# Affymetrix Probe Level Analysis

Rafael A. Irizarry and Zhijin Wu
Department of Biostatistics, JHU
Johnson and Johnson, 12/5/3

# Contact Information

- e-mails: rafa@jhu.edu, zwu@jhsph.edu

- Personal webpages:
- http://www.biostat.jhsph.edu/~ririzarr
- http://www.biostat.jhsph.edu/~zwu

- 

(http://www.bioconductor.org)

# Acknowledgements


- Paco Martinez-Murillo, Forrest Spencer (JHU)
- Felix Naef (Rockefeller)
- Ben Bolstad, Sandrine Dudoit, Terry Speed (Berkeley)
- Jean Yang (UCSF)
- Robert Gentleman (Harvard)
- Wolfgang Huber (Germany)
- Johnson & Johnson

# Outline

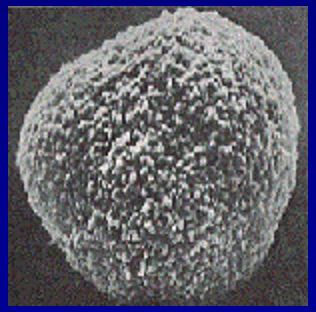
- Quick review of technology
- Overview of Issues
- Previous Work
- RMA
- Improvements to RMA

# Applications of microarrays

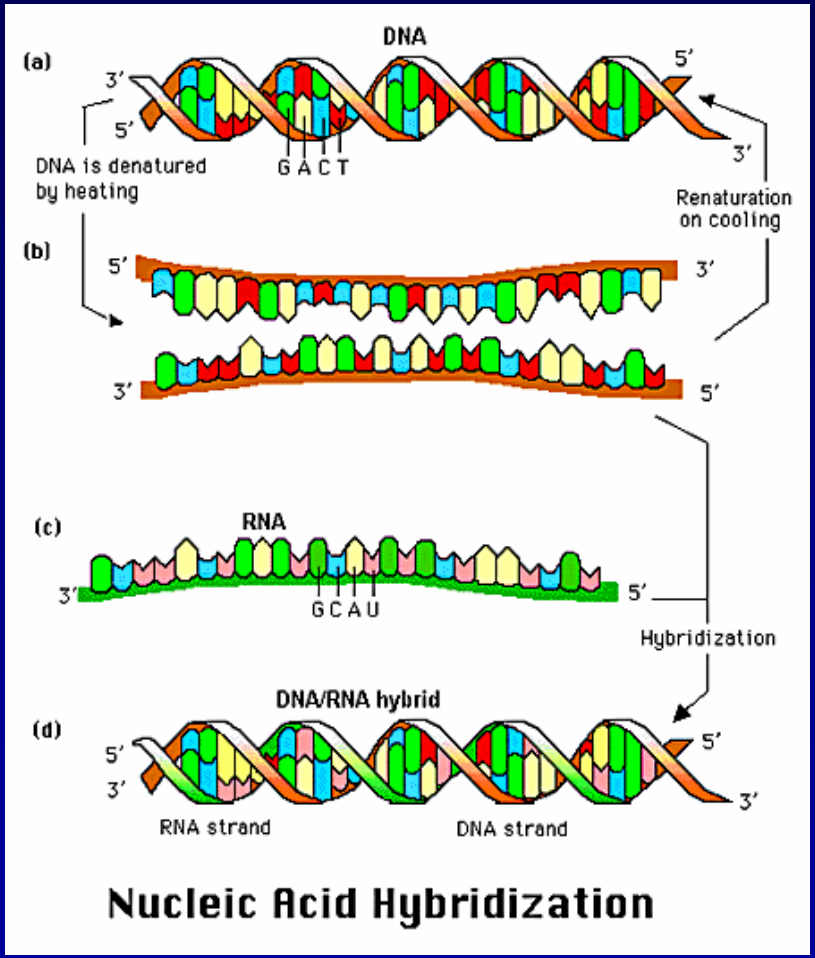
- Measuring transcript abundance
  - Differential Expression
  - Classifying samples
  - Detecting expression pattern
- Other applications:
  - Genotyping
  - TAG arrays

Brain

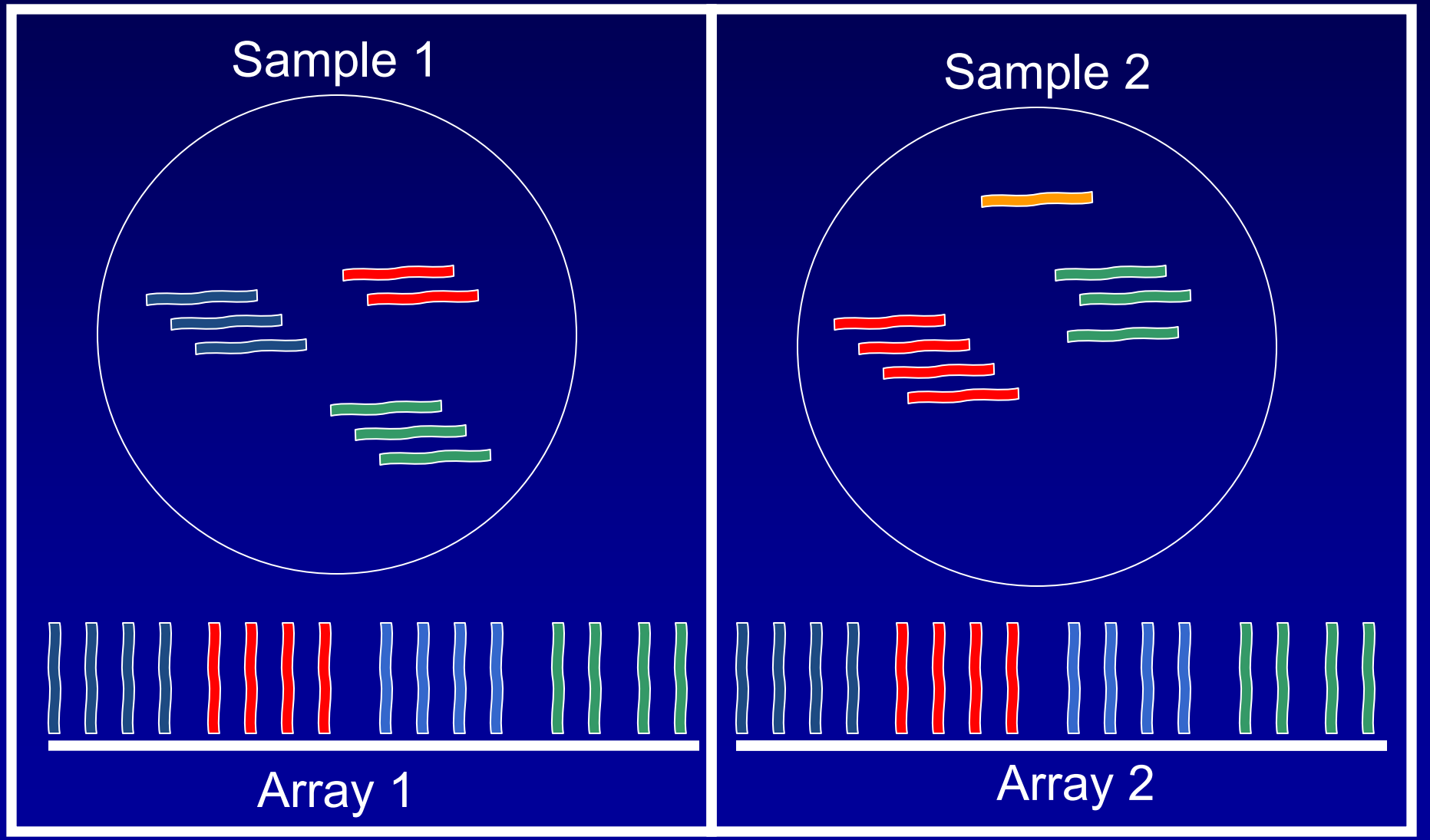
Liver



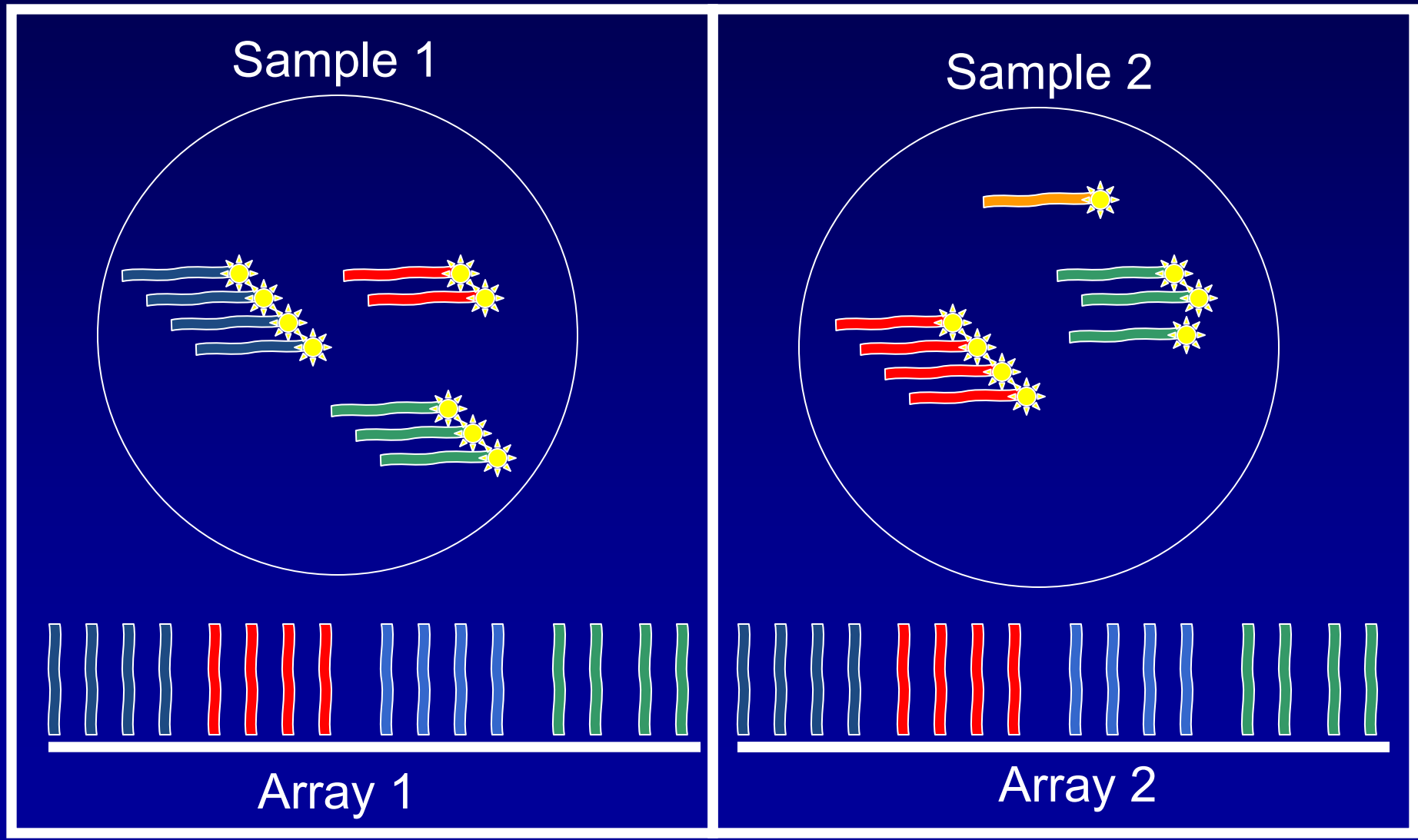
# How they work



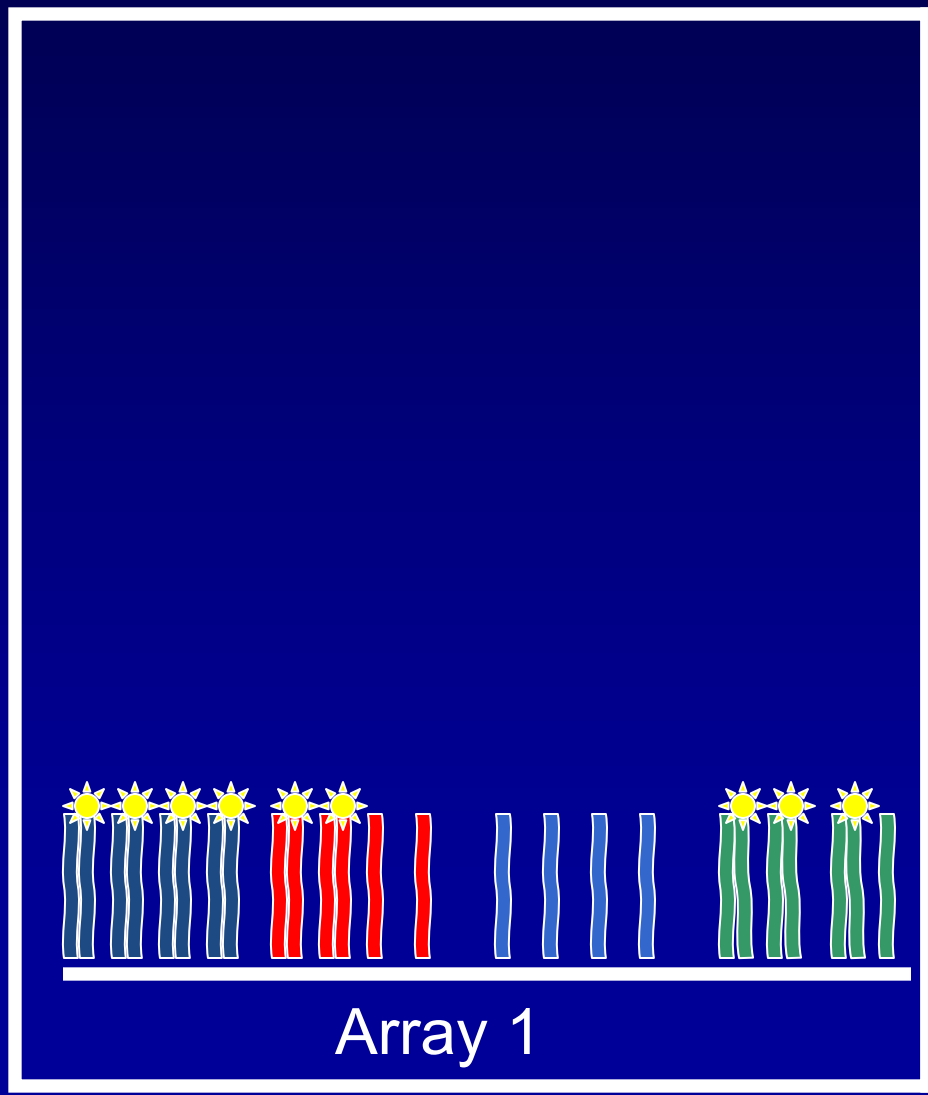
# Before Labelling

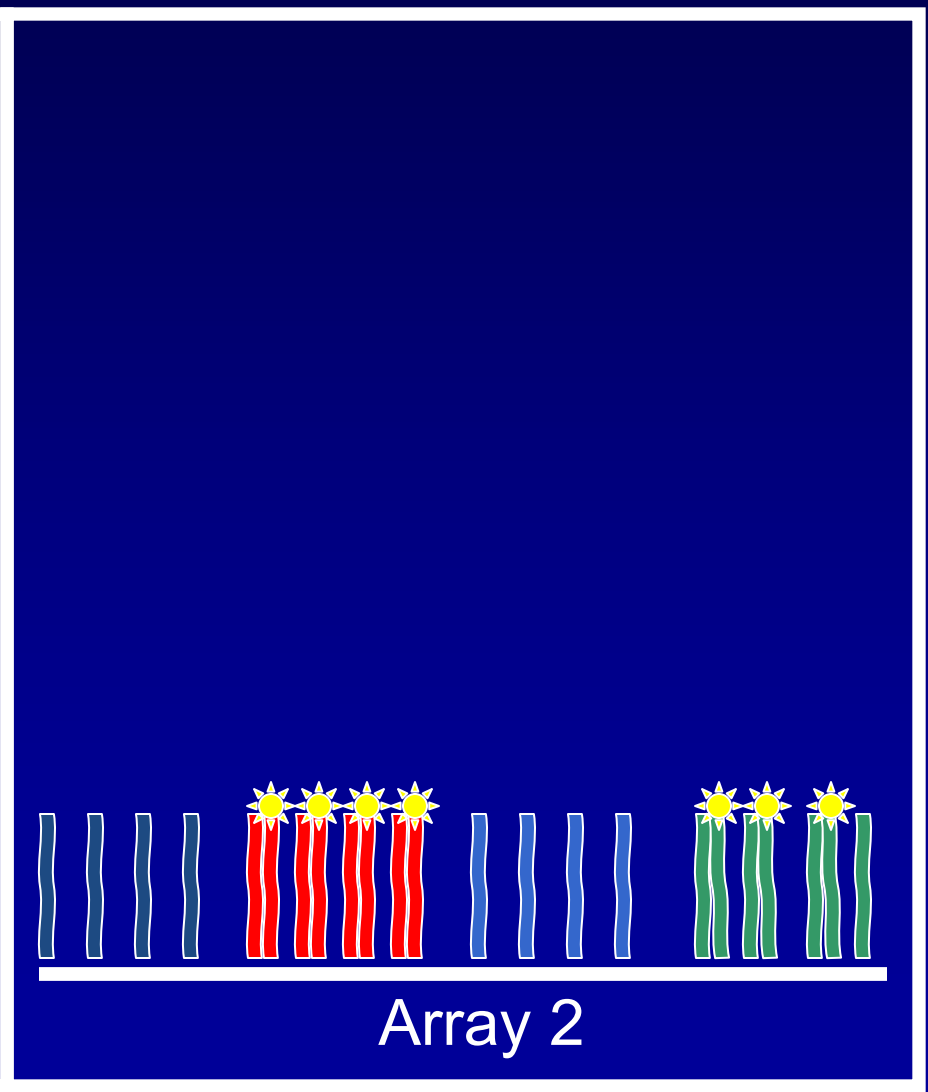


# Before Hybridization

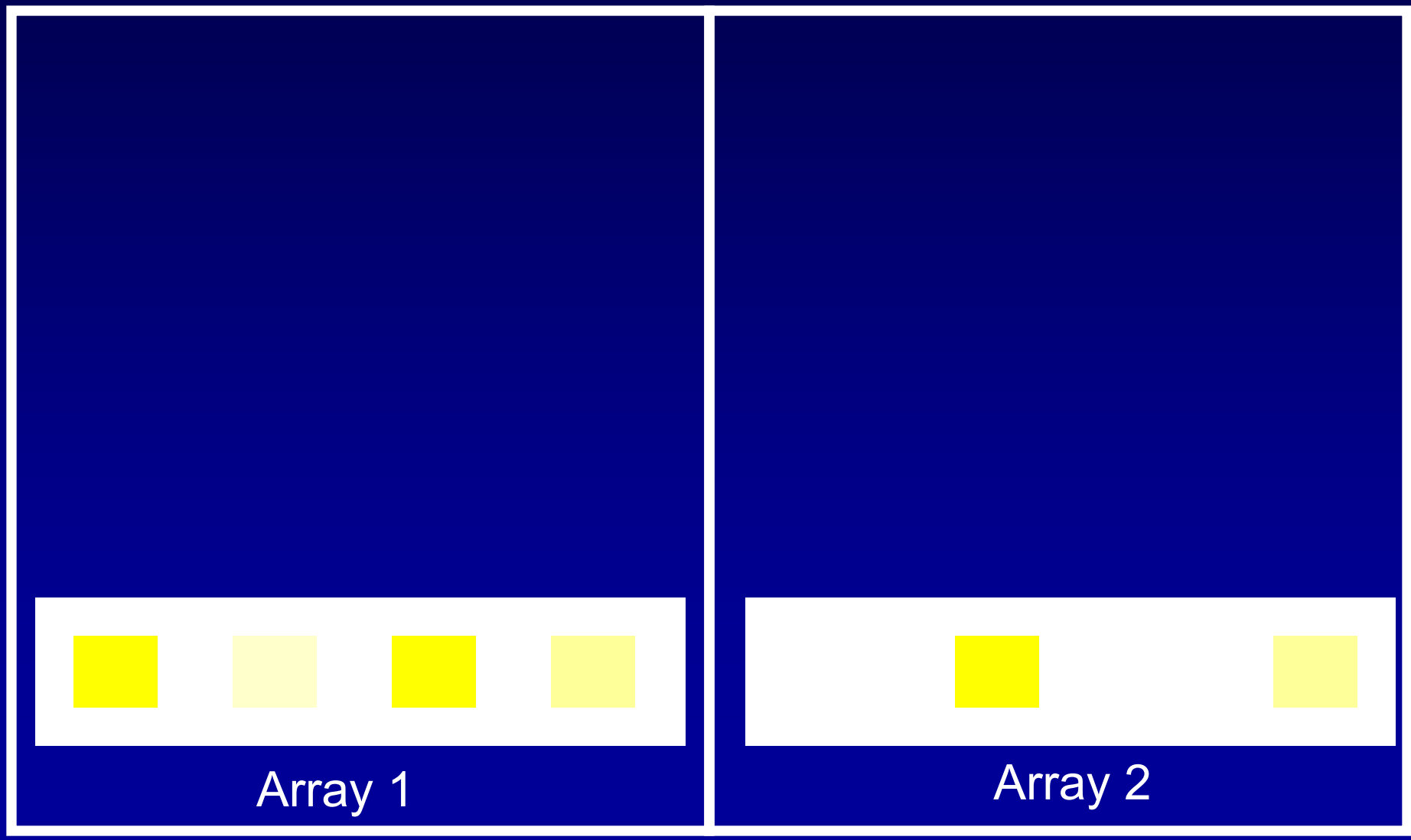


# After Hybridization

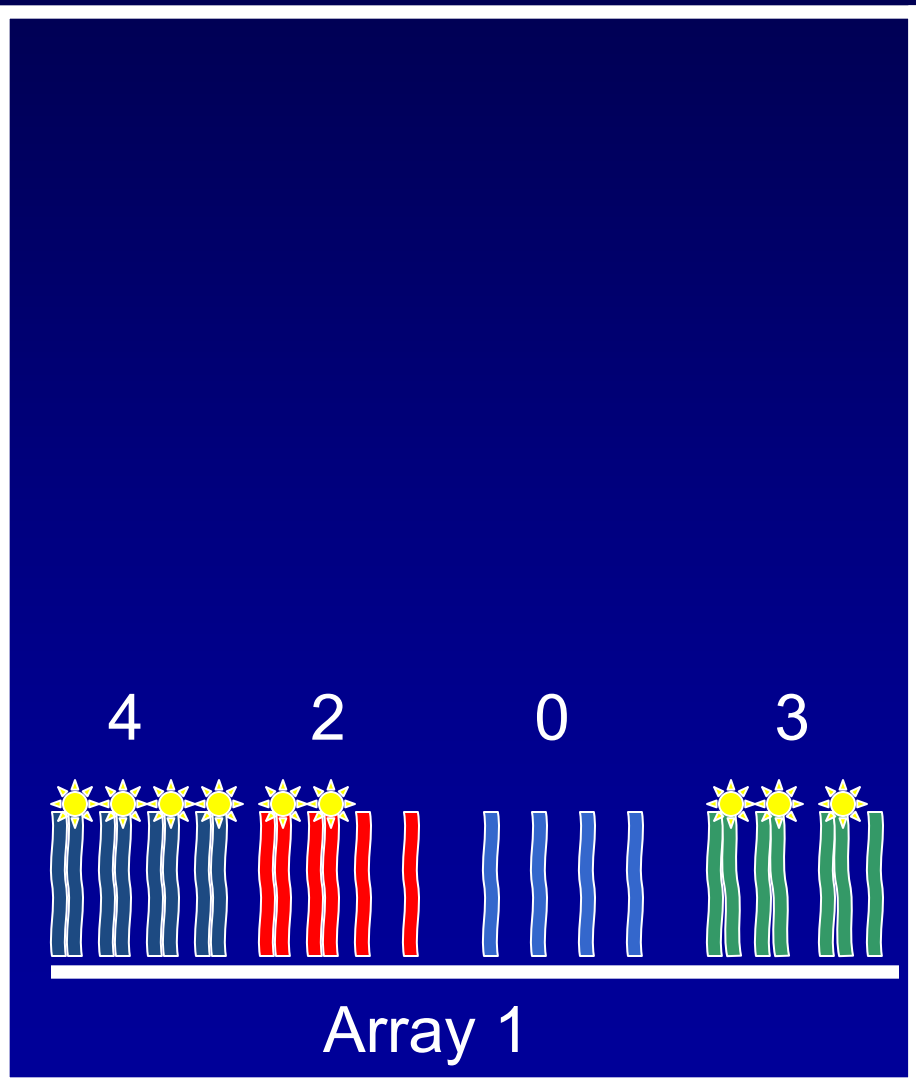


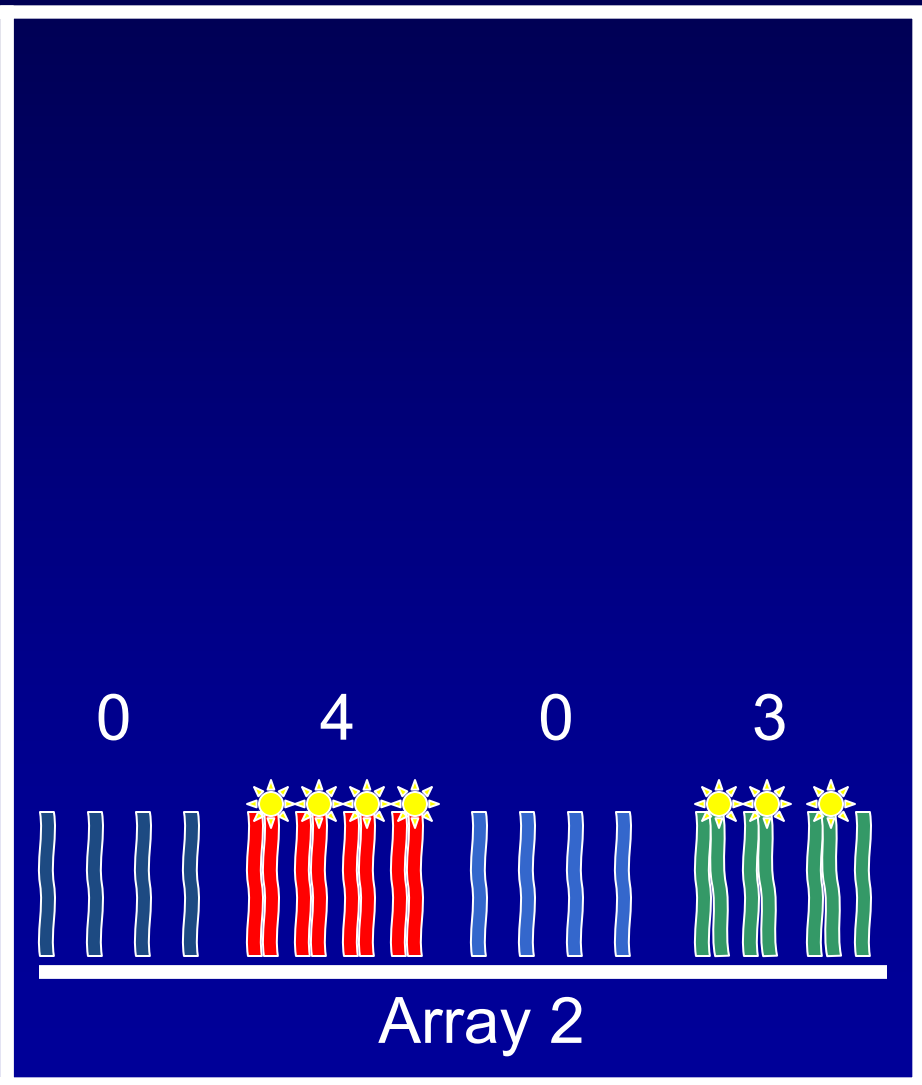



# Scanner Image

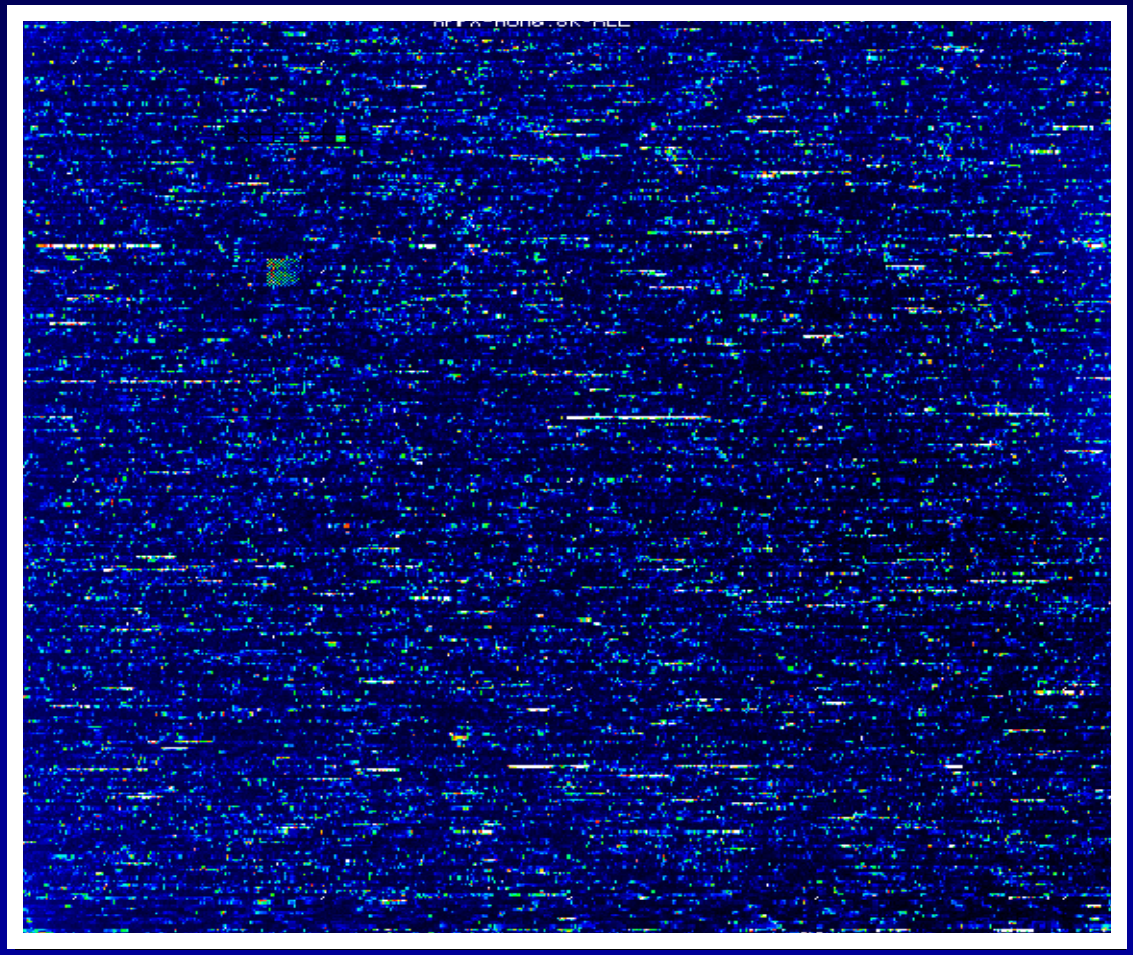


# Quantification

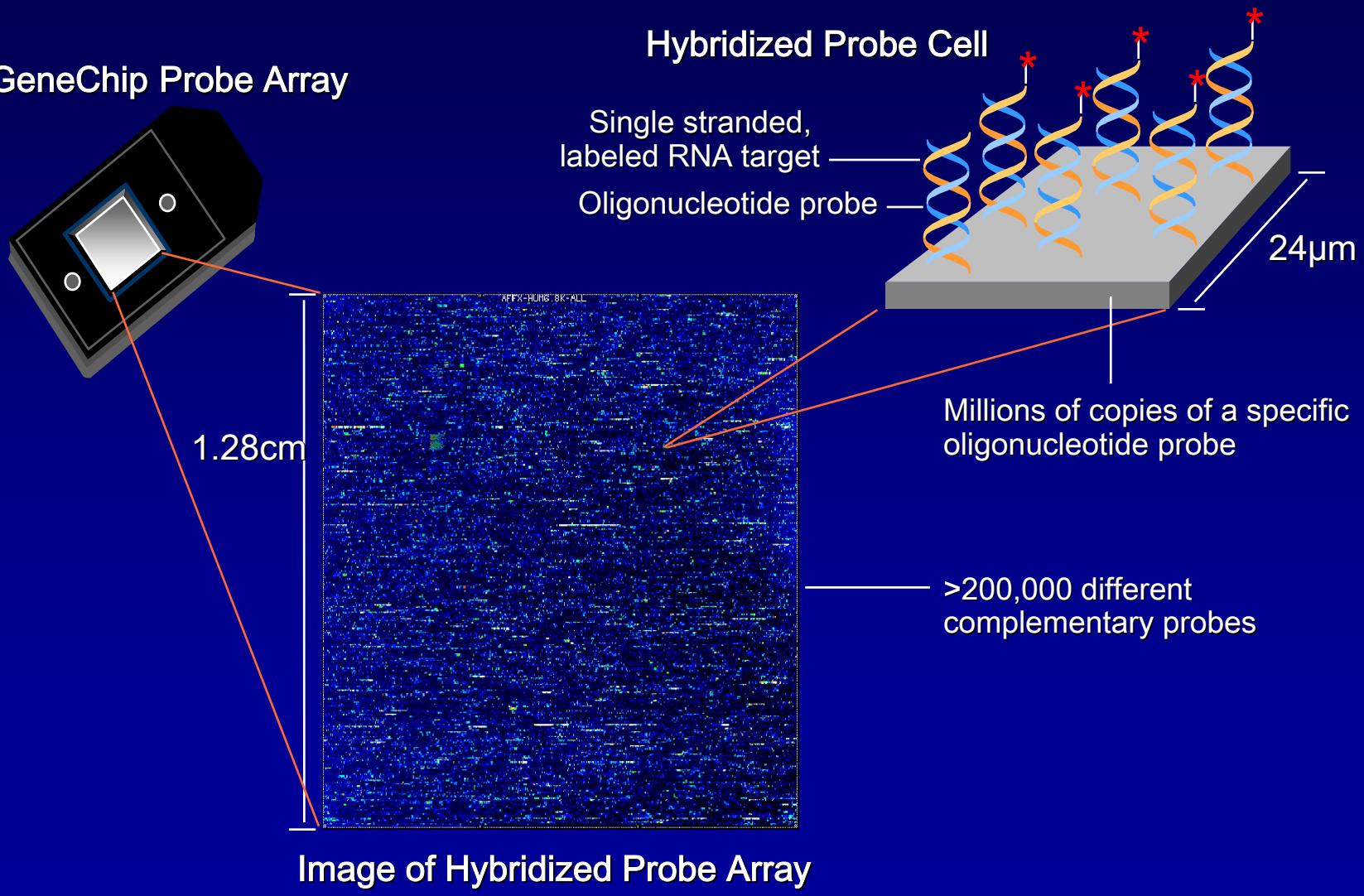


# Microarray Image



# Case Study:
# Preprocessing Affymetrix GeneChip Arrays




*Compliments of D. Gerhold*

# GeneChip® Expression Array Design

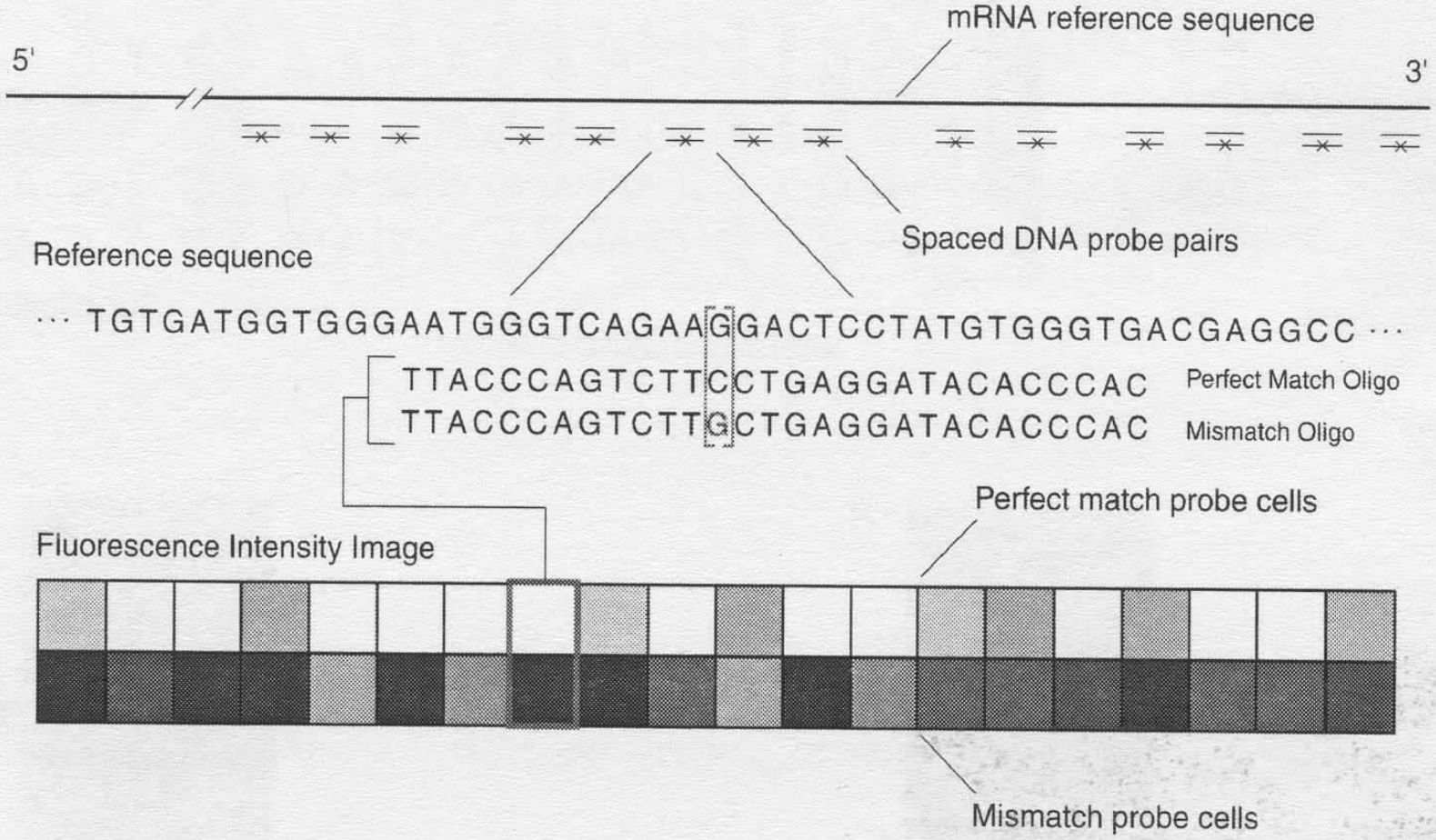
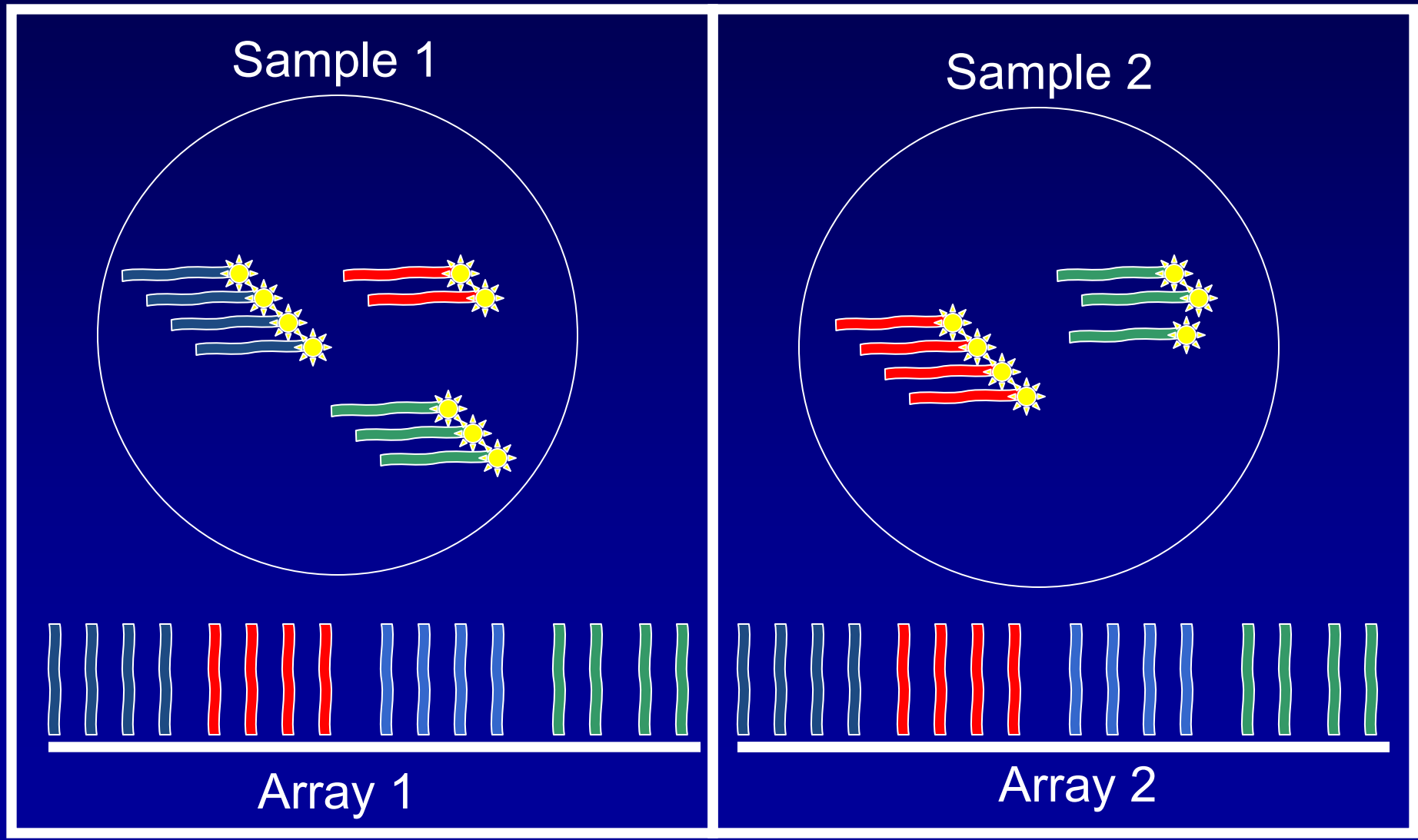
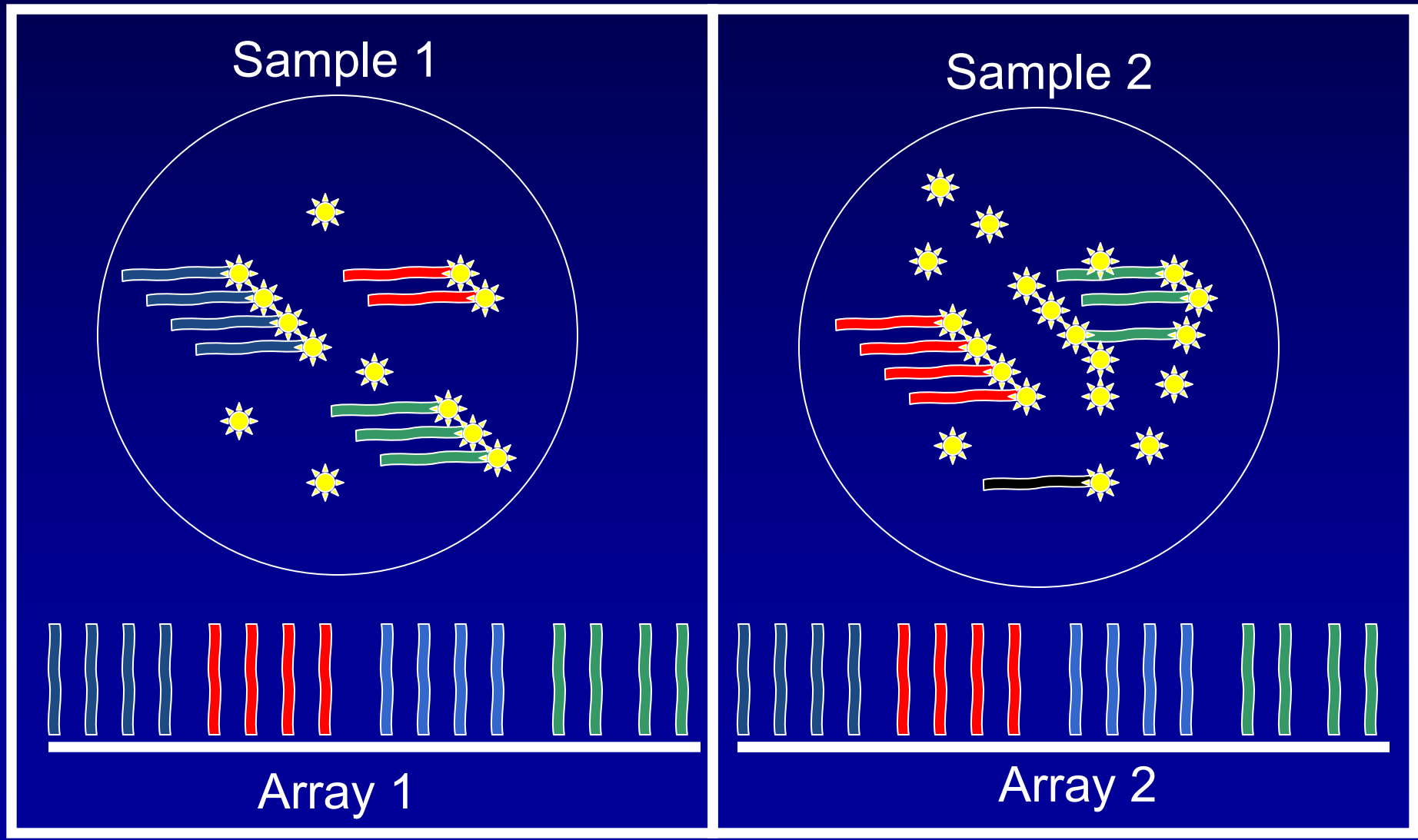



Figure 1-3 Expression tiling strategy

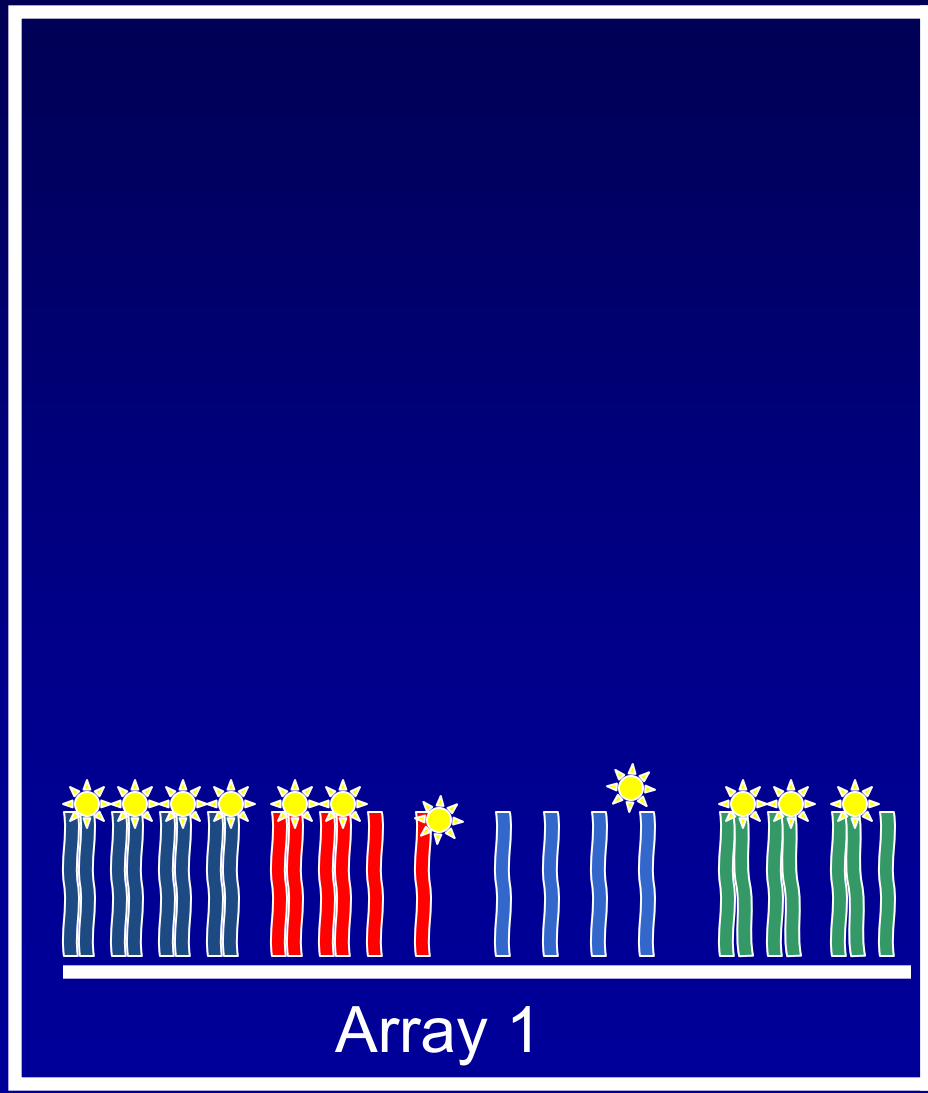
# Before Hybridization

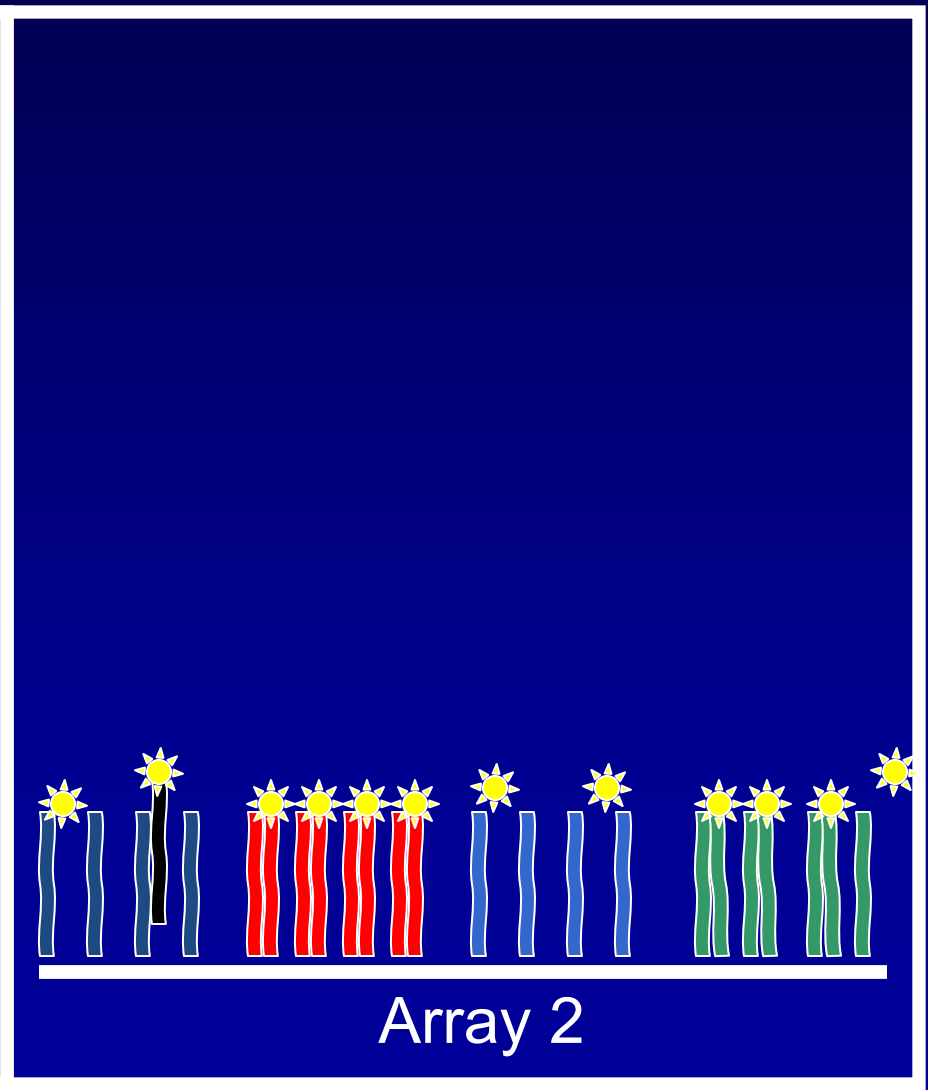


# More Realistic



# Non-specific Hybridization



# GeneChip® Expression Array Design

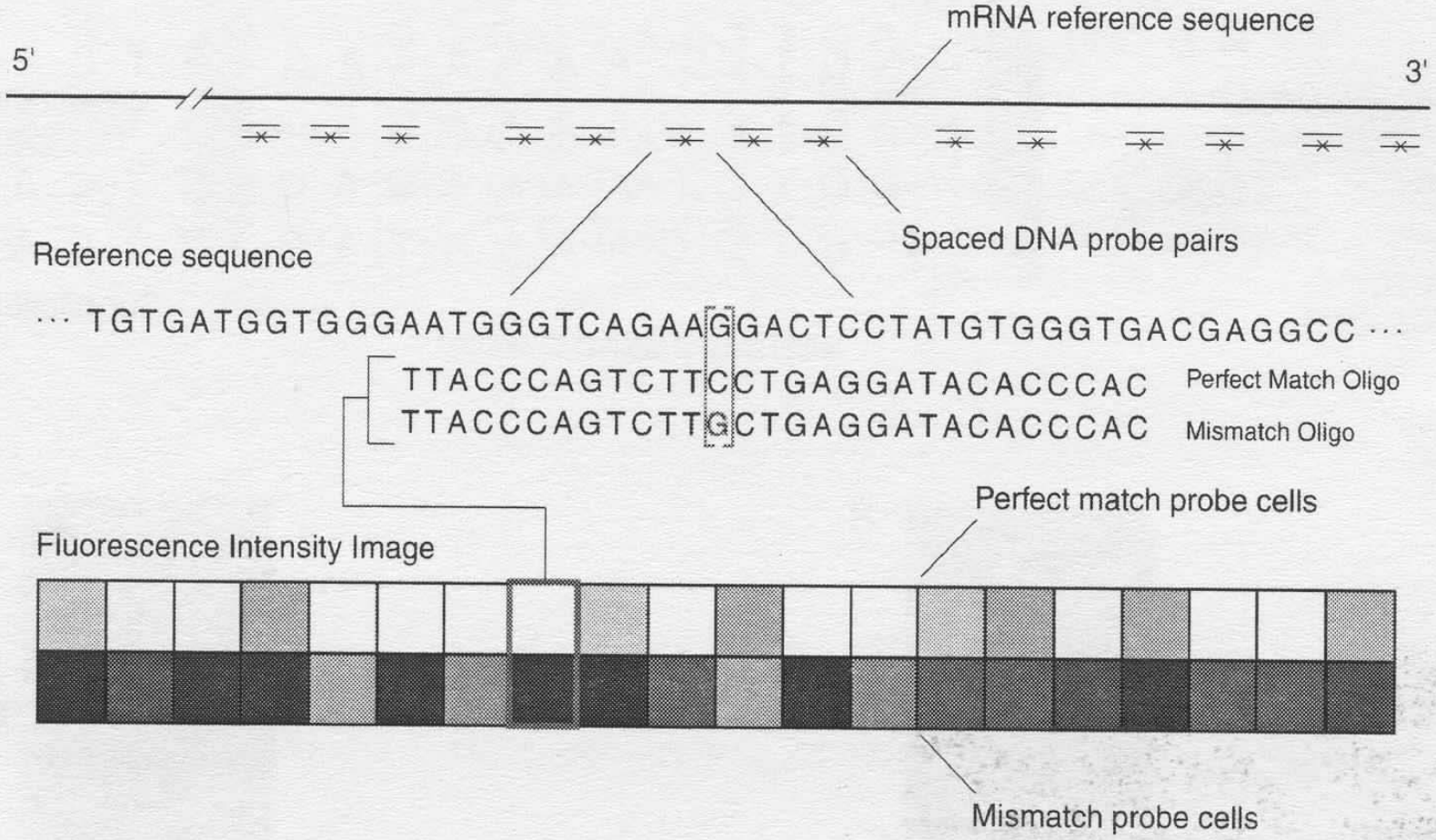



Figure 1-3 Expression tiling strategy

# Statistical Problem

- Each gene is represented by 11-20 pairs (PM and MM) of probe intensities
- Each array has 8K-20K genes
- Usually there are various arrays
- Obtain measure for each gene on each array:

Summarize probeset data

- Background adjustment and normalization are issues

# Default until 2002 (MAS 4.0)

- GeneChip® software used *Avg.diff*

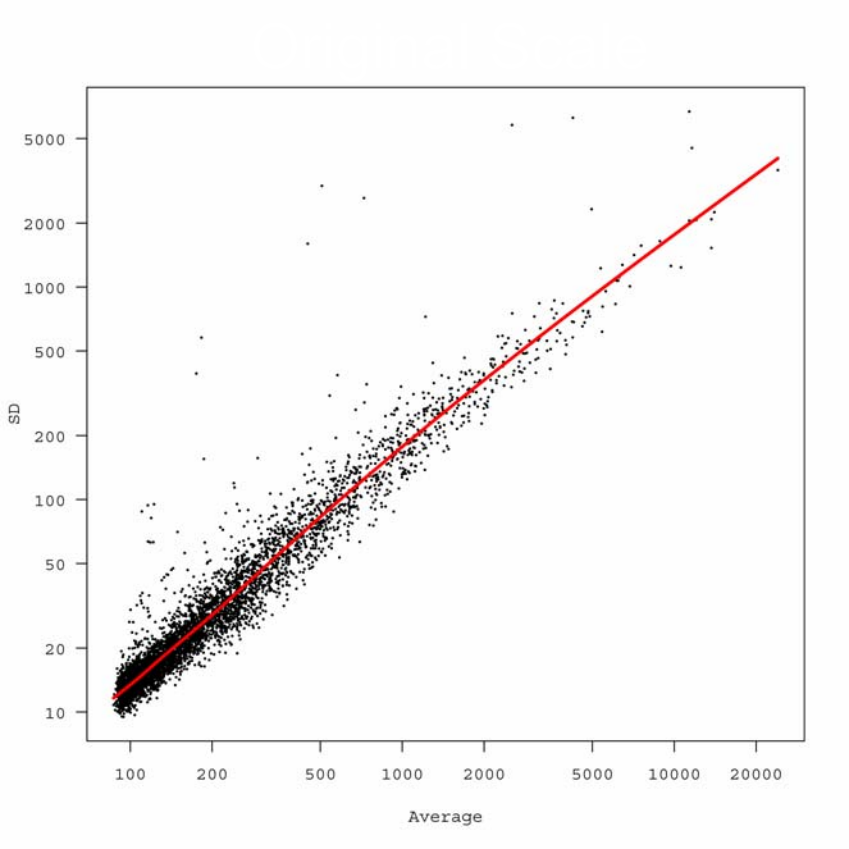
$$Avg.diff = \frac{1}{|\mathrm{A}|} \sum_{j \in \mathrm{A}} (PM_j - MM_j)$$

- with A a set of "suitable" pairs chosen by software.

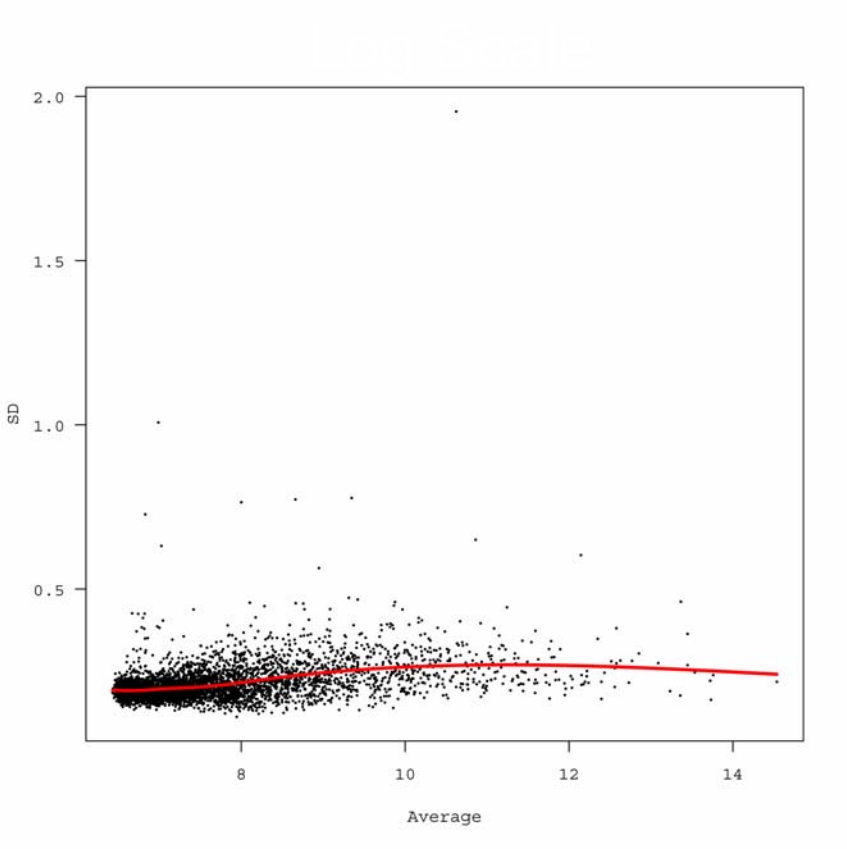
- Obvious Problems:
  - Many negative expression values
  - No log transform

# Why use log?

Original scale




Log scale



# Current default (MAS 5.0)

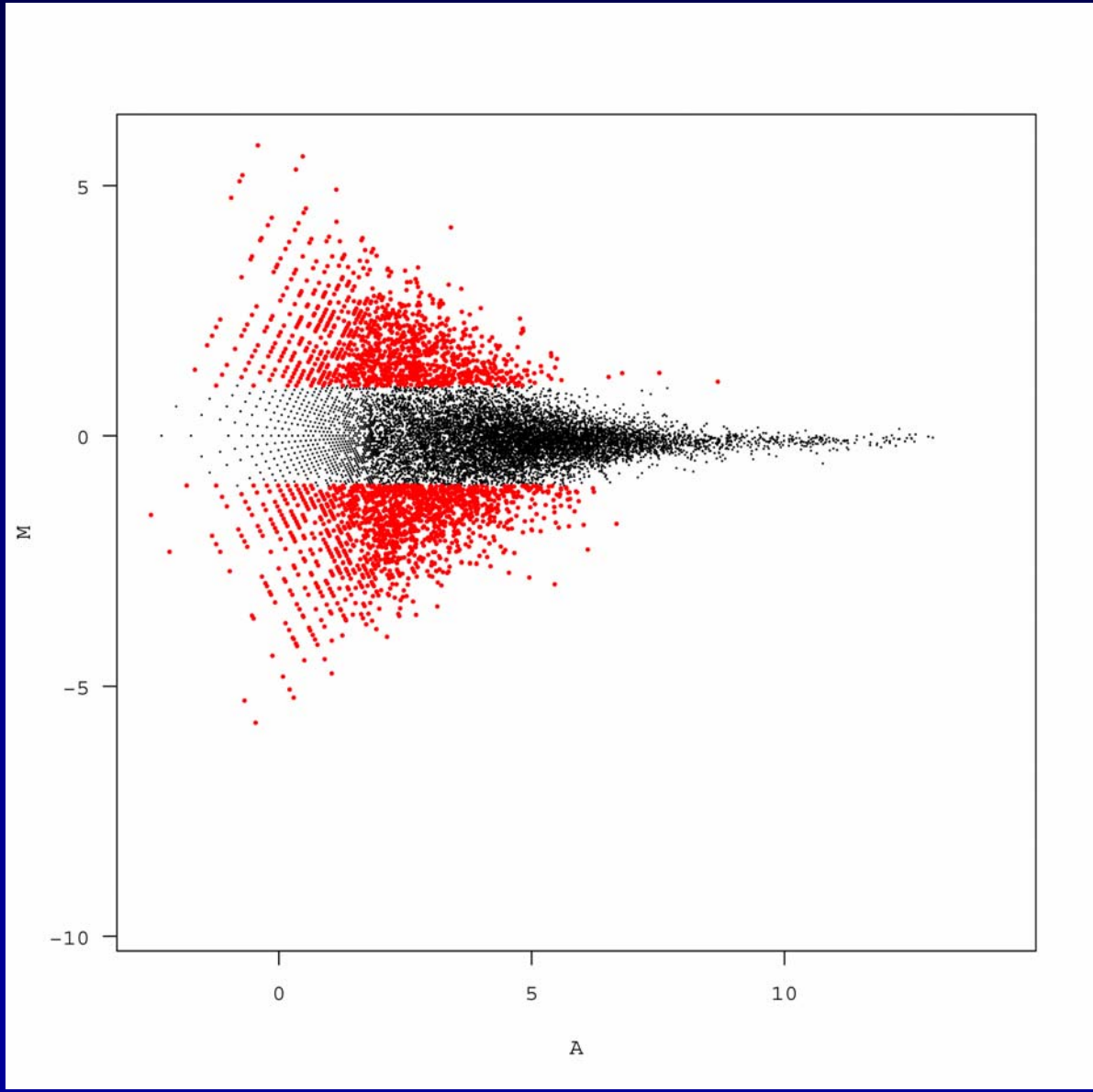
- GeneChip® new version uses something else

$$signal = TukeyBiweight\{\log(PM_j - MM_j^*)\}$$

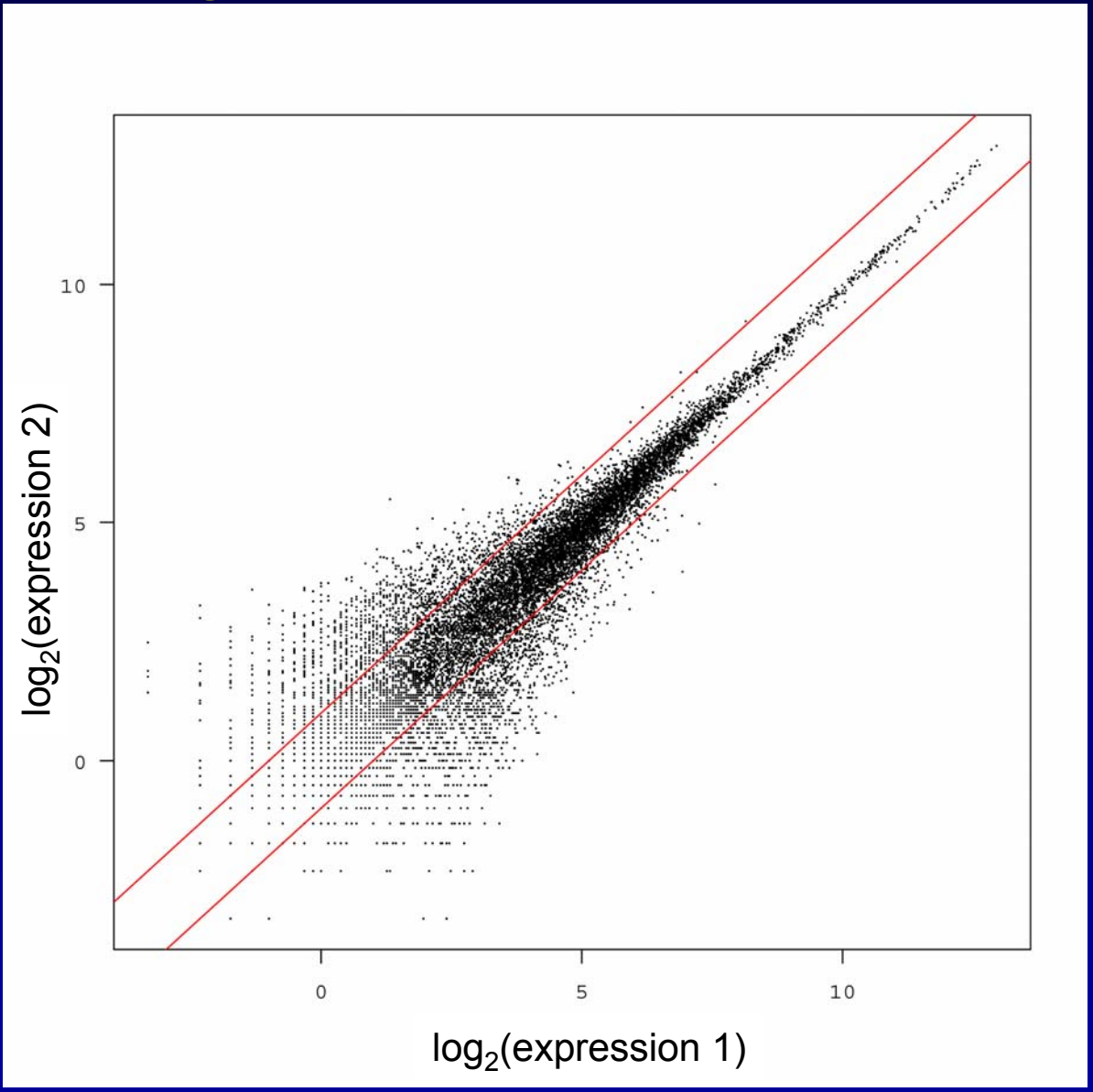
- with *MM** a version of MM that is never bigger than PM.

- Ad-hoc background procedure and scale normalization are used.

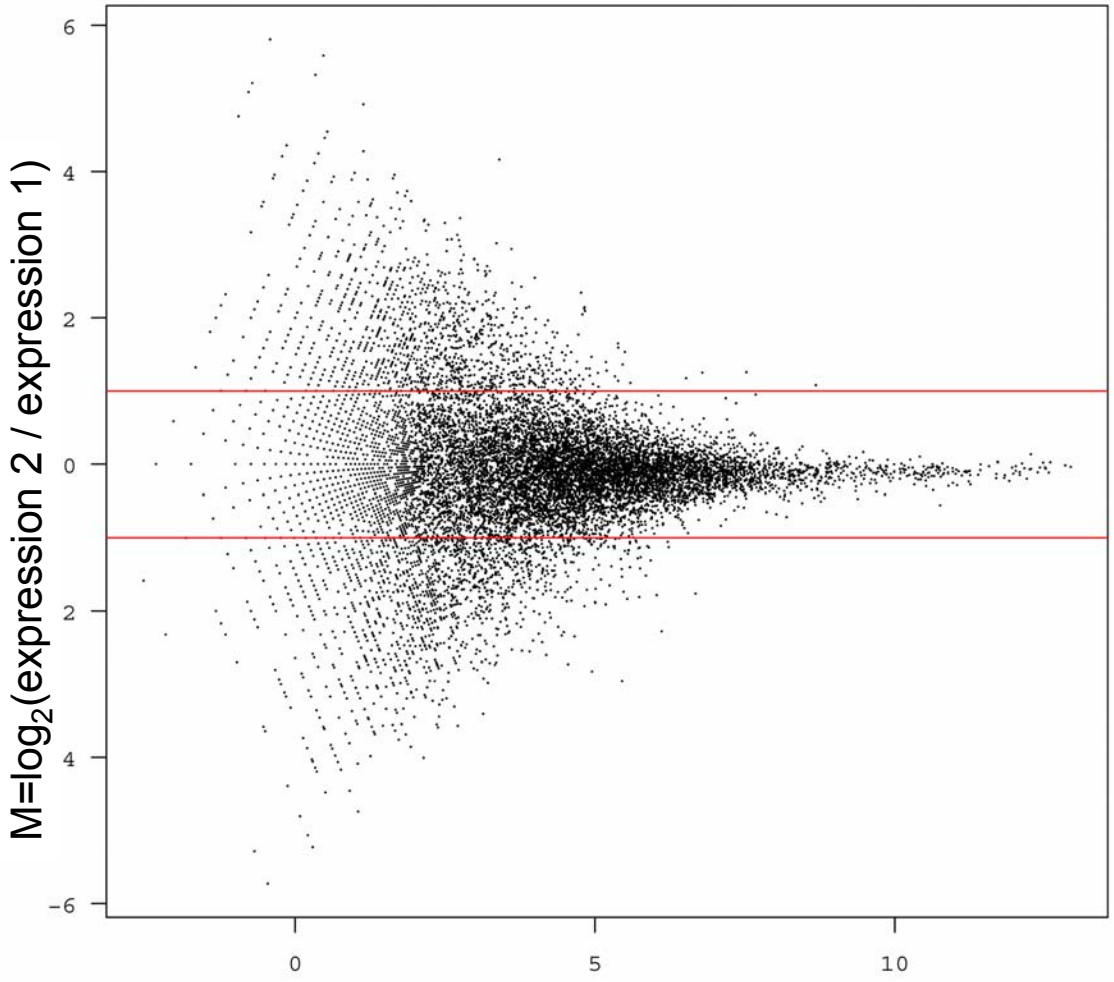
# Can this be improved?



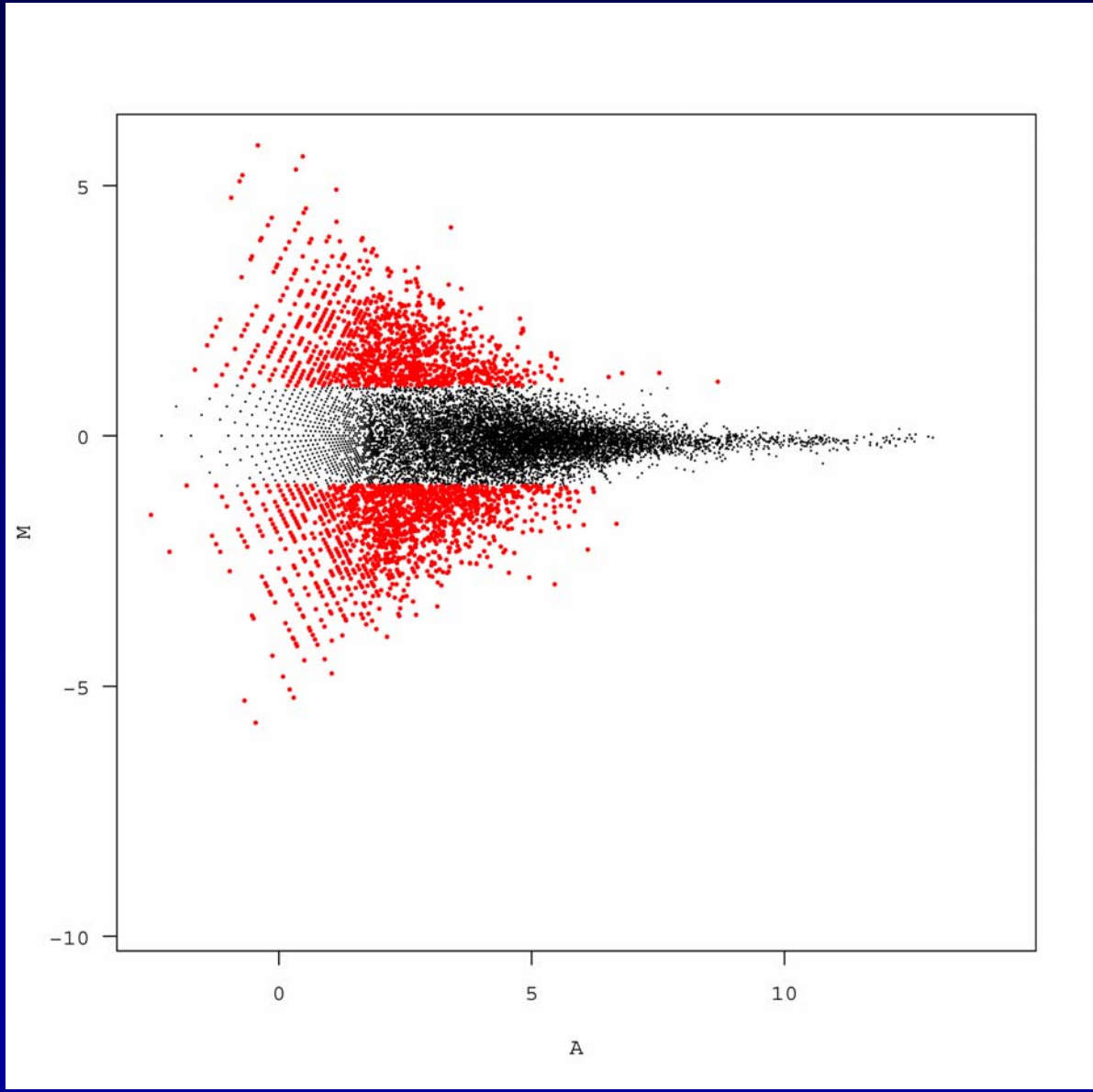
# Log-scale scatter plot



# MvA Plot


6
4
2
0
2
4
−6
0
5
10
M=log2(expression 2 / expression 1)
A= ½{ log2(expression 2) + log2(expression 1) } /2

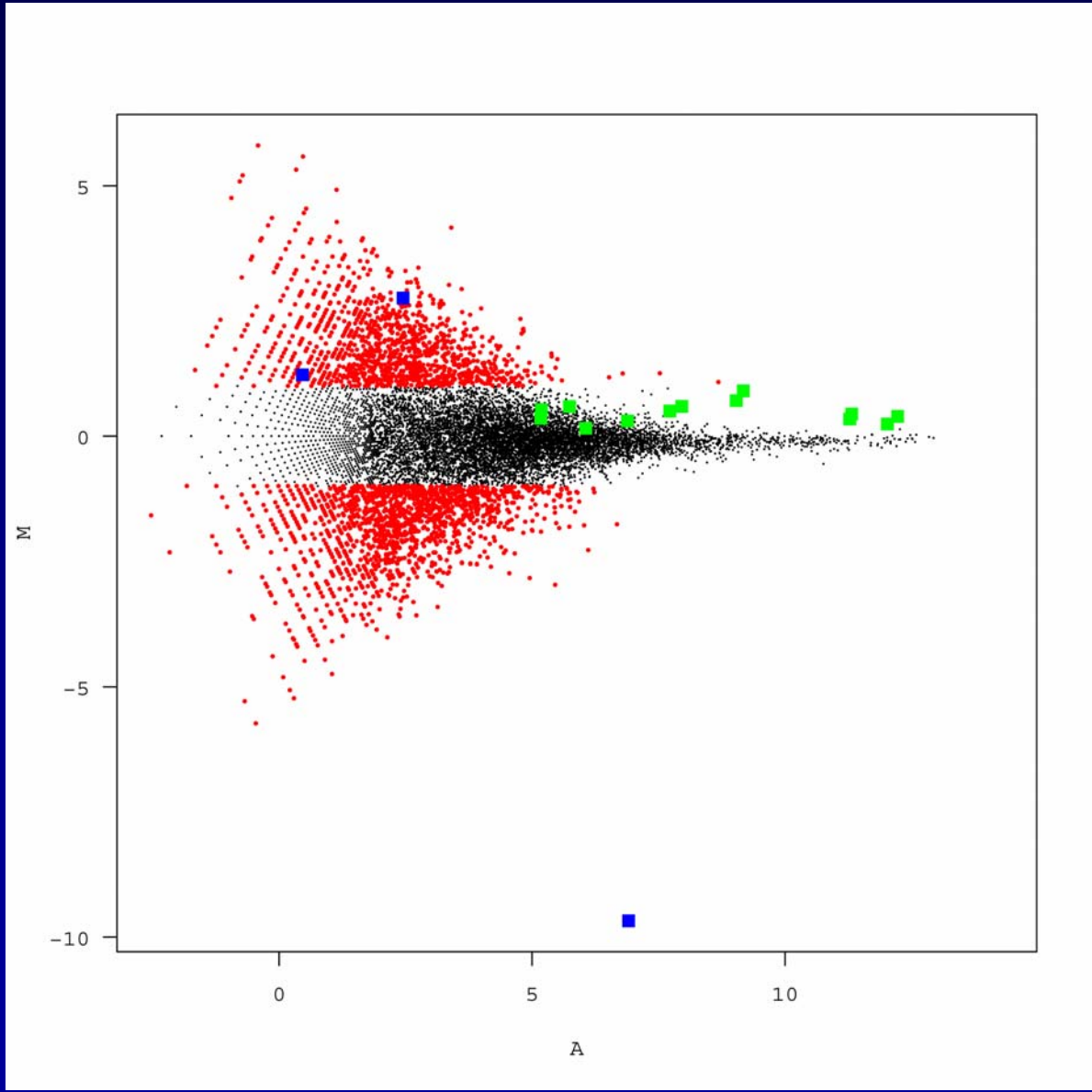
# Can this be improved?



# Precision/Accuracy

- It appears precision can be improved. How does it relate to accuracy?
- Spike-in experiments (Affymetrix and GeneLogic)
- Dilution Study (GeneLogic)

# Use Spike-In Experiment



# First academic alternative: dChip

Li and Wong fit a model

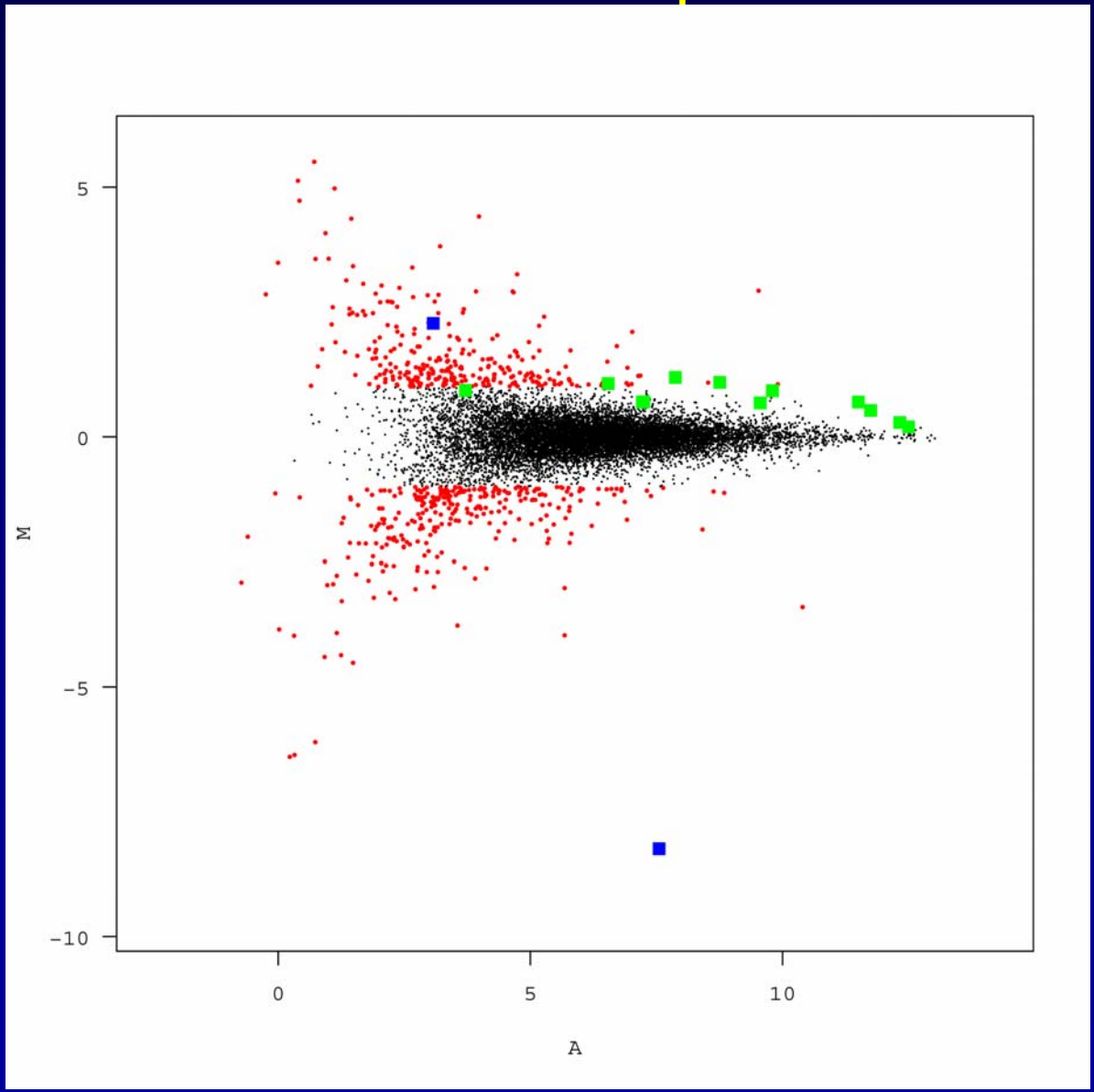
$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}, \varepsilon_{ij} \propto N(0, \sigma^2)$$

Here $\theta_i$ represents expression on chip *i*
and $\phi_j$ represents the *probe effect*

A non-linear normalization technique is used and the model assumptions are used to remove outliers.

# dChip is better

## but still room for improvement

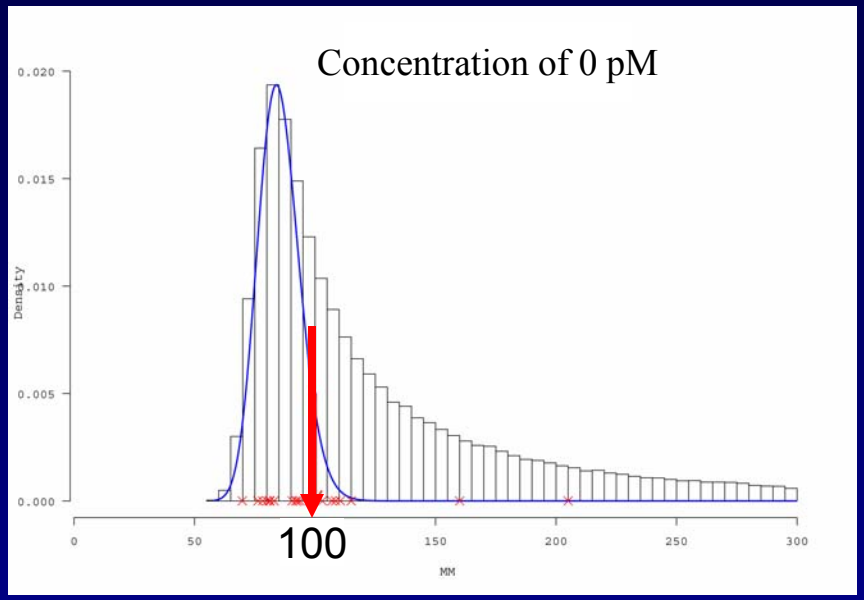


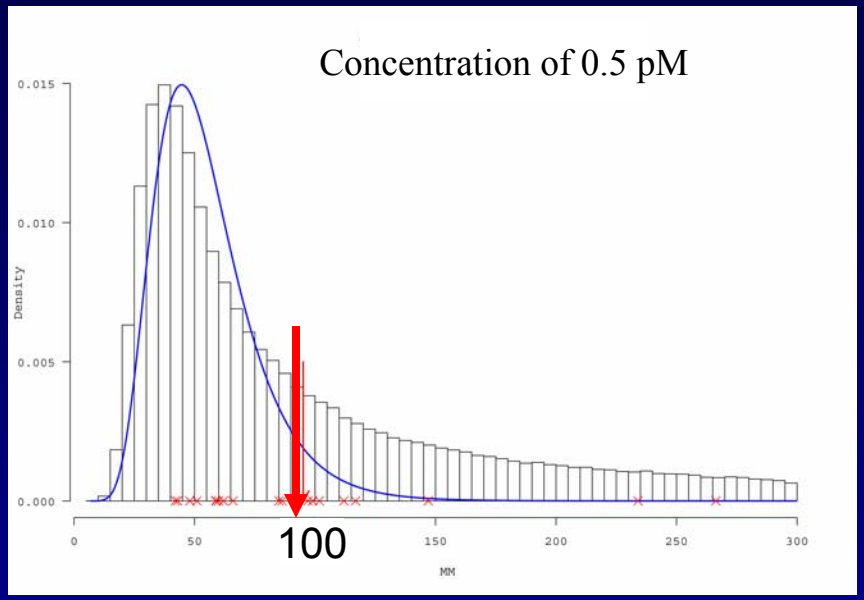
# Three steps

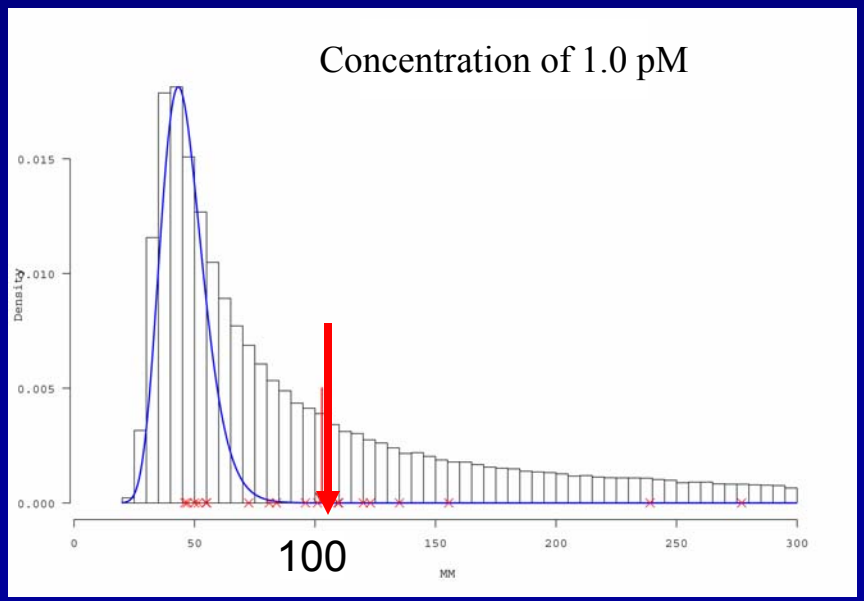
From the spike-in data we learn that:

- We need to background adjust
- Normalize
- Summarize appropriately (in the log-scale)

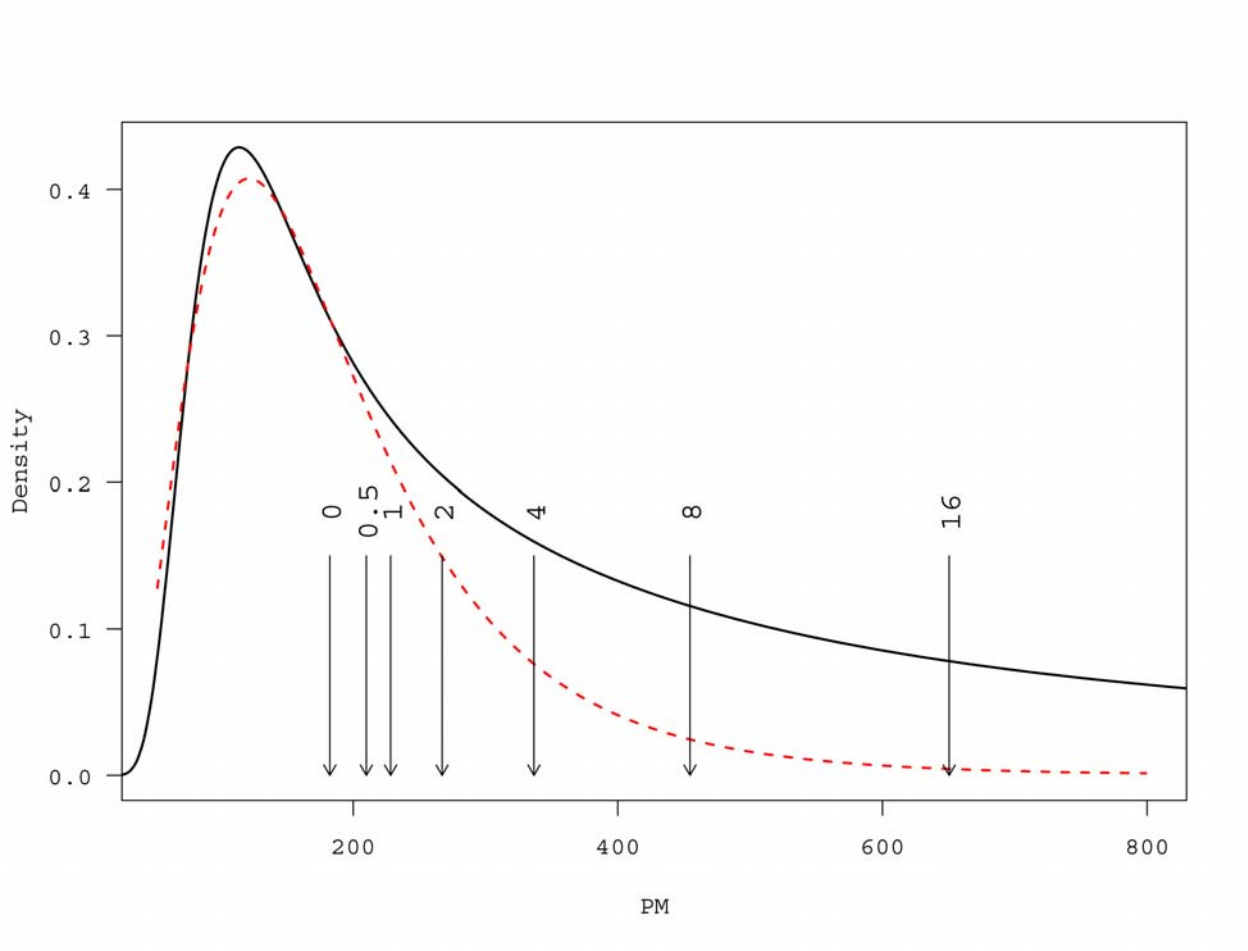
# Why background correct?



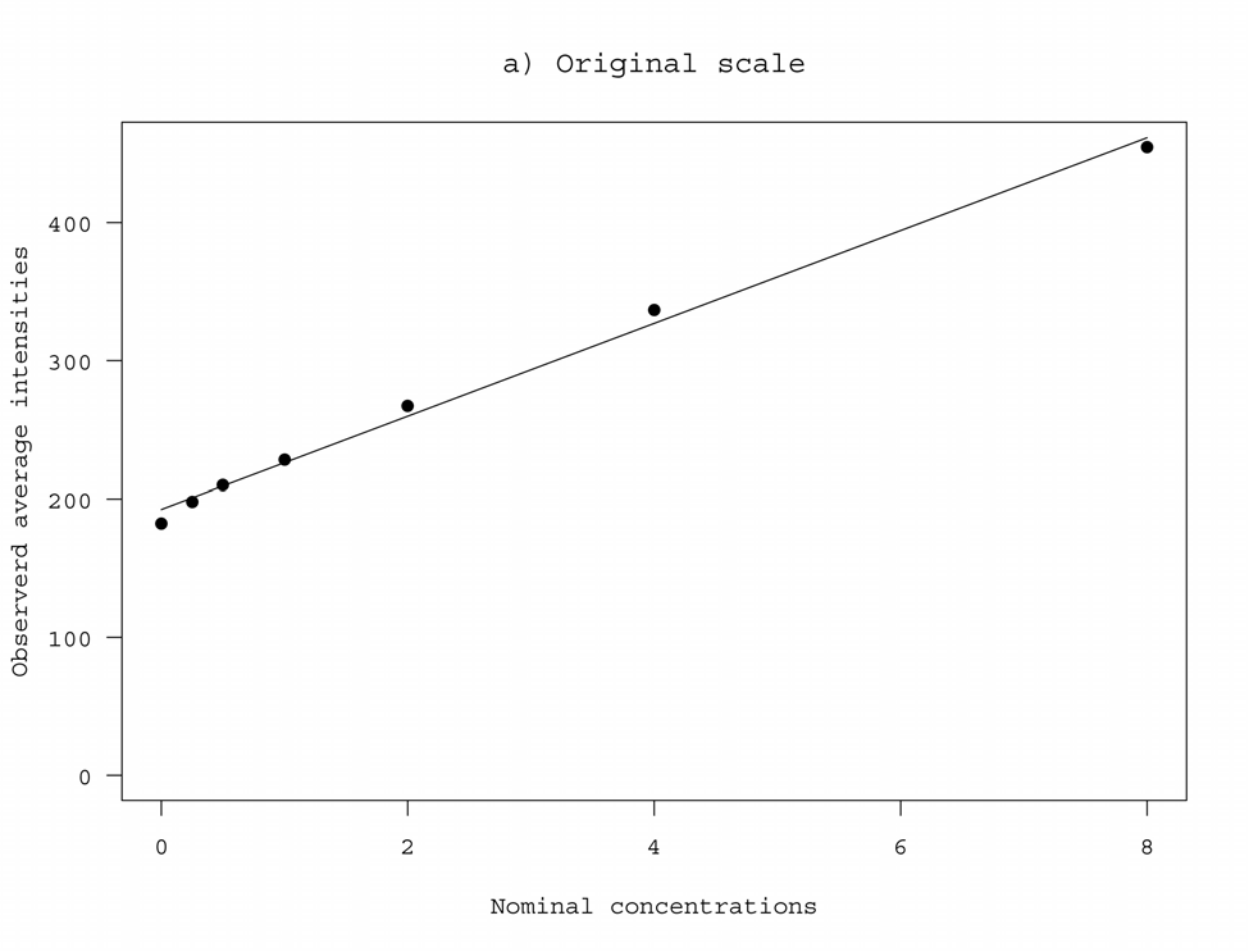



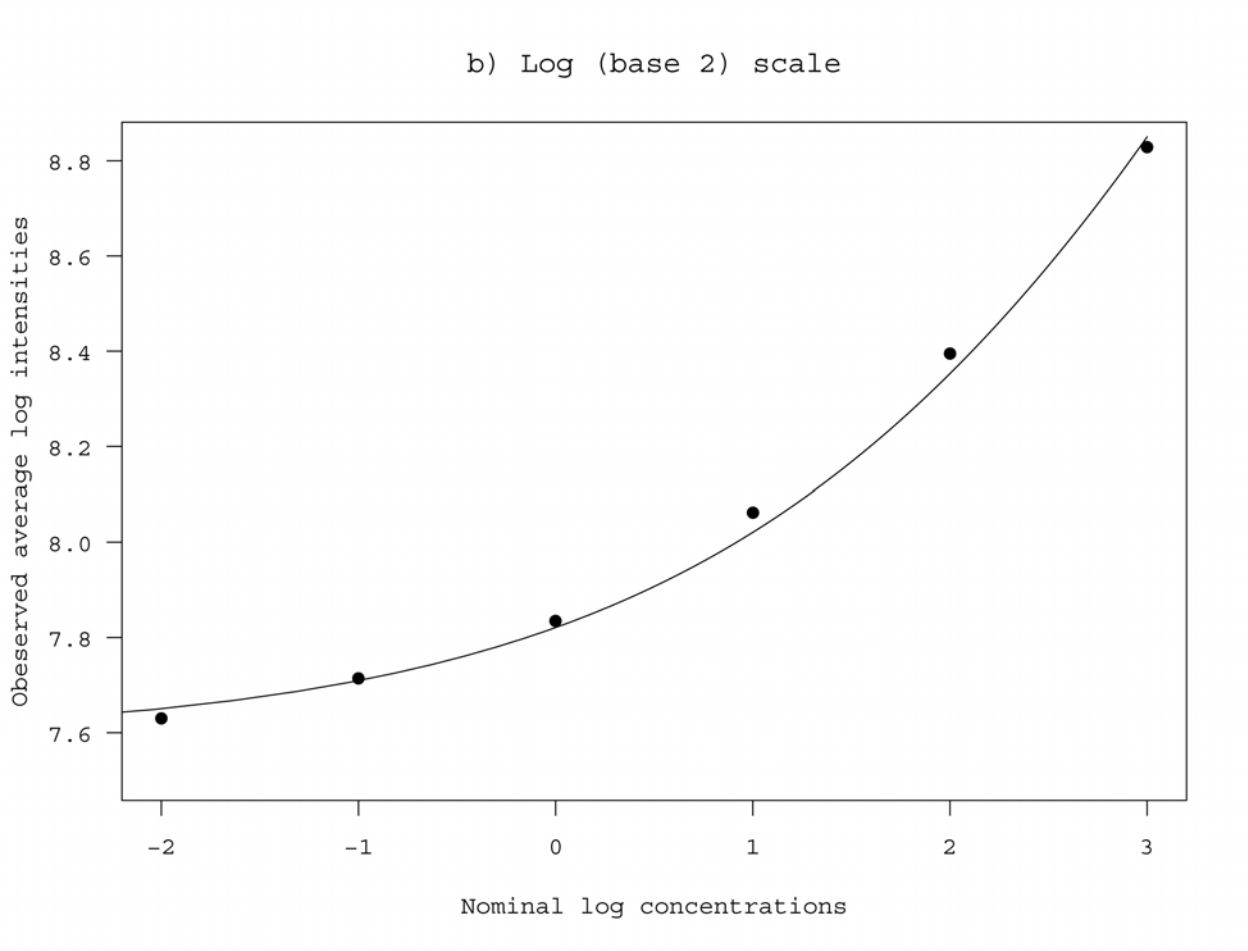
# Why background correct?



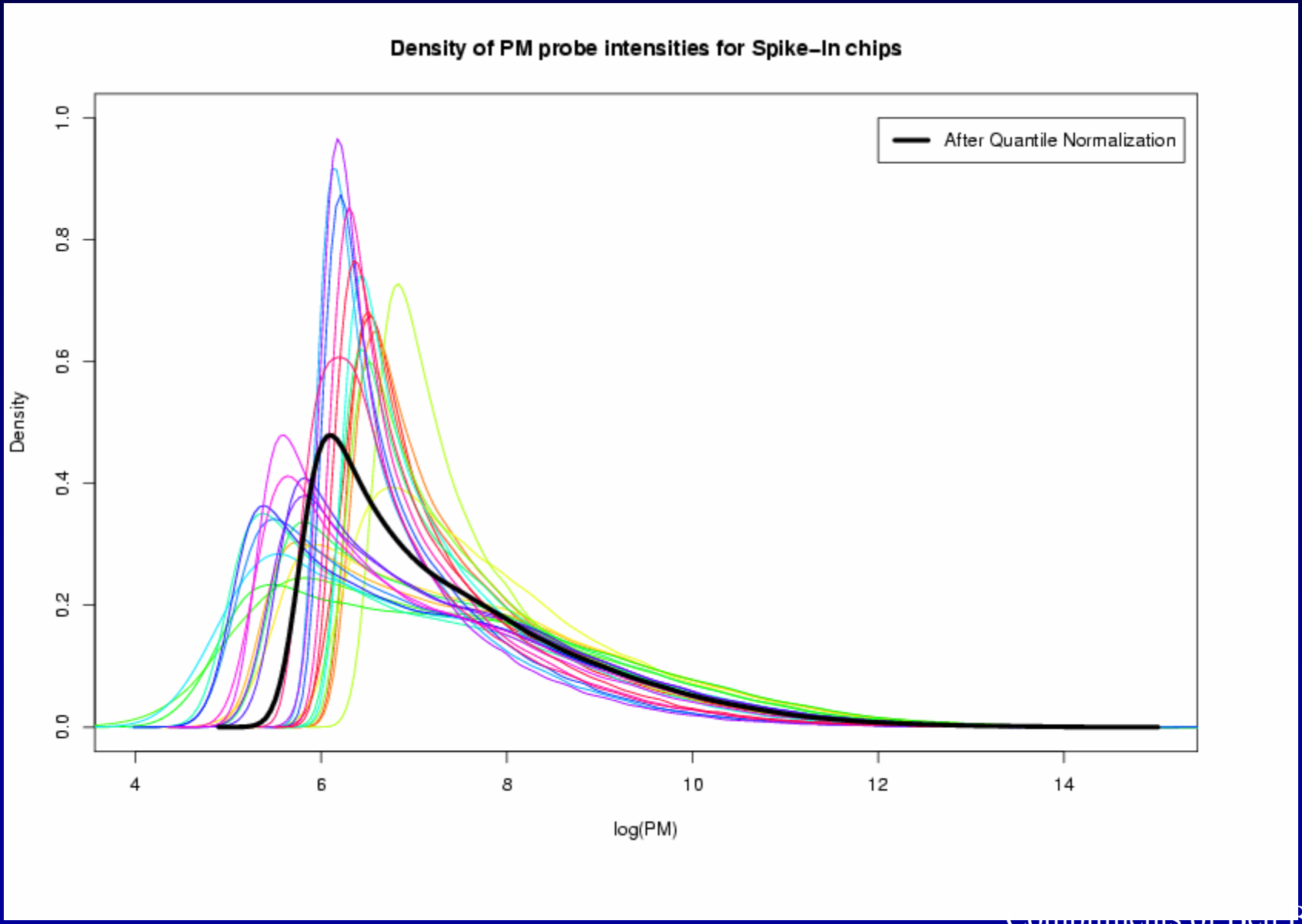
# Why background correct?



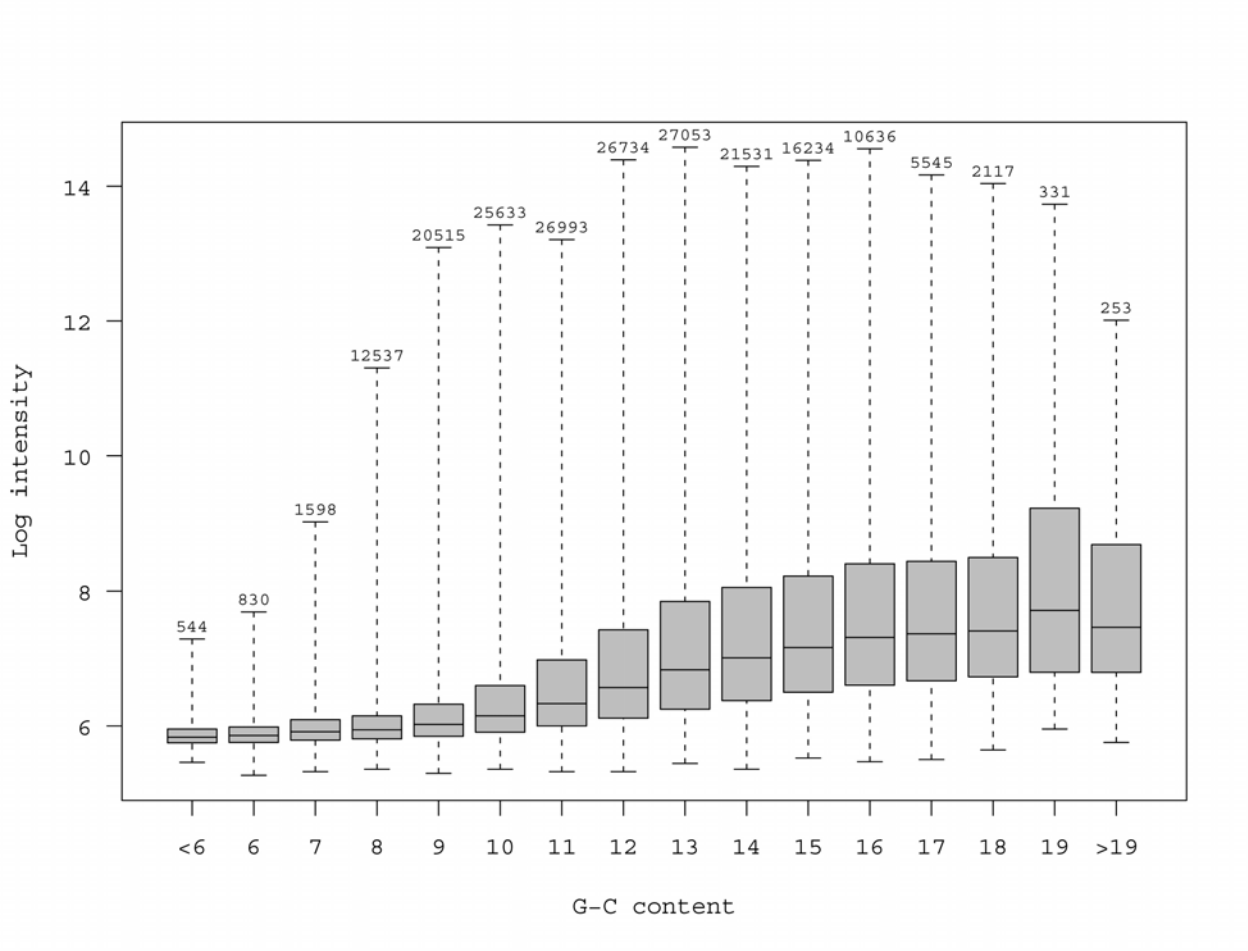
# Why background correct?



# Why normalize?



# Why correct for non-specific hyb?




One MM not enough? Look for more!

# RMA

- Robust Multiarray Analysis (RMA) is a 3-step approch:
  - ignores MM and remove global background
  - quantile normalize
  - use median polish to estimate log expression robustly
- Irizarry et al: Biostatistics (2003)
- Irizarry et al: NAR (2003)
- affy R package (www.bioconductor.org)

# Background adjustment

# Deterministic Model

PM = B + N + S

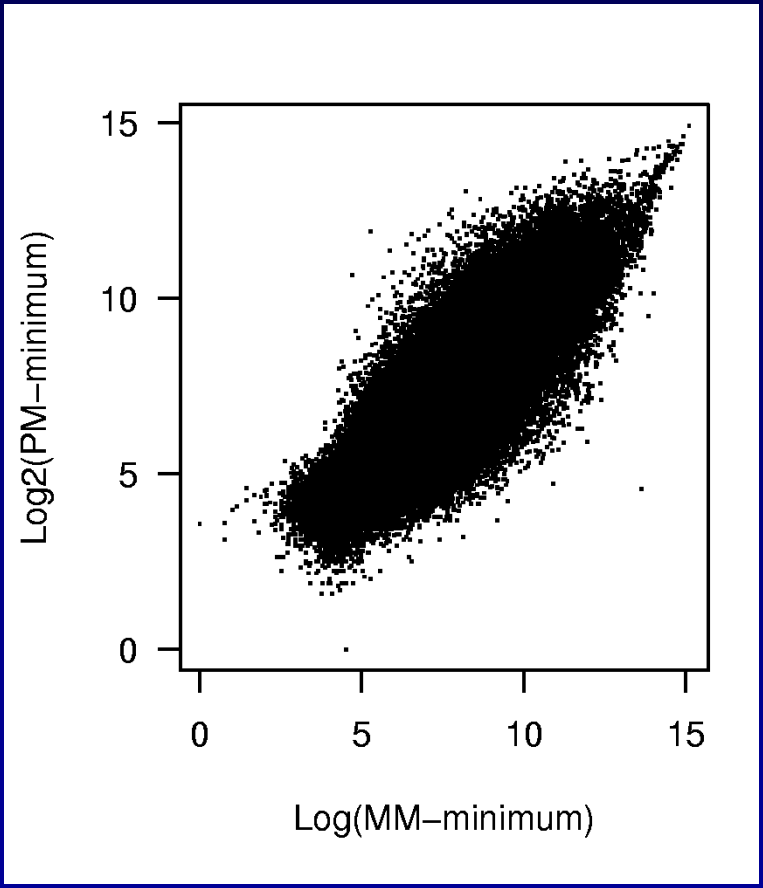
MM = B + N

PM – MM = S

# Do MMs measure non-specific binding?

Look at Yeast DNA hybridized to Human Chip(HGU95)

log (PM-B) v log (MM-B)




Not perfectly: This explains large variance

# Stochastic Model
## (Additive background/multiplicative error)

$PM = B_{PM} + N_{PM} + S,$

$MM = B_{MM} + N_{MM}$

$\log(N_{PM}), \log(N_{MM}) \sim$ Bivariate Normal ($\rho \approx 0.7$)

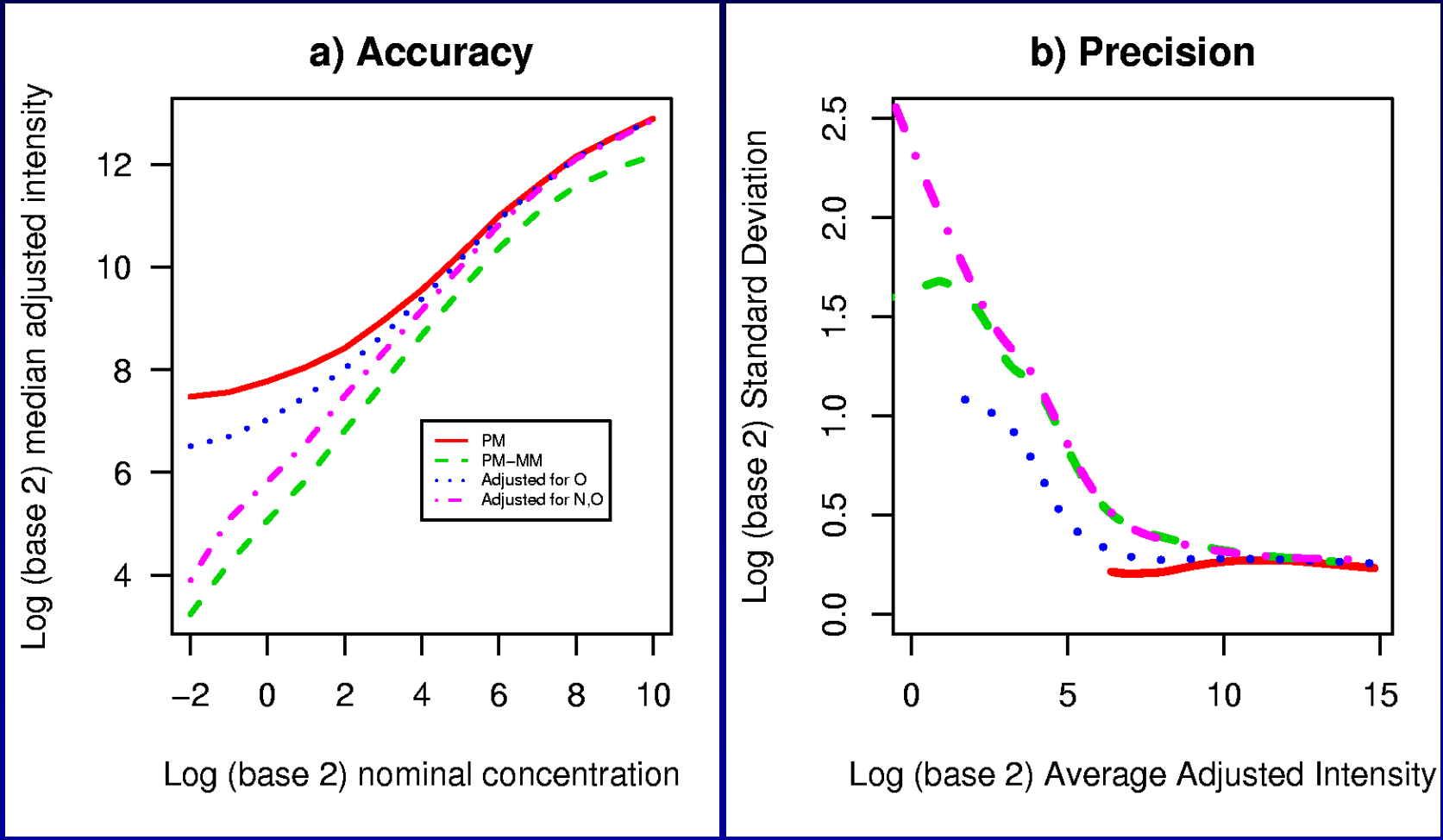
$S = \exp(\Theta + \alpha + \varepsilon)$

$\Theta$ *is the quantity of interest (log scale expression)*

$E[PM - MM] = S,$ but

$Var[\log(PM - MM)] \sim 1/\exp(\Theta)$ (can be very large)

# Can we just ignore background?




PM is a biased estimate of $\Theta$

# Alternative Approach

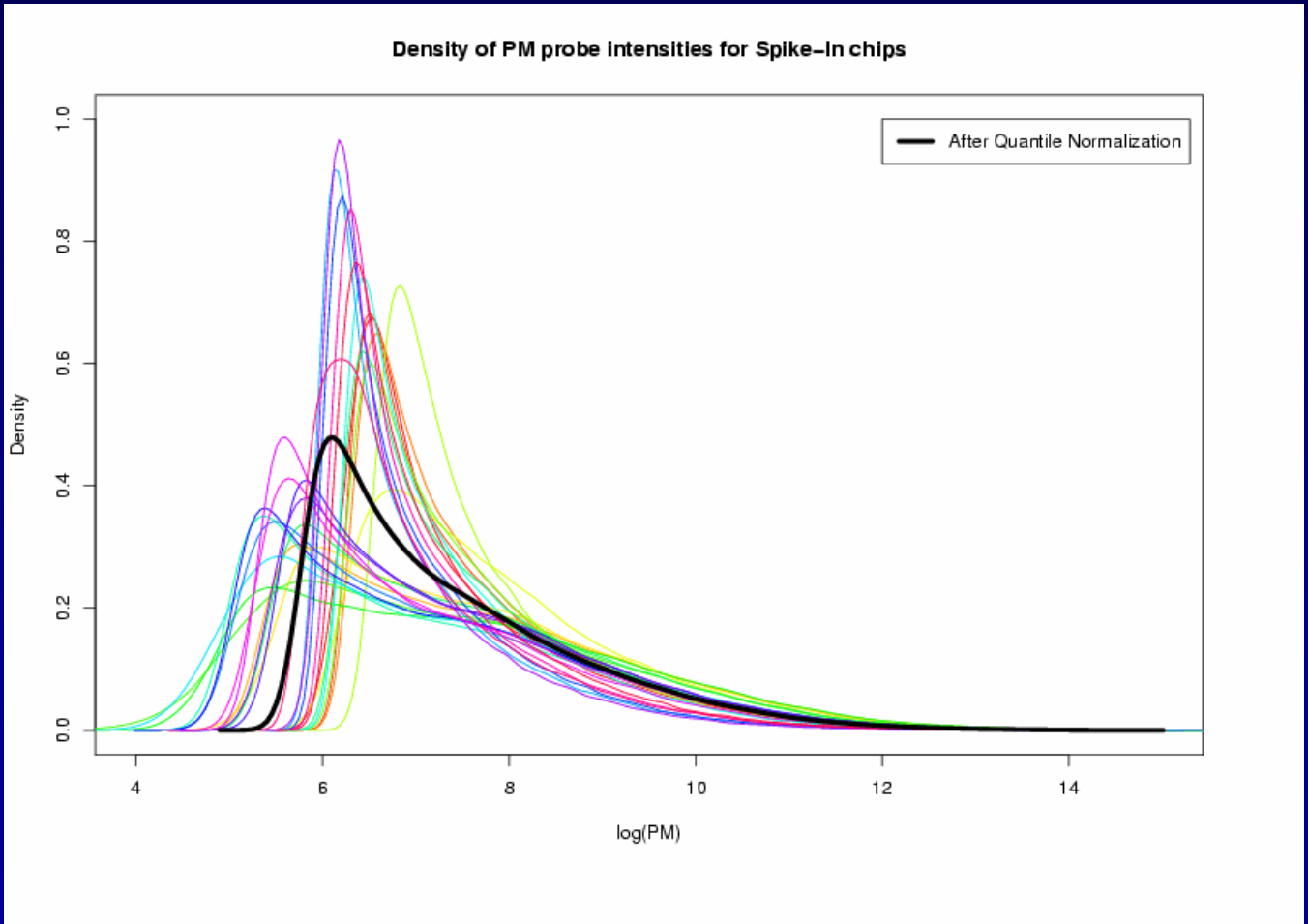
Predict log(*S*) from *PM,MM*

For example:

1) *E[ log(S)* | *PM, MM ]*

2) Estimate Θ and obtain standard error: Formal hypothesis testing

# Quantile normalization



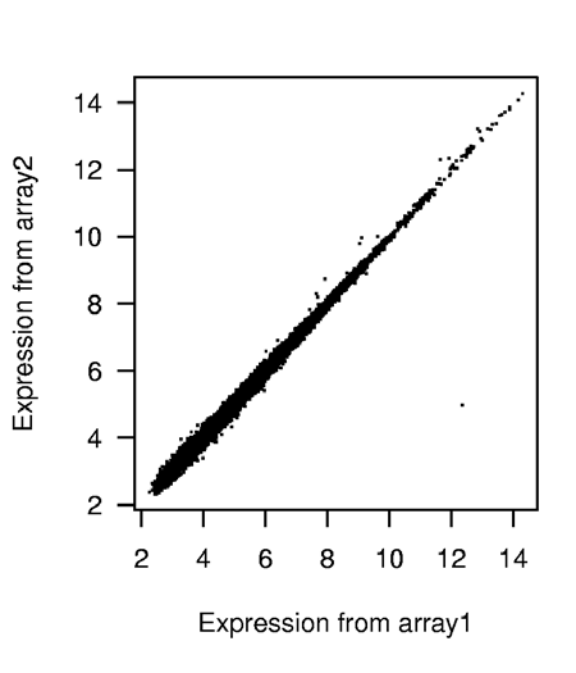
# Summarization

- Do it in the log-scale
- Account for the probe effect
- Use robust procedure

# Probe-effect

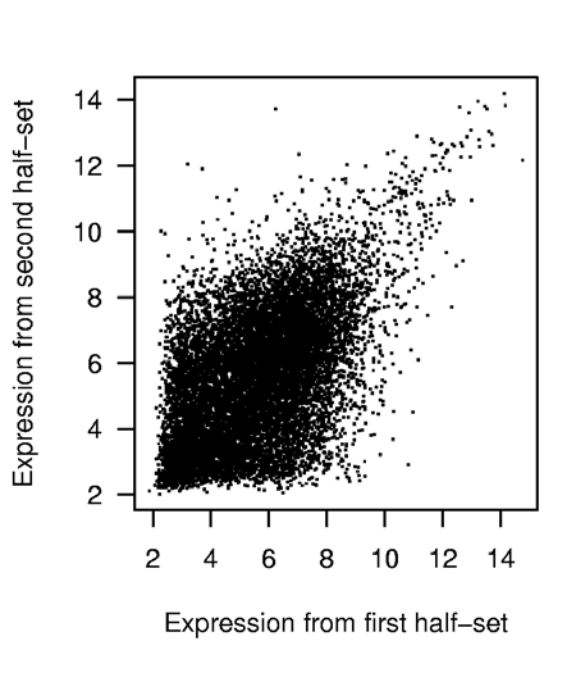
- Li and Wong (2001) first observed the very strong probe effect
- Within the same probeset, a large range of intensities (orders of magnitude) is observed. But across arrays, variance of intensities, for the same probe, is relatively small
- This probe effect explains high correlation between replicate arrays

# Expression from 2 replicate arrays



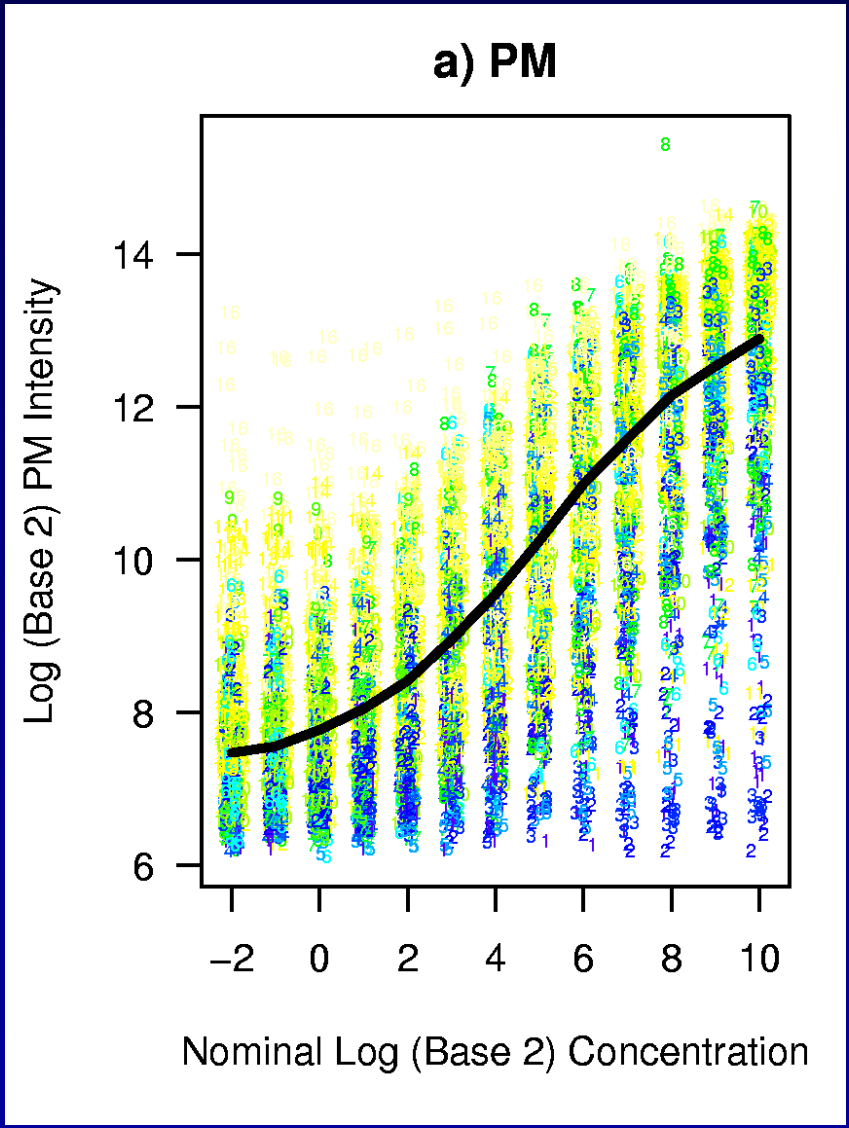

Correlation is higher than 0.99

# Expression from probesets divided into 2 (at random)

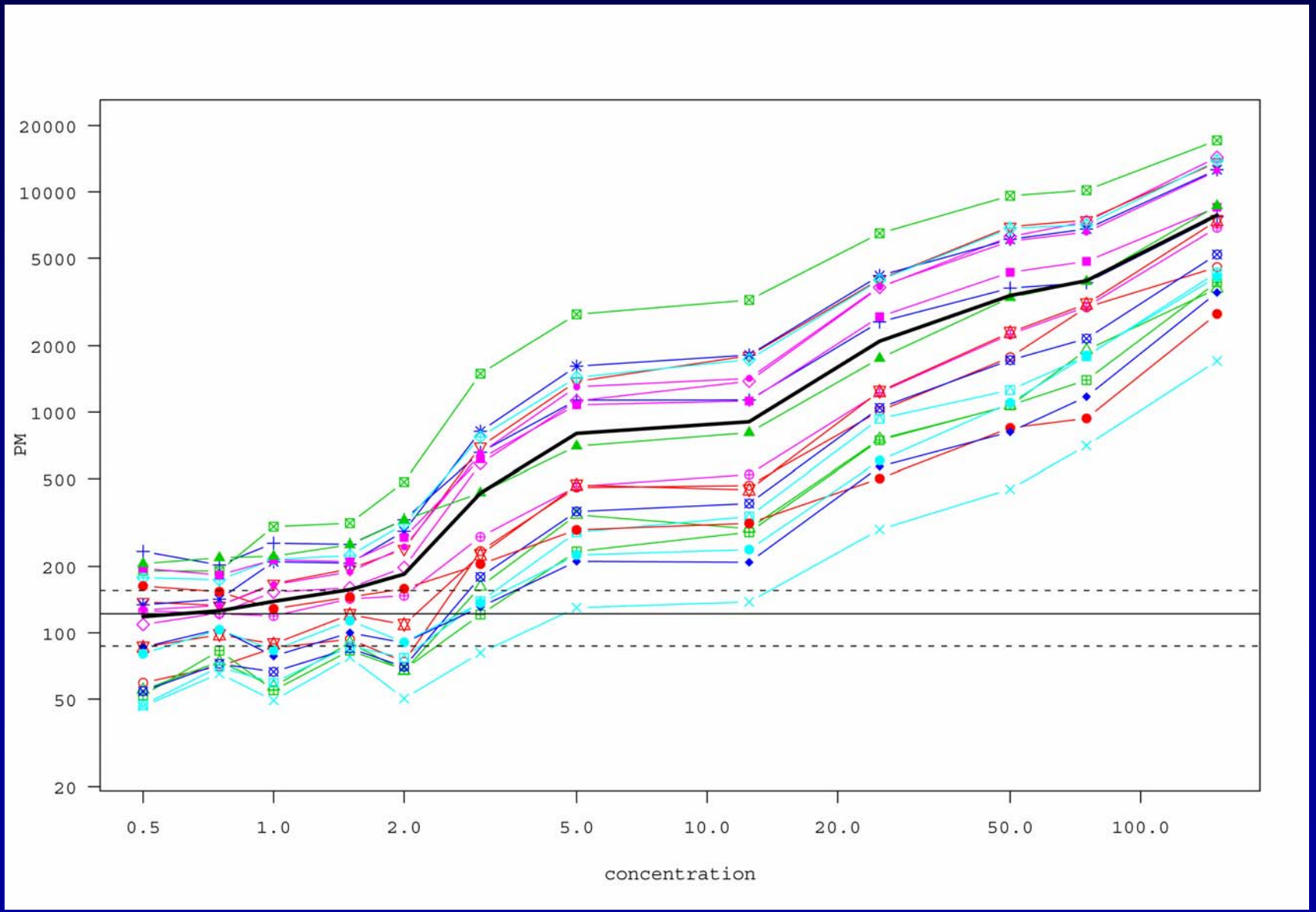



Correlation drops to 0.55

# Probe effect seen in spike-ins



# Why fit log scale additive model?



# RMA

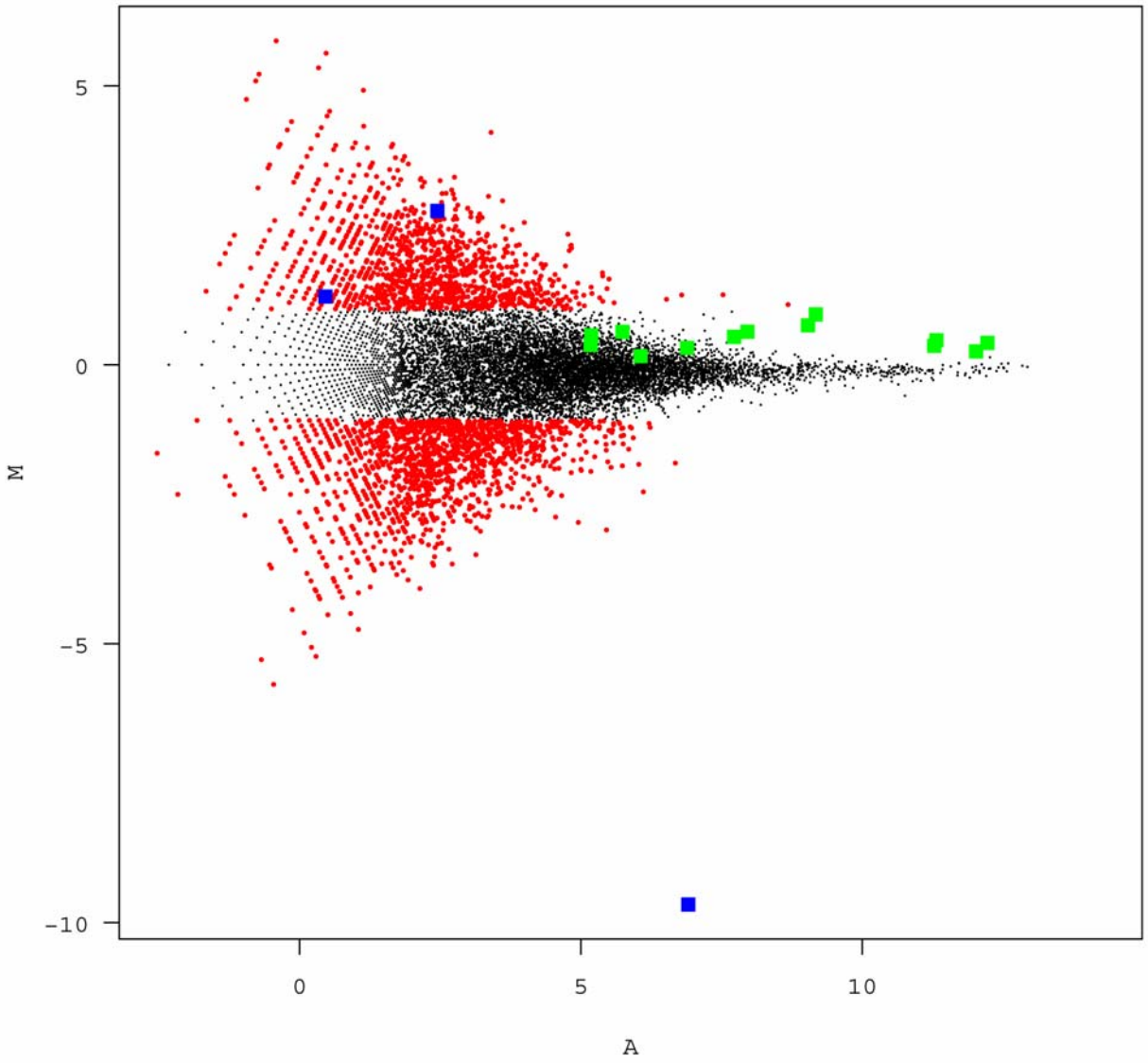
- Instead of subtracting MM,

  Assume $PM = B + S$
- To estimate $S$, use expectation: $E[S|B+S]$, with B normal and S exponential
- After quantile normalization, assume:

$$log_2 S_{ij} = \Theta_i + \alpha_j + \varepsilon_{ij}$$

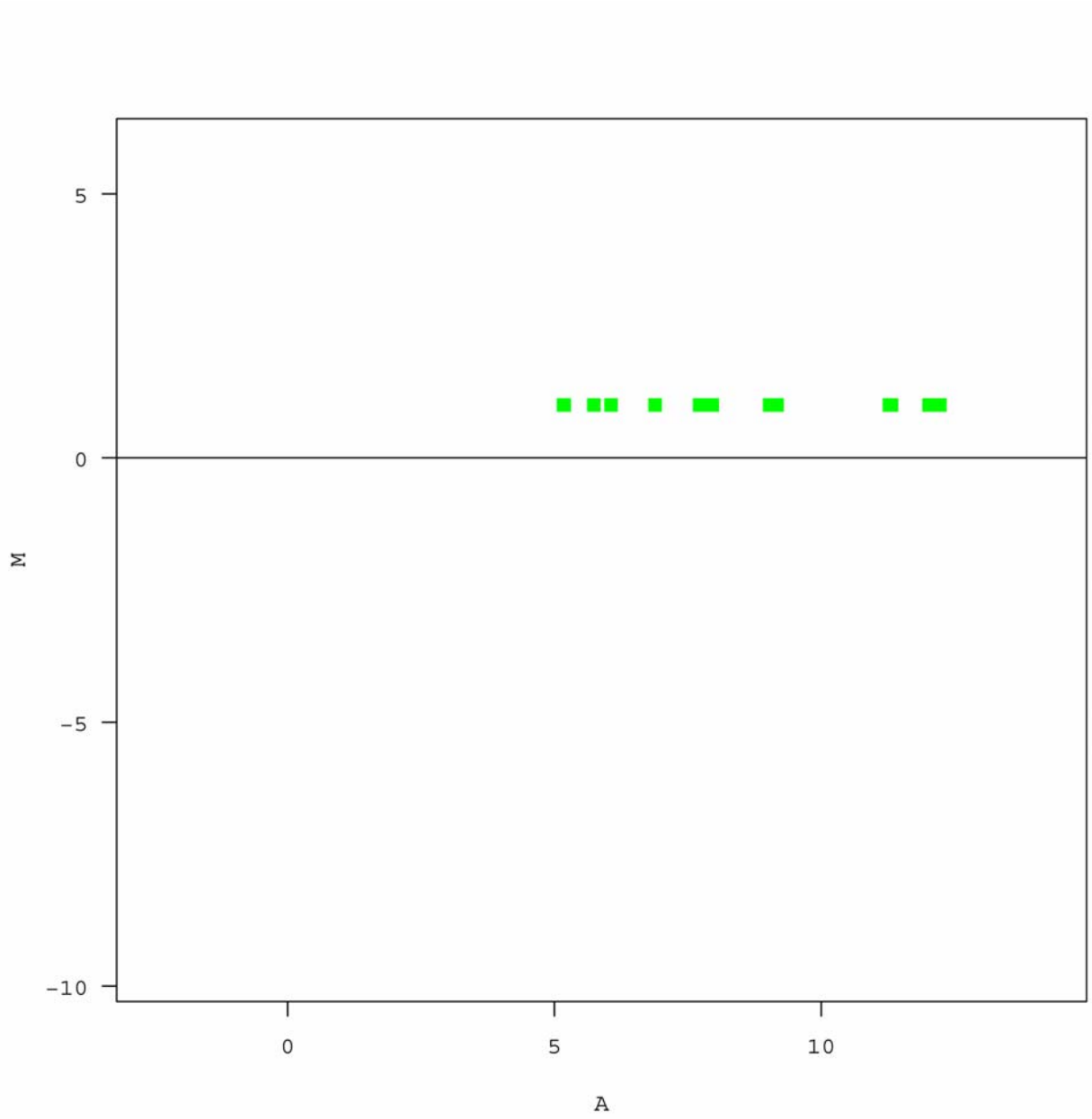
- Estimate $\Theta_i$ using robust procedure (median polish)
- We call this procedure RMA
- Does it make a difference?

# Does it make a difference?



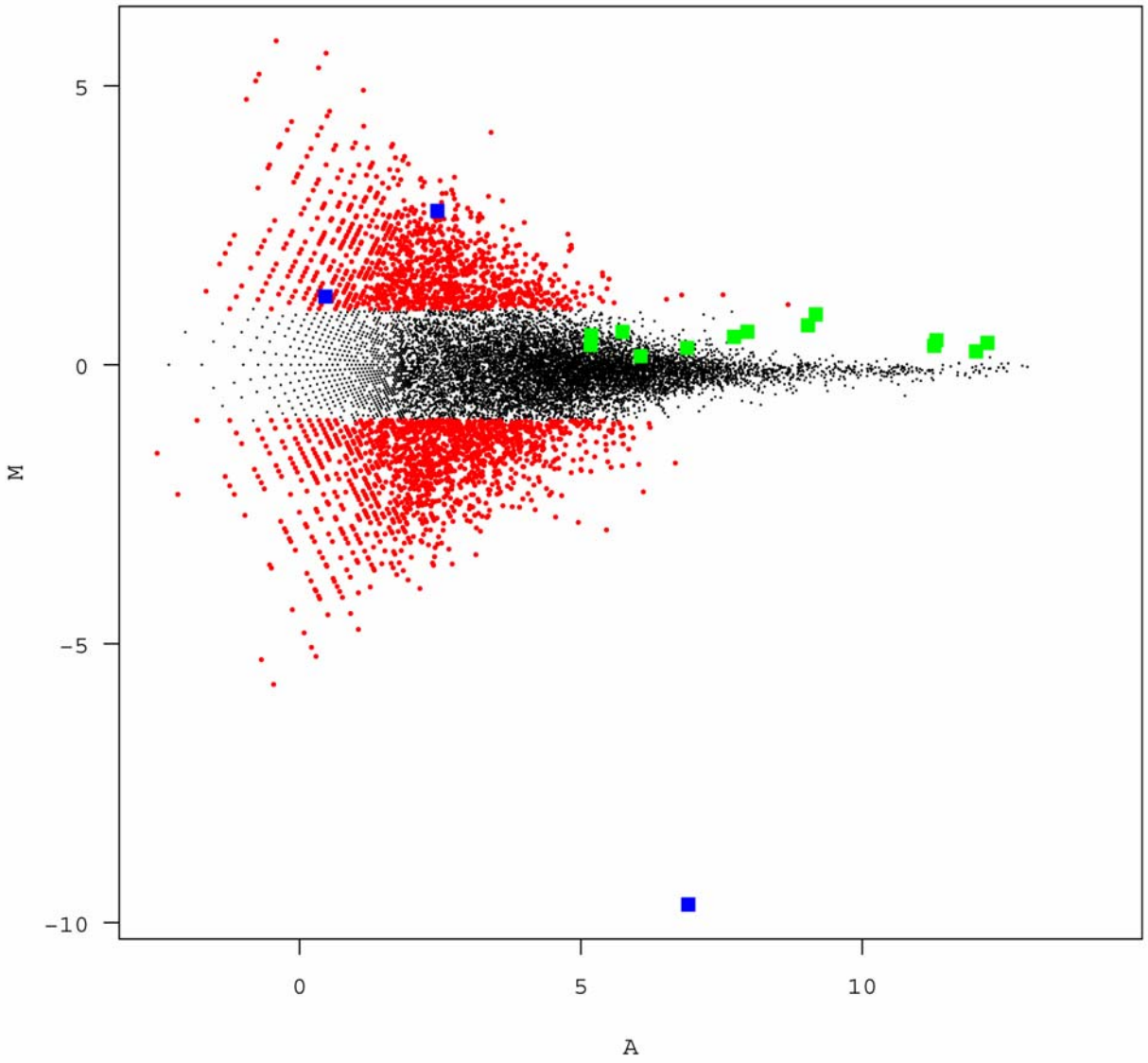

1
270
2074
3063
3935
4639
4652
5149
5372
5947
6448
6870
7037
7549
8429
9721

# Perfect



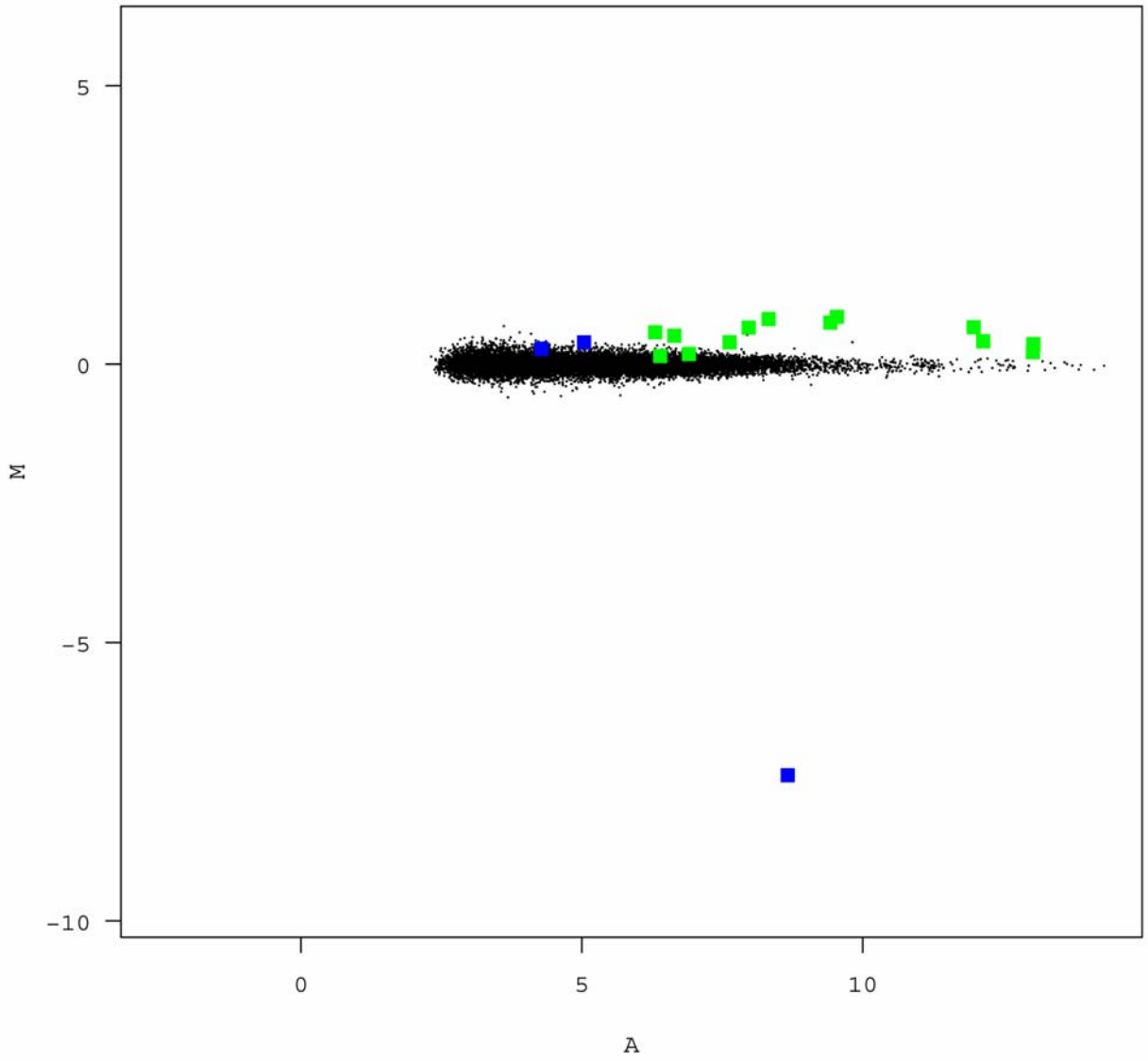

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

# MAS 5.0




1
270
2074
3063
3935
4639
4652
5149
5372
5947
6448
6870
7037
7549
8429
9721

# RMA



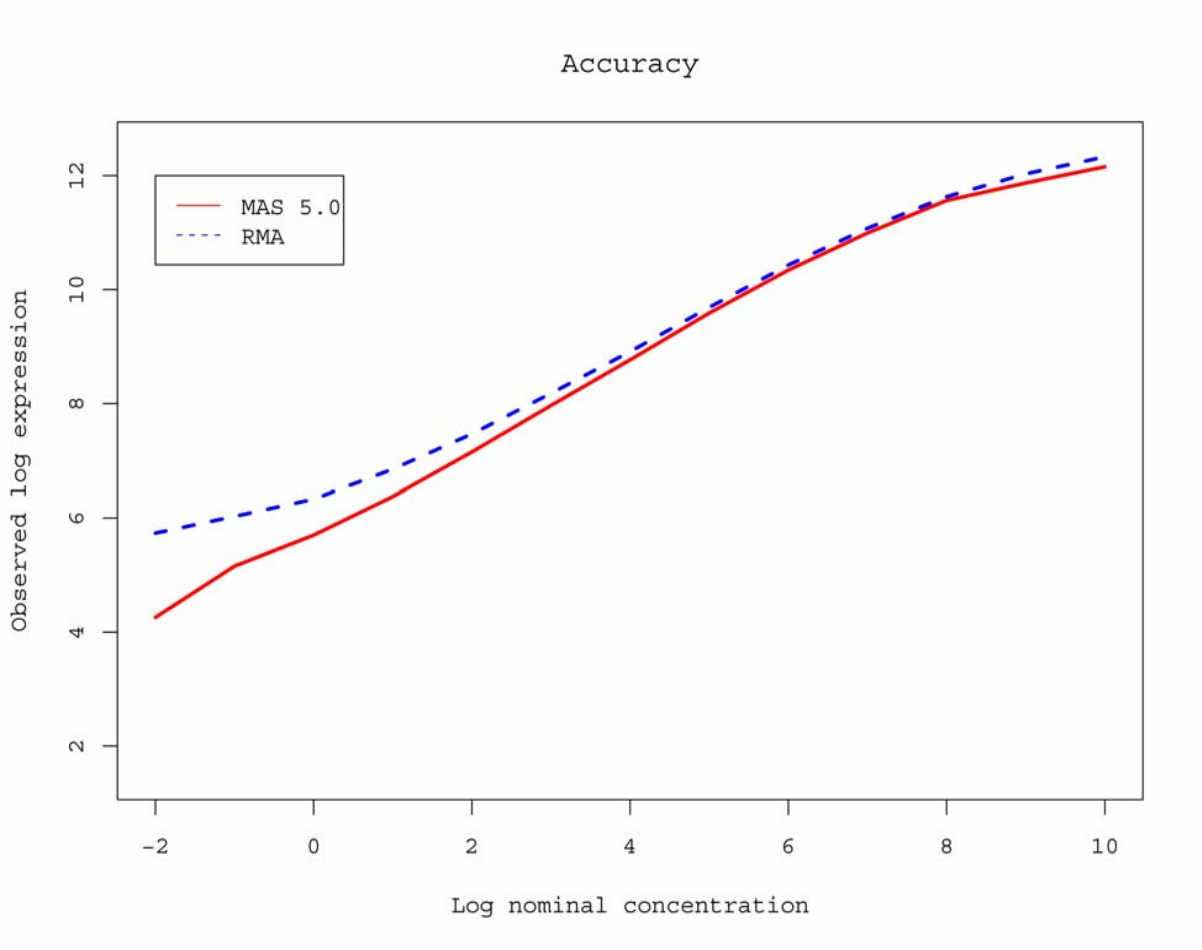

1
2
3
4
6
7
10
16
45
56
58
88
406
999
1643
2739

# Can RMA be improved?

Global Accuracy and Precision

| | Slope | Median SD | Percentile | Rank |
|---|---|---|---|---|
| MAS 5.0 | 0.69 | 0.63 | 82.43 | 2188 |
| RMA | 0.61 | 0.11 | 99.96 | 4 |

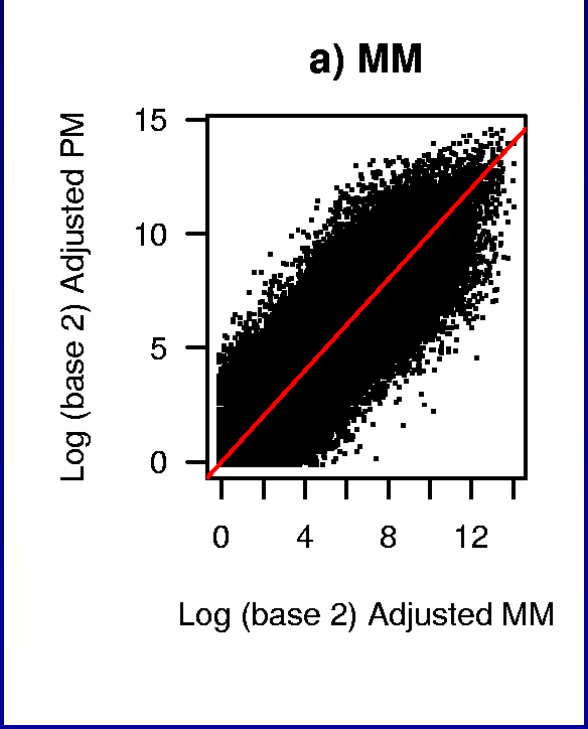
# Can RMA be improved?

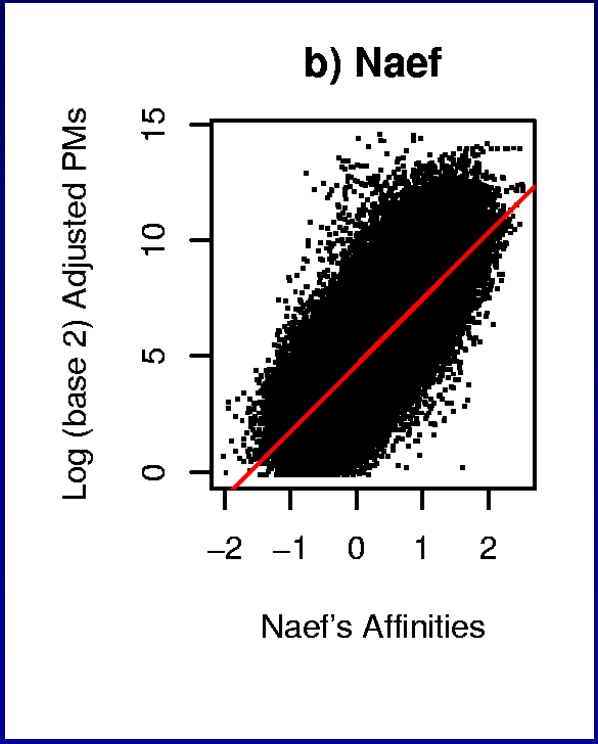


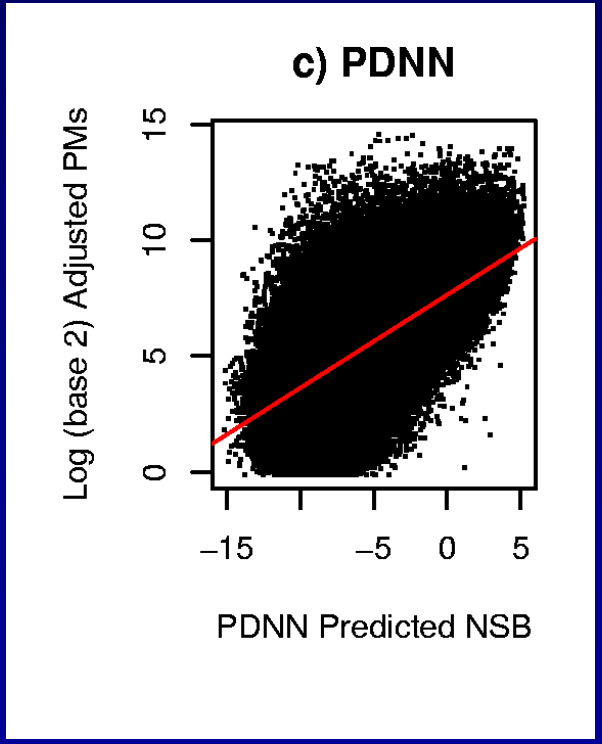
# Current Work

- Incorporate MM and sequence information to build an improved model and estimate
- Find alternative, faster, approaches to posterior mean
- Preliminary work: GCRMA

# Predict NSB with sequence info



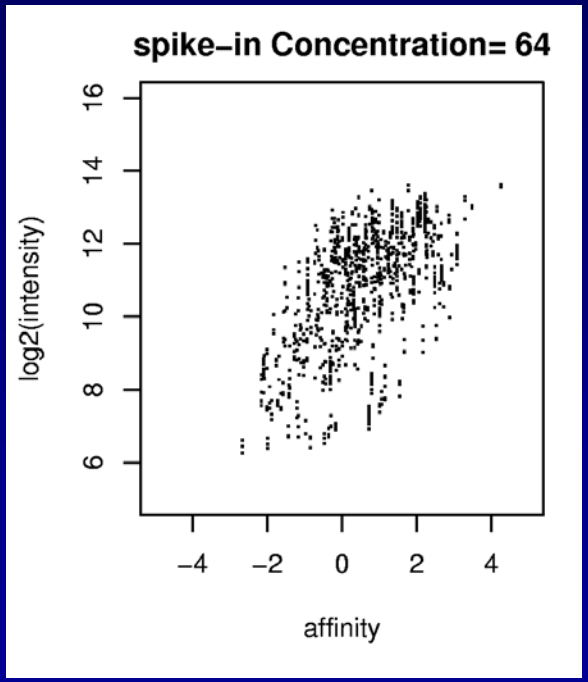



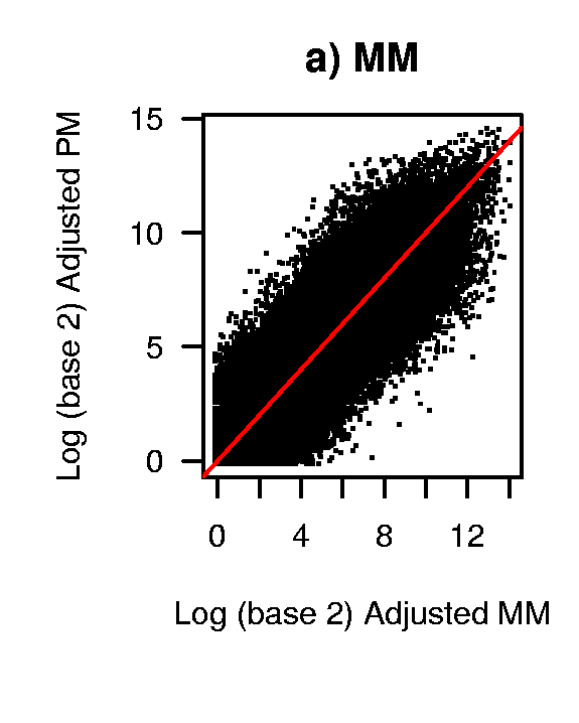
# Naef's model

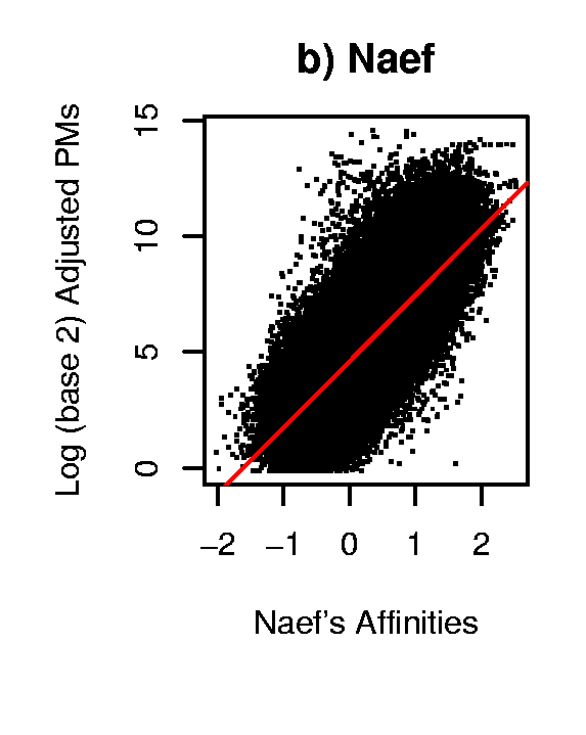
- Assume that being an A,T,G or C has a position dependent effect on probe effect
- Assume that this effect is a smooth function of position (Naef uses cubic polynomials we use splines)
- Use training data to get *affinities*

# Naef uses these to predict probe effect

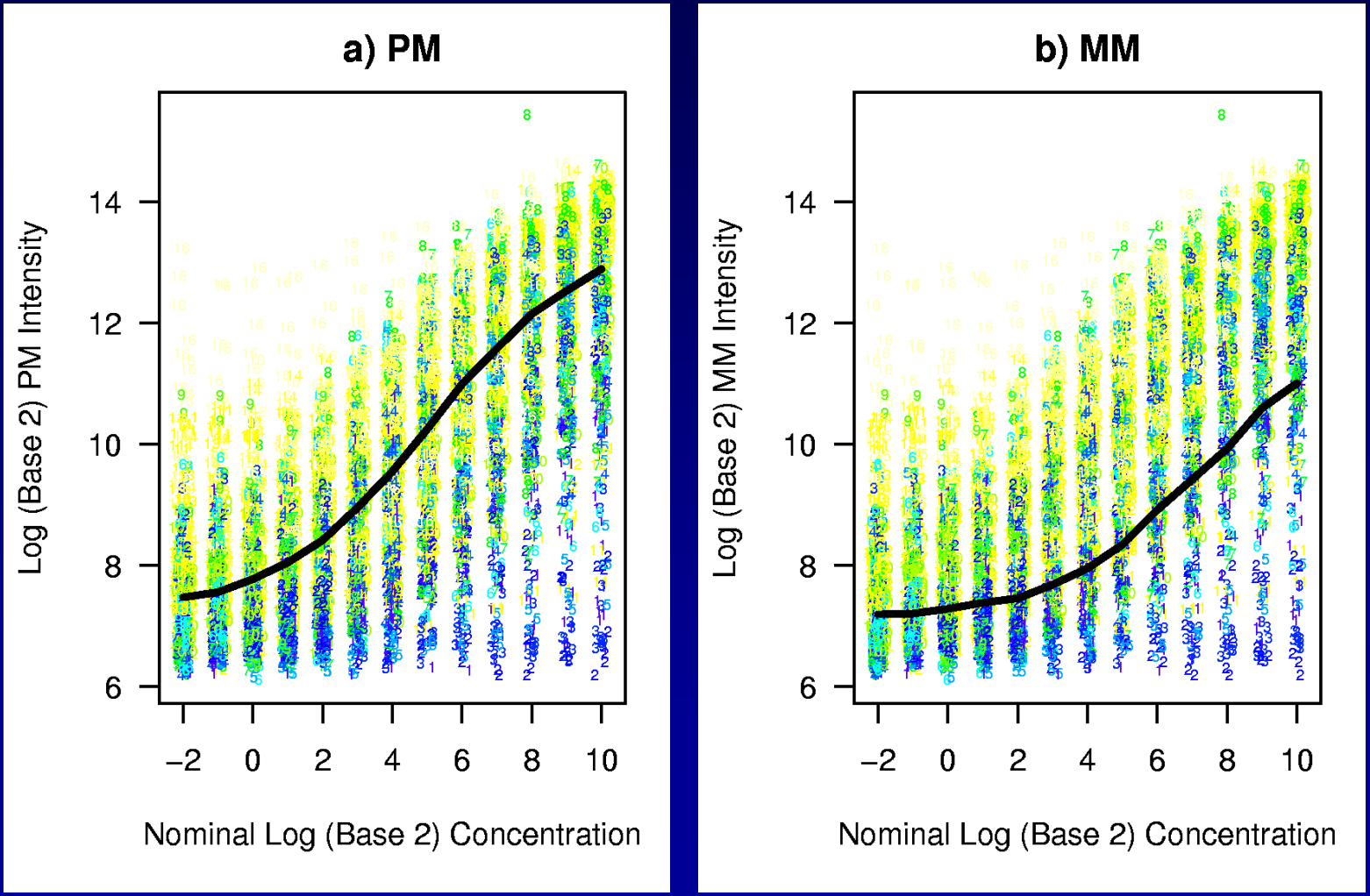


# We use them to predict NSB too



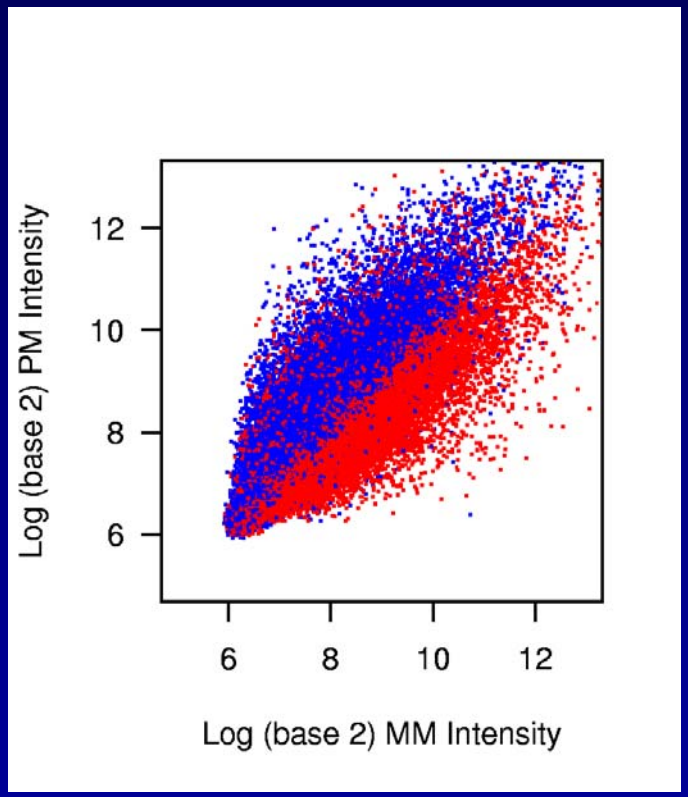



# Problems with MM

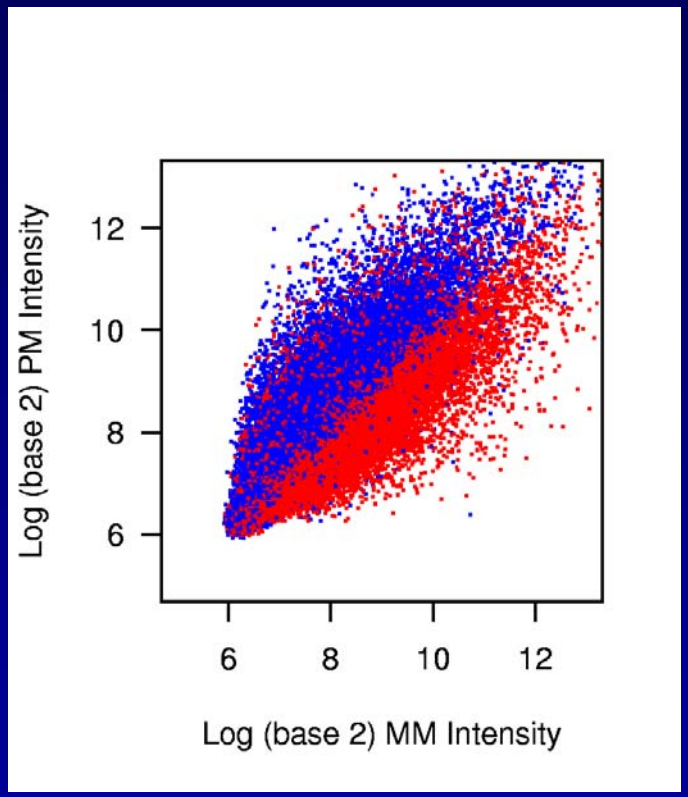


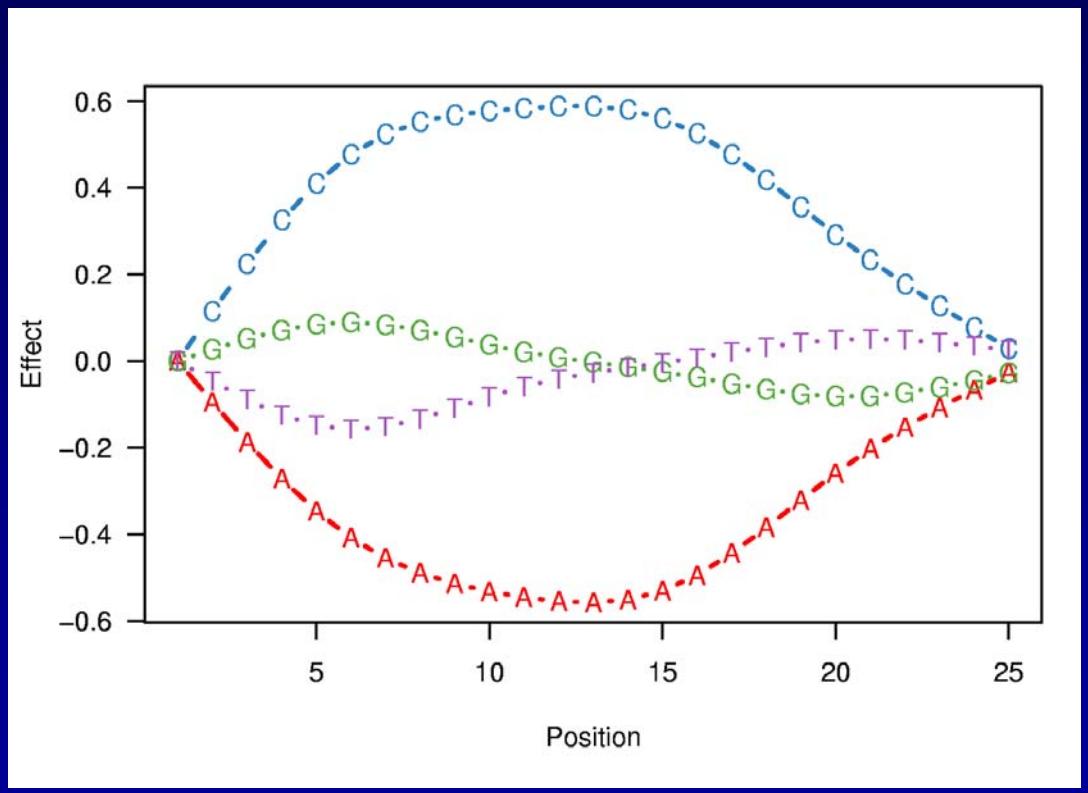

Also they take up half the space on the chip ($250)

# More problems with MM

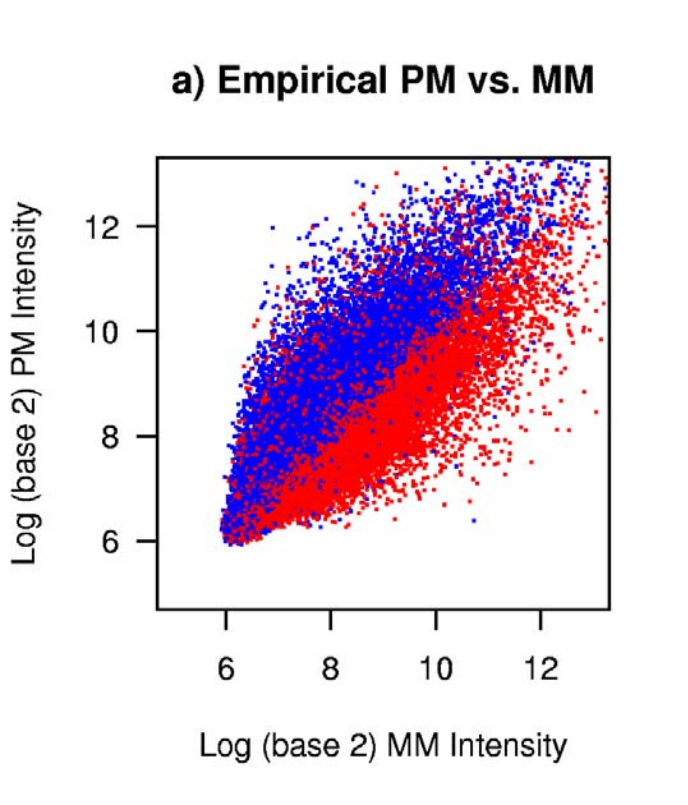


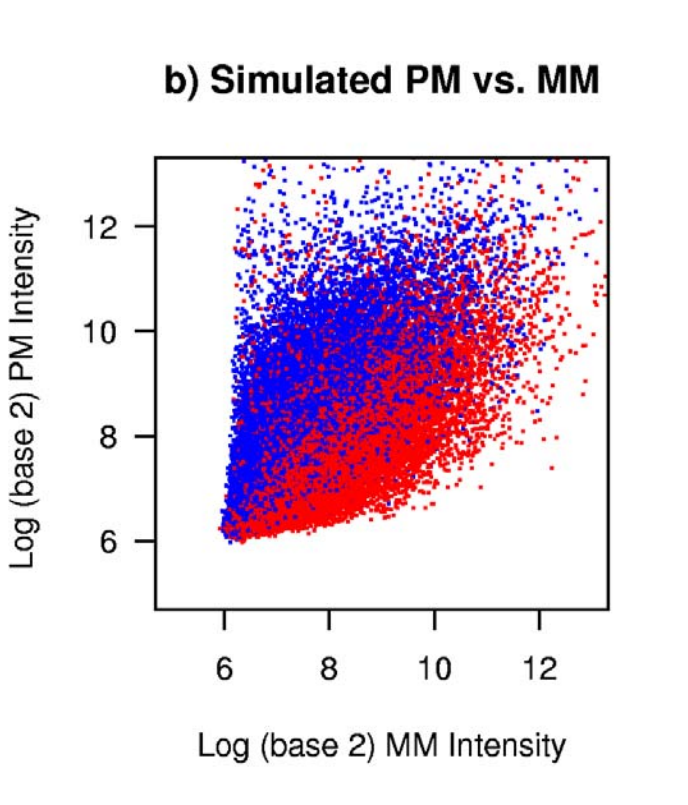
# More problems with MM



# Our model predicts this



# Adjustment options

- Define a loss function, assume S is random variable, find empirical Bayes esimtate, e.g. for log ratio based loss the solution is:

$$E[\ \log(S) \mid PM, MM\ ]$$

- GCRMA assumes S follows power-law or log(S) is uniform

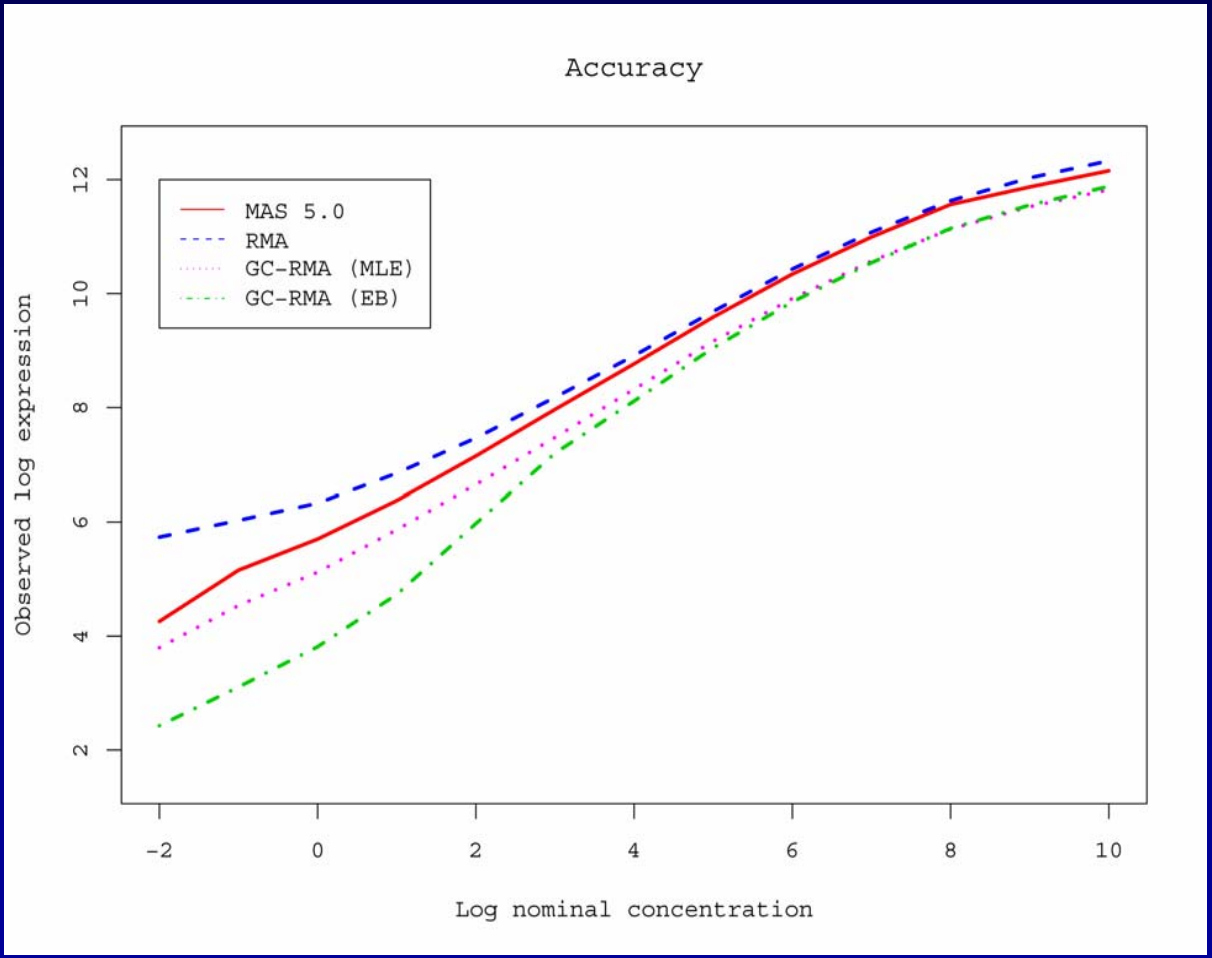
# Does it help?

Global Accuracy and Precision

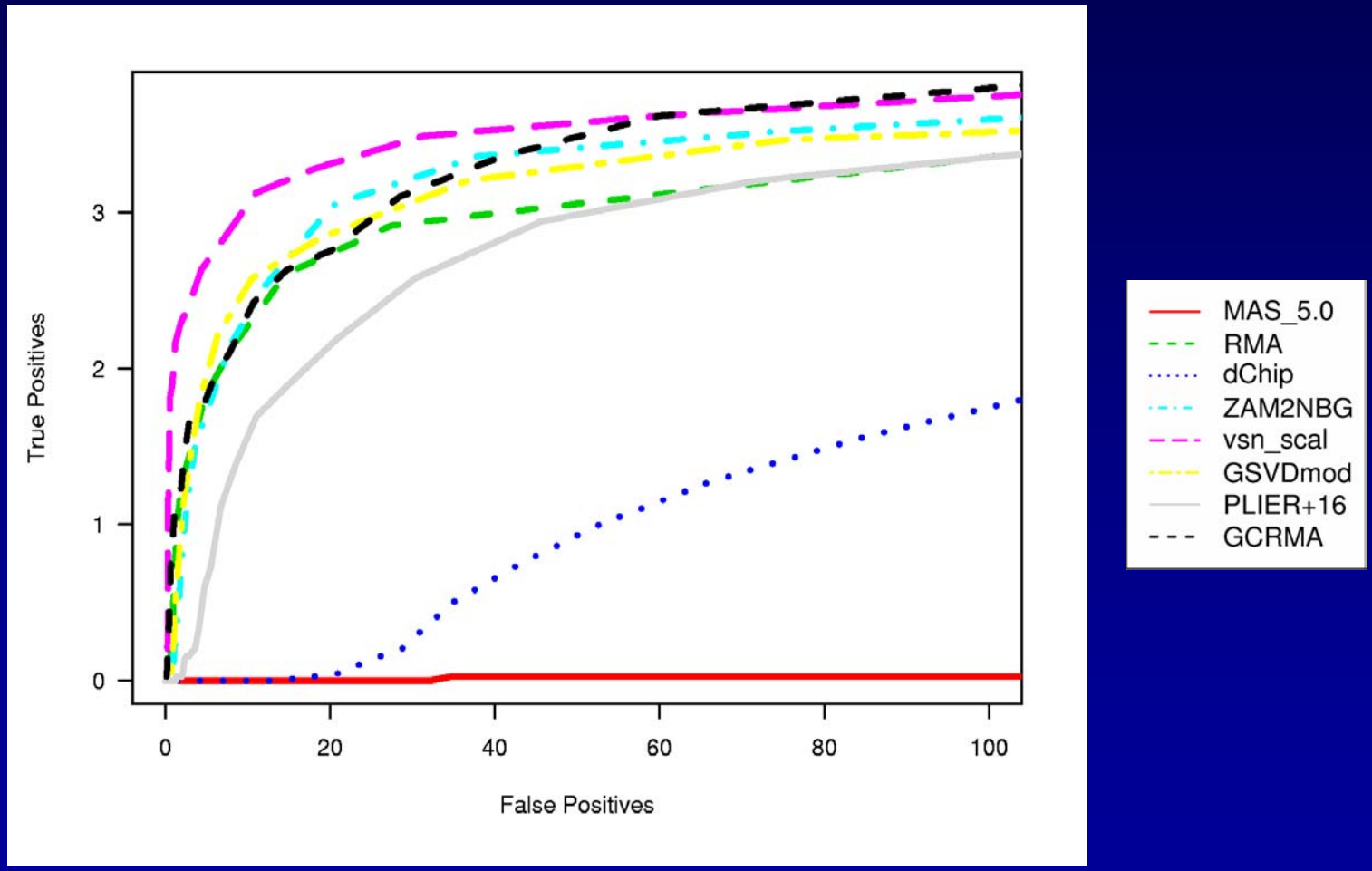
| | Slope | Median SD | Percentile | Rank |
|---|---|---|---|---|
| MAS 5.0 | 0.69 | 0.63 | 82.43 | 2188 |
| RMA | 0.61 | 0.11 | 99.96 | 4 |
| GCRMA | 0.85 | 0.08 | 99.98 | 2 |

Local slopes also improve

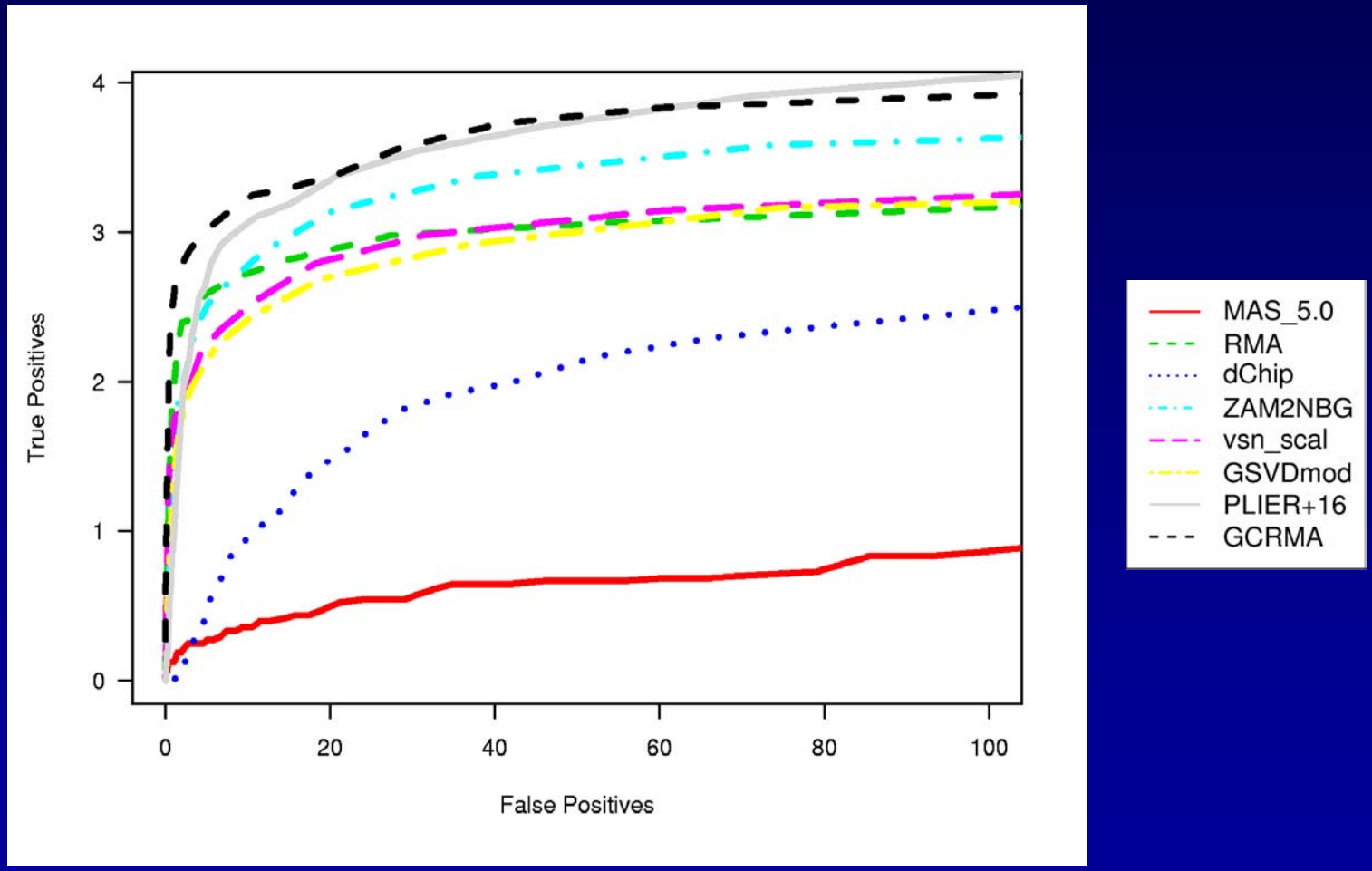
# Does it help?


Accuracy
MAS 5.0
RMA
GC-RMA (MLE)
GC-RMA (EB)
Observed log expression
2
4
6
8
10
12
-2
0
2
4
6
8
10
Log nominal concentration

# ROC for FC=2 spikes



# ROC for low concentration spikes



# Local Ranks

| | -2:-1 | -1:0 | 0:1 | 1:2 | 2:3 | 3:4 | 4:5 | 5:6 | 6:7 | 7:8 | 8:9 | 9:10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MAS_5.0 | 1715 | 2736 | 2282 | 1998 | 1898 | 1935 | 1887 | 2051 | 2352 | 2633 | 3976 | 4128 |
| RMA | 380 | 360 | 27 | 5 | 3 | 3 | 3 | 3 | 5 | 10 | 76 | 408 |
| GCRMA | 15 | 8 | 4 | 2 | 3 | 3 | 3 | 4 | 9 | 18 | 86 | 161 |
| % of data | 25 | 16 | 18 | 17 | 11 | 6 | 2 | 1 | 1 | 0 | 0 | 0 |

# Conclusion

- Data exploration useful tool for quality assessment and motivating models
- Statistical thinking helpful for interpretation
- Statistical models may help find signals in noise
- Physical models help improve accuracy

# Supplemental Slide

# Local Ranks

| | -2:-1 | -1:0 | 0:1 | 1:2 | 2:3 | 3:4 | 4:5 | 5:6 | 6:7 | 7:8 | 8:9 | 9:10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MAS_5.0 | 1715 | 2736 | 2282 | 1998 | 1898 | 1935 | 1887 | 2051 | 2352 | 2633 | 3976 | 4128 |
| dChip | 961 | 907 | 200 | 113 | 61 | 52 | 52 | 75 | 102 | 192 | 477 | 962 |
| RMA | 380 | 360 | 27 | 5 | 3 | 3 | 3 | 3 | 5 | 10 | 76 | 408 |
| GCRMA | 15 | 8 | 4 | 2 | 3 | 3 | 3 | 4 | 9 | 18 | 86 | 161 |
| % of data | 25 | 16 | 18 | 17 | 11 | 6 | 2 | 1 | 1 | 0 | 0 | 0 |