

Talks today

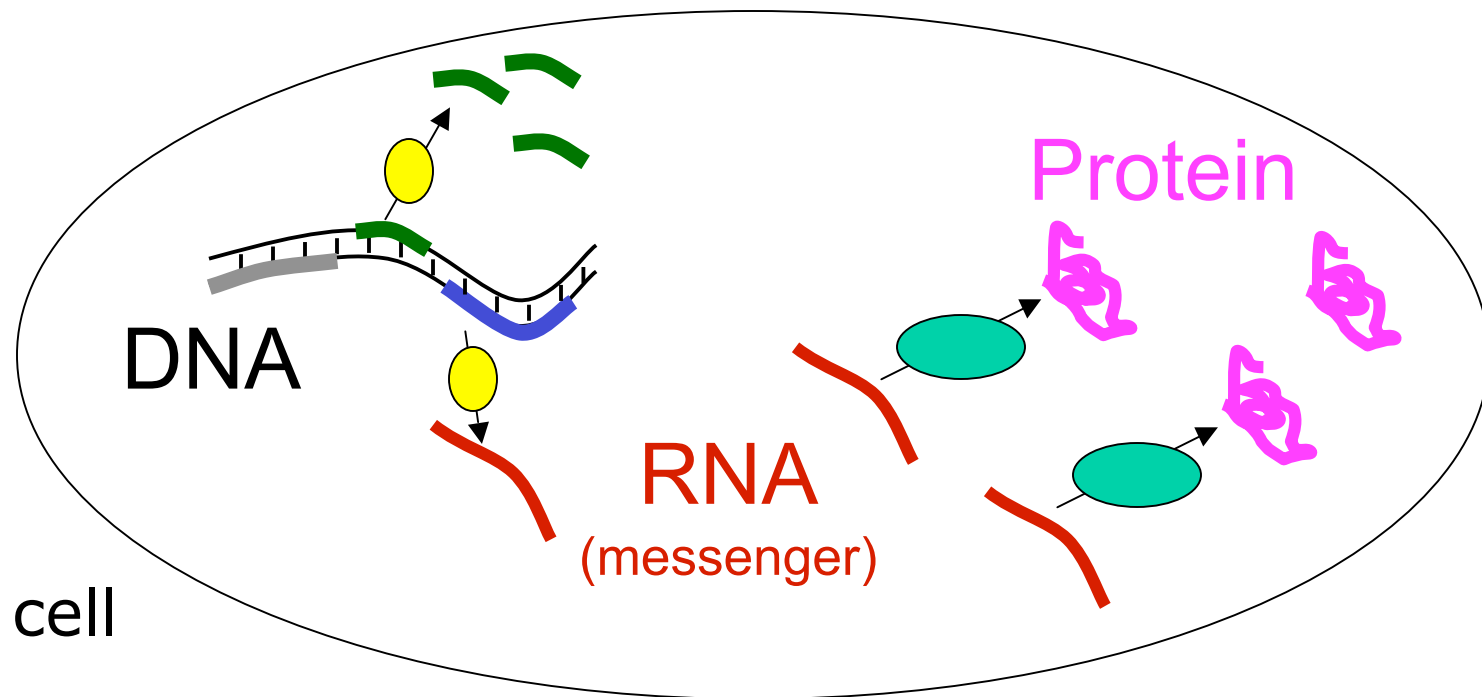
- **Dr. Phil Green** , Professor of Genome Sciences, University of Washington "Finishing the Gene-ome: Computationally Directed Gene Structure Verification in *C. elegans*" **1:30 in Health Sciences K-069**
- **Dr. Mark Chee** , Vice President of Genomics, Illumina, Inc. "Accessing Genetic Information: Technology for Large Scale SNP Genotyping" **3:30 in Hitchcock Hall 132**

DNA Microarray Overview

(Some slides from Dr. Holly Dressman, Duke University
http://genome.genetics.duke.edu/STAT_talk_301.ppt)

Gene Expression: The “Central Dogma”

DNA → RNA → Protein

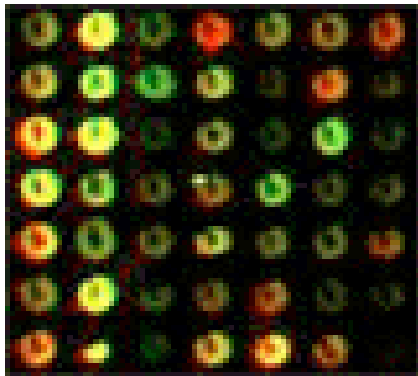


Gene Expression

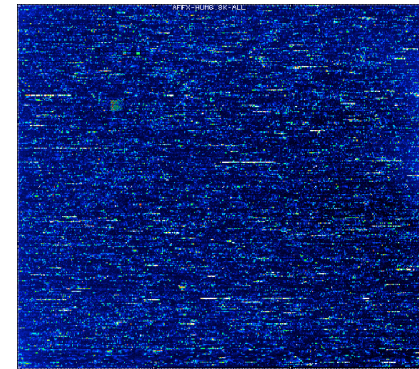
- Proteins do most of the work
- They're dynamically created/destroyed
- So are their mRNA blueprints
- Different mRNAs expressed at different times/places
- Knowing mRNA “expression levels” tells a lot about the state of the cell

Microarrays

A snapshot that captures the activity pattern of thousands of genes at once. [



Custom spotted arrays

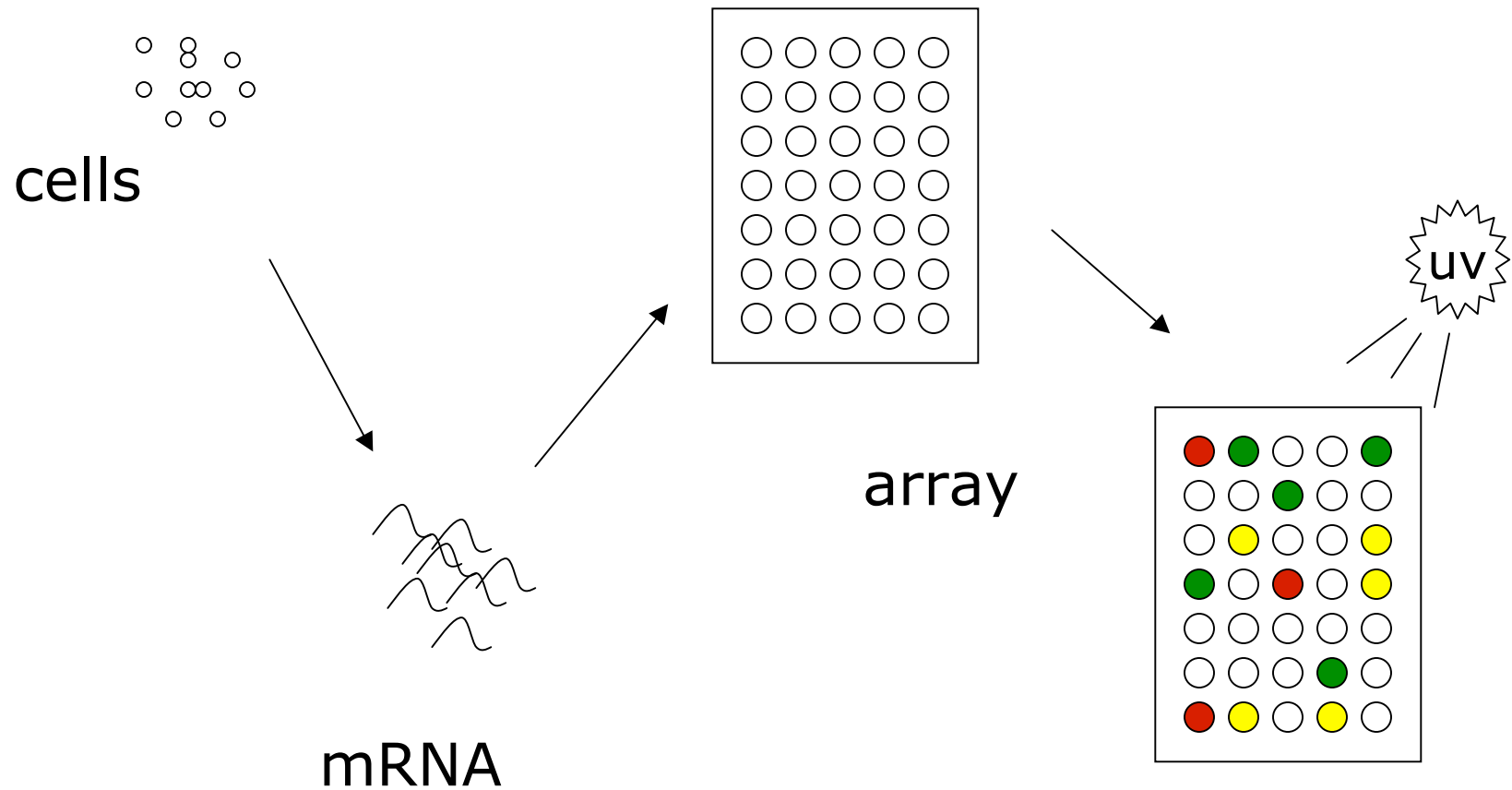


Affymetrix GeneChip

Expression Microarrays

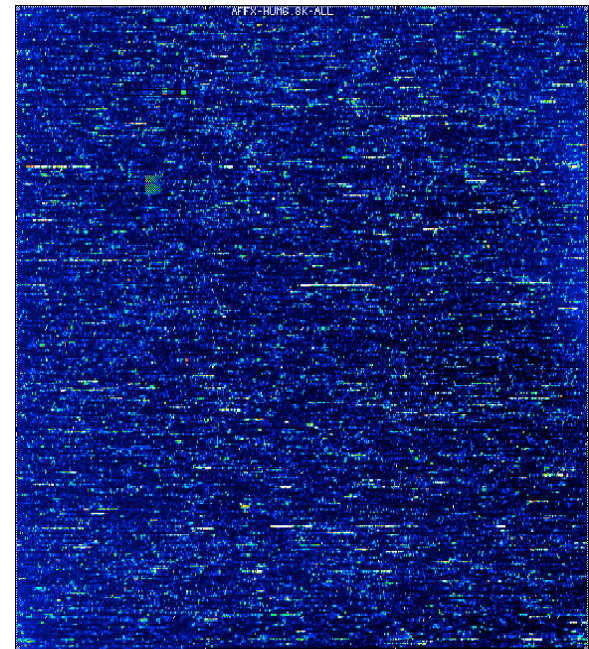
- Thousands to hundreds of thousands of spots per square inch
- Each holds millions of copies of a DNA sequence from one gene
- Take mRNA from cells, put it on array
- See where it sticks – mRNA from gene x should stick to spot x

An Expression Array Experiment



An Example Application

- 72 leukemia patients
 - 47 ALL
 - 25 AML
- 1 chip per patient
- 7132 human genes per chip



Golub, et al., Science 286:531-537 (1999).

Key Issue: What's Different?

- What genes are behaving differently between ALL & AML (or other disease/normal states)?
- Potential uses:
 - Diagnosis
 - Prognosis
 - Insight into underlying biology/biologies
 - Treatment

A Classification Problem

- Given an array from a new patient: is it ALL or AML?
- Many possible approaches:
LDA, logistic regression, NN, SVM, ...
- Problems:
 - Noise
 - Dimensionality

Other Applications

- Study gene function & regulation
 - Covarying $\sim\sim$ > coregulated?
 - Covarying $\sim\sim$ > common pathway?
- Refined categorization of diseases
 - E.g., "prostate cancer" is almost certainly not one disease. Are subtypes distinguishable at expression level?



Practical Applications of Microarrays

Gene Target Discovery

By allowing scientists to compare diseased cells with normal cells, arrays can be used to discover sets of genes that play key roles in diseases. Genes that are either overexpressed or underexpressed in the diseased cells often present excellent targets for therapeutic drugs.

Pharmacology and Toxicology

Arrays can provide a highly sensitive indicator of a drug's activity (pharmacology) and toxicity (toxicology) in cell culture or test animals. This information can then be used to screen or optimize drug candidates prior to launching costly clinical trials.

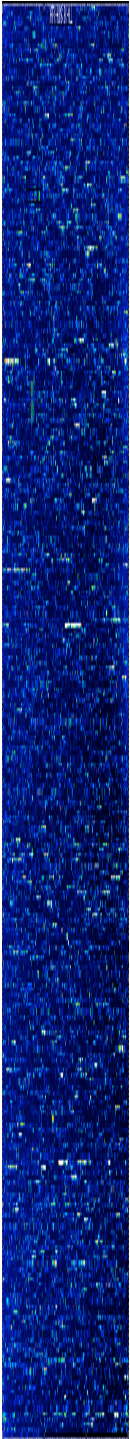
Diagnostics

Array technology can be used to diagnose clinical conditions by detecting gene expression patterns associated with disease states in either biopsy samples or peripheral blood cells.

Microarray Technologies

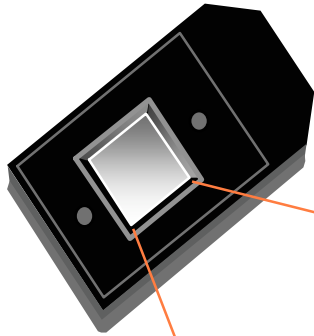
- Oligo Arrays
 - Affymetrix -
 - one color
 - Short oligos
 - match/mismatch
 - Agilent, inter alia
 - 2 color
 - Longer oligos
- Spotted cDNA arrays

GeneChip® Probe Array



GeneChip[®] Probe Arrays

GeneChip Probe Array



1.28cm

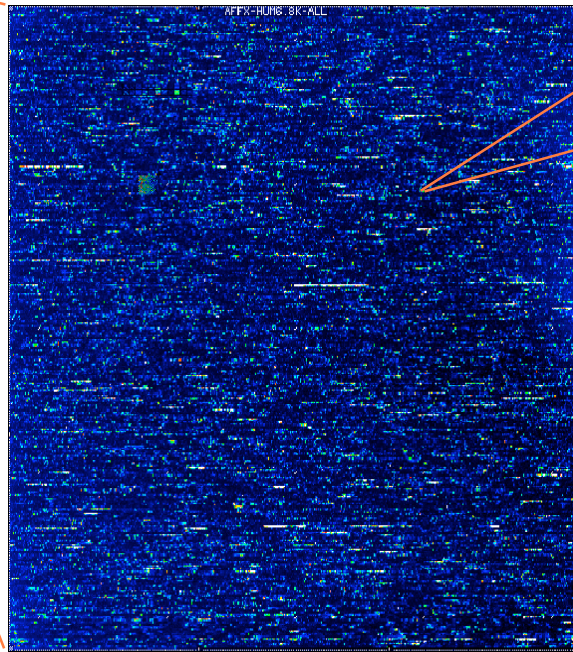
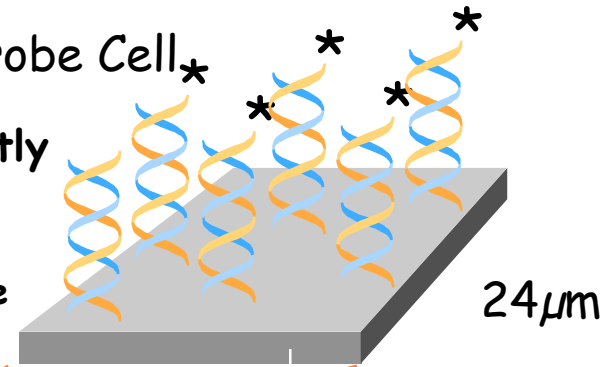


Image of Hybridized Probe Array

Hybridized Probe Cell*

Single stranded, fluorescently labeled DNA target

Oligonucleotide probe



Each probe cell or feature contains millions of copies of a specific **oligonucleotide** probe

Over 250,000 different probes complementary to genetic information of interest

How unique is a 20-mer?

- VERY CRUDE model: DNA is random—every position is equally likely to be A, C, G, or T, independent of every other
- Then probability of a random 20-mer is

$$\begin{array}{|c|} \hline 1 \\ \hline \\ \hline 4 \\ \hline \end{array}^{20} = \begin{array}{|c|} \hline 1 \\ \hline \\ \hline 2 \\ \hline \end{array}^{40} = \begin{array}{|c|} \hline 1 \\ \hline \\ \hline 2 \\ \hline \end{array}^{10} \begin{array}{|c|} \hline \\ \hline \\ \hline \\ \hline \end{array}^4 = \begin{array}{|c|} \hline \\ \hline \\ \hline \\ \hline \end{array} \frac{1}{1024} \begin{array}{|c|} \hline \\ \hline \\ \hline \\ \hline \end{array}^4 \times \begin{array}{|c|} \hline \\ \hline \\ \hline \\ \hline \end{array}^{10^3} \begin{array}{|c|} \hline \\ \hline \\ \hline \\ \hline \end{array}^4 = 10^{-12}$$

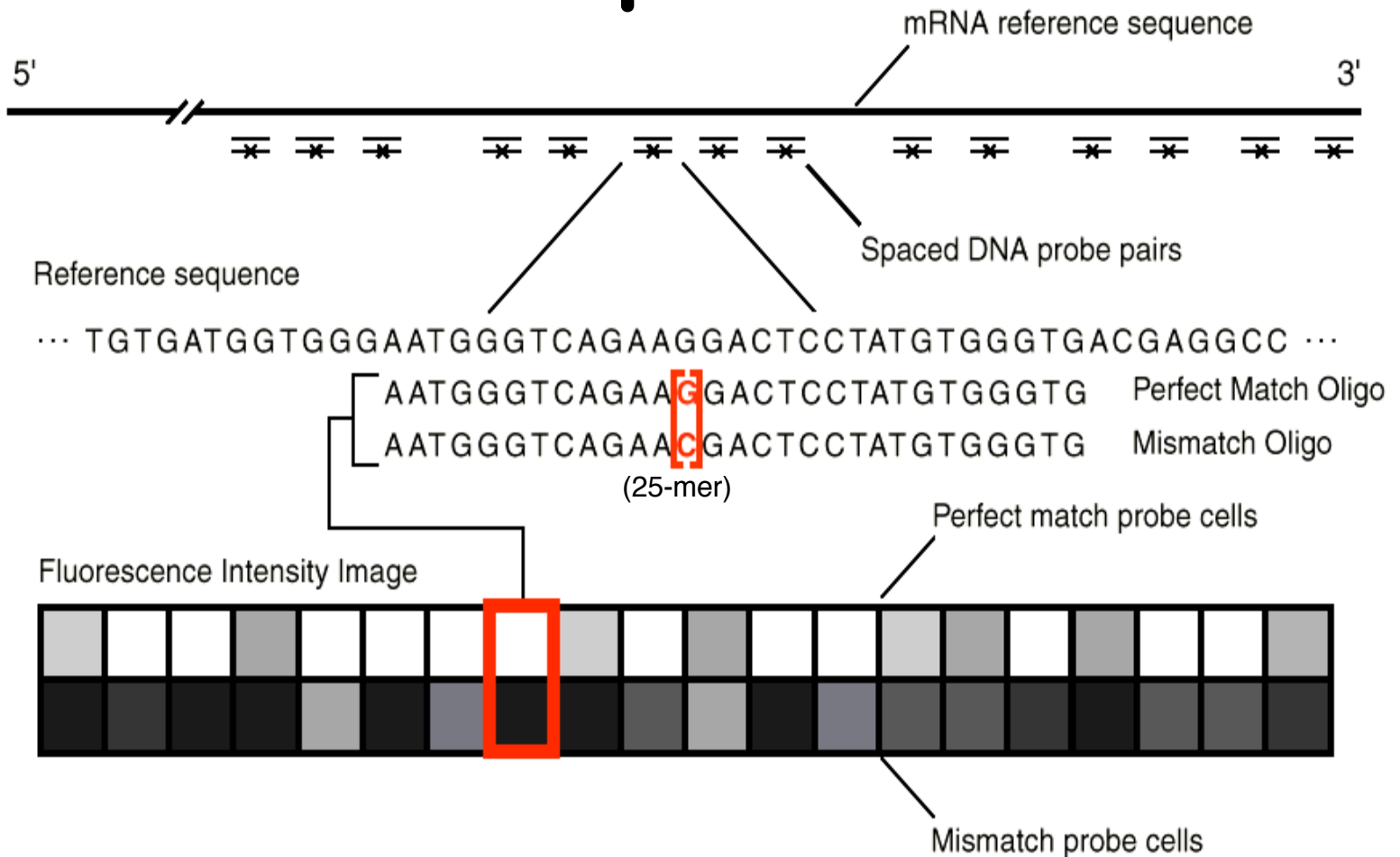
- So, a random 20-mer occurs in random human-sized DNA sequence with probability about $3 \times 10^9 \times 10^{-12} = .003$

How Random is a Genome?

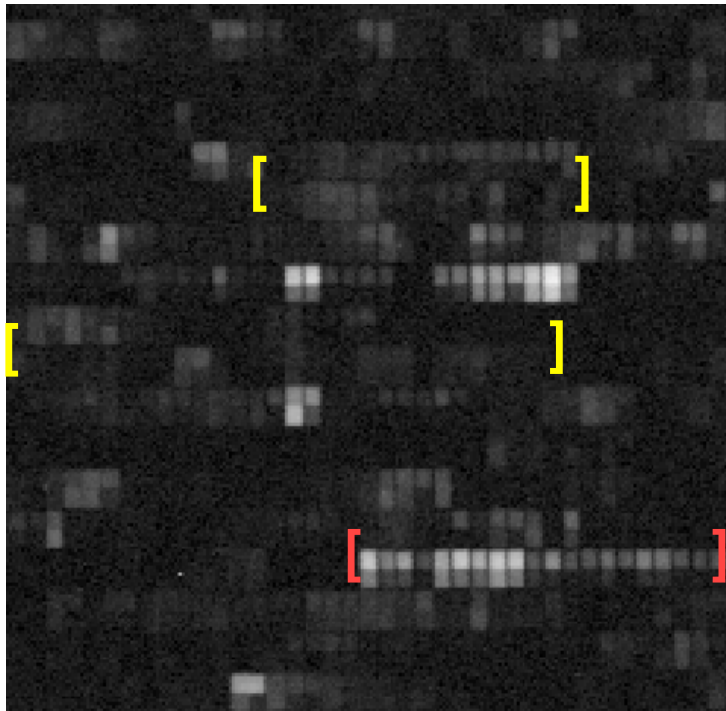
- G/C content can vary from ~40-60% across and within organisms ("isochores")
- Adjacent pairs not independent
- Adjacent triples not independent (esp. in genes)
- ...
- Many large-scale repeats, e.g.
 - similar genes, domains within genes
 - transposons & other junk
 - within primates, ~ 5% of all DNA is composed of (noisy) copies of a 300bp ALU sequence
- Nevertheless, crude model is useful guide

Probe Tiling Strategy

Gene Expression

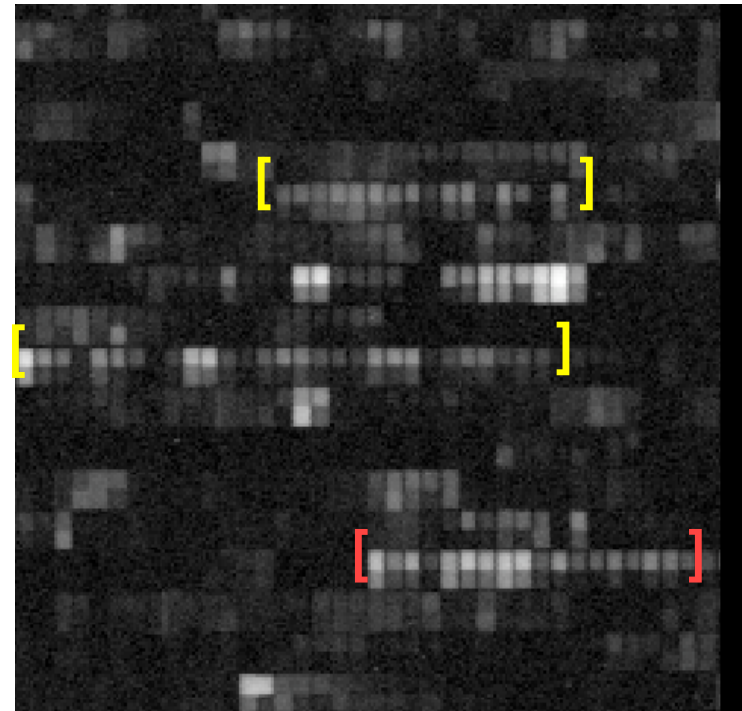


Gene Expression Tiling Strategy



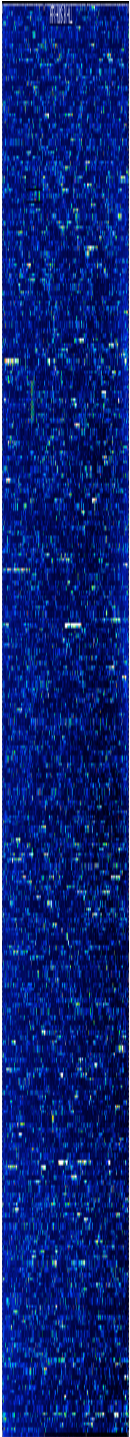
Uninduced

40 separate hybridization events are involved in determining the presence or absence of a transcript

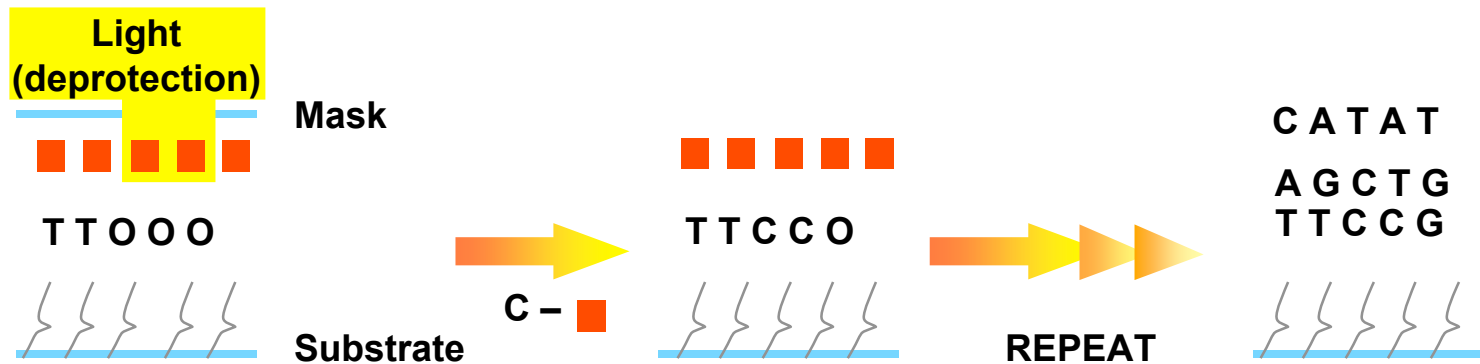
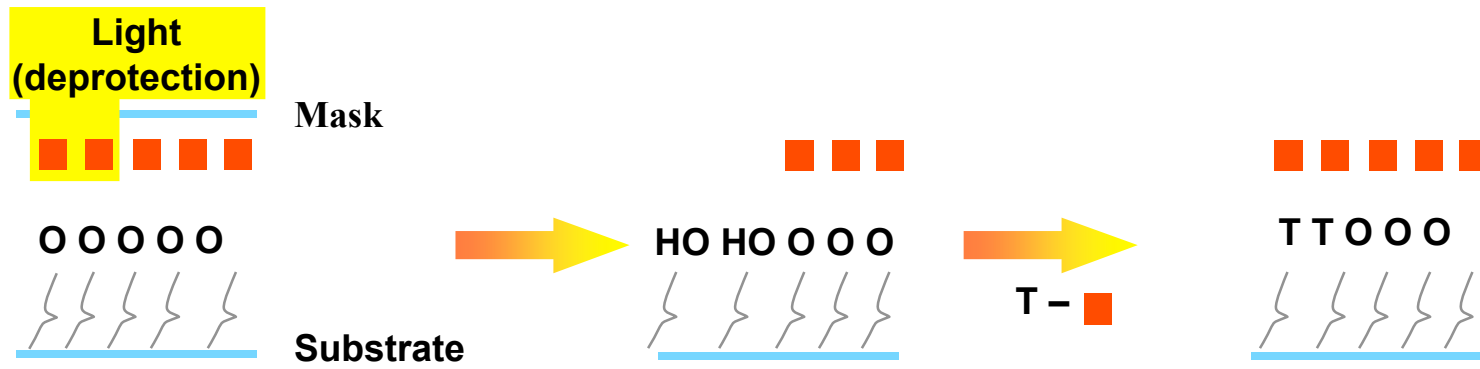


Induced

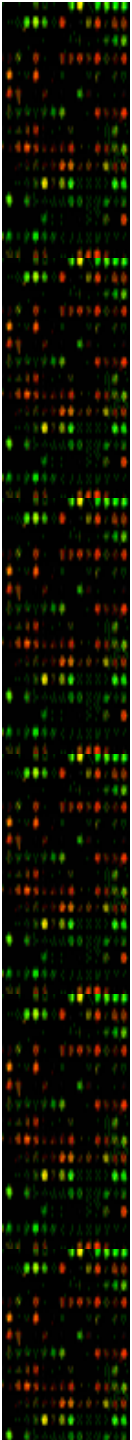
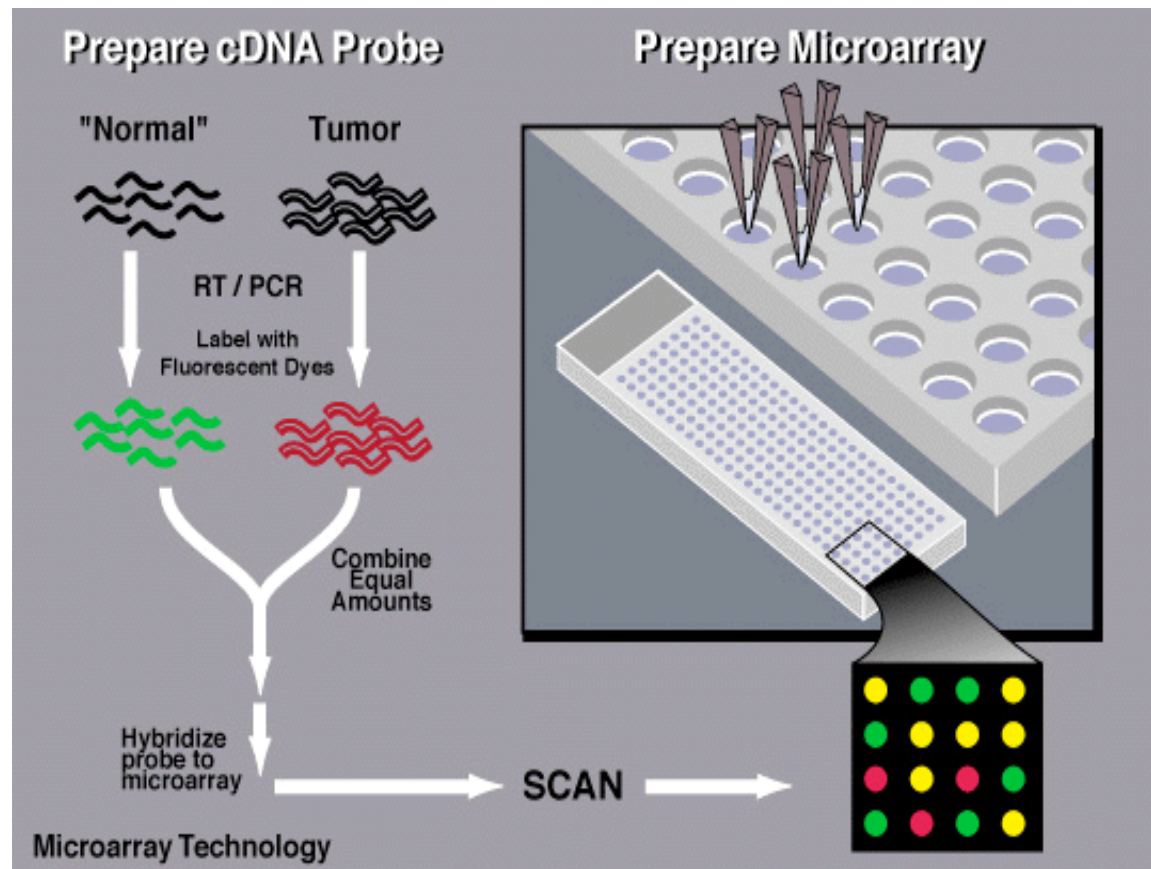
80 separate hybridization events are involved determining differential gene expression between two samples

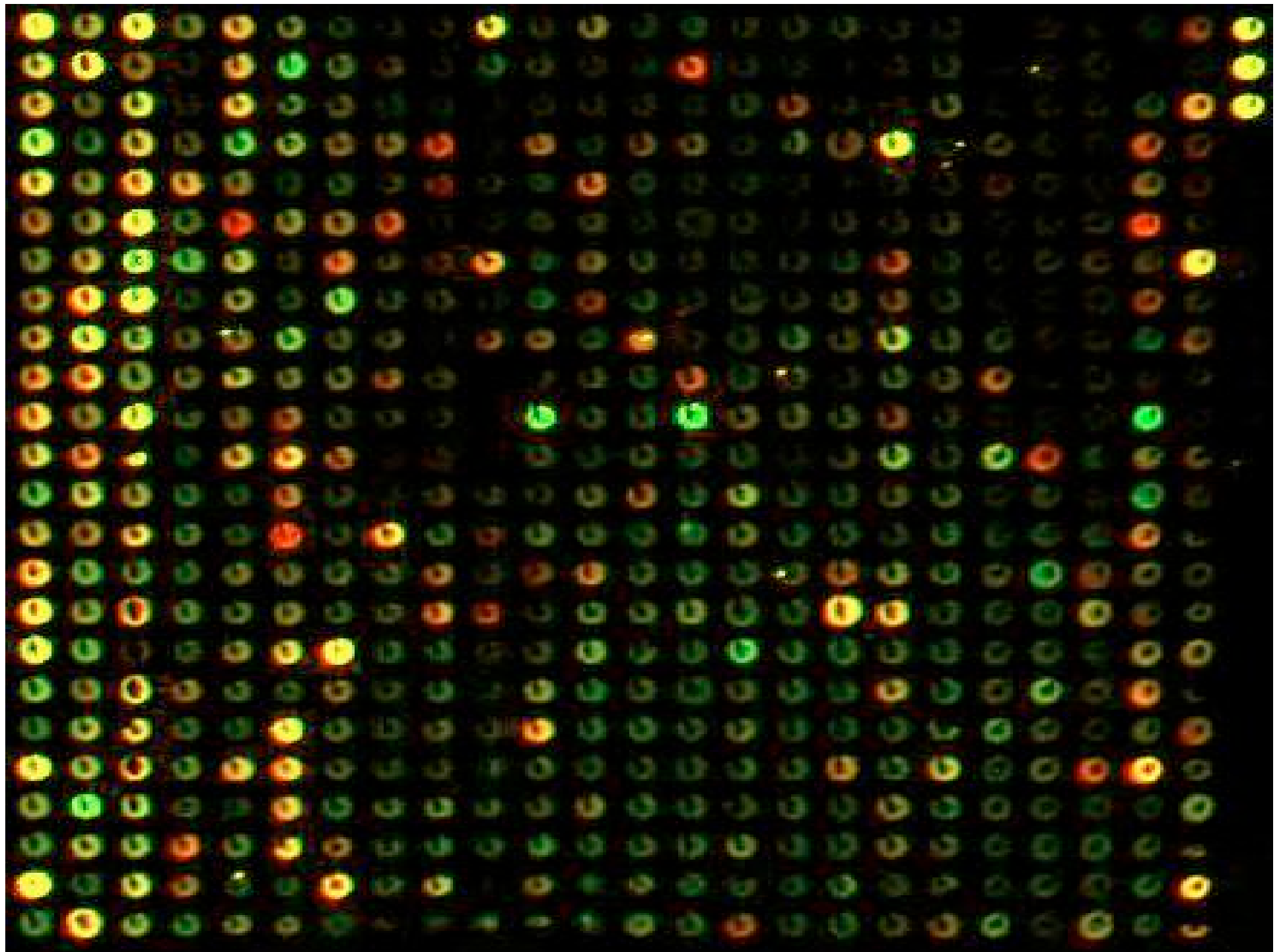


Synthesis of Ordered Oligonucleotide Arrays



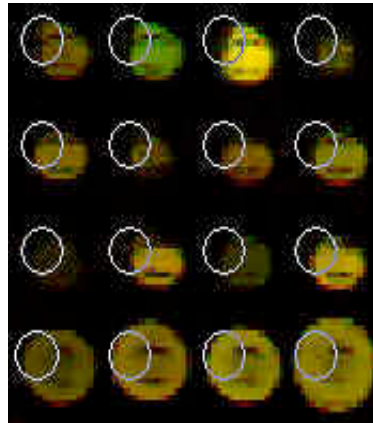
Spotted Microarray Process



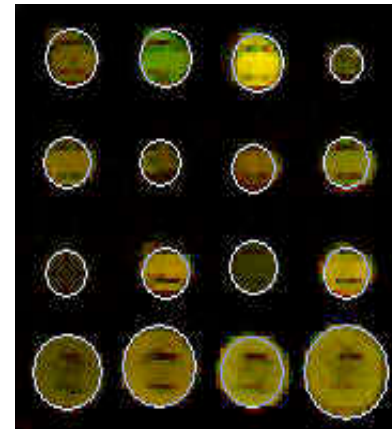


GenePix Pro Features

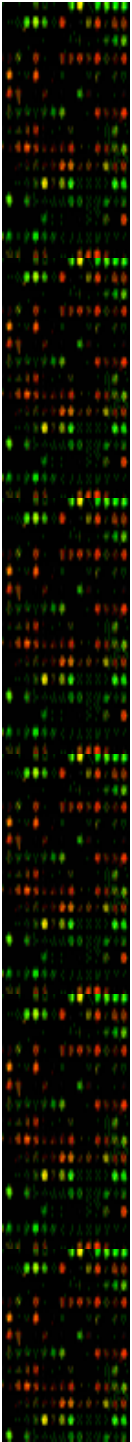
- Auto Align



Before Auto Align

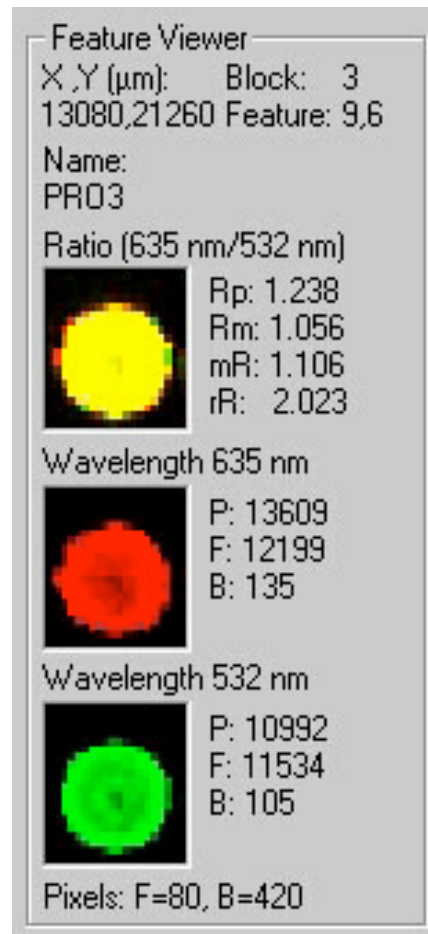


After Auto Align



GenePix Pro Features

•Feature Viewer



P = pixel intensity

F = feature intensity

B = background intensity

Rp = ratio of pixel intensities

Rm = ratio of means

mR = median of ratios

rR = regression ratio



Spotted glass slide microarrays

Advantages

Low cost per array

Custom gene selection

Any species

Competitive hybridization

Open architecture

Disadvantages

Clone management

Clone cost

Quality control



Affymetrix GeneChip system

Advantages

- Stream line production
- Large number of genes and ESTs/chip
- Several number of species

Disadvantages

- System cost
- GeneChip cost
- Proprietary system
- Limits on customizing

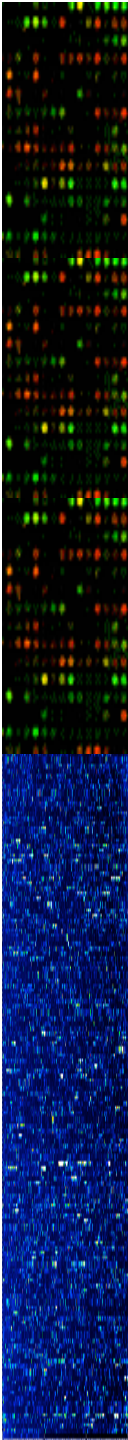
Micro Array Noise Sources

- Lot-to-lot variation (chips, reagents,...)
- Experiment-to-experiment variation
 - cell state, culture purity
 - sample preparation, hybridization conditions
- Spot-to-spot variation
 - unequal dye incorporation
 - dye nonlinearity/saturation
 - uneven spot sizes
 - self- & cross-hybridization
 - Image capture & processing (spot finding, quantization, sensors)
- ...



Challenges in analyzing Microarray Data

- Amount of DNA in spot is not consistent
- Spot contamination
- cDNA may not be proportional to that in the tissue
- Low hybridization quality
- Measurement errors
- Spliced variants
- Outliers
- Data are high-dimensional "multi-variant"
- Biological signal may be subtle, complex, non linear, and buried in a cloud of noise
- Normalization
- Comparison across multiple arrays, time points, tissues, treatments
- How do you reveal biological relationships among genes?
- How do you distinguish real effect from artifact?



Factors to consider in designing microarray experiments

- Need to do lots of control experiments-validate method
- Do replicate spotting, replicate chips, and reverse labeling for custom spotted chips
- Do pilot studies before doing "mega chip" experiments
- Don't design experiment without replication; nothing will be learned from a single failed experiment
- Design simple (one-two factor) experiments, i.e. treatment vs. untreated
- Understand measurement errors
- In designing Databases; they are useful ONLY if quality of data is assured
- Involve statistical colleagues in the design stages of your studies

The Microarray Biz. (circa 3/2001)

- Despite concerns above...
- "In early 1997, scientists never envisioned looking at more than 25 to 50 gene-expression levels simultaneously. Today everybody tells us that they want to look at the whole genome." -- T.Kreiner, Affymetrics
- 45% annual growth rate 1999-2000
- Est. \$1B by 2005 for diagnostics alone.