



# Bio - Microarrays as a Disruptive Technology and Why Computation Will Be Central to Managing the Disruption

Isaac S. Kohane  
<http://www.chip.org>

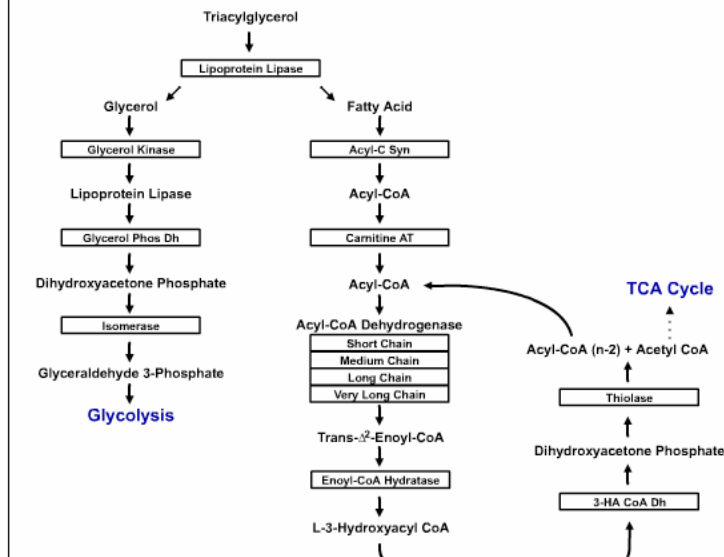


## Define Disruptive Technology

- **Disruptive Technology** is a term popularized by Harvard Business School professor Clayton Christensen in his book *The Innovator's Dilemma*.
  - ✓ Christensen believes that the main reason that successful and apparently well-run and well established organizations loose market share, and sometimes go out of business, is that they fail to recognize the distinction between sustaining and disruptive technologies.
  - ✓ "It has to meet at least two of three criteria: Be ten times cheaper than any alternative, have ten times higher performance, and ten times higher functionality. All three is best" Rick Rotondo, Director of Disruptive Technologies, MeshNetworks.
    - ☞ Transistor vs. vacuum tube
  - ✓ Are microarrays disruptive?
  - ✓ And if so, Have biologists recognized the disruption?



## Fatty Acid Degradation



3-  
1)



## Why are microarrays disruptive to biology?

- With small numbers of genes, the uncertainty is bounded
  - ✓ There are expectations about how small numbers of genes interact
- 1000's of genes can be done in time taken to do a handful previously
- A commodity available to non-researcher
  - ✓ Interpretation is "standardized"
- Output is foreign to basic biology researchers
  - ✓ Does not load easily into standard tools
  - ✓ Analysis is non-standard

IFF (Uncompressed) decompressor are needed to s



## What are the characteristic of a microarray How do you recognize one?

- Small form-factor capable of measuring significant fraction of the ---ome (Genome, Transcriptome, Proteome)
- Minimal labor in data acquisition
- Automated data path to a digital electronic format
- Sustainable high throughput processing



## Is this a microarray?

QuickTime?and a  
TIFF (Uncompressed) decompressor  
are needed to see this picture.

200 protein  
measurements  
(antibody based)

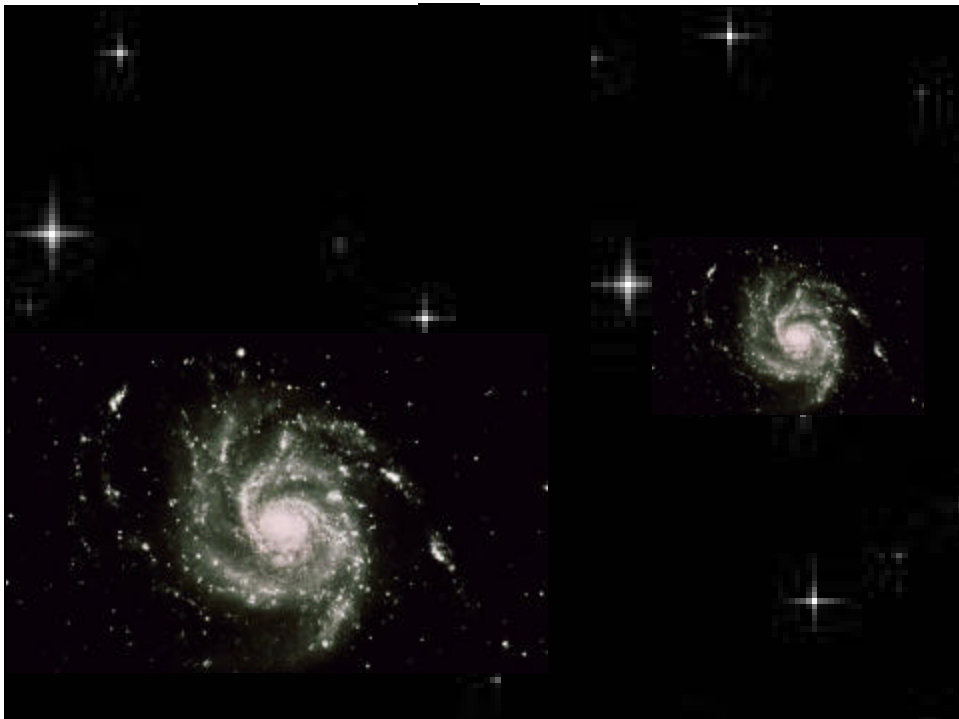
$10^{6+++}$  proteins

QuickTime?and a  
TIFF (Uncompressed) decompressor  
are needed to see this picture.



## What are we looking for in a microarray

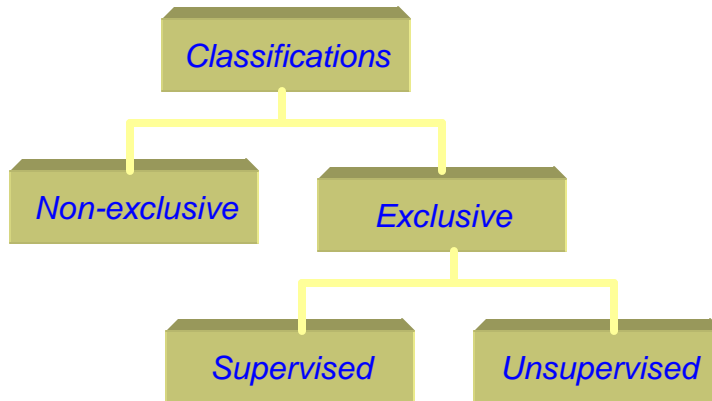
- Rarely: a single gene responsible for a process
- Infrequently: look at a specified genes pathway
- Frequently: a set of genes working in coordinated fashion
  - ✓ The assumption: there is structure at the biological scale
  - ✓ Why should this assumption hold?
  - ✓ What evidence do we have that it does?







## Taxonomy of machine learning

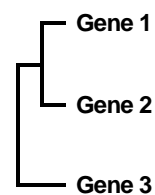


## Phylogenetic-type tree

	RNA Expr Gene 1	RNA Expr Gene 2	RNA Expr Gene 3
Experiment 1	0.7	0.3	7.3
Experiment 2	1.2	1.9	6.5
Experiment 3	1.1	0.9	8.1

### Correlation Coefficient

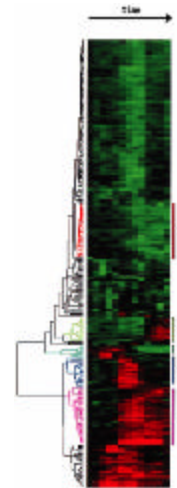
Gene 1			
Gene 2	.88		
Gene 3	-.19	-.62	



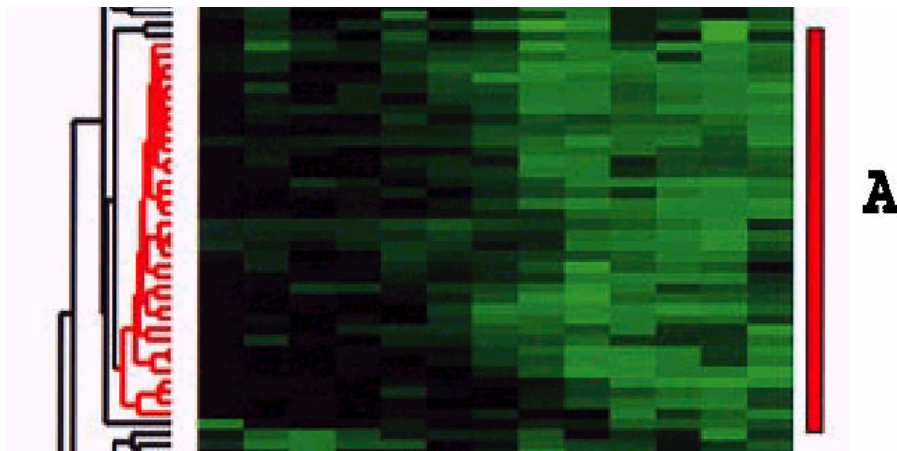


## Phylogenetic-type tree / Correlation coefficient

- Similarly functioning genes cluster together
  - ✓ E.g. cell cycle
  - ✓ E.g. structural proteins
- Also found 10 temporal clusters in 8613 genes in the response of human fibroblasts to serum
- Suggests role for hundred's of unknown genes that co-cluster with known genes



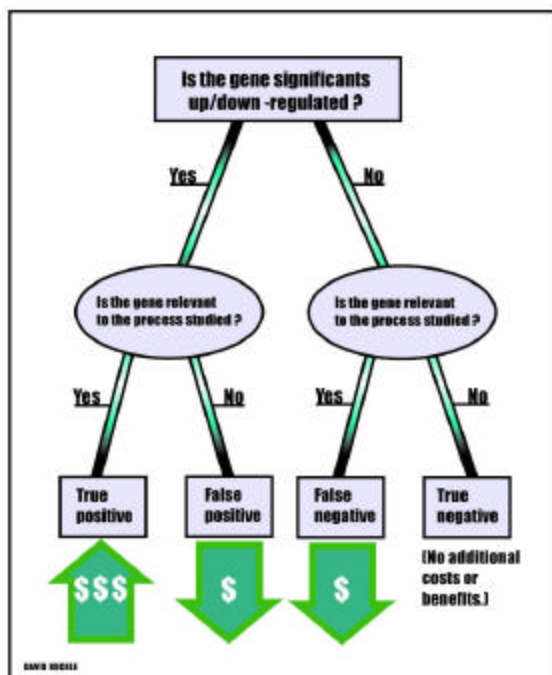
lyer VR, Science 1999;283(25):83-7.





## The point

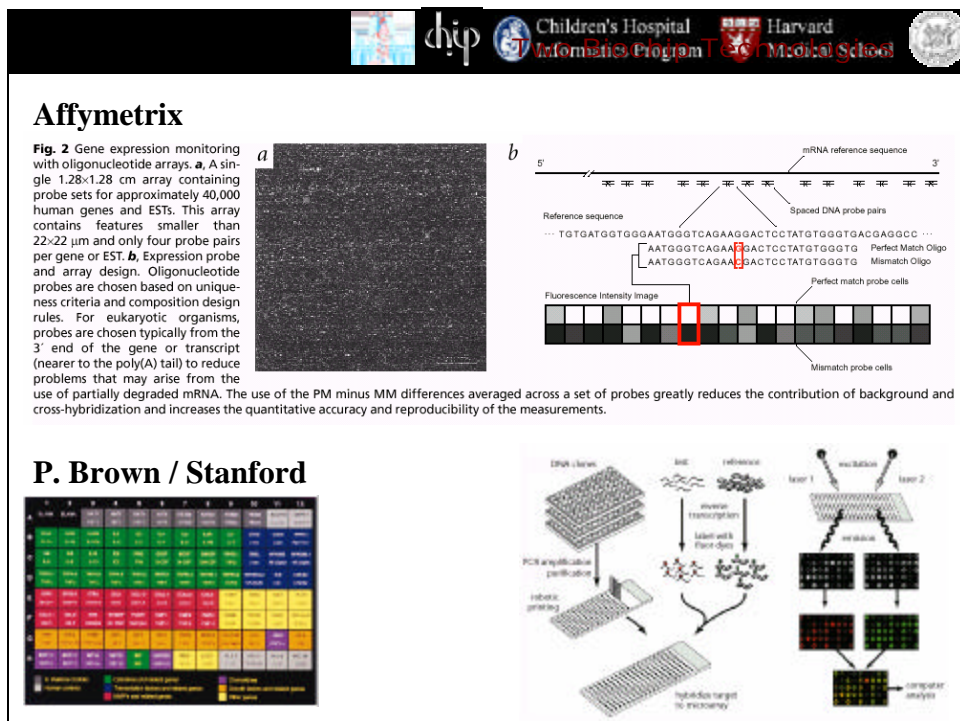
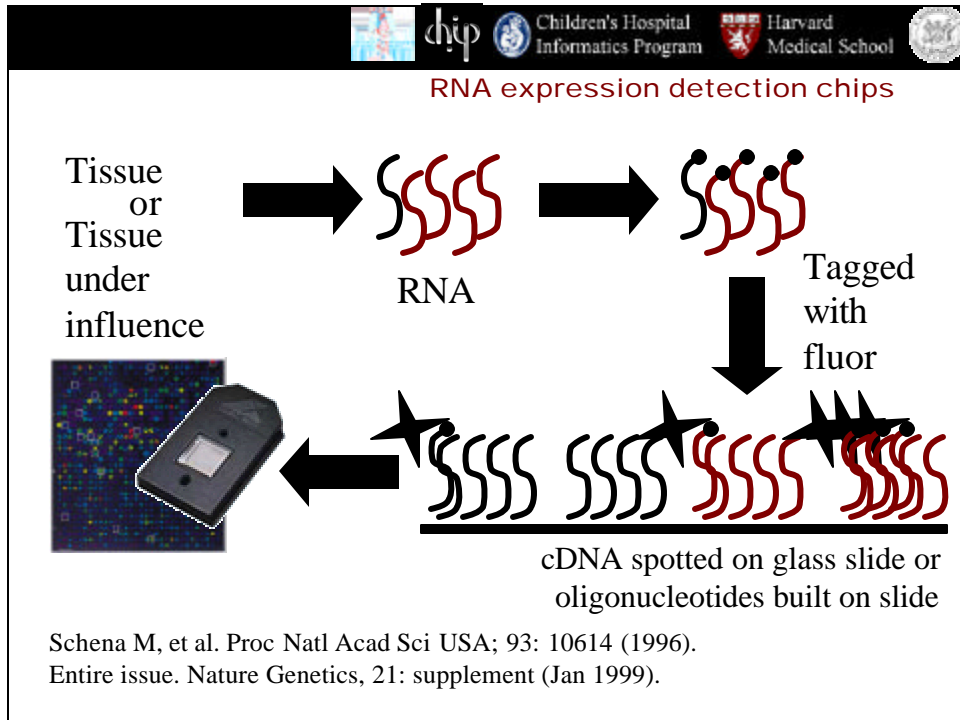
- If you are not looking for global patterns
- If you are not looking for "guilt" by association across a significant portion of the "-ome"
  - ✓ Then microarray technology may be inappropriate and misleading.
  - ✓ Why is that a problem?



## Decision-Theoretic Approach to the Solution

- Finding the low-hanging fruit
- Understanding the costs of false positives and negatives







Children's Hospital  
Informatics Program



Harvard  
Medical School



## Robotic Spotter

QuickTime? and a TIFF (Uncompressed) decompressor are needed to see this picture.



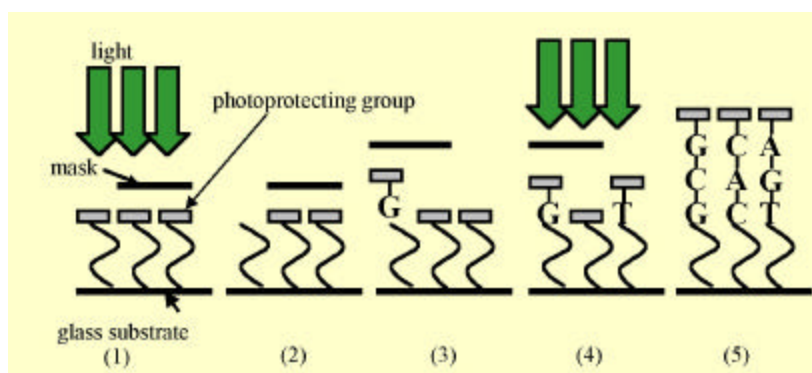
Children's Hospital  
Informatics Program

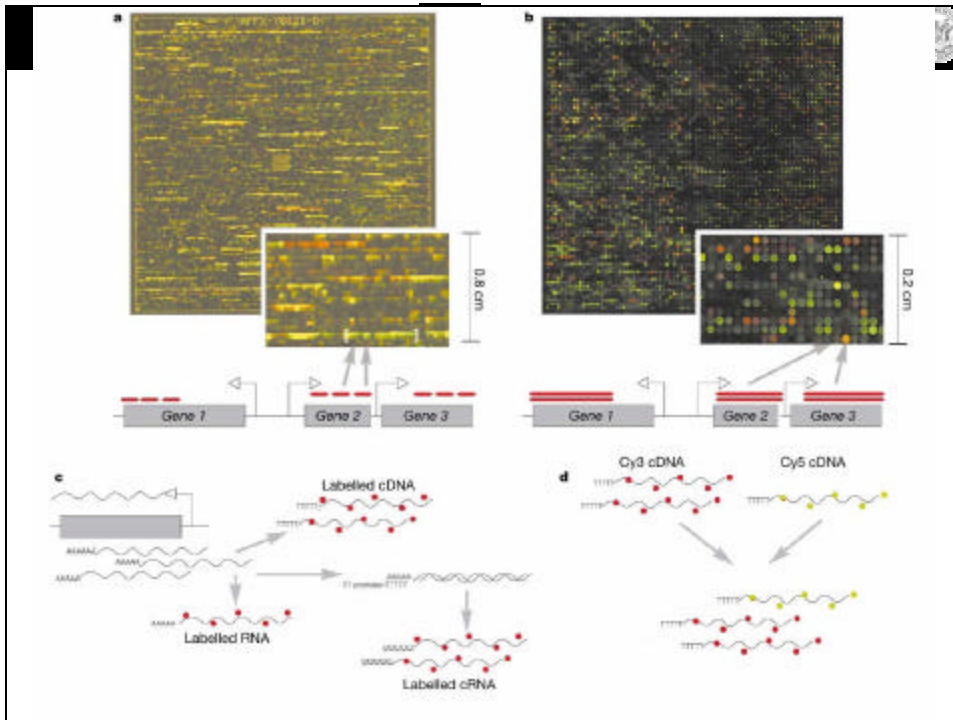





Harvard  
Medical School



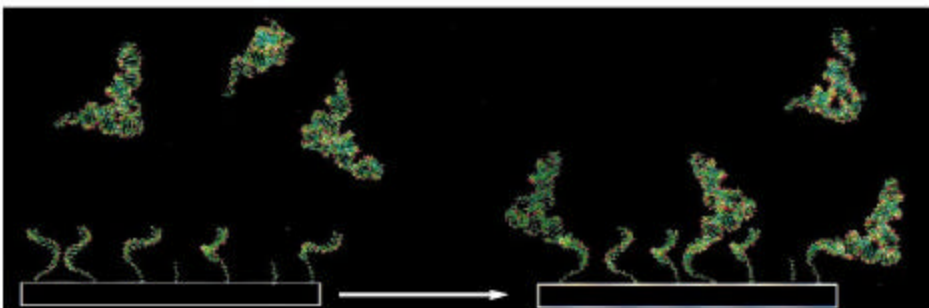
## Photolithography





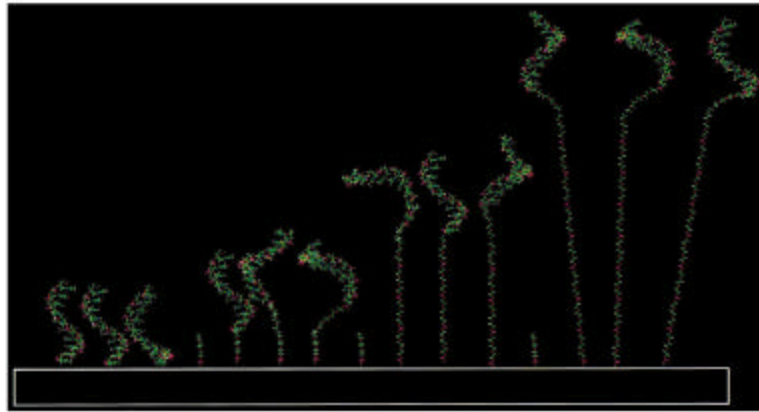




## Secondary structure of both probe and target



**Fig. 2** Long target sequences are likely to fold in on themselves as a result of intramolecular Watson-Crick base pairing. This structure hides parts of the target from the oligonucleotide probes. Large targets are also likely to be inhibited by their bulk from approaching the surface. Illustrated here is tRNA<sup>phe</sup> in solution, hybridizing to tethered decanucleotides.

Southern, Vekrellis, 1999

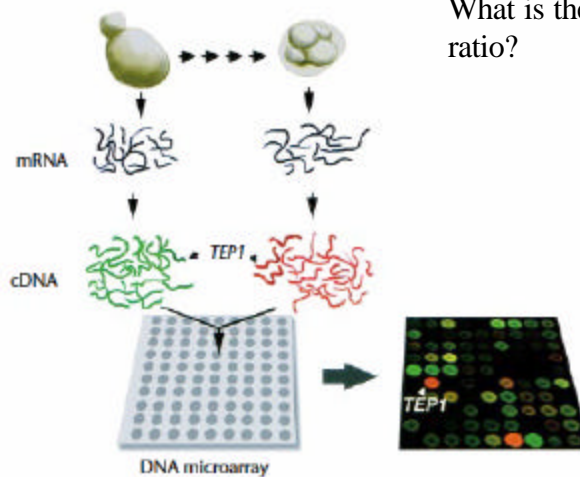


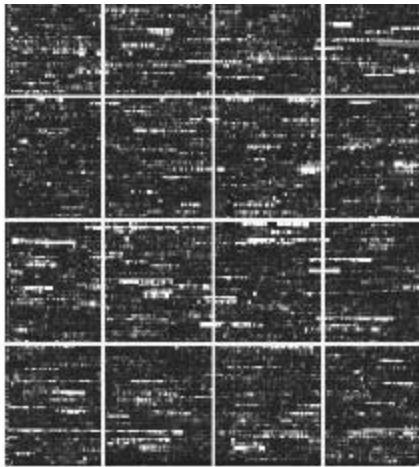
**Fig. 1** The density of oligonucleotides on the surface is approximately 10 pmol per mm<sup>2</sup> on aminated polypropylene, approximately 0.1 pmol per mm<sup>2</sup> on glass after ammonia deprotection—equivalent to one molecule per 39 square angstroms. The oligonucleotides are just about within reach of each other on glass, but rather closely packed on polypropylene supports. Spacers help to overcome steric interference, which can take a number of forms: the ends of the probes closest to the surface are less accessible than the ends furthest away; tethered molecules may crowd each other. Oligonucleotides on long spacers are better able to extend away from their neighbours and from the surface to allow interaction with the target. In this and other figures, the molecules are shown in a stretched conformation. It is likely that the molecules are in a dynamic state which may include this as one extreme, but in which the average state is somewhat more condensed. The linkers illustrated are oligoethylene glycols 26, 60 and 105 atoms in length.



## The ratio assay

What is the ratio?





Take the average of the lowest 2% of cell intensity values in this sector. This is the background of the sector. Subtract the Background from the average intensities of all cells in this sector.



## Calculated Background Noise

$$Q = 1/N \sum_{i \in \text{allbgcells}} \frac{\text{stdev}_i}{\sqrt{\text{pixel}}} * SF * NF$$

Noise for  
a given  
probe  
array

Total # of  
background  
cells -  
lowest 2%  
for each  
sector

Standard  
deviation of the  
intensities of the  
pixels making  
up background  
cell i

Total # of pixels  
in cell i

Scaling  
Factor

Normaliza-  
tion Factor



## Average Difference and Reliability

	Intensity			Intensity	
PM	 2,346	This probe pair is most likely <b>Positive</b> since the intensity of PM is much higher than MM	PM	 293	This probe pair is most likely <b>Negative</b> since the intensity of MM is much higher than PM
MM	 684		MM	 1,872	

ie significance is determined by calculating both the ratio (PM / MM) and the difference (PM - MM) associated with each probe pair. These values are then compared against two threshold values: the **Statistical Difference Threshold** (SDT) and the **Statistical Ratio Threshold** (SRT). This is expressed mathematically as follows:

A probe pair is **Positive** if:

1. PM - MM  $\geq$  SDT

And

2. PM / MM  $\geq$  SRT

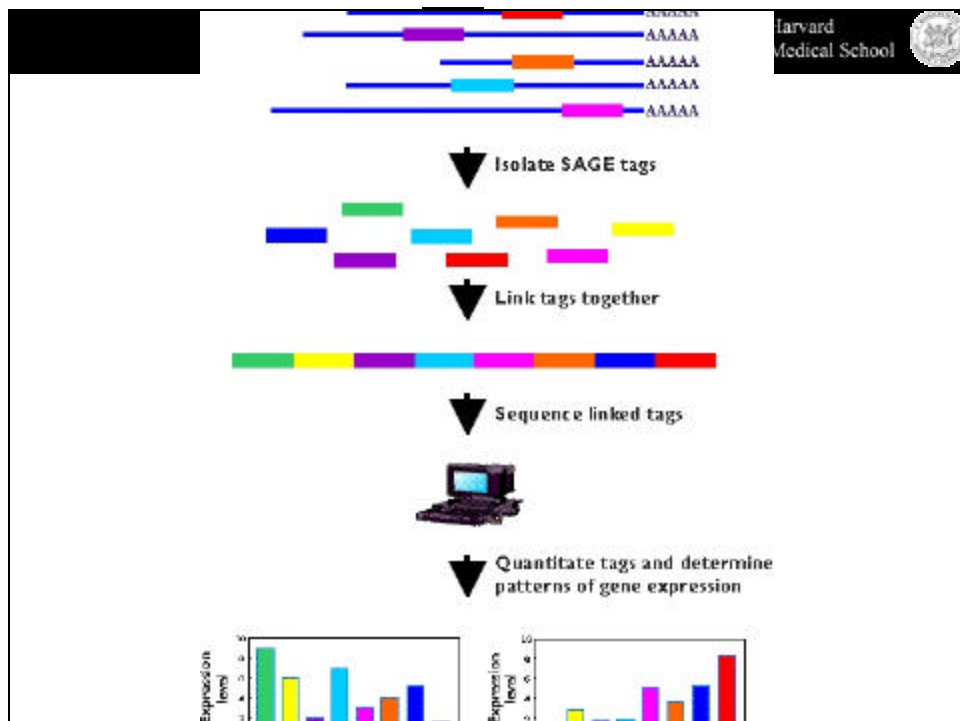
A probe pair is **Negative** if:


1. MM - PM  $\geq$  SDT


And


2. MM / PM  $\geq$  SRT

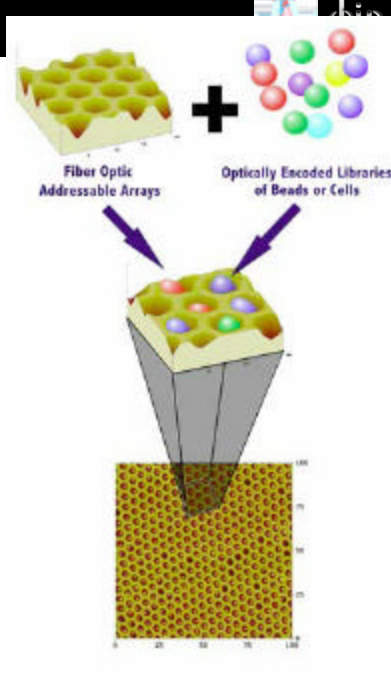
Note: not all probe pairs will be scored as Positive or Negative.




Children's Hospital  
Informatics Program


Harvard  
Medical School








Fiber Optic Addressable Arrays

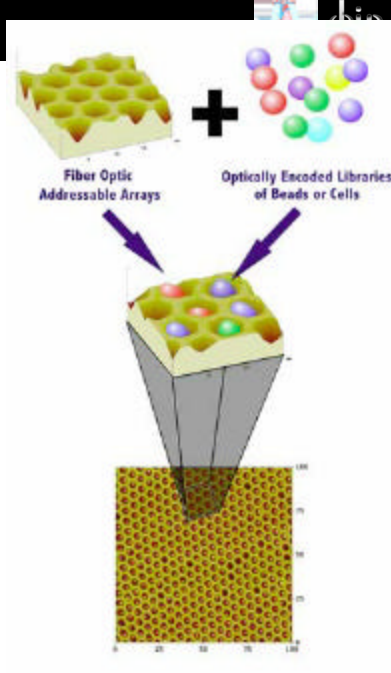
Optically Encoded Libraries of Beads or Cells

Other massively parallel expression measurement technologies coming...


Children's Hospital  
Informatics Program


Harvard  
Medical School





Fiber Optic Addressable Arrays

Optically Encoded Libraries of Beads or Cells

Other massively parallel expression measurement technologies coming...





## Inkjet technology

- Speed
  - ✓ On-the-fly, non-contact printing
  - ✓ Fast turnaround time (FTAT)
- Flexibility
  - ✓ *In situ* synthesis and deposition
  - ✓ Customizable
  - ✓ Hypothesis driven array design
- Quality
  - ✓ Feature control: positioning and size



## Inkjet Arrays: Quality and Flexibility

	# TIJ Fires	Feature Sizes (mm)	Inkjet
	1	70	
	2	90	
	3	90x110	
	4	100x130	
			Pins

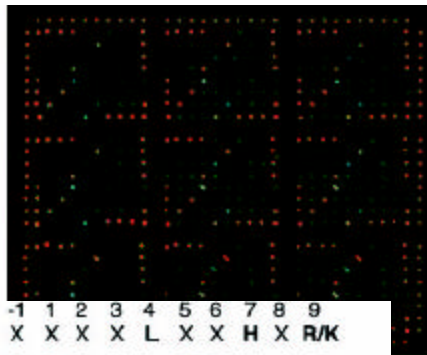
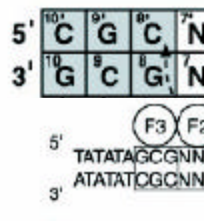




QuickTime?and a TIFF (Uncompressed) decompressor are needed to see this picture.

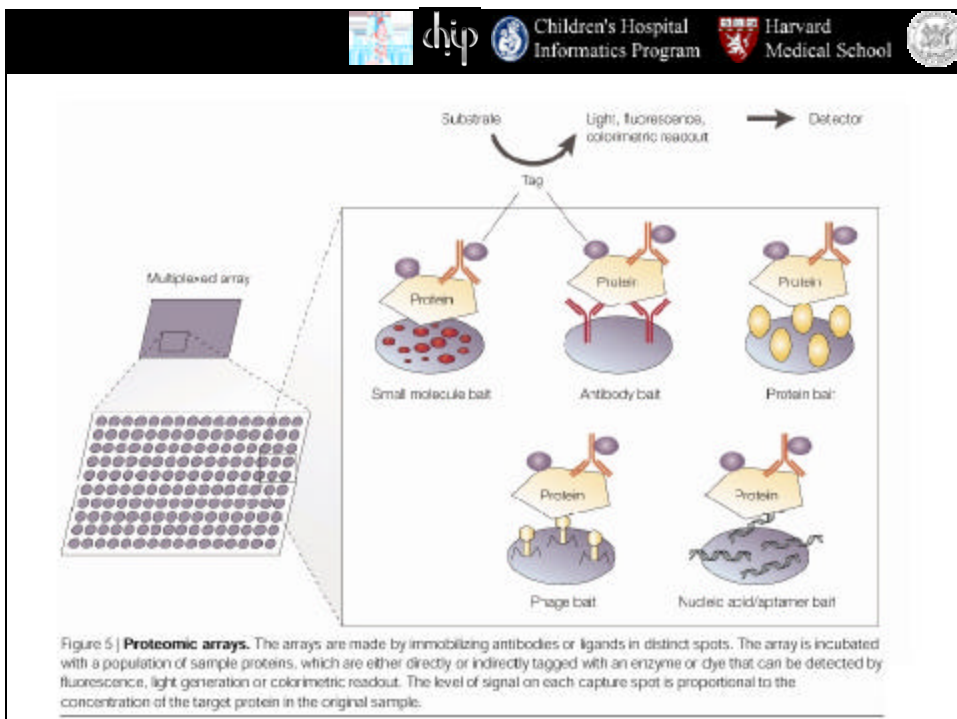
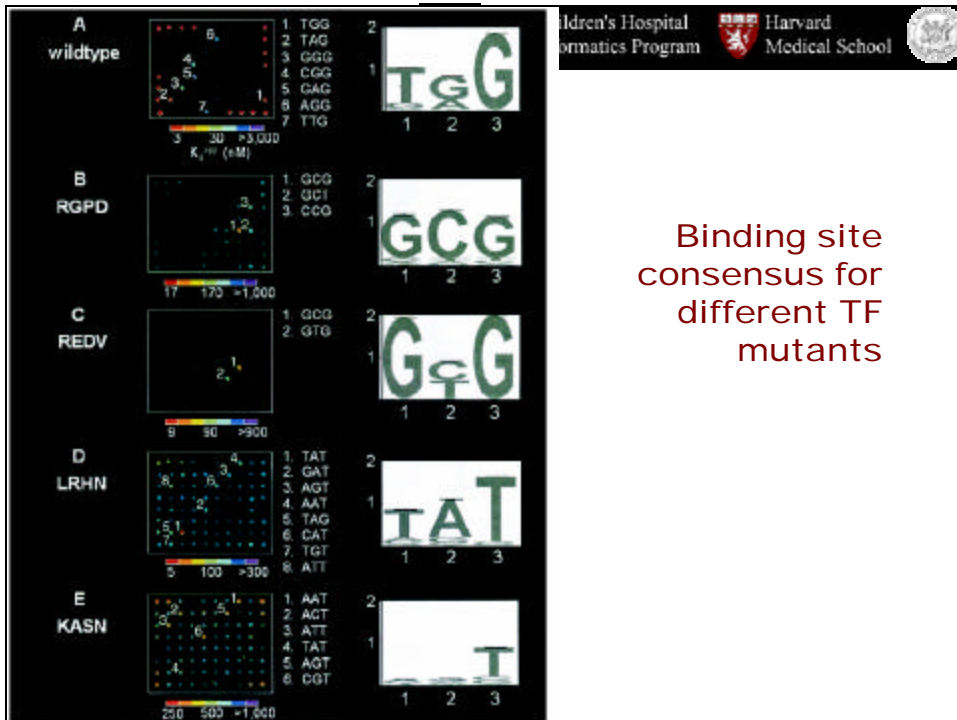


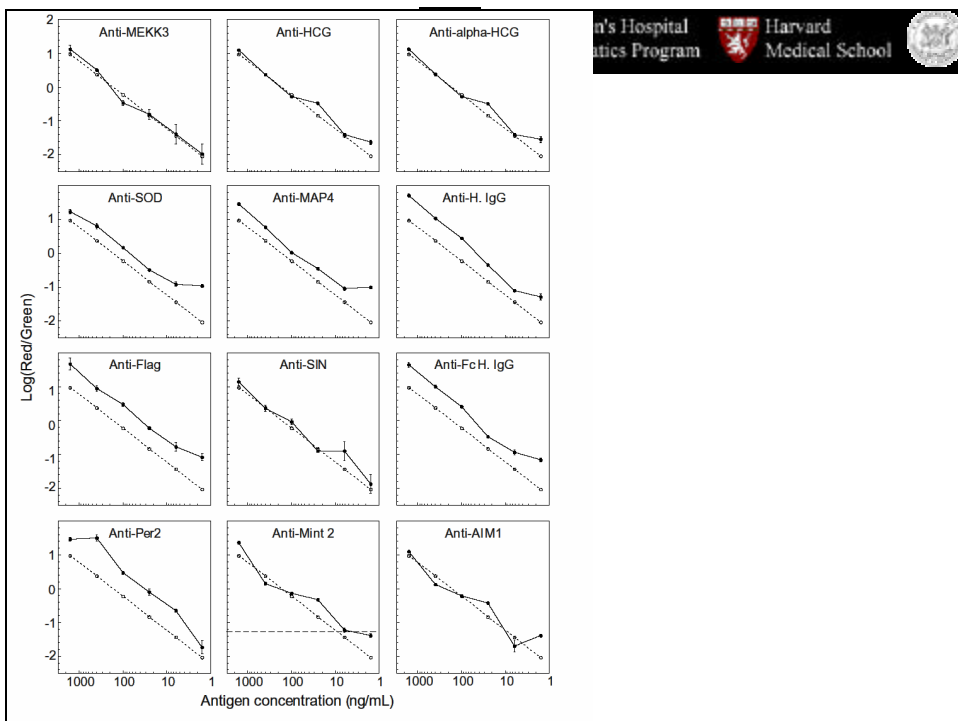
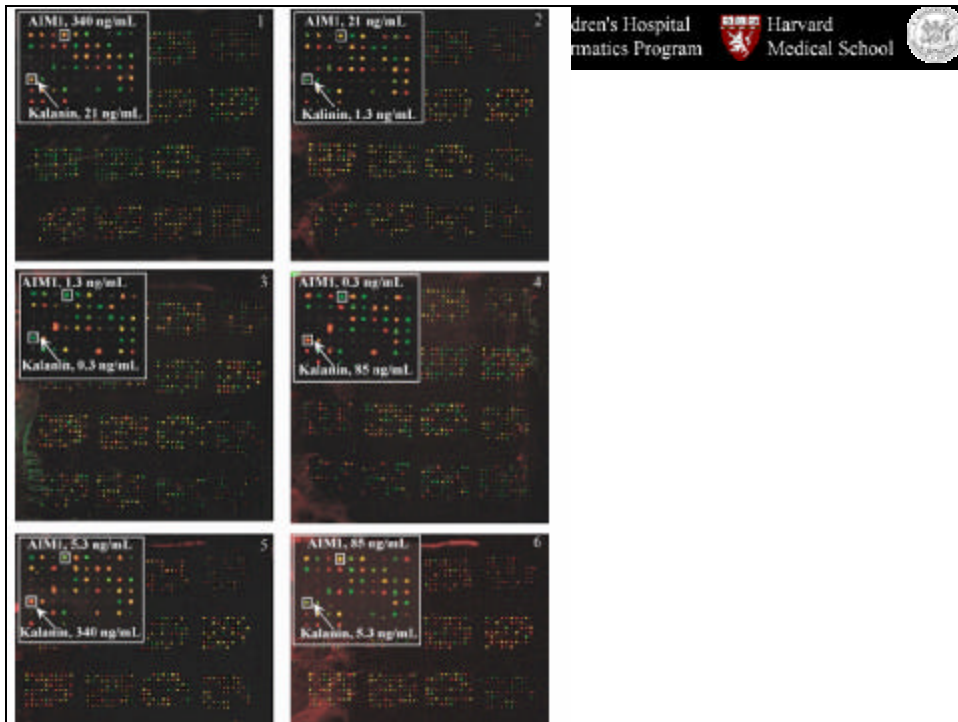
A.



wildtype  
RGPD  
REDV  
LRHN  
KASN

-1	1	2	3	4	5	6	7	8	9
X	X	X	X	L	X	X	H	X	R/K
R	S	D	H	L	T	T	H	I	R
R	G	P	D	L	A	R	H	G	R
R	E	D	V	L	I	R	H	G	K
L	R	H	N	L	E	T	H	M	R
K	A	S	N	L	V	S	H	I	R





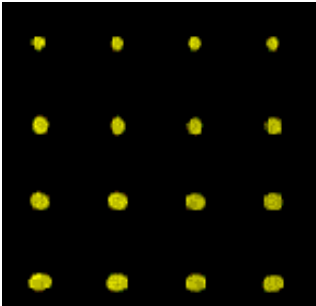
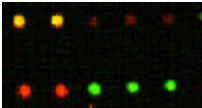
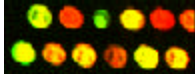


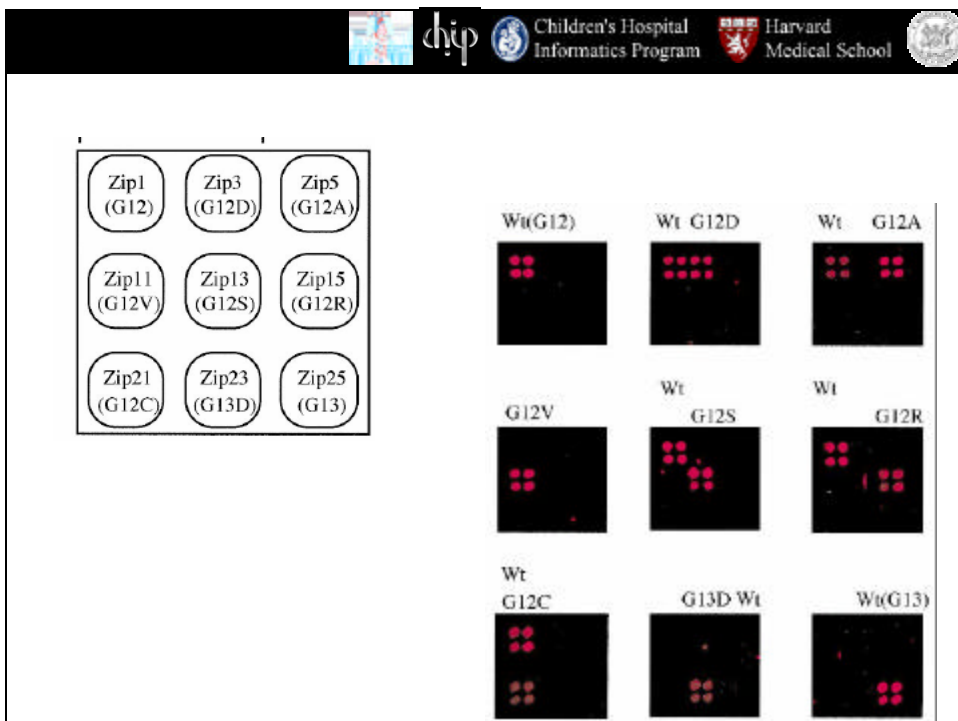
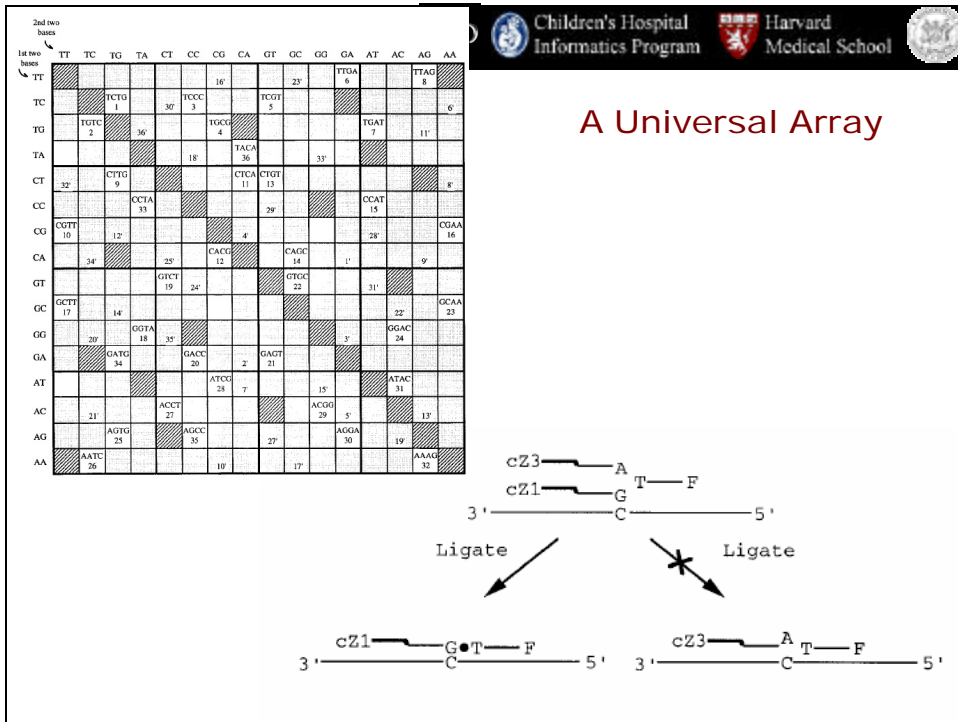
## Inkjet technology

- Speed
  - ✓ On-the-fly, non-contact printing
  - ✓ Fast turnaround time (FTAT)
- Flexibility
  - ✓ *In situ* synthesis and deposition
  - ✓ Customizable
  - ✓ Hypothesis driven array design
- Quality
  - ✓ Feature control: positioning and size



## Inkjet Arrays: Quality and Flexibility

	# TIJ Fires	Feature Sizes (mm)	Inkjet
	1	70	
	2	90	
	3	90x110	
	4	100x130	
			Pins
			



[illegible]

- Indeed, what is there that does not appear fabulous when it comes to our knowledge for the first time? How many things, too, are looked upon as quite impossible until they have been actually effected?
  - ✓ Pliny the Elder (23 AD - 79 AD), Natural History





## Are Functional Genomicists Plinians?

“They would observe a phenomenon with skill that comes from experience and write highly detailed notes about what they saw. They did not usually employ a method in which they formally tested anything.”





## Imminent Genomic Plinian Eruption?

- Over-promising.
- Slowness to acknowledge the limitations of our measurement techniques.
- The challenge of linking genomic data to biological and clinical significance.
- The lack of formal hypothesis testing.
- The lack of sufficient multidisciplinary expertise
- Are functional genomicists over-promising?



## Intervention fold-differences

	RNA Expr Gene 1	RNA Expr Gene 2	RNA Expr Gene 3
Before Intervention	0.7	0.3	7.3
After Intervention	1.2	2.1	7.4

### Fold Difference

Gene 2	7.0
Gene 1	1.7
Gene 3	1.0





Children's Hospital  
Informatics Program



Harvard  
Medical School



## What are we to make of small FD?

- Starvation and longevity in rats (Lee, Science, 1999)

AA062328	↓ 3.4	DnaJ Homolog 2	Chaperone
X63023	↓ 1.9	Cytochrome P-450-III A	Detoxification
U03283	↓ 1.8	Cyp1b1 Cytochrome P450	Detoxification
U14390	↓ 1.8	Aldehyde Dehydrogenase-3	Detoxification
X76850	↓ 1.8	MAPKAP2	Unknown
D26123	↓ 1.7	Carbonyl Reductase	Detoxification
L4406	↓ 1.7	Hsp105-beta	Chaperone
U40930	↓ 1.5	Oxidative Stress-Induced Protein	Unknown
U66887	↓ 1.8	RAD50	Double strand break repair
AA059718	↓ 1.7	DNA Polymerase Beta	Base excision repair
W42234	↓ 1.6	XPE	Nucleotide excision repair
D43694	↓ 1.8	Math-1	Differentiation
D16464	↓ 1.7	HES-1	Differentiation
W13191	↓ 1.6	Thyroid Hormone Receptor Alpha-2	Thyroid hormone receptor



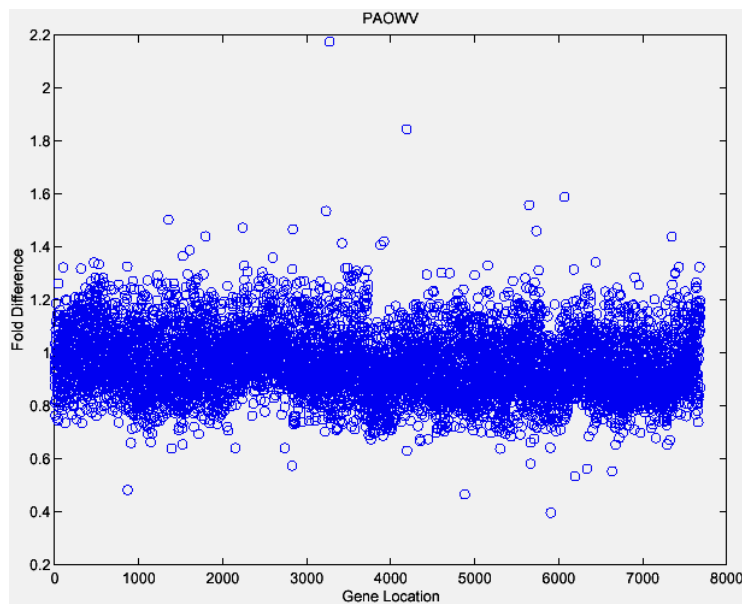
Children's Hospital  
Informatics Program



Harvard  
Medical School

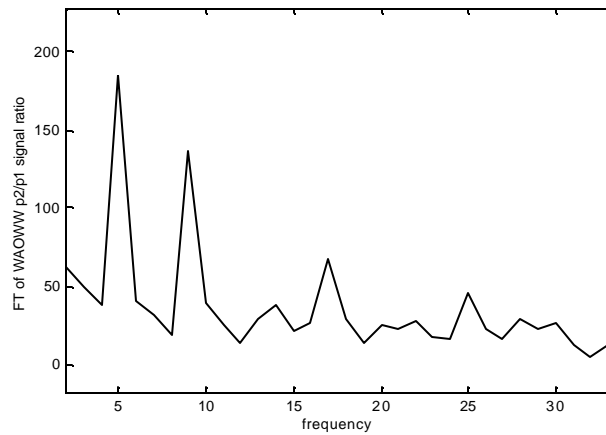


Positional noise





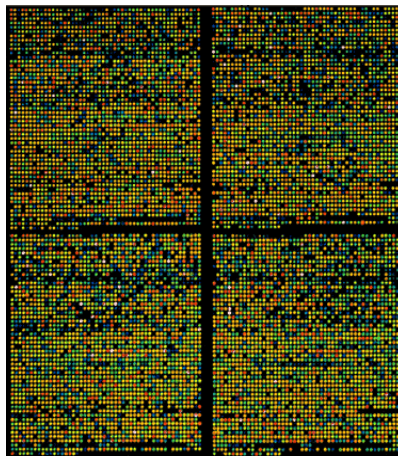
## Frequency content of array



Why?



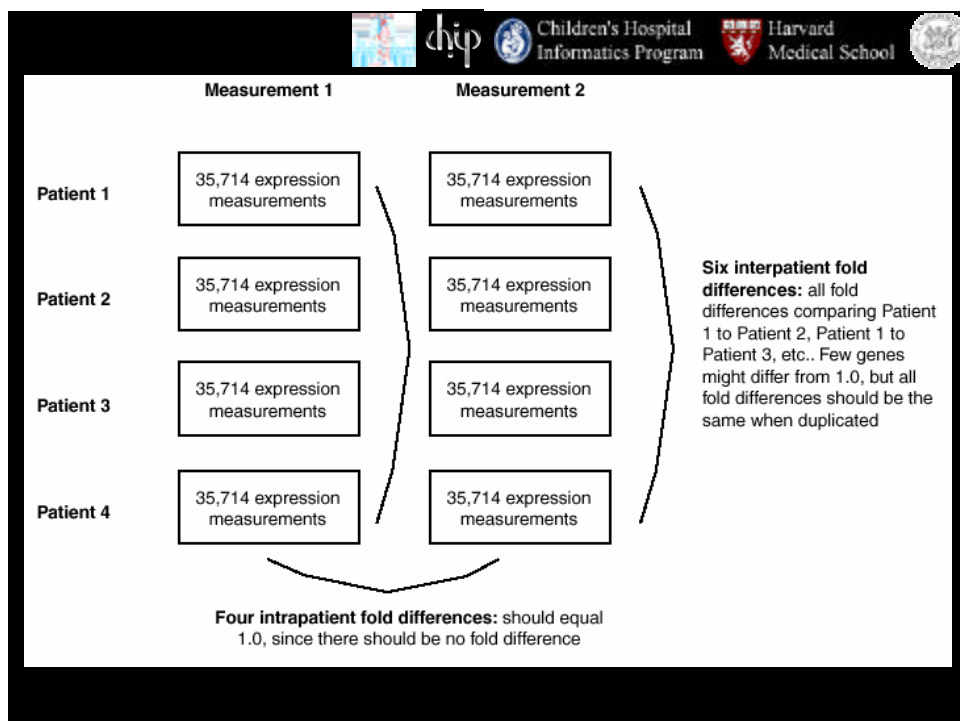
## Incye's GEM Microarray





## Case History

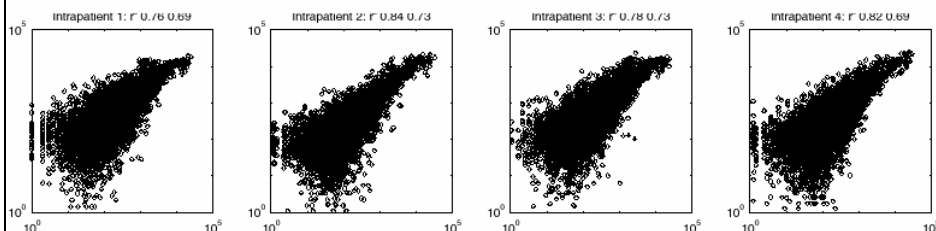
- Four different patients with glucose intolerance, without diabetes
- RNA from each muscle biopsy sample placed on two sets of Affymetrix Hu35K microarrays
- **Goal:** find genes differentially expressed between patients





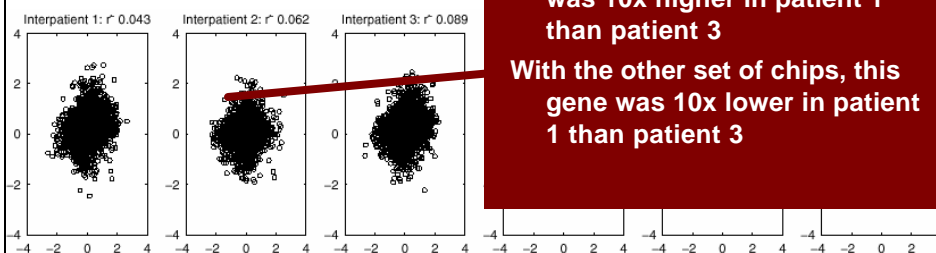
## Seemingly high reproducibility

- Correlation coefficients of **intrapatient** repeated measurements ranged from 0.76 to 0.84
- Dropped to 0.69 to 0.73 with log transformation



## Poor reproducibility

- Very poor correlation coefficients for **interpatient** fold differences
- Ranged from 0.01 to 0.09



With one set of chips, this gene was 10x higher in patient 1 than patient 3

With the other set of chips, this gene was 10x lower in patient 1 than patient 3

- Restricting to "present" calls is not enough
- If the entire experiment had not been reproduced, we would never know the fold differences were in question



### Why does this happen?

$$\frac{0.6}{0.3} = 2 \text{ fold up}$$

$$\frac{600}{300} = 2 \text{ fold up}$$

± 0.3 noise

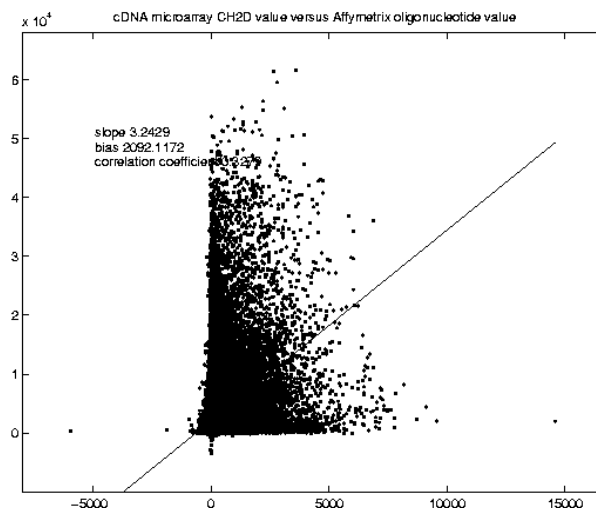
$$\frac{0.3}{0.6} = 2 \text{ fold down}$$

$$\frac{599.7}{300.3} \approx 2 \text{ fold up}$$

*PSB 2001*



### Lack of Platform Comparability



*Bioinformatics  
2001*

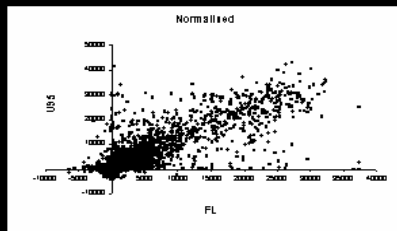
60 Cell Lines (NCI-60) Oligonucleotide vs cDNA



## Poor Correlation Across Different Generations of the Same Platform

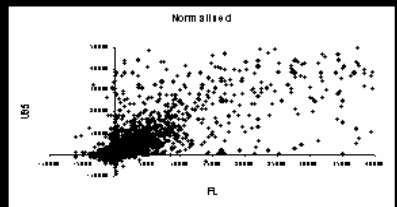
HuGeneFL  
vs  
U95A

Sample I



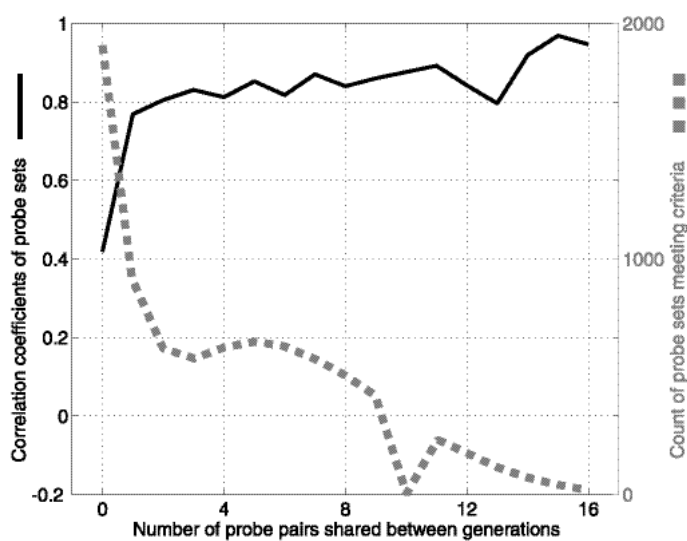
$R^2 = 0.70$

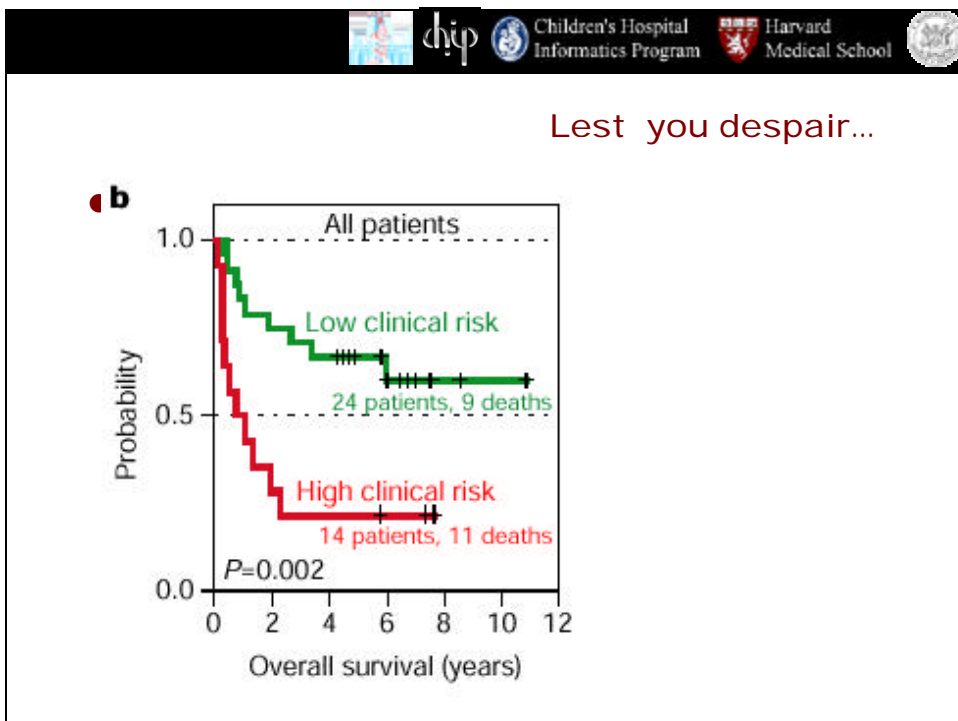
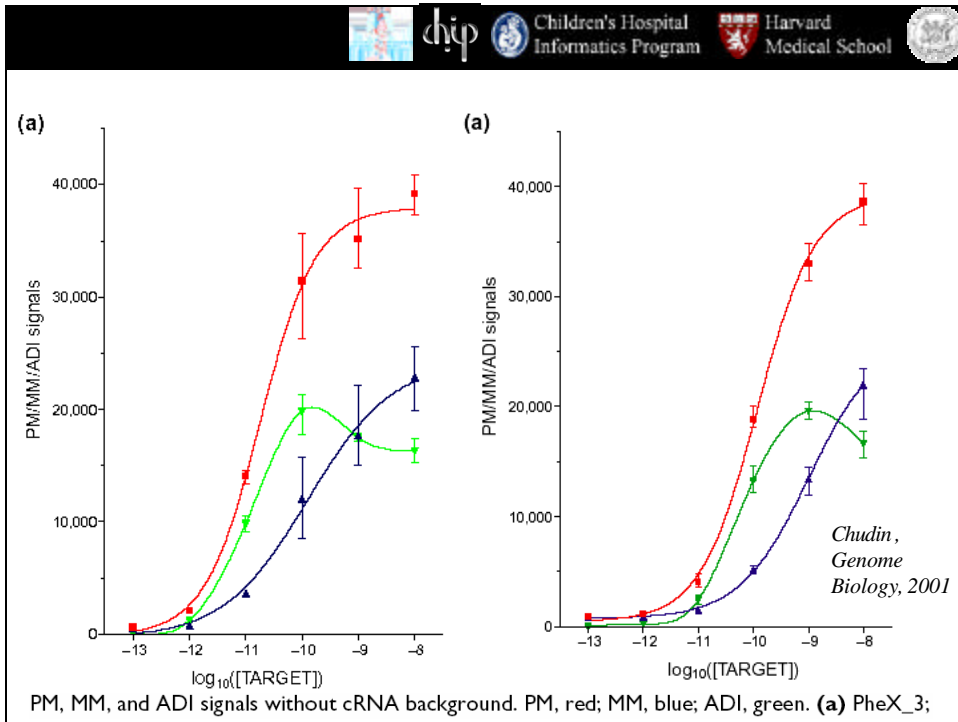
Sample II



$R^2 = 0.59$

*Under review*

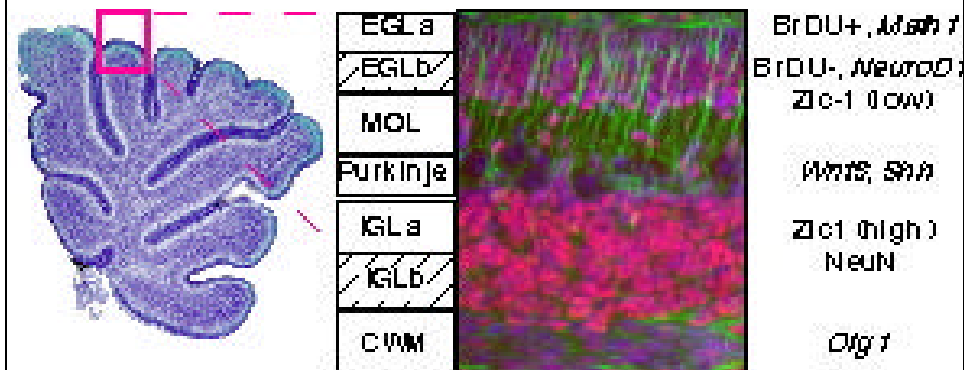




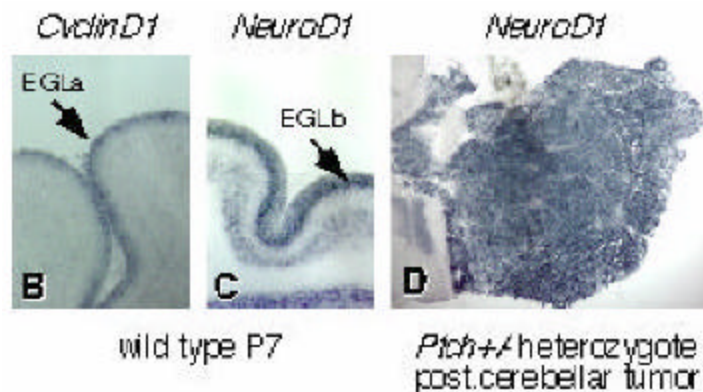


## Finding the Needle in the Haystack: A Case History

- Cerebellum has pivotal roles in the coordination of posture and locomotion
- Laminar organization of the cerebellar cortex has facilitated understanding its basic circuitry, functions and ontogeny



## Sonic Hedgehog (Shh), development and tumorigenesis





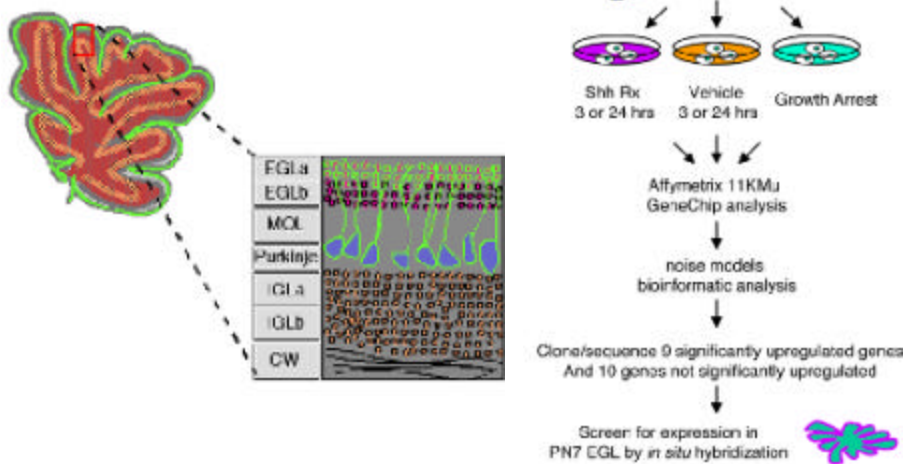


## The challenge

- Find novel members (the needles) of the Shh pathway in a large haystack.
- The haystack
  - ✓ Large number of probes
  - ✓ Whole cerebellum
    - ☞ (whole organ, multiple cell types)
- But the signal of interest is confined to thin superficial layer
- Can we find the signal in time and space?

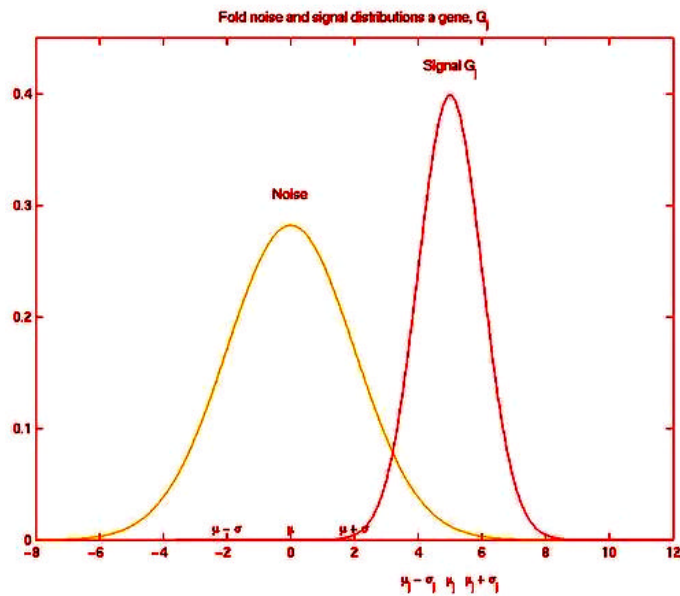


## Overall schema





## A noise model with triplicate measurements



## What average are we calculating?

$S_1 S_2 S_3$

$V_1 V_2 V_3$

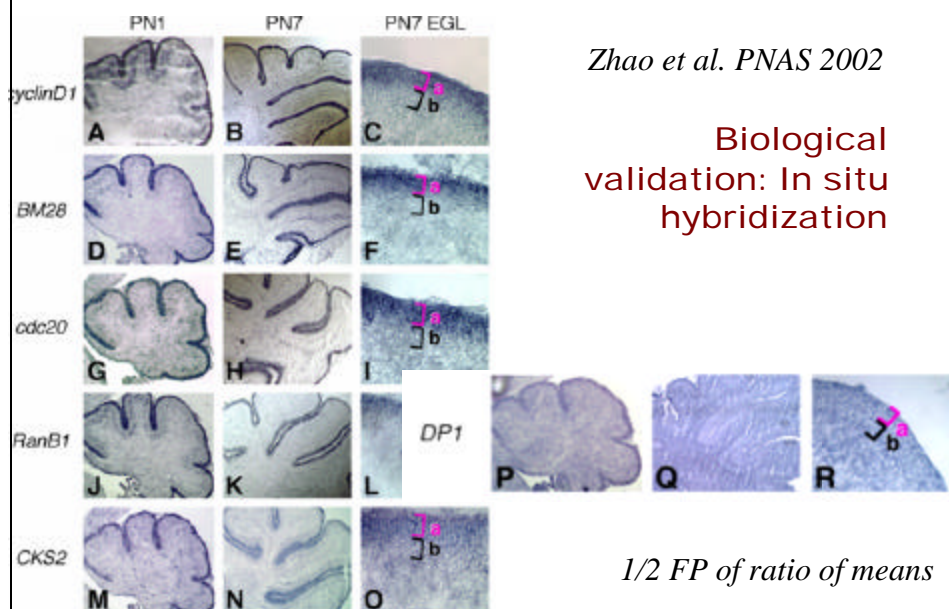
$$\frac{\frac{1}{3} \sum_{i=1}^3 S_i}{\frac{1}{3} \sum_{i=1}^3 V_i}$$



## How do you calculate average interest rate?

- If 4 different interest rates over four years
- $(1+r1)(1+r2)(1+r3)(1+r4)P \rightarrow Q^4P$
- $Q=$

$$\sqrt[4]{(1+r1)(1+r2)(1+r3)(1+r4)}$$





## Relevance to Human Disease

- How can we leverage this developmental view of the mouse?
- Human medulloblastoma microarray data
  - ✓ Pomeroy et al, Nature 2002
- Find the mouse homologues of the genes up and down regulated in the tumors.
- Principal Component Analysis to find the main sources of variance in the developmental time series

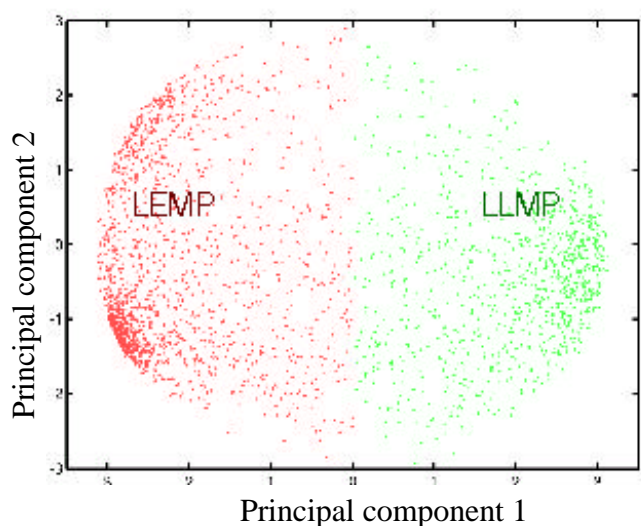
... Developmental Time Series ...

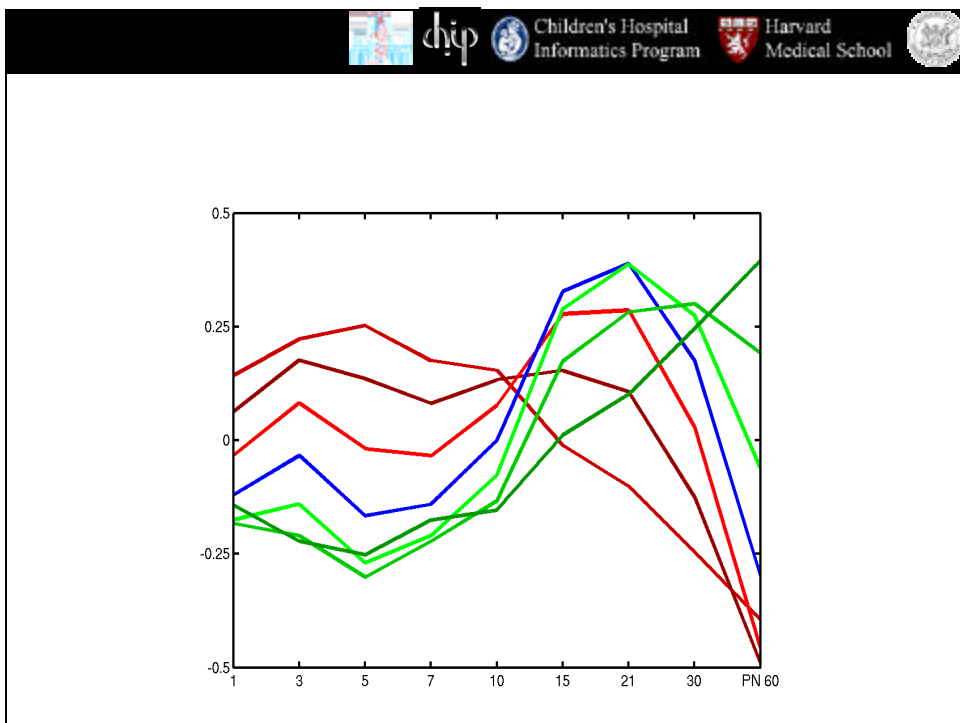
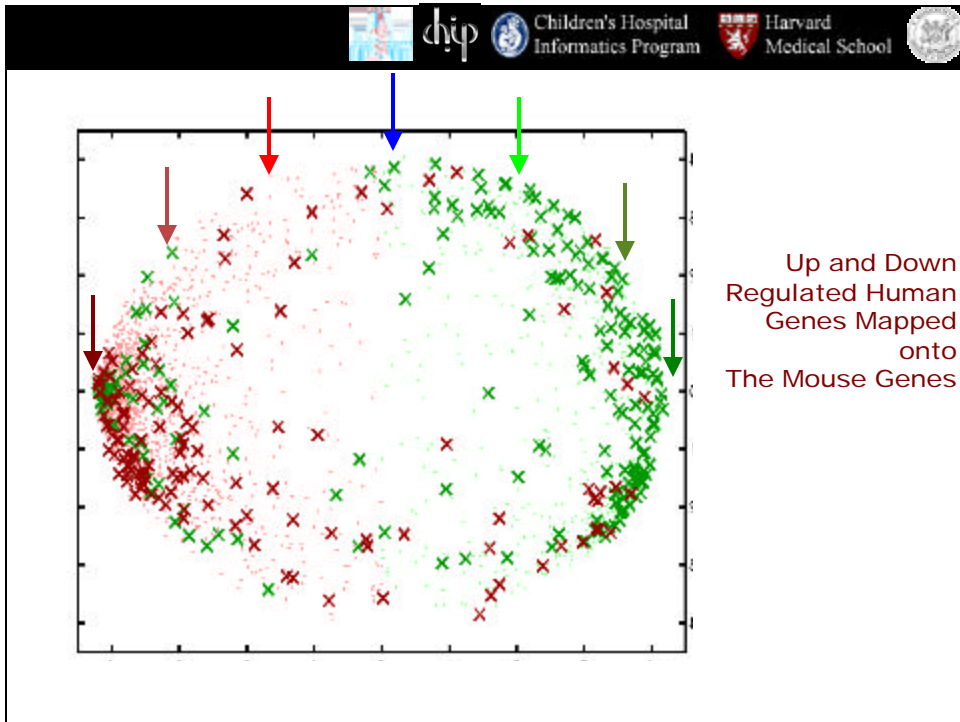


QuickStart v. 1.0  
See the QuickStart document  
at <http://www.chip.org>



## Mouse Cerebellar time course by first two principal components

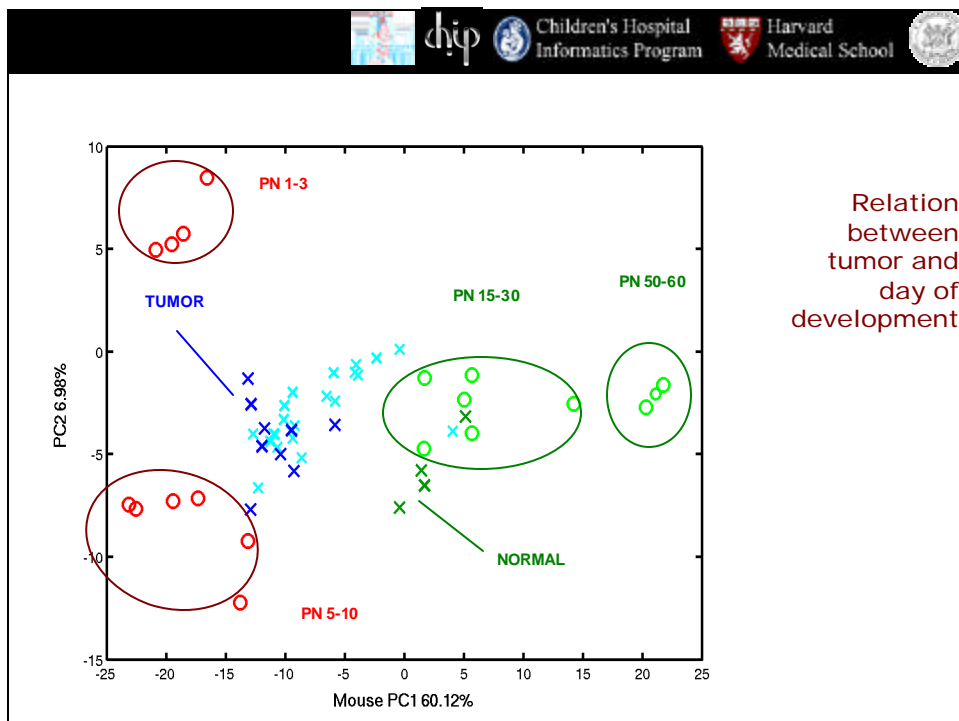






## The New Histopathology and History Repeated

- Lobstein (1829) and Cohnheim (1887) were amongst the first to theorize similarities between human embryogenesis and the biology of cancer cells
- The brain tumor classification system devised by Bailey and Cushing in 1926, from which modern taxonomies derive, emphasizes histological resemblance to cells of the developing central nervous system





## Summary

- Microarrays provide high throughput snapshot of the 'ome
- Expression arrays are the most mature and still very noisy
- Much remains to be done in both measurement technology and analysis
- Careful use of the data permits
  - ✓ Generation of a natural classification of disease
  - ✓ With further insight into mechanism