Pareto posterior fronts for gene filtering

A. O. Hero

University of Michigan - Ann Arbor

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

Collaborators:	G. Fleury,	ESE - Paris
	S. Yoshida, A. Swaroop	UM - Ann Arbor
	T. Carter, C. Barlow	Salk - San Diego

Outline

1. Gene microarrays
2. Gene filtering problem
3. Posterior Pareto analysis
4. Application: development and aging in retina

Kellog Sensory Gene Microarray Node: Objectives

Establish genetic basis for development, aging, and disease in the retina

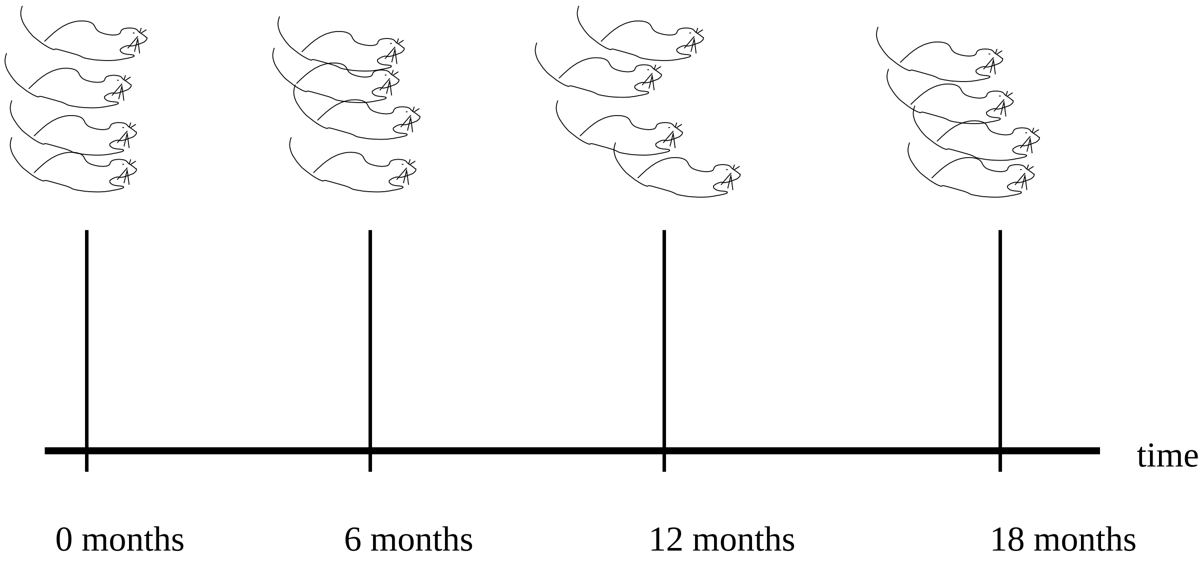


Figure 1: *Sample gene trajectories over time.*

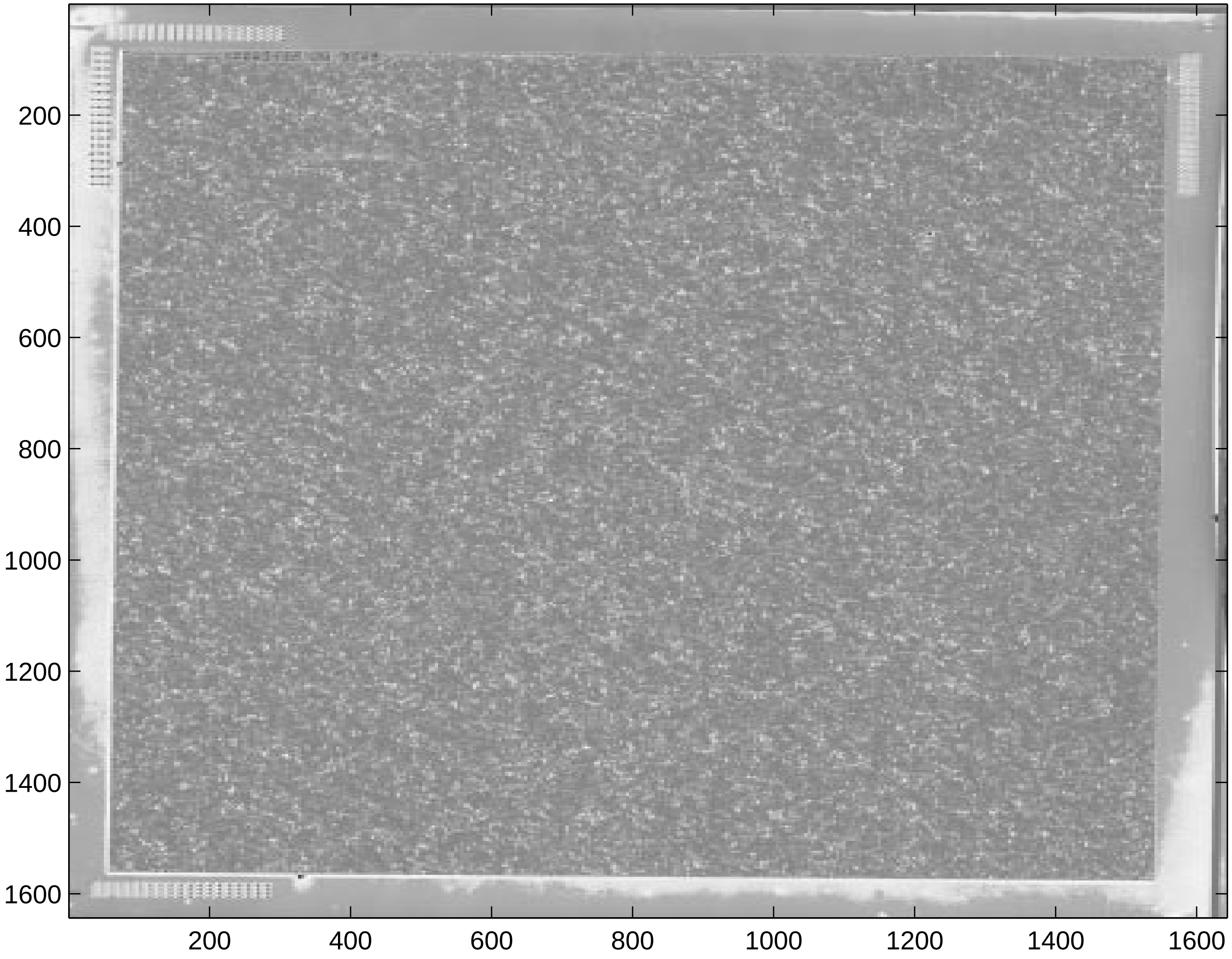


Figure 2: *Affymetrix GeneChip microarray.*

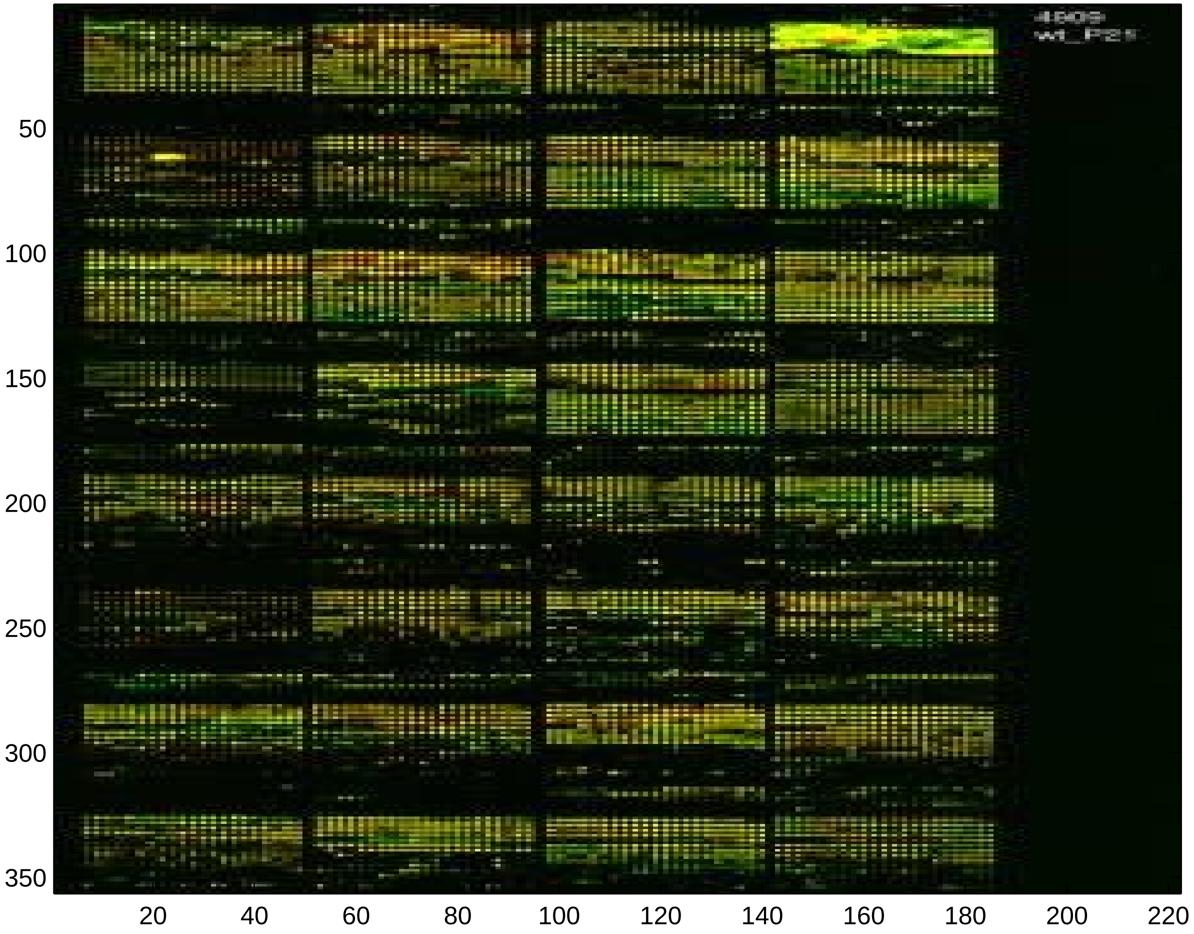


Figure 3: *cDNA spotted array.*

Pareto Gene Filtering (Fleury&etal:ICASSP02, Hero&etal:GENSIPS02)

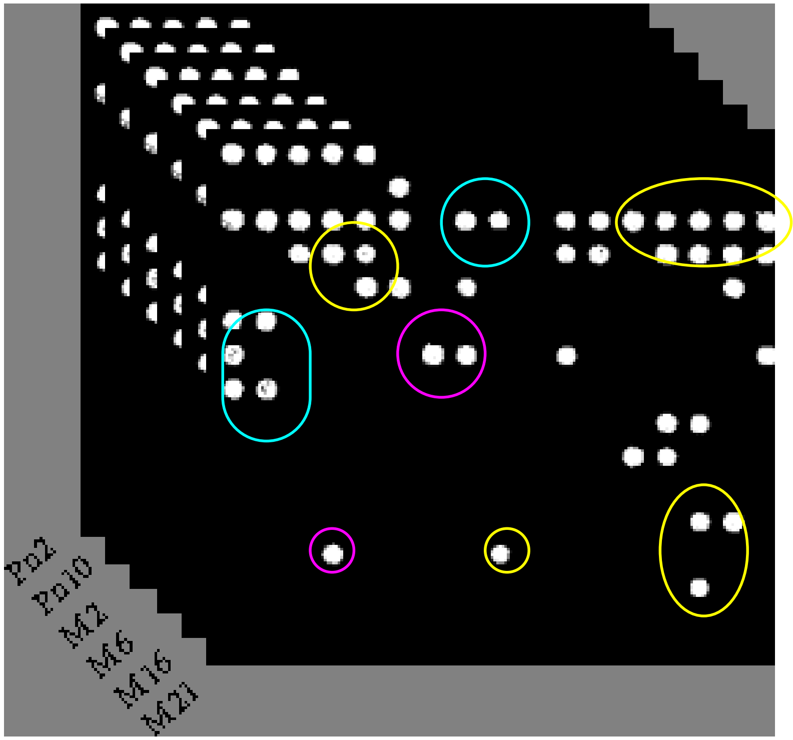


Figure 4: *Clustering on the Data Cube.*

Objective: Classify time trajectory of gene i into one of K classes

Gene Trajectory Classification

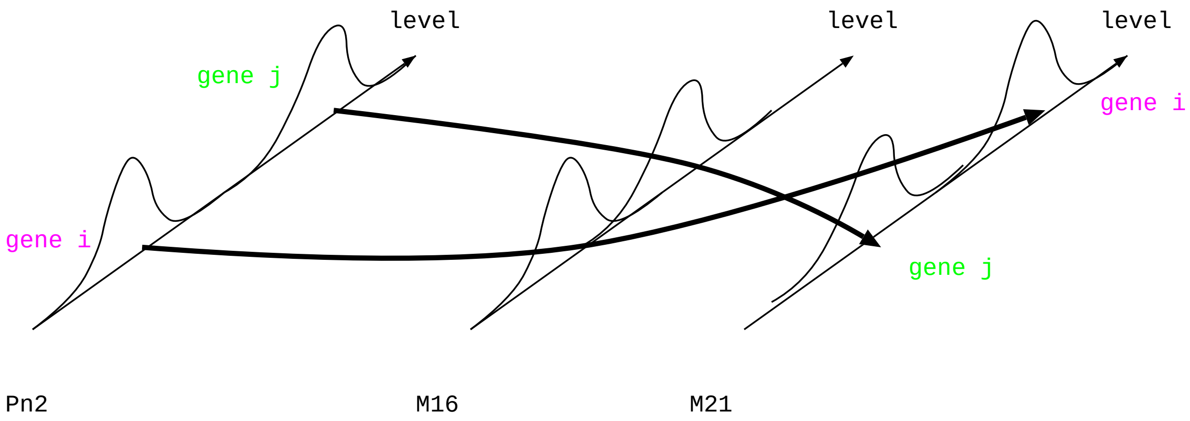


Figure 5: *Gene i is old dominant while gene j is young dominant*

Objective: extract gene trajectories (n) from sequence of repeated (m) microarray experiments over time samples (t)

$$y_{tm}(n), \quad n = 1, \ldots, N,\ t = 1, \ldots, T,\ m = 1, \ldots, M.$$

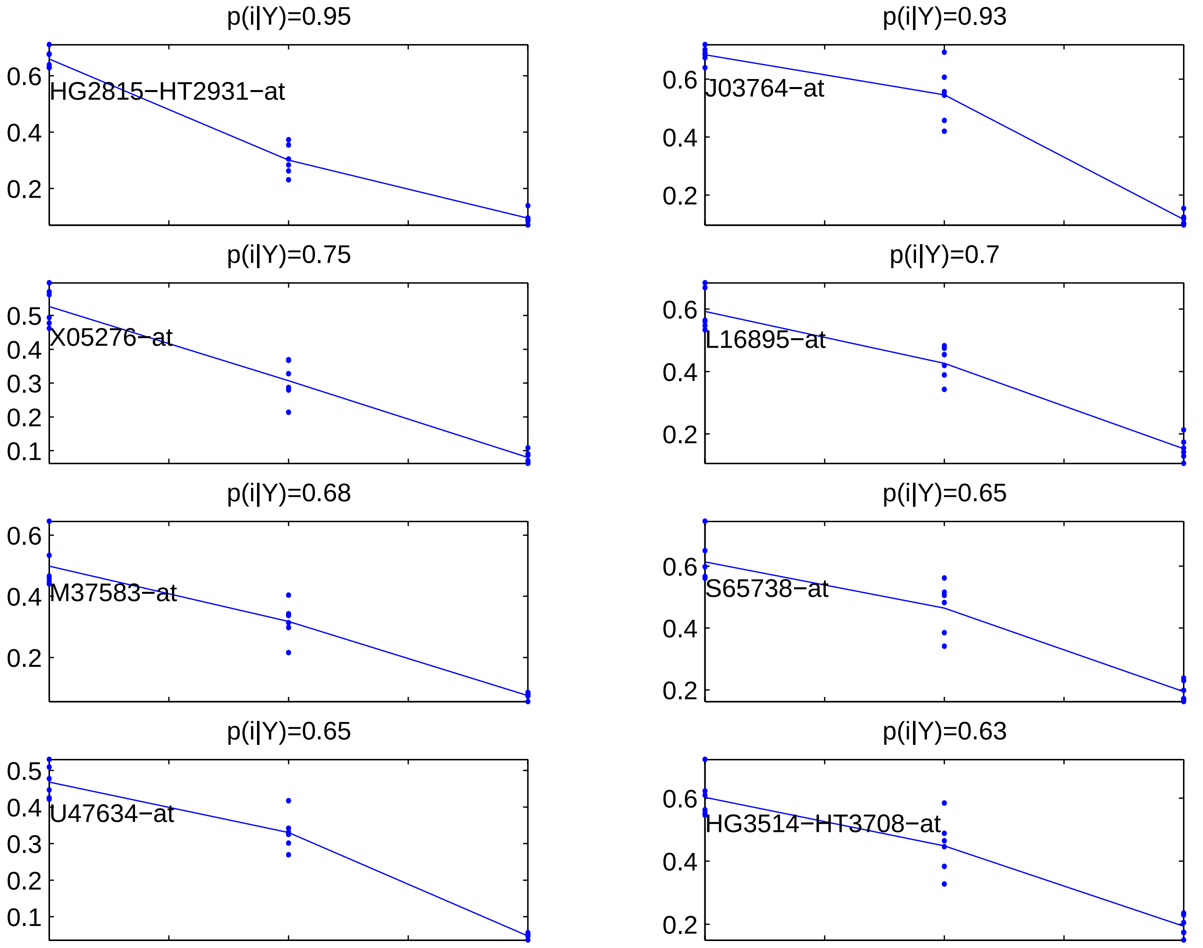


Figure 6: *8 ranked monotone decreasing gene profiles.*

Clustering and filtering Methods

Principal approaches:

- Hierarchical clustering (kdb trees, CART, gene shaving)
- K-means clustering
- Self organizing (Kohonen) maps
- Vector support machines

Validation approaches:

- Significance analysis of microarrays (SAM)
- Bootstrapping cluster analysis
- Leave-one-out cross-validation
- Replication (additional gene chip experiments, quantitative PCR)

Gene Filtering via Multiobjective Optimization

Gene selection criteria: for n-th gene $\xi_1(Y(n)),\ \ ,\ldots,\ \ \xi_P(Y(n))$

Possible $\xi_p(Y(n))$'s for finding uncommon genes

- Squared mean change from $t=1$ to $t=T$:

$$\xi_1(Y(n)) = |\bar{y}_{T*}(n) - \bar{y}_{1*}(n)|^2$$

- Standard deviation at $t=1$:

$$\xi_2(Y(n)) = \overline{(y_{1m}(n) - \bar{y}_{1*}(n))^2}$$

- Standard deviation at $t=T$:

$$\xi_3(Y(n)) = \overline{(y_{Tm}(n) - \bar{y}_{T*}(n))^2}$$

Some possible scalar functions:

- t-test statistic (Goss etal 2000): $T(n) = \frac{\xi_1(Y(n))}{\frac{1}{2}\xi_2(Y(n))+\frac{1}{2}\xi_3(Y(n))}$
- R^2 statistic (Hastie etal 2000): $R^2(n) = \frac{T_n}{1+T_n}$
- H statistic (Sinha etal 1998): $H(n) = \frac{\xi_1(Y(n))}{\sqrt{\xi_2(Y(n))\xi_3(Y(n))}}$

Objective: find genes which maximize or minimize the selection criteria

Aggregated Criteria

Let $\{W_p\}_{p=1}^P$ be experimenter's cost "preference pattern"

$$\sum_{p=1}^{P} W_p = 1,\ W_i \geq 0$$

Find optimal gene via:

$$\max_n \sum_{p=1}^{P} W_p \xi_p(Y(n)), \quad or \quad \max_n \prod_{p=1}^{P} (\xi_p(Y(n)))^{W_p}$$

Q. What are the set of optimal genes for all preference patterns?

A. These are *non-dominated* genes (Pareto optimal)

Defn: Gene *i* is dominated if there is a $j \neq i$ s.t.

$$\xi_p(Y(i)) \leq \xi_p(Y(j)),\ p = 1,\ldots,P$$

Pareto Optimal Fronts

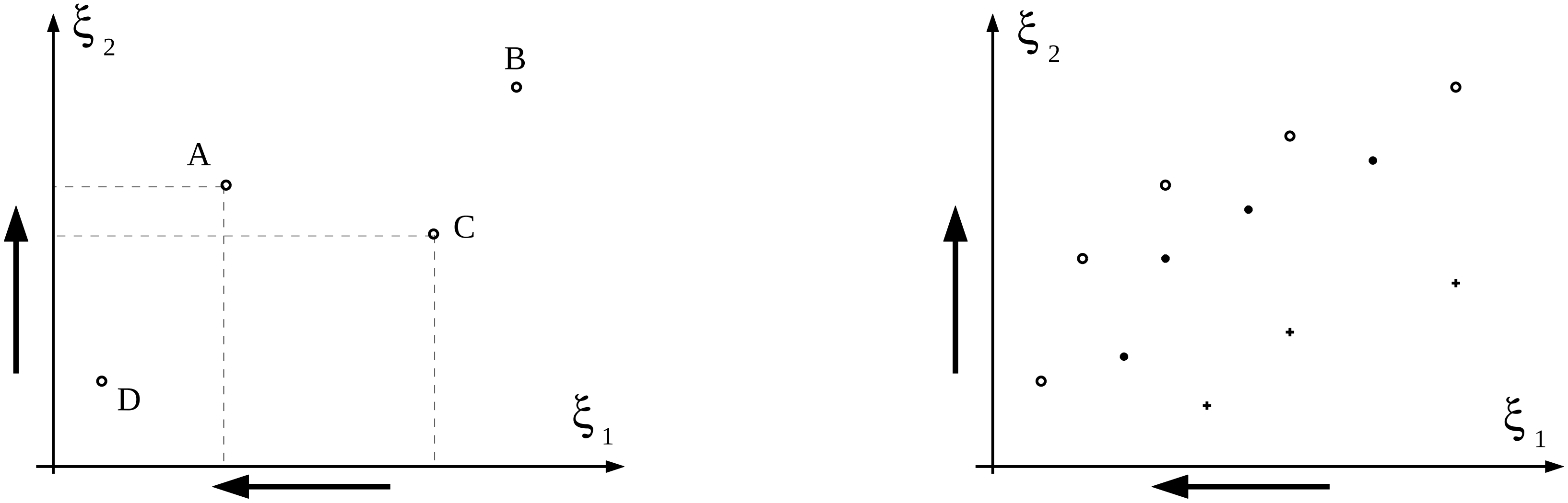


Figure 7: *a). Non-dominated property, and b). Pareto optimal fronts, in dual criteria plane.*

Pareto Gene Filtering vs. Paired T-test

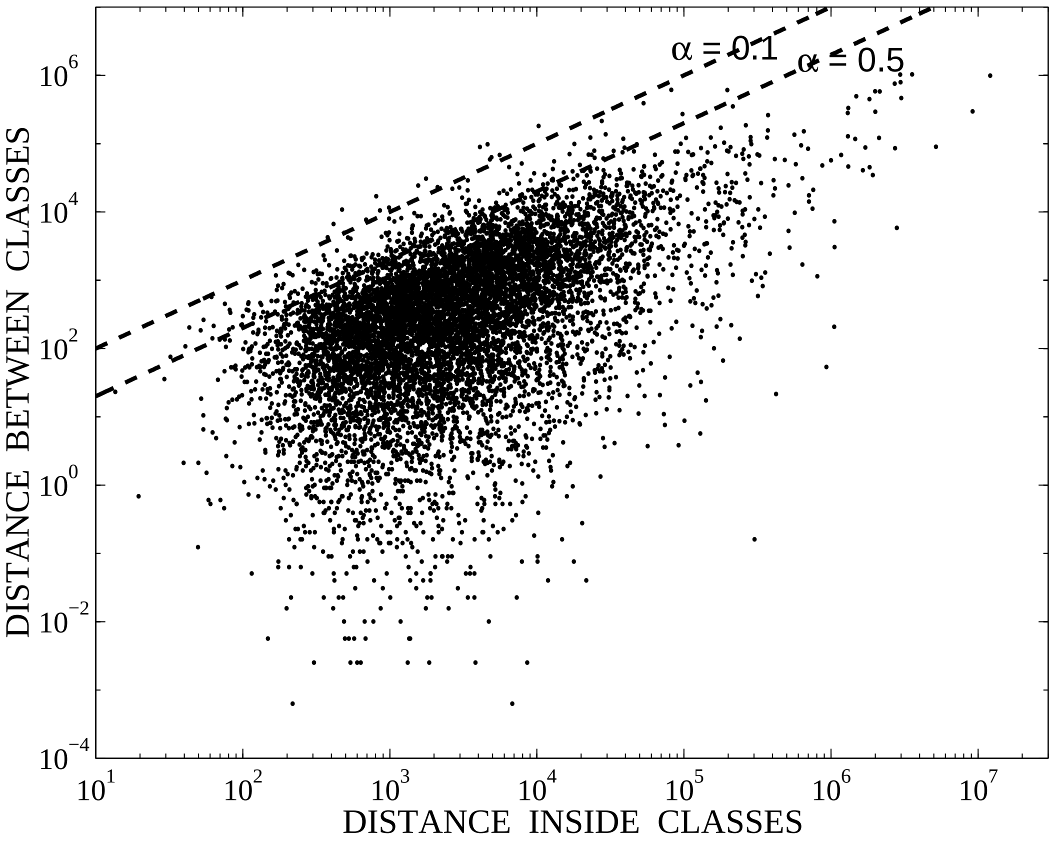


Figure 8: $\xi_1 =$ *mean change vs* $\xi_2 =$ *pooled standard deviation for 8826 mouse retina genes. Superimposed are T-test boundaries*

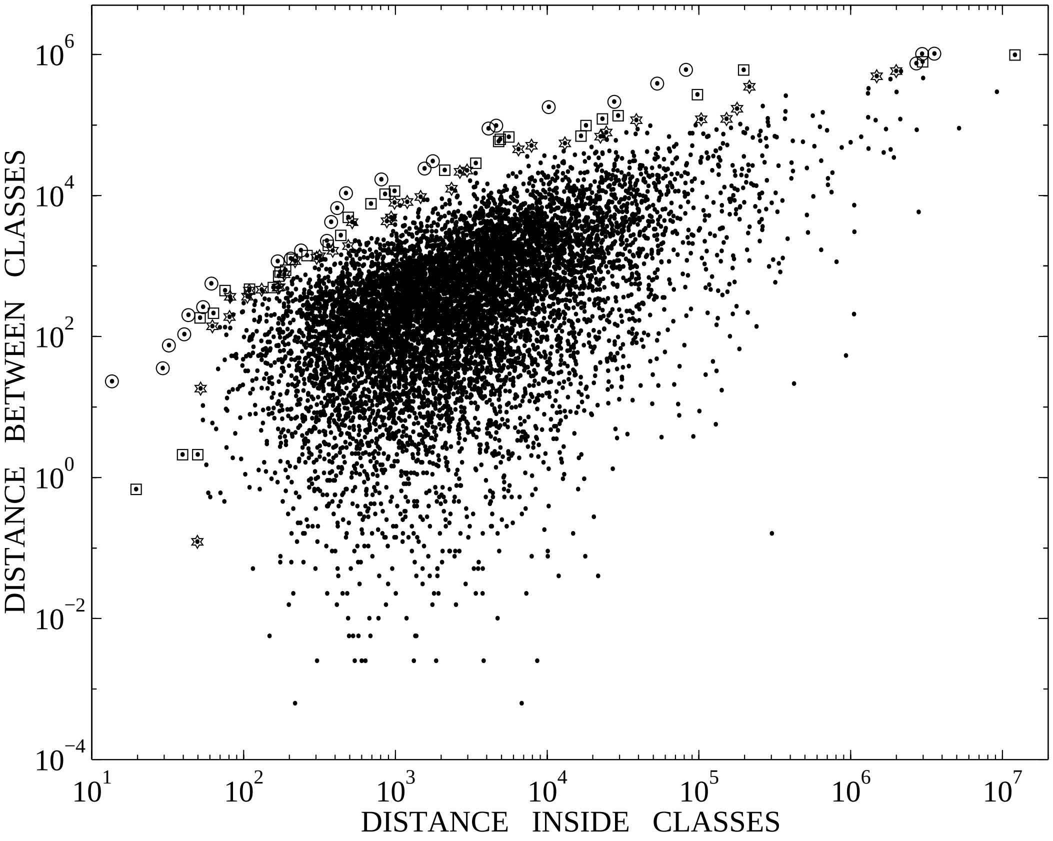


Figure 9: *First (circle) second (square) and third (hexagon) Pareto optimal fronts.*

Application: Development and Aging in Mouse Retina

Mouse Retina Experiment:

- Retinas of 24 mice sampled and hybridized
- 6 time points: Pn2, Pn10, M2, M6, M16, M21
- 4 mice per time sample
- Affymetrix GeneChip layout with 12422 poly-nucleotides
- Affymetrix attribute analyzed: “AvgDiff”
- Used Affymetrix filter to eliminate all genes labeled “A”

Objective: Find interesting gene trajectories within the set of remaining 8826 genes

Multi-objective Non-parametric Pareto Filtering

Define *trend vector*: $\psi(n) = [b_1, \ldots, b_6]$, $b_i \in \{0,1\}$

- Old dominant filtering criteria:
  - Maximum end-to-end increase ($T = 6$)

$$\xi_1(Y(n)) = \bar{y}_{T*}(n) - \bar{y}_{1*}(n) = \mathrm{max}$$

  - high consistency over $6^4 = 4096$ possible combinations of trajectories

$$\xi_2(Y(n)) = \frac{\text{\# trajectories having } \psi(n) = [1, \ldots, 1]}{4096}$$

Occurence Histogram

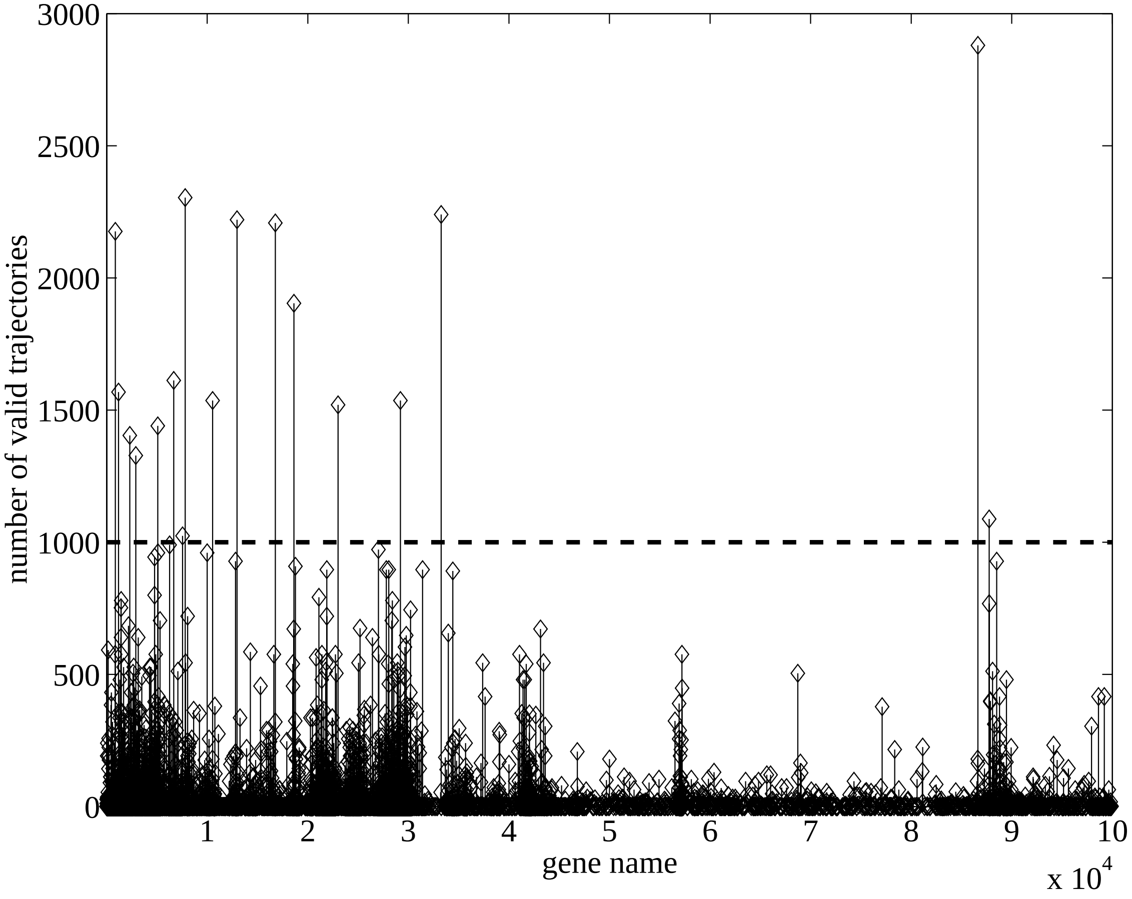
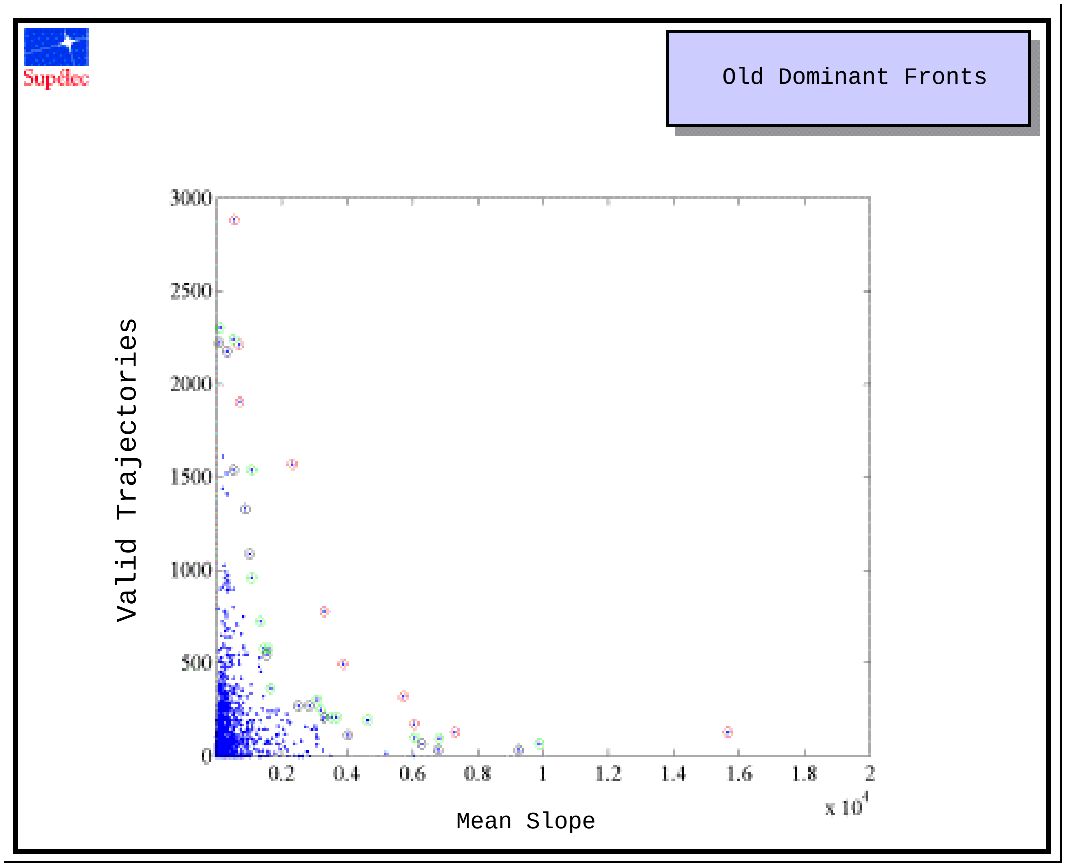


Figure 10: *Monotonicity occurrence histogram with threshold.*

Old Dominant Pareto Fronts



Old Dominant Fronts

Figure 11: *Pareto fronts for old dominant genes.*

Three-objective Pareto Filtering

Objective Extract “aging genes”

- Strictly increasing filtering criteria:
  - Maximum end-to-end increase

$$\xi_1(Y(n)) = \bar{y}_{T*}(n) - \bar{y}_{1*}(n) = \max$$

  - High consistency over $6^4 = 4096$ possible combinations of trajectories

$$\xi_2(Y(n)) = \frac{\#\text{ trajectories having } \psi_i = [1,\ldots,1]}{4096} = \max$$

  - no plateau

$$\xi_3(Y(n)) = [\bar{y}_{t+1,*}(n) - 2\bar{y}_{t*}(n) + \bar{y}_{t-1,*}(n)]^2 = \min$$

Pairwise Pareto Fronts

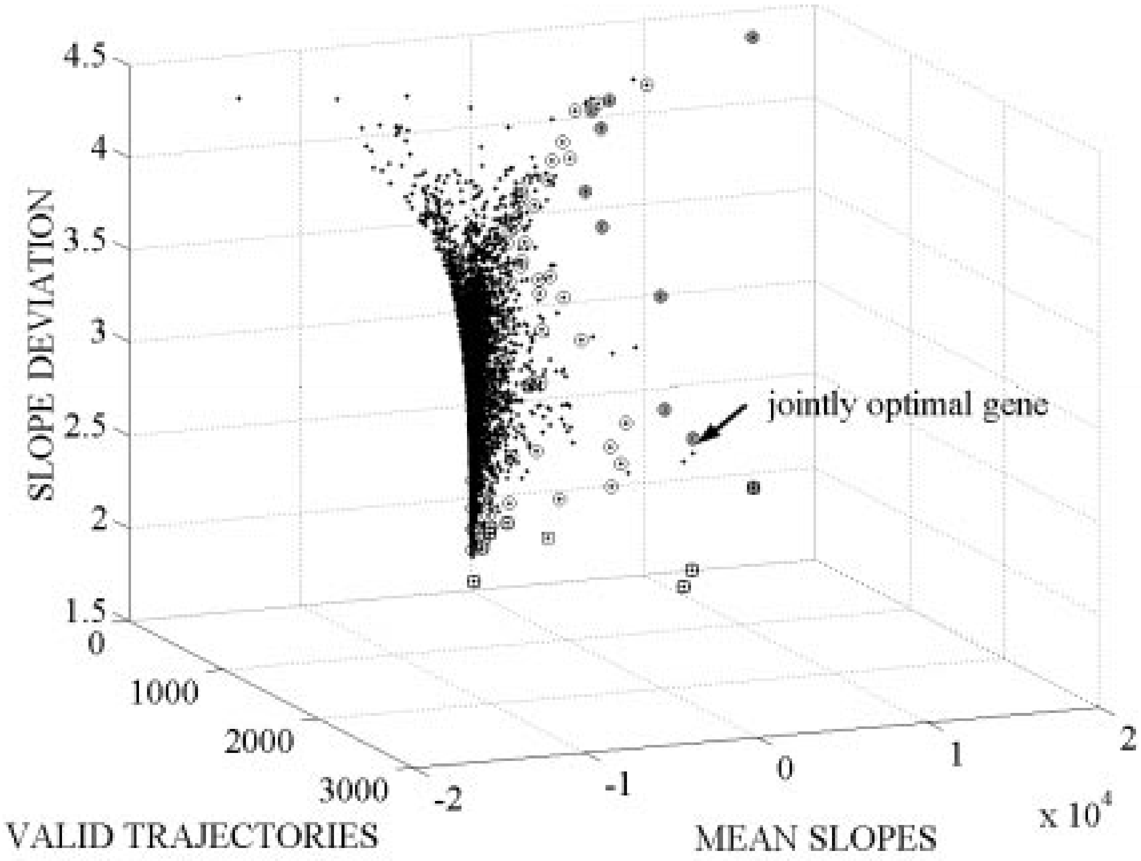


Figure 12: *First Pareto fronts for each pair of criteria taken from the set (ξ_1, ξ_2 and ξ_3). Each one of this front is denoted by squares, circles and stars, respectively.*

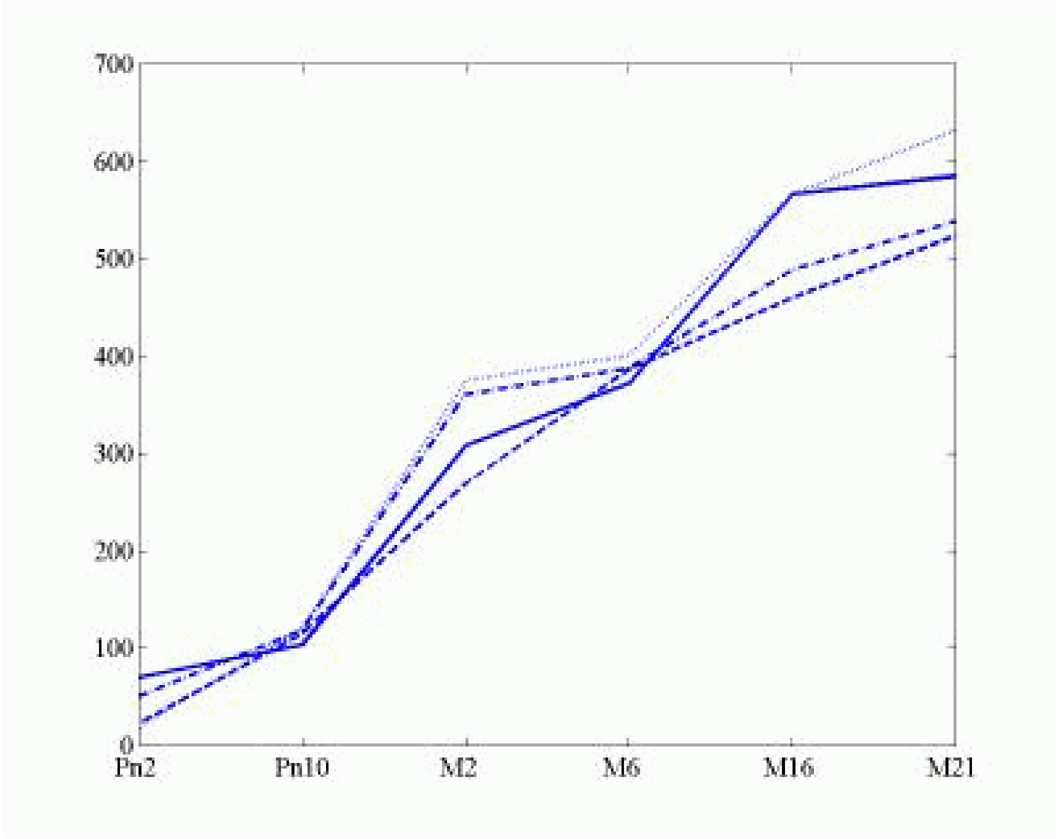


Figure 13: *Top ranked gene profile is Mus musculus* $5^{'}$ *end cDNA (Unigene 86632)*

Confidence measures?

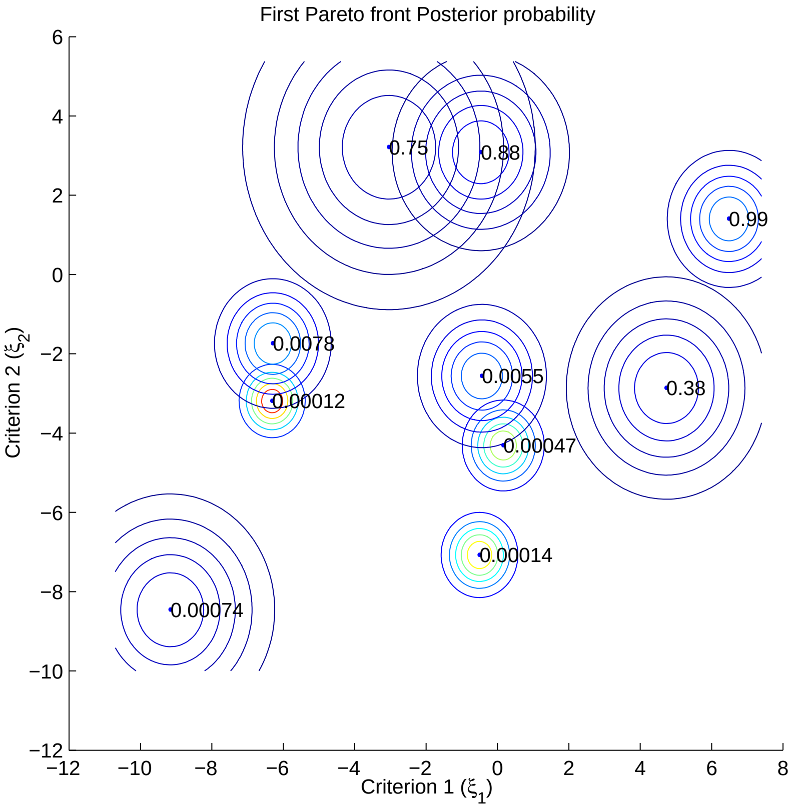


Figure 14: *PPF analysis over dual criteria to be maximized.*

Cross-validation approach

- Leave-one-out cross validation

Let $Y^{-m}(n)$ denote one possible set of $T \times (M - 1) = 6 \times 3$ samples

Cross-validation Algorithm:

```
Do $m = 1, \ldots, 4^6$:

            Compute  $\left(\xi_1(Y^{-m}(n)), \xi_2(Y^{-m}(n))\right)$

   Find Genes in First 3 Pareto fronts: $G^{-m}$
   End
   $\textit{Resistant}$ Genes = $\cap_{m=1}^{4^6} G^{-m}$
```

Posterior Pareto Front (PPF) approach

Given prior on mean expression levels $\overline{\xi}_p(n) = E[\xi_p(Y(n))]$ find

$$\begin{aligned} p(i|Y) &\overset{\text{def}}{=} P(\text{gene } i \text{ on Pareto front}|Y) \\ &= P(\overline{\xi}_1(i) \geq \max_j \overline{\xi}_1(j) \text{ or } \ldots \text{ or } \overline{\xi}_P(i) \geq \max_j \overline{\xi}_P(j)|Y) \\ &= \sum_{k=1}^{P} P(E_k(i)|Y) - \sum_{k_1<k_2} P(E_{k_1}(i), E_{k_2}(i)|Y) + \ldots \\ &\quad +(-1)^{p+1} \sum_{k_1<\ldots<k_p} P(E_{k_1}(i), \ldots, E_{k_p}(i)|Y) \\ &\quad +(-1)^{P+1} P(E_{k_1}(i), \ldots, E_{k_P}(i)|Y) \end{aligned}$$

E_i denotes the event $\xi_1(\mu(i)) \geq \max_j \xi_1(\mu(j))$

Gaussian observations with noninformative prior

1. Assume conditionally linear Gaussian model $\varepsilon_{tm}(n) \sim N\left(0, \sigma_t^2(n)\right)$

$$y_{tm}(n) = \mu_t(n) + \varepsilon_{tm}(n)$$

2. Assume non-informative prior

$$f_{\mu_t(n), \sigma_t^2(n)}(u, s) \; = \; \frac{c}{s^{a/2}}, \;\; u \in \mathbf{R}, \; s \in \mathbf{R}^+$$

3. Adopt *Profile contrasts* as selection criteria:

$$\begin{bmatrix} \overline{\xi}_1(n) \\ \vdots \\ \overline{\xi}_P(n) \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1T} \\ \vdots & \ddots & \vdots \\ a_{P1} & \cdots & a_{PT} \end{bmatrix} \begin{bmatrix} \mu_1(n) \\ \vdots \\ \mu_T(n) \end{bmatrix}$$

Example contrasts

$$
\begin{aligned}
A_2 &= \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}, \\
A_2' &= \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \\
A_3 &= \begin{bmatrix} -1 & 0 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \\
A_3' &= \begin{bmatrix} -1 & 1 & 0 \\ -1 & -1 & 2 \\ 1 & 1 & 1 \end{bmatrix},
\end{aligned}
$$

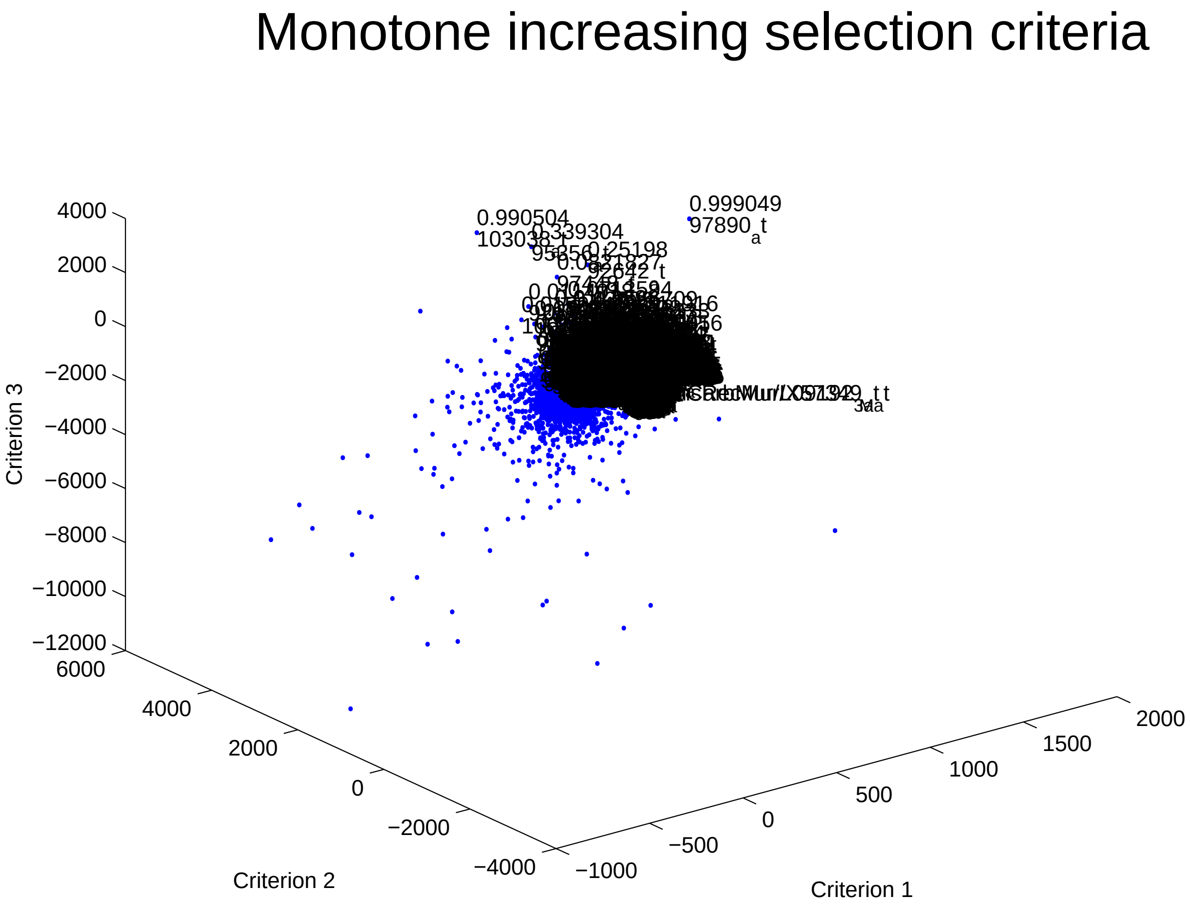


Figure 15: *Avgdiff indices for Affy mouse study. Scatterplot of slope contrasts with posterior Pareto probabilities over positive orthant.*

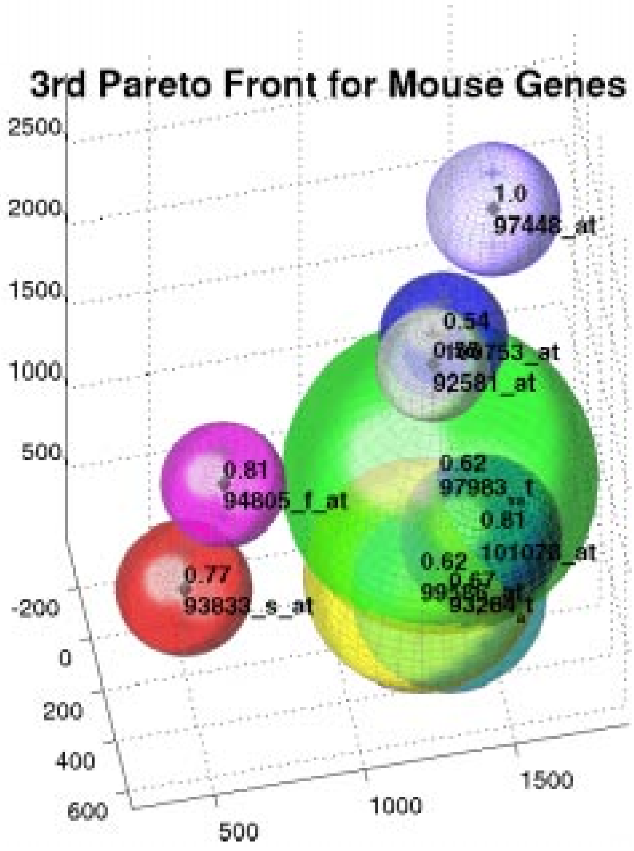


Figure 16: *Third posterior Pareto front for (affy mouse study).*

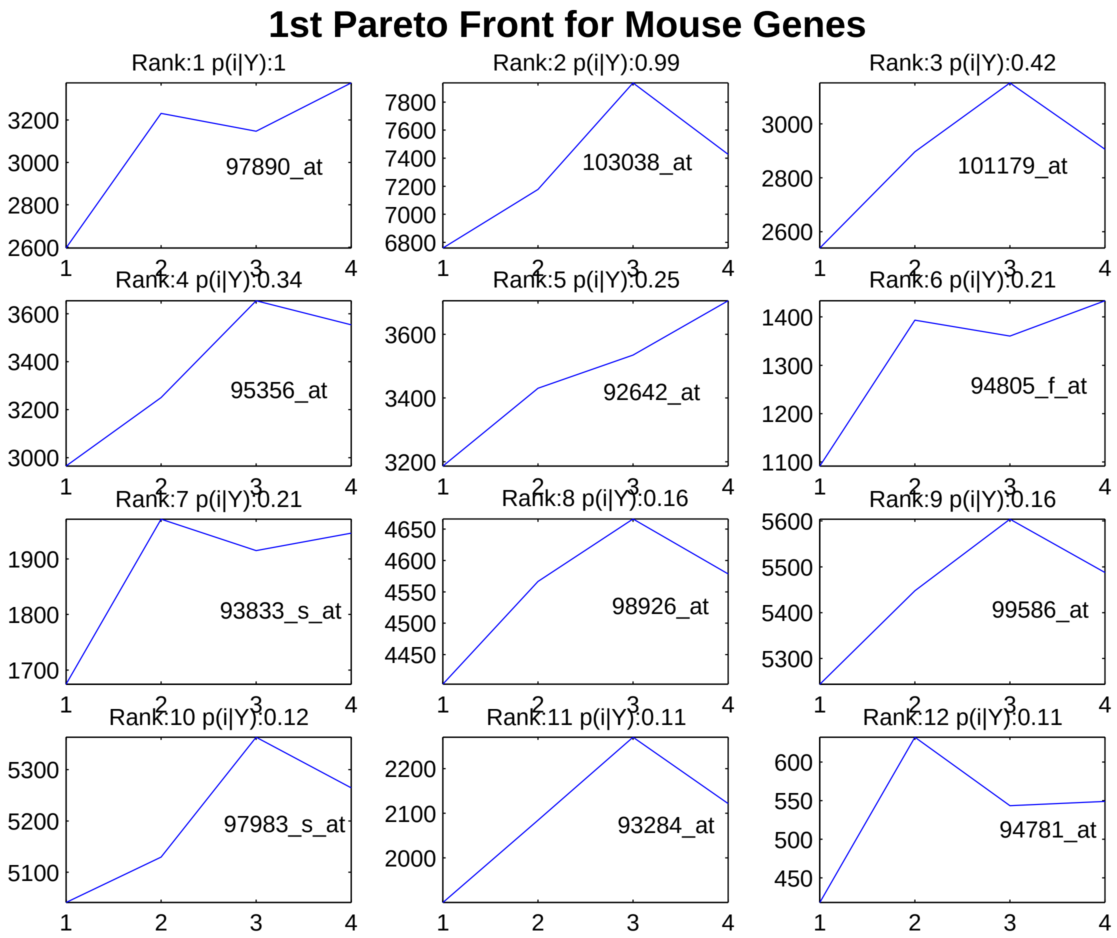


Figure 17: *Ranked first posterior Pareto front gene trajectories (Affy mouse study).*

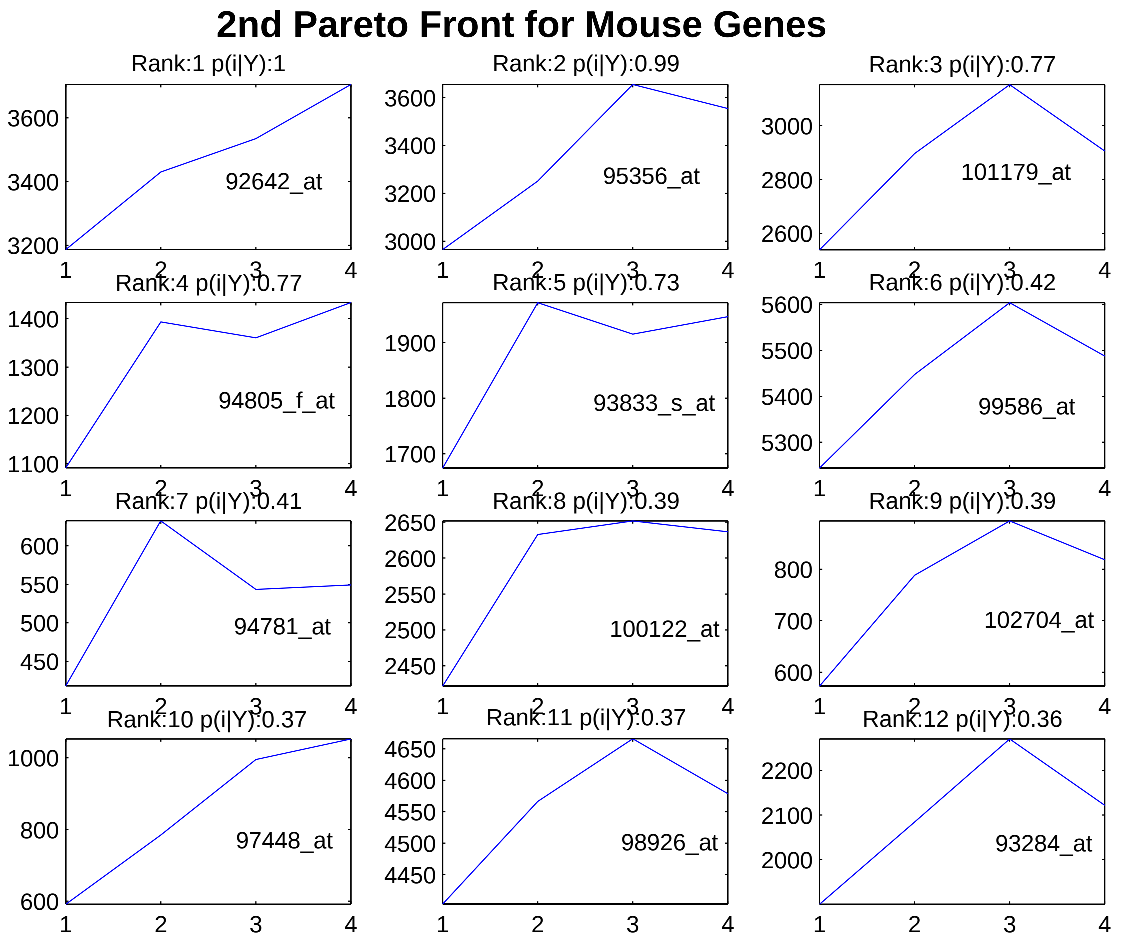


Figure 18: *Ranked second posterior Pareto front gene trajectories (Affy mouse study).*

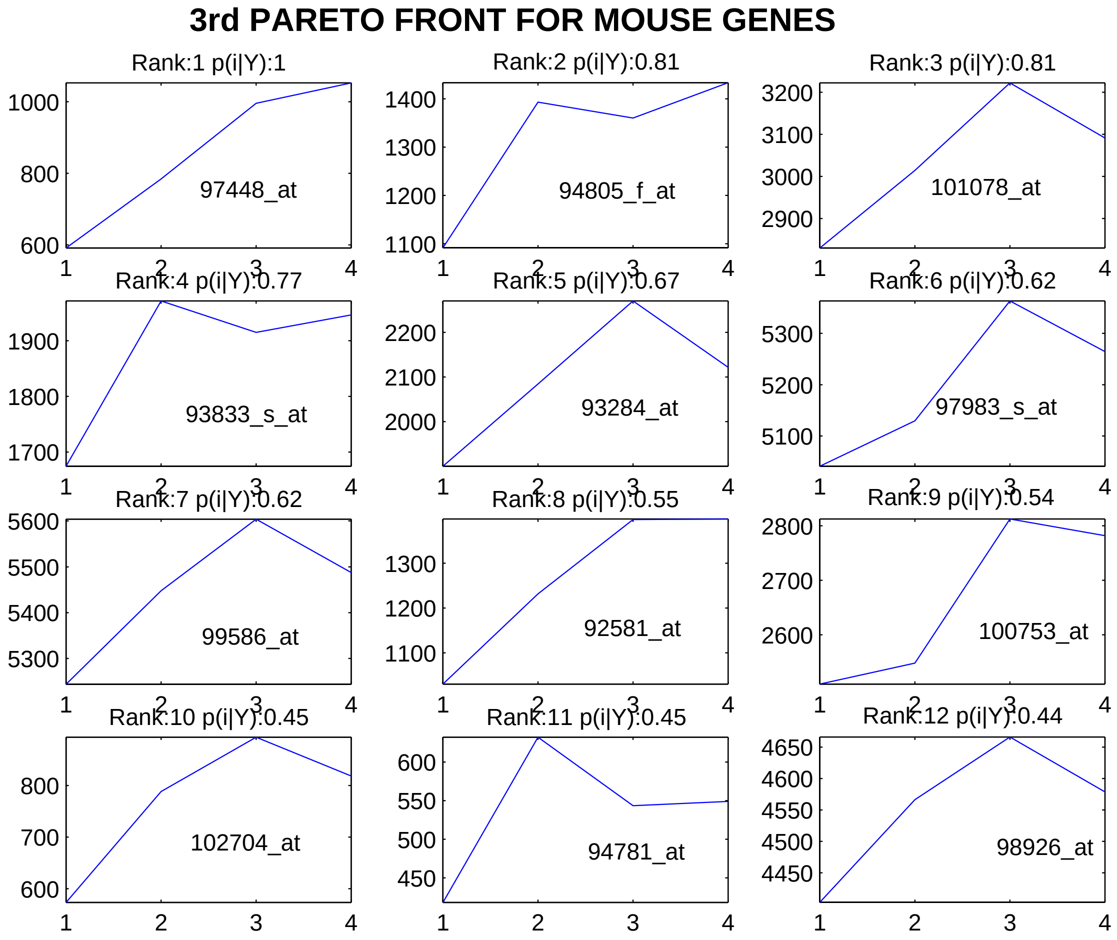


Figure 19: *Ranked third posterior Pareto front gene trajectories (Affy mouse study).*

Application: Fred Wright's data

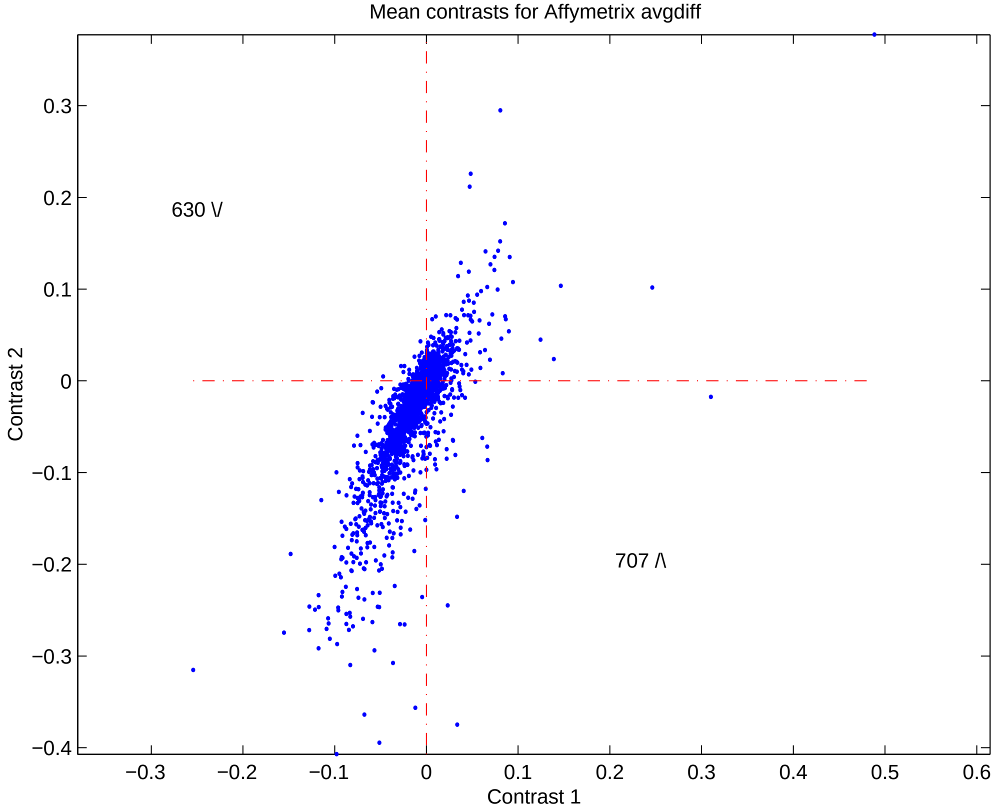


Figure 20: *Scatterplot of slope contrasts (Sample mean contrasts defined from the first two rows of A_3') for avgdiff indices for Fred Wright's HuGeneFL mixture study. Annotations are the number of non-monotone genes with convex cup (upper left) and convex cap (lower right) profiles.*

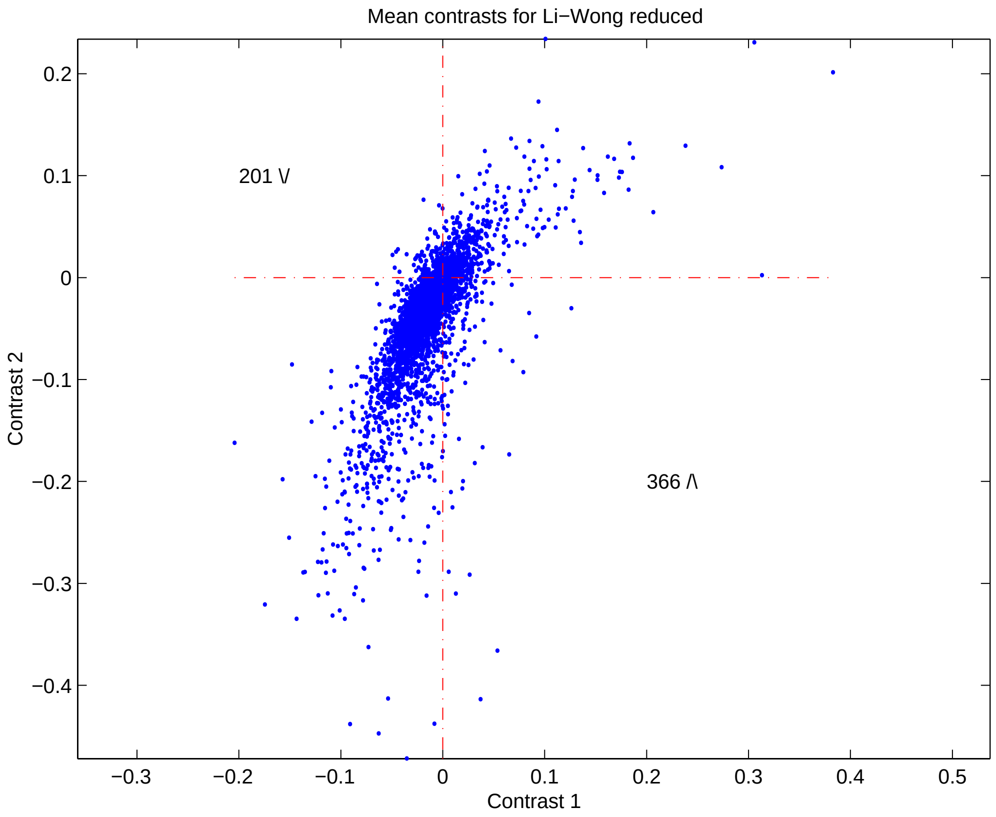


Figure 21: *Scatterplot of slope contrasts (Sample mean contrasts defined from the first two rows of* A_3'*) for Li-Wong reduced indices for Fred Wright's HuGeneFL mixture study. Annotations are the number of non-monotone genes with convex cup (upper left) and convex cap (lower right) profiles.*

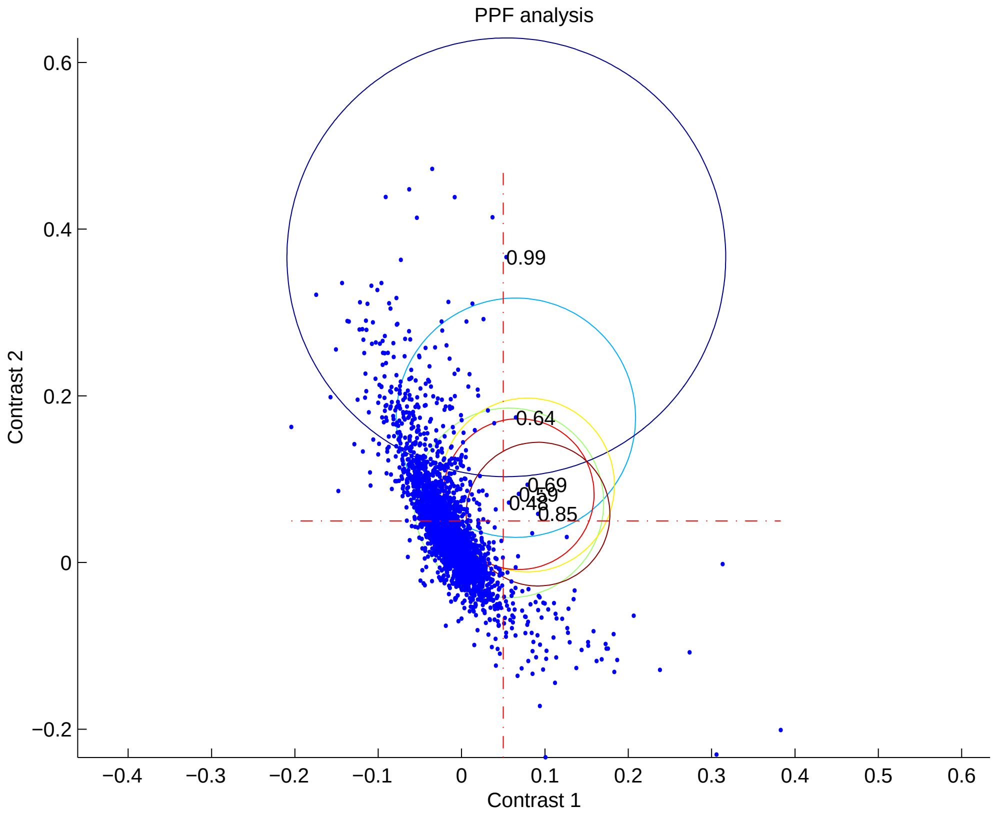


Figure 22: *The 6 top scoring genes resulting from PPF analysis of the most non-monotone convex cap profiles for Fred Wright's data using Li-Wong reduced indices.*

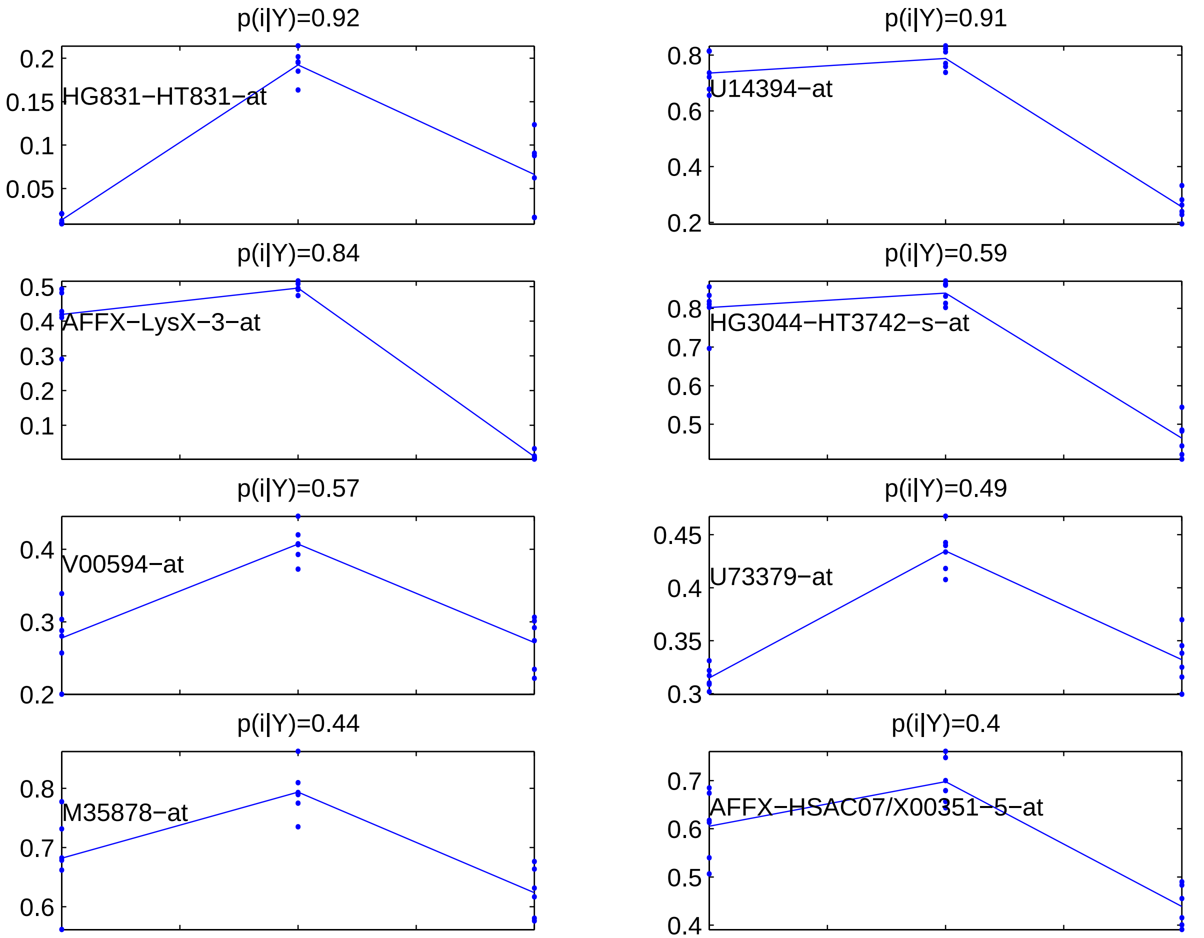


Figure 23: *First 8 rank ordered convex cup genes profiles from Li-Wong indices.*

Conclusions

1. New methods of data mining are needed to perform robust and flexible gene filtering
2. Cross-validation can account for statistical sampling uncertainty
3. Non-informative priors can be used to find posterior front probability
4. Genetic priors: phylogenetic trees, BLAST database, etc?