Lecture 2:
DNA Microarray Overview

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

Announcements

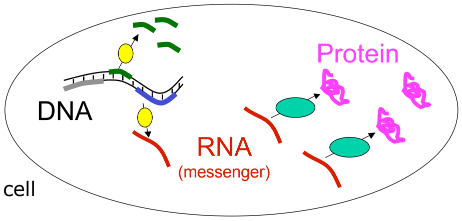
- Go to class web page
  http://www.cs.washington.edu/527
  - Add yourself to class list
  - Check out HW1, including last year's
- CSE 590CB Org. meeting today,
  3:30 MEB 243
  http://www.cs.washington.edu/590cb

Talks

- **Dr. Martin Tompa,** UW "Tools for Prediction of Regulatory Elements in Microbial Genes"
  - Combi Seminar**:** Wed 10/6 1:30, K-069
  - CSE Seminar: Tue 10/12 3:30, EE-105
- **Dr. Michal Linial**, Hebrew University, "The Protein Family Tree: Making Biological Sense From Sequence Data"
  - CSE Seminar:Thu 10/7 3:30 pm, EE-105

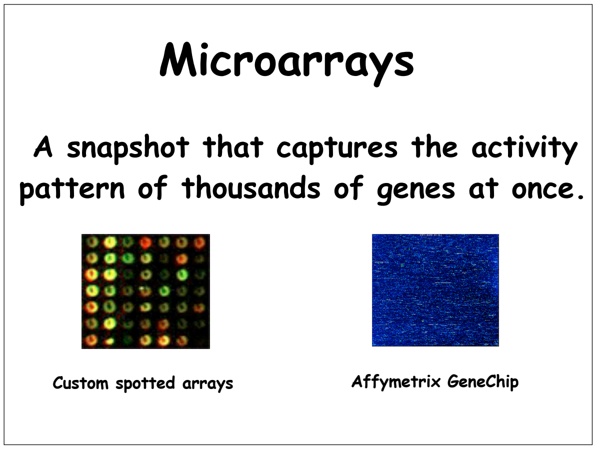
Gene Expression:
The "Central Dogma"

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

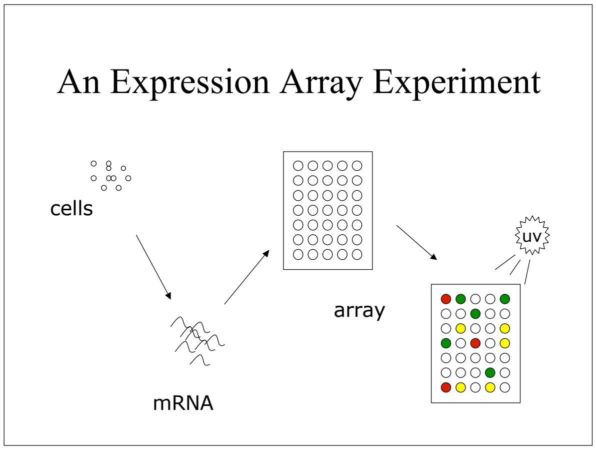
A snapshot that captures the activity pattern of thousands of genes at once.

Custom spotted arrays

Affymetrix GeneChip

Expression Microarrays

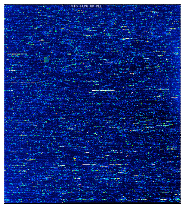
- The Array
  - Thousands to hundreds of thousands of spots per square inch
  - Each holds millions of copies of a DNA sequence from one gene
- Its Use
  - Take mRNA from cells, put it on array
  - See where it sticks – mRNA from gene x should stick to spot x



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?

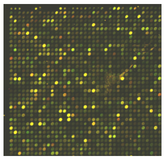
- What genes are behaving differently between ALL & AML
  (or other disease/normal states)?
- Potential uses:
  - Diagnosis
  - Prognosis
  - Insight into underlying biology/biologies
  - Treatment

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

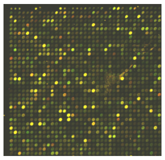
An Example Application

- Yeast "Sporulation"
- 7 time points over ~18 hours
- One array per time point
- All 6200 yeast genes on each



Chu, DeRisi, Eisen, Mulholland, Botstein, Brown, Herskowitz, "The Transcriptional Program of Sporulation in Budding Yeast," Science, 282 (Oct 1998) 699-705

An Example Application



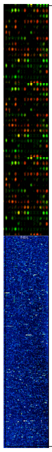
- Yeast “Sporulation”
- 7 time points over ~18 hours
- One array per time point
- All 6200 yeast genes on each

3-10x increase in number of genes known to be involved in sporulation, many with recognizable analogs in humans, presumably key players in egg/sperm formation

Chu, DeRisi, Eisen, Mulholland, Botstein, Brown, Herskowitz, “The Transcriptional Program of Sporulation in Budding Yeast,” Science, 282 (Oct 1998) 699-705

Other Applications

- Study gene function & regulation
  - Covarying ~~> coregulated?
  - Covarying ~~> 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
- Diseased vs normal cell comparison suggests sets of genes having key roles.
- Over/underexpressed genes in the diseased cells can suggest drug targets

Pharmacology and Toxicology
- Highly sensitive indicator of a drug’s activity (pharmacology) and toxicity (toxicology) in cell culture or test animals.
- Screen or optimize drug candidates prior to costly clinical trials.

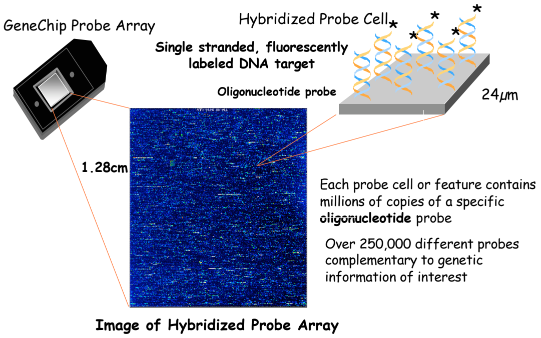
Diagnostics
- Potential to diagnose clinical conditions by detecting gene expression patterns associated with disease states in either biopsy samples or peripheral blood cells.

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



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

Image of Hybridized Probe Array

How unique is a 20-mer?

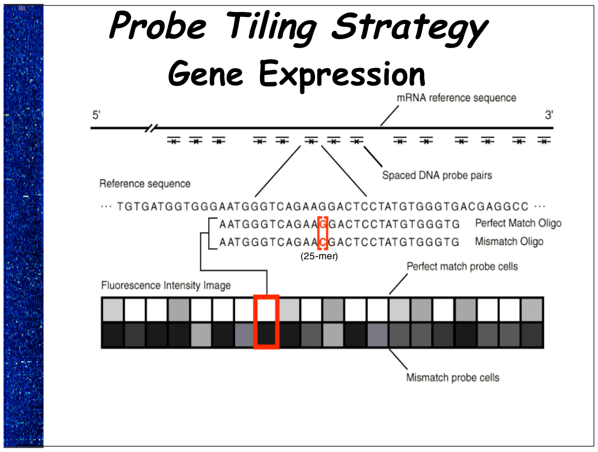
- VERY CRUDE model: DNA is random—every position is equally likely to be A, C, G, or T, independent of every other
- Then probability of a random 20-mer is

$$\left(\frac{1}{4}\right)^{20} = \left(\frac{1}{2}\right)^{40} = \left(\left(\frac{1}{2}\right)^{10}\right)^{4} = \left(\left(\frac{1}{1024}\right)\right)^{4} \approx \left(10^{-3}\right)^{4} = 10^{-12}$$

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

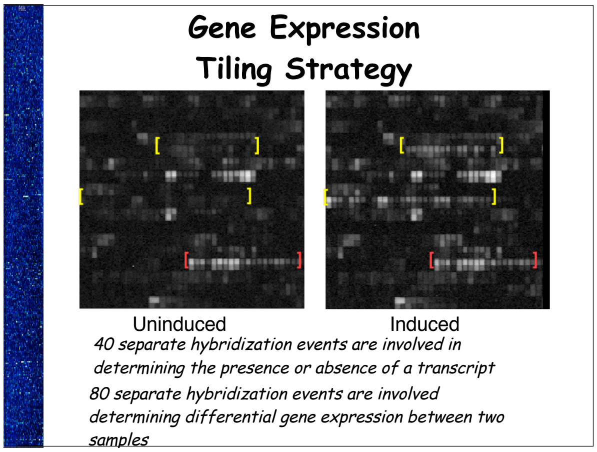
How Random is a Genome?

- 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 (noisey) copies of a 300bp ALU sequence
- Nevertheless, crude model above is a useful guide



Probe Tiling Strategy
Gene Expression

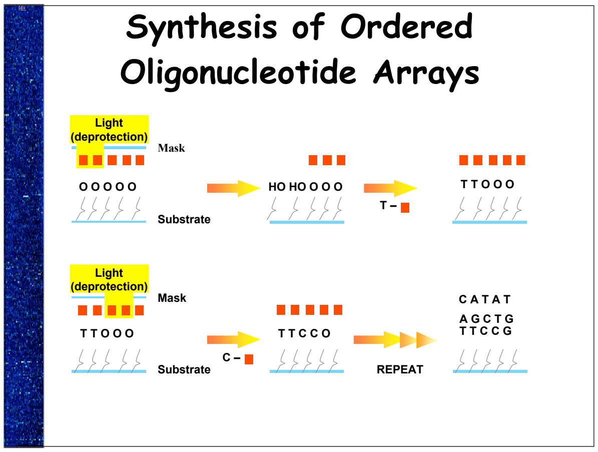
mRNA reference sequence
5'
3'
Spaced DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
AATGGGTCAGAAGGACTCCTATGTGGGTG Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTGGGTG Mismatch Oligo
(25-mer)
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells



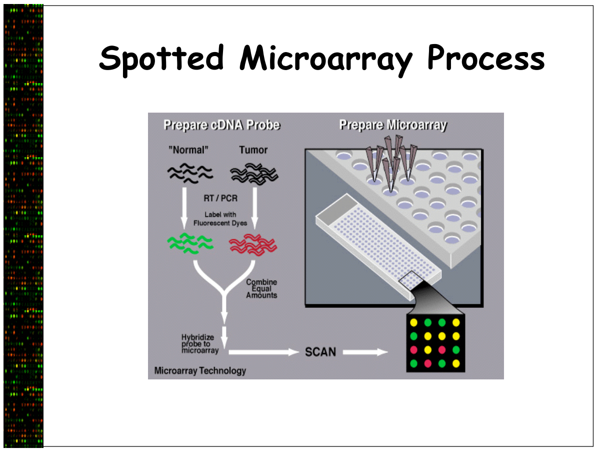
Gene Expression
Tiling Strategy

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

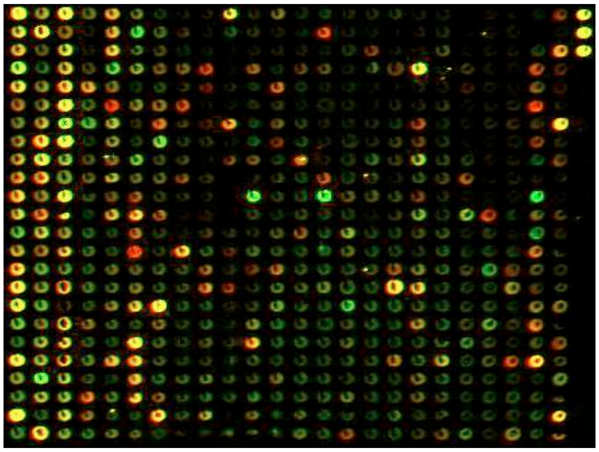
80 separate hybridization events are involved determining differential gene expression between two samples



Synthesis of Ordered Oligonucleotide Arrays

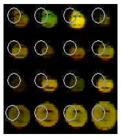


Spotted Microarray Process

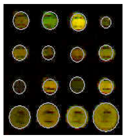


GenePix Pro Features

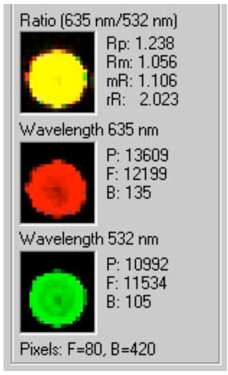
- Auto Align



Before Auto Align



After Auto Align



```
Ratio (635 nm/532 nm)
        Rp: 1.238
        Rm: 1.056
        mR: 1.106
        rR:   2.023
Wavelength 635 nm
        P: 13609
        F: 12199
        B: 135
Wavelength 532 nm
        P: 10992
        F: 11534
        B: 105
Pixels: F=80, B=420
```

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

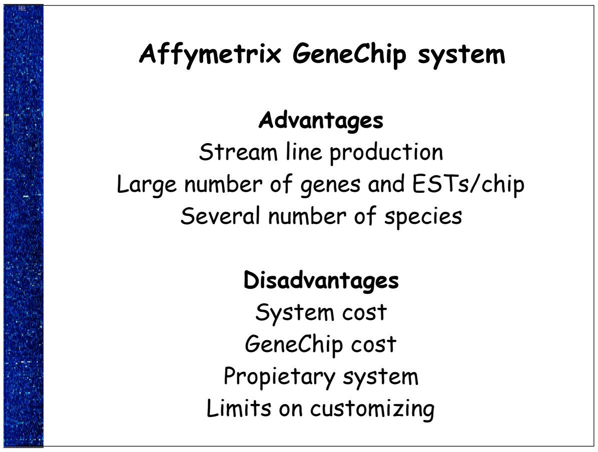
Spotted glass slide microarrays

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

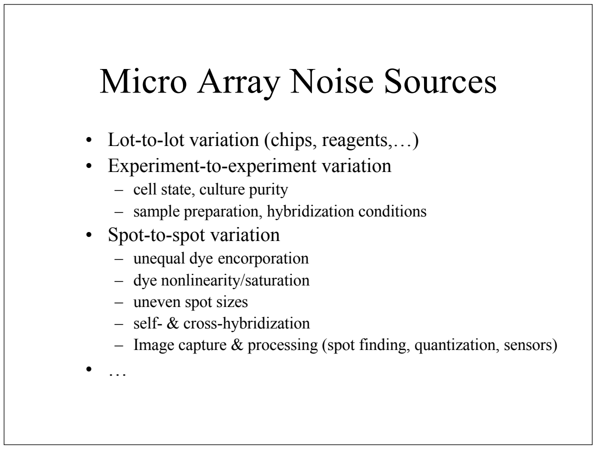
Disadvantages

System cost

GeneChip cost

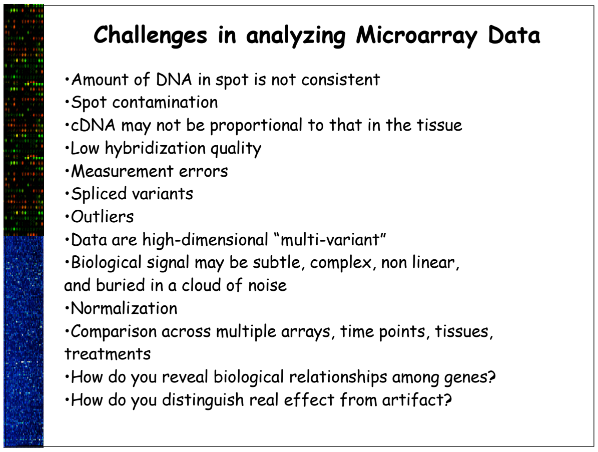
Propietary system

Limits on customizing



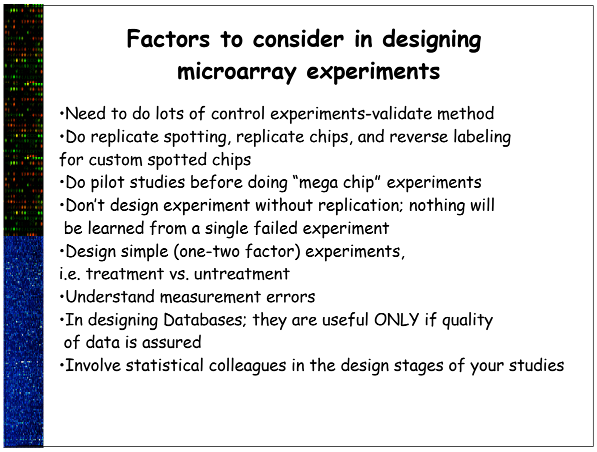
Micro Array Noise Sources

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



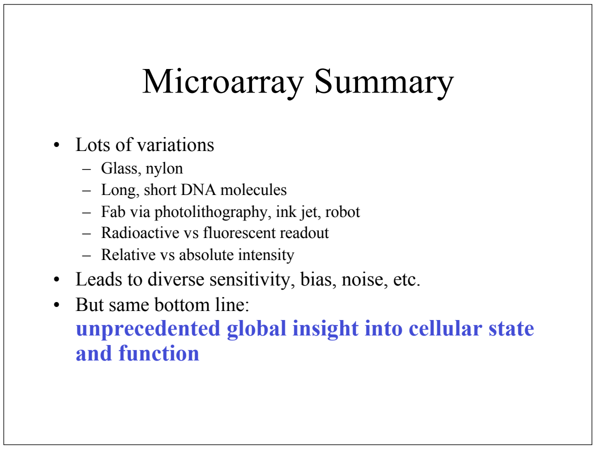
Challenges in analyzing Microarray Data

•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?



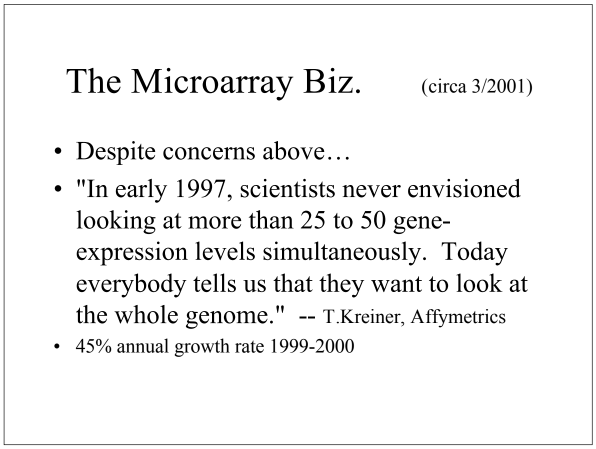
Factors to consider in designing microarray experiments

•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. untreatment
•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



Microarray Summary

- Lots of variations
  - Glass, nylon
  - Long, short DNA molecules
  - Fab via photolithography, ink jet, robot
  - Radioactive vs fluorescent readout
  - Relative vs absolute intensity
- Leads to diverse sensitivity, bias, noise, etc.
- But same bottom line:
  unprecedented global insight into cellular state and function



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