# Gene Expression Microarrays for Dummies

## *What We Learned this Summer*

Monnie McGee & Zhongxue Chen

Department of Statistical Science

Southern Methodist University

# Acknowledgments

SMU Microarray Analysis Group (SMUMAG)

| Faculty | Students |
| --- | --- |
| Jing Cao | Zhongxue Chen |
| Tony Ng | Kinfe Gedif |
| William Schucany | Drew Hardin |
| Xinlei Wang | Jobayer Hossain |
| | Ariful Islam |
| | Julia Kozlitina |

Data supplied by Boland Lab at Baylor.

# Outline

- Motivation
- Central Dogma of Biology
- Types of Microarrays
- Central Dogma of Microarray Analysis
- Robust Multi-Chip Average
- Improvements (?) to RMA
- Future Work

# Colon Cancer Cell Line Data

- Microarrays of four cell lines
  - HCT116: Microsatellite Instability Model
  - HCT111 Plus 3: MSI plus a corrective gene
  - SW48: CIMP line (silencing of genes)
  - SW480: Chromosomal Instability (CIN) line
- Four treatments to each line (including no treatment)
- Two “control” cell lines (RKO & HT29)
- Total of 18 microarrays

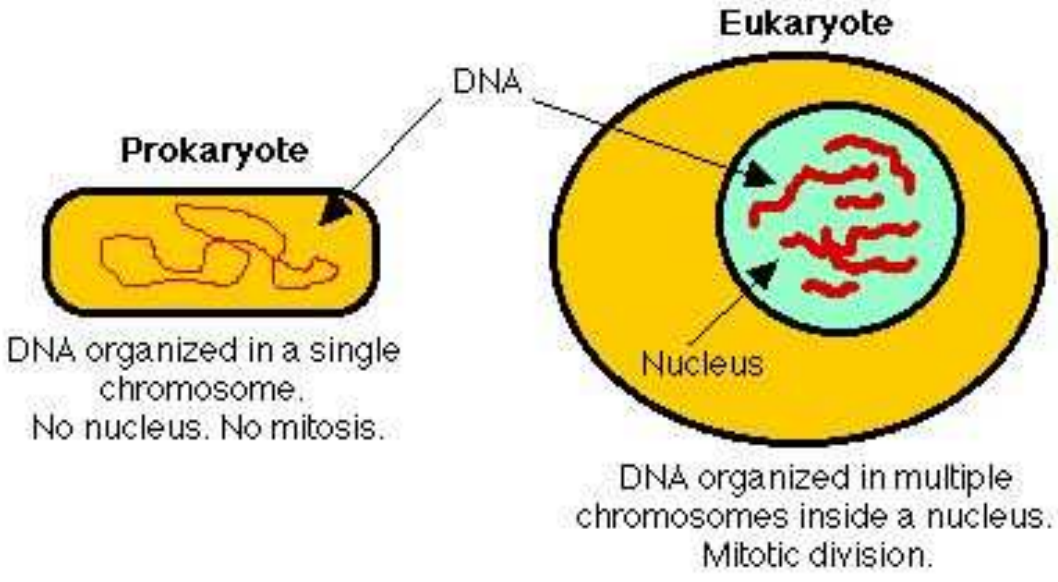
# Colon Cancer Cell Line Data

- Microarrays of four cell lines
  - HCT116: Microsatellite Instability Model
  - HCT111 Plus 3: MSI plus a corrective gene
  - SW48: CIMP line (silencing of genes)
  - SW480: Chromosomal Instability (CIN) line
- Four treatments to each line (including no treatment)
- Two “control” cell lines (RKO & HT29)
- Total of 18 microarrays

Question: What genes are differentially expressed among the various cell lines?

# Two Cell Types

Cells are the fundamental working units of all organisms.




Prokaryotes vs. Eukaryotes

Image drawn by Thomas M. Terry for The Biology Place.

# Key Macromolecules

- Lipids
  - Mostly structural in function
  - Construct compartments that separate inside from outside
- DNA
  Encodes hereditary information
- Proteins
  - Do most of the work in the cell
  - Form 3D structure and complexes critical for function

# DNA and Base Pairs

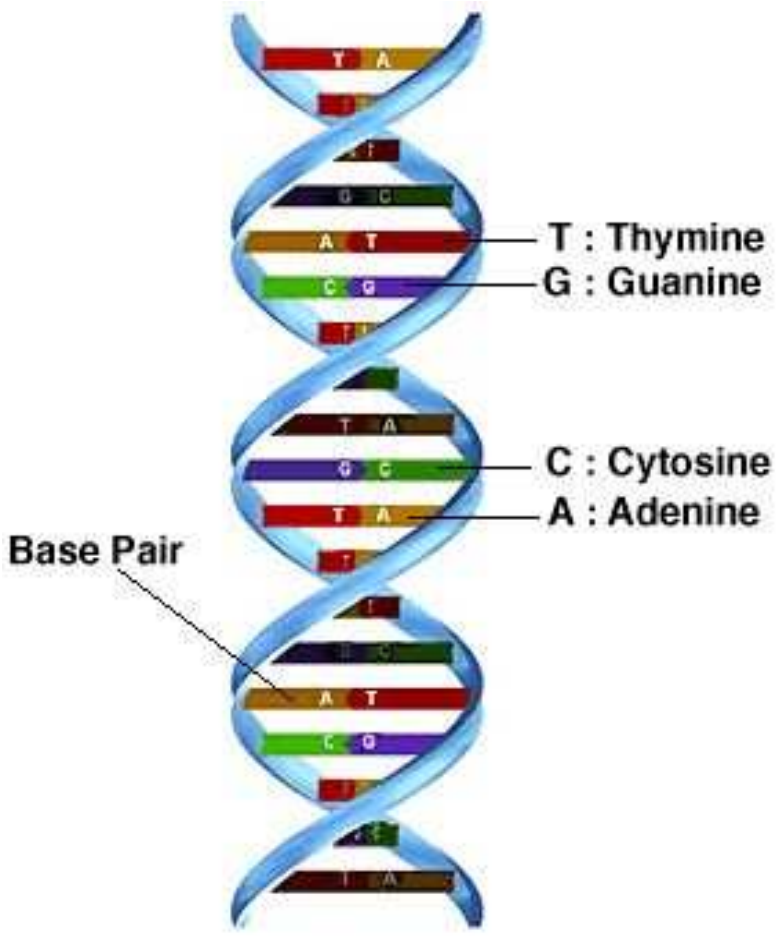



Image Courtesy of ExploreMore Television

# Central Dogma of Biology

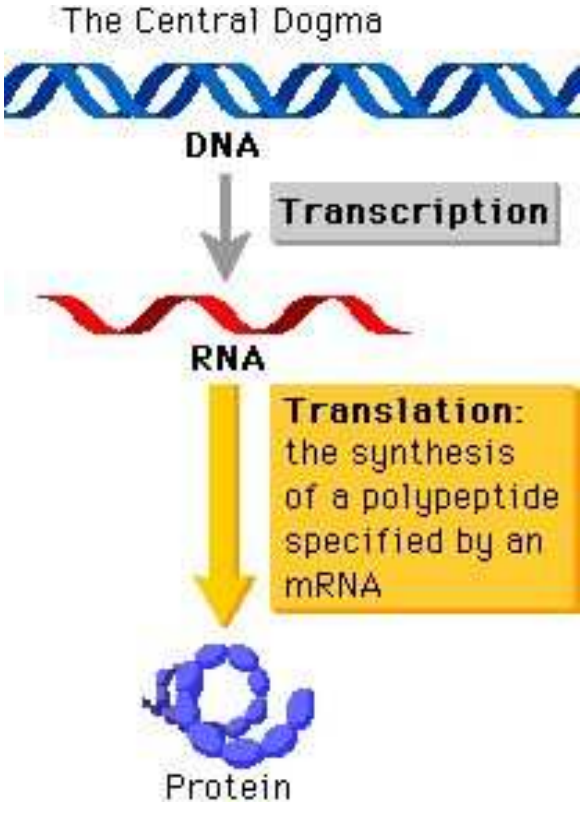



Image Courtesy of BioCoach

# Transcription

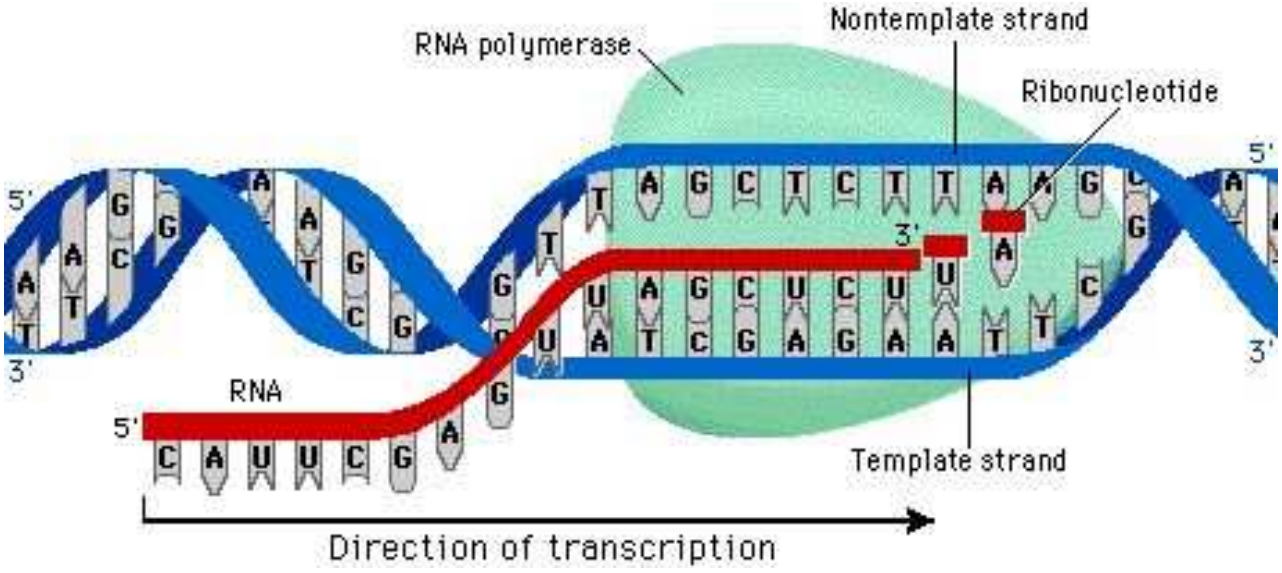



Image Courtesy of BioCoach

Movie of Complete Transcription

# Measuring Gene Expression

Gene expression can be quantified by measuring either mRNA or protein.

- mRNA Measures

  Quantitative Northern blot, qPCR, qrt-PCR, short or long oligonucleotide arrays, cDNA arrays, EST sequencing, SAGE, MPSS, MS, bead arrays, etc.

- Protein Measures

  Quantitative Western blots, ELISA, 2D-gels, gas or liquid chromatography, mass-spec, etc.

# Why Microarray Analysis?

- Large-scale study of biological processes
- Activity in cell at a certain point in time
- Account for differences in phenotypes on a large-scale genetic level
- Sequences are important, but genes have effect through expression

# Why Microarray Analysis?

- Large-scale study of biological processes
- Activity in cell at a certain point in time
- Account for differences in phenotypes on a large-scale genetic level
- Sequences are important, but genes have effect through expression

Rough measurement on a grand scale which has utility

# Measuring Gene Expression

Basic idea: Quantify concentration of a gene's mRNA transcript in a cell at a given time

# Measuring Gene Expression

Basic idea: Quantify concentration of a gene's mRNA transcript in a cell at a given time

How?

- Immobilize DNA probes onto glass (or other medium)
- Hybridize labeled target mRNA with probes
- Measure how much binds to each probe (i.e. forms DNA)

# Microarray Measurements

All raw measurements are fluorescence intensities

- Target cDNA (or mRNA) is radioactively labeled
- Molecules in dye are excited using a laser
- Measurement is a count of the photons emitted
- Entire slide or chip is scanned, and the result is a digital image
- Image is processed to locate probes and assign intensity measurements to each probe

# Microarray Technologies

- Two Channel Spotted Arrays
  - Robotic Microspotting
  - Probes are 300 to 3000 base pairs in length
  - Long-oligo arrays: probes are uniformly 60 to 90 bp
  - Commerical arrays using inkjet technology
- Single-channel Arrays
  - High-density short oligo (25 bp) arrays (Affymetrix, Nimblegen)

# Spotted Arrays

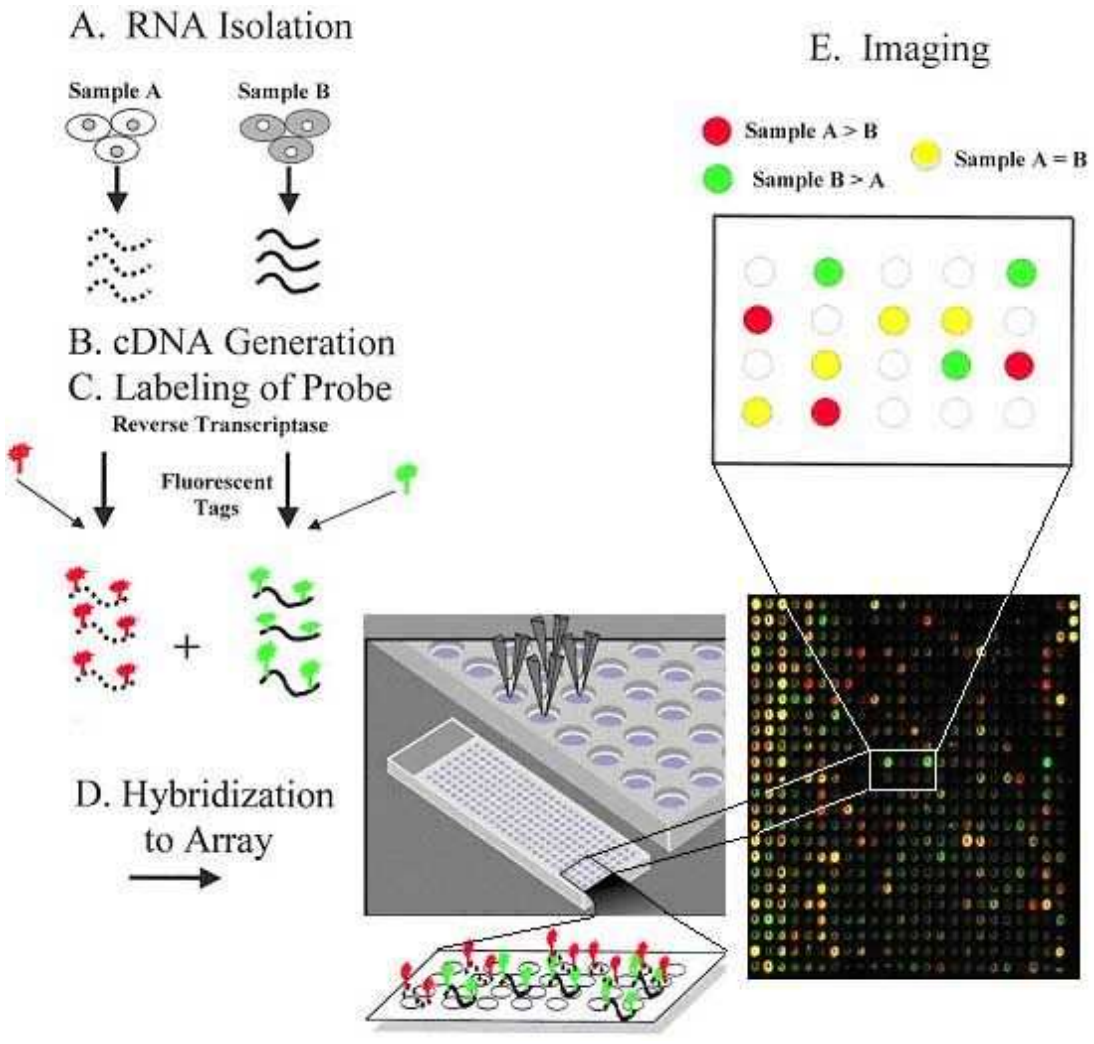
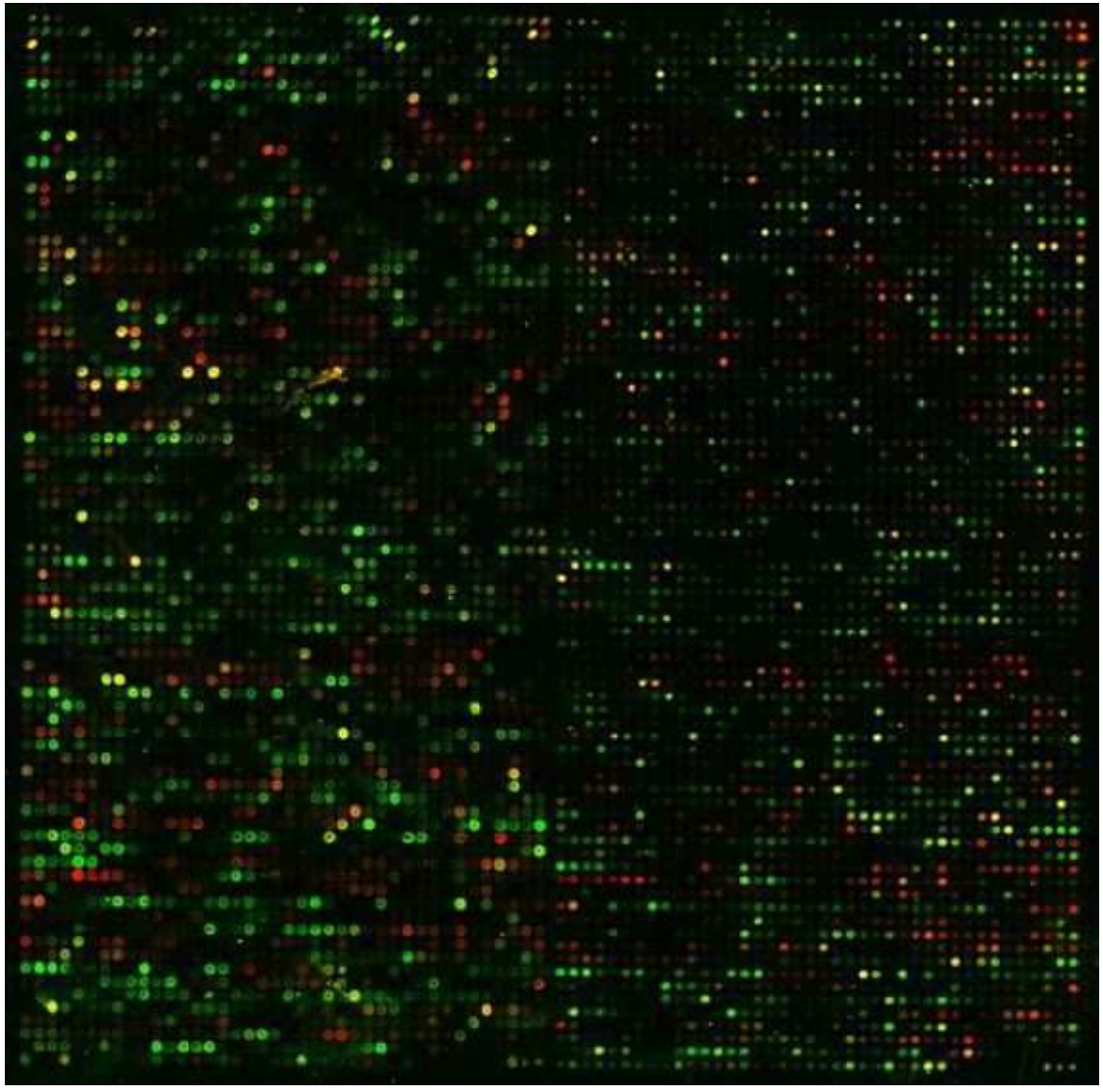



Diagram courtesy of Columbia Department of Computer Science

# Yeast Array Image



Yeast Array courtesy of Russ Altman, Stanford University

# The Affymetrix Chip


Human Genome U133 Plus 2.0 Array

Courtesy of Affymetrix

Some Definitions

- Probes = 25 bp sequences
- Probe sets = 11 to 20 probes corresponding to a particular gene or EST
- Chip contains 54K probe sets

# *In situ* Synthesis of Probes

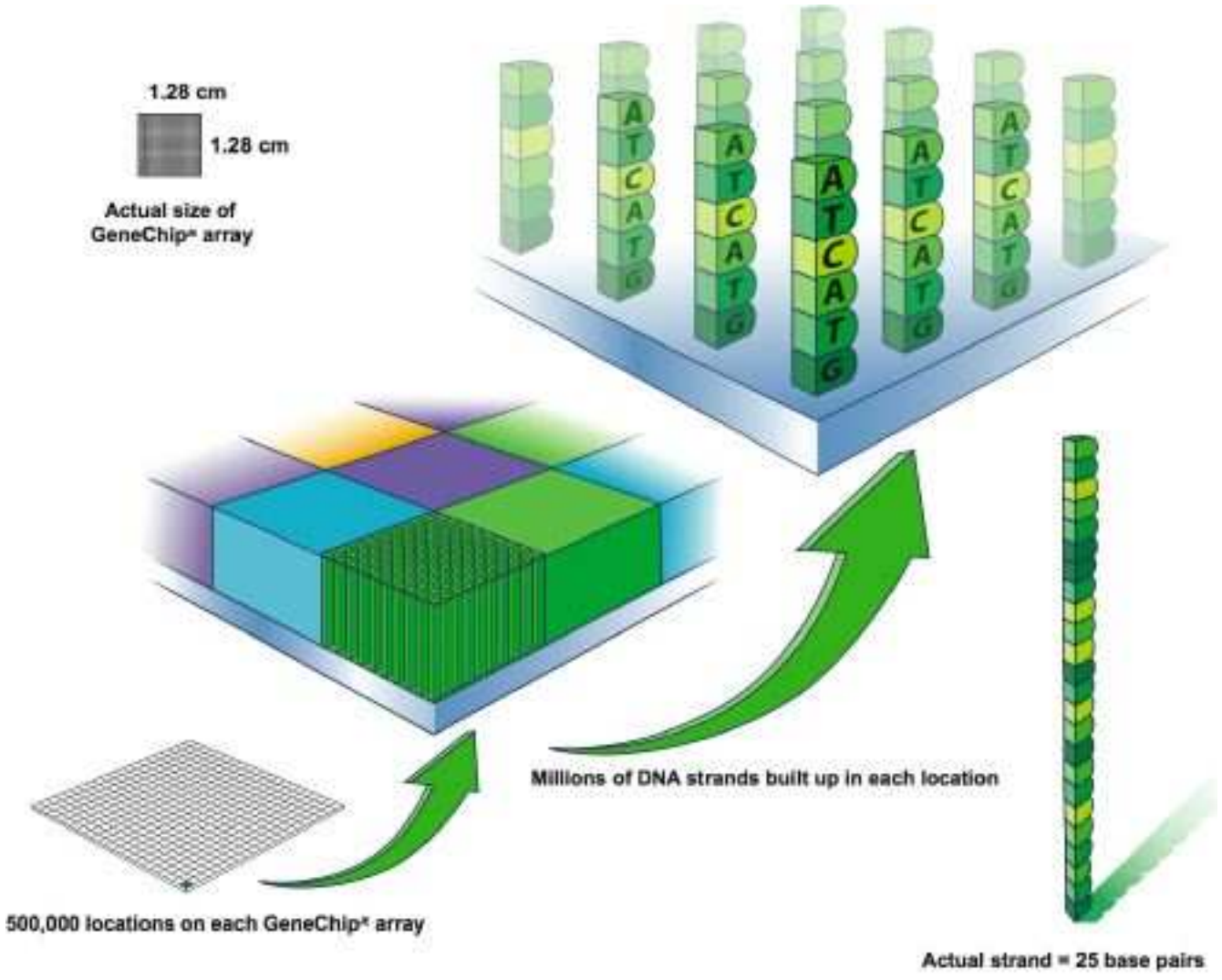



Image Courtesy of Affymetrix

# mRNA Hybridizes to Probes

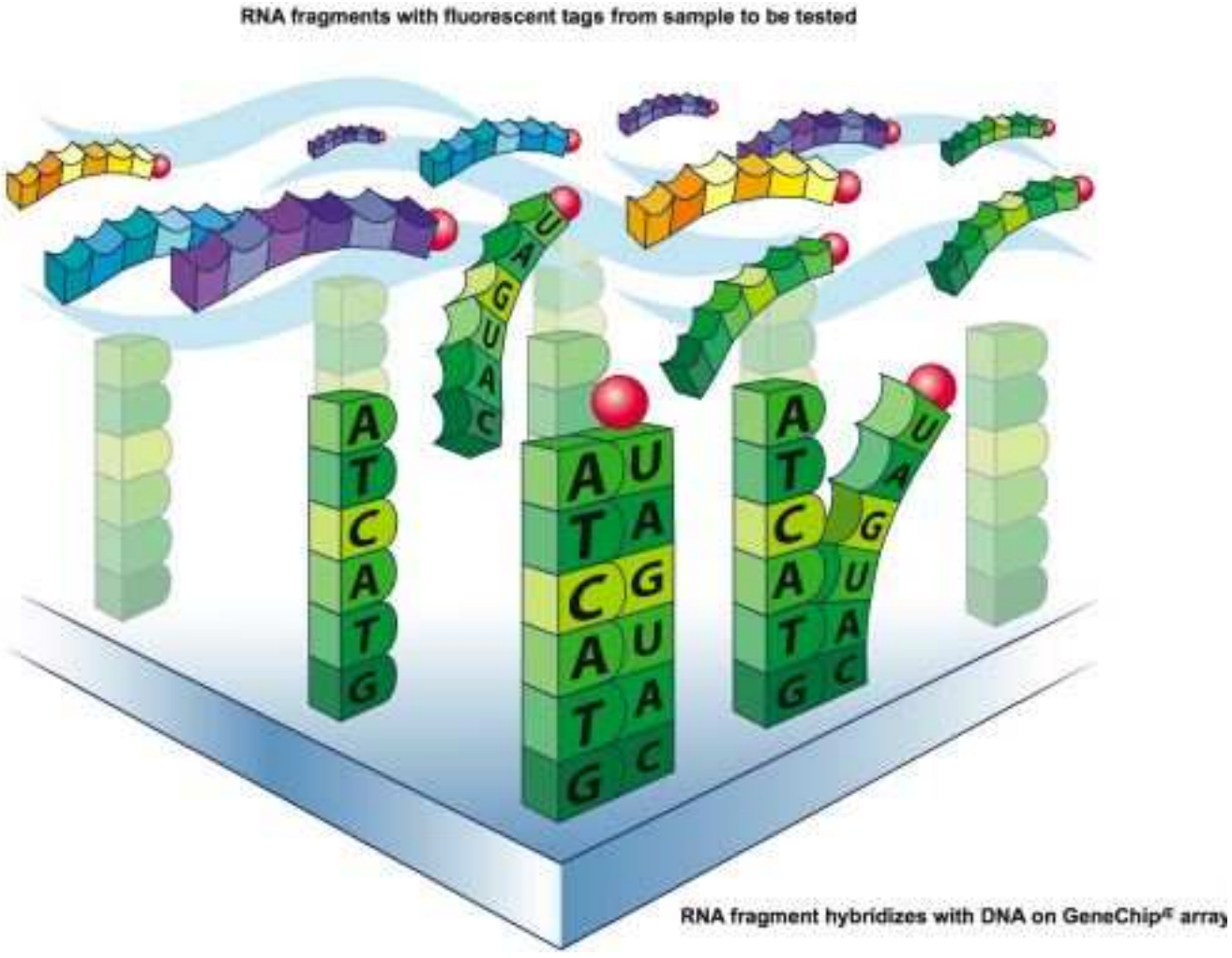



Image Courtesy of Affymetrix

# Perfect Match vs. Mismatch

- PM Probe = 25 bp probe perfectly complementary to a specific region of a gene
- MM Probe = 25 bp probe agreeing with a PM apart from the middle base.
- The middle base is a transition (A $\Longleftrightarrow$ G, C $\Longleftrightarrow$ G) of that base

# Perfect Match vs. Mismatch

- PM Probe = 25 bp probe perfectly complementary to a specific region of a gene
- MM Probe = 25 bp probe agreeing with a PM apart from the middle base.
- The middle base is a transition (A $\Longleftrightarrow$ G, C $\Longleftrightarrow$ G) of that base

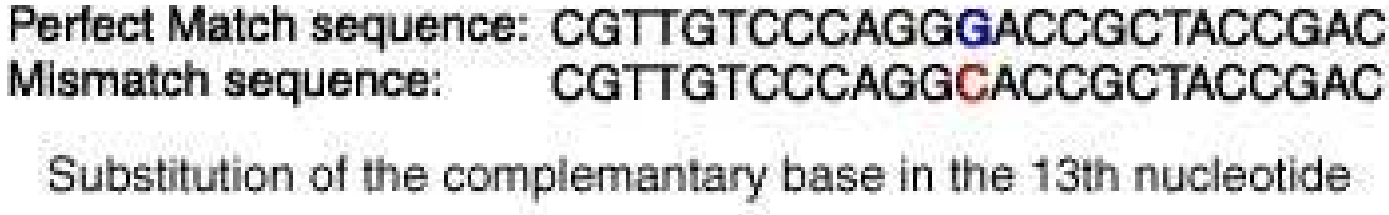



Image Courtesy of Affymetrix

# PM and MM Example

Target Transcript for Human recA gene:

ctcagcttaagtcatggaattctagaggatgtatctcacaagtaggatcaag

| | |
|---|---|
| ctcagcttaagtcatggaattctag | PM1 |
| ctcagcttaagtgatggaattctag | MM1 |
| tcagcttaagtcatggaattctaga | PM2 |
| tcagcttaagtctt ggaattctaga | PM2 |
| attctagaggatgtatctcacaagt | PM3 |
| attctagaggatctatctcacaagt | MM3 |
| aggatgtatctcacaagtaggatca | PM4 |
| aggatgtatctct caagtaggatca | MM4 |

Source: Naef and Magnasco (2003). Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Physical Review*, **68**.

# Image of *E. Coli* Gene Chip

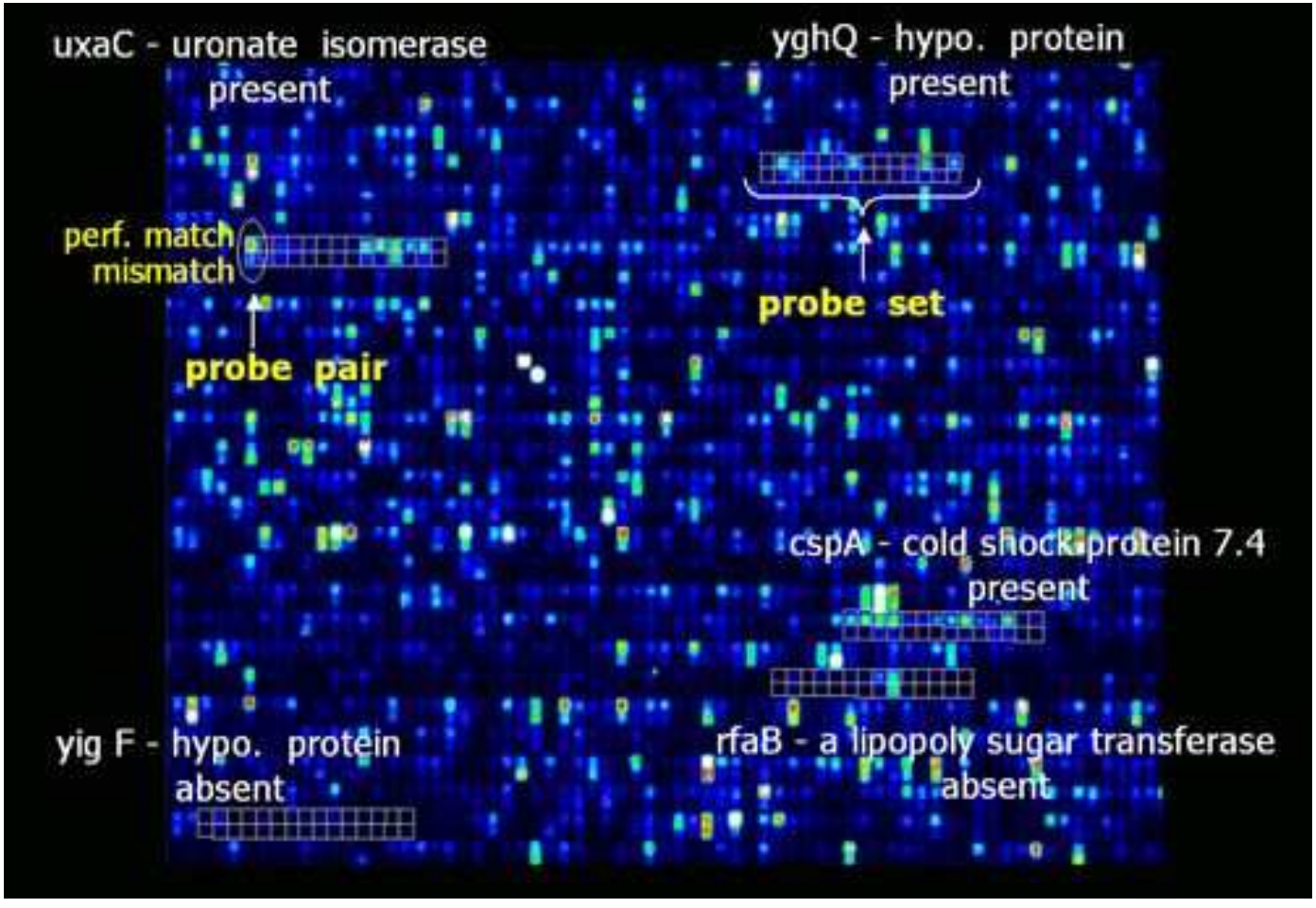



Image Courtesy of Lee Lab at Cornell University

# Analysis Tasks

- Identify up- and down-regulated genes.
- Find groups of genes with similar expression profiles.
- Find groups of experiments (tissues) with similar expression profiles.
- Find genes that explain observed differences among tissues (feature selection).

# Central Dogma of MA Analysis

Computing Expression Values for each probe set requires three steps:

- Background correction (image correction for cDNA)
- Normalization
- Summarization

# Central Dogma of MA Analysis

Computing Expression Values for each probe set requires three steps:

- Background correction (image correction for cDNA)
- Normalization
- Summarization

One Approach: Robust Multichip Analysis (RMA)
Irizarry *et. al.*, *Nucleic Acids Research*, 2003

# Background Correction in RMA

$$X = S + Y$$

where

$X$ = observed probe–level intensity

$S \sim E(\alpha)$ = true signal

$Y \sim TN(\mu, \sigma^2)$ = background noise

Reference: Irizarry *et. al.*, *Biostatistics*, 2003

# RMA for the Right–Brained ...

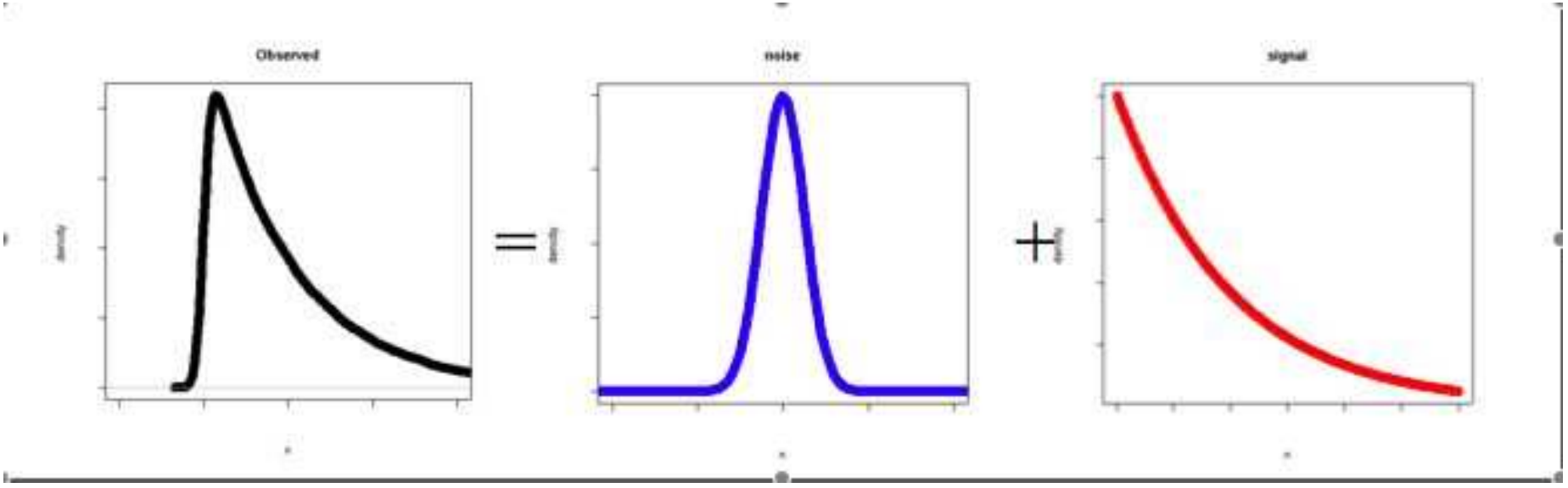



Image courtesy of Terry Speed

# Colon Cancer Cell Line Data

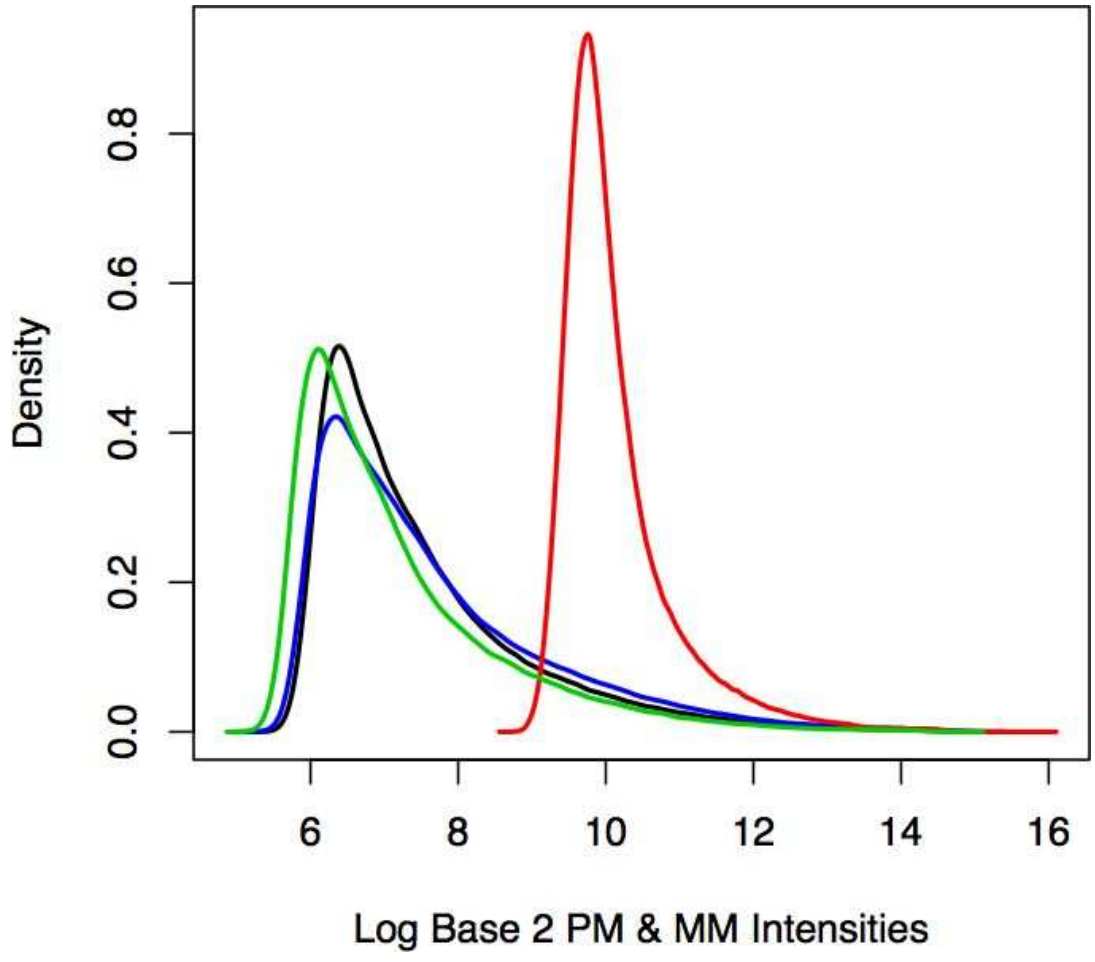
- Microarrays of four cell lines
  - HCT116: Microsatellite Instability Model
  - HCT111 Plus 3: MSI plus a corrective gene
  - SW48: CIMP line (silencing of genes)
  - SW480: Chromosomal Instability (CIN) line
- Four treatments to each line (including no treatment)
- Two “control” cell lines (RKO & HT29)
- Total of 18 microarrays

# Colon Cancer Cell Line Data

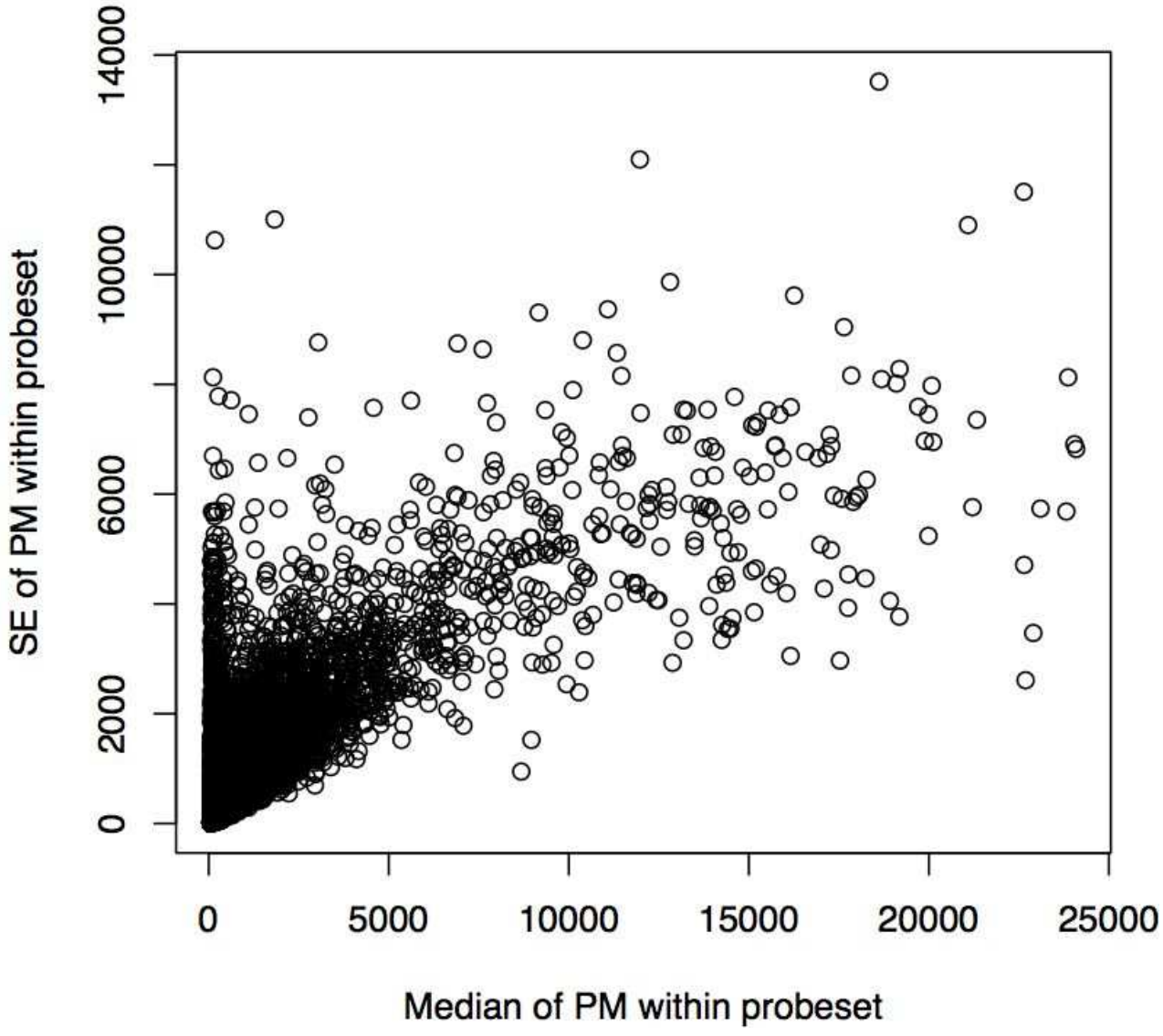
- Microarrays of four cell lines
  - HCT116: Microsatellite Instability Model
  - HCT111 Plus 3: MSI plus a corrective gene
  - SW48: CIMP line (silencing of genes)
  - SW480: Chromosomal Instability (CIN) line
- Four treatments to each line (including no treatment)
- Two “control” cell lines (RKO & HT29)
- Total of 18 microarrays

Question: What genes are differentially expressed among the various cell lines?

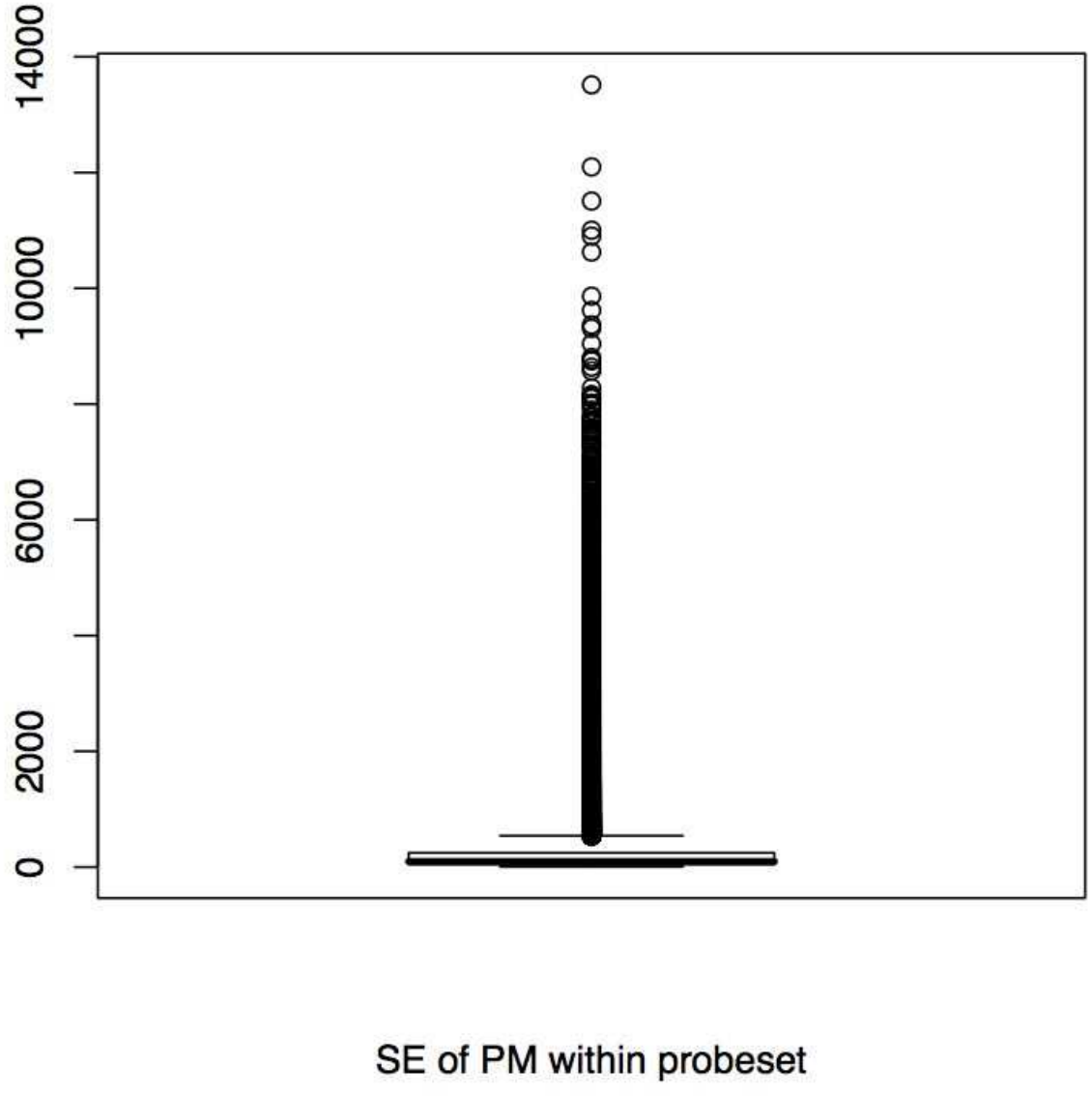
# Log Base 2 SW 480 Intensities



# Exploratory Data Analysis



# Exploratory Data Analysis (cont'd)



# Parameter Estimation

- Background Corrected intensity is $E_{ij} = E(S_{ij}|X_{ij})$, where $i = 1 \ldots G$, and $j = 1, \ldots, J$.
- We need to estimate $\mu$, $\sigma$, and $\alpha$.

# Parameter Estimation

- Background Corrected intensity is $E_{ij} = E(S_{ij}|X_{ij})$, where $i = 1 \ldots G$, and $j = 1, \ldots, J$.
- We need to estimate $\mu$, $\sigma$, and $\alpha$.

How does RMA estimate the parameters?

- $\mu$ = Mode of observations to the left of the overall mode
- $\sigma$ = Sample standard deviation for observations to left of overall mode
- $\alpha$ = Mode of observations to the right of the overall mode

# Simulation Experiment

- 100 replications for $n = 100,000$.
- True parameter values of $\mu = 50, 100$, $\sigma = 10, 20$, and $\alpha = 50, 250$.
- Estimate of $\sigma$ is the same as RMA
- Four methods for estimating $\alpha$: Mean, Median, $75^{th}$ percentile, and $99.95^{th}$ percentile of PM values larger than overall mode
- Five methods of estimating $\mu$

# Estimating $\mu$

Estimate $\mu$ with

- Affy method
- Overall mode ($s$) of PM intensities
- Mode of data to the left of $2s$
- Either of the above plus a one-step correction, defined by the formula:

$$\phi\left(\frac{s-\mu}{\sigma}-\alpha\sigma\right)=\alpha\sigma\left[\Phi\left(\frac{s-\mu}{\sigma}-\alpha\sigma\right)\right]$$

# Results

MSE for $\alpha$, when $\mu = 50$, $\sigma = 10$, $\alpha = 50$

Using RMA: 1754

# Results

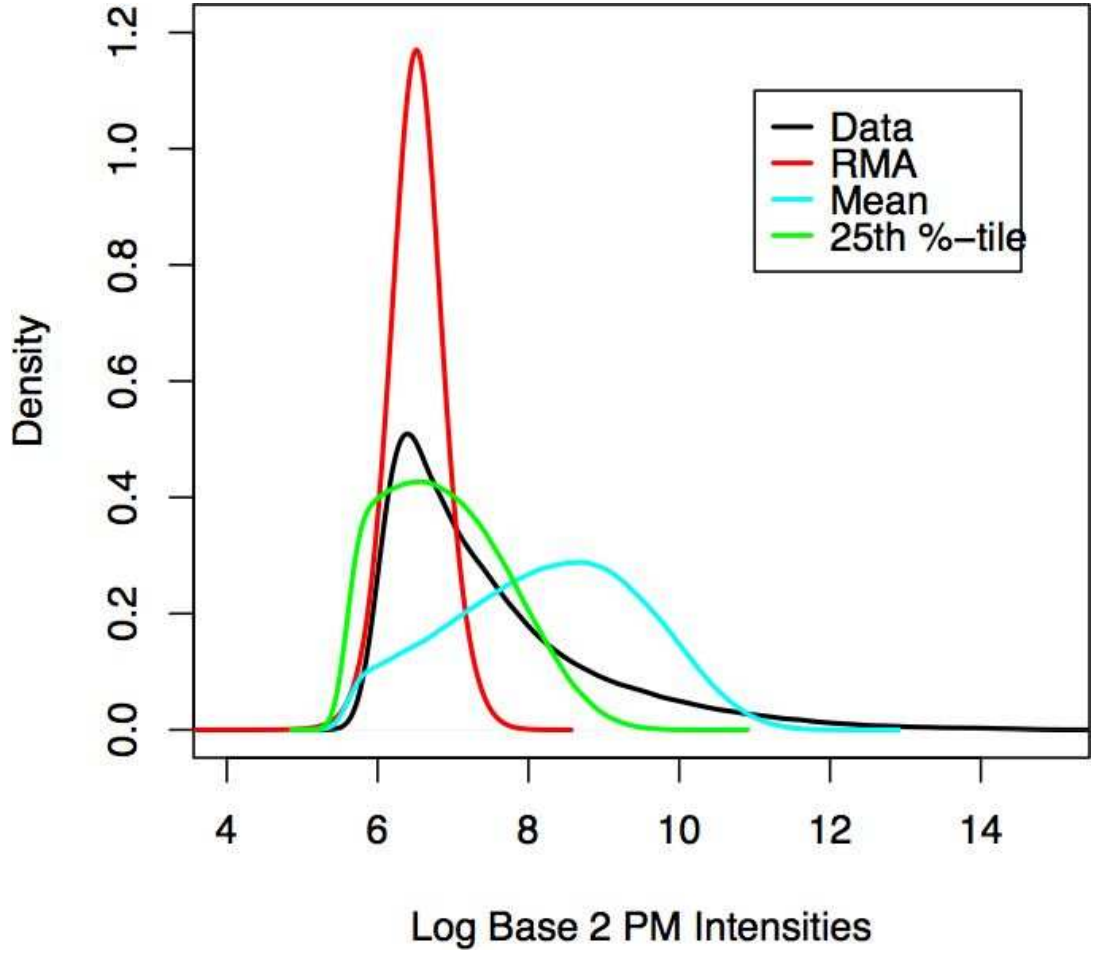
MSE for $\alpha$, when $\mu = 50$, $\sigma = 10$, $\alpha = 50$

Using RMA: 1754

| $\hat{\mu}$ | $\hat{\alpha}$ Given By | | | |
|---|---|---|---|---|
| | Mean | Median | 75% | 99.95% |
| $s$ | 0.413 | 1.117 | 71.45 | 3.111 |
| $s+1$ | 95.97 | 233.9 | 31.72 | 2.378 |
| $2s$ | 0.163 | 0.457 | 103.2 | 4.124 |
| $2s+1$ | 58.69 | 185.3 | 18.18 | 1.926 |

# Performance of Estimates

PM intensities compared to original curve for $\hat{\mu} = 2s + 1$ and various estimates of $\alpha$.




Data: SW 480 cell line with short term treatment.

# An Aside on RMA

RMA has been shown to give results which are

- More precise
- More accurate

compared to more principled approaches.

Hein, *et. al.* BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data, *Biostatistics*, 2005

# Other Ideas Yet Untried

- Fourier or Bootstrap Deconvolution
  (Hall & Qiu 2005, Cordy & Thomas 1997)

# Other Ideas Yet Untried

- Fourier or Bootstrap Deconvolution
  (Hall & Qiu 2005, Cordy & Thomas 1997)
- Nonparametric Approach

# Other Ideas Yet Untried

- Fourier or Bootstrap Deconvolution
  (Hall & Qiu 2005, Cordy & Thomas 1997)
- Nonparametric Approach
  - Find smallest $k_1$% of PM intensities

# Other Ideas Yet Untried

- Fourier or Bootstrap Deconvolution
  (Hall & Qiu 2005, Cordy & Thomas 1997)
- Nonparametric Approach
  - Find smallest $k_1$% of PM intensities
  - Obtain $k_2$% of corresponding MM intensities

# Other Ideas Yet Untried

- Fourier or Bootstrap Deconvolution
  (Hall & Qiu 2005, Cordy & Thomas 1997)
- Nonparametric Approach
  - Find smallest $k_1$% of PM intensities
  - Obtain $k_2$% of corresponding MM intensities
  - MM intensities are an estimate of background noise

# Other Ideas Yet Untried

- Fourier or Bootstrap Deconvolution
  (Hall & Qiu 2005, Cordy & Thomas 1997)
- Nonparametric Approach
  - Find smallest $k_1$% of PM intensities
  - Obtain $k_2$% of corresponding MM intensities
  - MM intensities are an estimate of background noise
- Model PM intensities as Nonstandard Mixtures
  (*Statistical Science*, 1989)

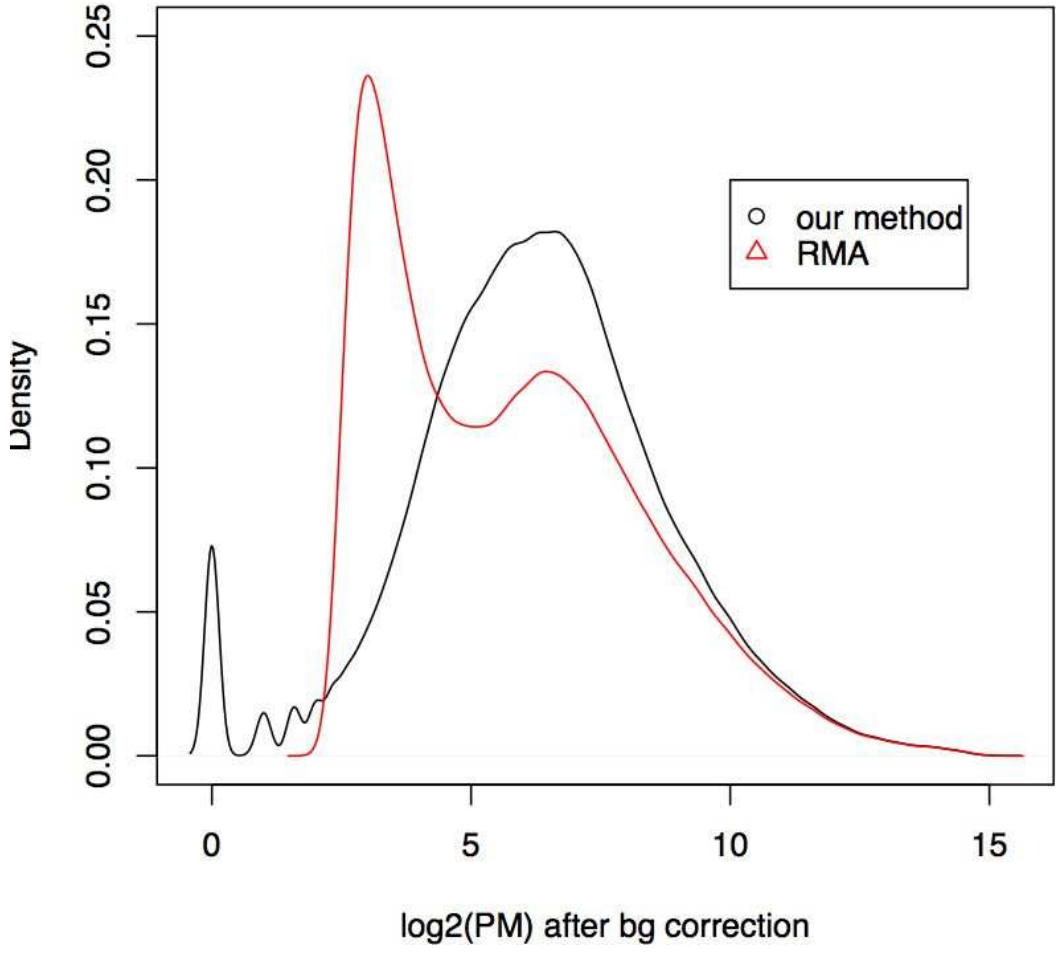
# Other Ideas Yet Untried

- Fourier or Bootstrap Deconvolution
  (Hall & Qiu 2005, Cordy & Thomas 1997)
- Nonparametric Approach
  - Find smallest $k_1$% of PM intensities
  - Obtain $k_2$% of corresponding MM intensities
  - MM intensities are an estimate of background noise
- Model PM intensities as Nonstandard Mixtures
  (*Statistical Science*, 1989)

$$X = S + Y, \text{ where } S \sim (1-p)\delta_0 + pF(x)$$

# Some Preliminary Results

Nonparametric Correction with $k_1 = 0.005$ and $k_2 = 0.975$ vs. RMA Correction




Data: SW480 Cell Line with Short-Term Treatment

# More Work To Do ...

- Does our background correction method result in the “right” answers?
  - Analyze Spike-In Data
  - ROCs
- Methods of Simulating Microarray Data
- Estimating background with non-differentially expressed (or control) genes
- Spatial Correlation in Affymetrix GeneChip Arrays
- Modeling Intensities with a Compound Mixture of Normal Distributions
- Creating pseudo-replicate arrays

# Unanswered Biological Questions

- Gene function annotation
  30,000 genes in human genome
- Biological networks: protein interaction
  Dynamic data of variable quality
- Comparative genomics
  Mapping concepts from organism to organism on a large scale

# Statistical Challenges

- Enormous amount of Data

# Statistical Challenges

- Enormous amount of Data
- Current methods are somewhat *ad-hoc*

# Statistical Challenges

- Enormous amount of Data
- Current methods are somewhat *ad-hoc*
- Data integration and visualization

# Statistical Challenges

- Enormous amount of Data
- Current methods are somewhat *ad-hoc*
- Data integration and visualization
- Data has variable specificity

# Statistical Challenges

- Enormous amount of Data
- Current methods are somewhat *ad-hoc*
- Data integration and visualization
- Data has variable specificity
- Dynamic nature of data

# Statistical Challenges

- Enormous amount of Data
- Current methods are somewhat *ad-hoc*
- Data integration and visualization
- Data has variable specificity
- Dynamic nature of data
- Multiple Comparisons

# References


1. Affymetrix Technical Note: Design and Performance of the GeneChip Human Genome U133 Puls 2.0 and Human Genome U133A Plus 2.0 Arrays (2003). www.affymetrix.com .
2. Cordy, C. B. and Thomas, D. R. (1997). Deconvolution of a distribution function. *Journal of the American Statistical Association*, **92**, 1459–65.
3. Hall, Peter and Qiu, Peihua (2005). Discrete-transform approach to deconvolution problems. *Biometrika*, **92**, 135–148.
4. Hein, A. K., Richardson, S., Causton, H., Ambler, G. K., and Green, P. J. (2005). BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data. *Biostatistics*, **6**, 349–373.
5. Irizarry, R. A. , Bolstad, B. M. , Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, **31** (4) e15.
6. Irizarry, R. A. , Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
7. Panel on Nonstandard Mixtures of Distributions (1989). Statistical Models and Analysis in Auditing. *Statistical Science*, **4**, 2-33.