

# **Exploration, Normalization, Summaries, and Software for Affymetrix Probe Level Data**

Rafael A. Irizarry

*Department of Biostatistics, JHU*

March 12, 2003

# Outline

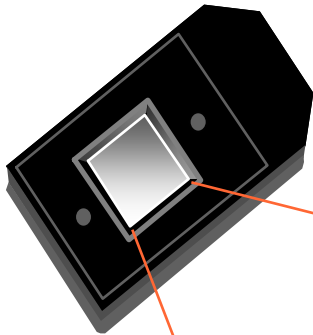
- Review of technology
- Why study probe level data?
- Probe level data → expression measures
- Case study

<http://www.biostat.jhsph.edu/~ririzarr>

rafa@jhu.edu

# Affymetrix GeneChip Arrays

GeneChip Probe Array



1.28cm

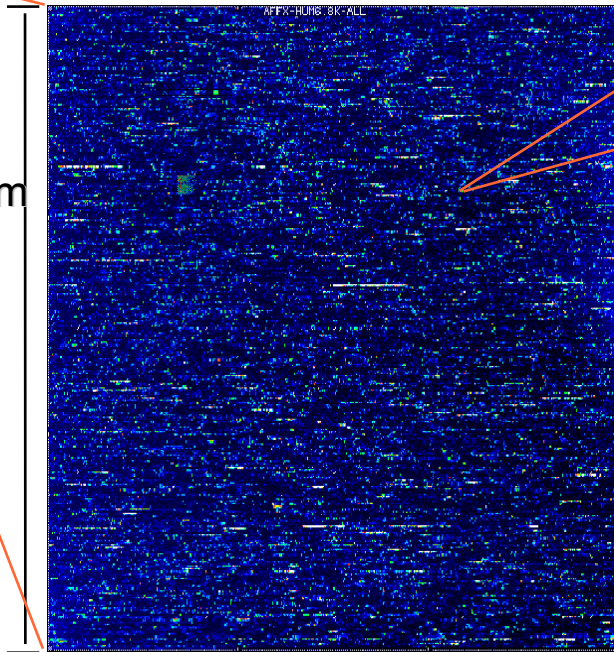
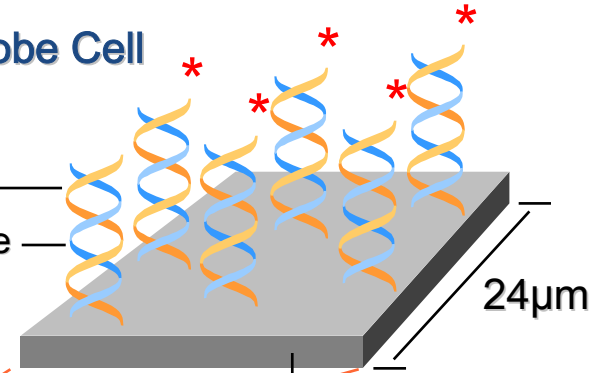


Image of Hybridized Probe Array

Hybridized Probe Cell

Single stranded, labeled RNA target  
Oligonucleotide probe



Millions of copies of a specific oligonucleotide probe

>200,000 different complementary probes

# GeneChip® Expression Array Design

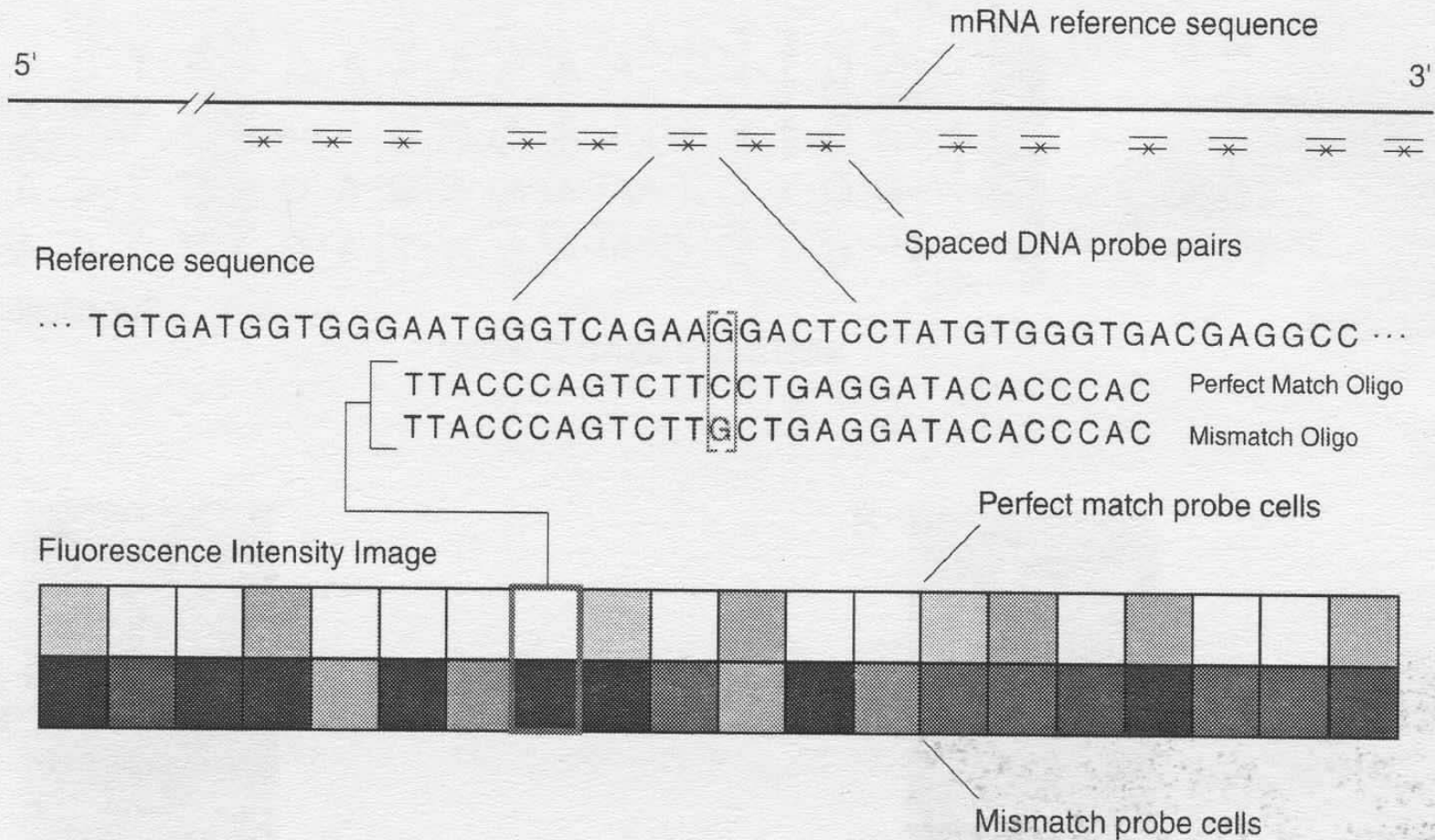


Figure 1-3 Expression tiling strategy

# Why Keep Probe Level Data?

- Quality control
  - Spatial Effects
  - RNA degradation (Leslie Cope)
- Detection of defective probes
- Transcript sequence “estimates” change
- Ways to reduce to expression measure  
keep improving

# Software: Bioconductor R package

## **affy**: AffyBatch class

**exprs**

Matrix of cel intensities, probes x samples (cel files)

**se.exprs**

Matrix of SDs for probe intensities

**phenoData**

Sample level covariates, instance of class **phenoData**

**cdfName**

Hash table R environment with location (CDF file) info

**annotation**

Name of annotation data

**description**

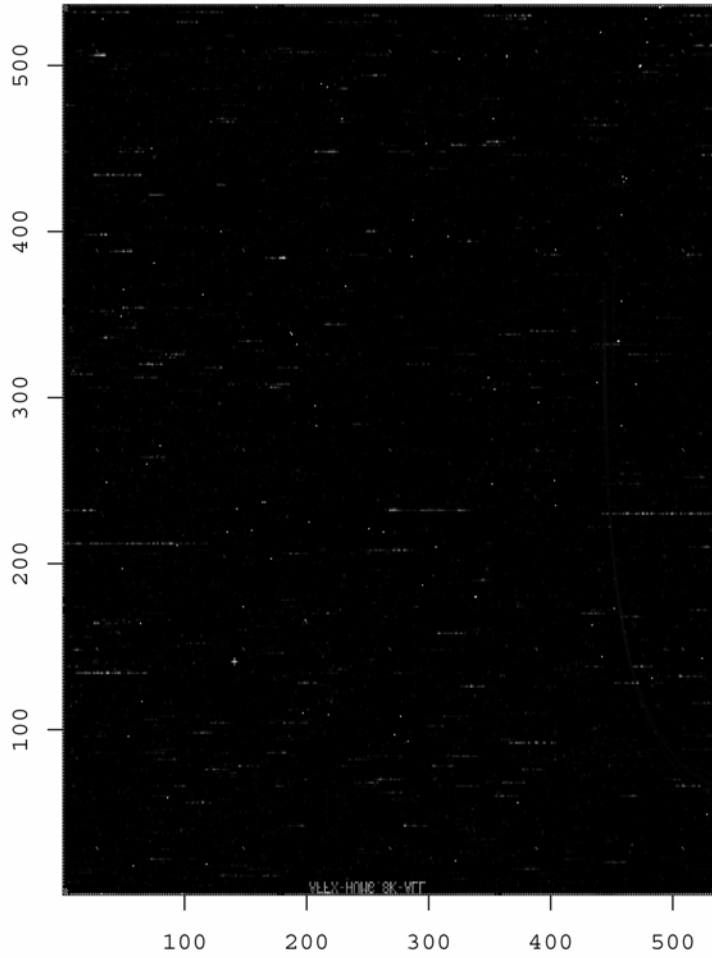
Object of class MIAME

**nrow, ncol, notes**

# rows and columns on array and any further notes

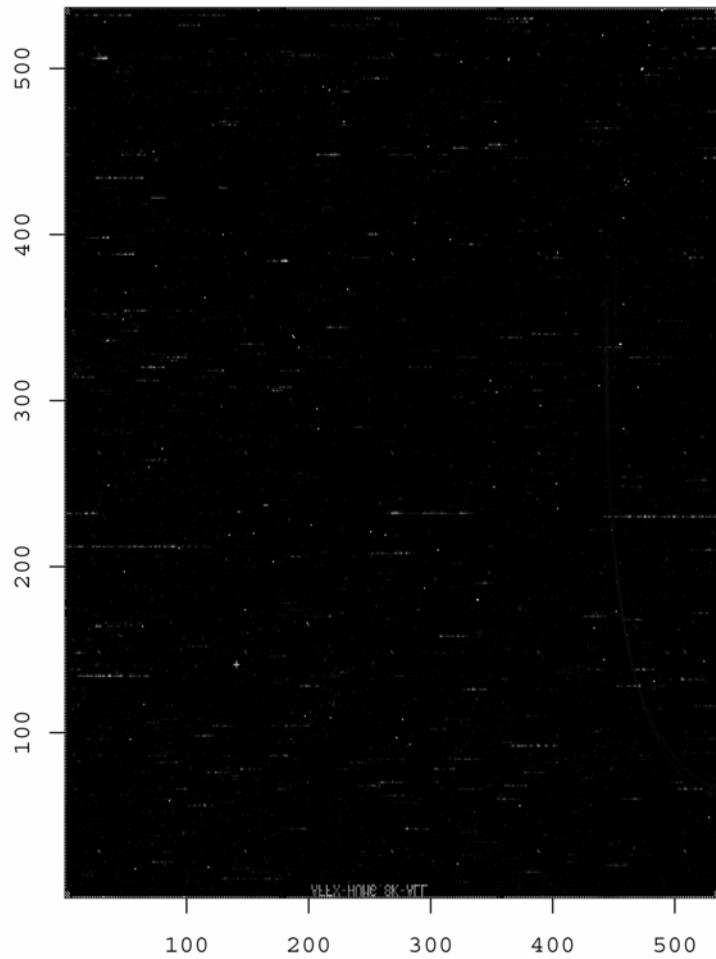
# QC

raw values

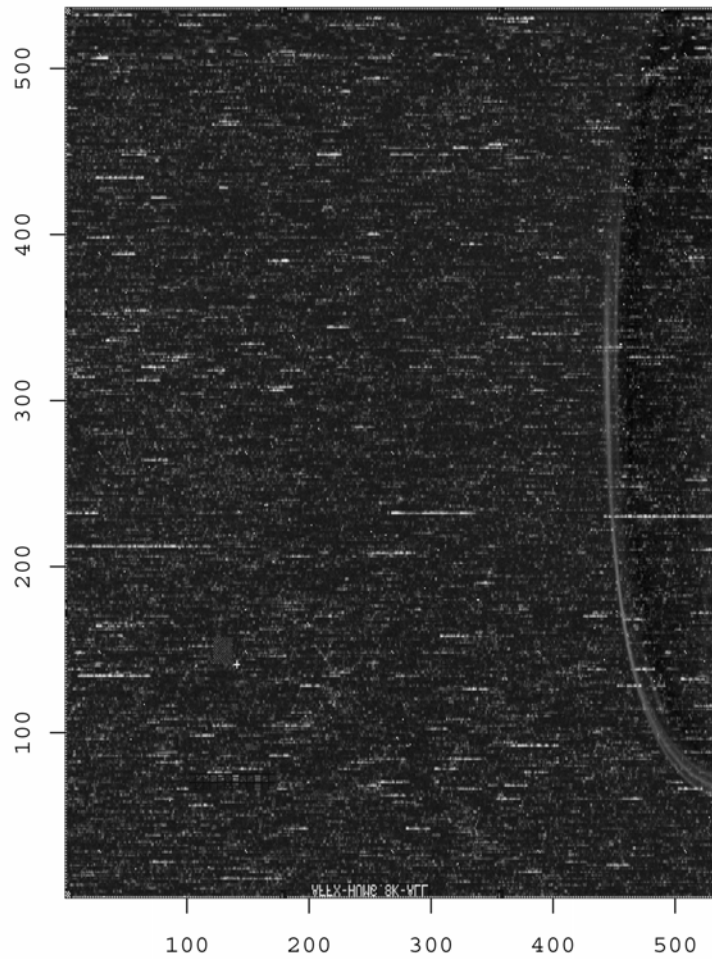


# QC

raw values



log2-transformed values



# Case Study

Probe level data → expression measures

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

# Default until 2002 (MAS 4.0)

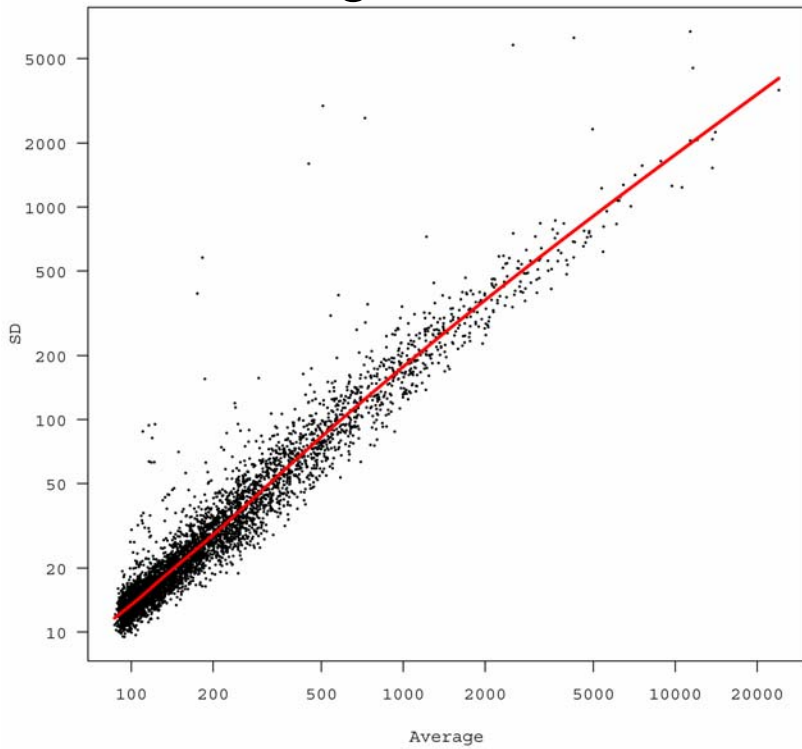
- GeneChip<sup>®</sup> software used *Avg.diff*

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

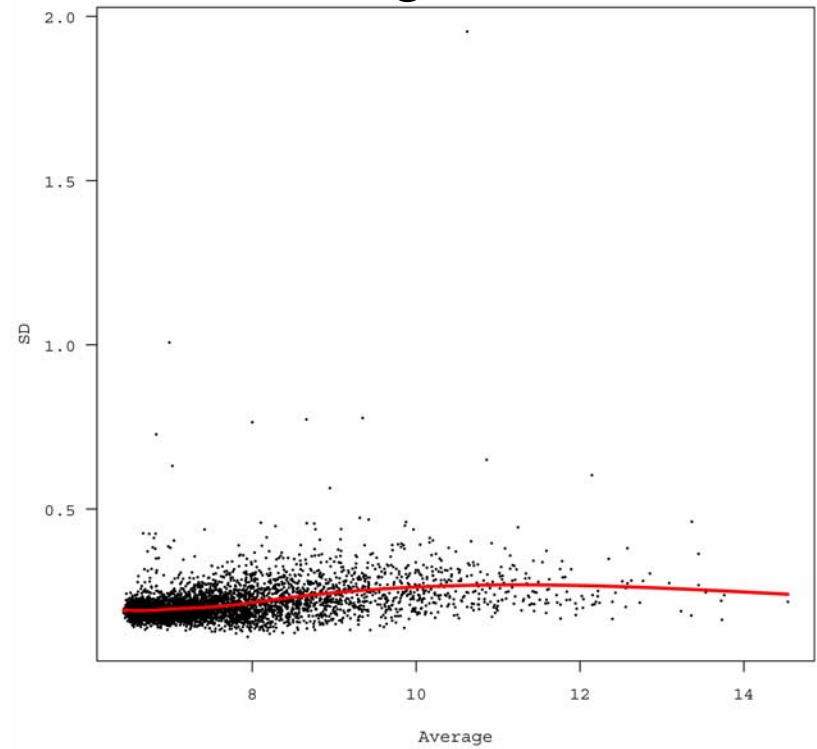
- with A a set of “suitable” pairs chosen by software.
- Obvious Problems:
  - Many negative expression values
  - No log transform

# Why use log?

## Original Scale



## Log Scale



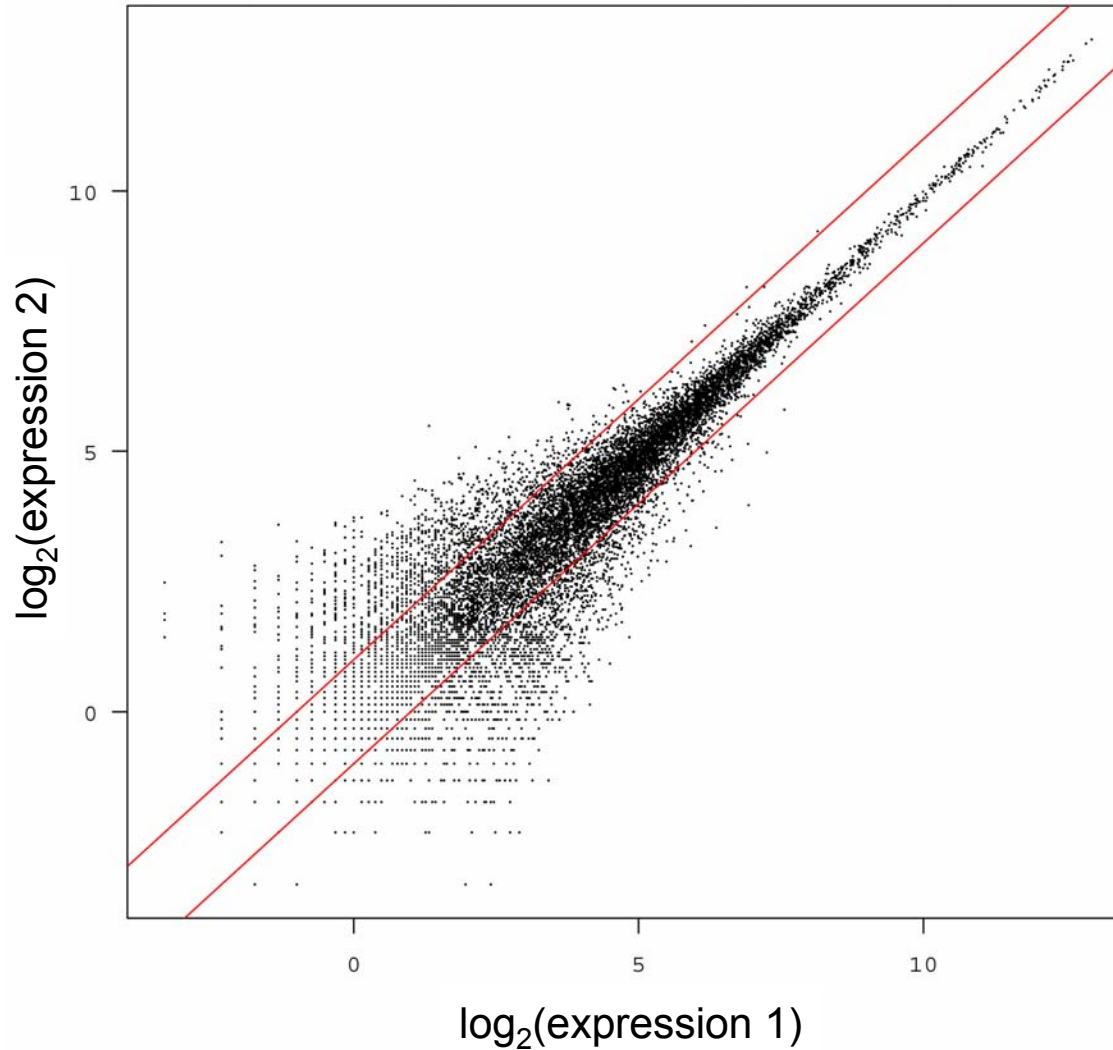
# Current default (MAS 5.0)

- GeneChip<sup>®</sup> new version uses something else

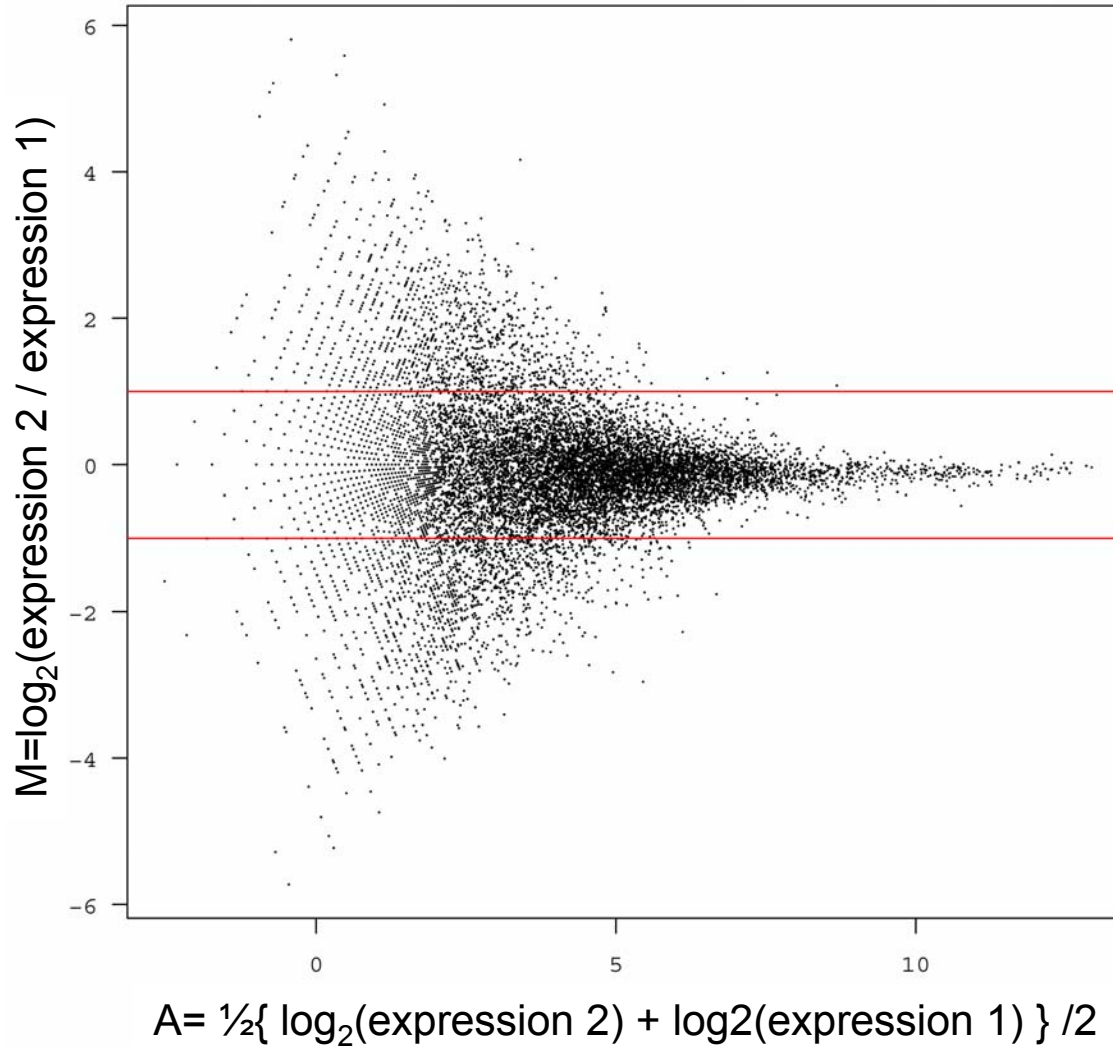
$$signal = TukeyBiweight\{\log(PM_j - MM_j^*)\}$$

- with  $MM^*$  a version of MM that is never bigger than PM.
- Ad-hoc background procedure and scale normalization are used.

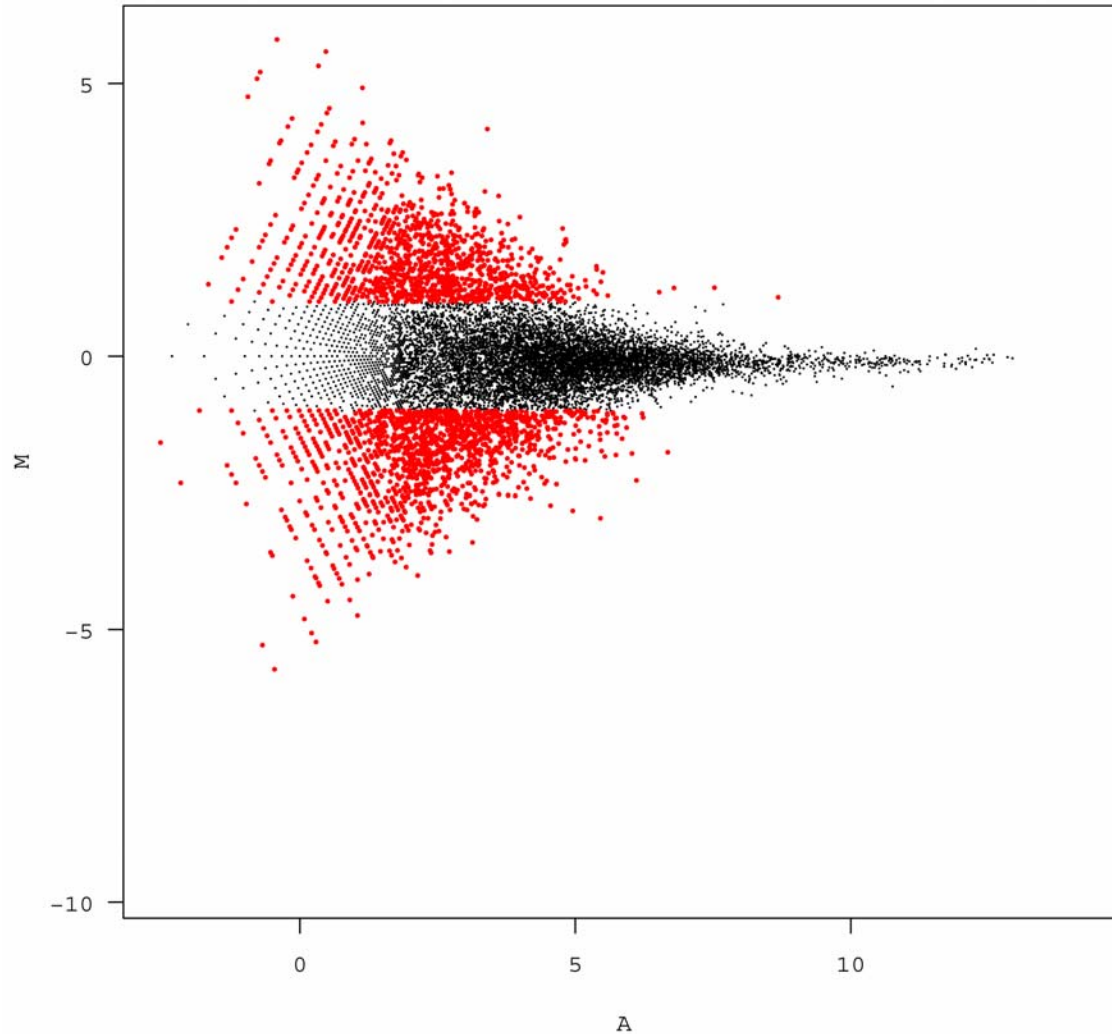
# Expression Data



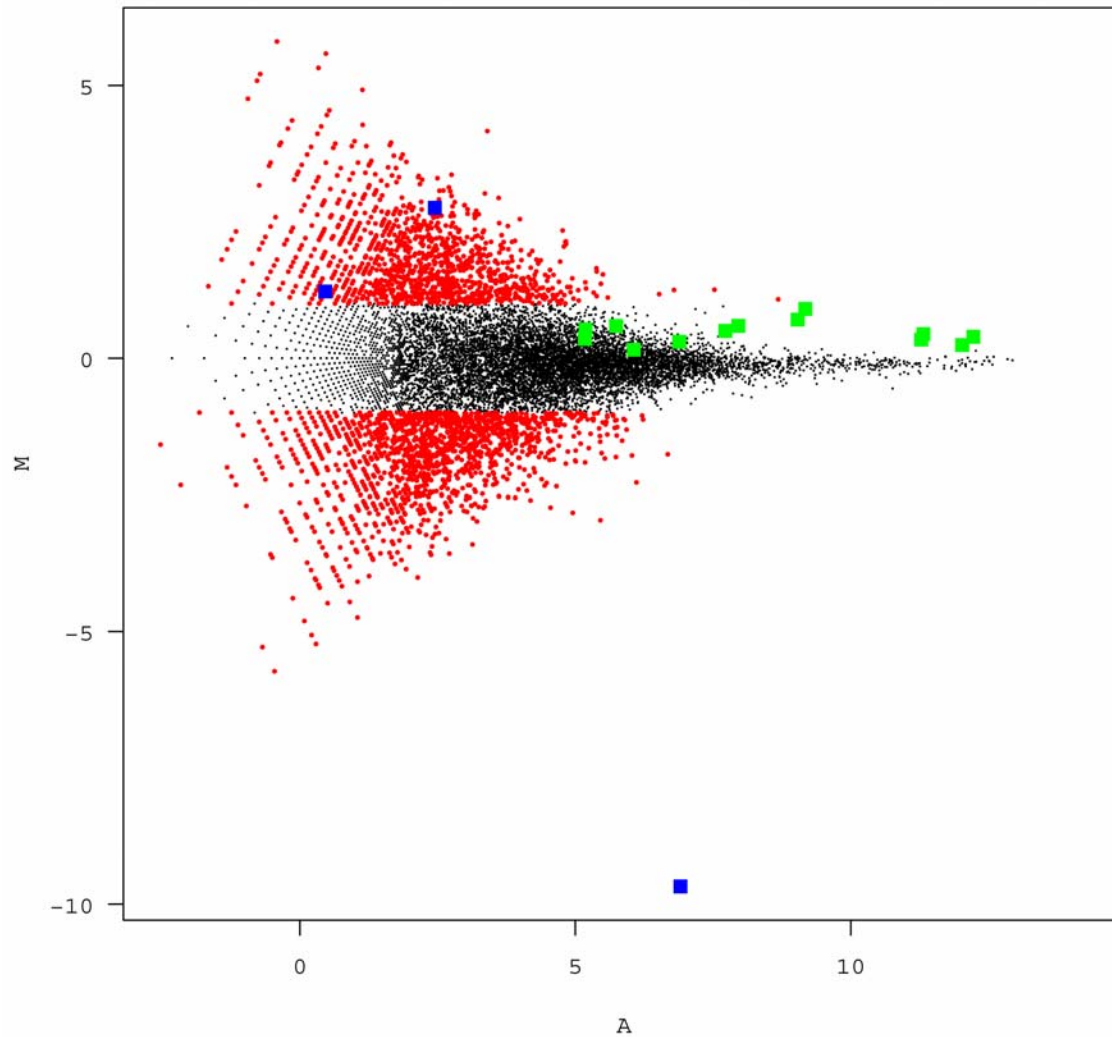
# MA plot



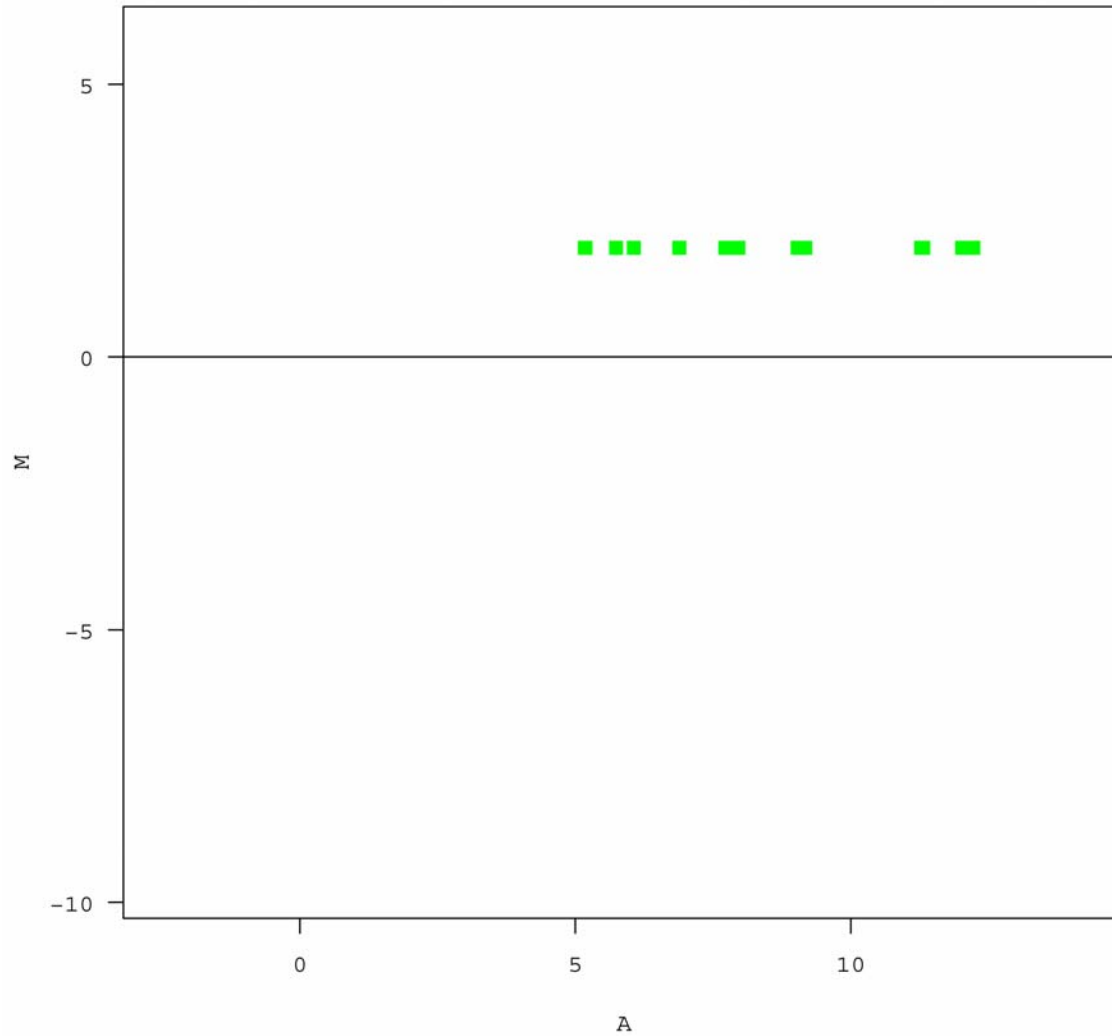
# Can this be improved?



# Use Spike-In Experiment



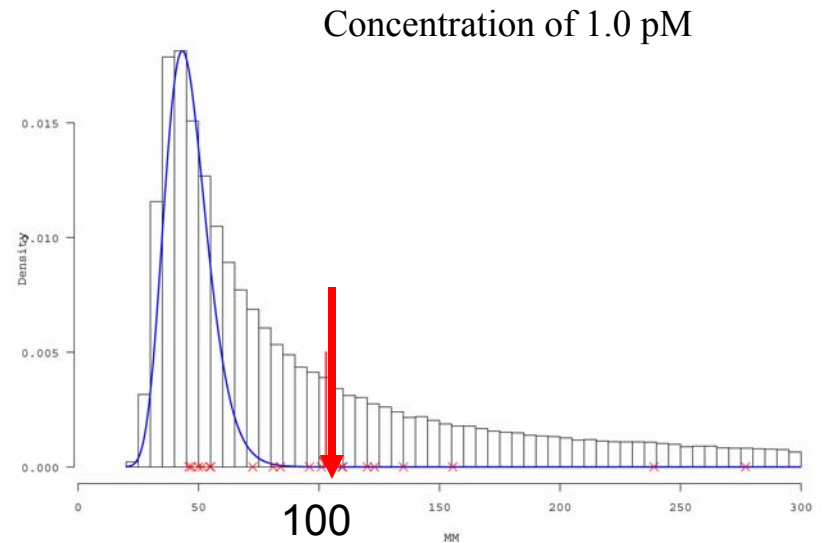
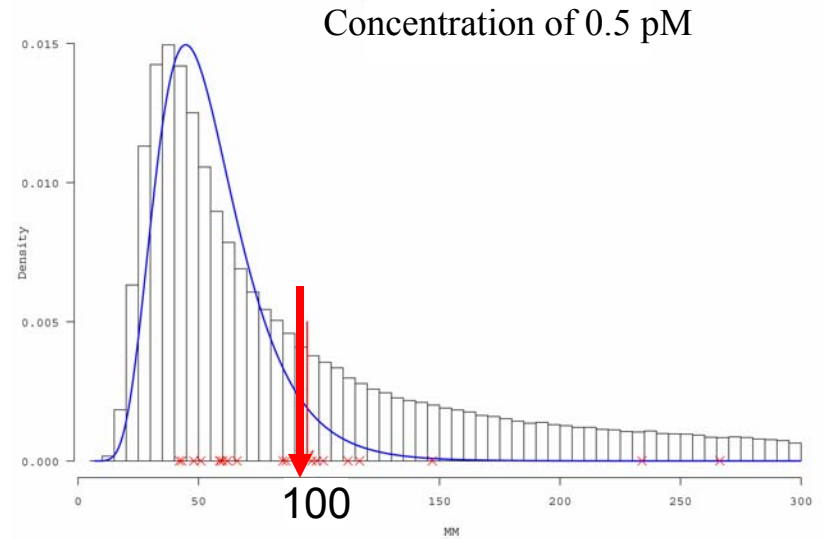
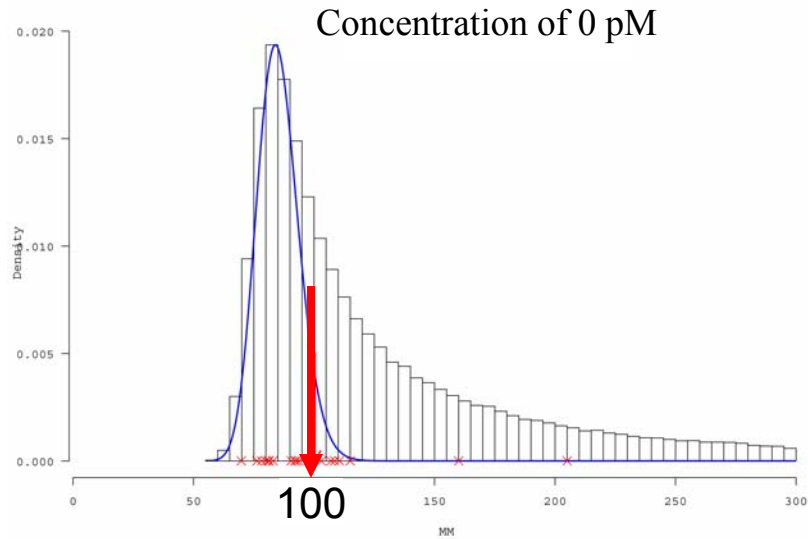
# Use Spike-In Experiment



# Spike-In Data

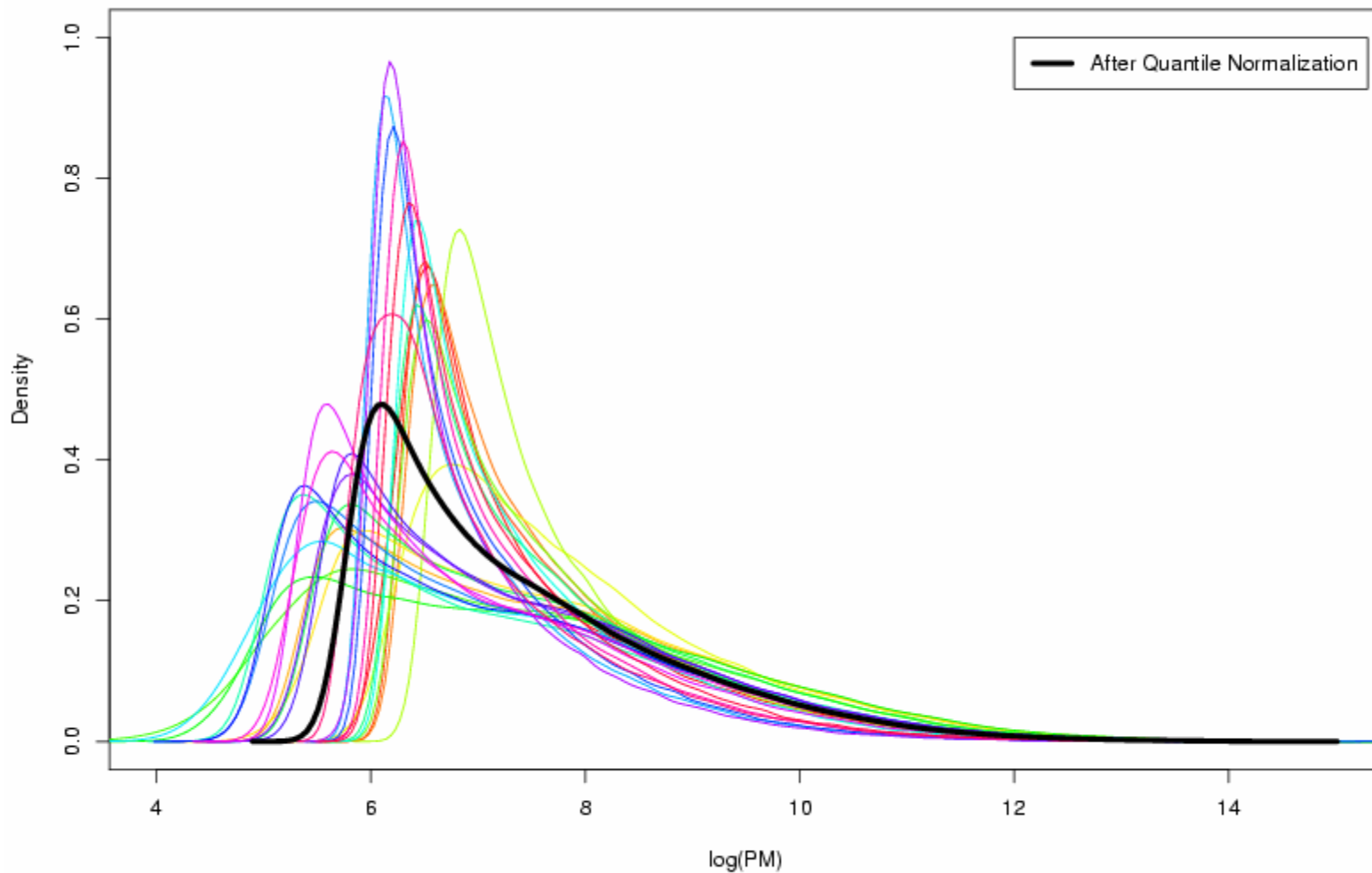
- Go back to probe level and find better ways to
  - Background correct
  - Normalize
  - Correct for probe-specific background
  - Summarize
- Next 4 slides: transcripts spiked-in at increasing concentrations across arrays

# Why background correct?

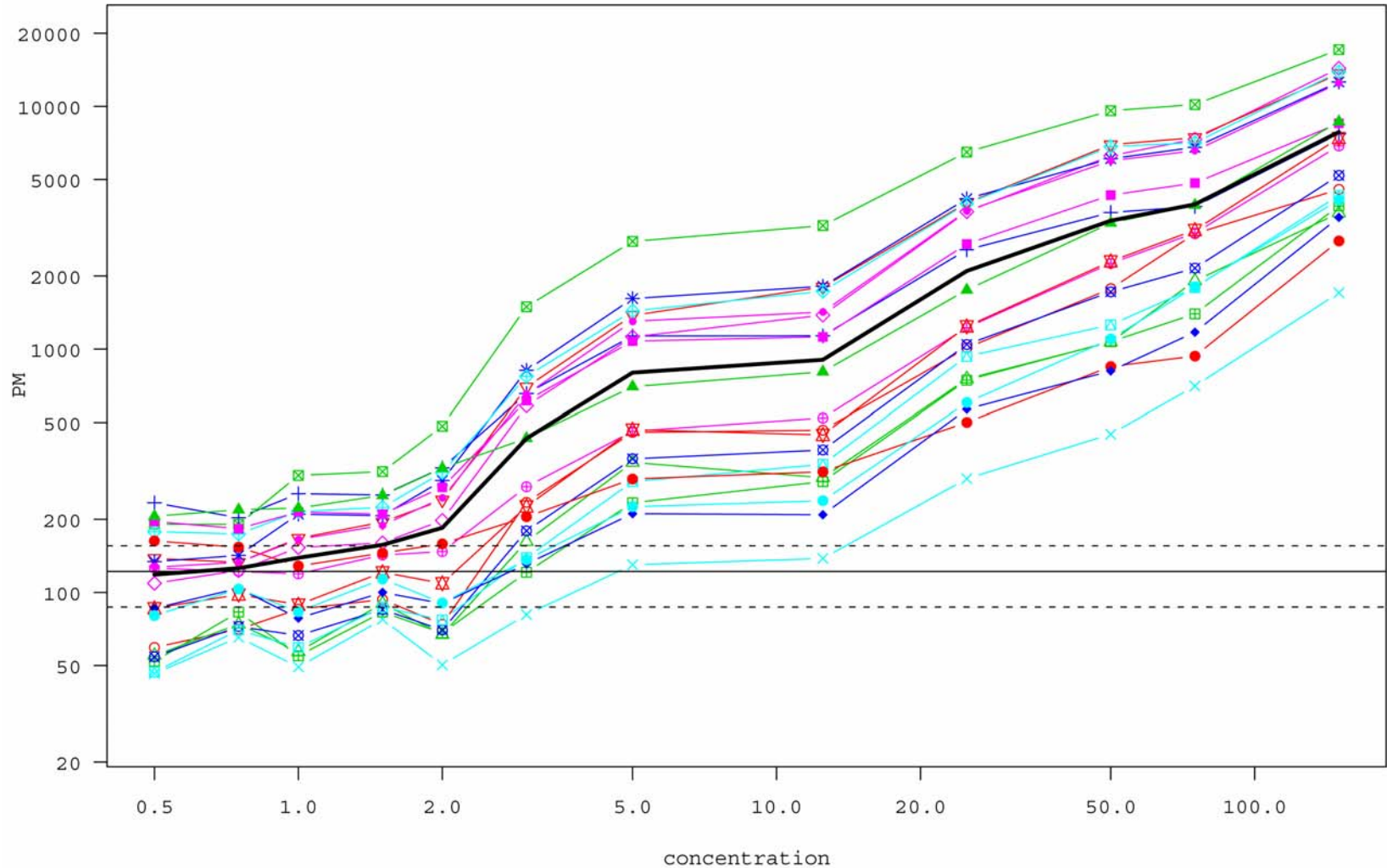


# Why normalize?

Density of PM probe intensities for Spike-In chips



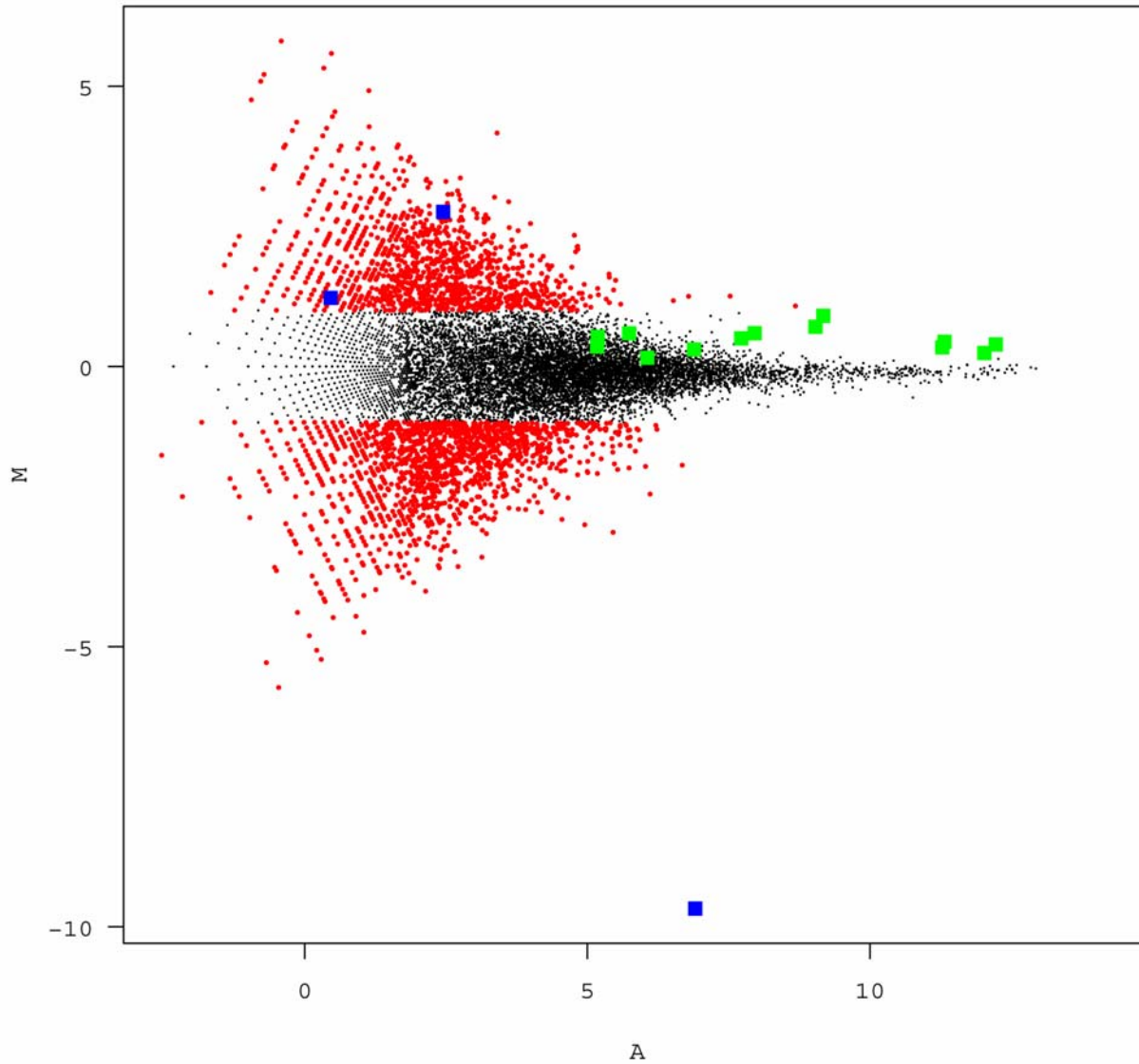
# Why fit log scale additive model?



# Statistical Model

- Instead of subtracting MM
- Assume  $PM = B + S$
- To estimate  $S$ , use expectation:  $E[S|B+S]$
- After quantile normalization, assume:
$$\log_2 S_{ij} = E_i + P_j + \varepsilon_{ij}$$
- Estimate  $E_i$  using robust procedure
- We call this procedure **RMA**
- Does it make a difference?

# MAS 5.0



Ranks

1

270

2074

3063

3935

4639

4652

5149

5372

5947

6448

6870

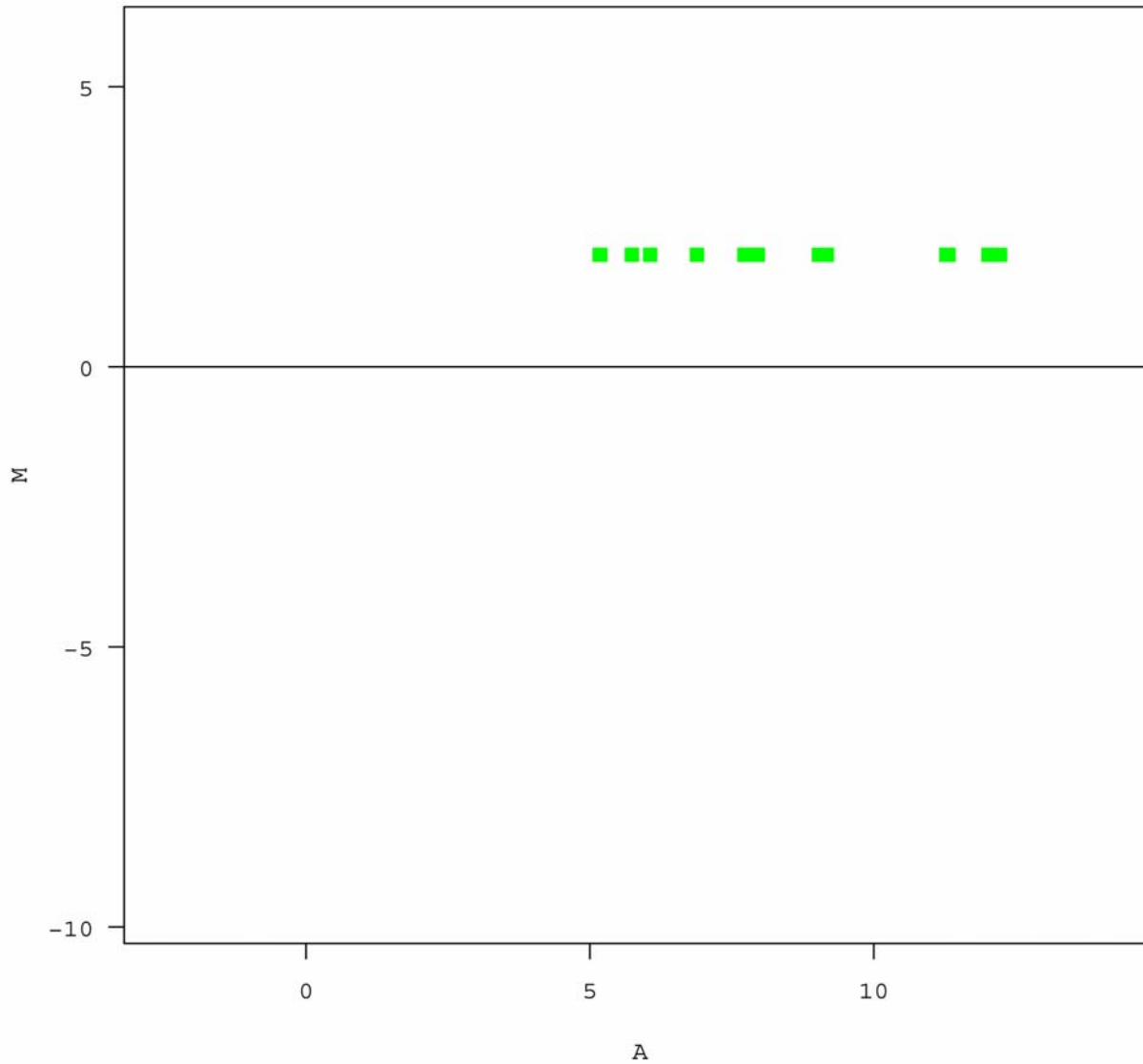
7037

7549

8429

9721

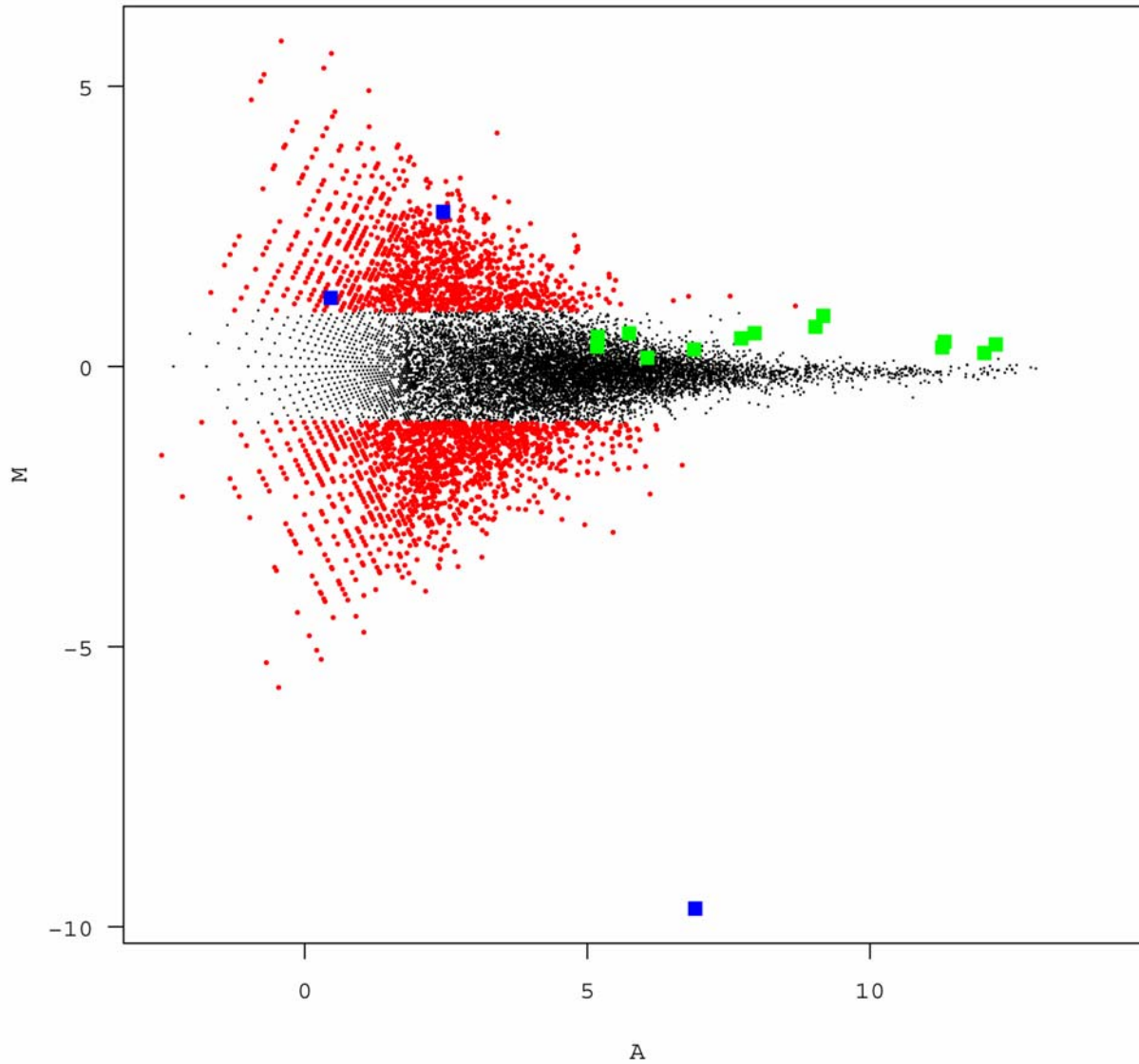
# Perfect



Ranks

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

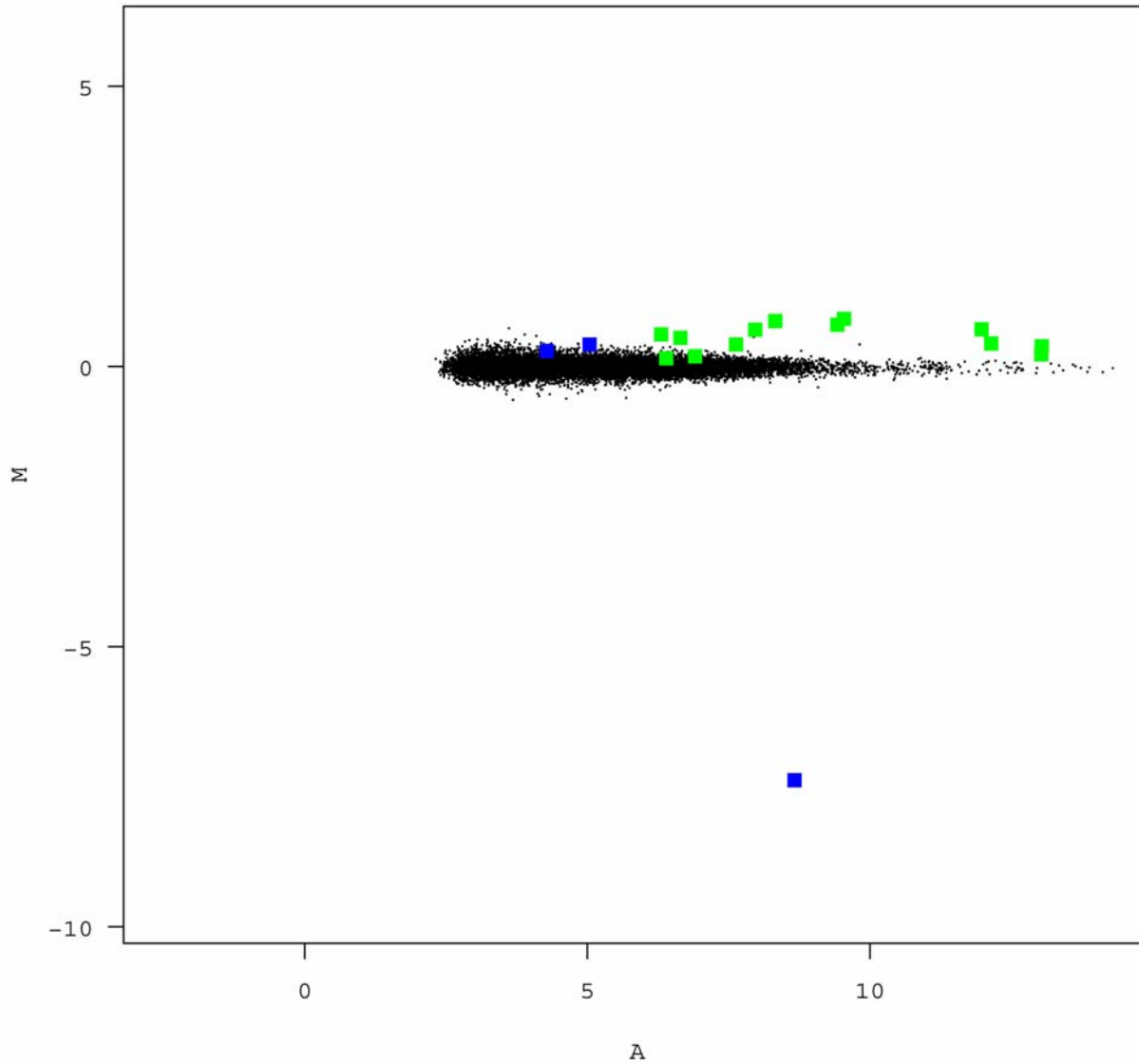
# MAS 5.0



Ranks

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

# RMA



Ranks

1

2

3

4

6

7

10

16

45

56

58

88

406

999

1643

2739

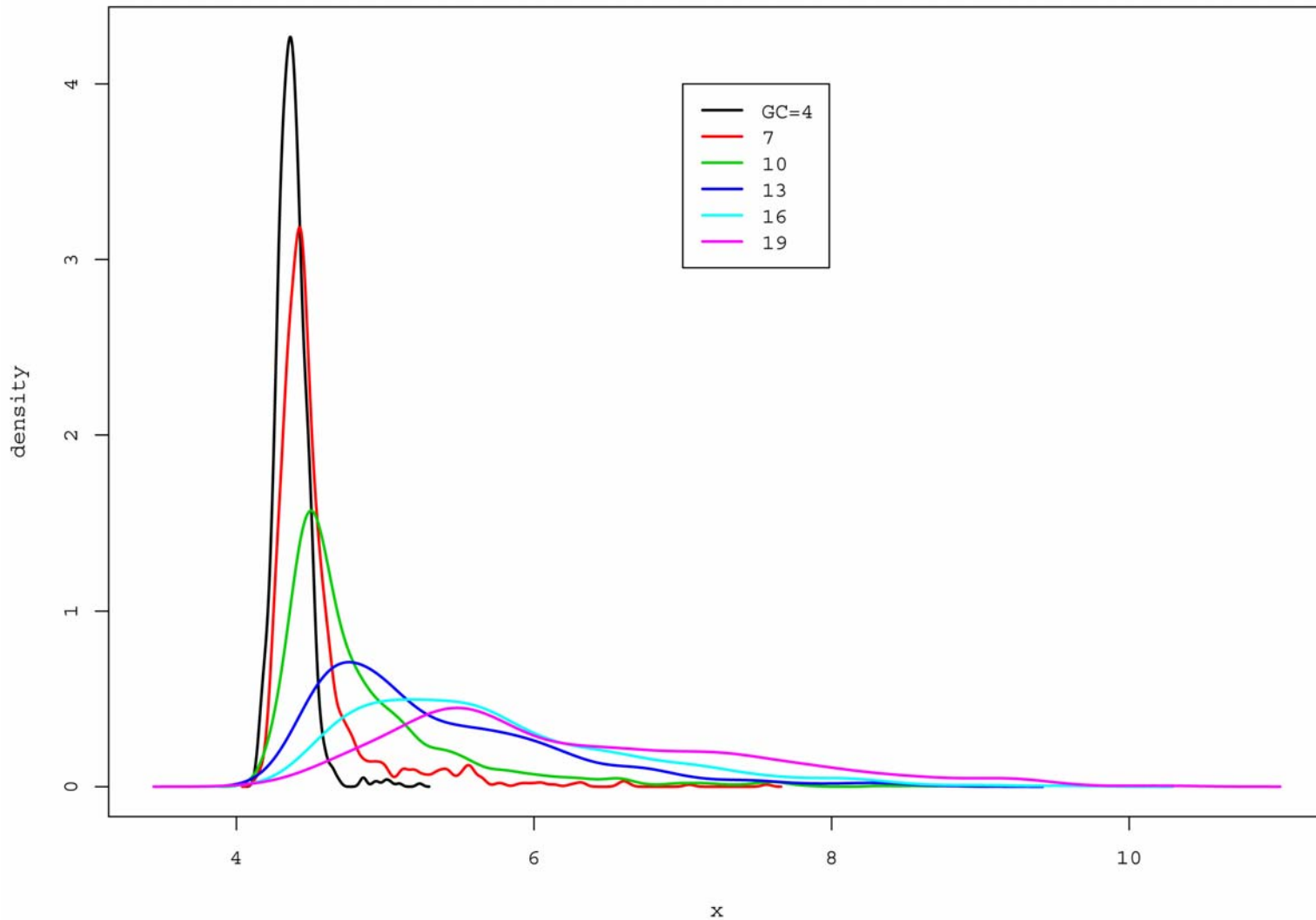
# References

- Irizarry et al: Biostatistics (2003)
- Bolstad et al: Bioinformatics (2003)
- Irizarry et al: NAR (2003)
- affy R package ([www.bioconductor.org](http://www.bioconductor.org))

# Current Work

- Improve Background Model
- Probe specific background correction
- Use sequence information to do this
- MatchAffy Software (Gentleman)

# GC-content Specific Background Correction



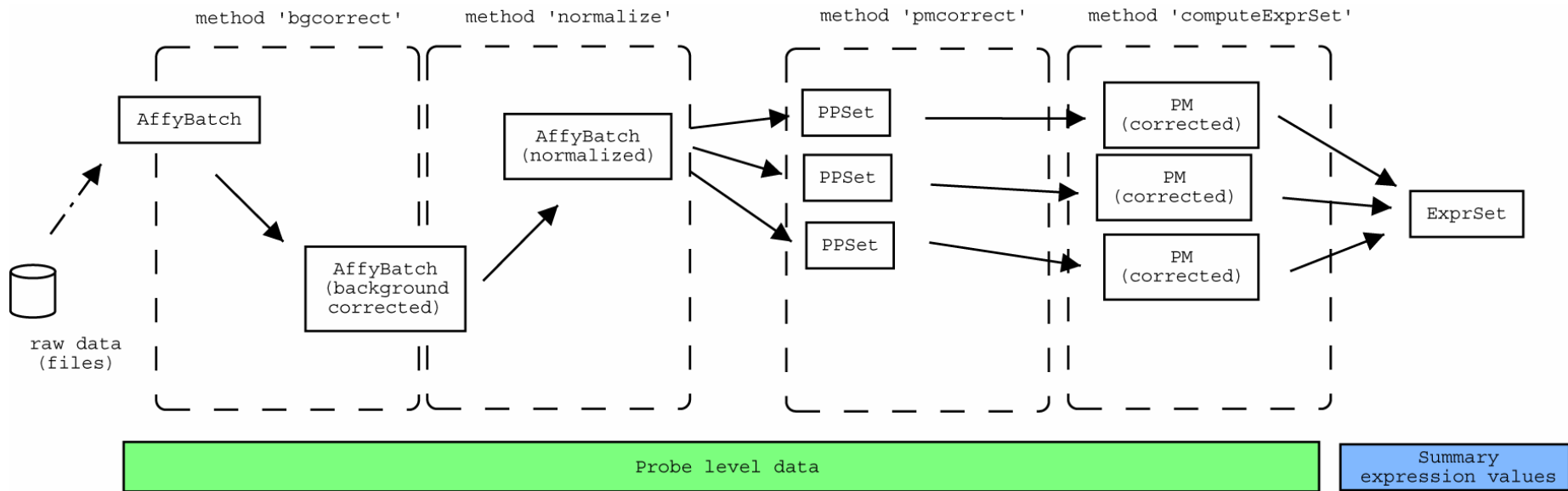
# Modularity

Background correction

Normalization

Probe Specific BG correction

Summary statistic



# Conclusion

- Working with probe level data can improve bottom line results
- Important to have flexible software

# Acknowledgements

Ben Bolstadr

Francois Collin

Leslie Cope

Laurent Guatier

Robert Gentleman

Bridget Hobbs

Terry Speed

Zhijin Wu

Affymetrix

Bioconductor

Genelogic

JHMI Microarray Core Facility

R

# Bottom Line

Assessment	MAS 5	RMA
FC>2, FP	3072.0	1.0
FC>2, TP	3.7	1.0
TP, FP=25	1.0	11.1
AUC, FP<100	6%	54%

FC = Fold Change

TP = True Positive

FP = False Positive

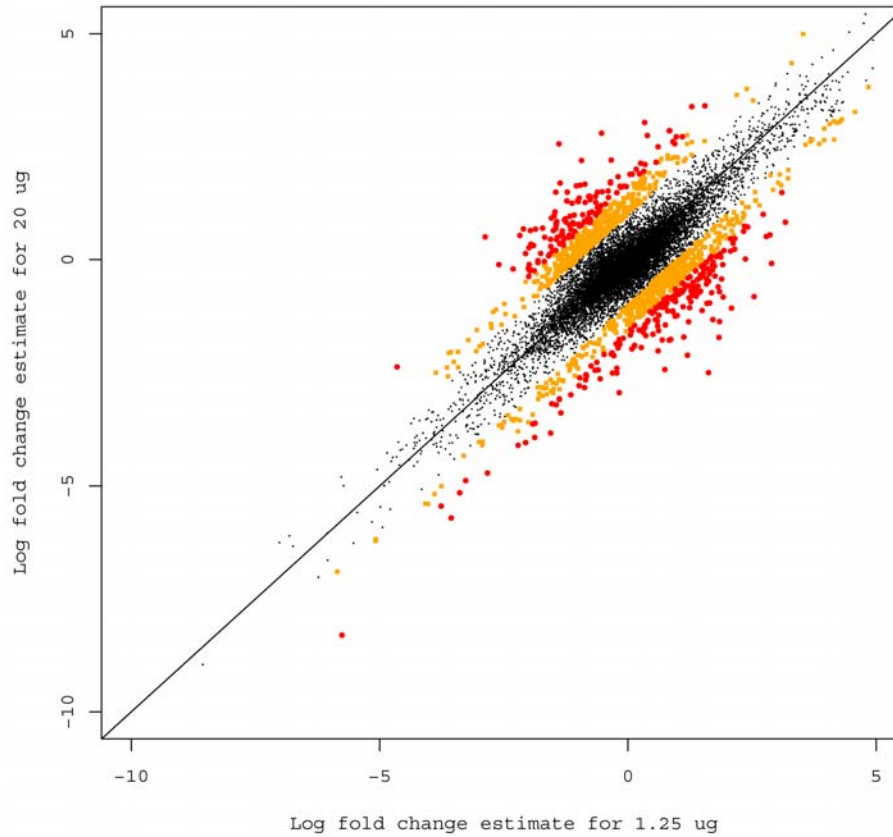
AUC = Area under the ROC curve

# \*What makes the difference in AUC?

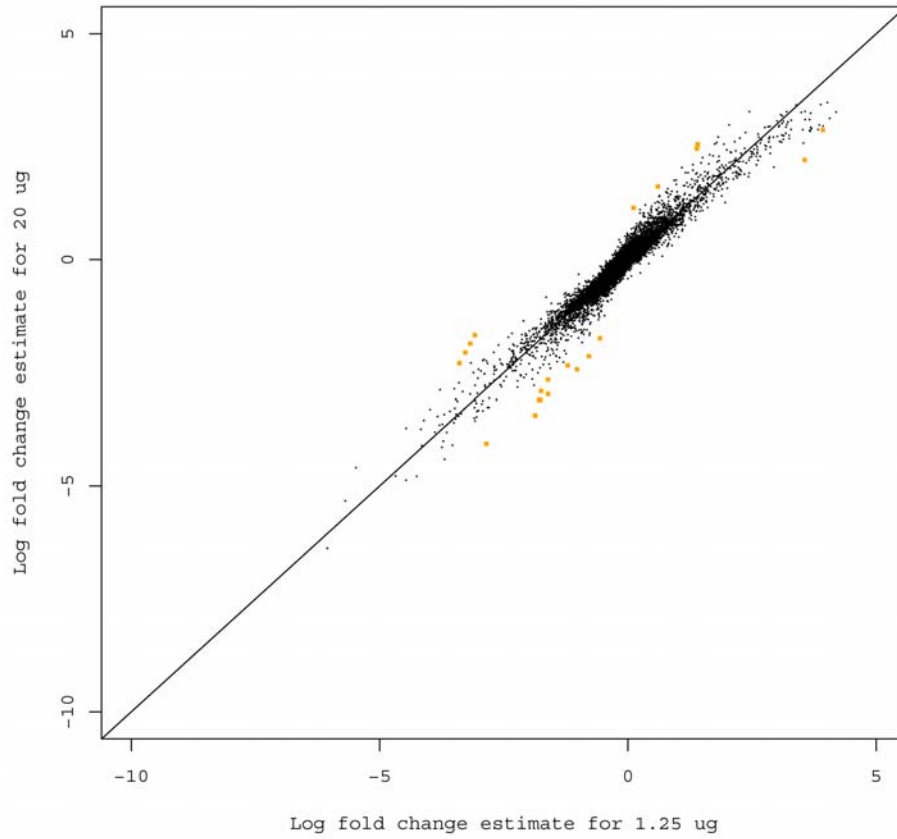
	Operation	Improvement
Preprocessing	Log	60%
	Background (no MM)	78%
	Normalization	343%
	Robustness	5%
	Test-Statistic	20%*

\* 58% improvement over t-test

# MAS 5.0



# RMA

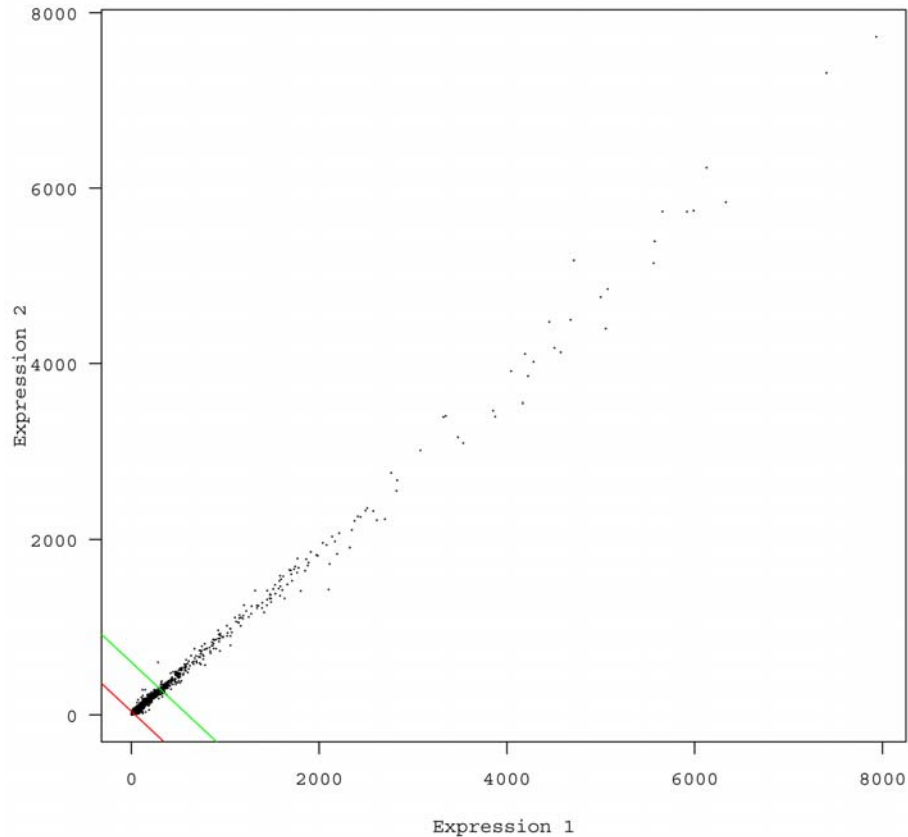


# What makes the difference (easy)?

Operation	Improvement
Log	276%
Background (no MM)	105%
Normalization	18%
Robustness	0.3%
Test-Statistic	NA

# Can this be improved?

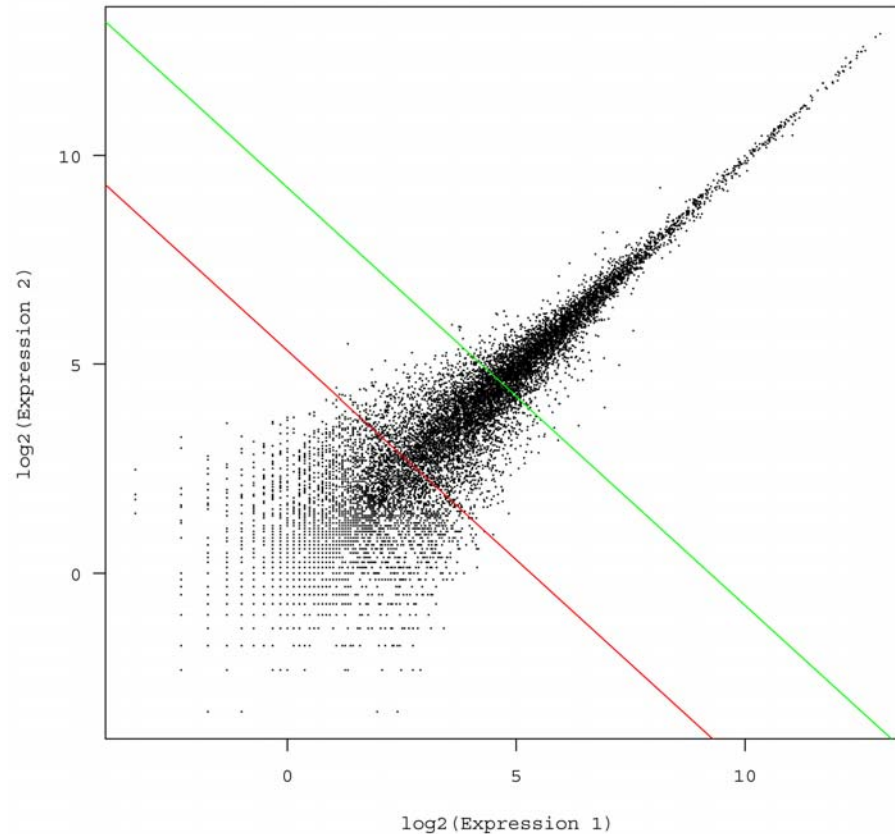
$R=0.99$



95% of data below green line, 50% below red

# Can this be improved?

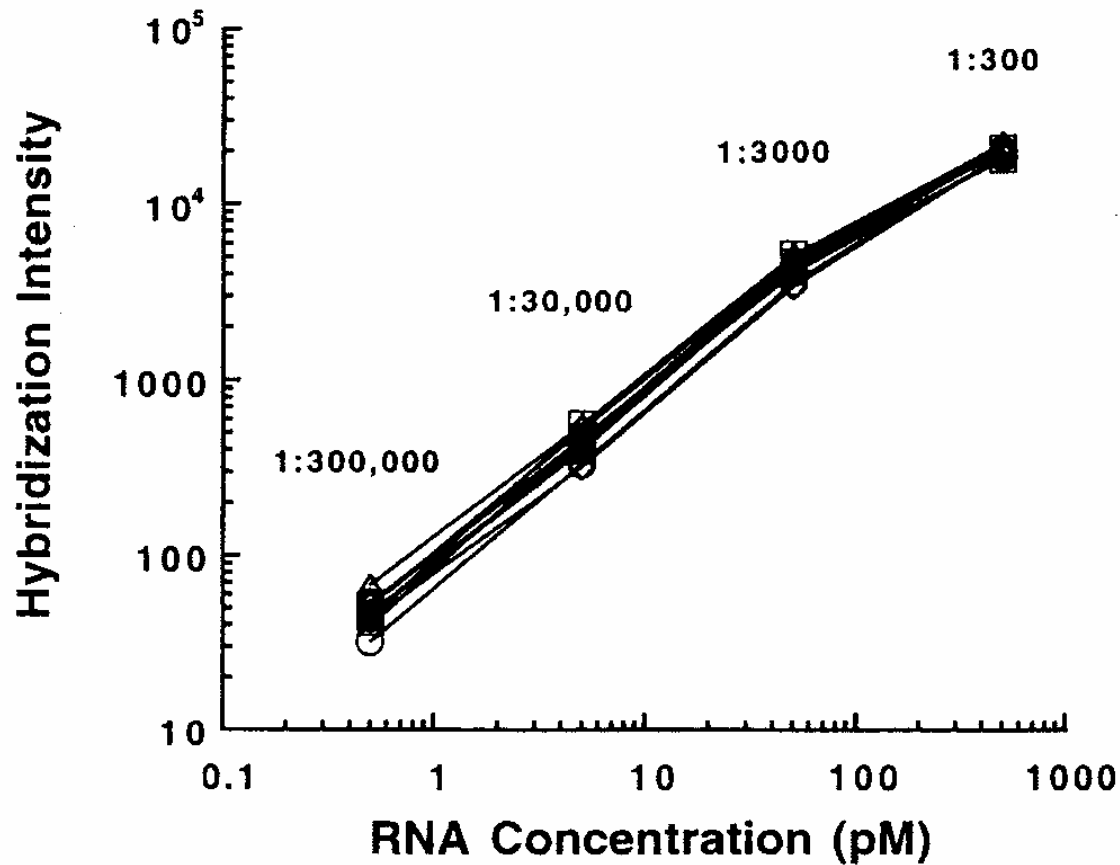
$R=0.80$



95% of data below green line, 50% below red

# What is the evidence?

Lockhart et al: Nature Biotechnology (1996)



# Affymetrix files

- Main software from Affymetrix company *MicroArray Suite - MAS*, now version 5.
- **DAT** file: Image file,  $\sim 10^7$  pixels,  $\sim 50$  MB.
- **CEL** file: Cell intensity file, probe level PM and MM values.
- **CDF** file: Chip Description File. Describes which probes go in which probe sets and the location of probe-pair sets (genes, gene fragments, ESTs).