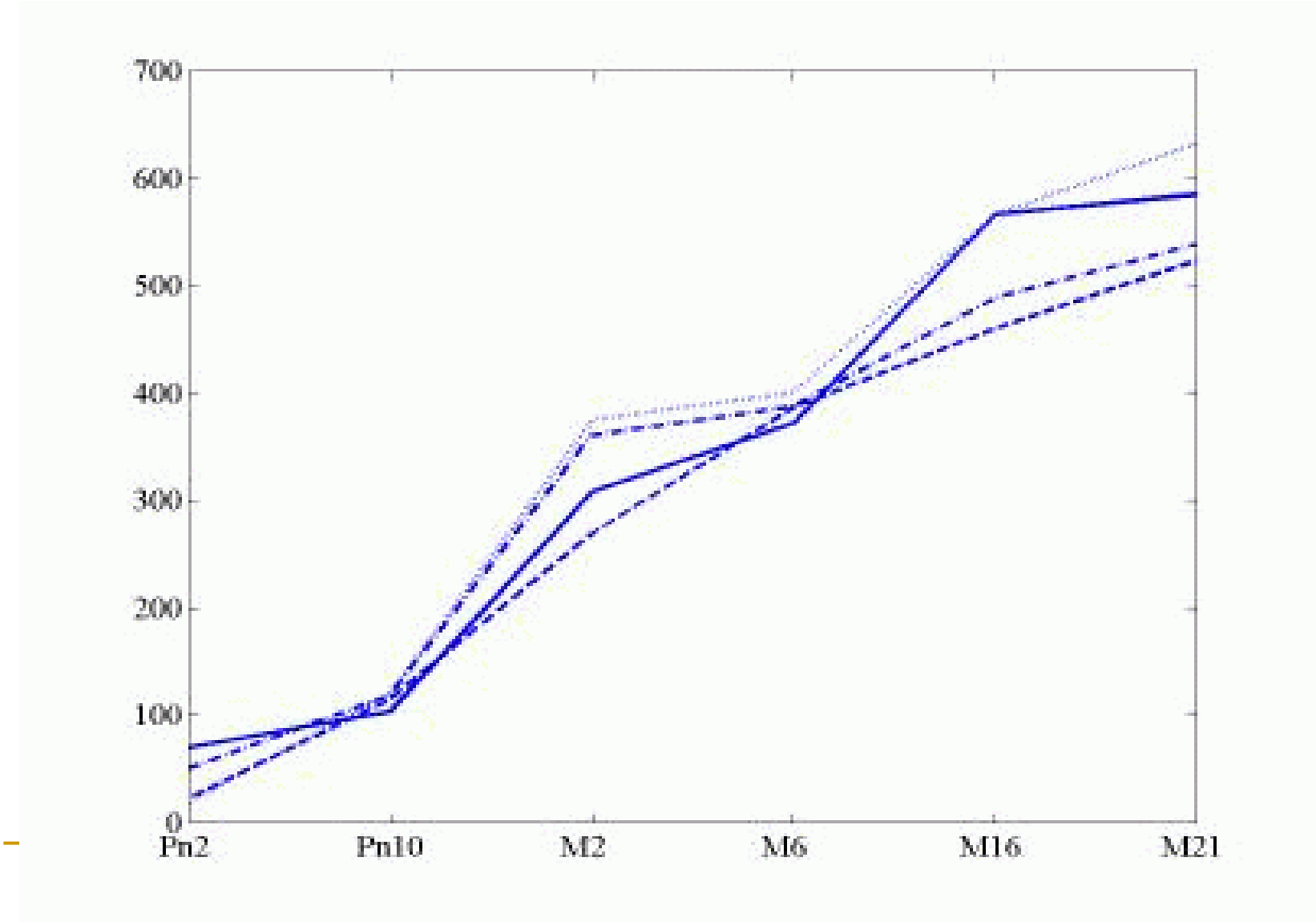
Monotonic-Profile Ranking Criteria

- 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

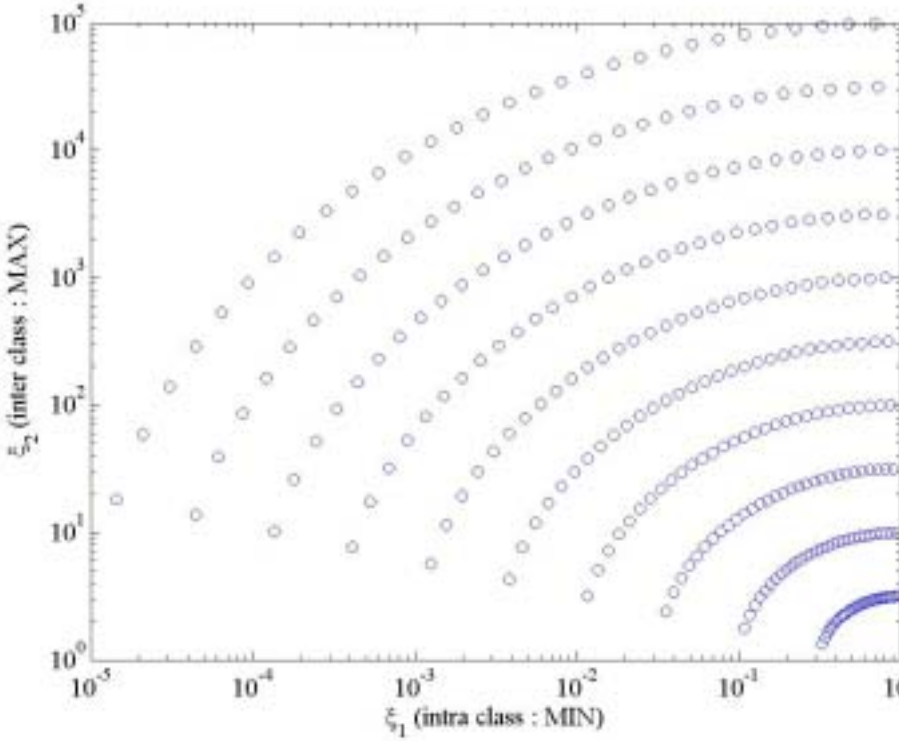


Profile of Pareto Optimal Aging Gene

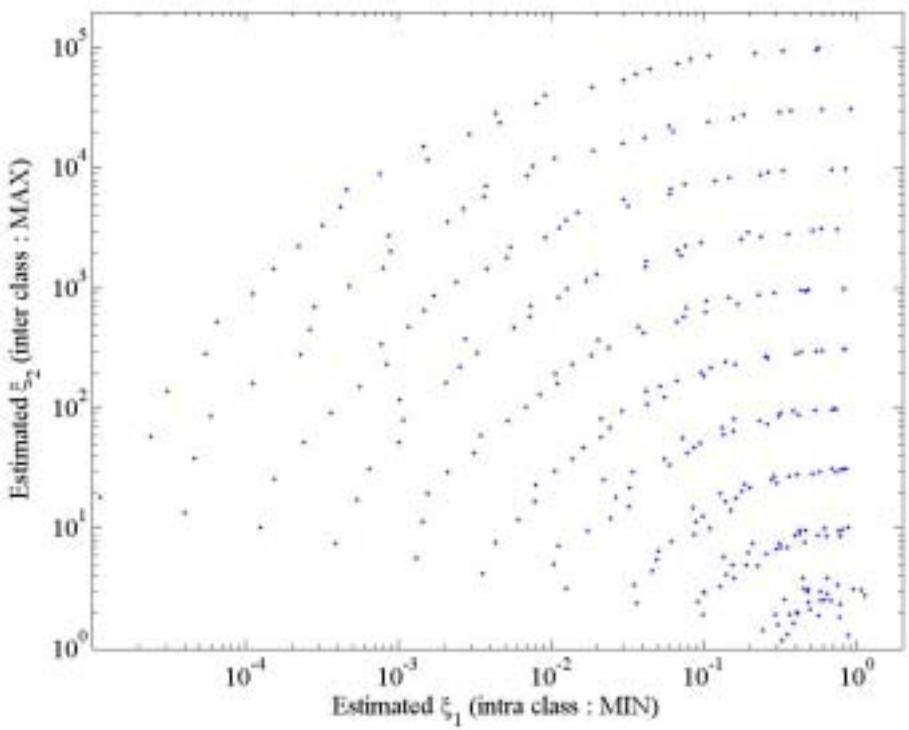


Effects of Sampling Error on Ranking

Hypothetical dual criterion planes



Ensemble mean scattergram (Ground truth)



Sample mean scattergram (Measured)

Ref: Fleury and Hero:JFI03

Accounting for Sampling Errors in PFA

- Key Concepts:
  - Pareto Depth Distribution
  - Pareto Resistant Gene

- 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:

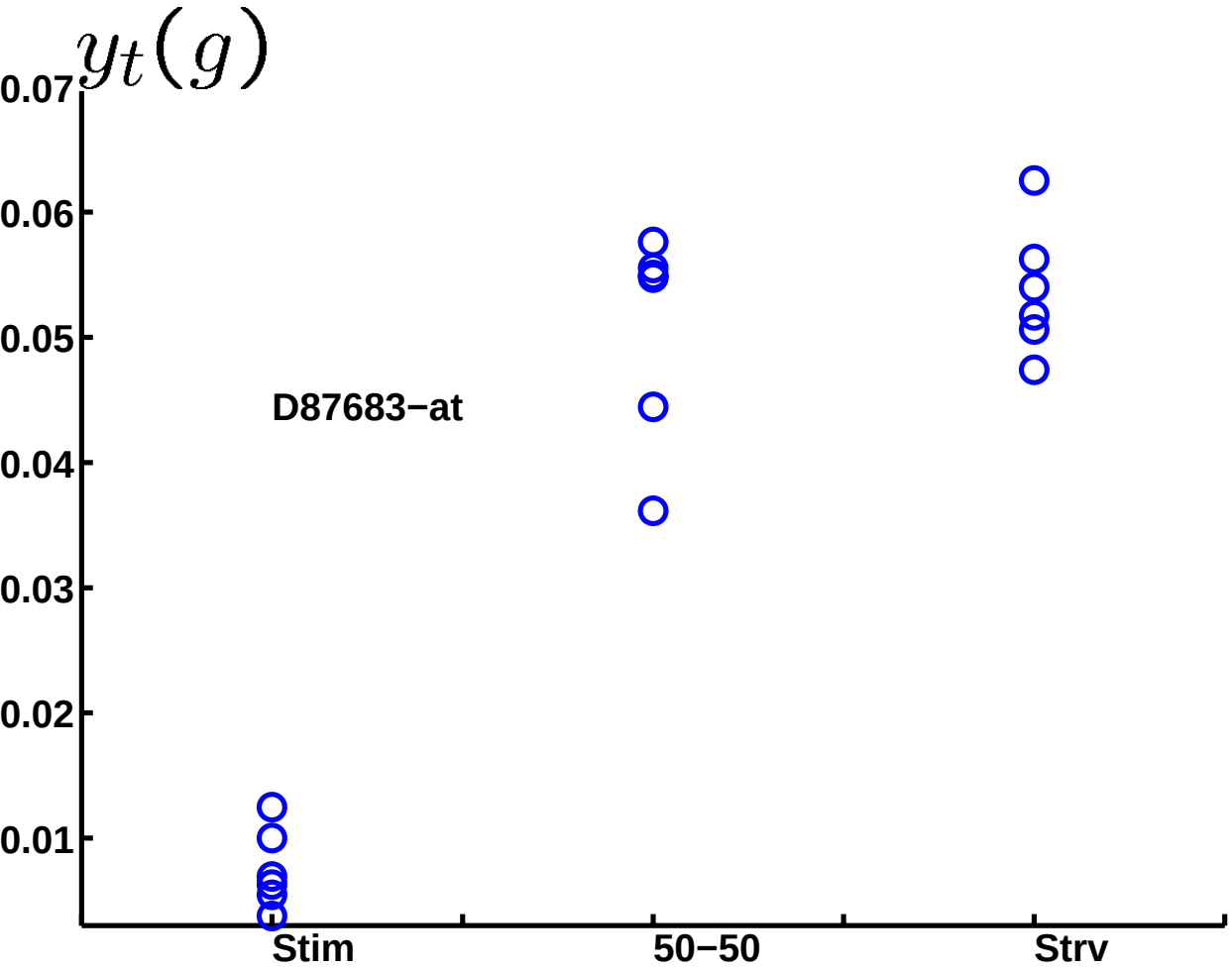
$$\xi_q(i) < \xi_q(g) \text{ and } \xi_p(i) \leq \xi_p(g),\ p \neq q.$$

- Posterior probability: gene g is on Pareto front

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

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

Application to Dilution Experiment



Human fibroblast data

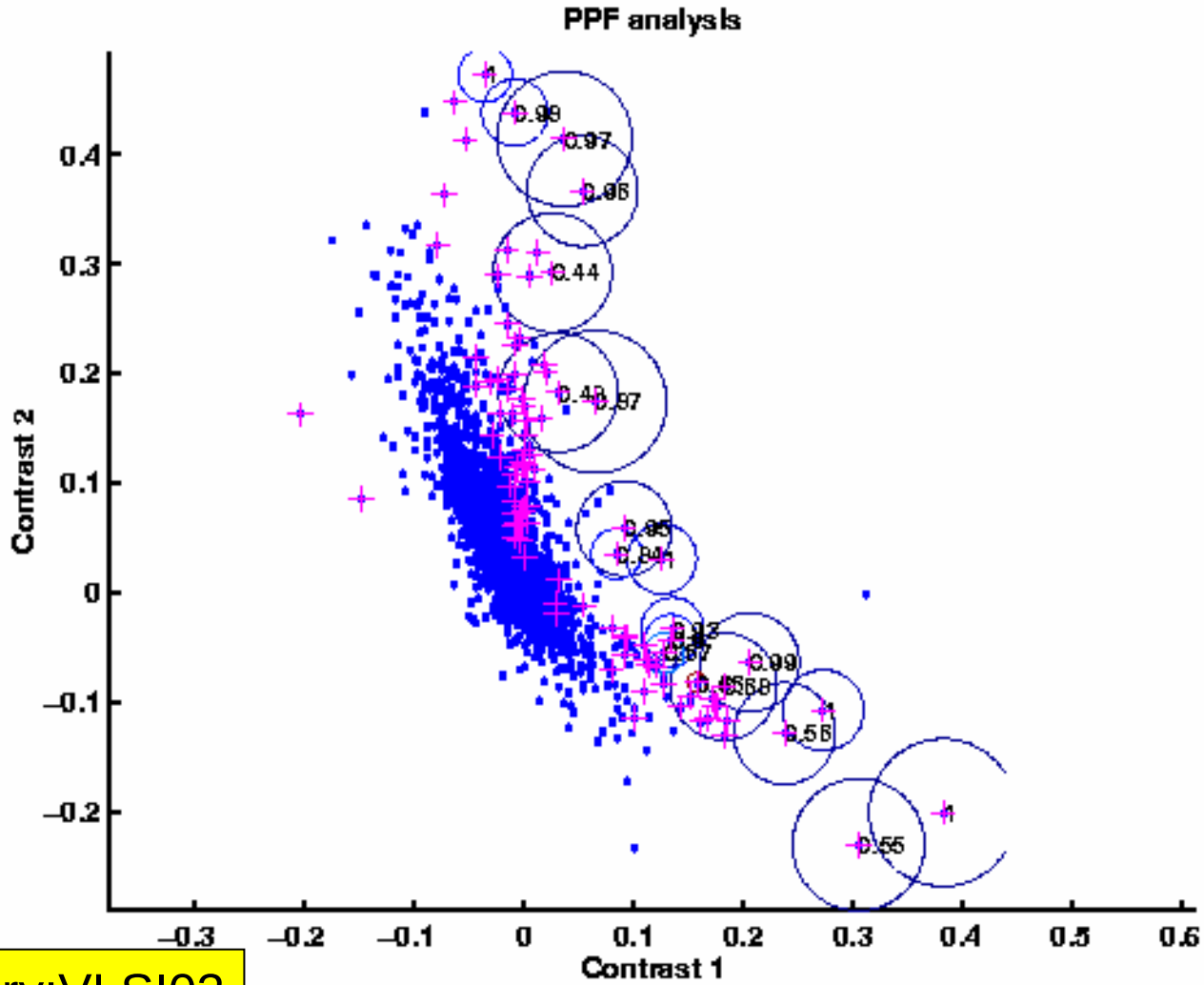
- **18 Fibroblast Cells**
- **3 treatments**
- **6 replicates**
- **7129 probesets**

$$\xi_1(g) = \bar{y}_2(g) - \bar{y}_1(g)$$
$$\xi_2(g) = (\bar{y}_2(g) + \bar{y}_1(g))/2 - \bar{y}_3(g)$$

Lemon&etal:(

Scattergram for Dilution Experiment

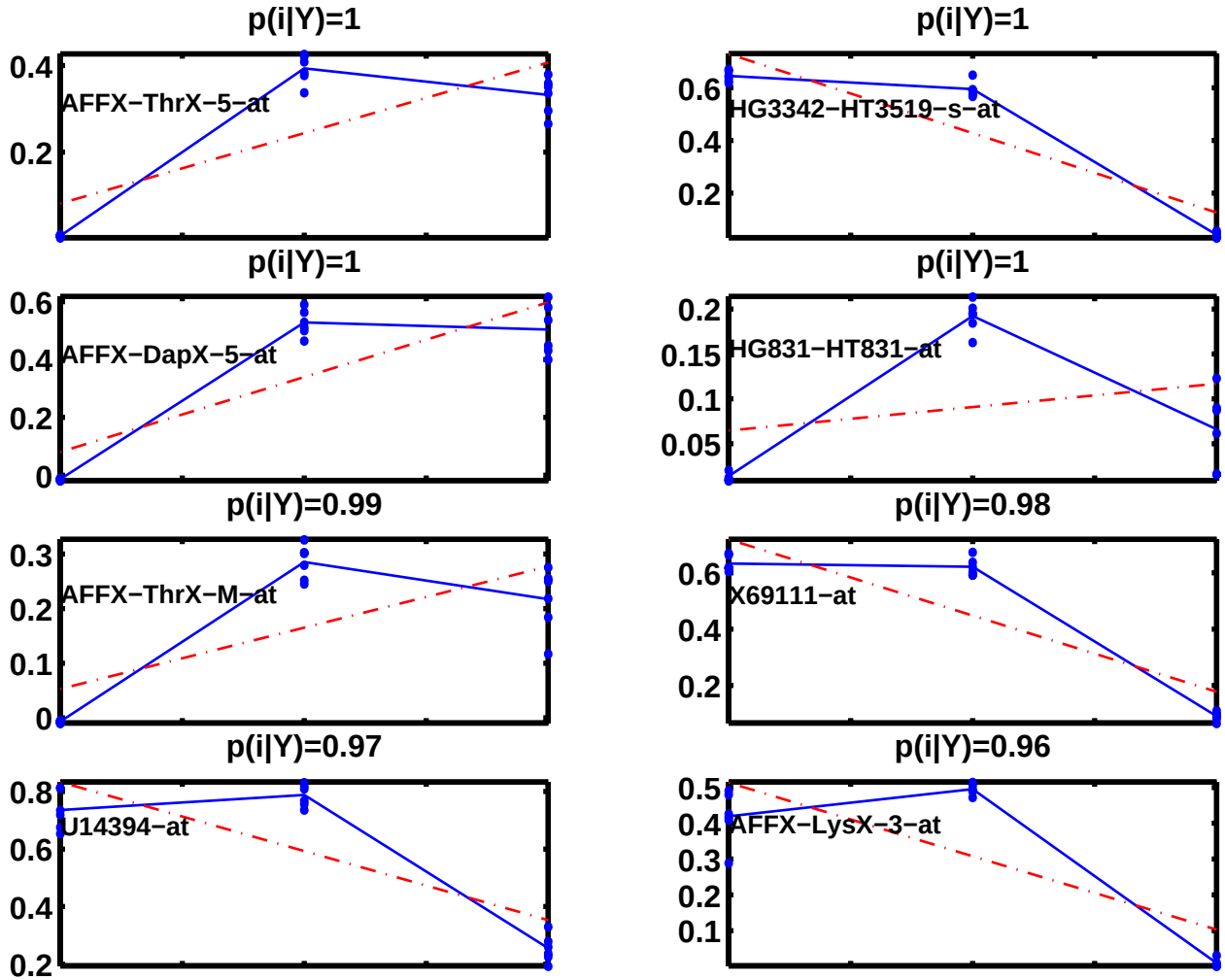
ξ_2



ξ_1

Hero&Fleury:VLSI03

Most Non-monotonic Trajectories



Hero&Fleury:VLSI03

Pareto Depth Sampling Distribution

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

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

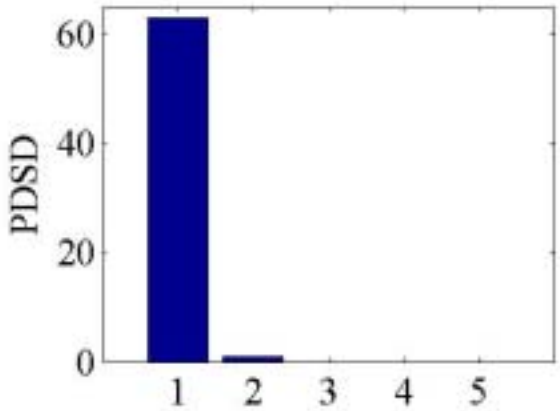
- (Re)sampling distribution of Pareto depth

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

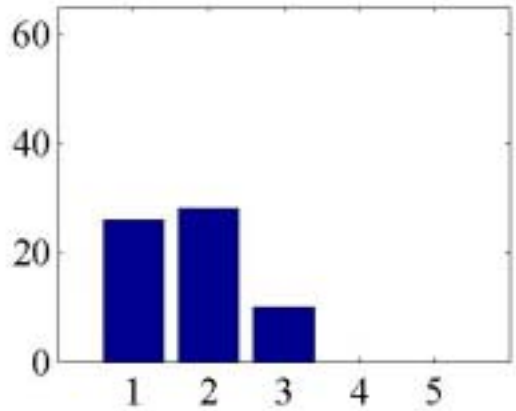
Ref: Fleury and Hero:JFI03

PDSD Examples for 4 different genes

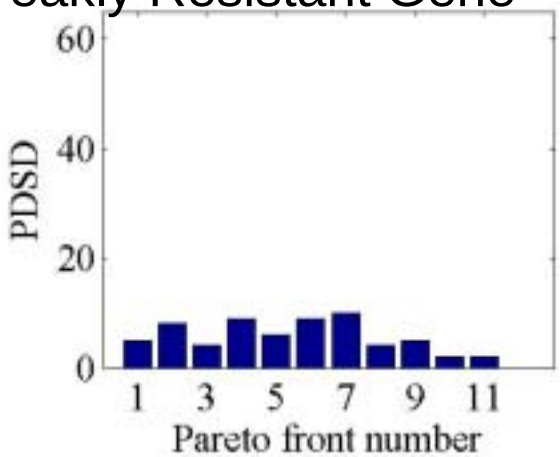
Stongly Resistant Gene



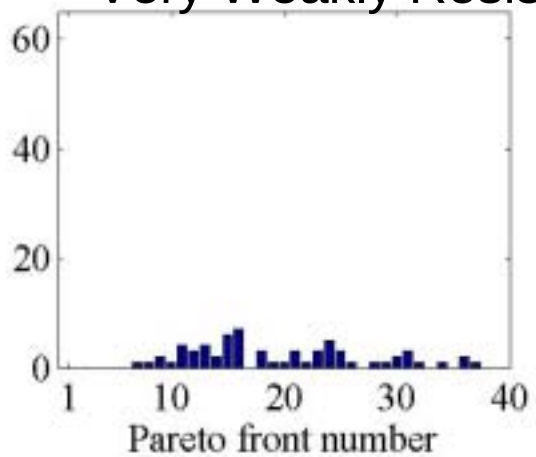
Moderately Resistant Gene



Weakly Resistant Gene

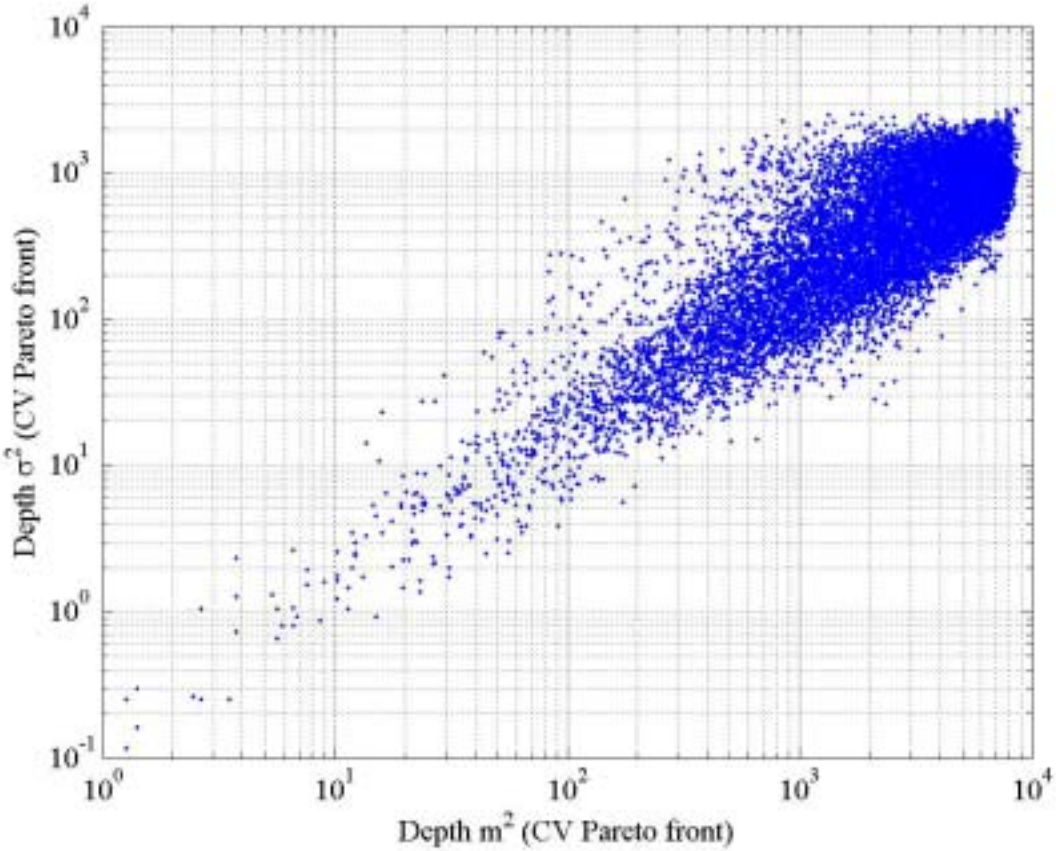


Very Weakly Resistant Gene



Ref: Fleury and Hero:JFI03

2nd Moment of Inertia (MOI) of PDSD:



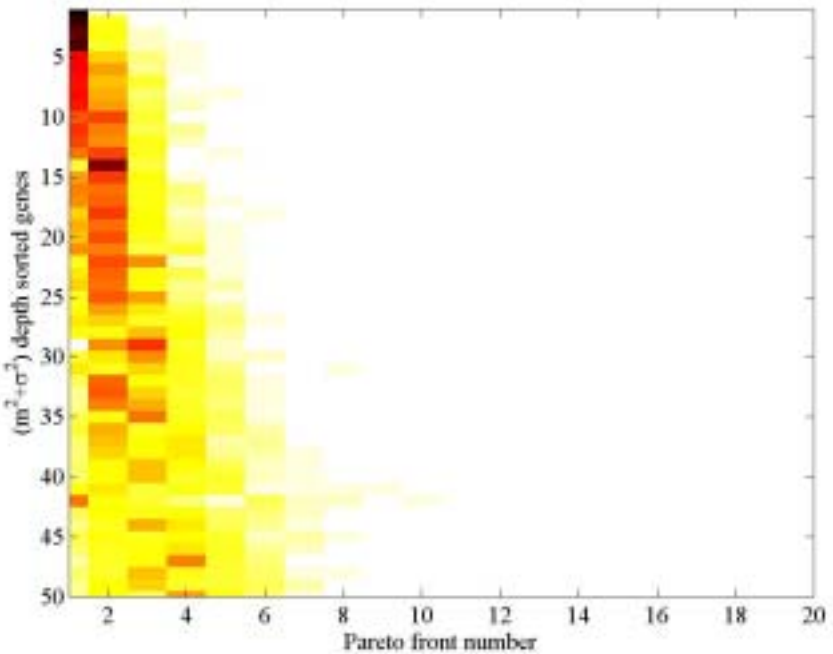
2nd MOI defined as:

$$T_g = \sum_{k=1}^{N} k^2 \mathrm{Pdsd}_g(k) = \sigma^2(g) + \mu^2(g)$$

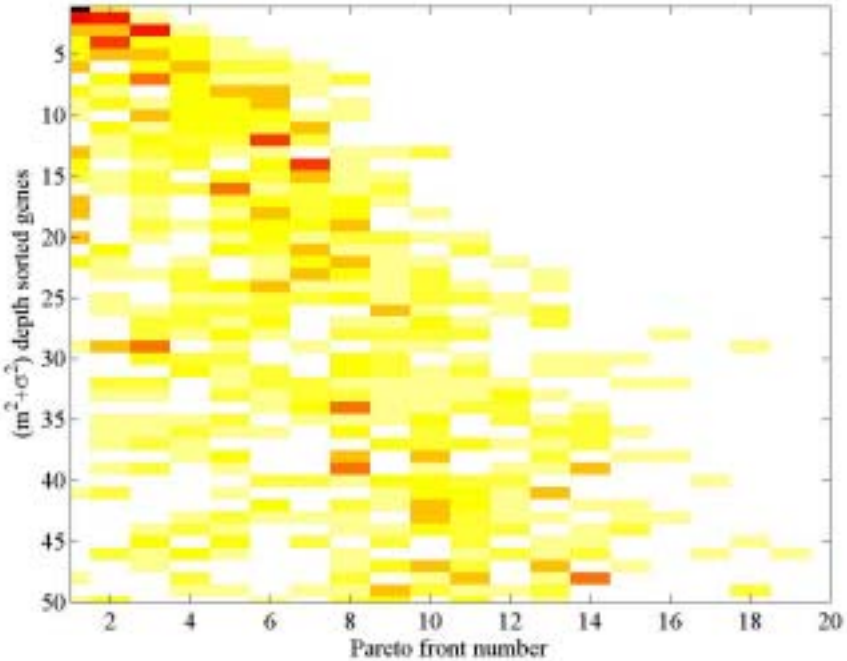
Ref: Fleury and Hero:JFI03

PDSD Gene Ranking Illustration

- Top 50 genes according to PDSD 2nd MOI



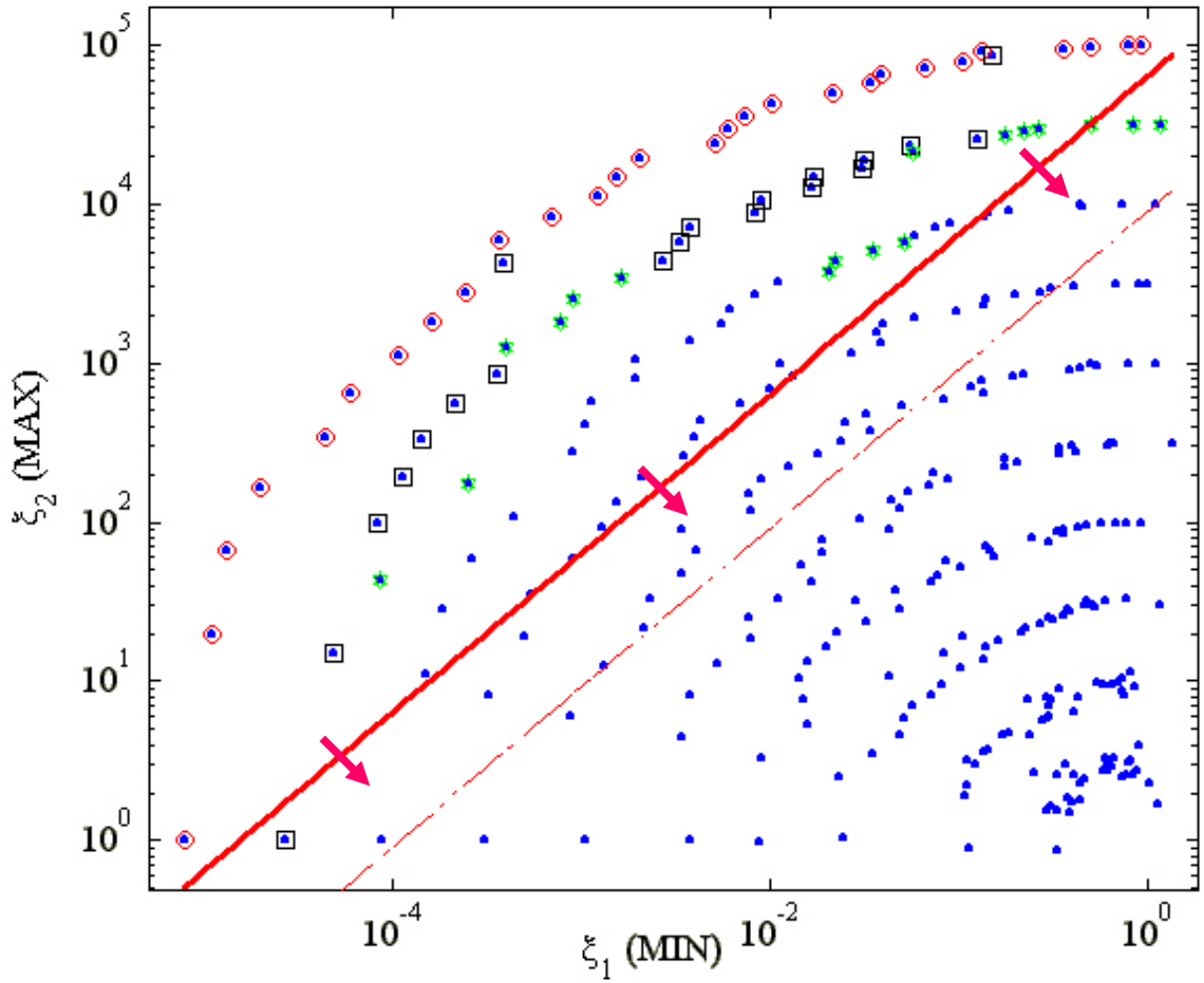
Affy Human Y/O Retina Data Set



Affy Mouse Retina Aging Data Set

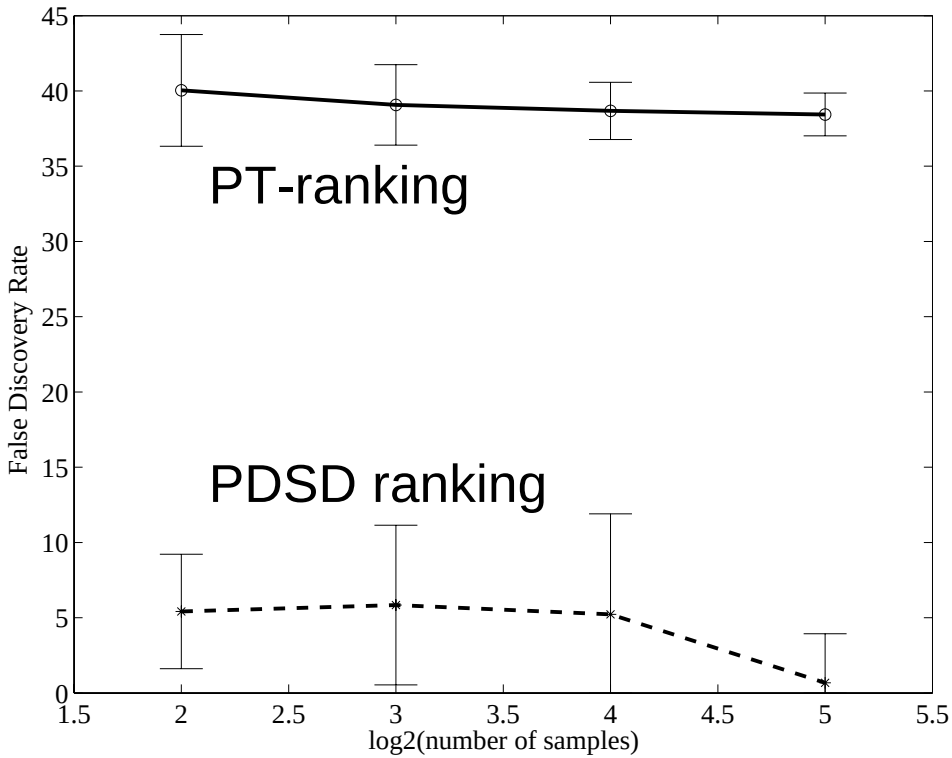
Ref: Fleury and Hero:JFI03

Quantitative Comparison: Pareto Front vs. Paired T Test ranking

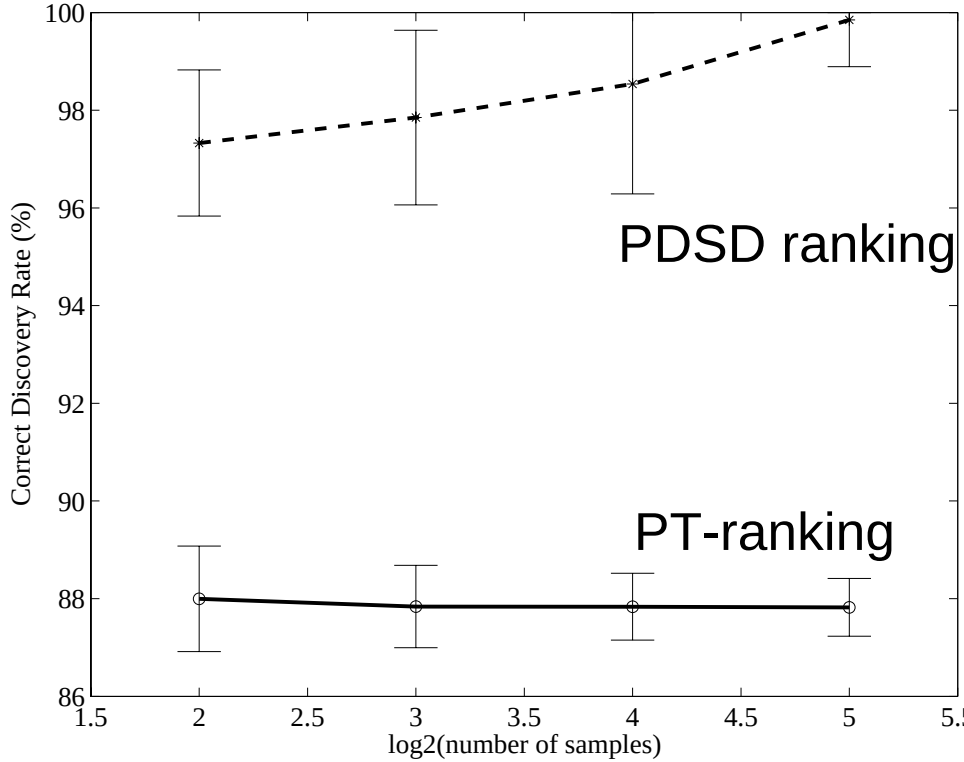


Ref: Fleury and Hero:JFI03

False Discovery Rate Comparisons



False Discovery Rate



Correct Discovery Rate

Ref: Fleury and Hero:JFI03

VI. Conclusions

- Biological and Statistical significance via FDRCI
- Provides P-values for screening and ranking
- Multiple criteria lead to Pareto Front Analysis
- PFA accounts for sampling errors via PDD's
- PFA has better FDR than screening-oriented methods for discovery of partial orderings
- Ongoing projects: Clustering, dimension reduction, learning the gene expression manifold

Dawning of Post-Genomic Era