Data Mining For Genomics

Alfred O. Hero III

The University of Michigan, Ann Arbor, MI

ISTeC Seminar, CSU

Feb. 22, 2003

1. Biotechnology Overview
2. Gene Microarray Technology
3. Mining the genomic database
4. The post-genomic era

I. Biotechnology Overview

- **Genome:** All the DNA contained in an organism. The operating system/program for gene 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.

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?

Clone ID	GenBank	GeneName	Symbol	UniGene	LocusLink	Chr.	Molecular Function
MRA-0298	BC013125	Similar to Rhodopsin	LOC212541	Mm.2965	212541	6	
MRA-0299	unknown						
MRA-0300	NM_008938	Peripherin 2	Prph2	Mm.5032	19133	17	
MRA-0301	NM_008831	Prohibitin	Phb	Mm.2355	18673	11	
MRA-0302	bad seq						
MRA-0303	bad seq						
MRA-0304	unknown						
MRA-0305	M19381	Calmodulin 1	Calm1	Mm.34246	12313	7	calcium ion binding::
MRA-0306	M28727	Tubulin, alpha 2	Tuba2	Mm.197515	22143	2	GTP binding::
MRA-0307	BF469955	RIKEN cDNA 1110018F16 gene	1110018F16Rik	Mm.40490	68594	3	
MRA-0308	J00376	Crystallin, alpha A	Cryaa	Mm.1228	12954	17	
MRA-0309	BB284055	Expressed sequence AI597479	AI597479	Mm.28817	98404	1	
MRA-0310	NM_007378	ATP-binding cassette, sub-family A (ABC1), member 4	Abca4	Mm.3918	11304	3	ATP binding::phospholipid transporter::ATP-binding cassette (ABC) transporter::

Link to sequence

Link to NCBI database

Source: Yu, Swaroop, etal (2002)

Clone ID	Biological Process	Cellular Components	Tissue Expressed
MRA-0298			eye;adult-retina;eyeball;retina;spinal ganglion;embryonic body between diaphragm re
MRA-0299			
MRA-0300	vision::	integral membrane protein::	eye;adult-retina;nervous system;retina;eyeball
MRA-0301			embryo, whole embryo;mammary;kidney;colon;nervous system;skin, melanoma;ton
MRA-0302			
MRA-0303			
MRA-0304			
MRA-0305	cell cycle::		embryo, whole embryo;hippocampus;testis;gonad;forelimb;branchial arches;mamma
MRA-0306	microtubule-based process::microtubule-based movement::	microtubule::	mammary;embryo, whole embryo;brain;spinal cord;spinal ganglion;head;neural retin
MRA-0307			nervous system;spleen;cortex;muscle;t cell;head;hippocampus;spinal cord;basal gangl
MRA-0308	sensory organ development::	cytoplasm::	eyeball;head;embryo, whole embryo;neural retina;eye;adult;spleen
MRA-0309			heart;liver;head;amygdala;mammary gland;pancreas;mammary;urinary bladder;embry
MRA-0310	vision::transport::phospholipid transfer to membrane::	integral plasma membrane protein::	eye;heart;brain;head;adult-retina;pineal-glands;embryonic body between diaphragm

Source: Yu, Swaroop, etal (2002)

Genome Sequencing Status

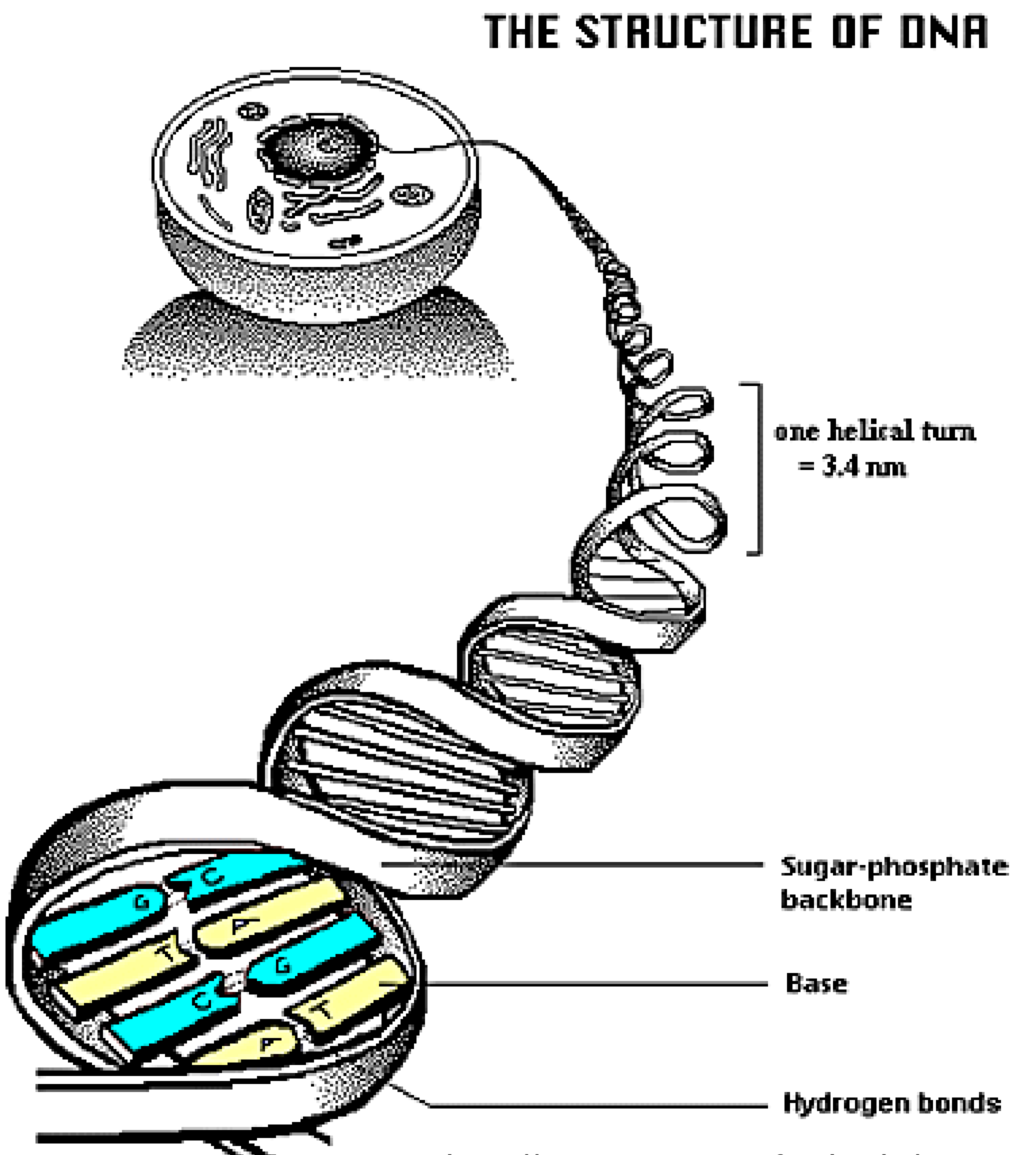
- Whole genome has been sequenced for over 1000 viruses and over 100 microbes
- Plant and animal genomes sequenced
  - Oat,soybean,barley,rice,wheat,corn
  - Mouse,zebrafish,human
- Plant and animal genomes in progress
  - Cotton,tomato,potato…
  - Rabbit,dog,chicken…

Source: NCBI Entrez-Genome, http://www.ncbi.nlm.nih.gov:80/entrez/

Sequencing Milestones

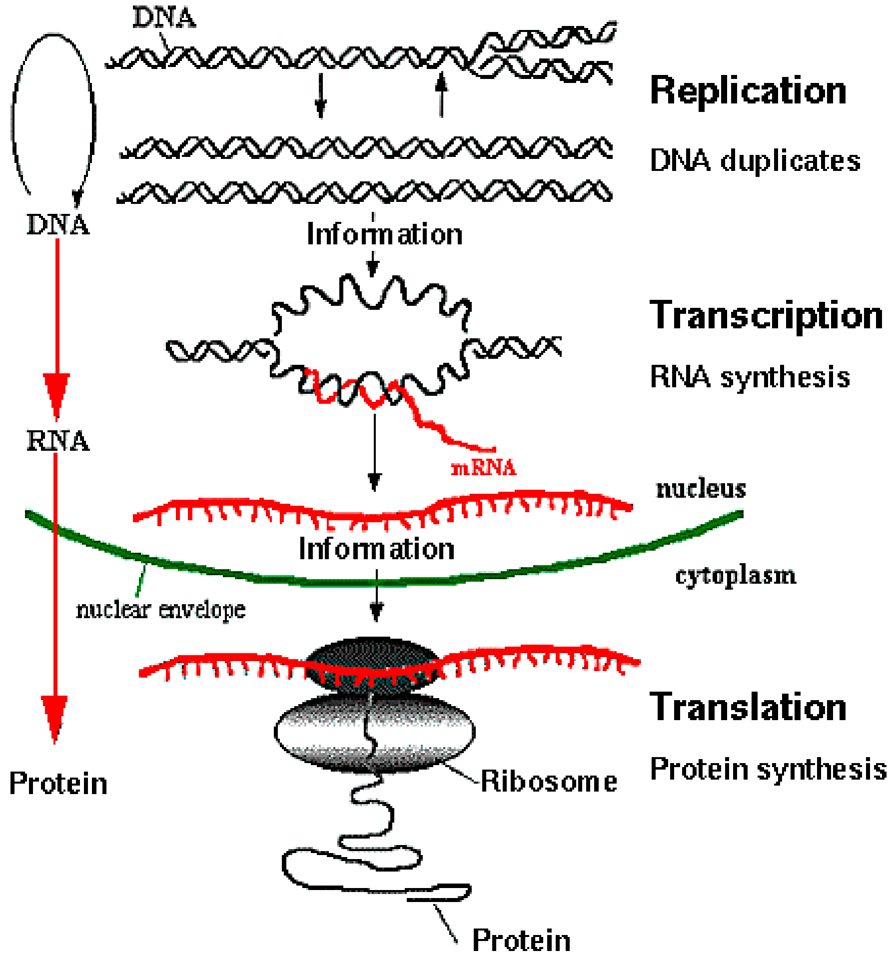
Organism	#of genes	% genes with inferred function	sequencing complete
E. Coli	4,288	60	1997
Yeast	6,600	40	1996
C. Elegans	19,000	40	1998
Drosophila	12,000-14,000	25	1999
Arabidopsis	25,000	40	2000
Mouse	26,000-40,000	10-20	2002
Human	26,383-39,114	10-20	2001

Source: http://www.biotech.ucdavis.edu/powerpoint/powerpoint.htm



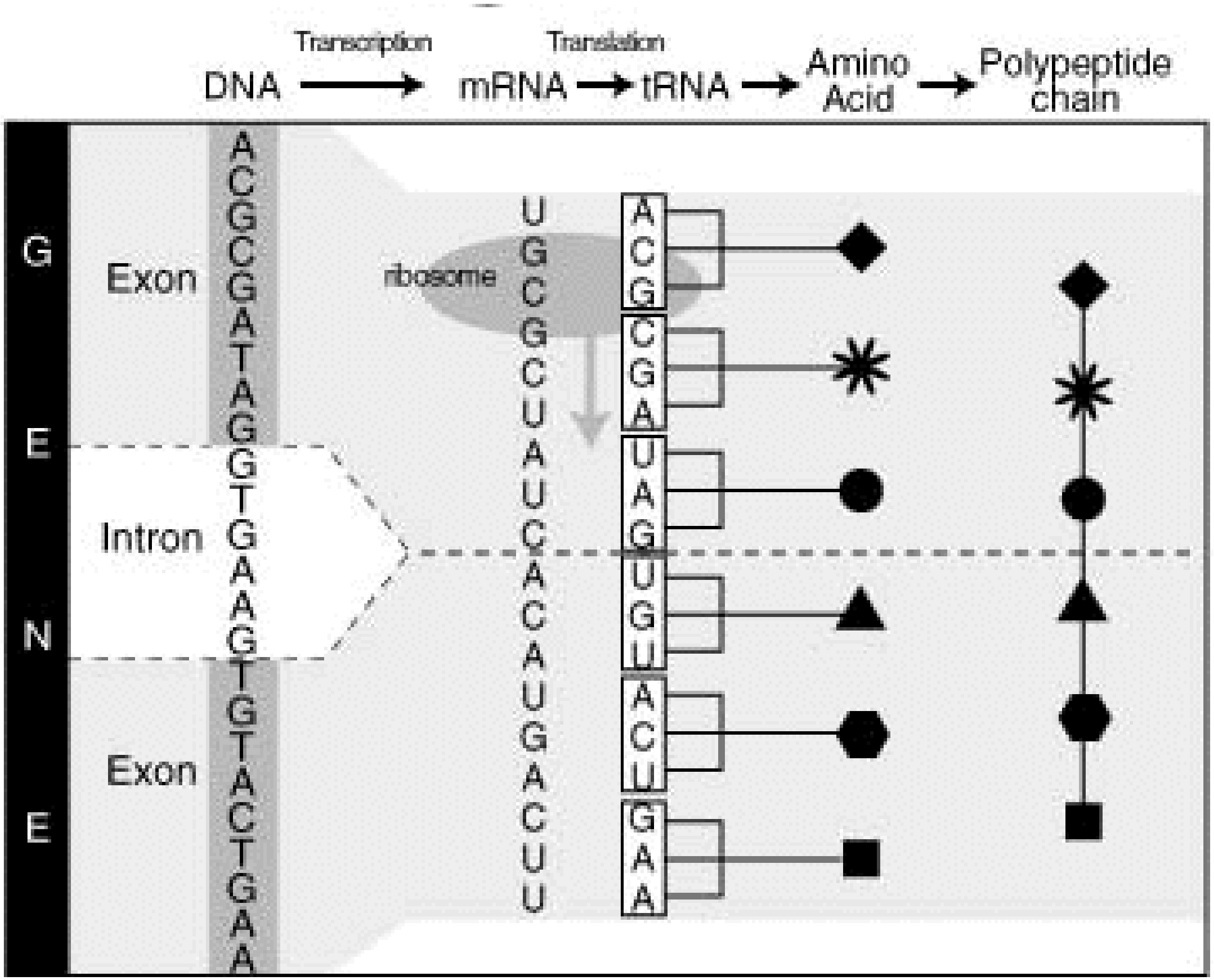
http://www-stat.stanford.edu/~susan/courses/s166/node2.html

Central Dogma of Molecular Biology



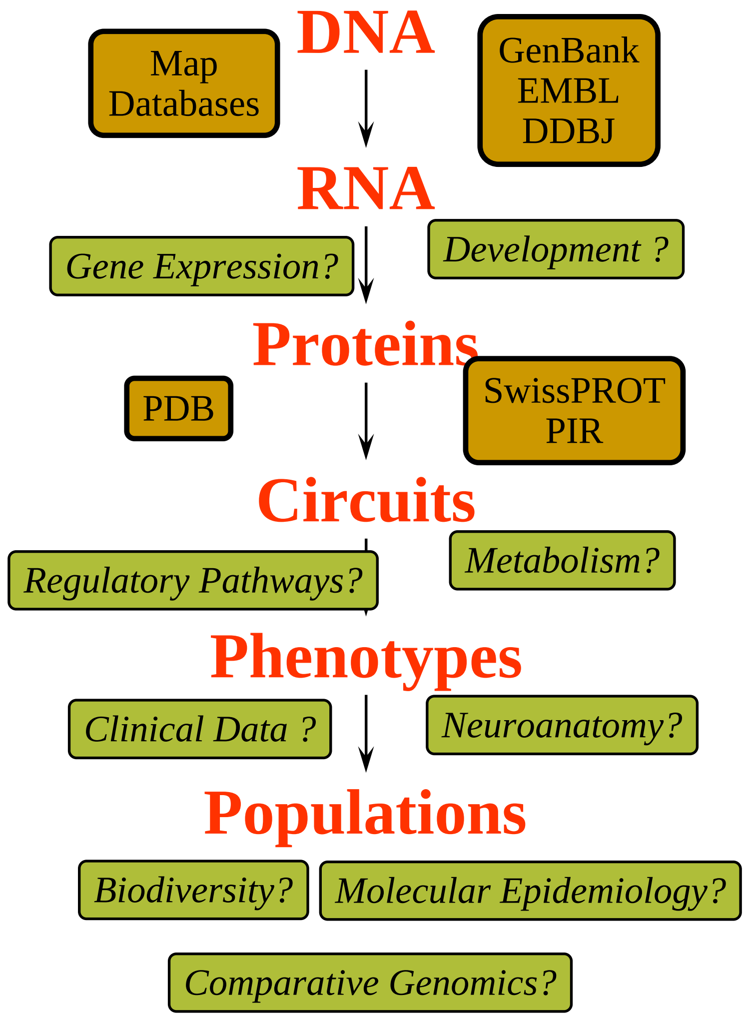
http://anx12.bio.uci.edu/~hudel/bs99a/lecture20

Central Dogma: From Gene to Protein

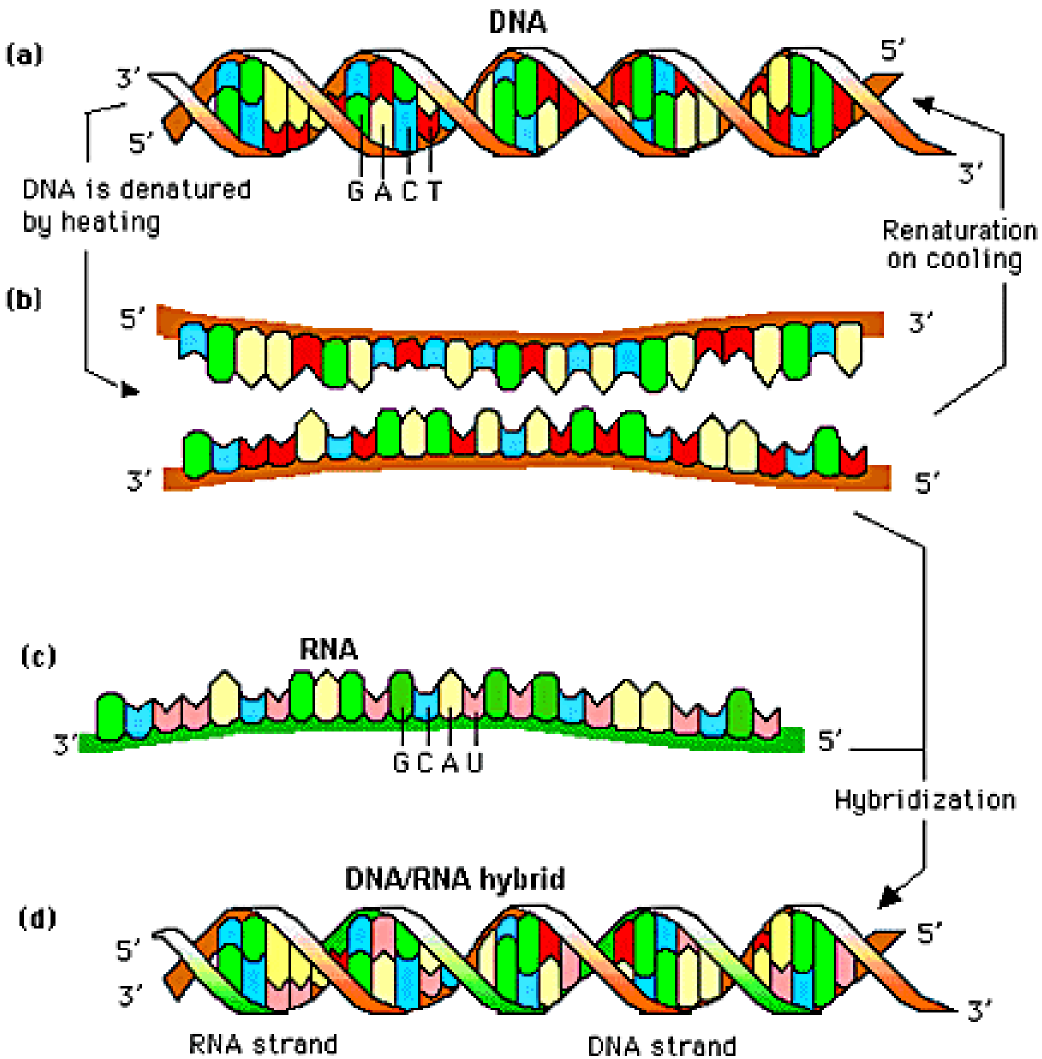


Source: NHGRI http://www.genome.gov/

Towards a unified theory….



Source: http://www.biotech.ucdavis.edu/powerpoint/powerpoint.htm



Nucleic Acid Hybridization

II. Gene Microarray Technologies

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

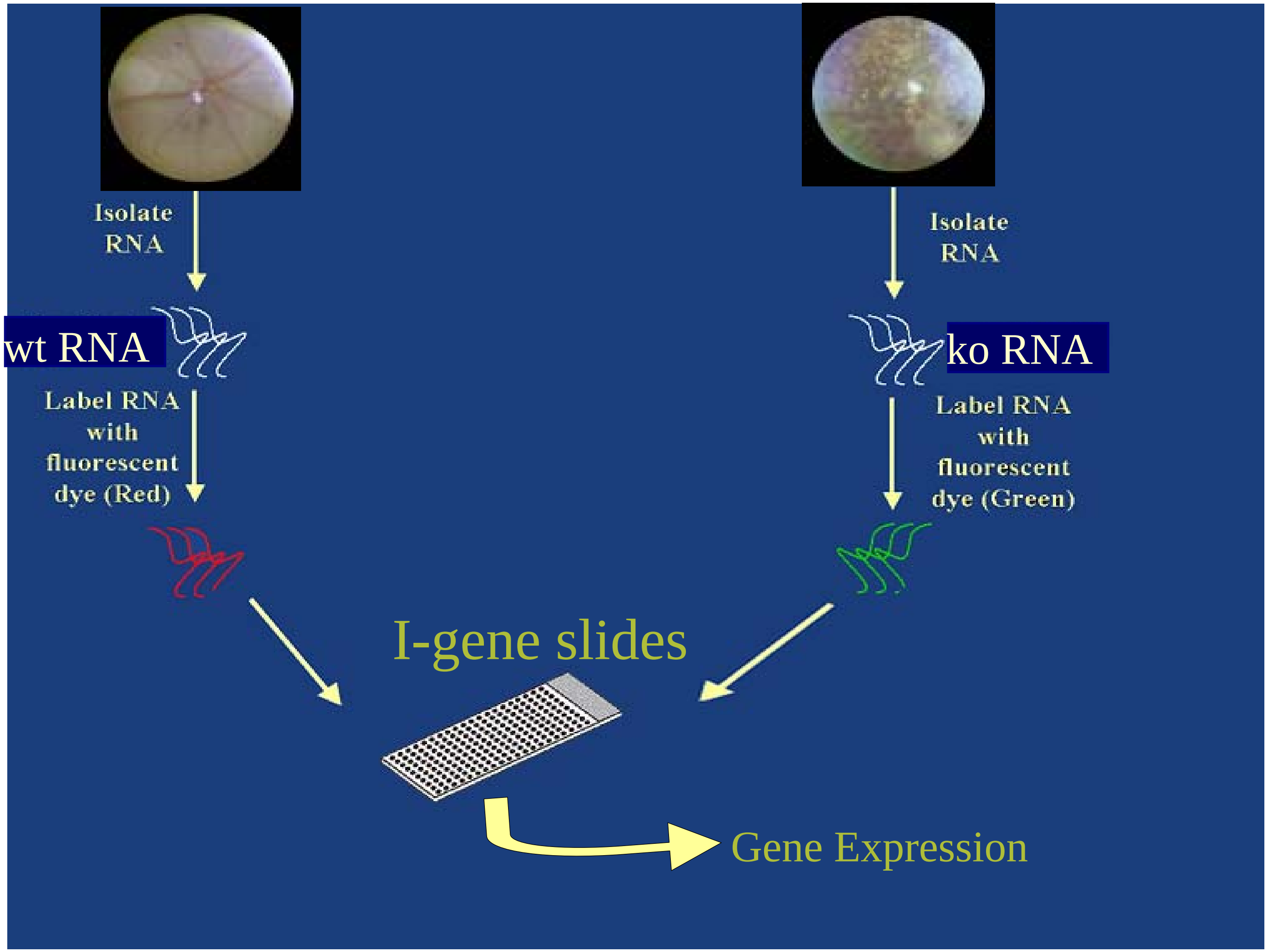
  1) Affymetrix GeneChip
  2) **cDNA spotted arrays**

- 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)

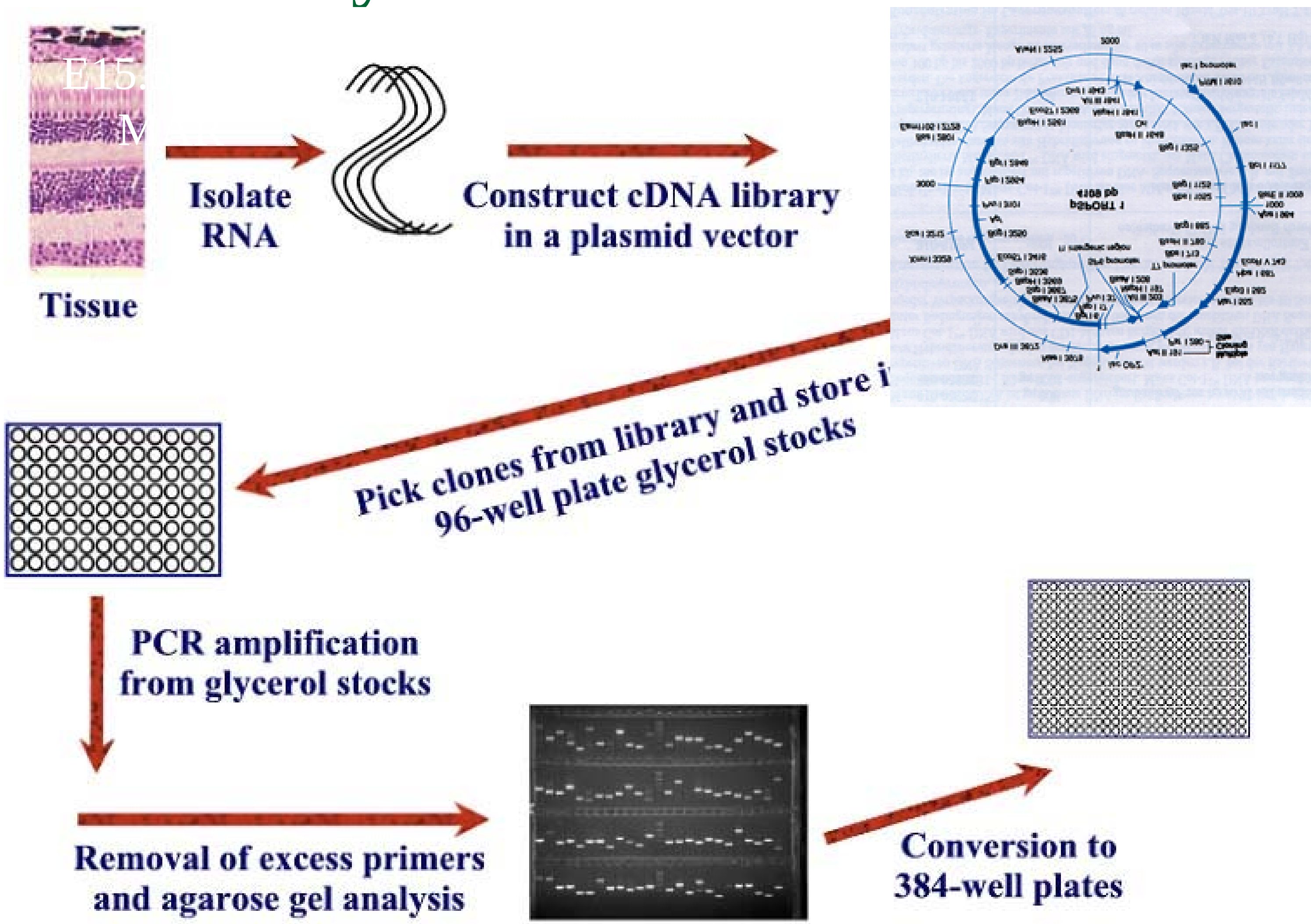
- Result: 10,000-50,000 genes can be probed at once

Specialized cDNA Array: Eye-Gene



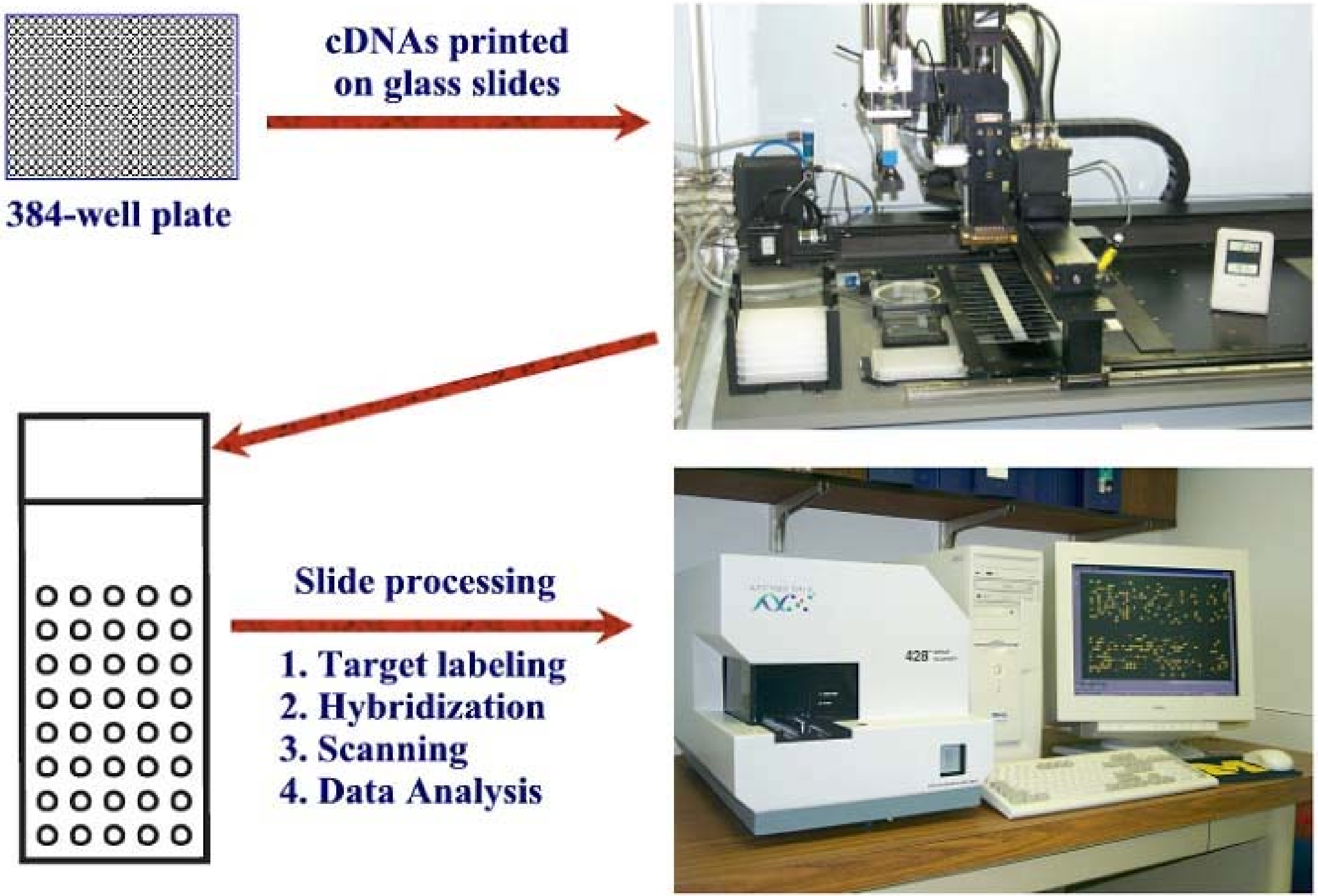
Source: J. Yu, UM BioMedEng Thesis Proposal (2002)

I-Gene Array: Probe Generation



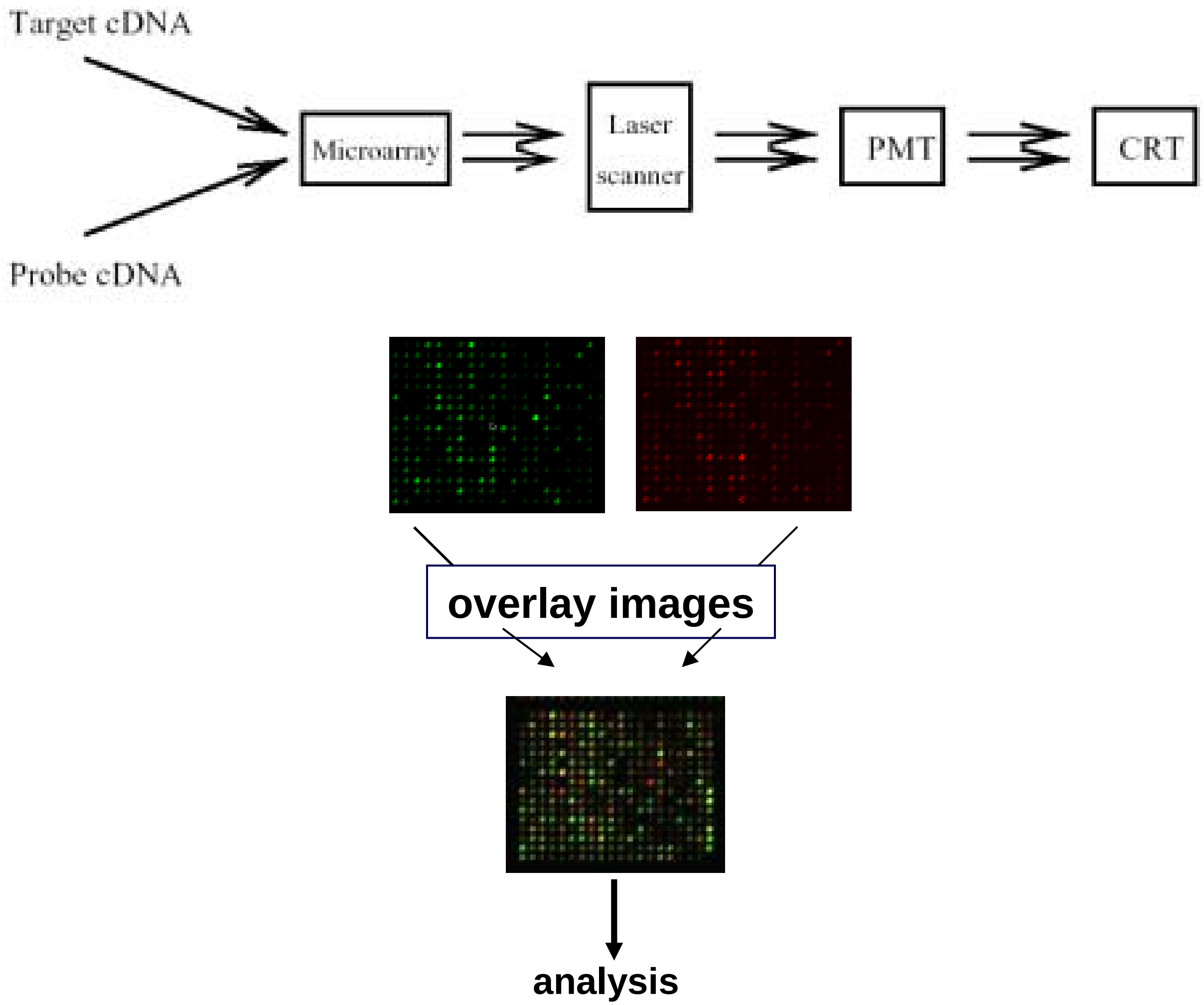
Farjo, R & Yu, J. Vision Research 42 (2002)

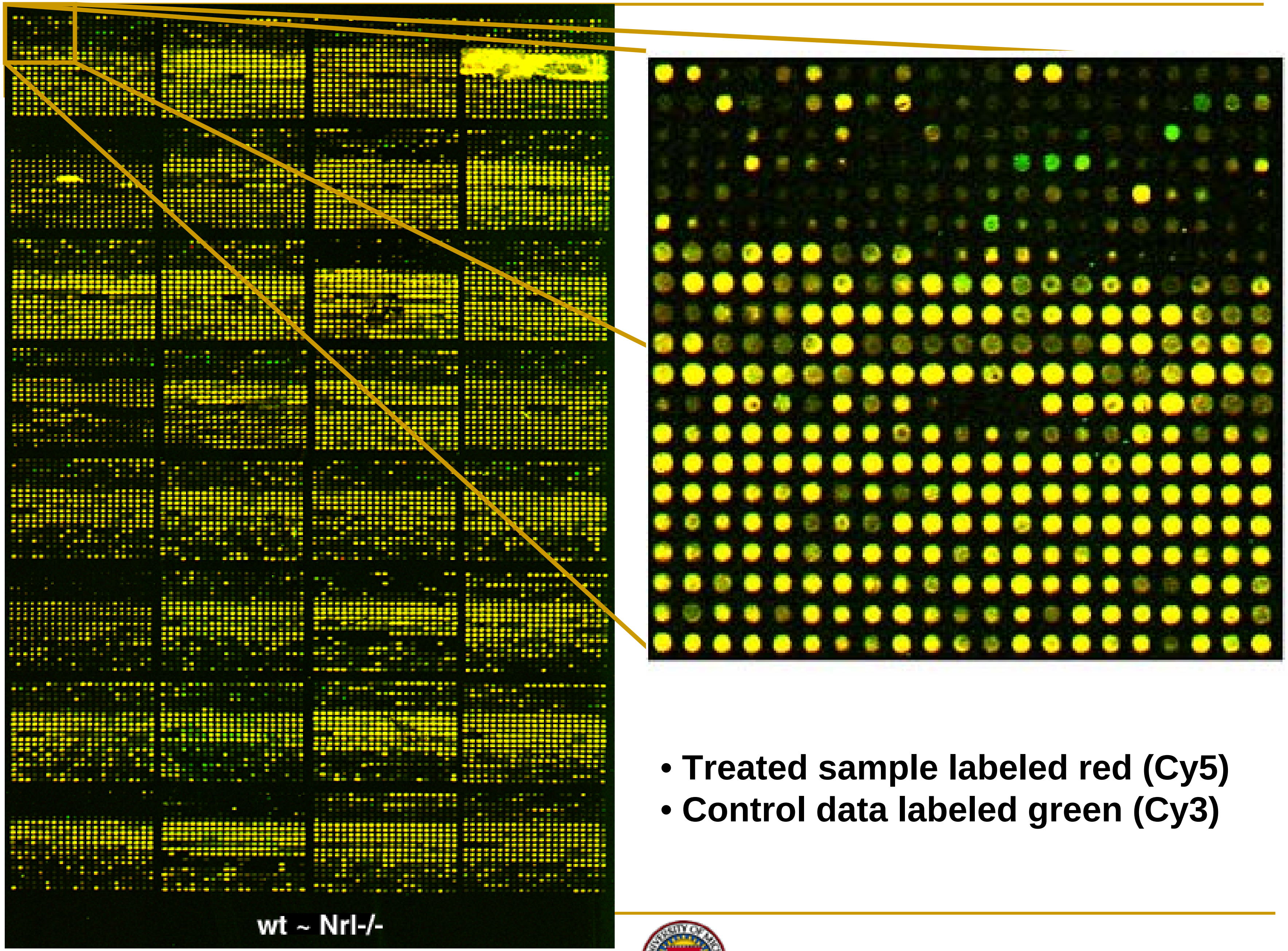
I-Gene Array: Printing and Processing



Farjo, R & Yu, J. Vision Research 42 (2002)

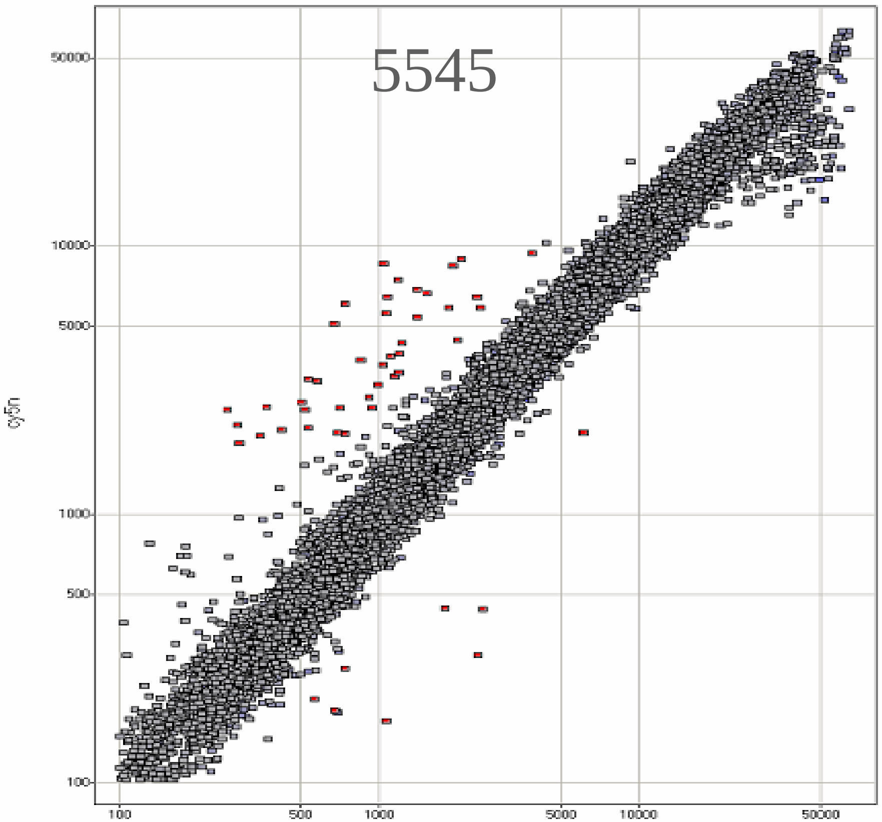
I-Gene Array: Image Formation

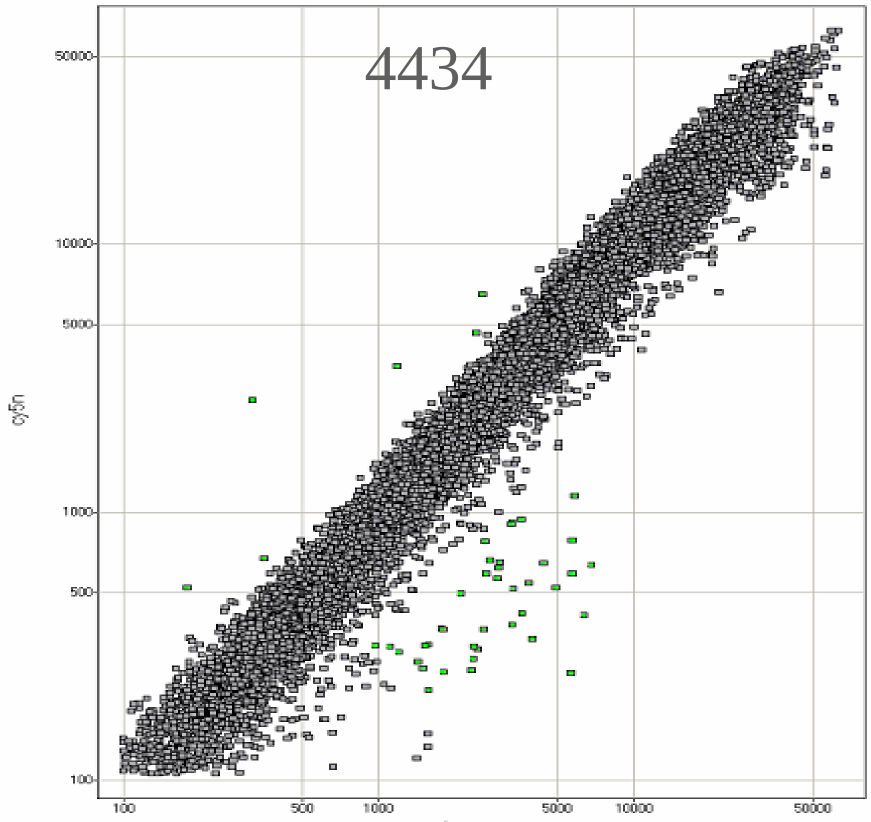




- Treated sample labeled red (Cy5)
- Control data labeled green (Cy3)

Single-Chip Raw Data Analysis





5545up	60	76	77	941	1307	1318	2188	2594	2982	3413	3504	4636	4711	5083	5100	5101	5124	5485	5486	5487
	5488	5509	5518	5519	5605	5961	6025	6372	6750	6766	7631	8008	8039	8434	10103	10943	10988	12295	12651	13104
5545dn	1083	3806	4651	4657	7567	7574	7964	10538	11042											
4434dn	40	60	76	77	898	907	941	1307	1318	2594	4636	4711	5083	5124	5485	5488	5509	5518	5519	5521
	5605	6372	6750	6766	7639	7983	8008	8039	10841	10843	10844	10862	10864	10943	10988	11443	11874	12060	12651	
4434up	3806	7574	7603	7964	10538	11042														

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

Problem: Experimental Variability

- **Population** – too 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!

III. Mining Statistical Genomic Data. Questions:

- How to estimate true Cy5 and Cy3 from raw data?
- How to compensate for experimental variability?
- How to extract expression profile ratios from a set of up to 50,000 probe responses?
- How to specify gene profile selection criteria for mining in this data?
- How to discover complex genetic pathways to disease, aging, etc?

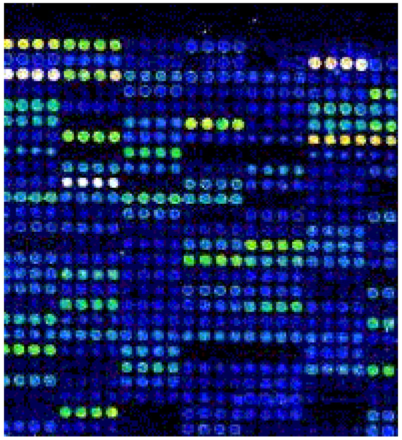
Mining Statistical Genomic Data. Answers:

- Spot Extraction: Cy5/Cy3 or Cy5-Cy3?
  - Image processing, image segmentation, non-linear anova models
- Comparing between microarray experiments
  - Statistical invariance, equalizing transformations, normalization
- Gene filtering and screening
  - Simultaneous statistical inference, T-tests, FDR
- Discovery of genetic pathways
  - Clustering, dependency graphs, HMM's

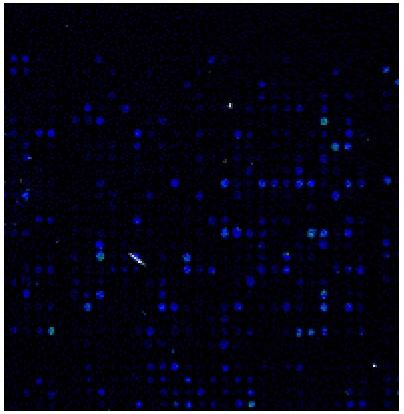
Spot Extraction Issues

- Technical noise and variability
- Laser gain and calibration
- Cy3/cy5 channel bleedthrough
- Image formation gain
- Spot-gridding algorithm
- Spot segmentation algorithm

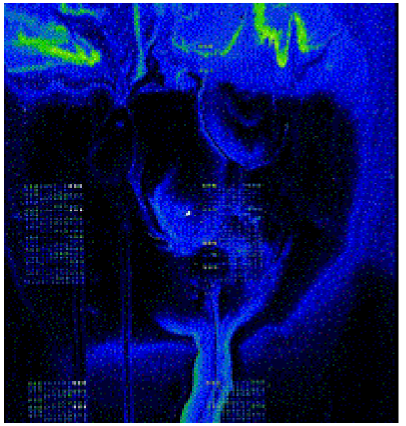
Technical Noise and Variability



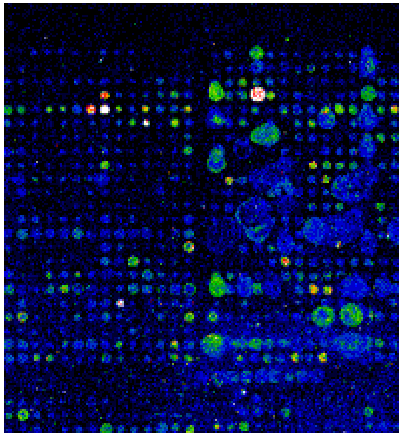
Good Signal



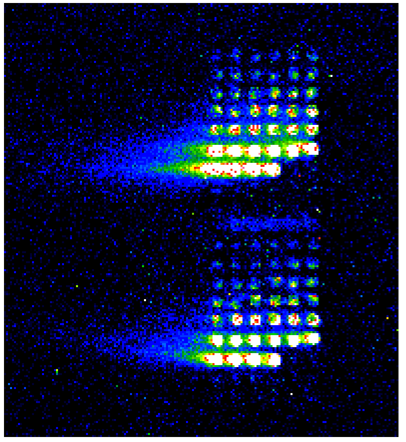
Weak Signal



Streaks



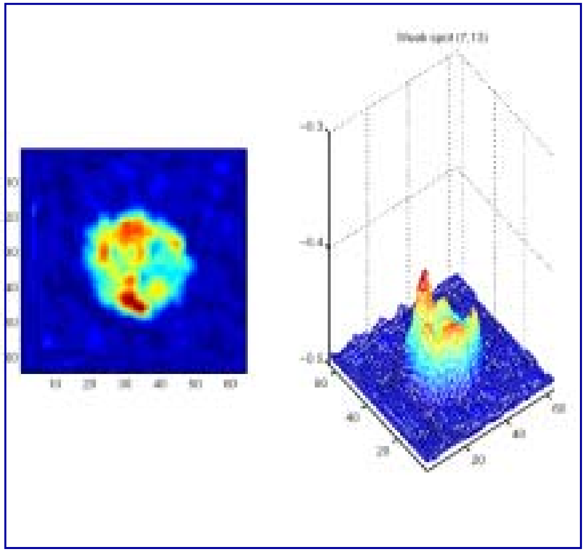
Irregular Spots

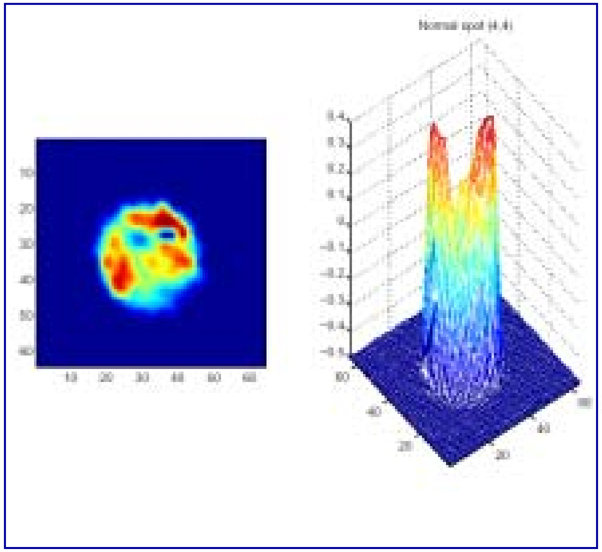


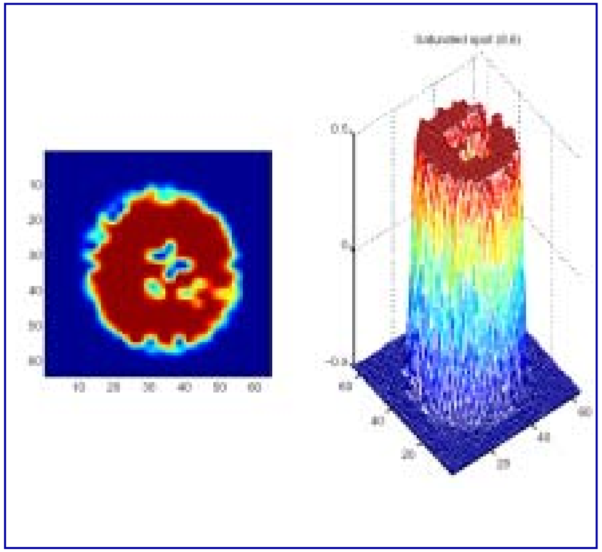
Comet Tails

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

Gain Effects







Weak

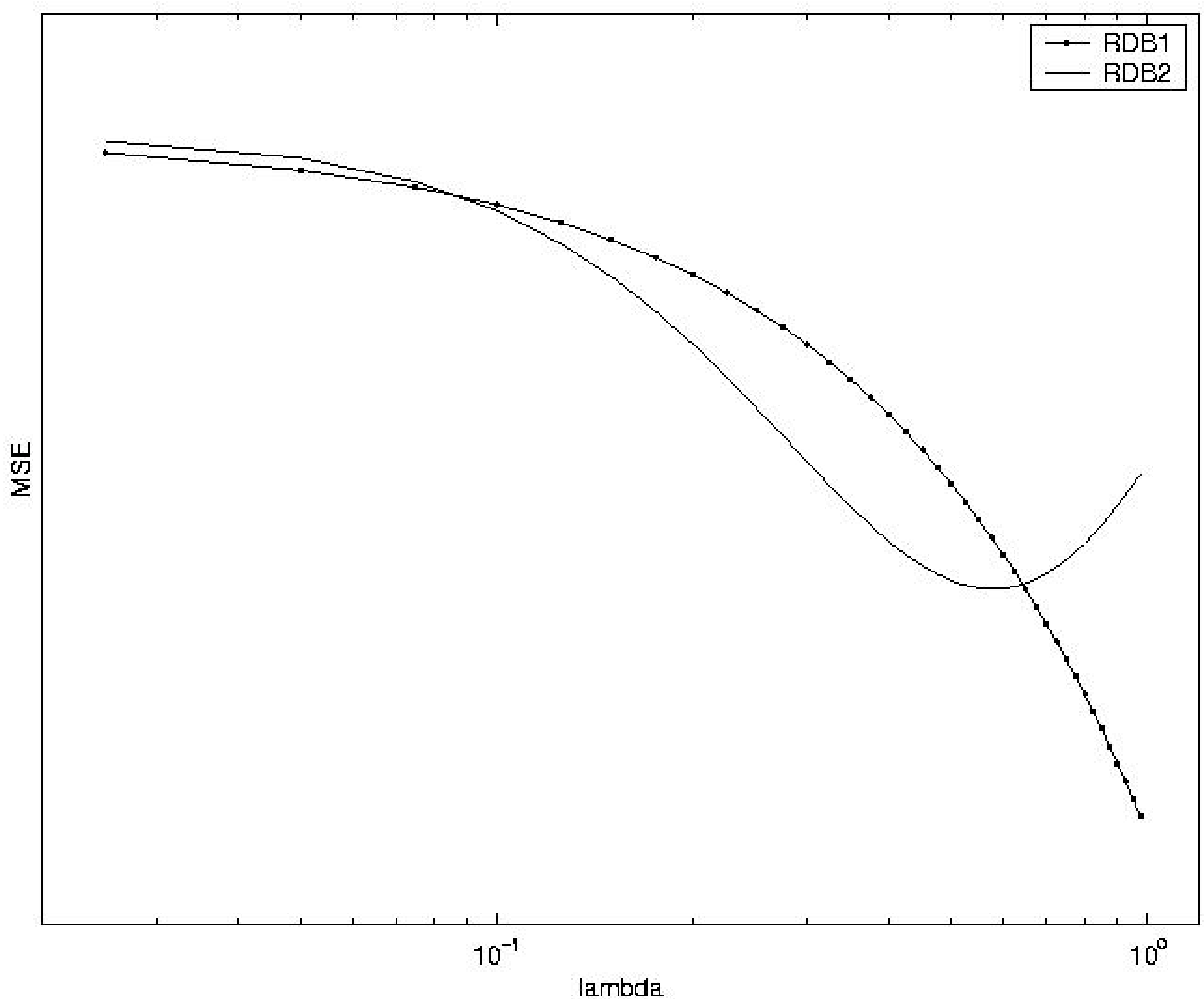
Normal

Saturated

Optimal gain can be studied by information theory

Rate Distortion Lower Bound

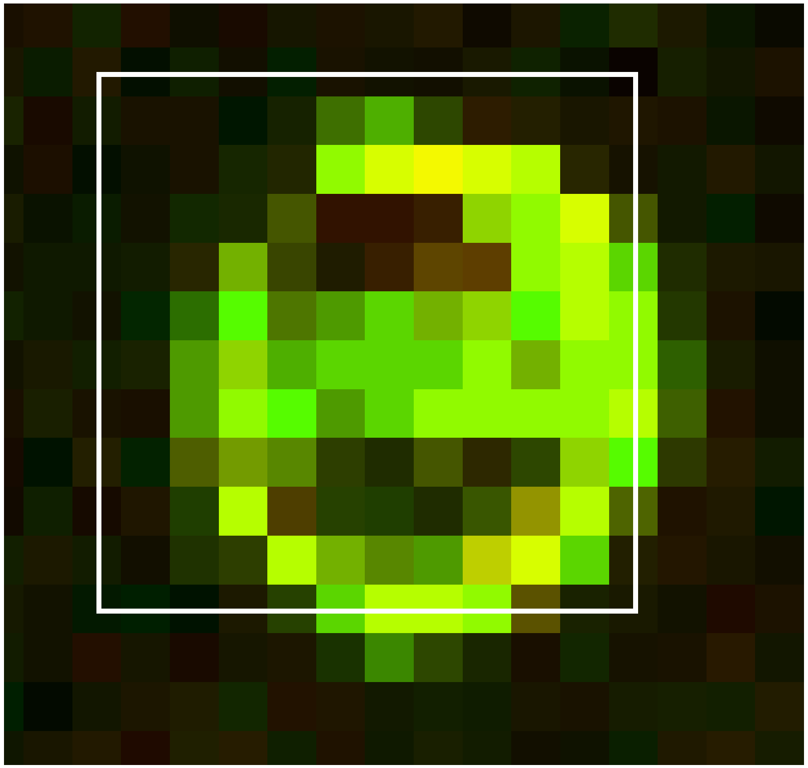
MSE



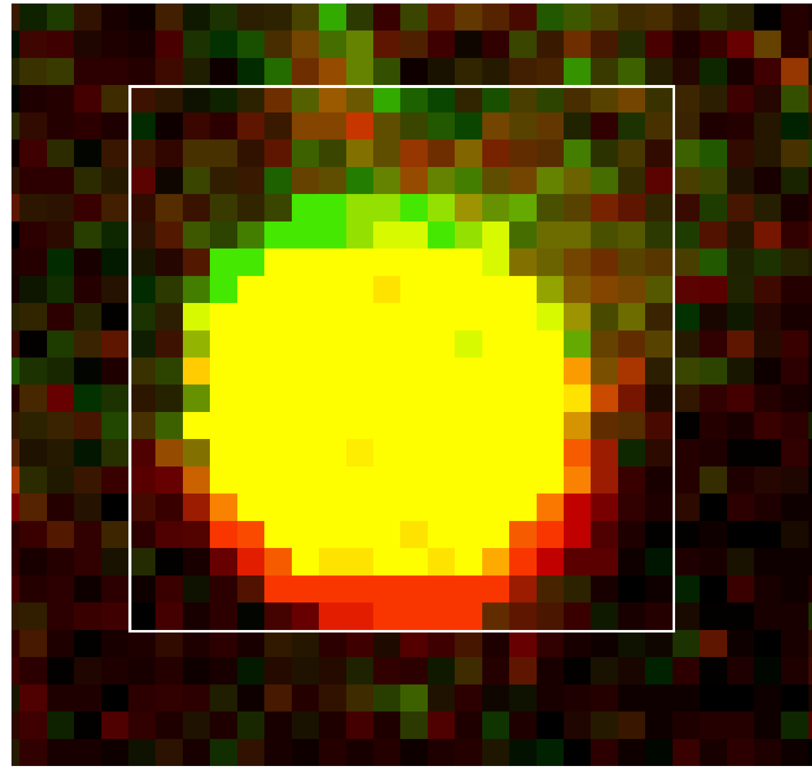
Gain

Hero Springer-03

Spot Segmentation Failure Modes



Grid misalignment



Laser Misalignment

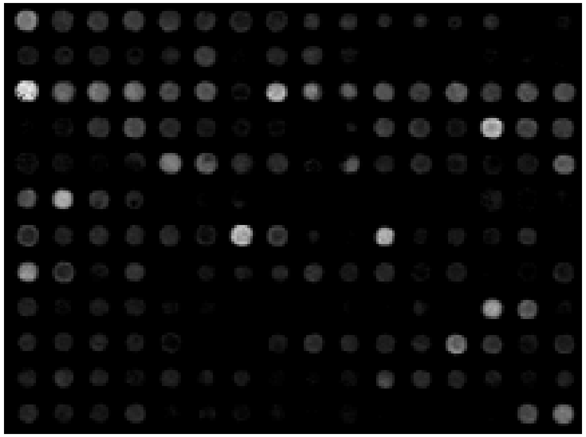
Source: C. Ball, Stanford Microarray Database

Steps in Conventional Segmentation

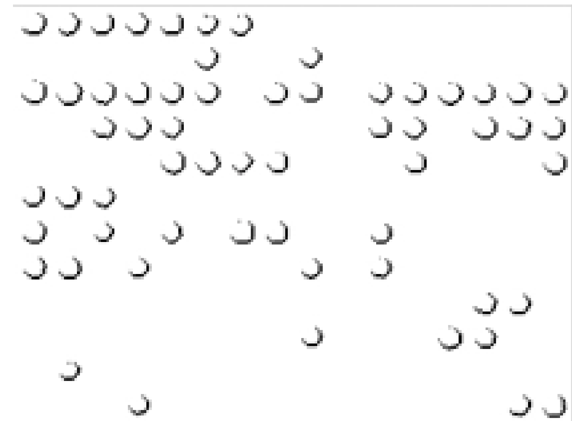
- **Addressing** – Locate "center of description" for each spot
- **Spot Segmentation** – Classification of pixels either as signal or background.
- **Spot Quantification** – Estimation of hybridization level/ratio of spot

Mathematical morphology unifies these steps

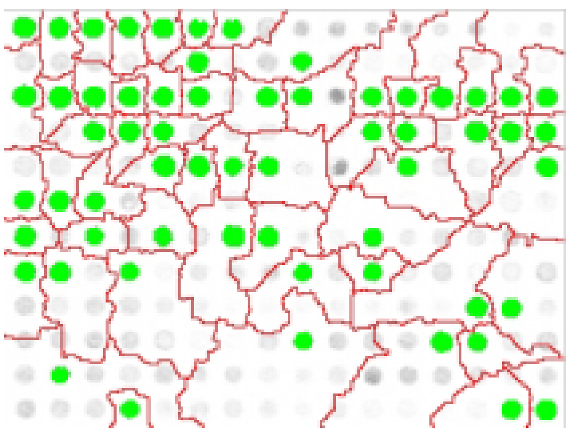
Segmentation via Morphological Operators



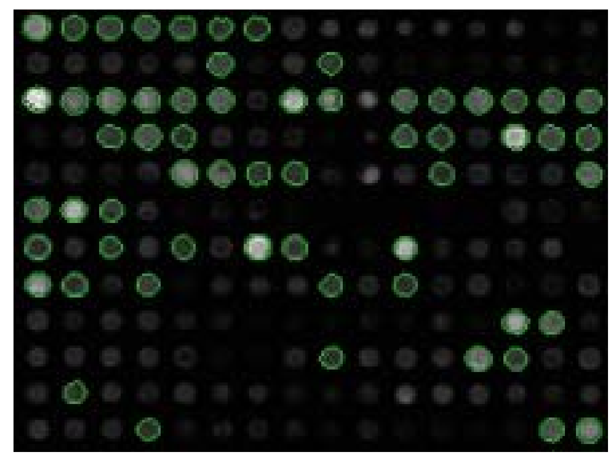
Original Image



Alternate-Sequential Filtered



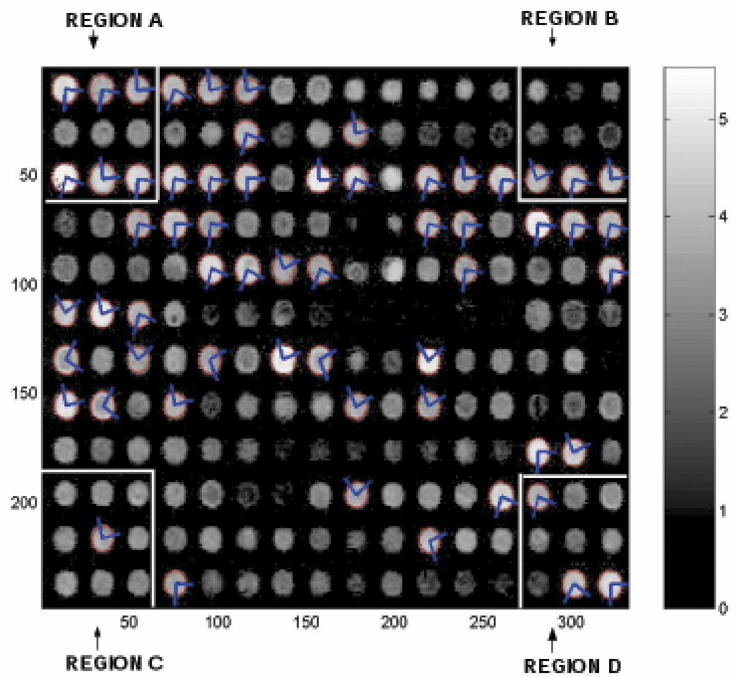
Watershed Transformed



Final Segmented Image

Siddiqui, Hero and Siddiqui, Asilomar-02

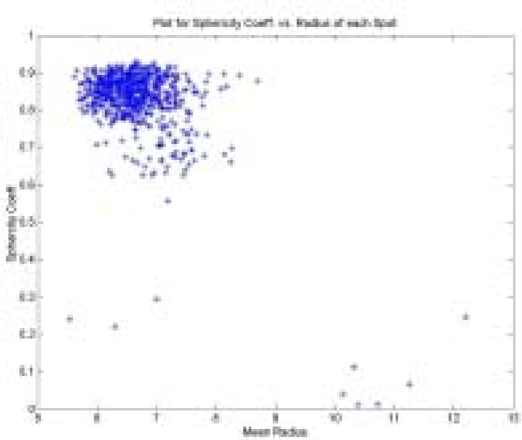
Spot EigenAnalysis



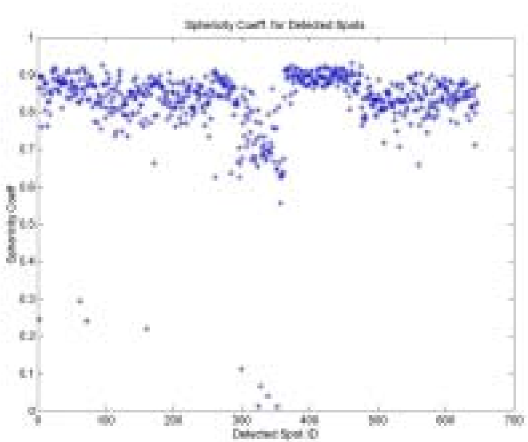
- Gray level covariance matrix over each spot boundary is calculated
- Eigen analysis of each covariance matrix is performed
- Trends in direction of eigenvectors indicate systematic bias in spot printing

Siddiqui, Hero and Siddiqui, Asilomar-02

Circularity Coefficient: mean(r^2)/(mean(r))^2



Radius vs. Sphericity Coeff.

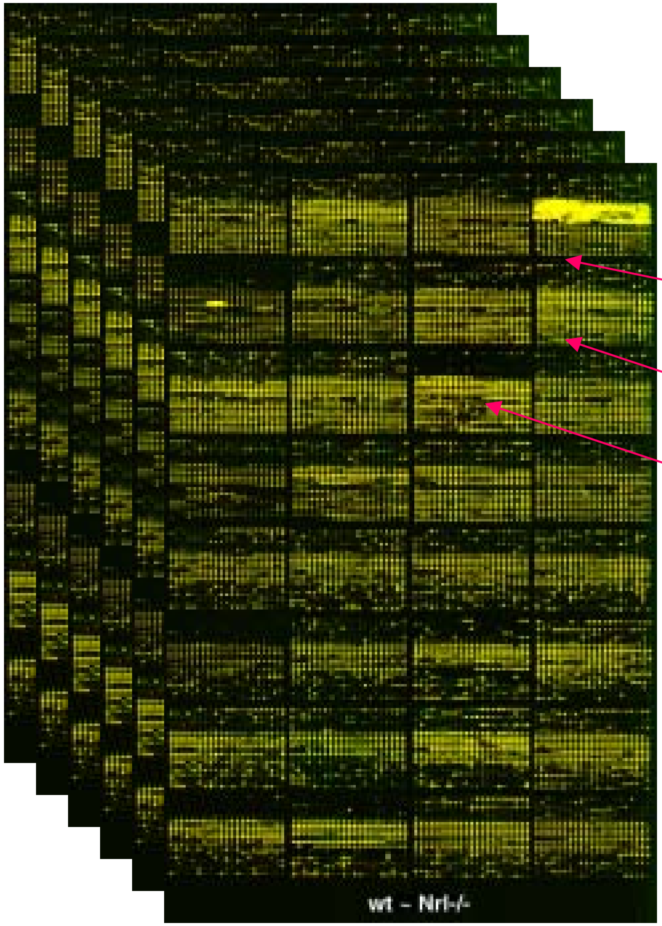


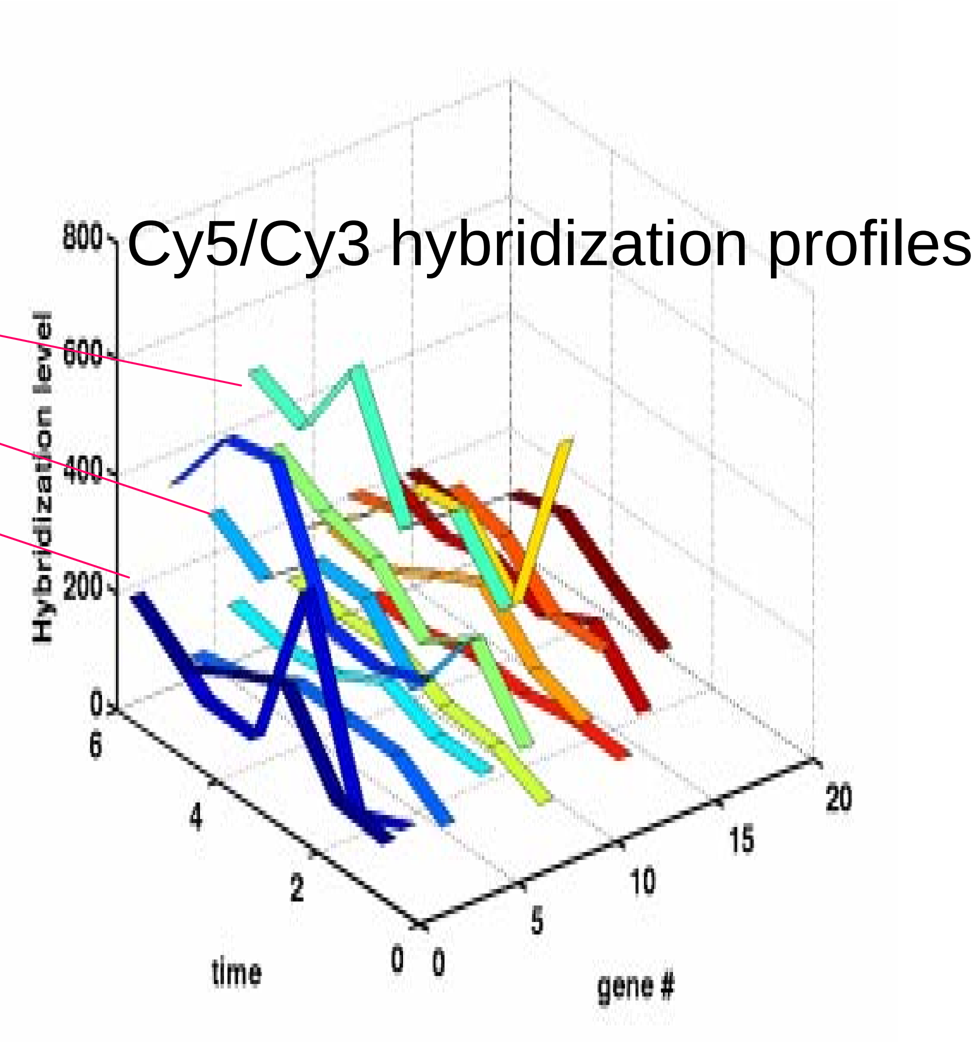
Sphericity Coeff

- Plot for Radius vs. Sphericity Coefficients (measure of circularity) of spots
- Spots with lower sphericity coefficients appear in lower half of plane
- Closer the sphericity coeff. is to 1, the better it is
- Deviation from circularity may give cause to discard data

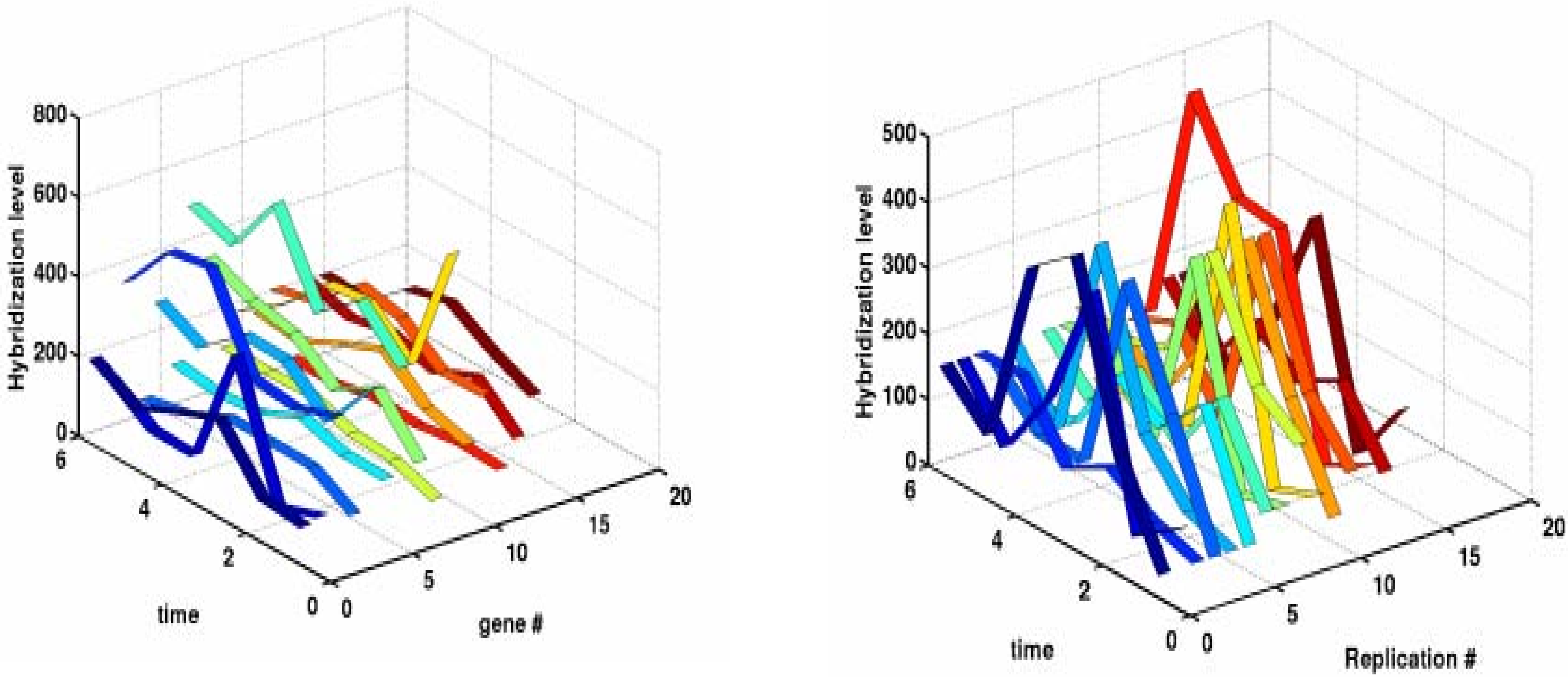
Siddiqui, Hero and Siddiqui, Asilomar-02

Another Dimension: Expression Profiles





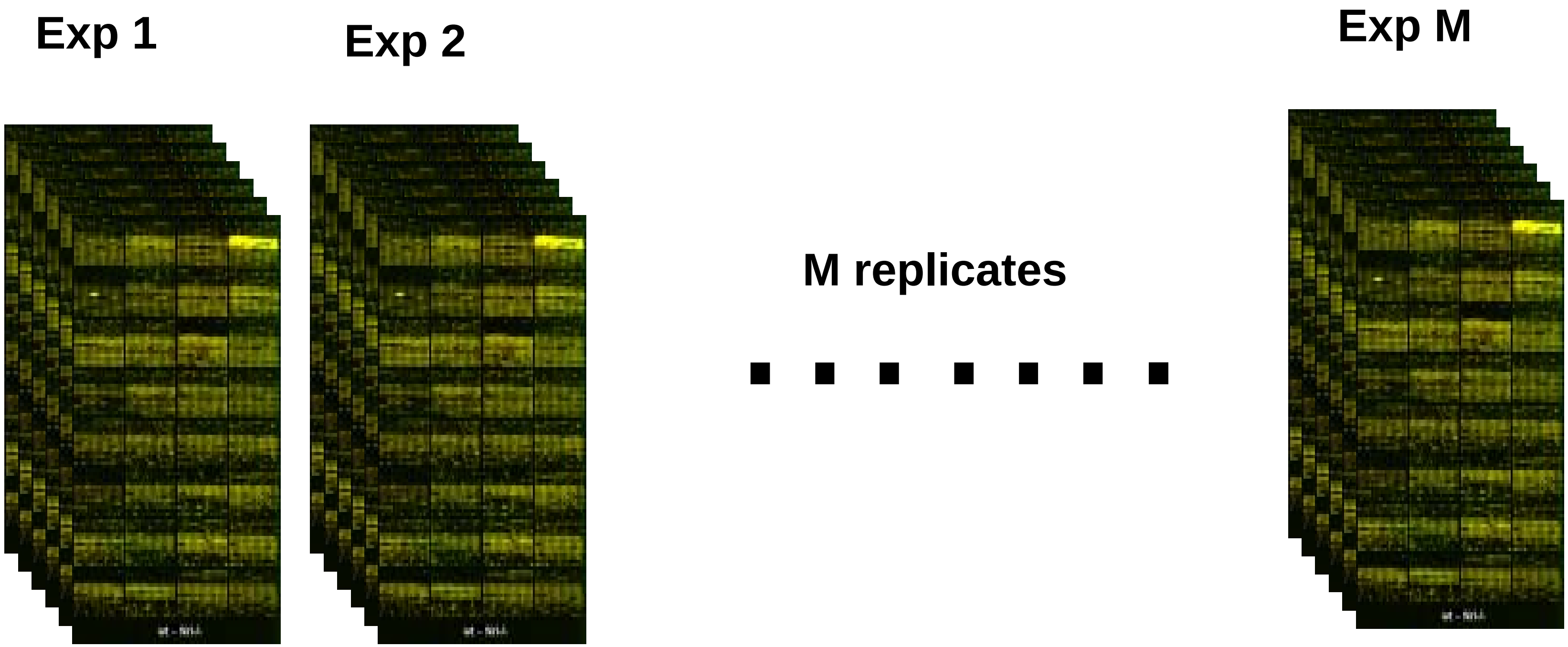
Problem: Intrinsic Profile Variability



Across-gene variability

Within-gene variability

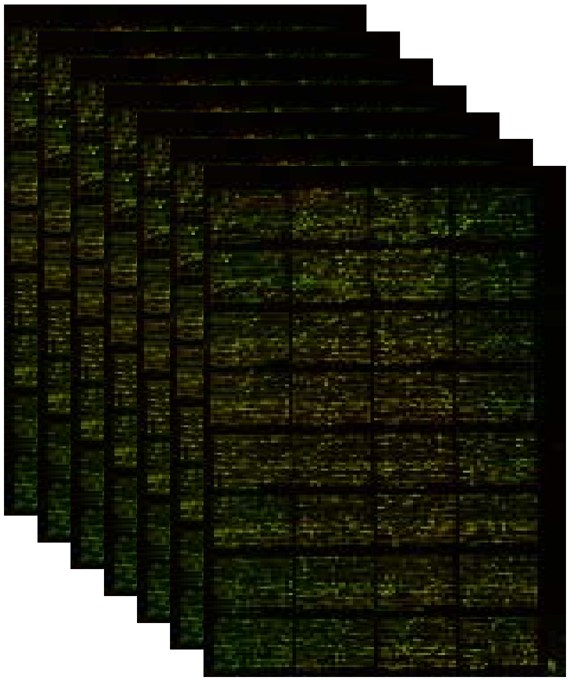
Solution: Experimental Replication



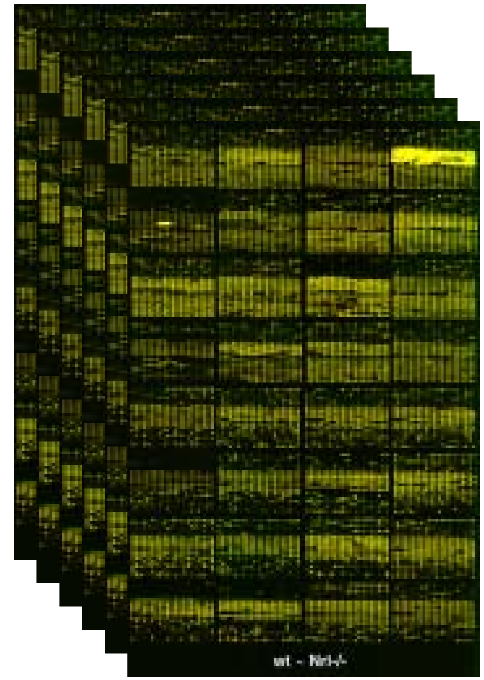
Issues:

- 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

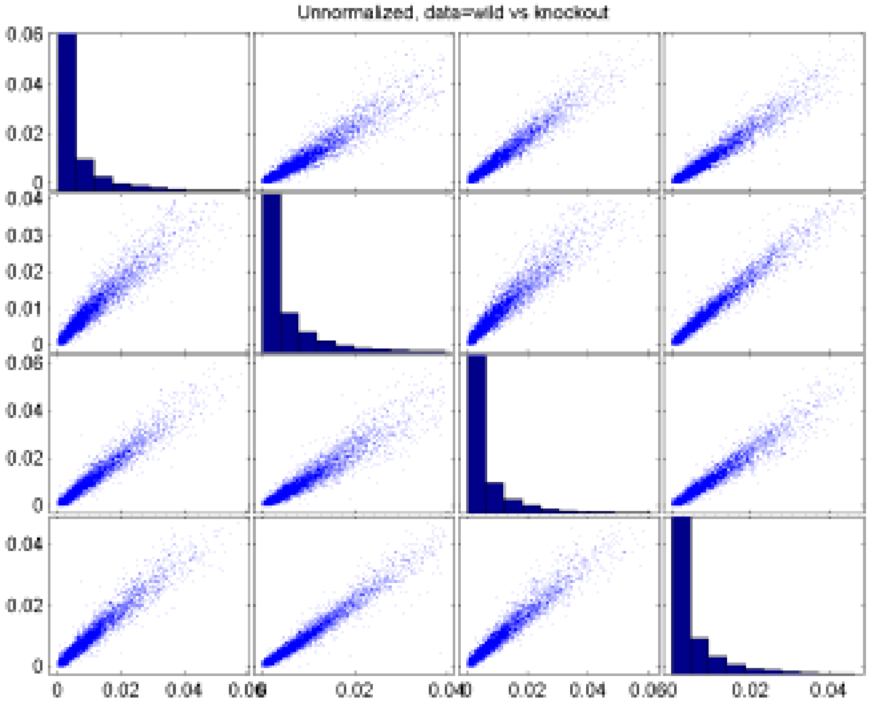


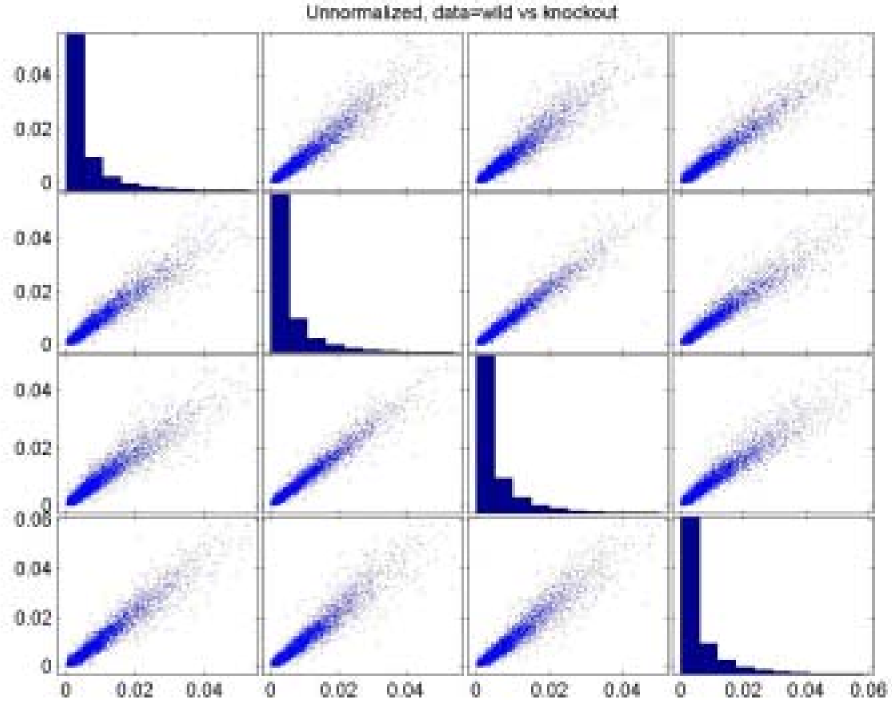
Experiment B

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

Un-Normalized Data Sets

Within-experiment intensity variations mask A-B differences:





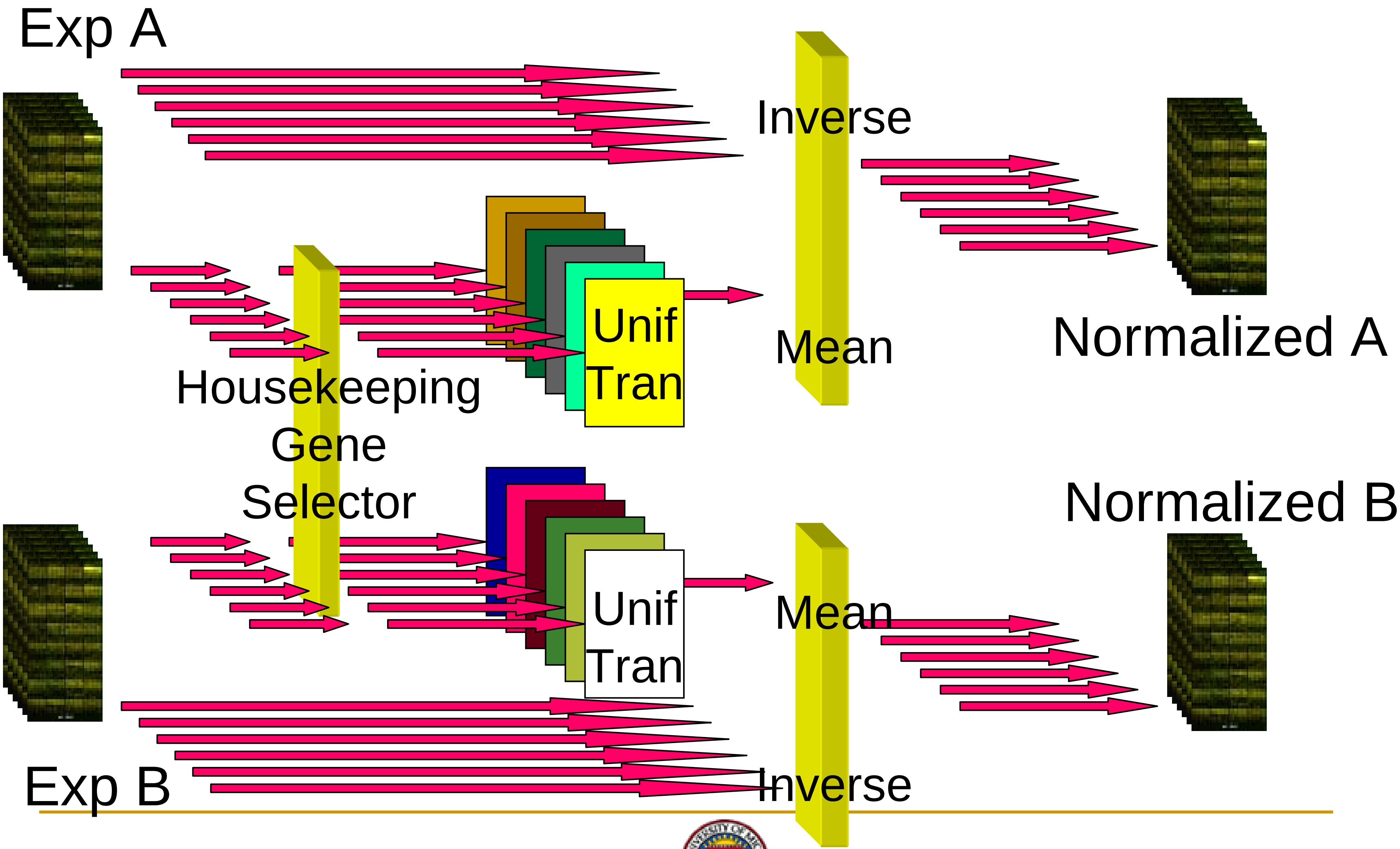
Experiment A (Wildtype) **Experiment B (Knockout)**

Hero&Fleury, ISSP-03

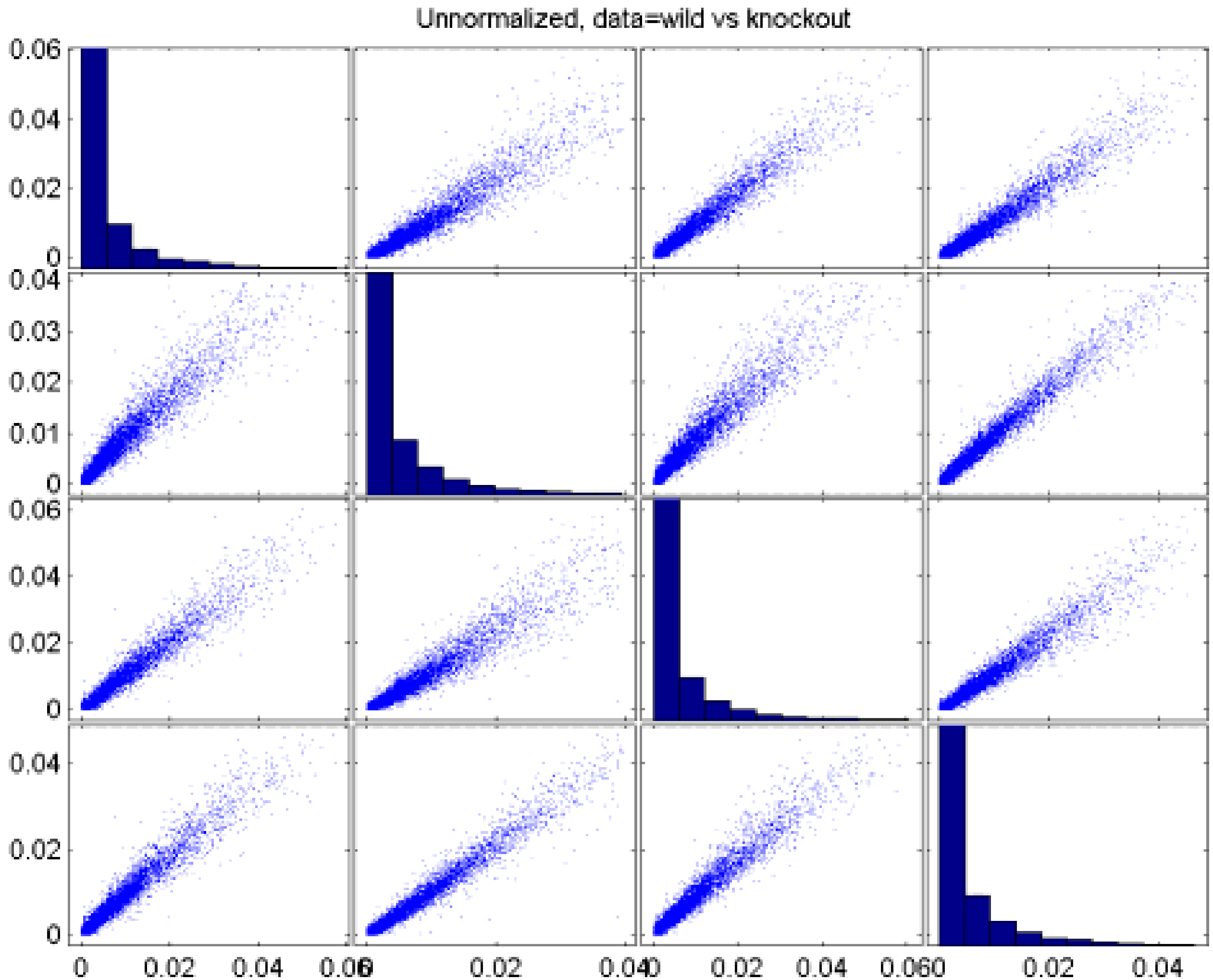
Two Approaches

- If quantitative gene profile comparisons are required:
  - must find normalization function to align all data sets within an experiment to a common reference.
- If only ranking of gene profile differences is required:
  - No need to normalize: can apply rank order transformation to measured hybridization intensities

A vs B Microarray Normalization Method

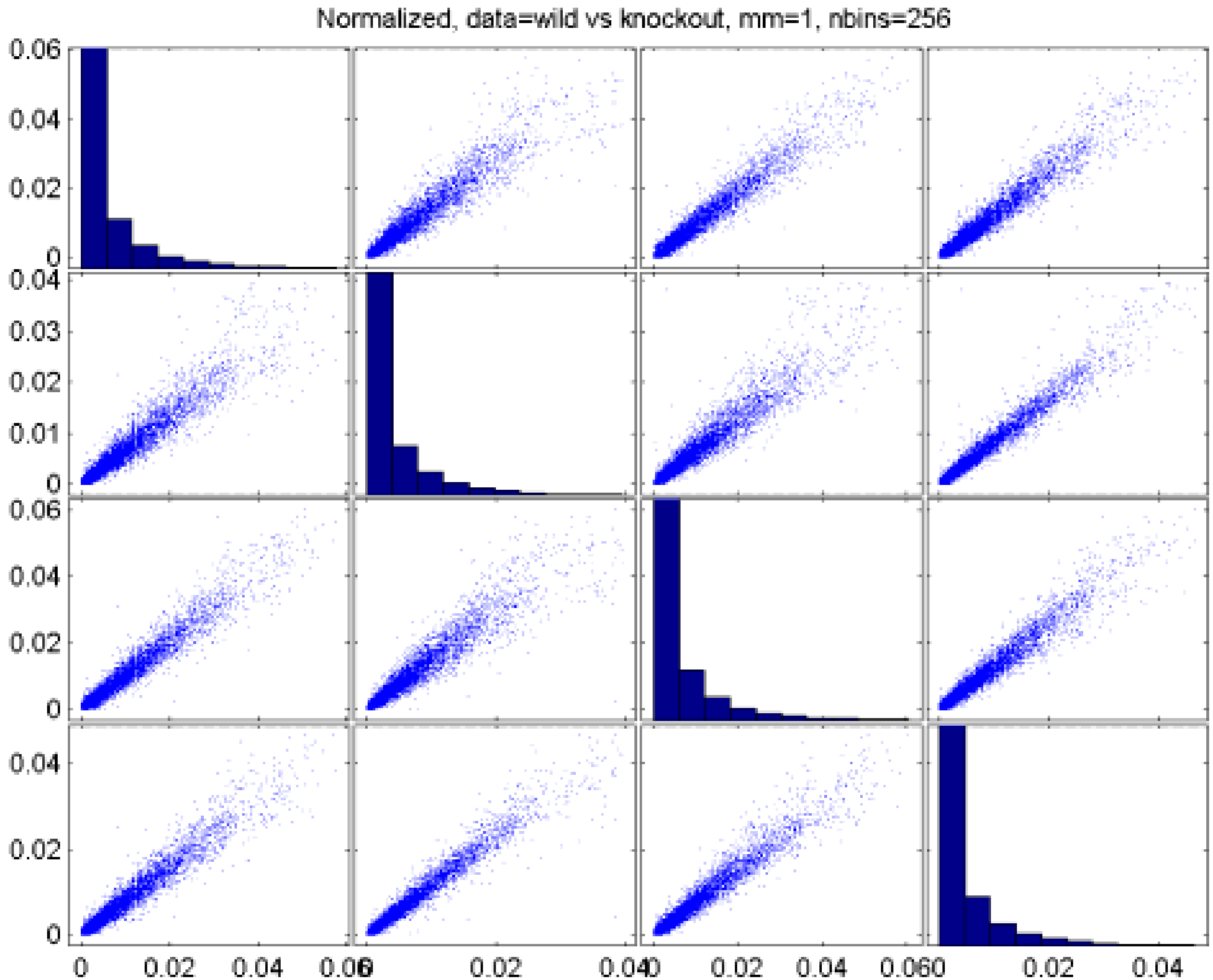


Un-Normalized Data Set (Wildtype)



Hero&Fleury, ISSP-03

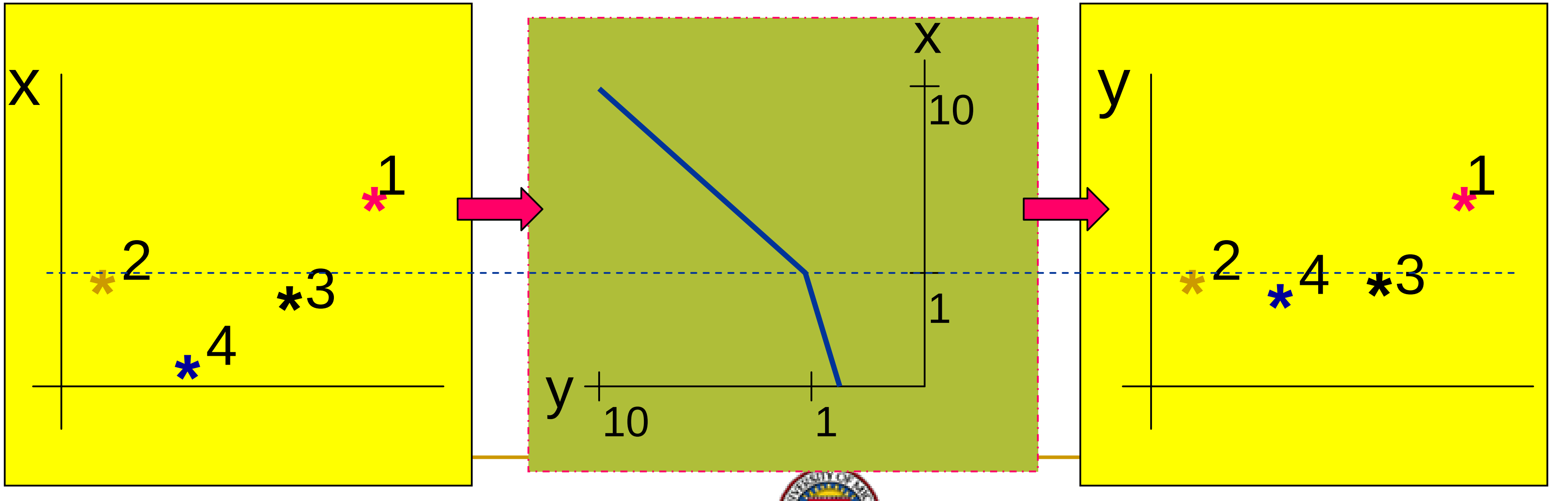
Normalized Data Set (Wildtype)



Hero&Fleury, ISSP-03

Profile Rank Order Statistics

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

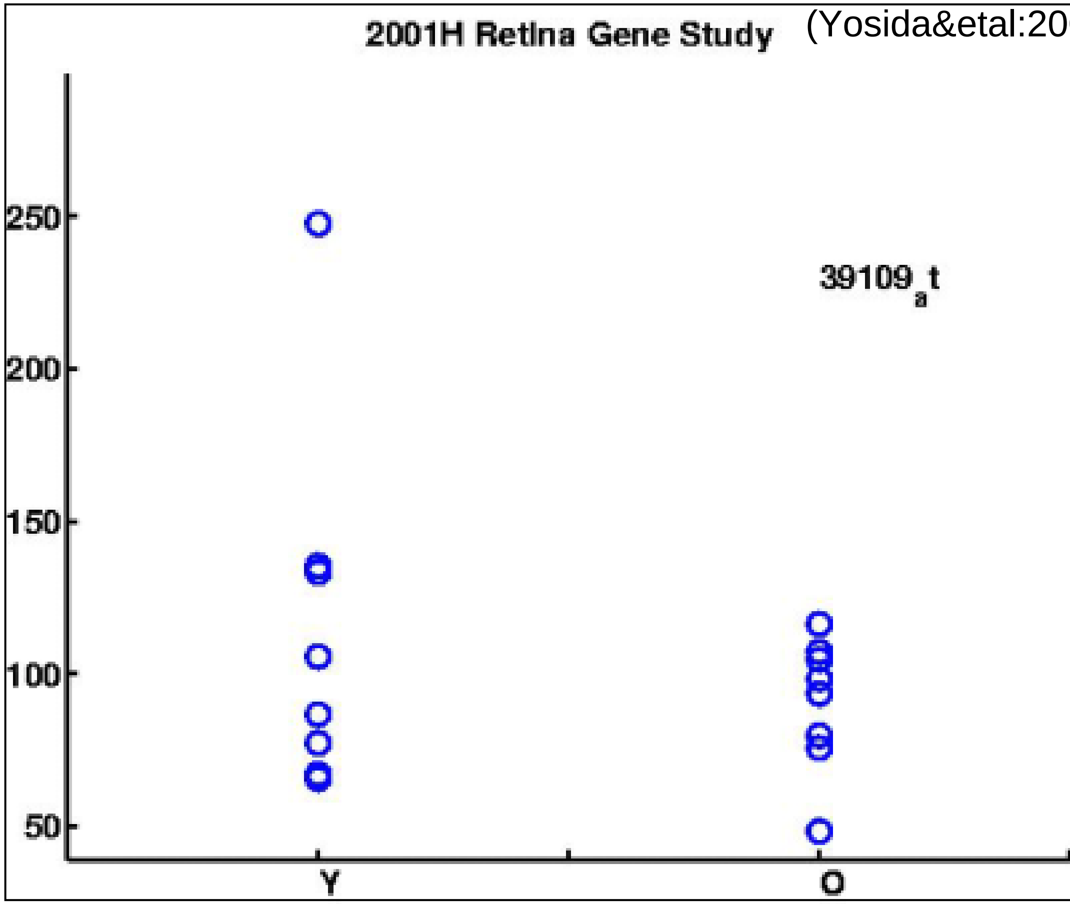


Mining Gene Expression Data

- Issues
  - Feature space
  - Feature selection criteria
  - Statistical robustification
  - Cross-validation
  - Experimental Validation

Y/O Human Retina Study

(Yosida&etal:2002)



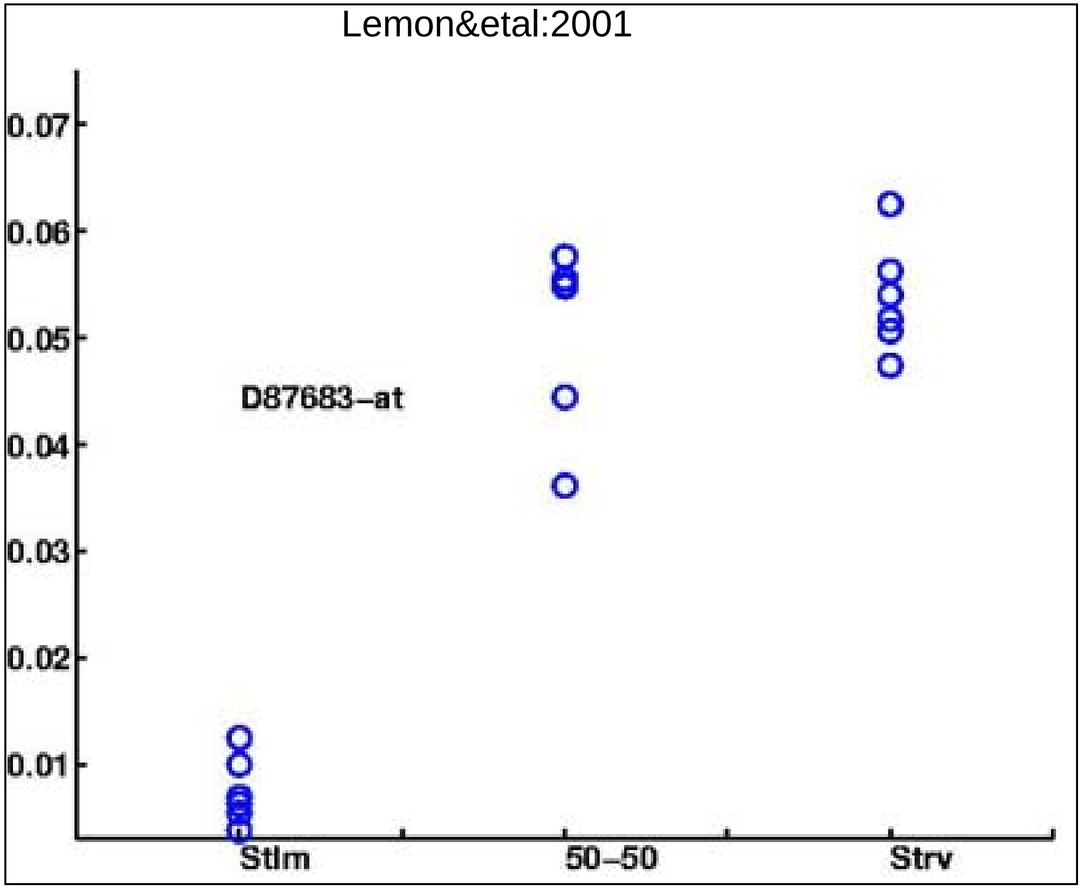
16 individuals in

2 groups of 8 subjects

Selection criteria:

$$\begin{aligned}\xi_1(g) &= \overline{O}(g) - \overline{Y}(g)\\ \xi_2(g) &= (\sigma_O^2(g) + \sigma_Y^2(g))/2\end{aligned}$$

Fred Wright's Human Fibroblast Data



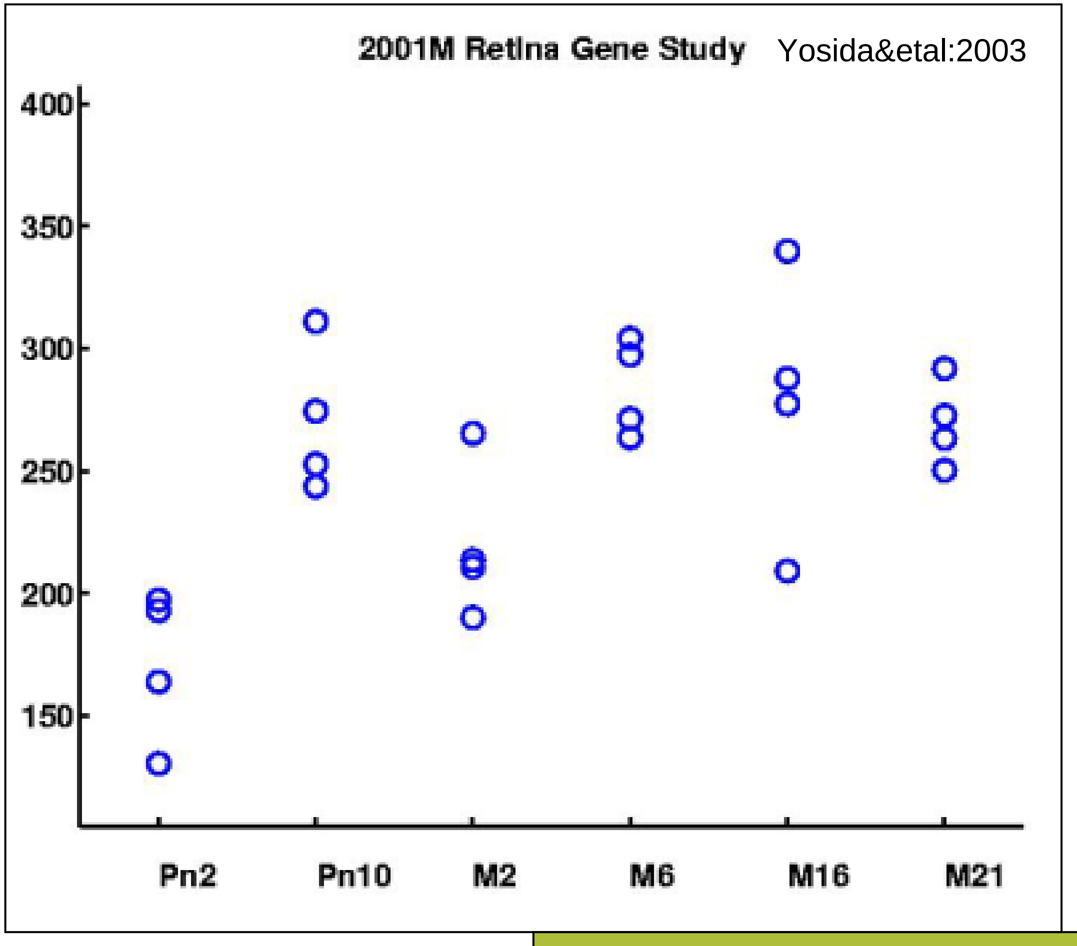
18 individuals in

3 groups of 6 subjects

Selection criteria:

$$\begin{aligned}\xi_1(g) &= (\mu_{100}(g) - \mu_{50})(g))(\mu_{50}(g) - \mu_0(g)) \\ \xi_2(g) &= (\sigma^2_{100}(g) + \sigma^2_{50}(g) + \sigma^2_0(g))/3\end{aligned}$$

Mouse Retinal Aging Data



24 mice in

6 groups of 4 subjects

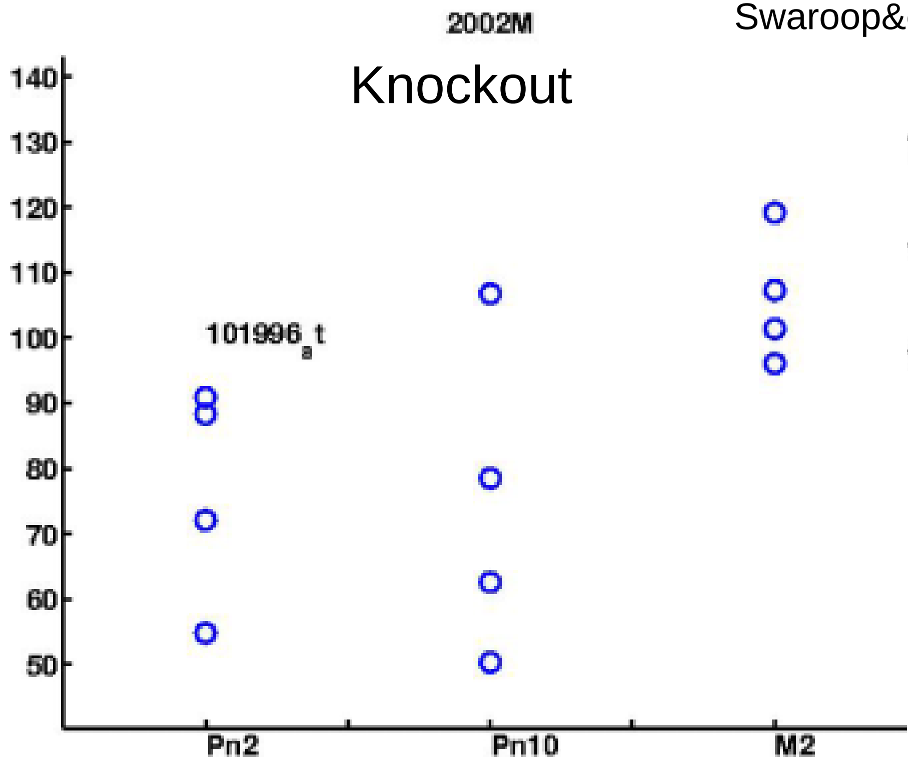
Selection criteria:

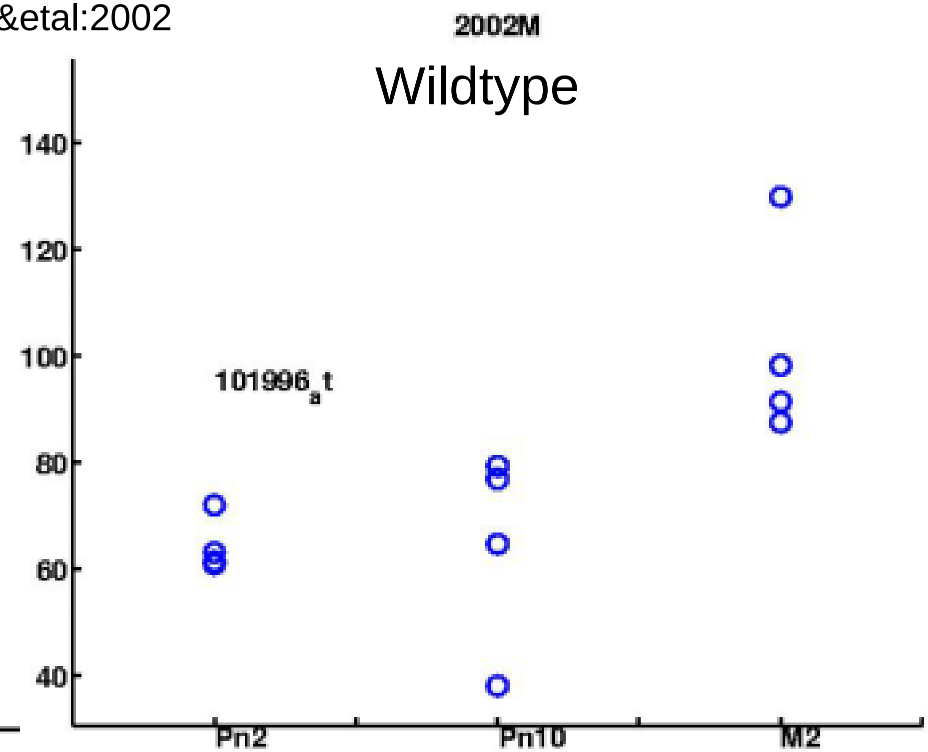
$$\begin{aligned}\xi_1(g) &= \Delta_{M21,M2}(g) = (\mu_{M21}(g) - \mu_{M2})(g))^2 \\ \xi_2(g) &= \max_{t=3,\dots,6} \{\mathrm{var}(\Delta_{t+1,t})(g)\}\end{aligned}$$

NRL Knockout vs Wildtype Retina Study

12 knockout/wildtype mice in 3 groups of 4 subjects

Swaroop&etal:2002





Selection criteria:

$$\begin{aligned} \xi_1(g) &= \Delta^2_{K,W}(g) = \|\mu_K(g) - \mu_W)(g)\|^2 \\ \xi_2(g) &= \max\{\mathrm{var}_K(g), \mathrm{var}_W(g)\} \end{aligned}$$

Data Mining with a Single Criterion

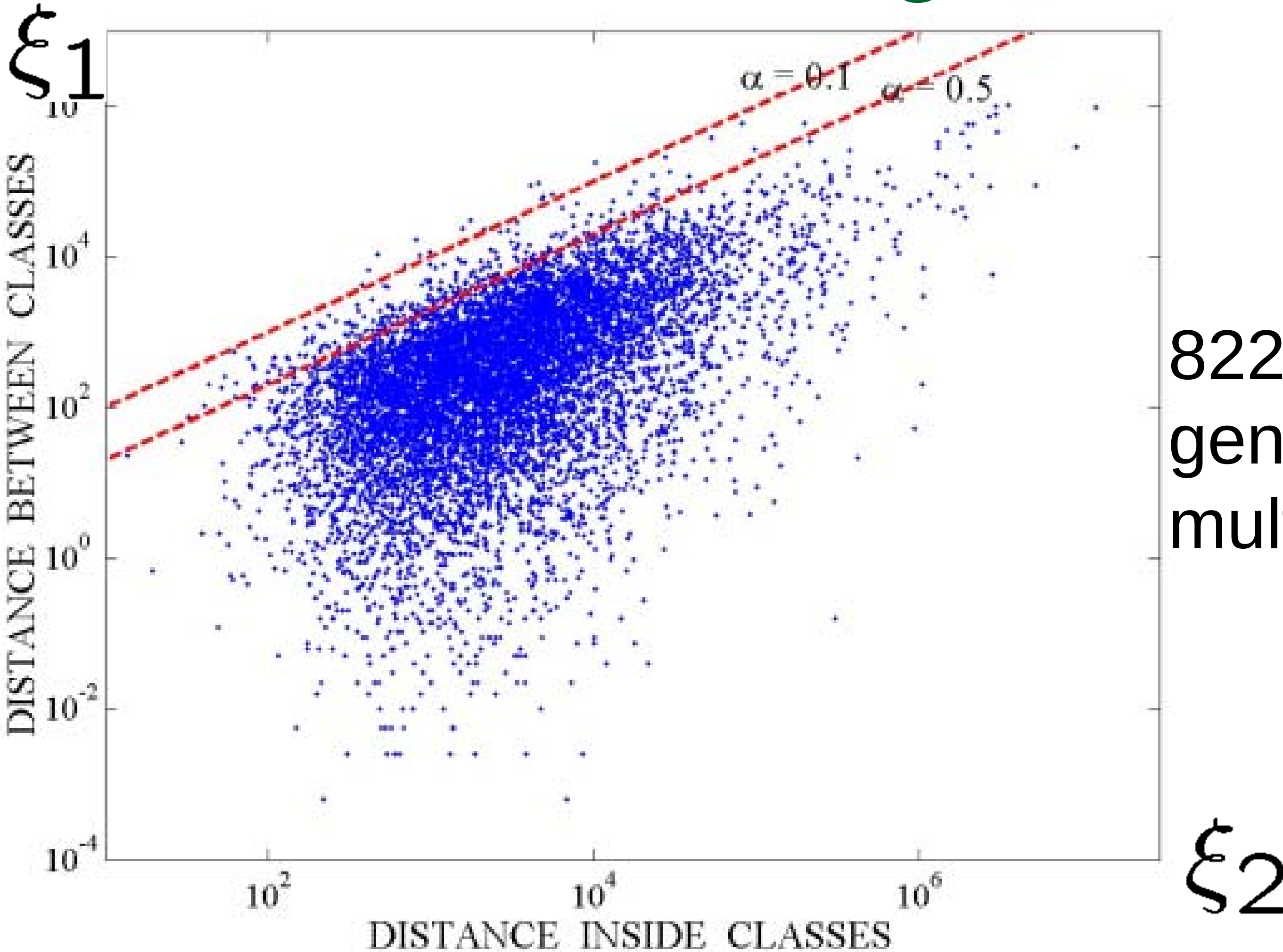
- Paired t-test with False Discovery Rate:

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

- For Y/O Human study:

$$T(g) = \frac{|\overline{O}(g) - \overline{Y}(g)|}{\sqrt{(\sigma^2_O(g) + \sigma^2_Y(g))/2}}$$

Multicriterion scattergram: T-test

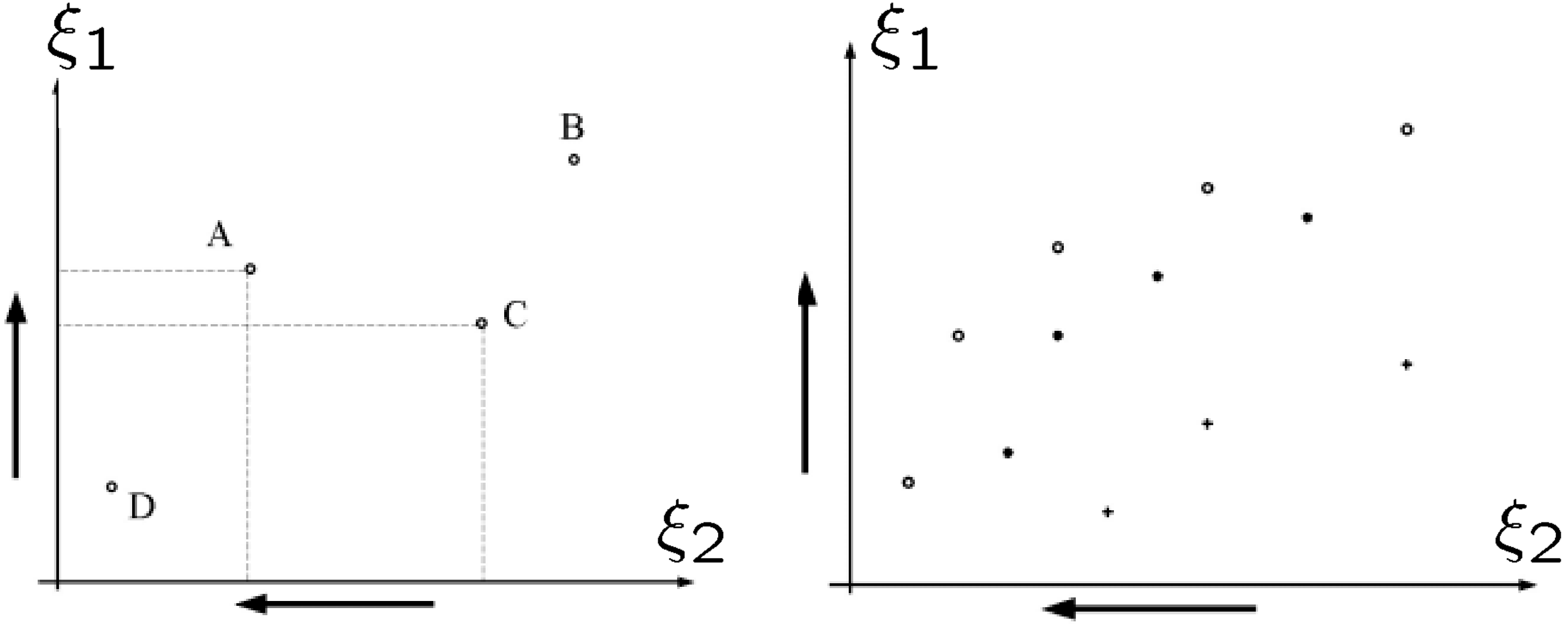


8226 Y/O retina genes plotted in multicriteria plane

Fleury&etal ICASSP-02

Multicriterion Selection Criteria

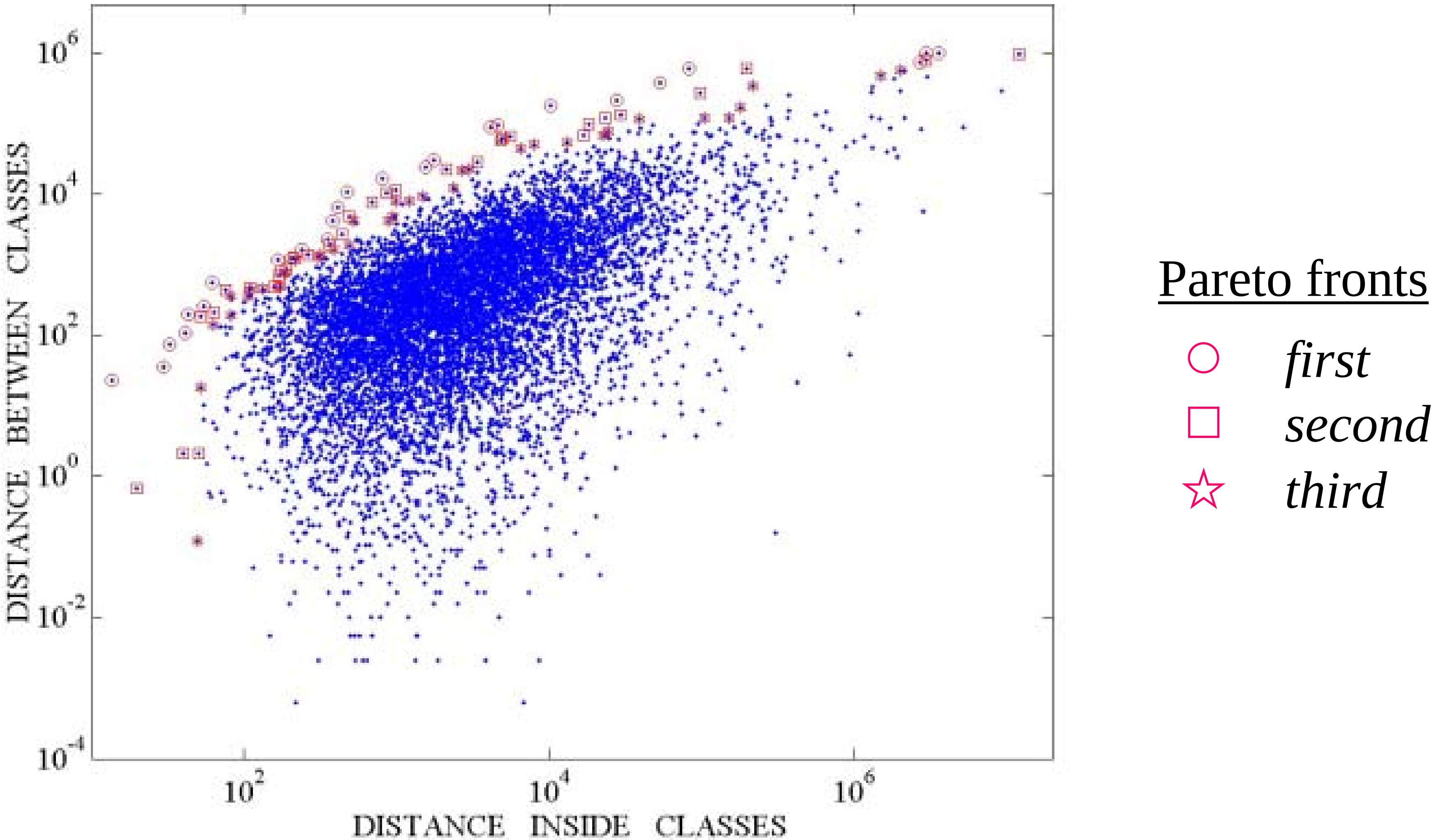
- Seek to find Pareto-optimal genes which strike a compromise between two criteria



A,B,D are Pareto optimal

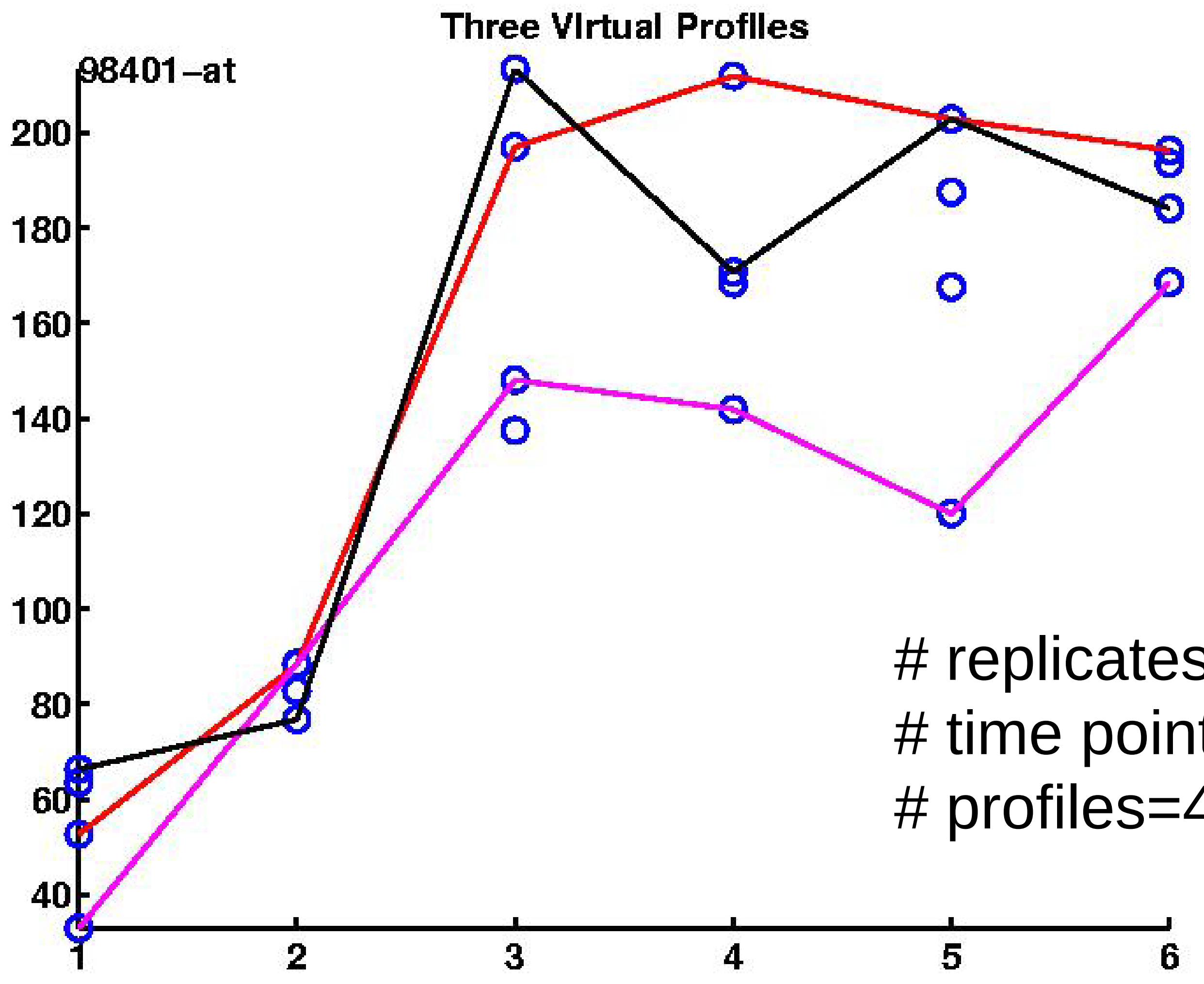
Pareto Fronts

Multicriterion scattergram: Pareto Fronts



Fleury&etal ICASSP-02

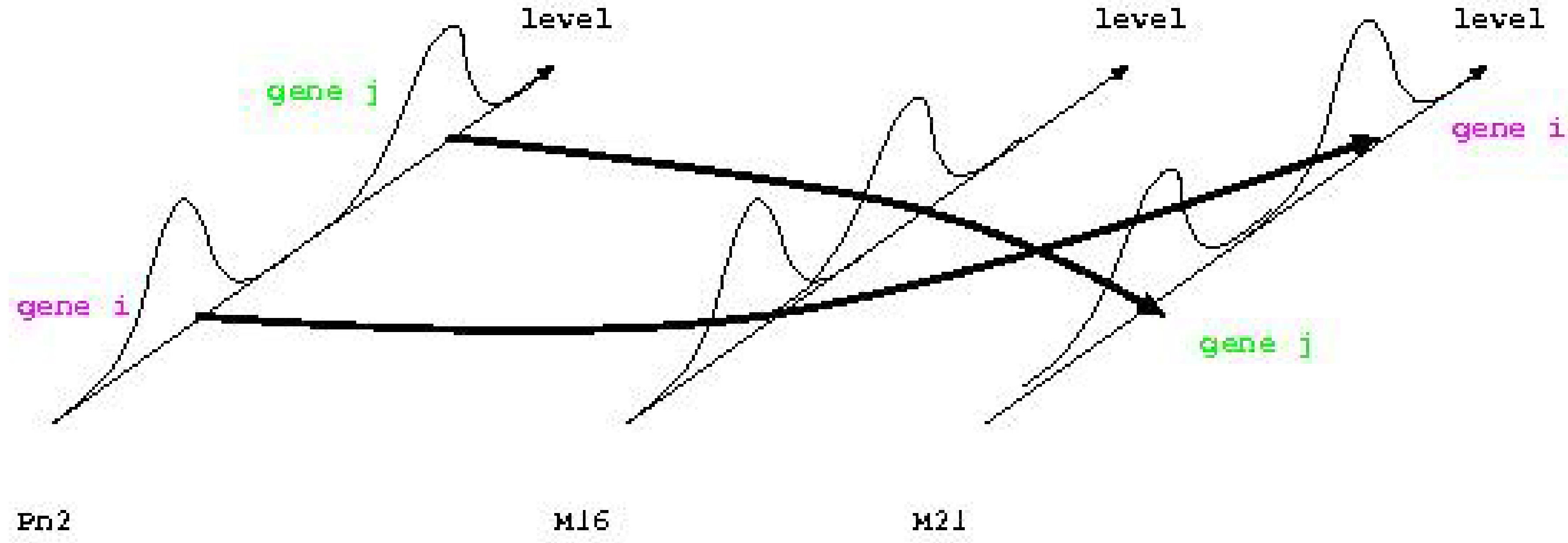
Cross-Validation Approach: Resampling



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

Bayesian approach: Posterior Analysis

$$P(i|Y) = P(\text{gene } i \text{ on PF} \mid \text{data } Y)$$



Pareto Front Likelihood table

PPF linear contrast	P(i|Y)	RPF linear contrast	P(i|Y)	RPF non-parametric	P(i|Y)
AFFX-ThrX-5-at	0.999	AFFX-DapX-5-at	1	U14394-at	0.944
HG3342-HT3519-s-at	0.998	AFFX-ThrX-5-at	1	U23435-s-at	0.694
AFFX-DapX-5-at	0.998	AFFX-ThrX-M-at	1	AFFX-PheX-M-at	0.685
HG831-HT831-at	0.996	HG3342-HT3519-s-at	1	AFFX-LysX-3-at	0.662
AFFX-ThrX-M-at	0.986	HG831-HT831-at	1	AFFX-LysX-M-at	0.648
X69111-at	0.984	U14394-at	1	AFFX-HSAC07/X00351-5-at	0.352
U14394-at	0.974	V00594-at	1	AFFX-ThrX-5-at	0.301
AFFX-LysX-3-at	0.962	X69111-at	1	AB000115-at	0.287
V00594-at	0.955	U45285-at	0.944	AFFX-DapX-5-at	0.245
U45285-at	0.932	AFFX-LysX-3-at	0.917	U53003-at	0.176
AB000115-at	0.899	AFFX-HSAC07/X00351-5-at	0.806	M92934-at	0.111
AFFX-HSAC07/X00351-5-at	0.866	AB000115-at	0.417	D29992-at	0.083
U73379-at	0.837	U73379-at	0.13	HG831-HT831-at	0.069
AFFX-DapX-M-at	0.678	V00594-s-at	0.074	S79522-at	0.042
Y09912-rna1-at	0.67	U75362-at	0.037	V00594-s-at	0.042
U75362-at	0.56	AFFX-PheX-5-at	0.028	D43636-at	0.032
AFFX-DapX-3-at	0.555	U03399-at	0.009	U22377-at	0.032
V00594-s-at	0.554			U75362-at	0.028
HG1980-HT2023-at	0.483			S70585-rna1-at	0.014
HG3044-HT3742-s-at	0.441			L02320-at	0.009
D43636-at	0.389			L05515-at	0.009
L27624-s-at	0.387			V00594-at	0.009
U03399-at	0.378			X69111-at	0.009
S69370-s-at	0.321			AFFX-PheX-5-at	0.005
AFFX-PheX-5-at	0.315			HG174-HT174-at	0.005

Hero&Fleury:VLSI03

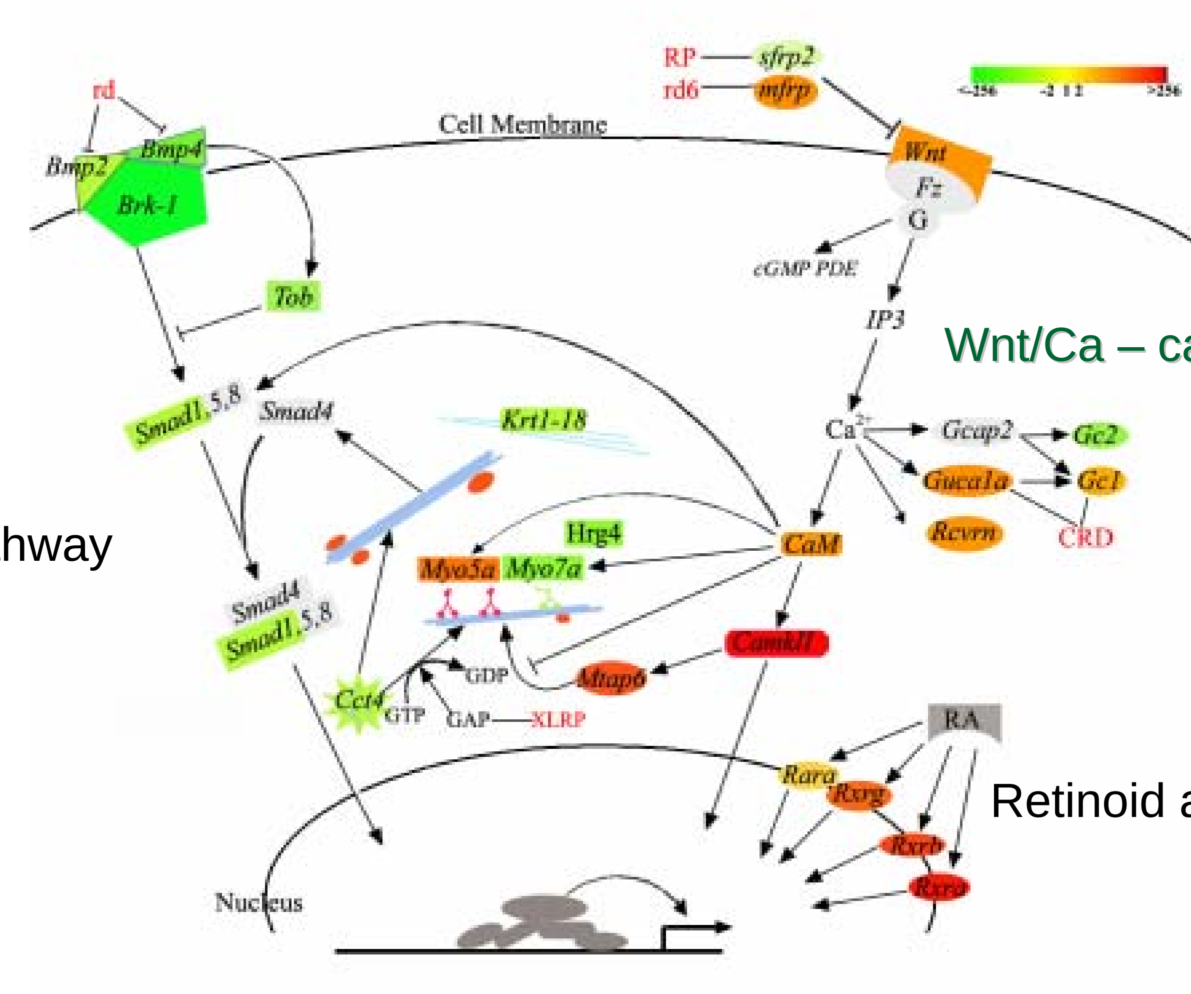
Robustification and Validation Issues

- Cross-validation recomputes Pareto fronts over all virtual profiles (Fleury&etal:2002).
- Bayesian Pareto front also robustifies prior uncertainty in data (Hero&Fleury:2002).
- Computational issues:
  - Cross-validated fronts: completely data-driven but computation is O(m^t)
  - Bayesian Pareto fronts: requires joint density of criteria and marginalization. Computation is linear in # replicates (m) and # time points (t).

The Post-Genomic Era

- Whole genomes of species will be mapped
- Genetic pathways to structure, metabolism, disease, will remain as open questions
- Pathway analysis: what are the important gene interactions?
  - Requires performing many more experiments than zero-interaction analysis
  - Computational load is exponentially increasing in number of genes in pathway
  - New algorithms and models are needed

Draft Pathways for Photoreceptor Function



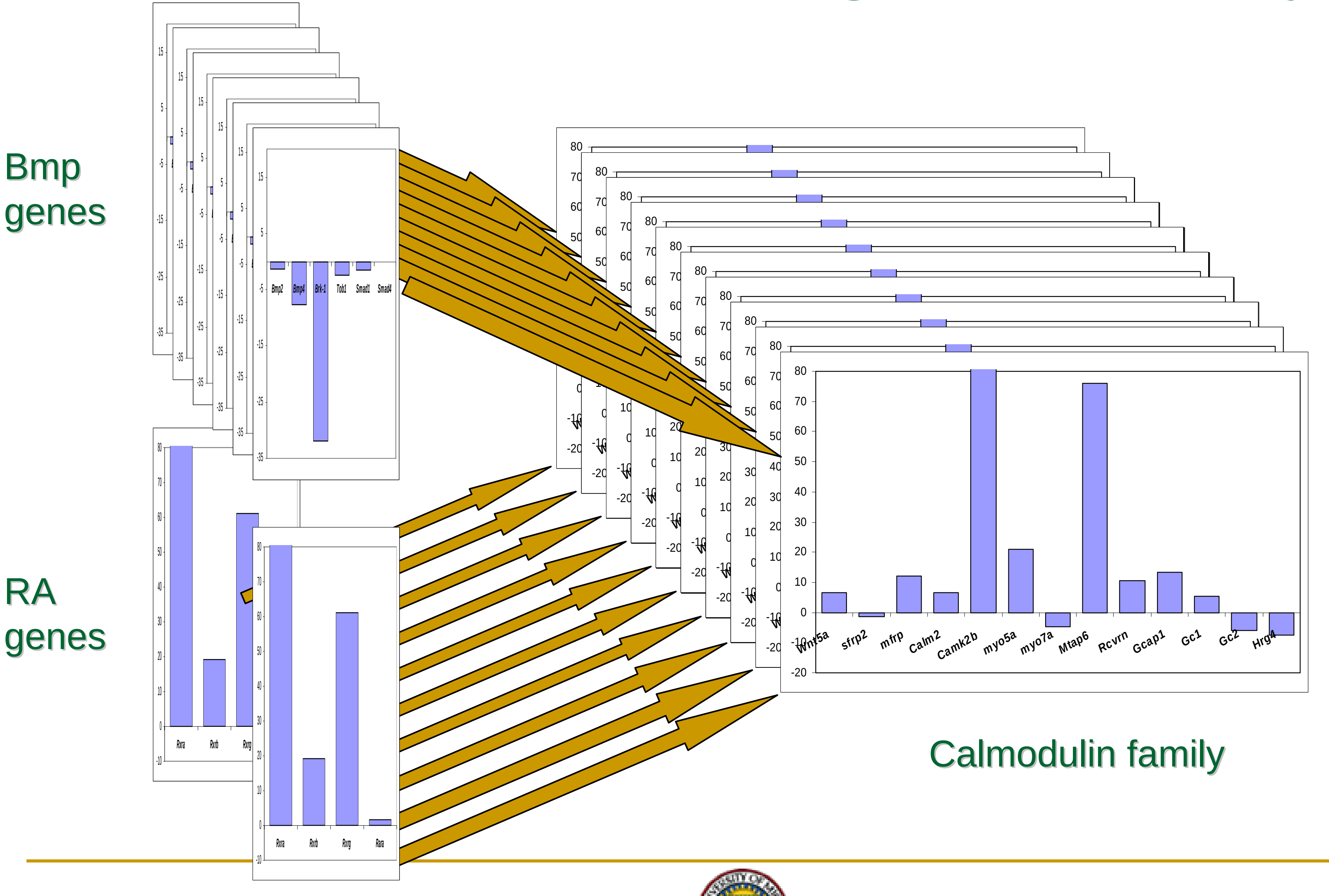
Wnt/Ca – calmodulin pathway

Bmp pathway

Retinoid acid pathway

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

Each Link: Gene Co-regulation Study



Conclusions

- Signal processing, math, computer science, statistics: ever-increasing role in genomics
- New frontiers:
  - Protein arrays
  - Mass Spect
  - Molecular Imaging
- Bottleneck will remain: computational and statistical inadequacies!

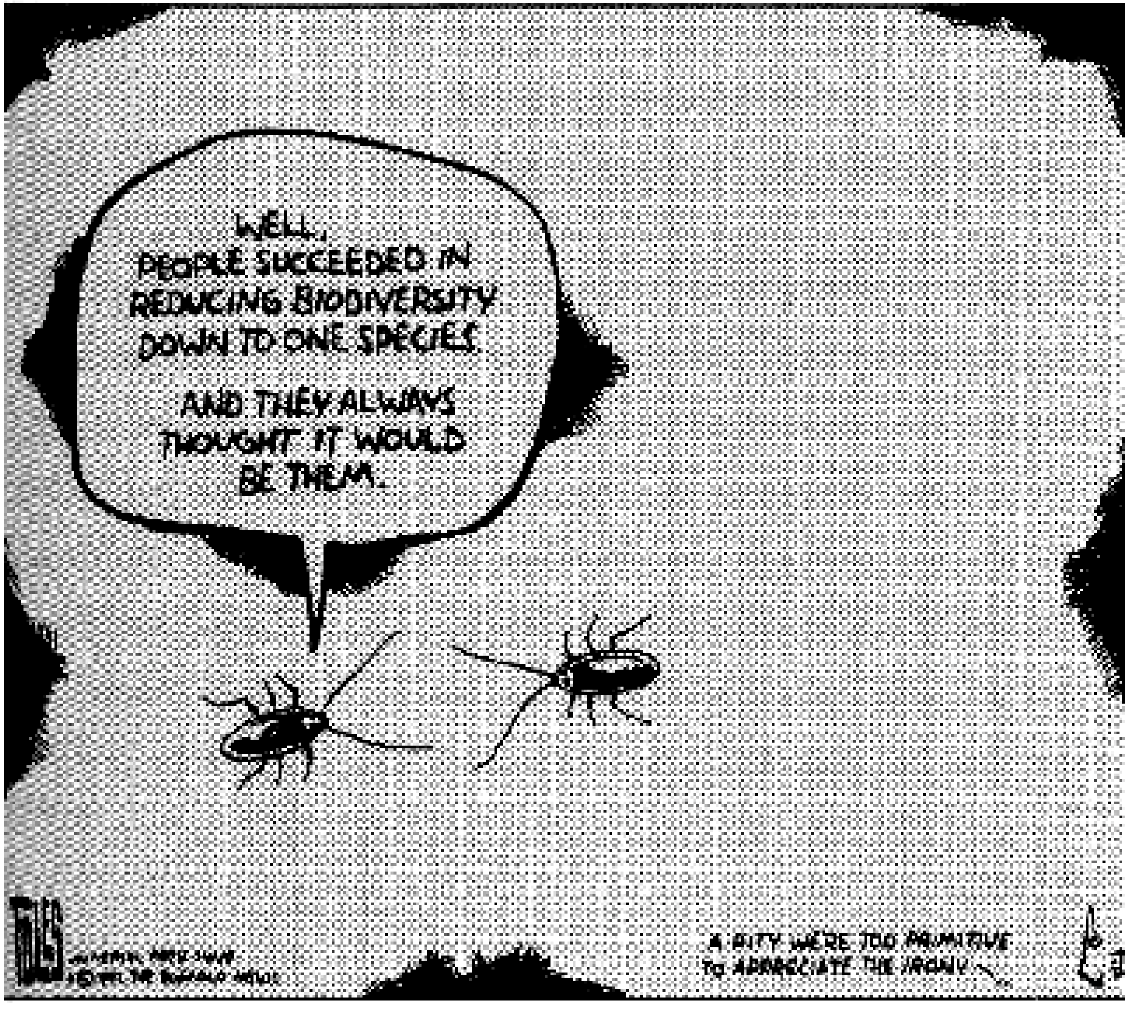
Dawning of Post-Genomic Era

GENETIC DAWN

SERGUEI
FRANCE

Post-Post-Genomic Era?

Or….


WELL, PEOPLE SUCCEEDED IN REDUCING BIODIVERSITY DOWN TO ONE SPECIES
AND THEY ALWAYS THOUGHT IT WOULD BE THEM.
A PITY WE'RE TOO PRIMITIVE TO APPRECIATE THE IRONY...

Oligonucleotide GeneChip Microarray