Bioinformatics and Genomics: A New SP Frontier?

A. O. Hero

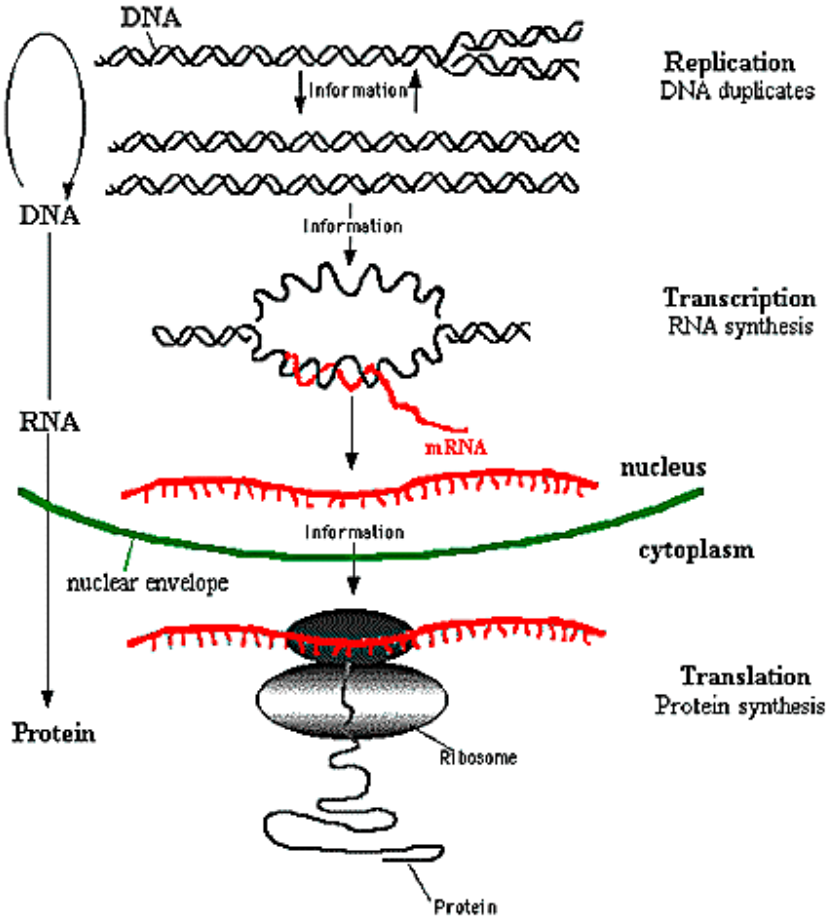
University of Michigan - Ann Arbor

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

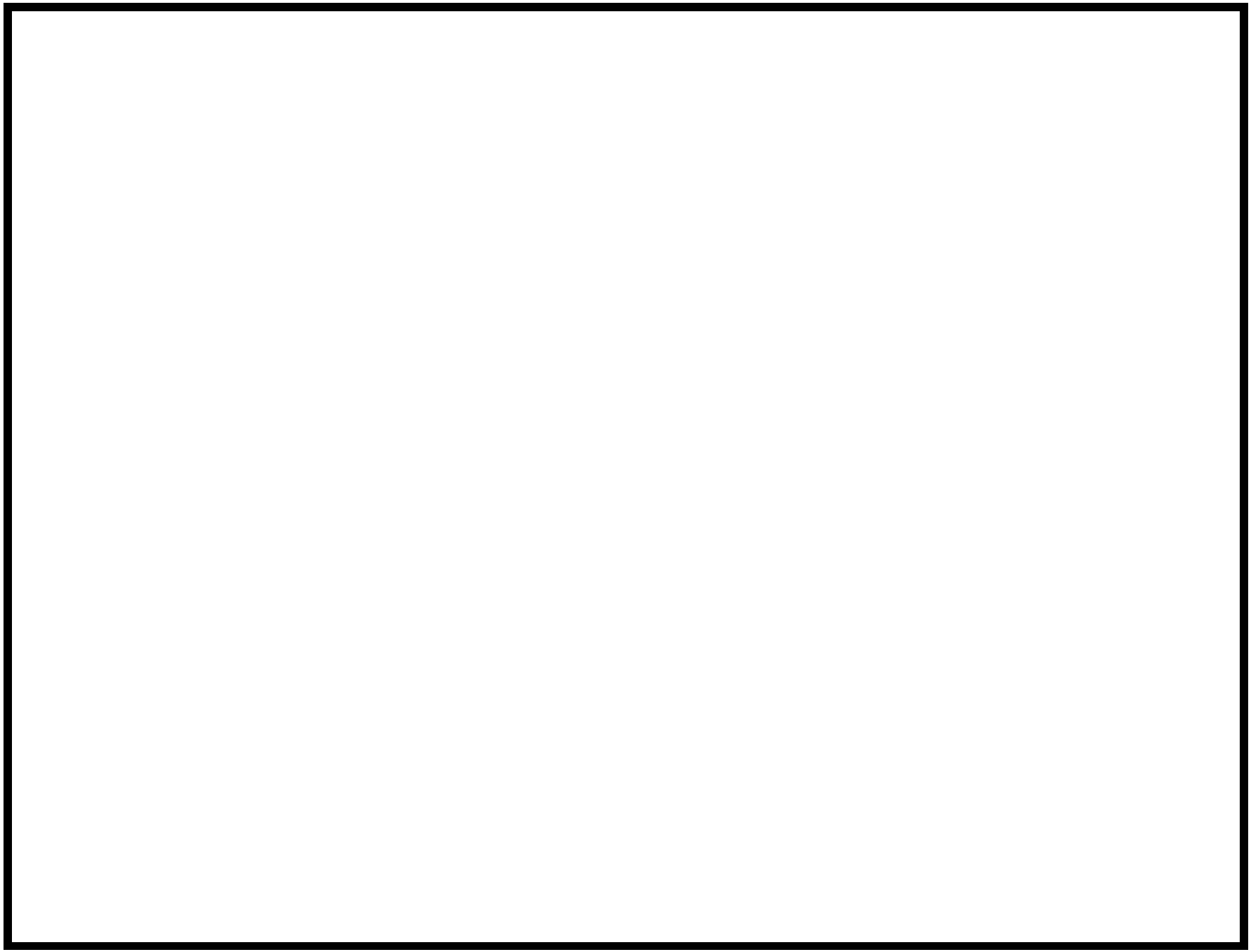
Outline

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



The Central Dogma of Molecular Biology

Figure 1:



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 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

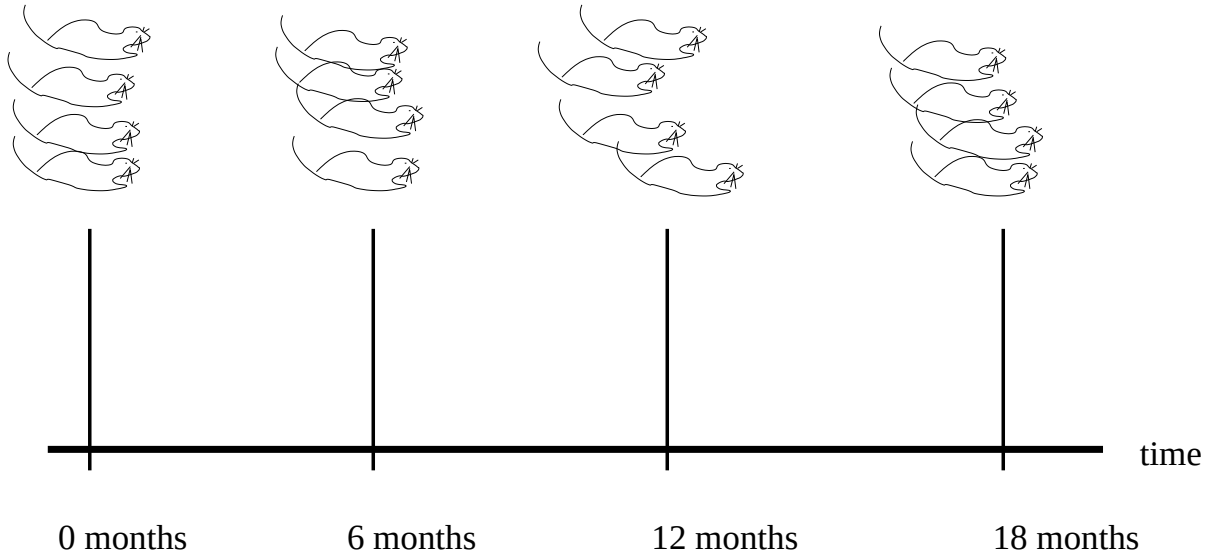


Figure 2: *Sample gene trajectories over time.*

II. Gene Microarrays

"Shotgun sequencing"

Figure 3: *Microarray experiment cycle.*

Microarray Technologies

Two principal microarray technologies:

1. Oligonucleotide chips (Affymetrix GeneChip)
2. cDNA spotted arrays (Synteni/Stanford chip)

These technologies share common experimental procedure...

1. Specific complementary ss DNA sequences (probes) are deposited at spots on a slide (*arraying*)
2. Dye-labeled DNA from sample is distributed over slide - complementary DNA binds to probes (*hybridization*)
3. Presence of bound DNA is read out by detecting spot flourescence by laser excitation (*scanning*)

Microarray Image Formation

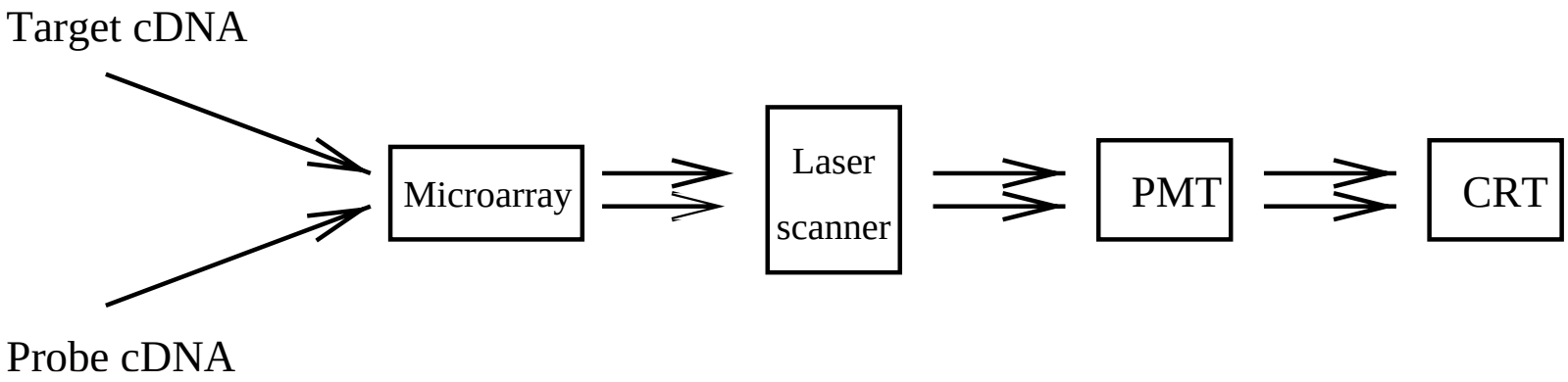


Figure 4: *Image formation process.*

...but these technologies also differ...

1. Gene fragment (Oligonucleotide) probes vs. whole gene (cDNA) probes
2. Probe etching to slide via photolithography (Affy) vs. spotting (Stanford)
3. Business model: corporate (Affy) vs. grassroots (Stanford)

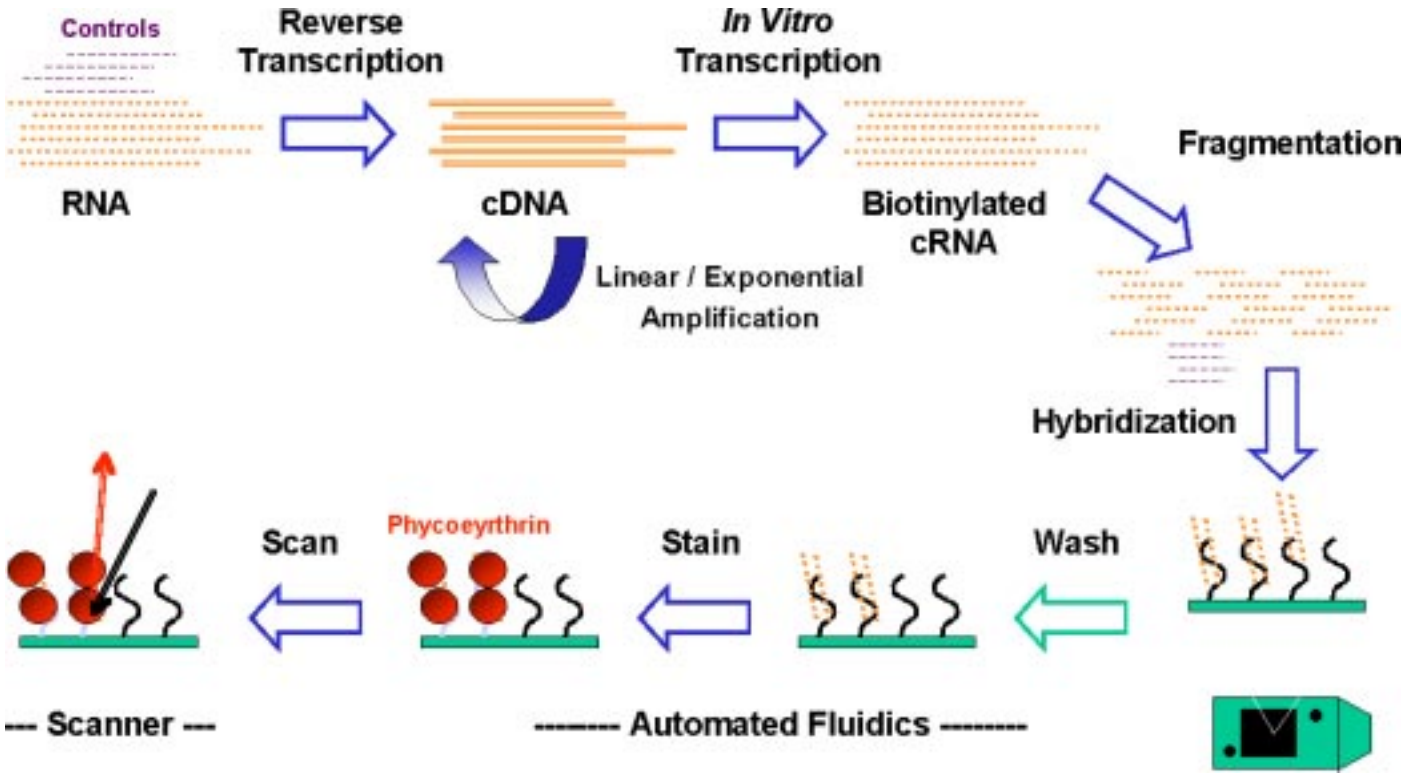


Figure 5: *Oligonucleotide (GeneChip) system (*`pathbox.wustl.edu`*).*

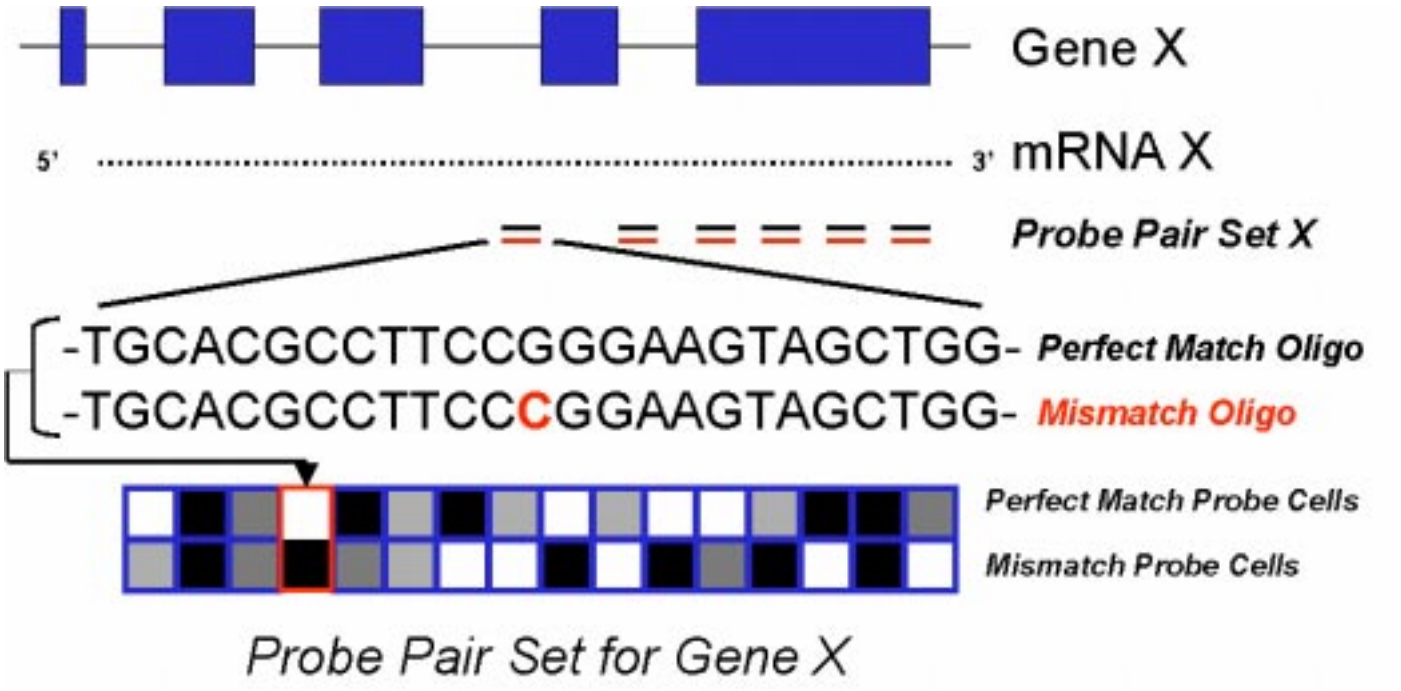


Figure 6: *Oligonucleotide PM/MM layout (*`pathbox.wustl.edu`*).*

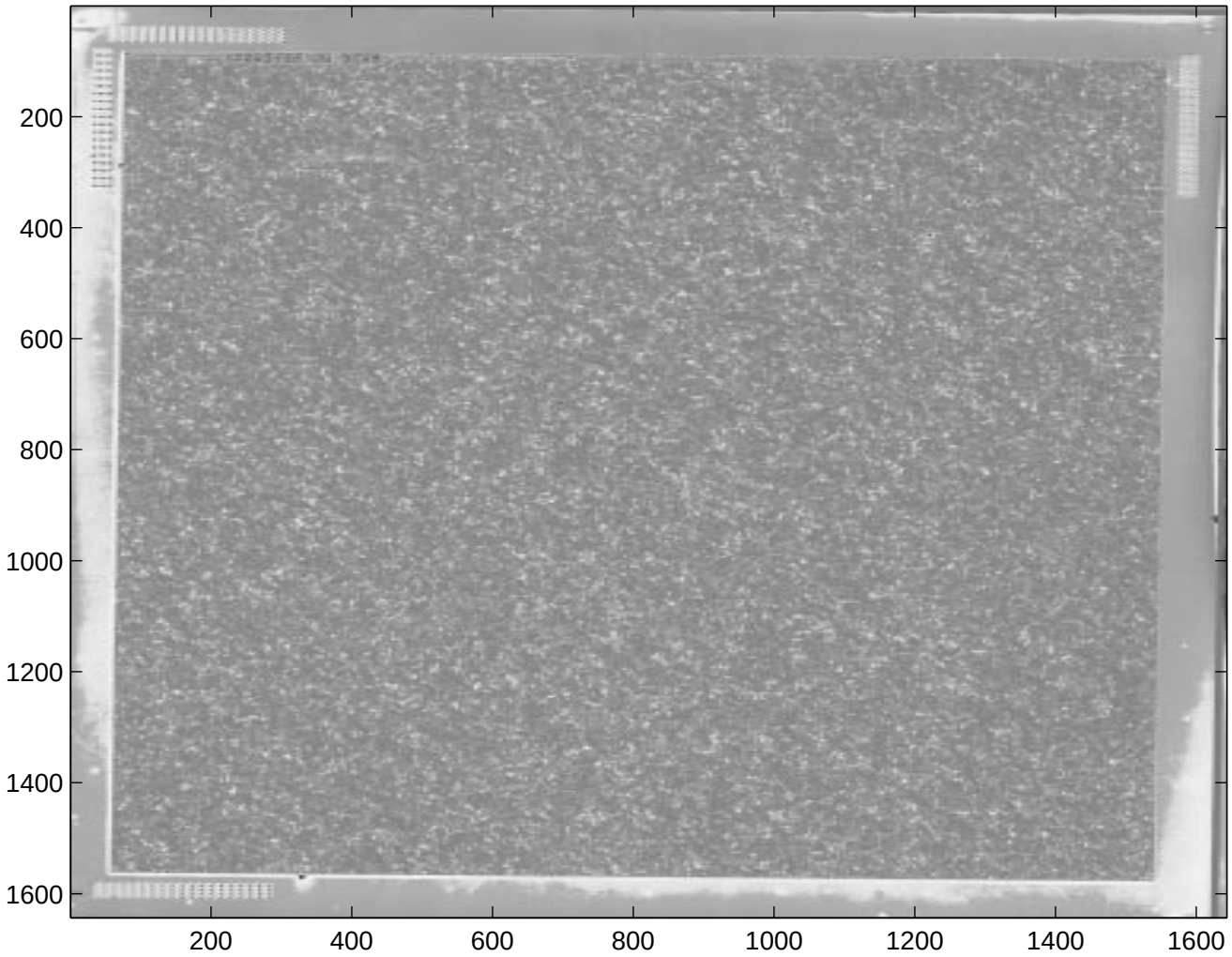


Figure 7: *Affymetrix GeneChip microarray.*

(Affymetrix) Output for Each Gene Probe

- **Avg-diff**: avg differences between 20 PM and MM pairs
- **Log-avg** : log of ratios between 20 PM and MM pairs
- **Positive probe pairs**: number of matches to PM
- **Negative probe pairs**: number of matches to MM
- **Absolute Call**: P,A,M

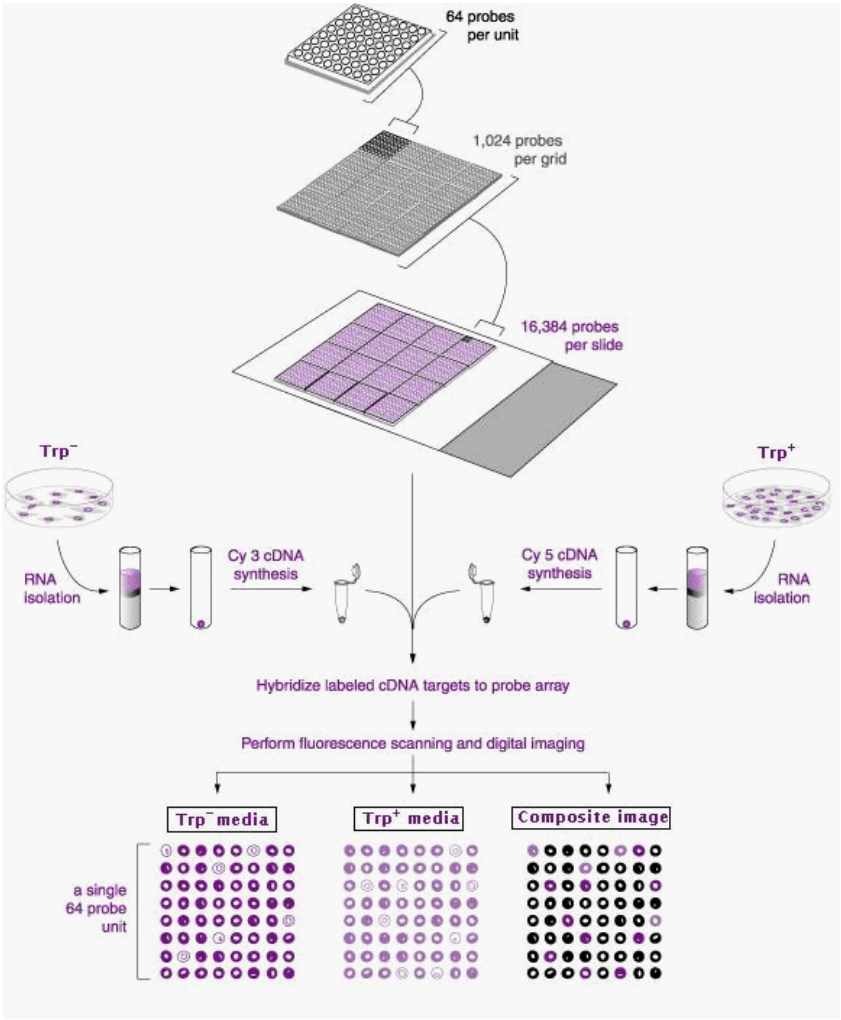


Figure 8: *cDNA (Stanford) system (*`teach.biosci.arizona.edu`*).*

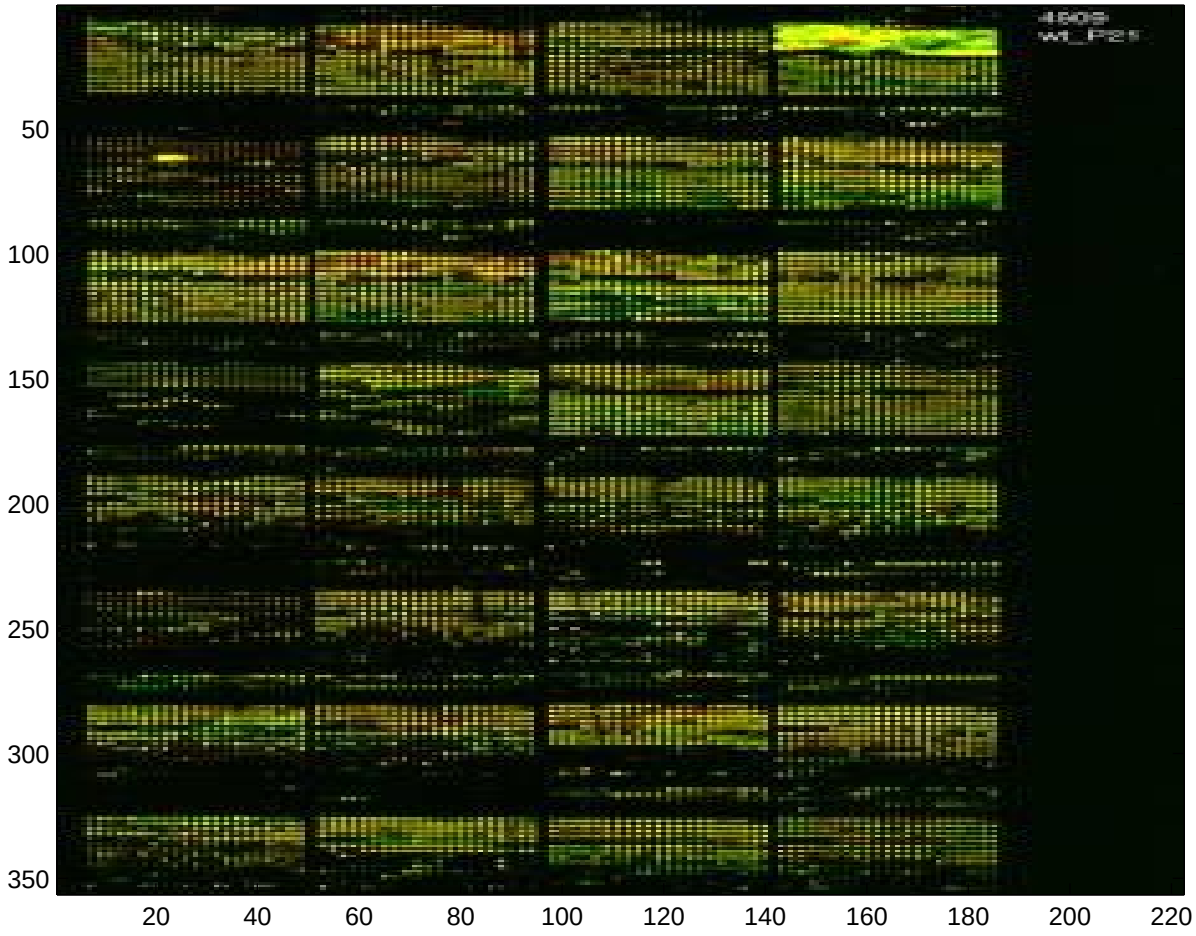


Figure 9: *cDNA spotted array.*

Control Factors Influencing Variability

- **Sample preparation**: reagent quality, temperature variations
- **Slide manufacture**: slide surface quality, dust deposition
- **Hybridization**: sample concentration, wash conditions
- **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
- **Cross hybridization**: similar but different genes bind to same probe
- **Raw Q**: noise and other random variations of a chip
- **Background**: avg of lowest 2% cell intensity values
- **% P**: percentage of transcripts present

Model-Based Signal Extraction

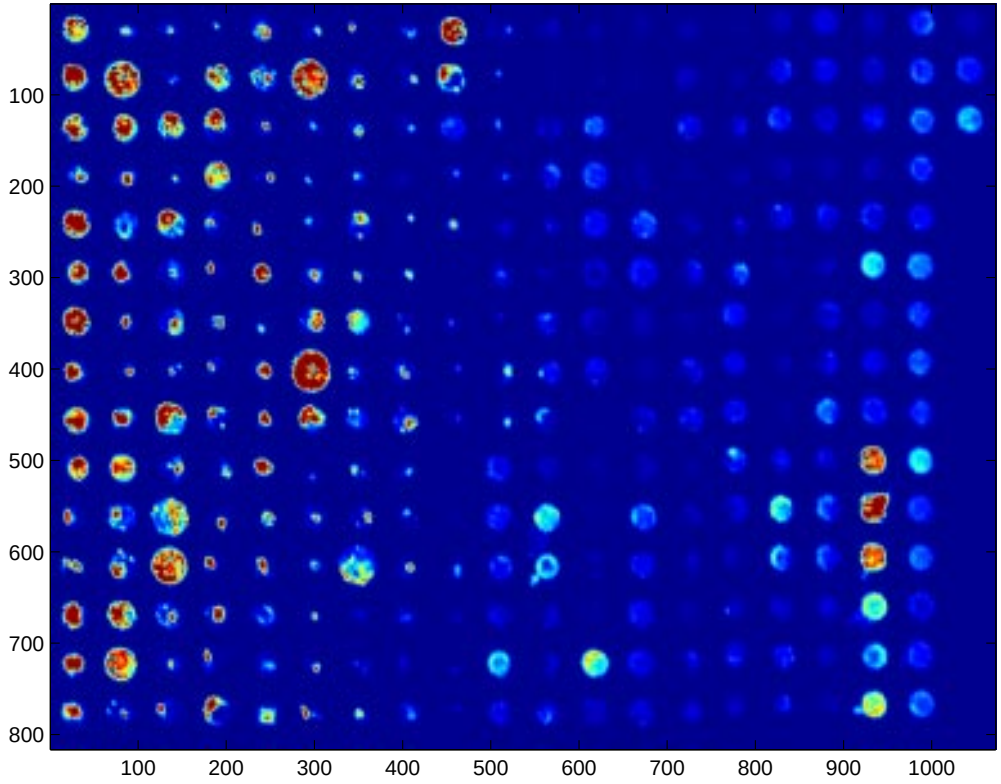


Figure 10: *Blowup of cDNA spotted array.*

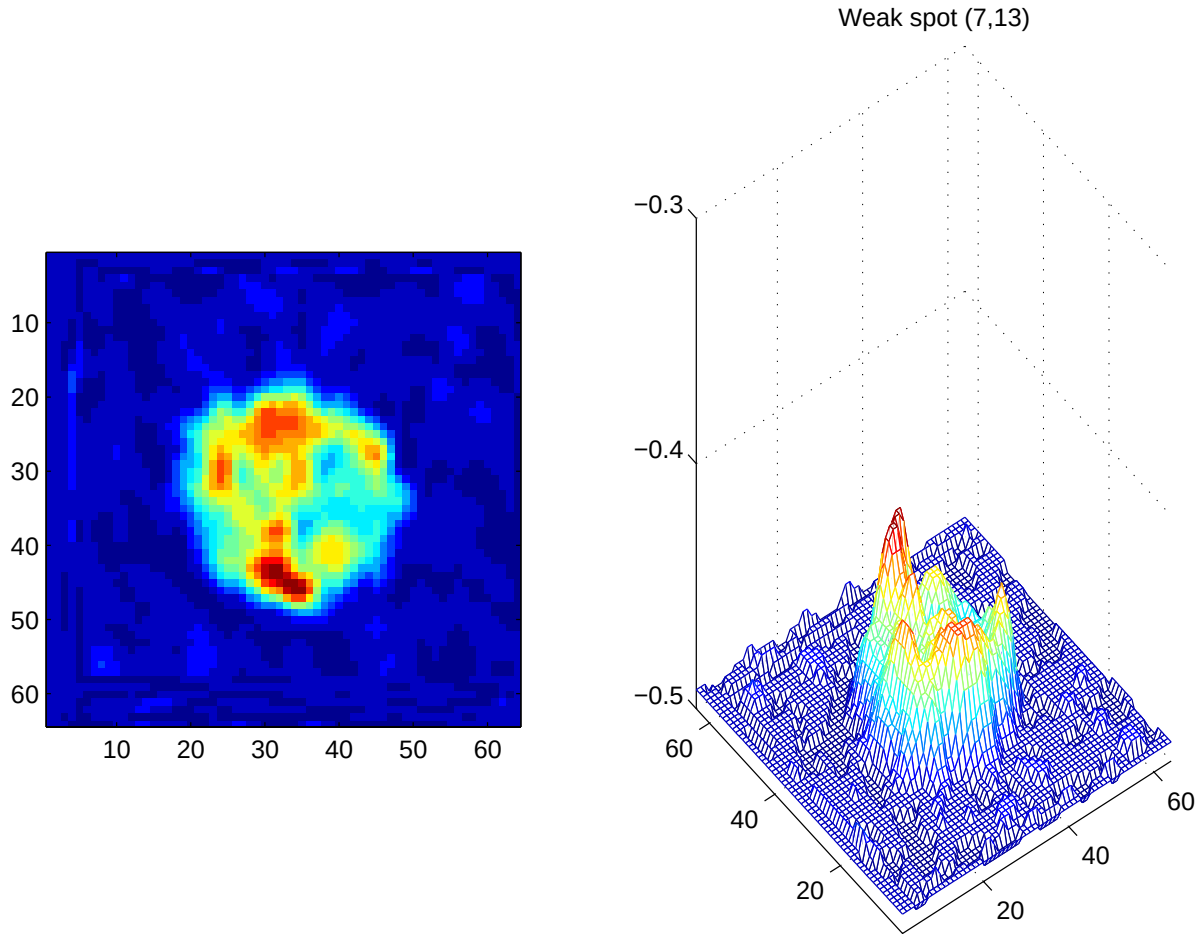


Figure 11: *Weak Spot.*

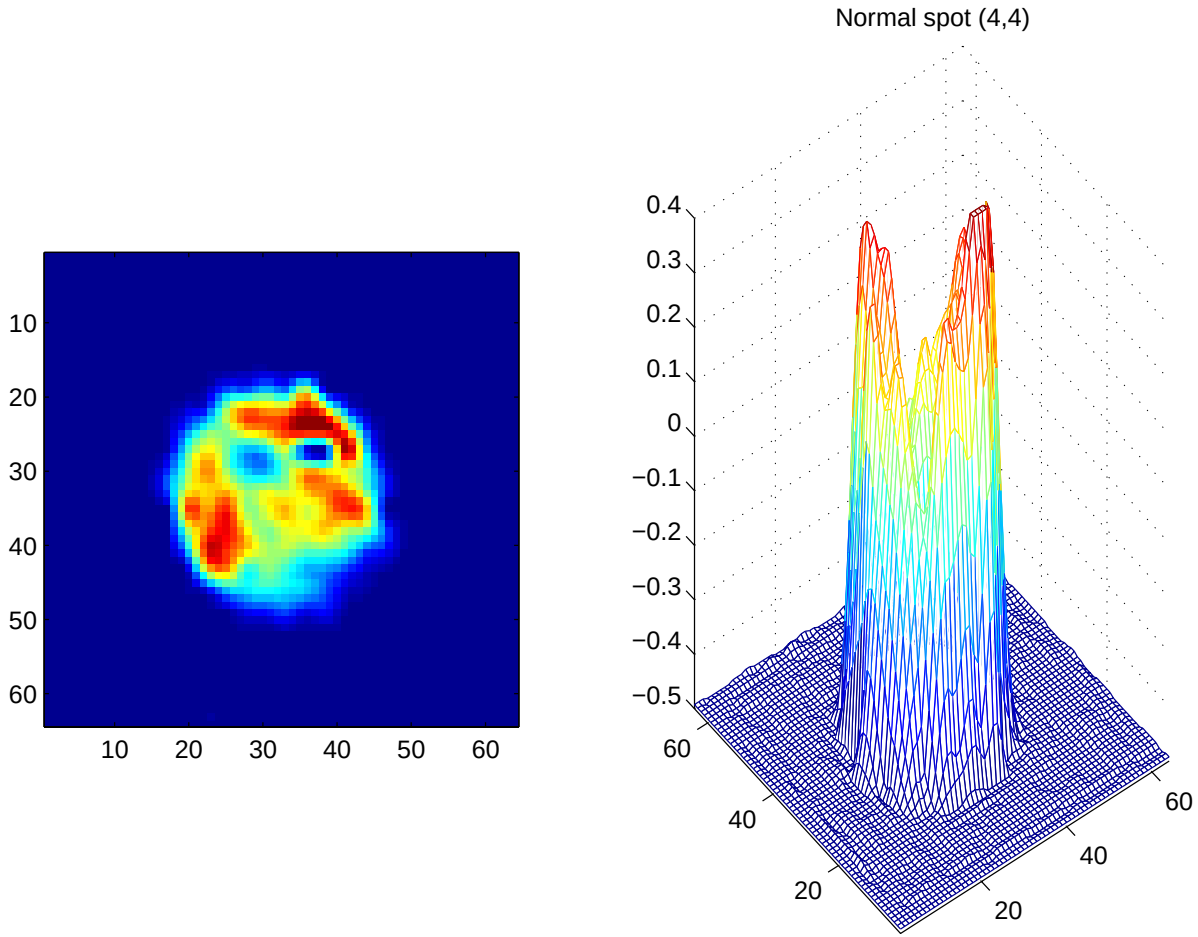


Figure 12: *Normal spot.*

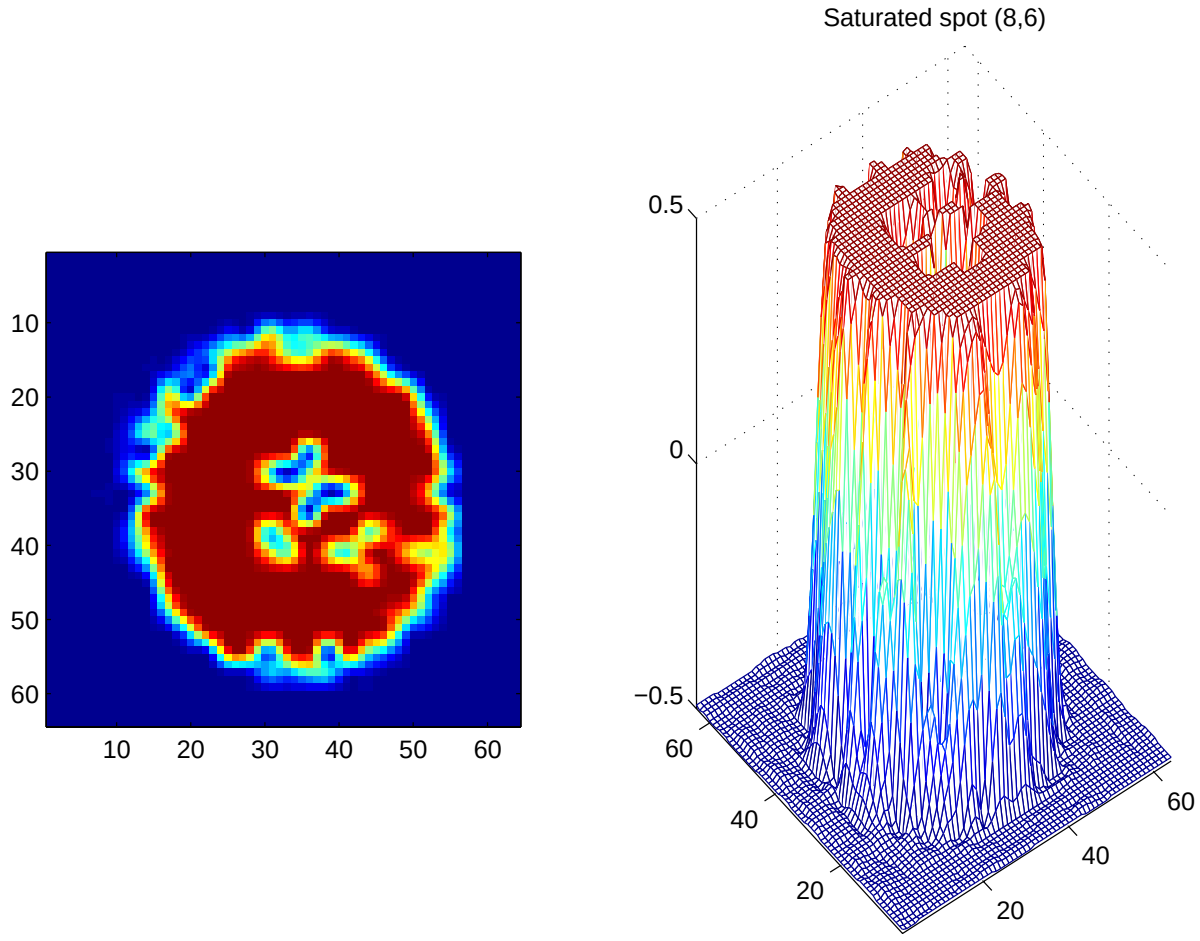


Figure 13: *Saturated spot.*

Filtered Poisson Measurement Model

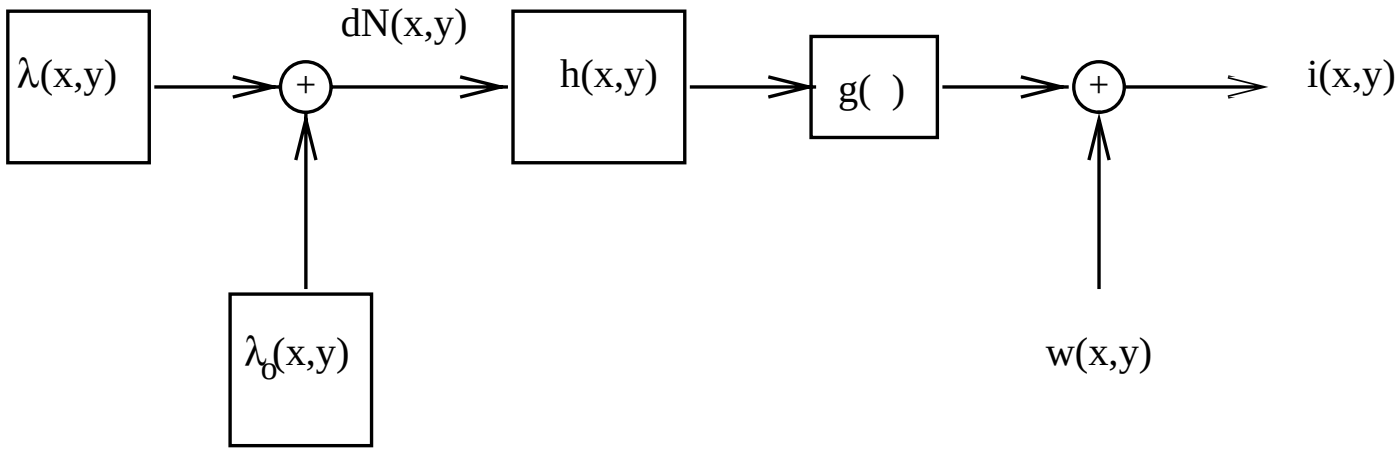


Figure 14: *Filtered Poisson model for microarray image.*

Mathematical Model

$$I(x,y) = g\left(r\int\int h(x-u,y-v)dN(u,v)\right) + w(x,y)$$

- $I(x,y)$: measured intensity
- $dN(u,v)$: inhomogeneous spatial Poisson process with intensity $\lambda_d + \lambda_o$
- $h(u,v)$: point spread function of image scanner
- g: spatially homogeneous non-linear response function
- $w(u,v)$: thermal electronic noise

Extraction of Gene Hybridization Levels

Objective: Estimate $\theta_j,\ j = 1,\ldots,\#_{probes}$

$$\lambda(x,y) = \sum_{j=1}^{\#_{probes}} \theta_j \Phi_j(x-u_j, y-v_j)$$

where

- $\Phi_j(u,v)$: (normalized) intensity of j-th spot

  Multi-component model for Φ_j

$$\Phi_j(u,v) = \sum_{k=1}^{\#basis} \alpha_{j,k}\phi_k(u,v)$$

- u_j, v_j: position of j-th spot

Compound Channel Representation

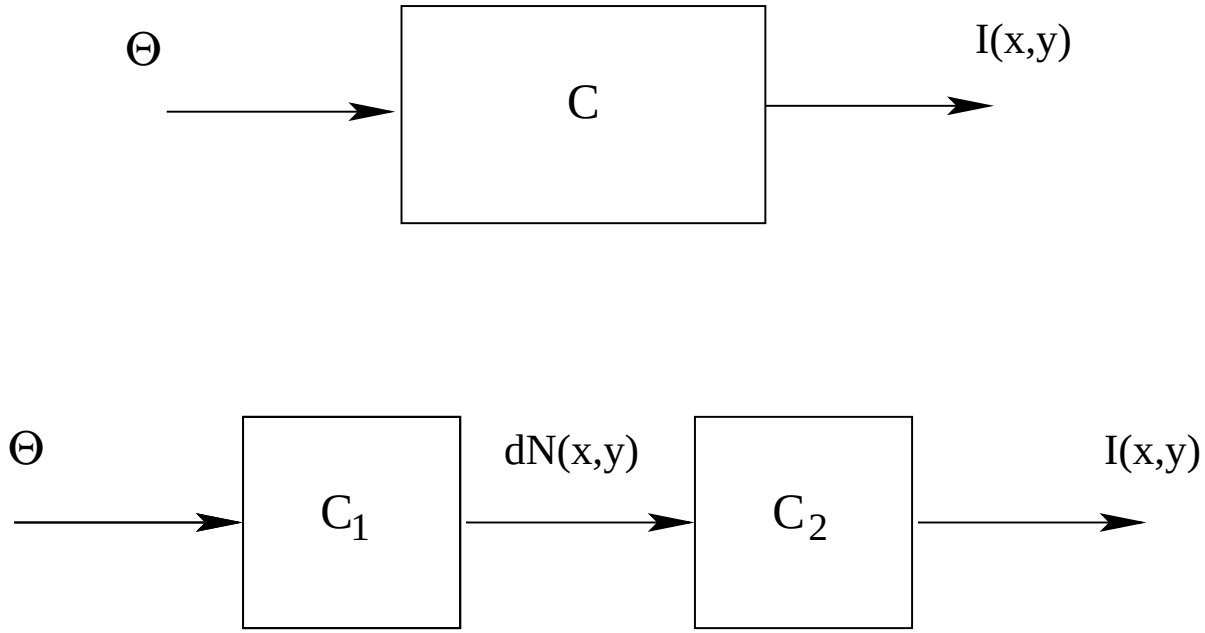


Figure 15: *Top: statistical representation of I as the output of channel C with input Θ. Bottom: decomposition of C into Poisson and Gaussian channels C_1 and C_2, respectively.*

Gabor Superposition - Spot Position MSE

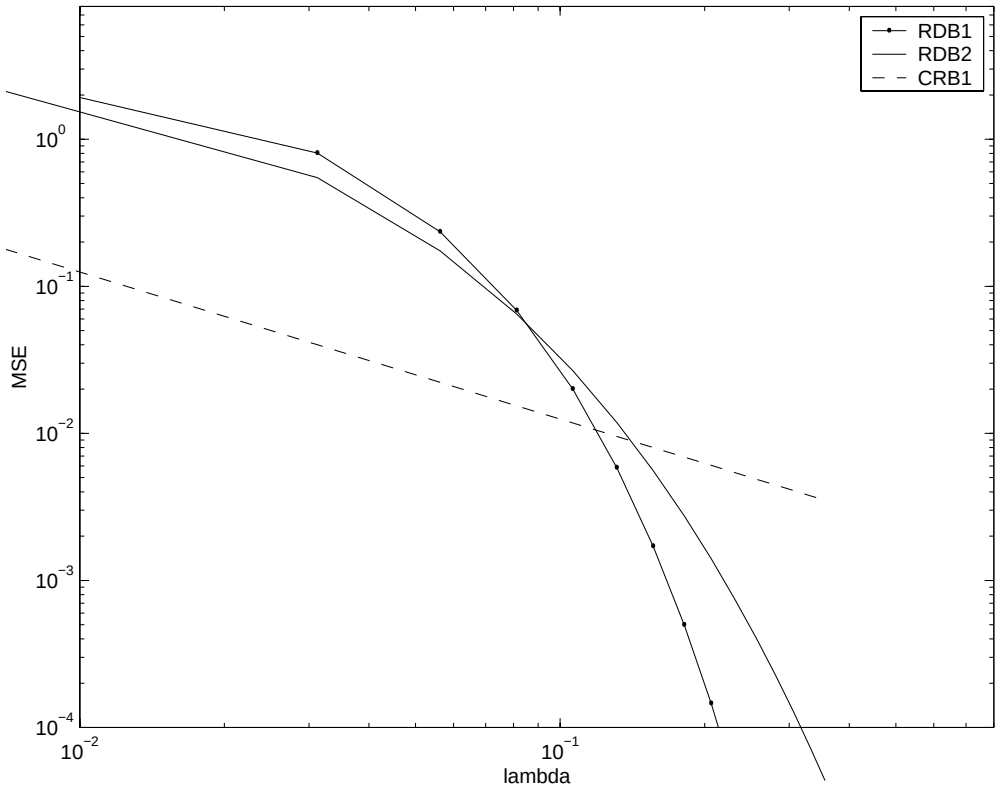


Figure 16: *MSE lower bounds on spot position.*

Gabor Superposition - Width MSE

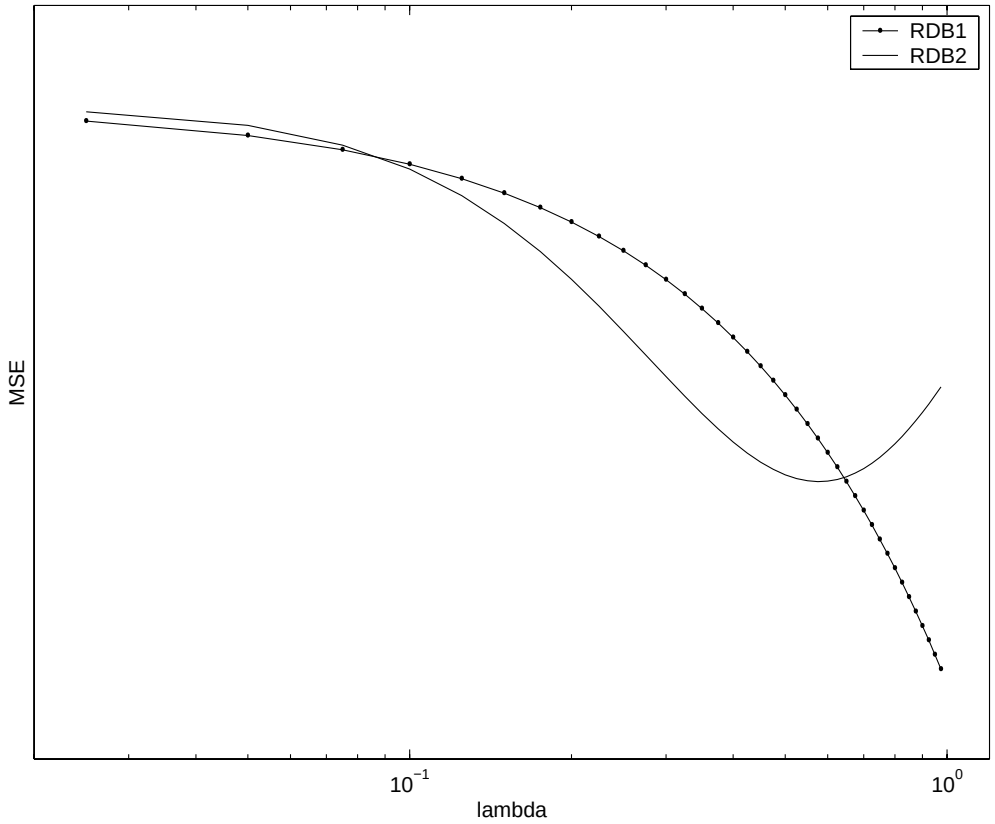


Figure 17: *Distortion-rate MSE lower bounds on Gabor widths of* $\Phi_j(x,y)$.

Optimal Gene Extraction

Imputed Log-likelihood function (Antoniadis&Hero:SP92):

$$l(\theta,\alpha,u,v) = \int\int \widehat{dN}(x,y)\ln(\lambda(x,y)+\lambda_o) - \int\int \lambda(x,y)dxdy$$

where

$$\begin{aligned}\widehat{dN}(u,v) &= E[dN(u,v)|I;\ \overline{\theta},\overline{\alpha},\overline{u},\overline{v}] \\ \lambda(x,y) &= \sum_{j=1}^{\#_{probes}} \theta_j \Phi_j(x,y)\end{aligned}$$

Assuming

- $g(u) = u$
- spot intensities Φ_j don't overlap
- $\lambda_o = 0$

$$\hat{\theta}_j = \int\int_{\text{cell}_j} \widehat{dN}(x,y)$$

Extraction of Differential Hybridization Levels

For $d = 1, 2$:

$$I_d(x,y) = r_d \int\int h_d(x-u, y-v)\, dN_d(u,v) + w_d(x,y)$$

Estimate of $\Delta\theta_j = \theta_{1j}/\theta_{2j}$ is

$$\hat{\theta}_j = \frac{\int\int_{\text{cell}_j} \widehat{dN}_1(x,y)}{\int\int_{\text{cell}_j} \widehat{dN}_2(x,y)}\, \rho$$

$$\rho = \hat{r}_2/\hat{r}_1 = \frac{\sum_{j=\text{hskpg}} \int\int_{\text{cell}_j} \widehat{dN}_2(x,y)}{\sum_{j=\text{hskpg}} \int\int_{\text{cell}_j} \widehat{dN}_1(x,y)}$$

Gene Clustering and Filtering

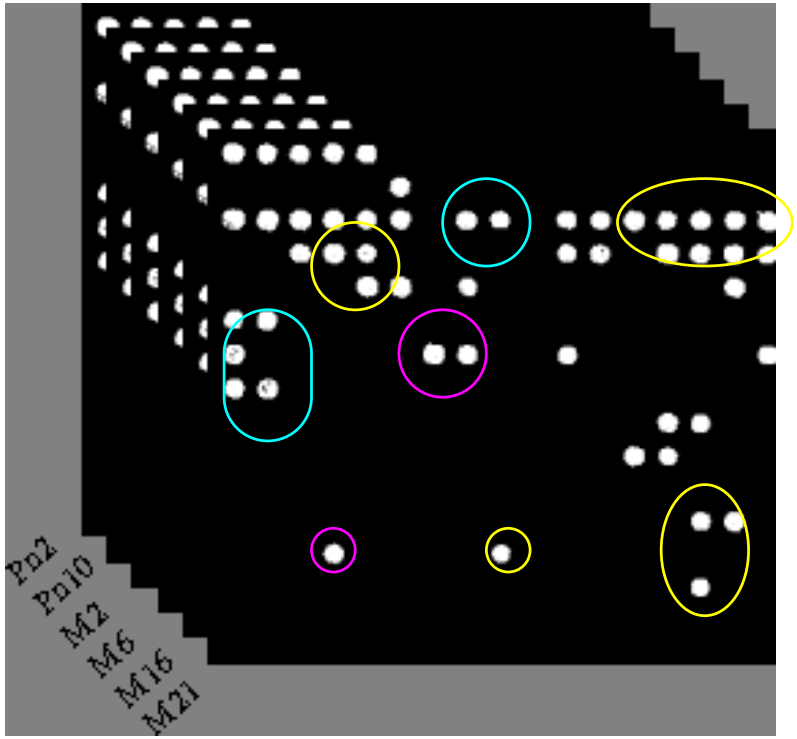


Figure 18: *Clustering on the Data Cube.*

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

Gene Trajectory Classification

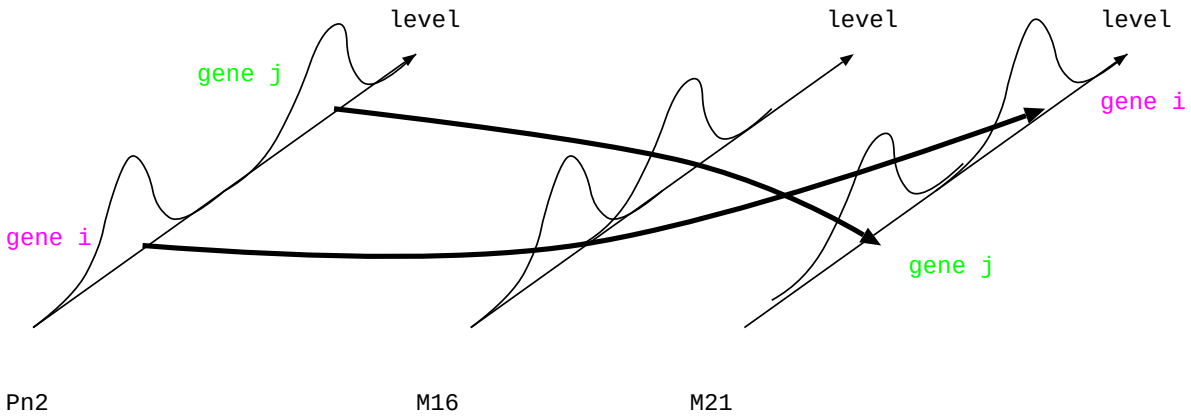


Figure 19: *Gene i is old dominant while gene j is young dominant*

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 $t = T$:

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

- Mean slope magnitude:

$$\xi_4(\theta_i) = \overline{|\Delta\theta_i(*,*)|}$$

- Mean slope dispersion:

$$\xi_5(\theta_i) = \overline{\left(|\Delta\theta_i(*,*)| - \overline{|\Delta\theta_i(\bullet,\bullet)|}\right)^2}$$

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.

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

Example: pairwise comparisons

i-th treatment generates two classes of responses X_i and Y_i:

$\{X_i(m)\}_{m=1}^{n_1}$ and $\{Y_i(m)\}_{m=1}^{n_2}$

- Pooled within-class dispersion

$$\xi_1(X_i, Y_i) = n_1 \overline{\left(X_i(*) - \overline{X_i(*)}\right)^2} + n_2 \overline{\left(Y_i(*) - \overline{Y_i(*)}\right)^2}$$

- Between class distance

$$\xi_2(X_i) = |\overline{X_i(*)} - \overline{Y_i(*)}|^2$$

Objective: Find i which achieves minimum ξ_1 and maximum ξ_2.

Pareto Optimal Fronts

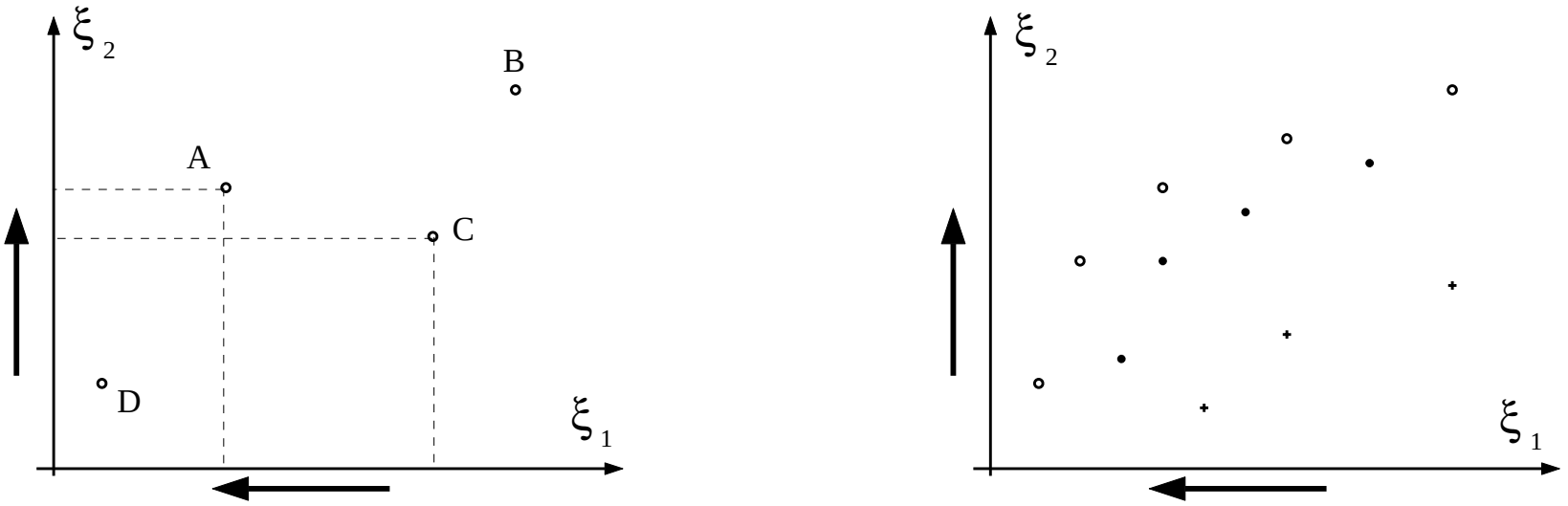


Figure 20: *a). Non-dominated property, and b). Pareto optimal fronts, in dual criteria plane.*

Pareto Gene Filtering vs. Paired T-test

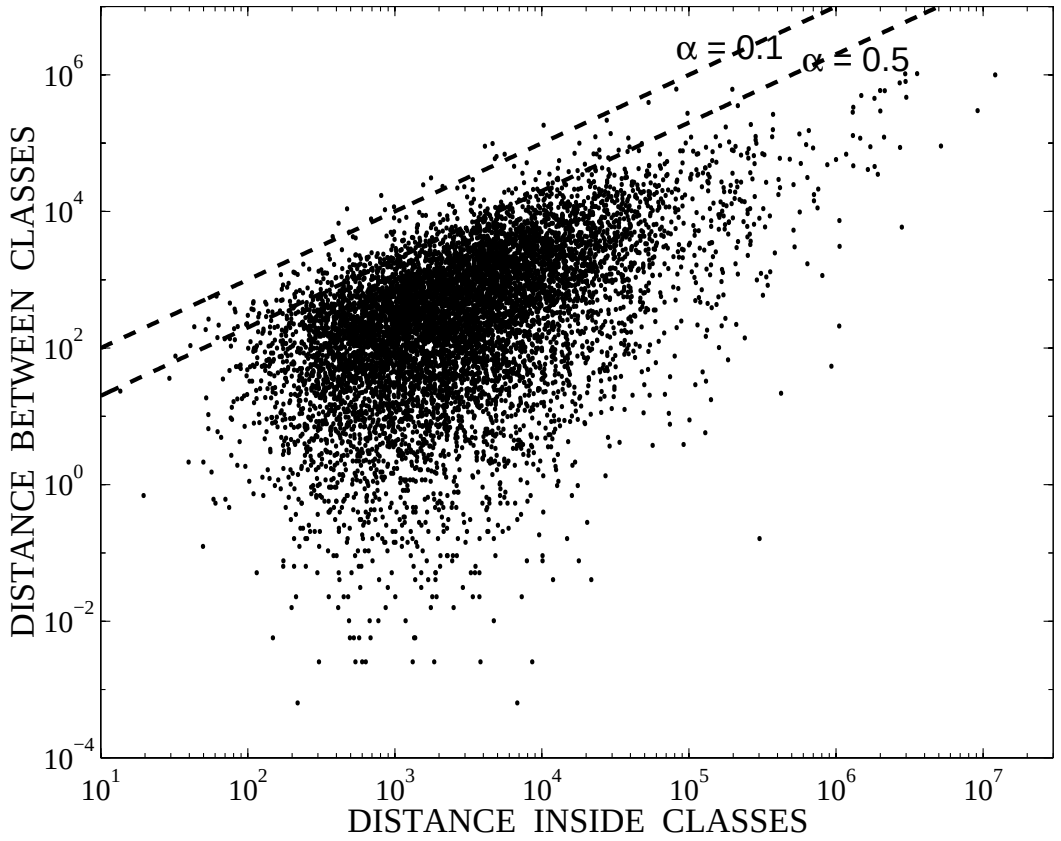


Figure 21: ξ_1 = *mean change vs* ξ_2 = *pooled standard deviation for 8826 mouse retina genes. Superimposed are T-test boundaries*

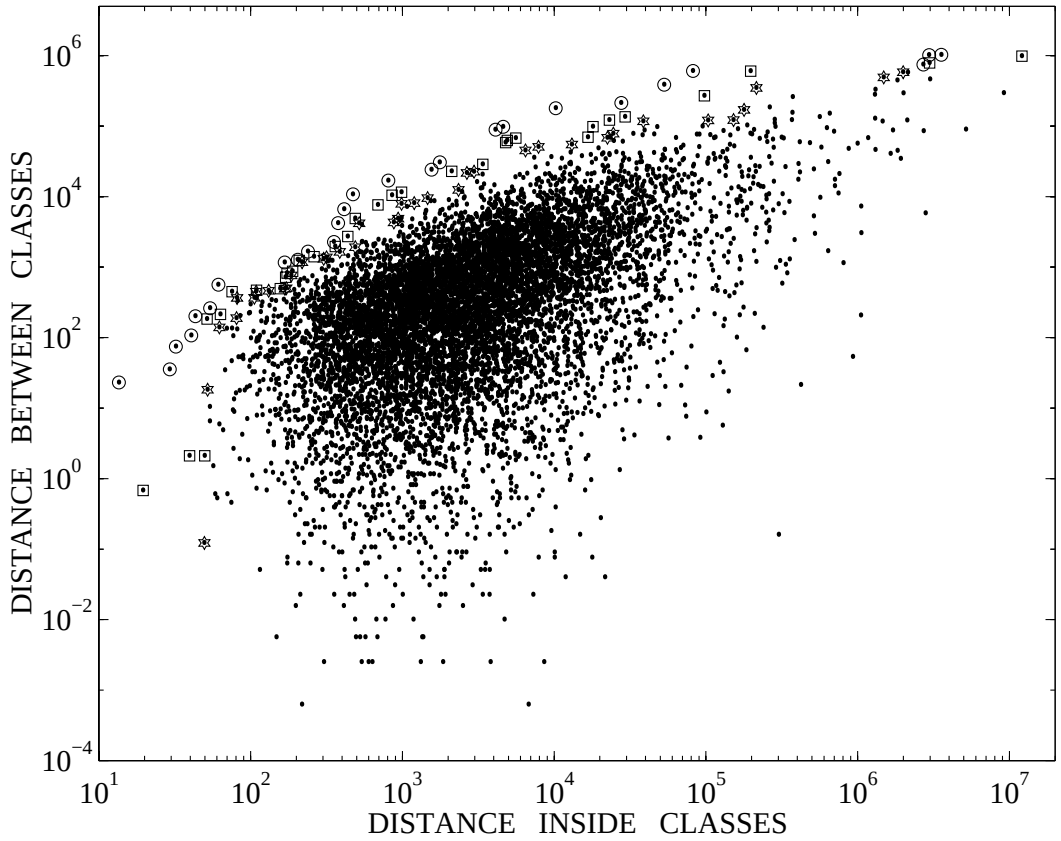


Figure 22: *First (circle) second (square) and third (hexagon) Pareto optimal fronts.*

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

Some Gene Trajectories

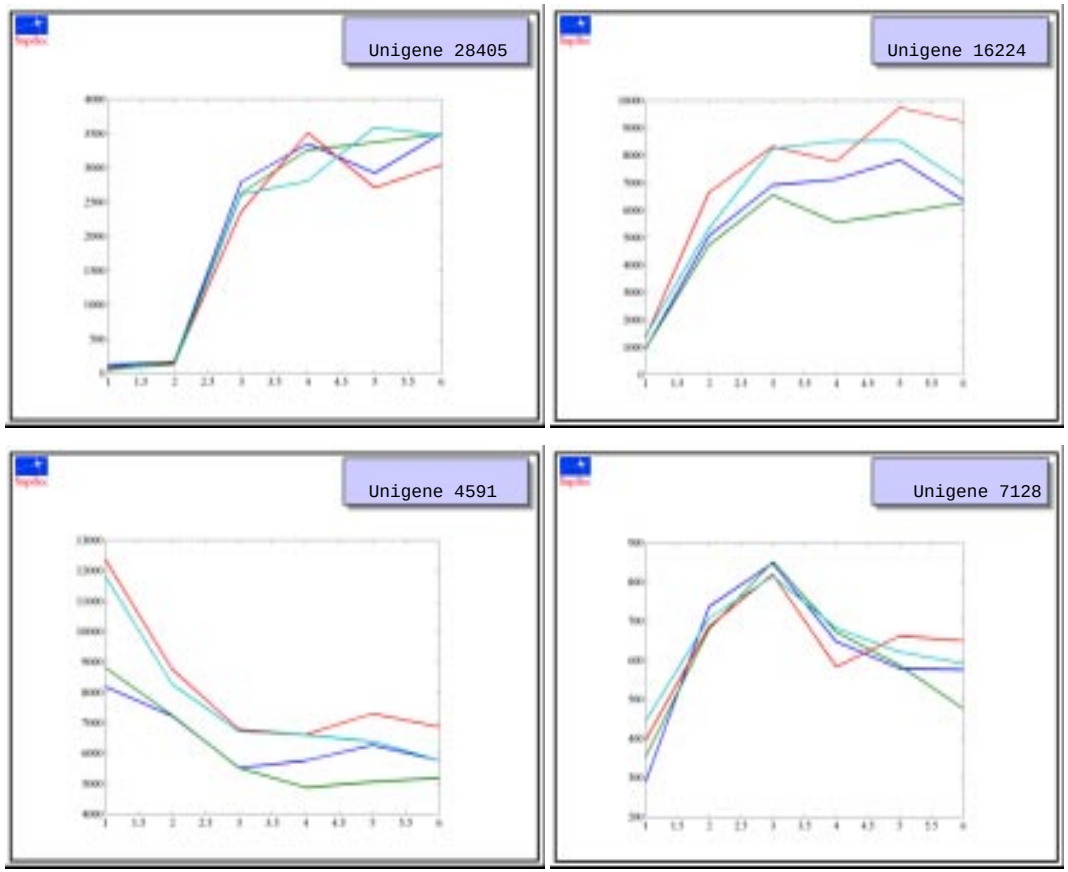


Figure 23: *Trajectories.*

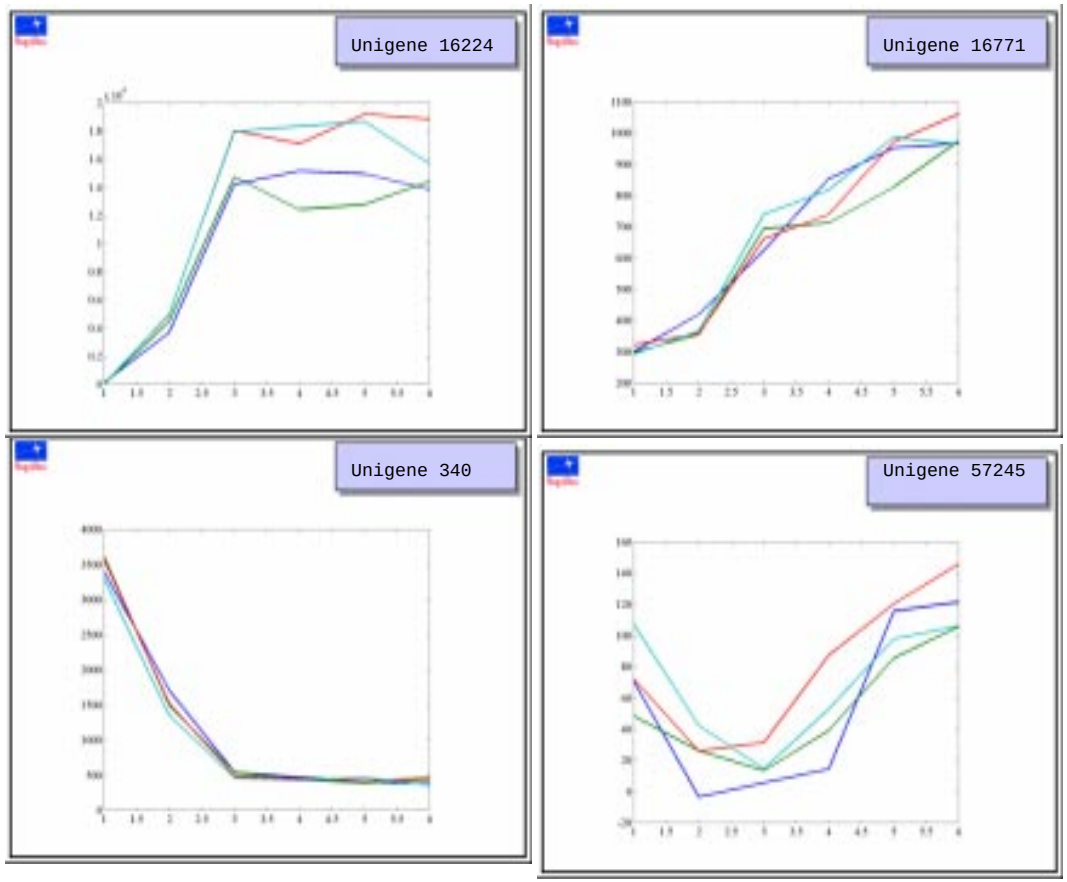


Figure 24: *Trajectories.*

Pairs of Trajectories for Replicated Segments

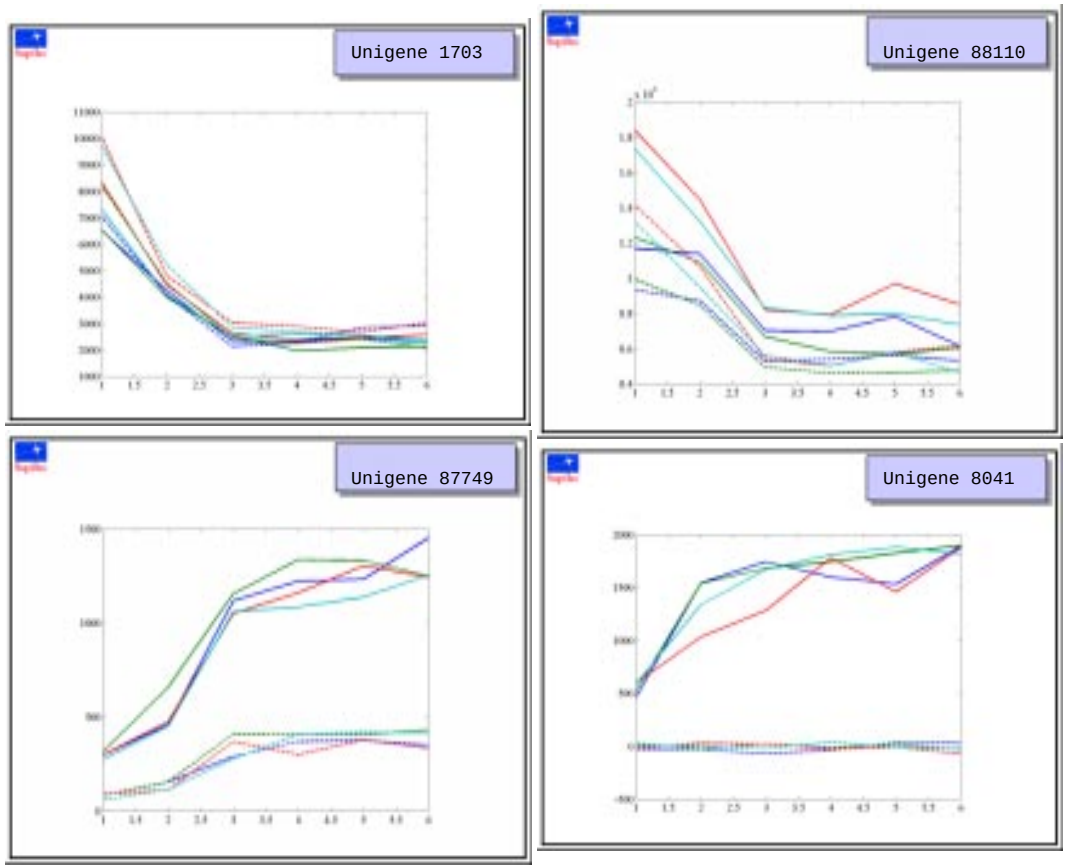


Figure 25: *Pairs of trajectories for replicated gene polynucleotide sequence.*

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

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

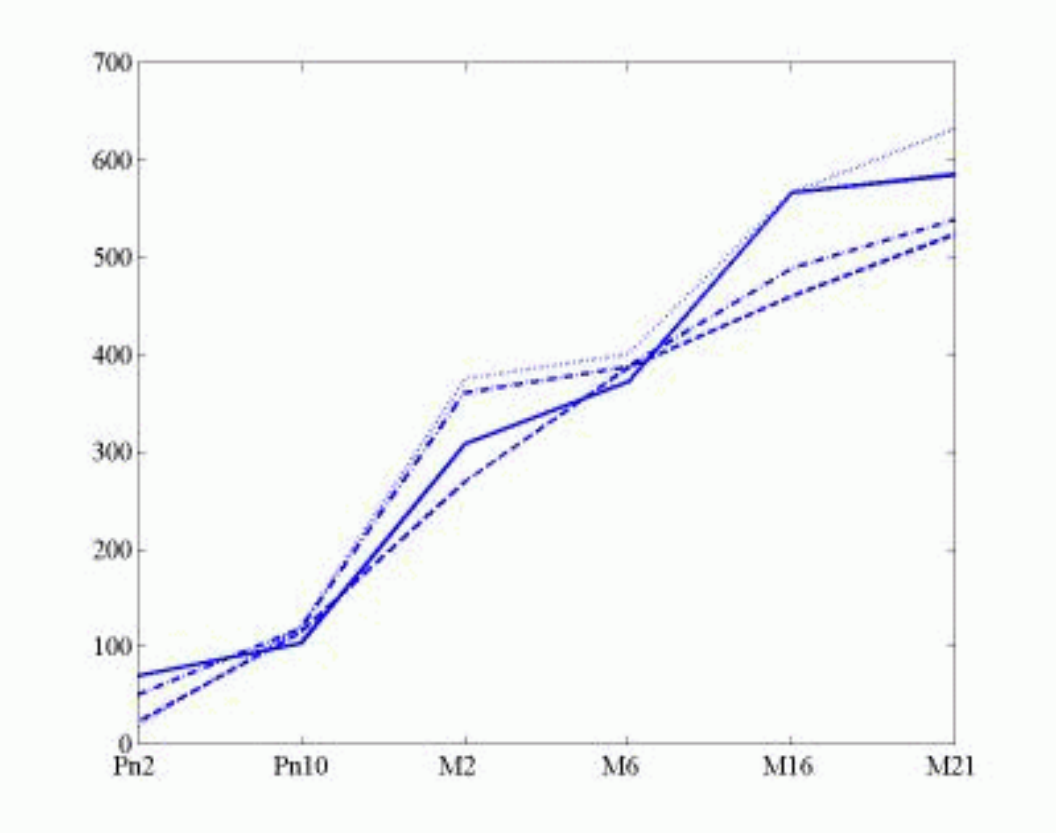


Figure 26: *4 candidate gene profiles from Mus musculus* $5'$ *end cDNA (Unigene 86632)*

Occurence Histogram

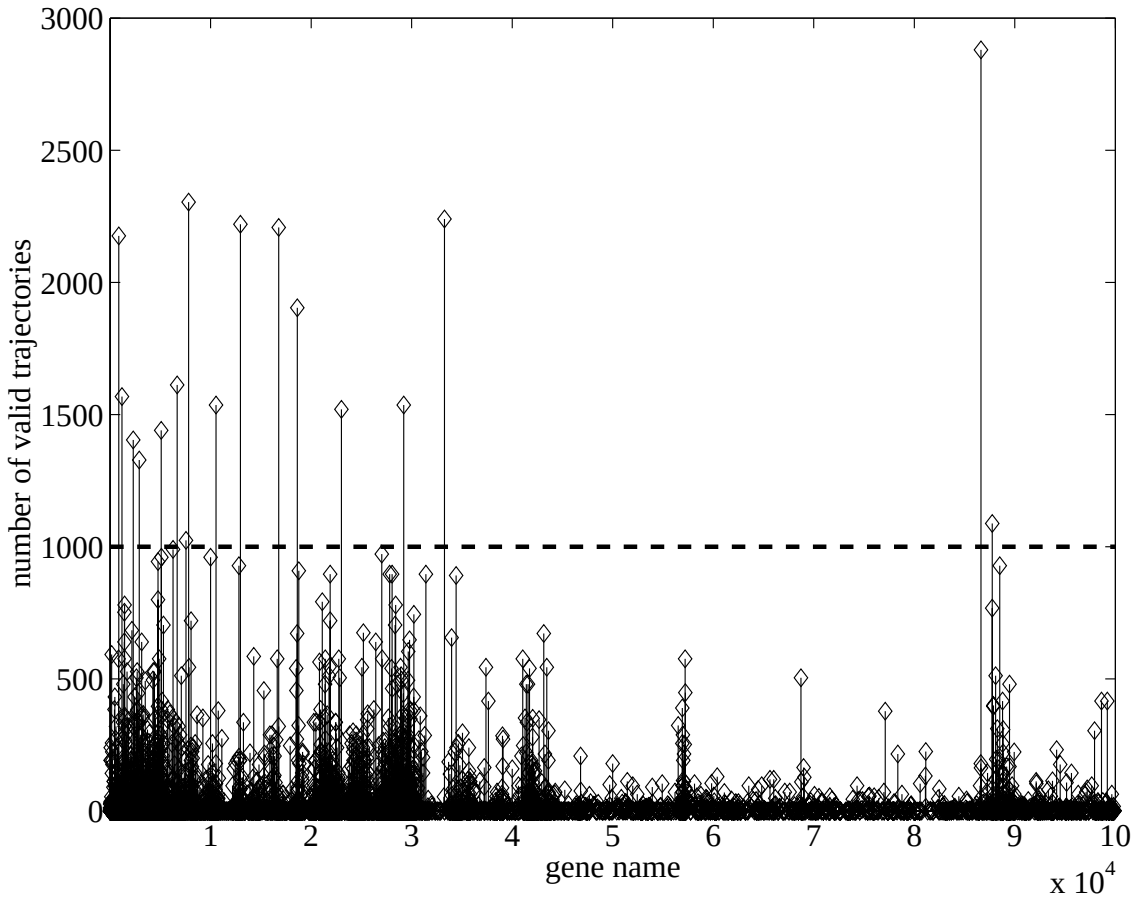


Figure 27: *Occurrence histogram with threshold.*

Old Dominant Pareto Fronts

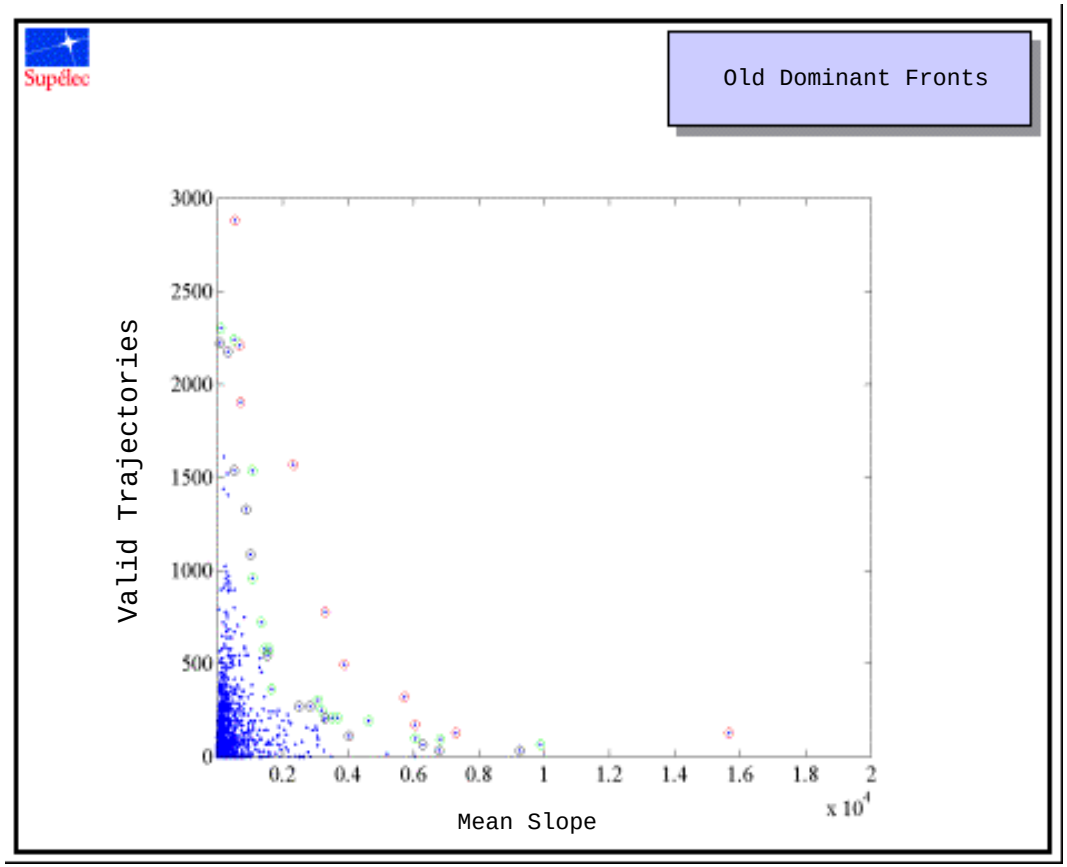


Figure 28: *Pareto fronts for old dominant genes.*

Old Dominant Genes in First Pareto Front

Unigene #	Affymetrix description
1186	Mouse Carbonic Anhydrase II cDNA
4263	Cystatin 3
16224	Guanylate cyclase activator 1a (retina)
16763	Mouse mRNA for aldolase A
16771	Mus musculus H-2K
18625	Aquaporin 1
28405	Mus musculus cDNA 3'end
42102	Mus musculus tubby like protein 1 mRNA
69061	Guanine binding protein α transducing 1
86632	Mus musculus 5'end cDNA

Table 1: *First Pareto Front gene description.*

Resistant Old Dominant Genes in first Three Fronts

- Leave-one-out cross validation

Let ψ_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 α transducing 1
86632	*Mus musculus 5'end cDNA*

Table 2: *Resistant genes remaining in first three Pareto fronts*

Young Dominant Filtering Criteria

- low 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

$$\xi_2(\psi_i) = \frac{\#\text{ trajectories having } \psi_i = [0,\ldots,0]}{4096}$$

Young Dominant Pareto Fronts

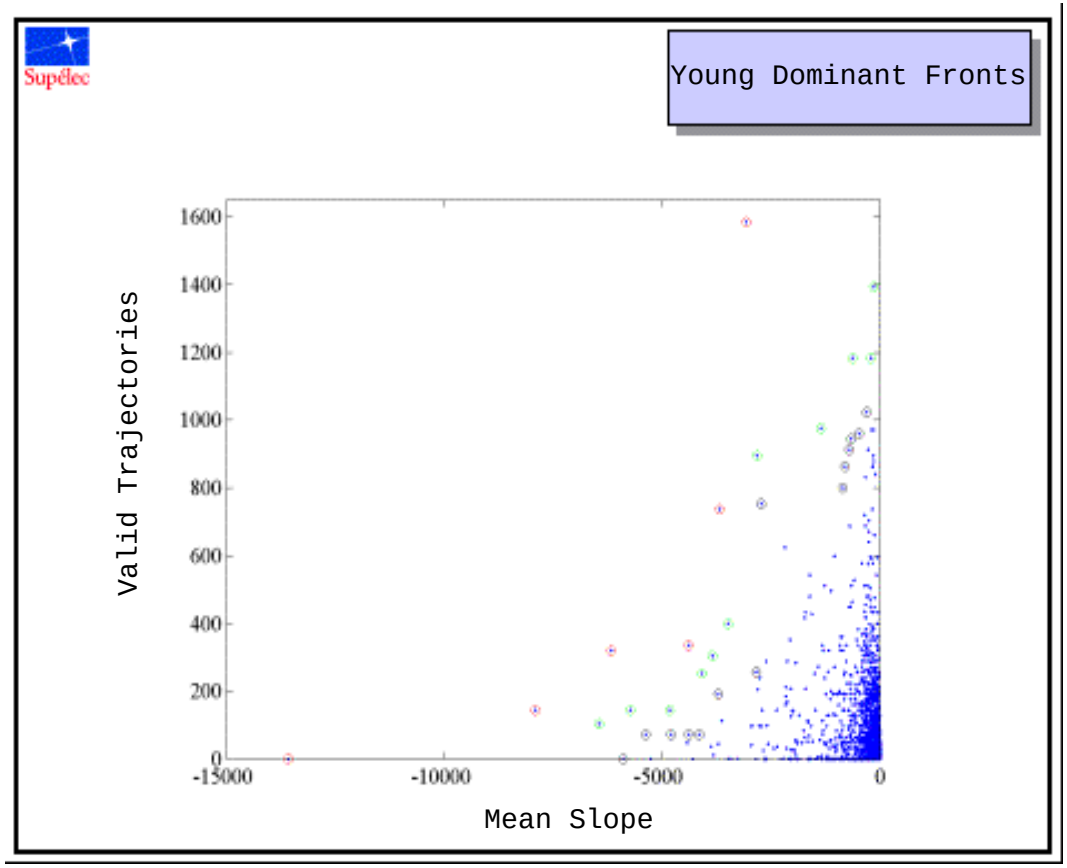


Figure 29: *Pareto fronts for young dominant genes.*

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

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

Pareto Fronts

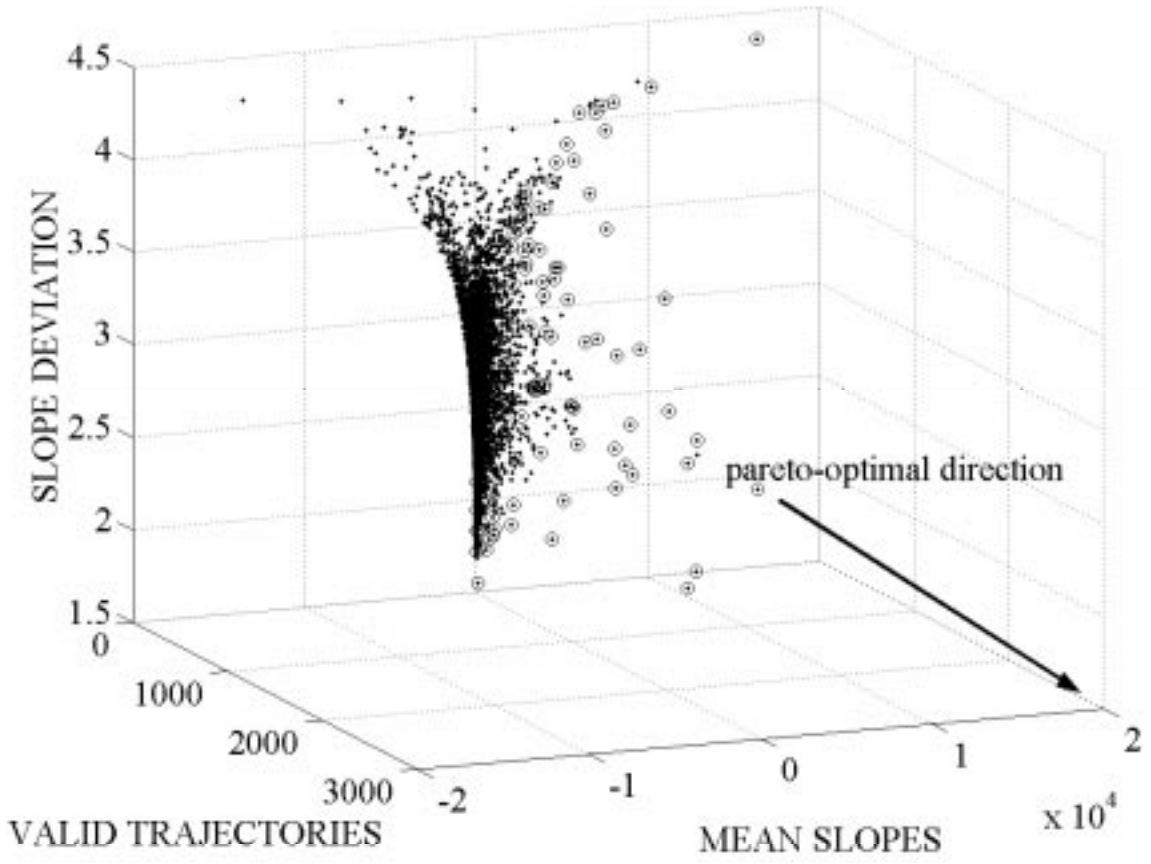


Figure 30: *First global Pareto front (o) for the three criteria (ξ_1, ξ_2 and ξ_3).*

Pairwise Pareto Fronts

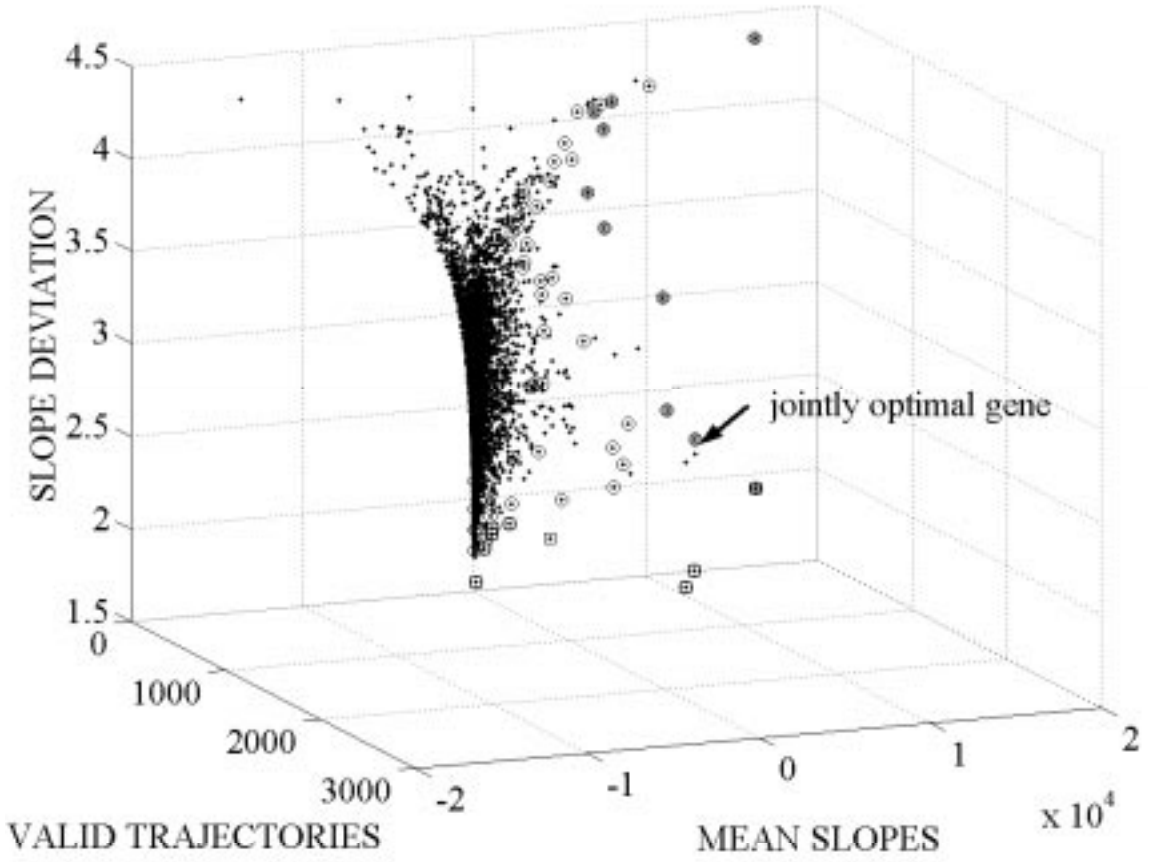


Figure 31: *First Pareto fronts for each pair of criteria taken from the set (ξ_1, ξ_2 and ξ_3). Each one of this front is denoted by squares, circles and stars, respectively.*

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 3: *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?