# Bioinformatics and Genomics: A New SP Frontier?

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. Bioinformatics background
2. Gene microarrays
3. Gene clustering and filtering for gene pattern extraction
4. Application: development and aging in retina

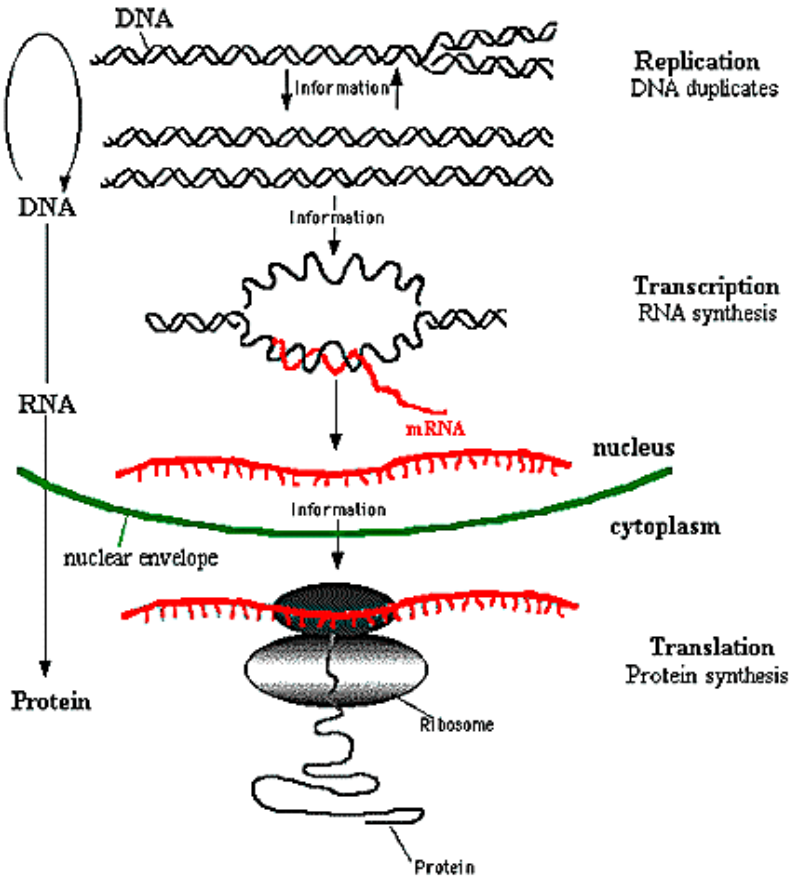


Figure 1: `http://www.accessexcellence.org`

## I. Bioinformatics background

- Every human cell contains 6 feet of double stranded (ds) DNA
- This DNA has 3,000,000,000 basepairs representing 50,000-100,000 genes
- This DNA contains our complete genetic code or *genome*
- DNA regulates all cell functions including response to disease, aging and development
- Gene expression pattern: snapshot of DNA in a cell
- Gene expression profile: DNA mutation or polymorphism over time
- Genetic pathways: changes in genetic code accompanying metabolic and functional changes, e.g. disease or aging.

**Genomics:** study of gene expression patterns in a cell or organism

## Possible Impact

- Understanding role of genetics in cell function and metabolism
- Discovering genetic markers and pathways for different diseases
- Understanding pathogen mechanisms and toxicology studies
- Development of genotype-specific drugs
- Development of genetic computing machines
- In situ genetic monitoring and drug delivery

## Kellog Sensory Gene Microarray Node: Objectives

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

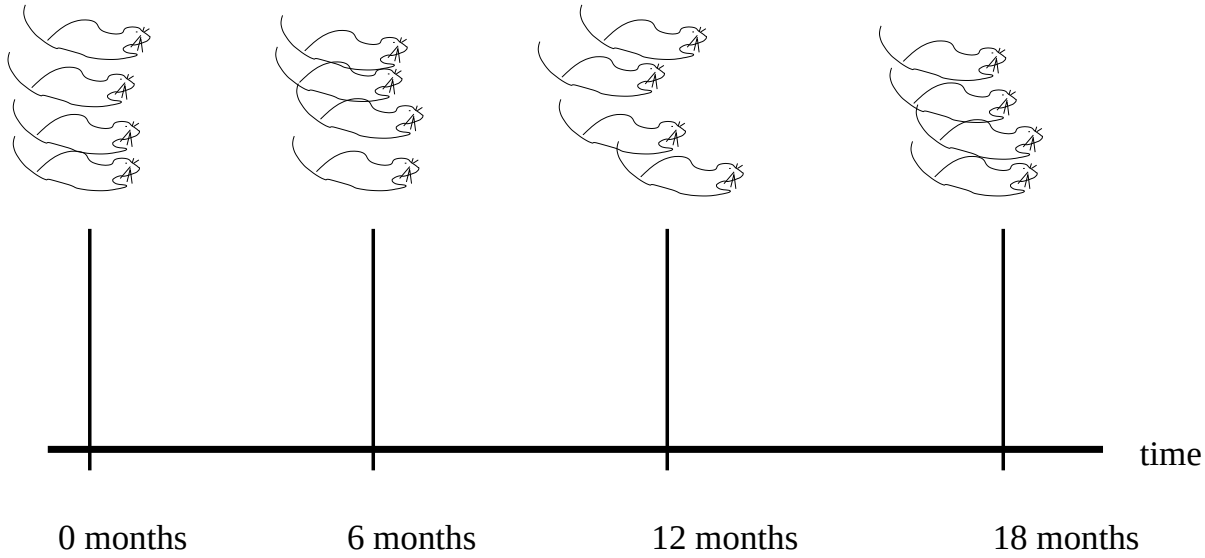


Figure 2: *Sample gene trajectories over time.*

## II. Gene Microarrays

Two kinds of “Shotgun sequencing:”

1. GeneChip Oligonucleotide Microarrays (Affymetrix)
2. cDNA Microarrays (Stanford)

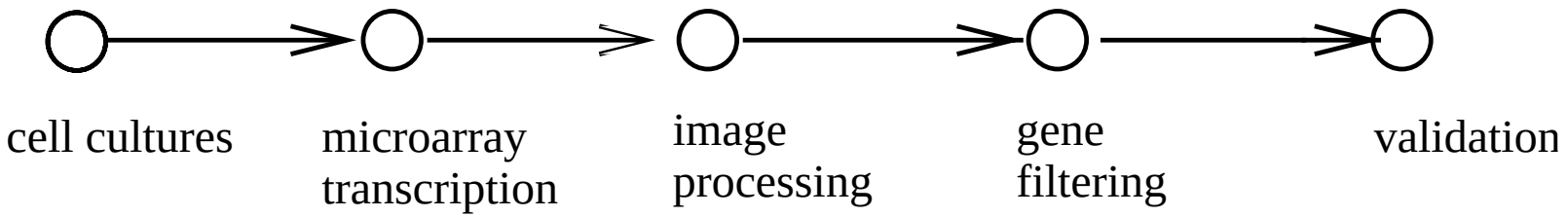


Figure 3: *Microarray experiment cycle.*

## Microarray Image Formation

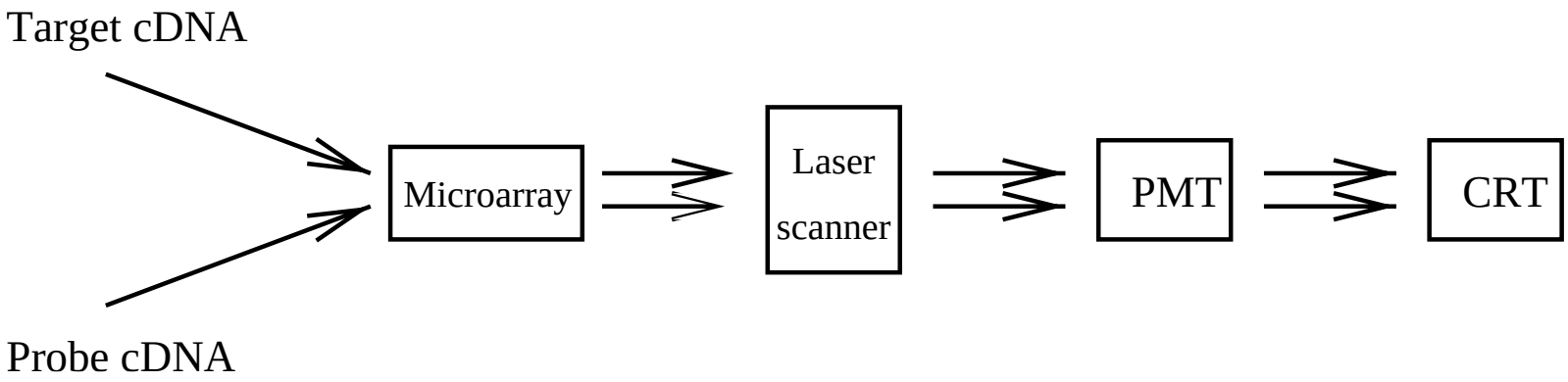


Figure 4: *Image formation process.*

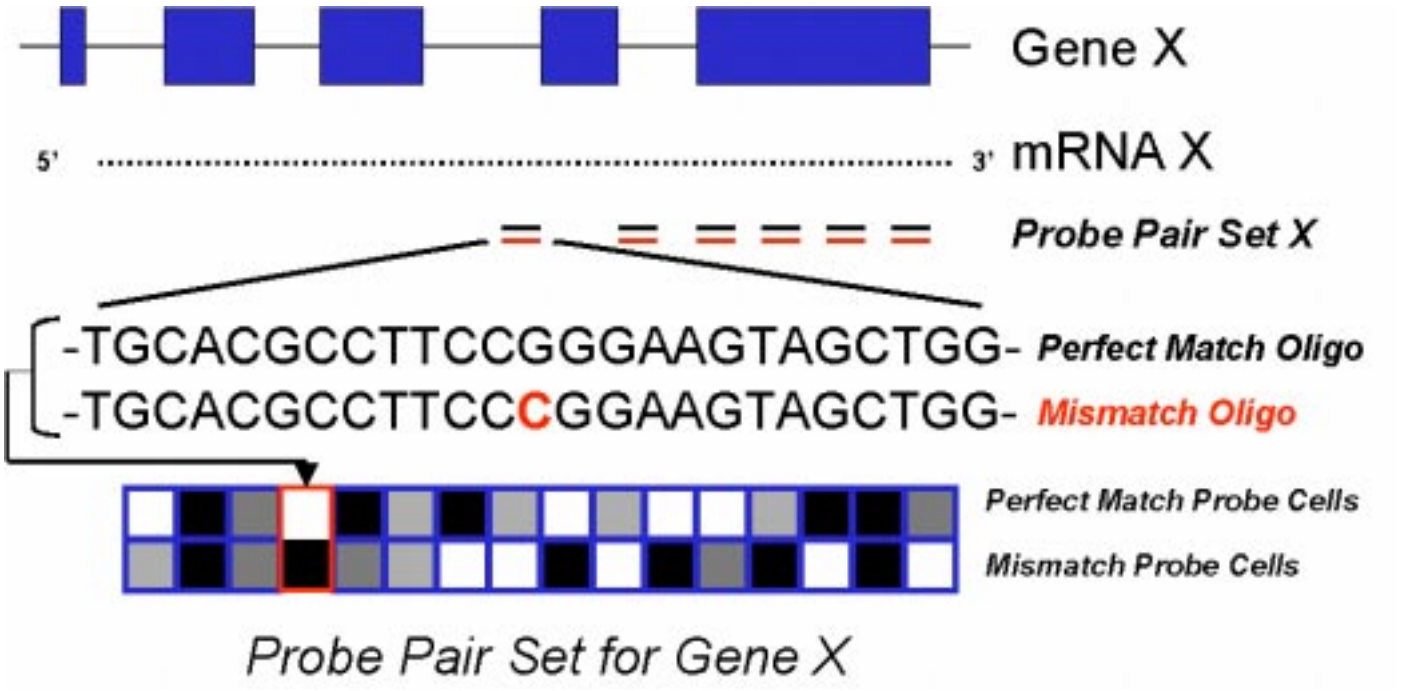


Figure 5: *Oligonucleotide PM/MM layout (*`pathbox.wustl.edu`*).*

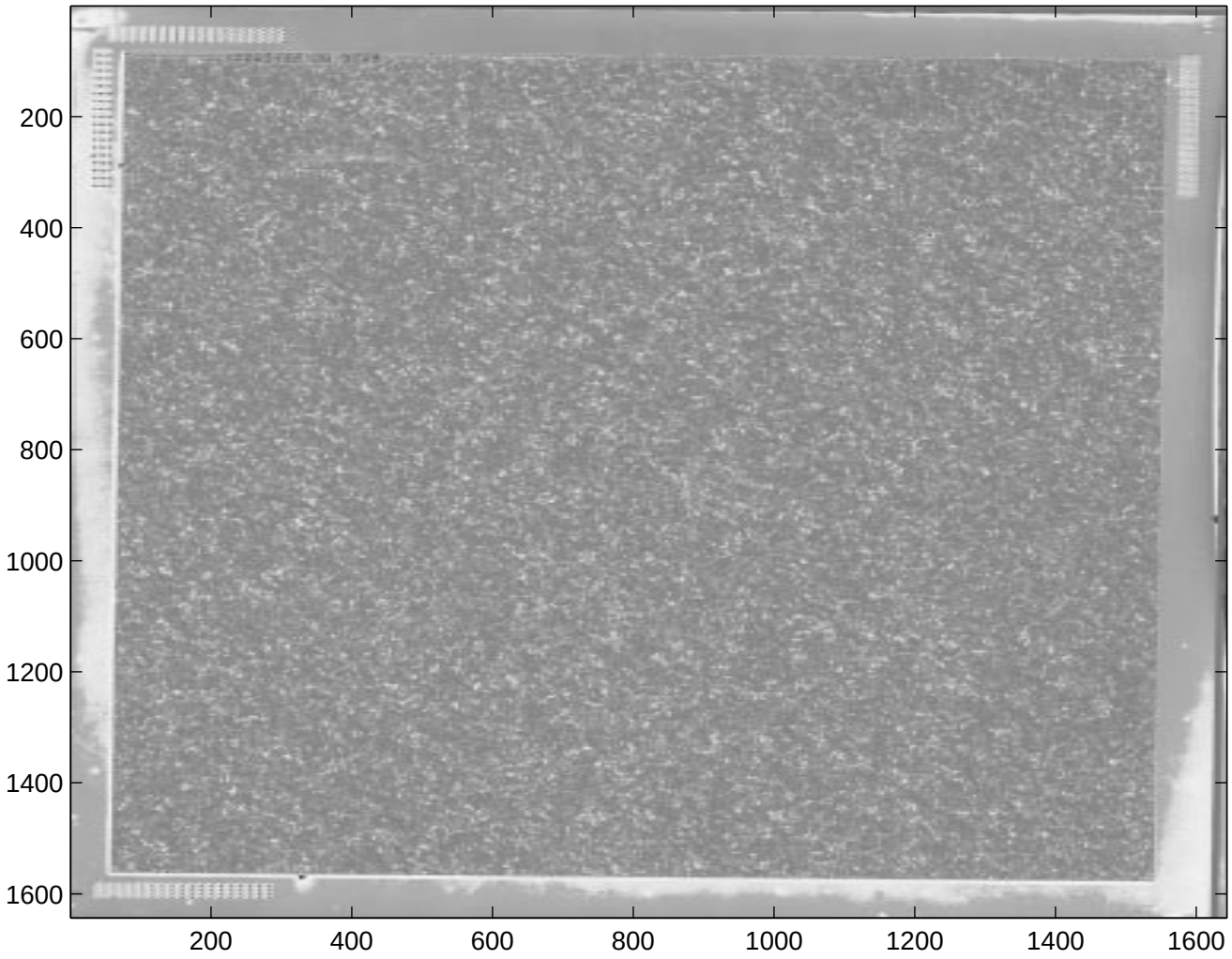


Figure 6: *Affymetrix GeneChip microarray.*

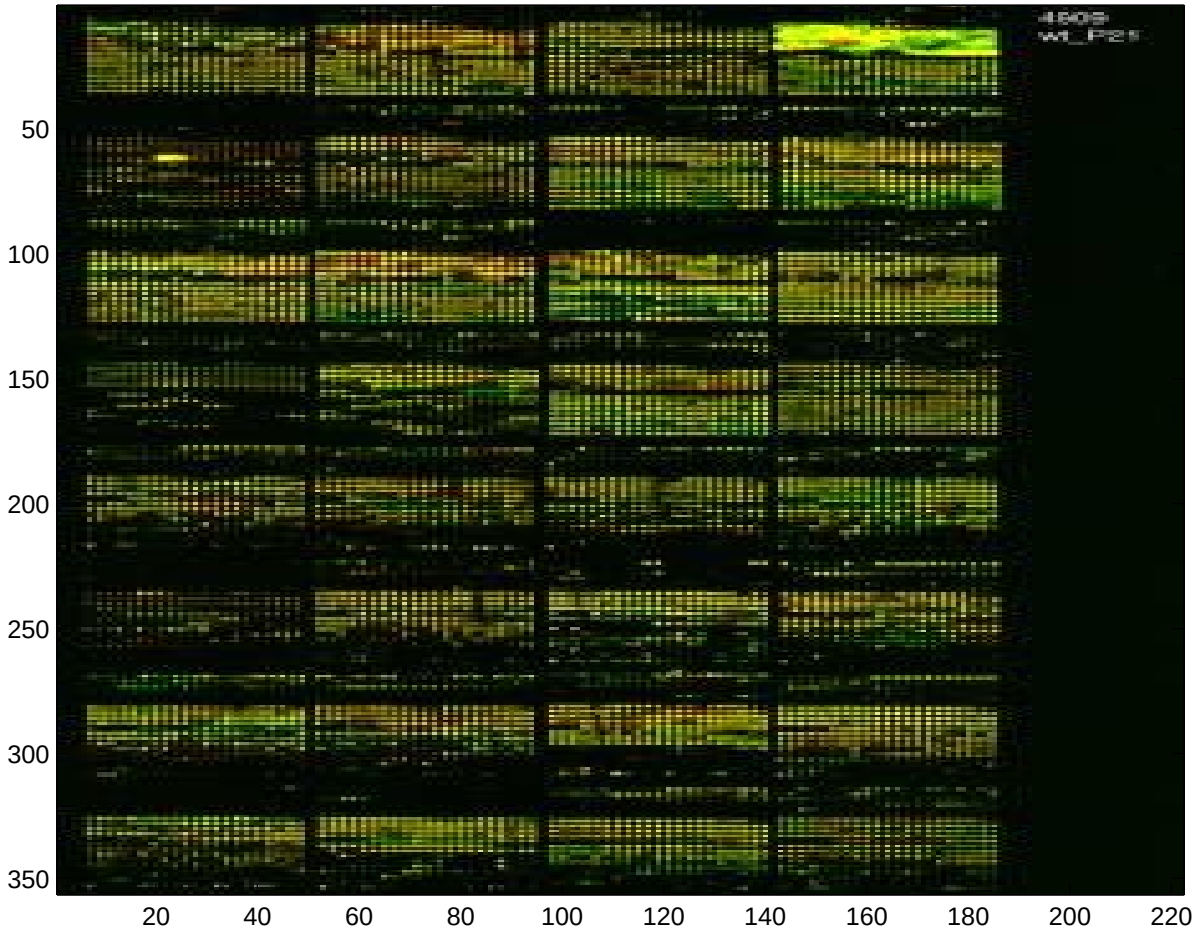


Figure 7: *cDNA spotted array.*

## Control Factors Influencing Variability

- **Sample preparation**: reagent quality, temperature variations
- **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 aberations, gain settings
- **Imaging and Extraction**: spot misalignment, discretization, clutter

$\rightarrow$ account for data variability

- **Scaling factors**: universal intensity amplification factor for a chip
- **Raw Q**: noise and other random variations of a chip
- **Background**: avg of lowest 2% cell intensity values
- **% P**: percentage of transcripts present

# Microarray Signal Extraction

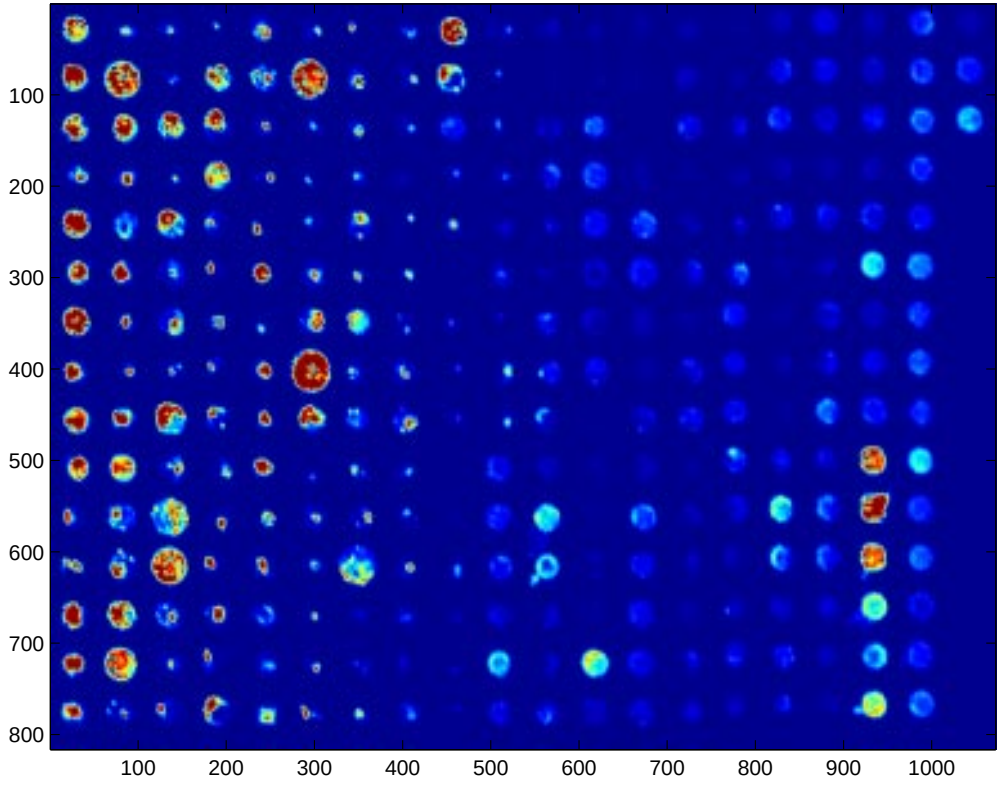


Figure 8: *Blowup of cDNA spotted array.*

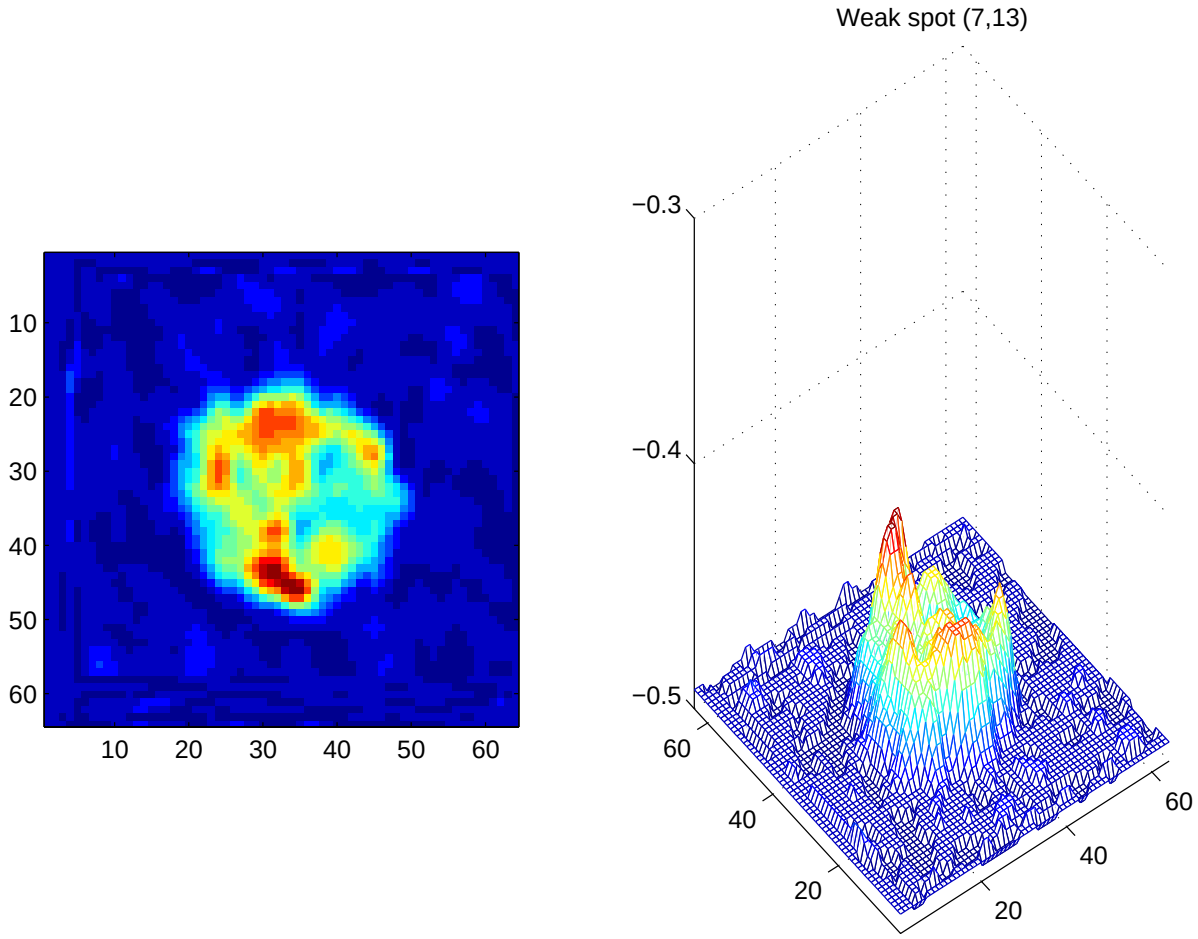


Figure 9: *Weak Spot.*

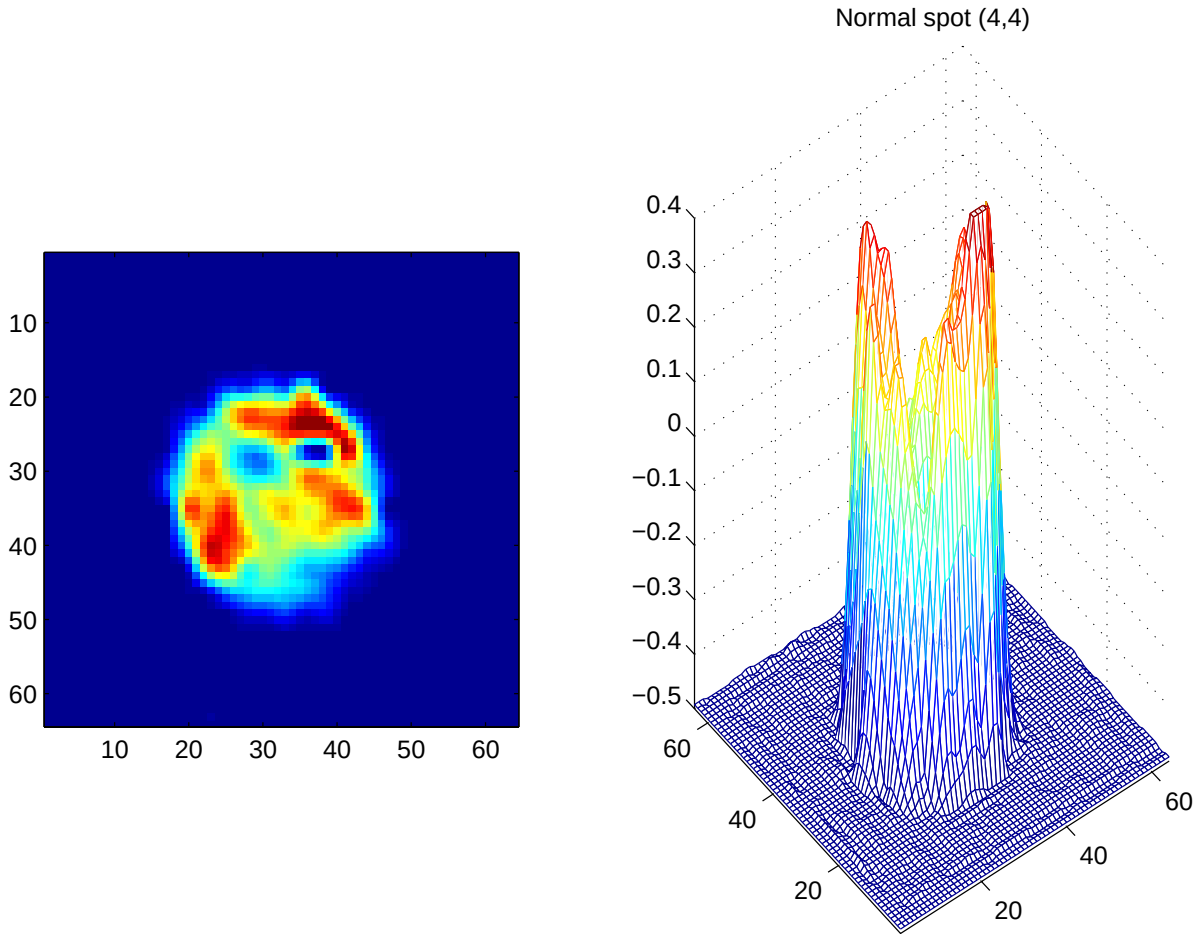


Figure 10: *Normal spot.*

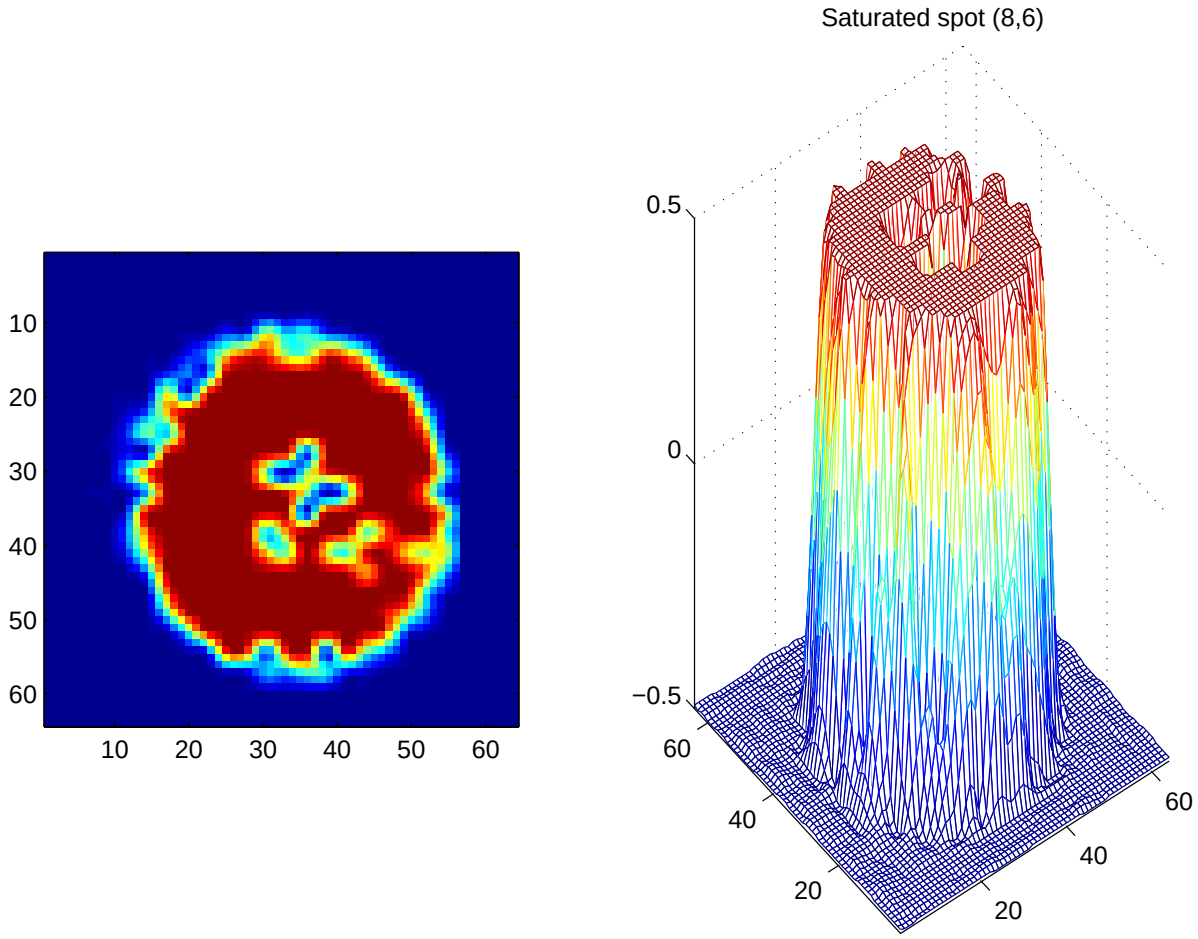


Figure 11: *Saturated spot.*

## Morphological Spot Segmentation (Siddiqui&Hero:ICIP02)

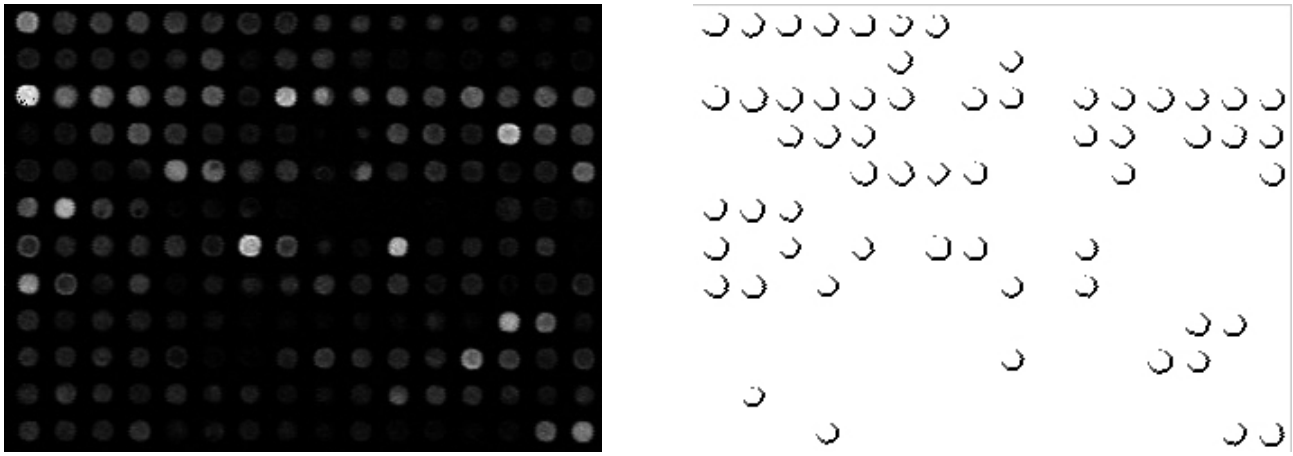

Figure 12: *(L) Original cDNA microarray image. (R) after alternating sequential filtering.*

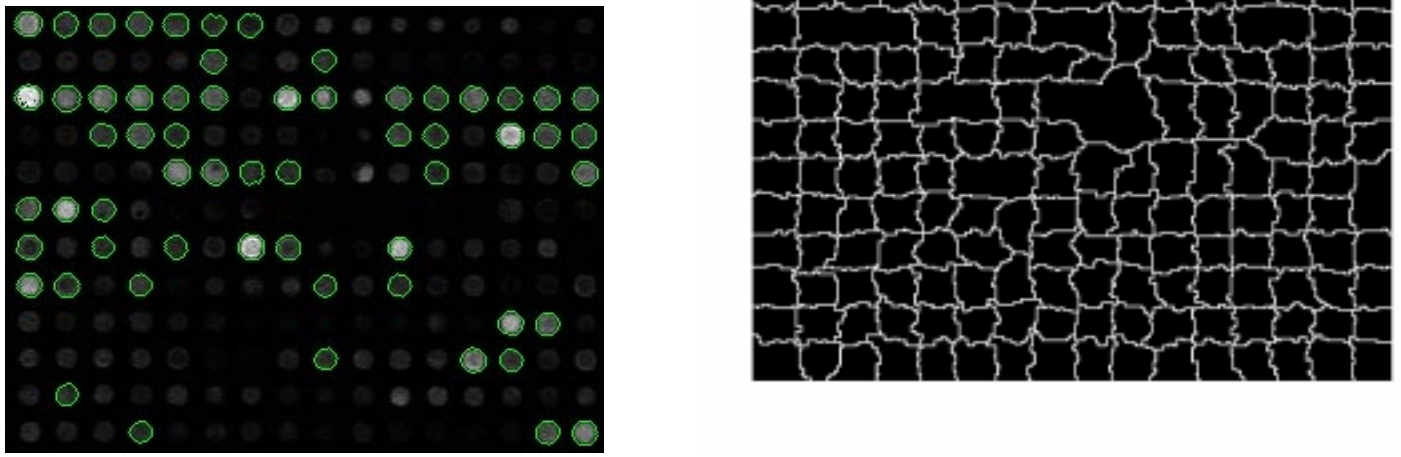

Figure 13: *(L) Final segmentation. (R) Spot watershed domains for noise averaging.*

## Model-based Signal Extraction (Hero:Springer02)

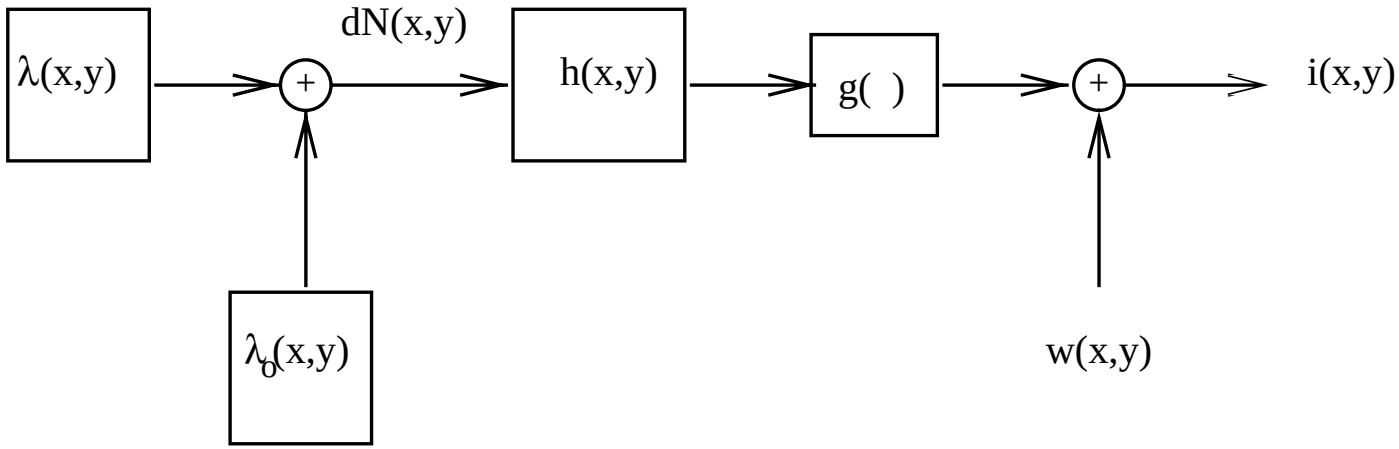


Figure 14: *Filtered Poisson model for microarray image.*

# Gabor Superposition - Width MSE

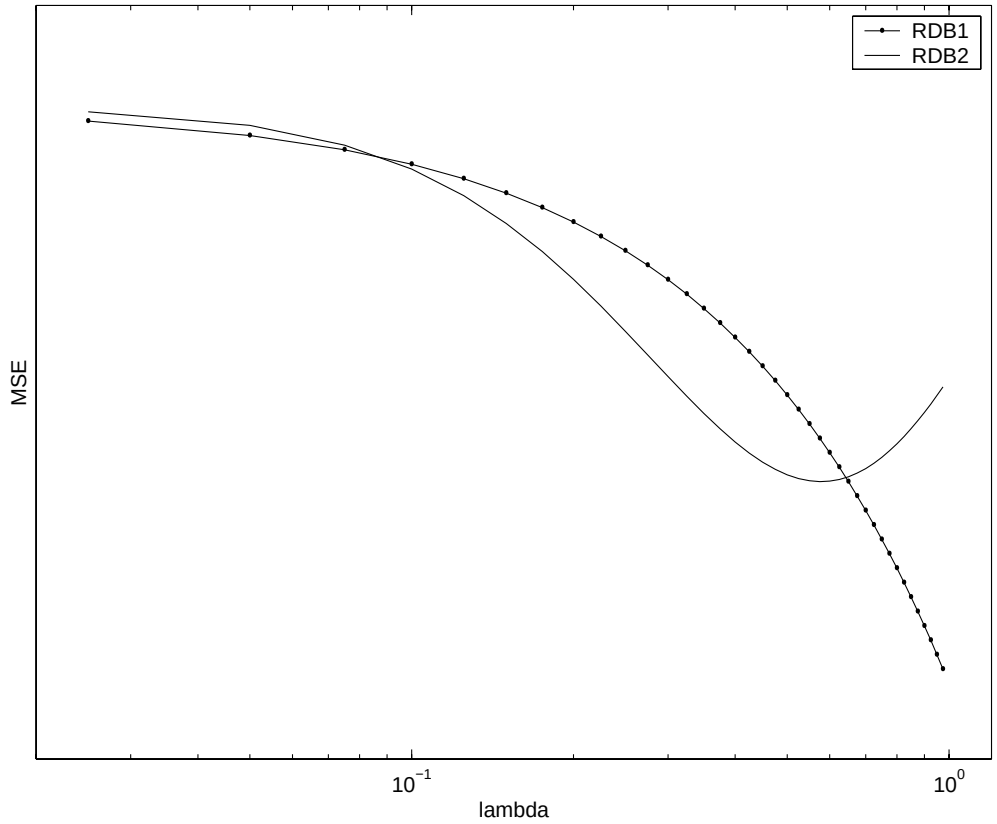


Figure 15: *Distortion-rate MSE lower bounds on Gabor widths of* $\Phi_j(x,y)$.

## Gene Clustering and Filtering (Fleury&etal:ICASSP02)

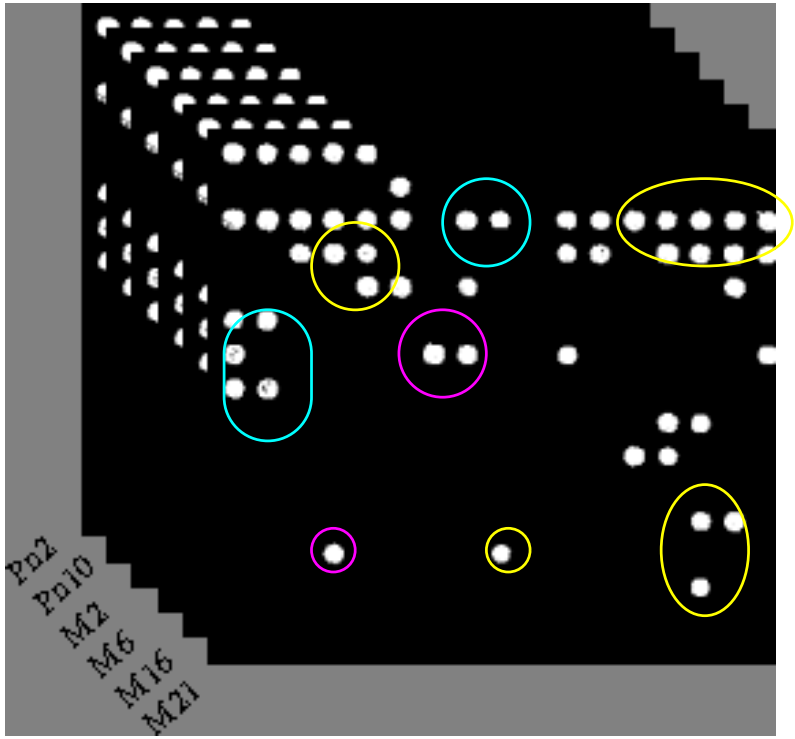


Figure 16: *Clustering on the Data Cube.*

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

# Gene Trajectory Classification

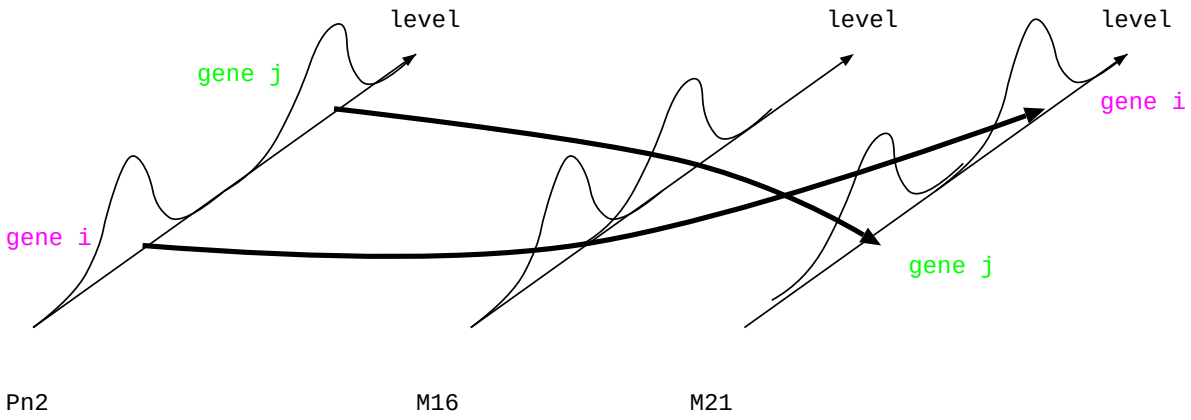


Figure 17: *Gene i is old dominant while gene j is young dominant*

Objective: classify gene trajectories from sequence of microarray experiments over time ($t$) and population ($m$)

$$\theta_i(m,t), \quad m = 1,\ldots,M,\ t = 1,\ldots,T$$

## 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 $i$-th gene $\xi_1(\theta_i),\ ,\ldots,\ \xi_P(\theta_i)$

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

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

$$\xi_1(\theta_i) = \left|\overline{\theta}_i(*,1) - \overline{\theta}_i(*,T)\right|^2$$

- Standard deviation at $t = 1$:

$$\xi_2(\theta_i) = \overline{\left(\theta_i(*,1) - \overline{\theta}_i(*,1)\right)^2}$$

- Standard deviation at $t = T$:

$$\xi_3(\theta_i) = \overline{\left(\theta_i(*,T) - \overline{\theta}_i(*,T)\right)^2}$$

**Some possible scalar functions**:

- $t$-test statistic (Goss etal 2000): $T_i = \frac{\xi_1(\theta_i)}{\frac{1}{2}\xi_2(\theta_i)+\frac{1}{2}\xi_3(\theta_i)}$
- $R^2$ statistic (Hastie etal 2000): $R_i^2 = \frac{T_i}{1+T_i}$
- $H$ statistic (Sinha etal 1998): $H_i = \frac{\xi_1(\theta_i)}{\sqrt{\xi_2(\theta_i)\xi_3(\theta_i)}}$

**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:

$$\max_i \sum_{p=1}^{P} W_p \xi_p(\theta_i), \quad \textit{or} \quad \max_i \prod_{p=1}^{P} (\xi_p(\theta_i))^{W_p}$$

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.

$$\xi_p(\theta_i) \leq \xi_p(\theta_j), \; p = 1, \ldots, P$$

# Pareto Optimal Fronts

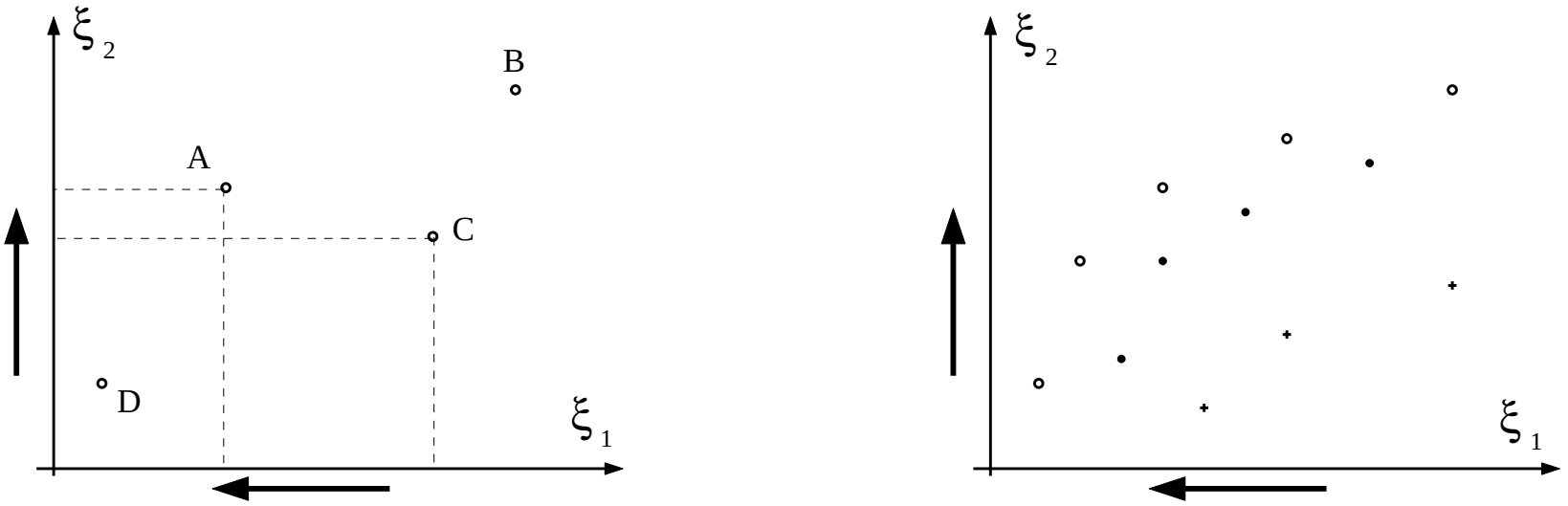


Figure 18: *a). Non-dominated property, and b). Pareto optimal fronts, in dual criteria plane.*

## Pareto Gene Filtering vs. Paired T-test

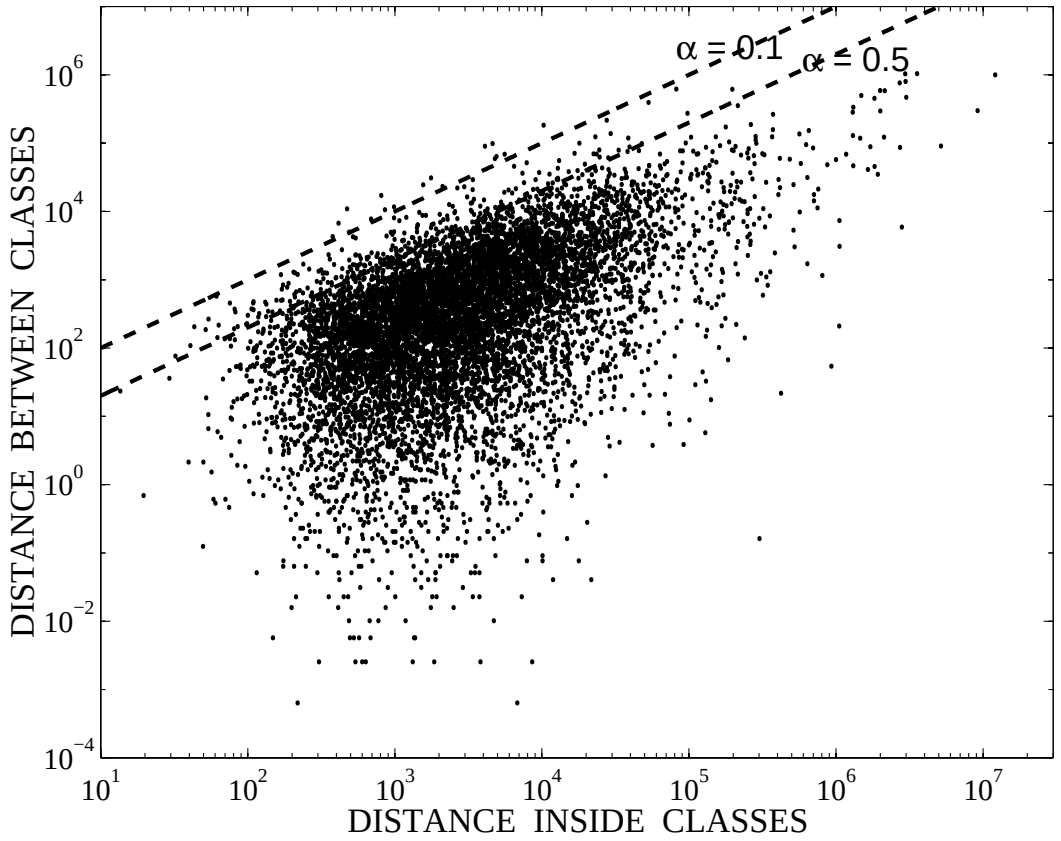


Figure 19: $\xi_1$ = *mean change vs* $\xi_2$ = *pooled standard deviation for 8826 mouse retina genes. Superimposed are T-test boundaries*

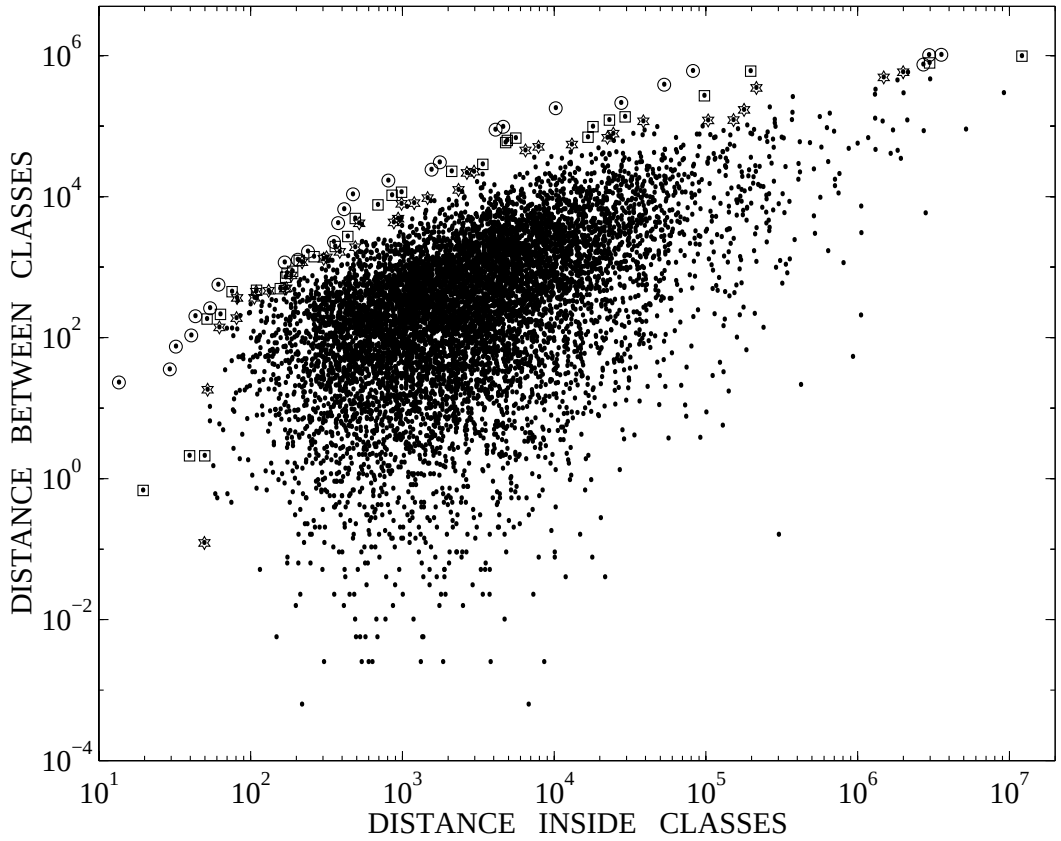


Figure 20: *First (circle) second (square) and third (hexagon) Pareto optimal fronts.*

## 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

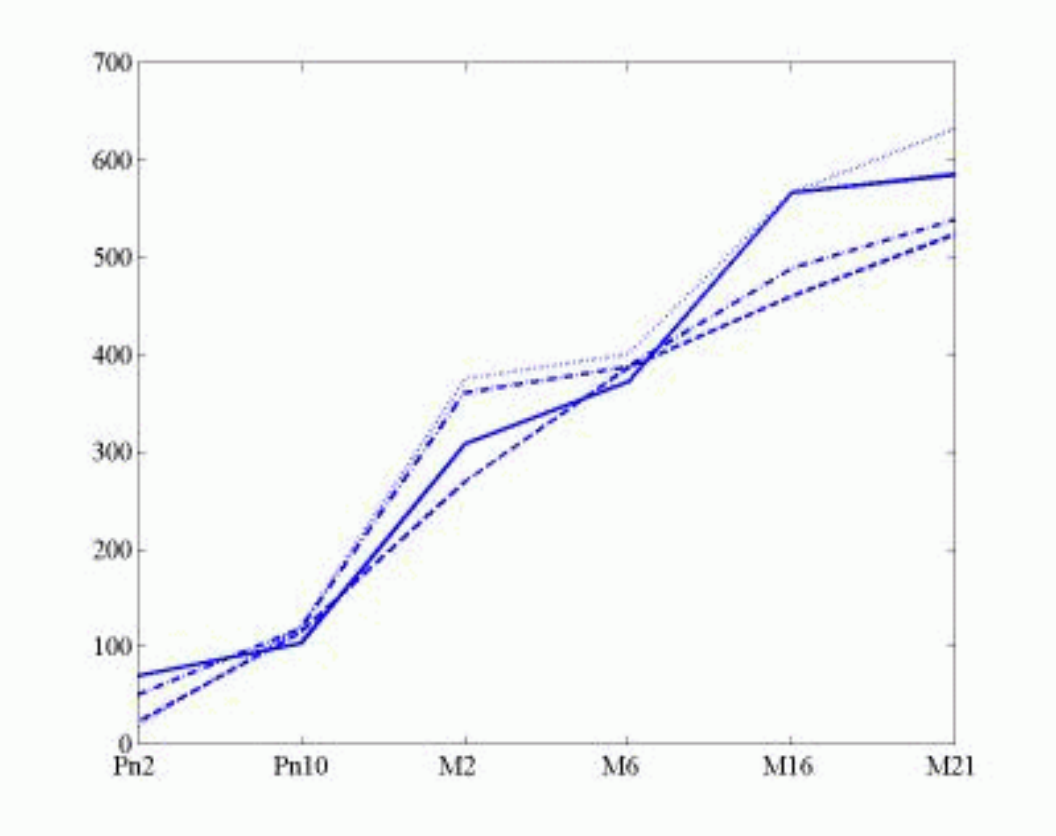


Figure 21: *4 candidate gene profiles from Mus musculus* $5'$ *end cDNA (Unigene 86632)*

## Multi-objective Non-parametric Pareto Filtering

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

- Old dominant filtering criteria:
  - high mean slope from $t = Pn1$ to $t = M21$

$$\xi_1(\psi_i) = \overline{b_i(*,*)}$$

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

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

# Occurence Histogram

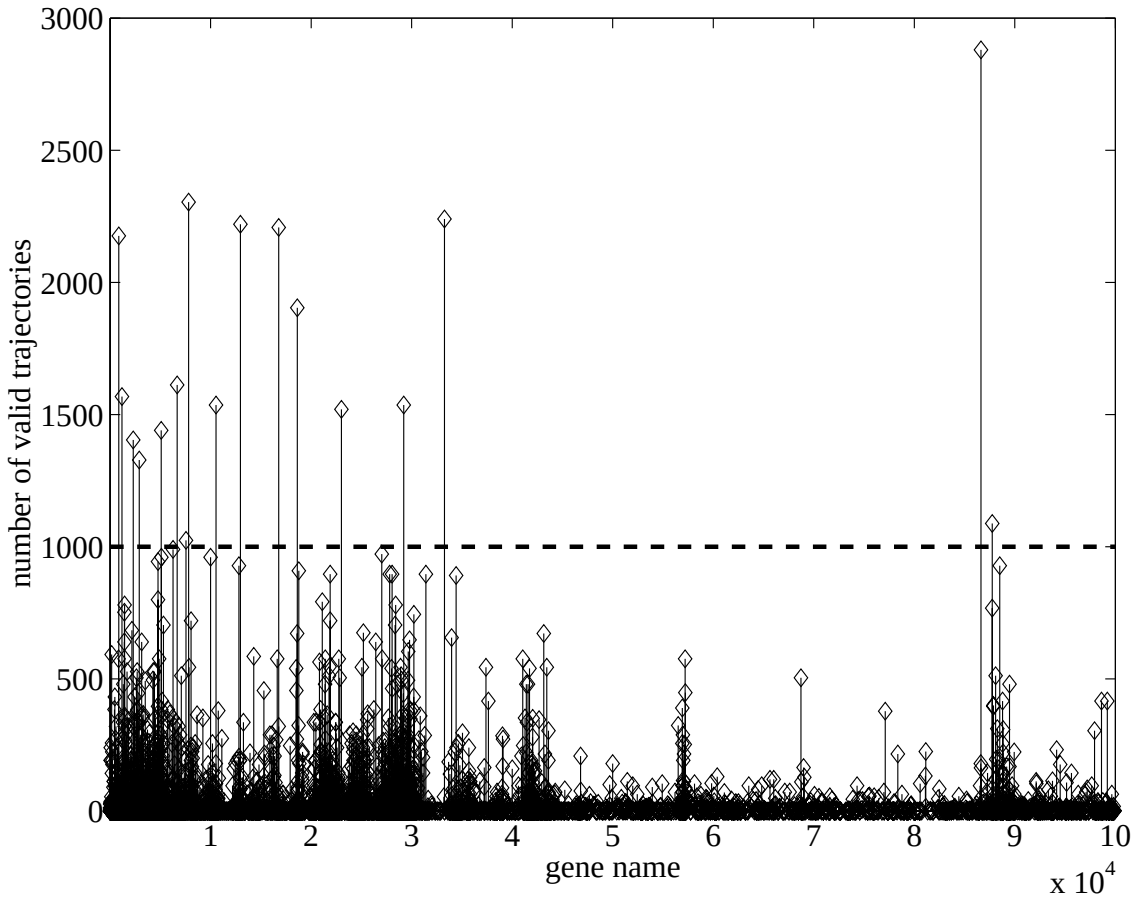


Figure 22: *Monotonicity occurrence histogram with threshold.*

# Old Dominant Pareto Fronts

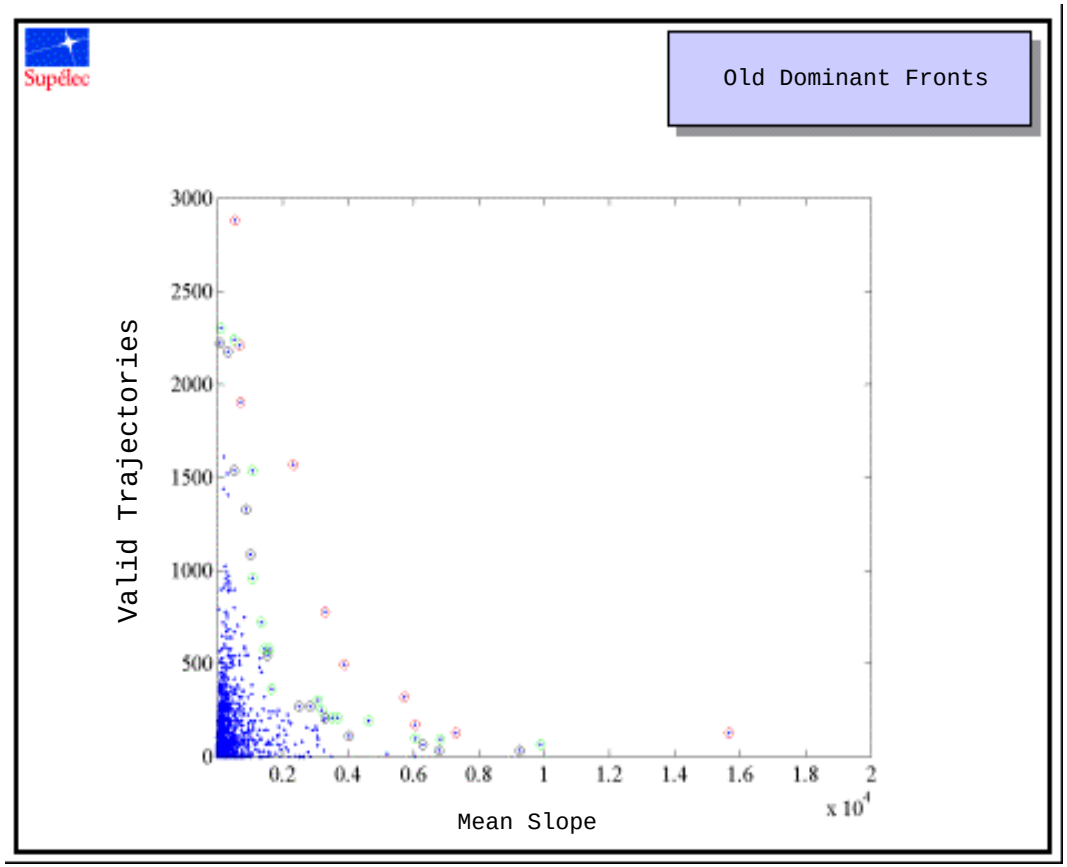


Figure 23: *Pareto fronts for old dominant genes.*

## Resistant Old Dominant Genes in first Three Fronts

- Leave-one-out cross validation

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

Cross-validation Algorithm:

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

`Compute` $\left(\xi_1(\psi_i^{-m}), \xi_2(\psi_i^{-m})\right)$

`Find Genes in First 3 Pareto fronts:` $G^{-m}$

`End`

`Resistant Genes =` $\cap_{m=1}^{4^6} G^{-m}$

| Unigene # | Affymetrix description |
|---|---|
| **1186** | *Mouse Carbonic Anhydrase II cDNA* |
| 1276 | Retinal S-antigen |
| 2965 | Mouse opsin gene |
| 3918 | ATP-binding casette 10 |
| **16224** | Guanylate cyclase activator 1a (retina) |
| **16763** | Mouse mRNA for aldolase A |
| **16771** | *Mus musculus H-2K* |
| 39200 | CGMP phosphodiesterase gamma |
| **42102** | Mus musculus tubby like protein 1 mRNA |
| **69061** | Guanine binding protein $\alpha$ transducing 1 |
| **86632** | *Mus musculus 5'end cDNA* |

Table 1: *Resistant genes remaining in first three Pareto fronts*

## Three-objective Pareto Filtering

**Objective** Extract "aging genes"

- Strictly increasing filtering criteria:
  - persistent positive trend

$$\xi_1(\psi_i) = \overline{\min_t b_i(*,t)} = \max$$

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

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

  - no plateau

$$\xi_3(\theta_i) = \overline{[\theta_i(*,t+1) - 2\theta_i(*,t) + \theta_i(*,t-1)]^2} = \min$$

## Pairwise Pareto Fronts

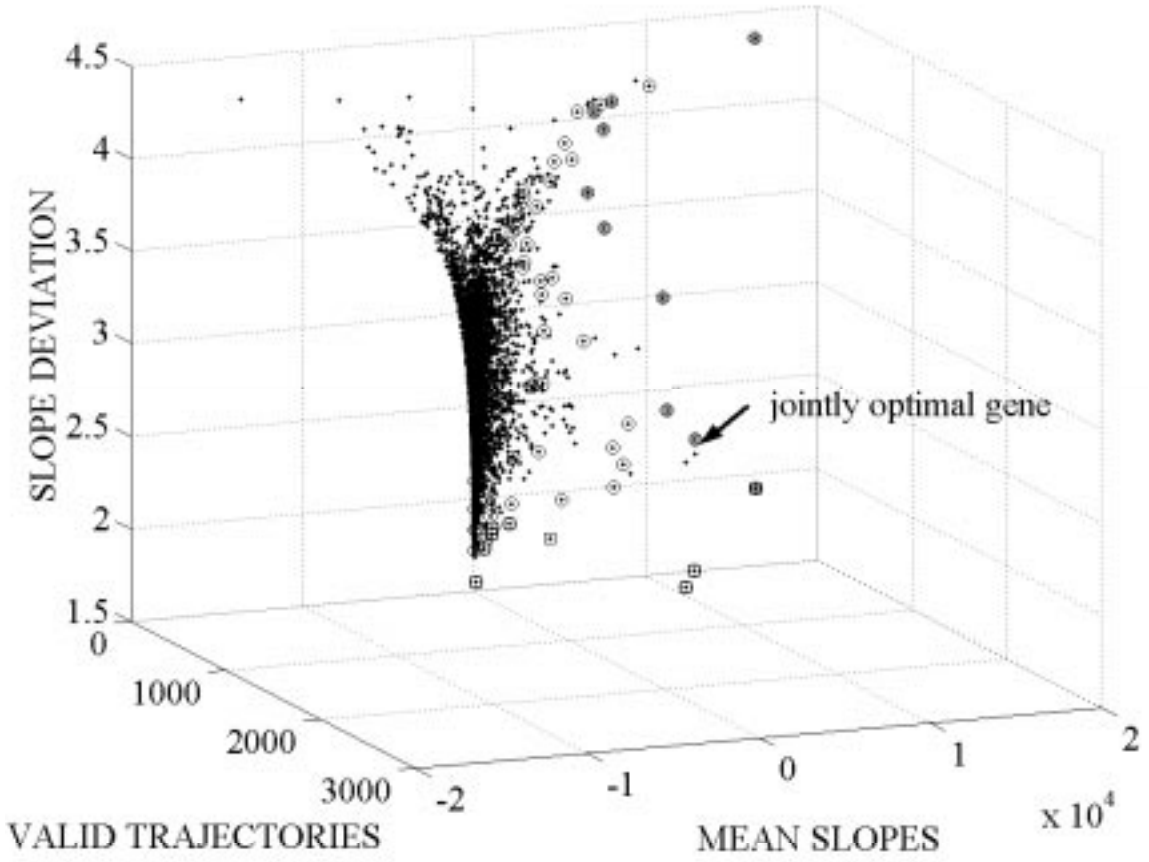


Figure 24: *First Pareto fronts for each pair of criteria taken from the set ($\xi_1$, $\xi_2$ and $\xi_3$). Each one of this front is denoted by squares, circles and stars, respectively.*

# Aging Genes Found by Pareto Filter

| Unigene # | Front | Description |
|---|---|---|
| 7800 | 1st | Inositol triphosphate receptor type 2 |
| **86632** | 2nd | Histocompatibility 2, L Region |
| 12956 | 2nd | Hyperpolarization-activated, cylcic nucleotide-gated K |
| 29213 | 3rd | RIKEN cDNA 1200015F23 gene |
| 33263 | 3rd | Histocompatibility 2, D region locus 1 |
| 29789 | 3rd | Expressed sequence A1430822 |
| 2289 | 3rd | RIKEN cDNA 1500015A01 gene |
| 6671 | 3rd | RIKEN cDNA 1110027O12 gene |
| **16771** | 4th | MHC class 1 antigen H-2K |
| 34421 | 4th | Q4 class 1 MHC |
| 6252 | 4th | Procollagen, type XIX, alpha 1 |
| 29357 | 4th | RIKEN cDNA 1300017C10 gene |

Table 2: *Resistant aging genes remaining in first four Pareto fronts*

## Conclusions

1. Signal processing has a role to play in many aspects of genomics
2. Careful physical modeling of image formation process can yield performance gains
3. New methods of data mining are needed to perform robust and flexible gene filtering
4. Cross-validation is needed to account for statistical sampling uncertainty
5. Joint intensity extraction and gene filtering?
6. Optimization algorithms for large data sets?
7. Genetic priors: phylogenetic trees, BLAST database, etc?