

# Low Level Analysis of Affymetrix GeneChip Data

Ben Bolstad

May 5, 2003



# Overview

- **Introduction**
- Brief Technology Overview
- Preprocessing Steps
  - Background correction/Signal adjustment
  - Normalization
  - Summarization
- Comparing the effect of different preprocessing methods on expression estimates
- Software
- Future/Ongoing work

# Introduction

- What is low level analysis and why do we do it?
  - Analysis and manipulation of probe intensity data
    - Expression calculation: Background, Normalization, Summarization
    - Determining presence/absence
    - Quality control diagnostics
  - Hopefully it will allow us to produce better, more biologically meaningful gene expression values
  - We want accurate (low bias) and precise (low variance) gene expression estimates

# Overview

- Introduction
- **Brief Technology Overview**
- Preprocessing Steps
  - Background correction/Signal adjustment
  - Normalization
  - Summarization
- Comparing the effect of different preprocessing methods on expression estimates
- Software
- Future/Ongoing work

# Brief Technology Overview

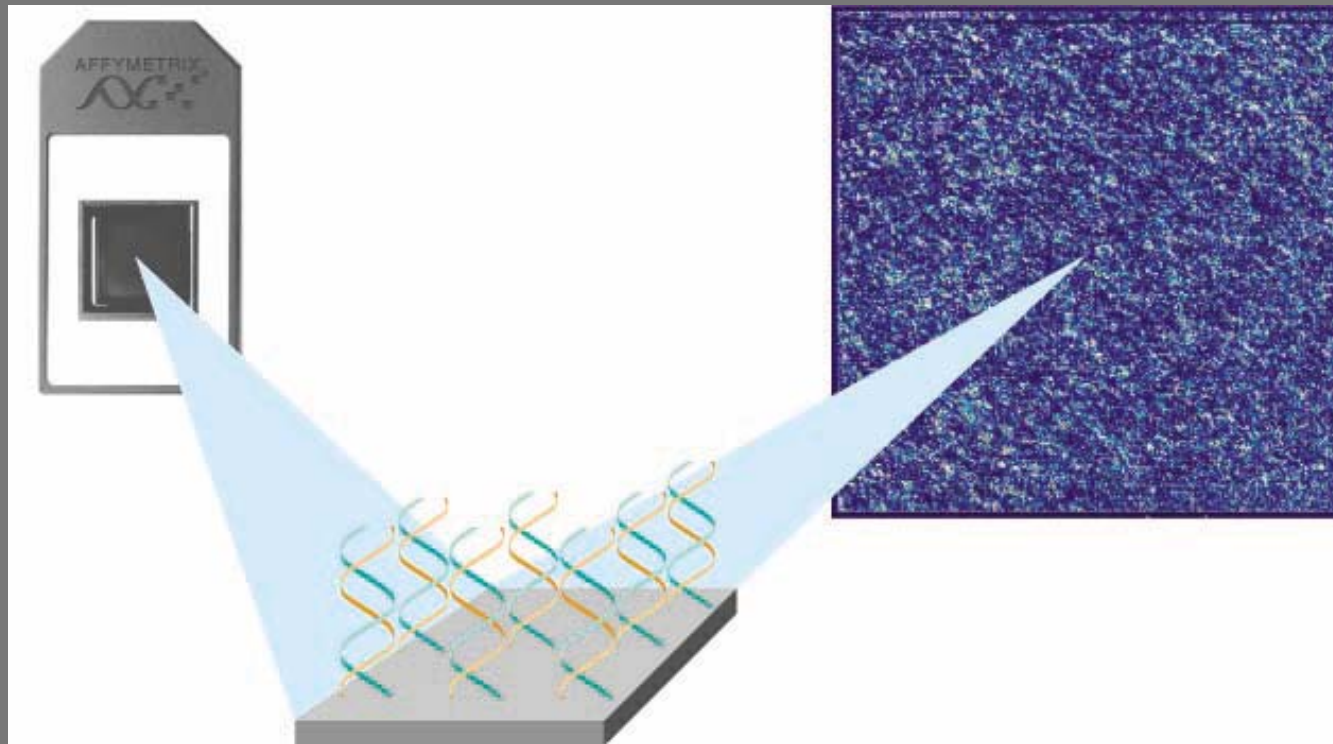
- High density oligonucleotide array technology as developed by Affymetrix  
<http://www.affymetrix.com>
- Known as the .....

# The GeneChip

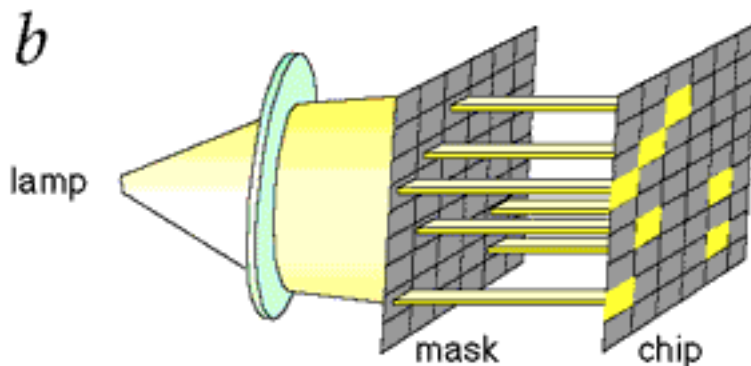
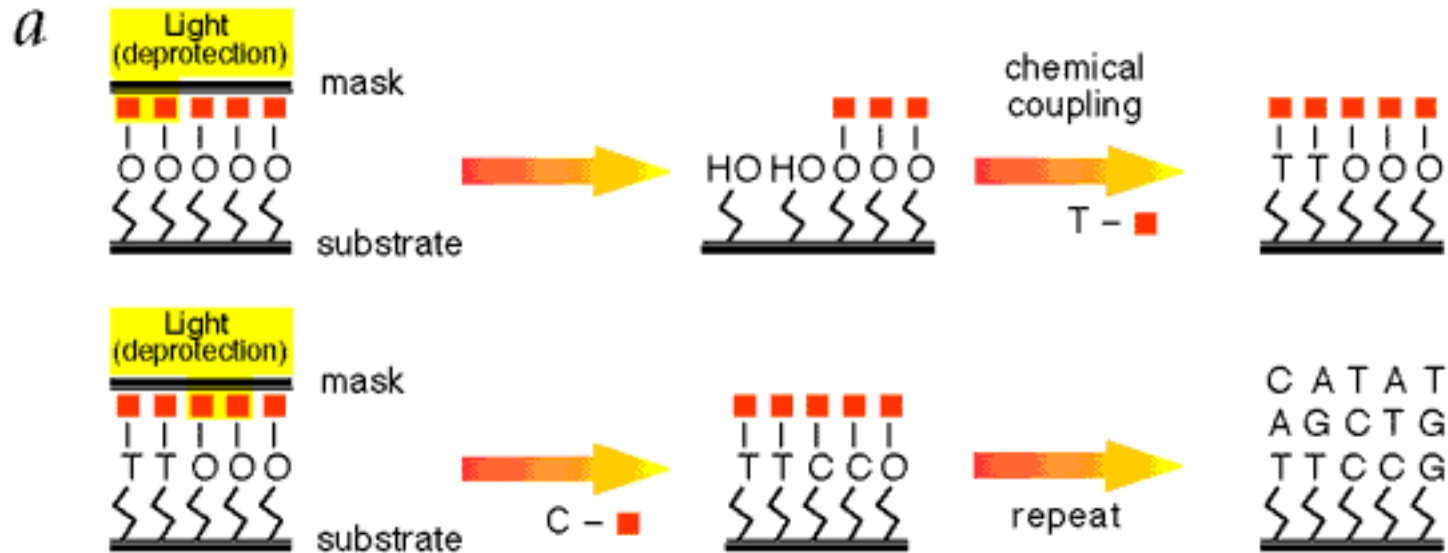


- 25 mer oligonucleotides interrogate genes, sequences of interest
- Affymetrix has rules for probe selection based upon: hybridization properties, specificity, potential for cross-hybridization
- Around 500,000 probes on an array
- Typically a group of 11-20 probes (pairs) all for same gene constitute a probeset

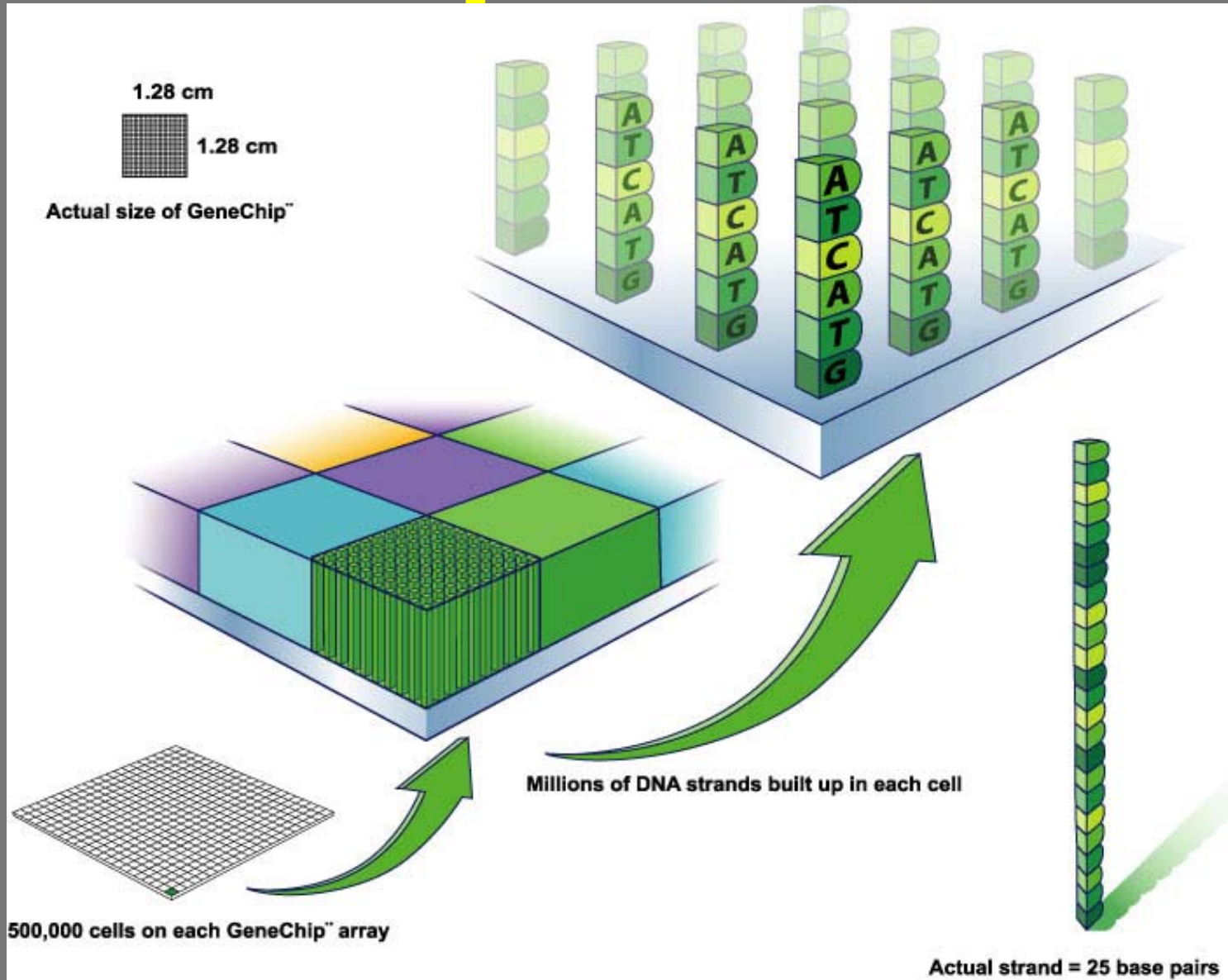
# From Chip To Data



# Constructing the Chip



# Focusing on a Single GeneChip Cell Location



# Two Probe Types

PM: the Perfect Match

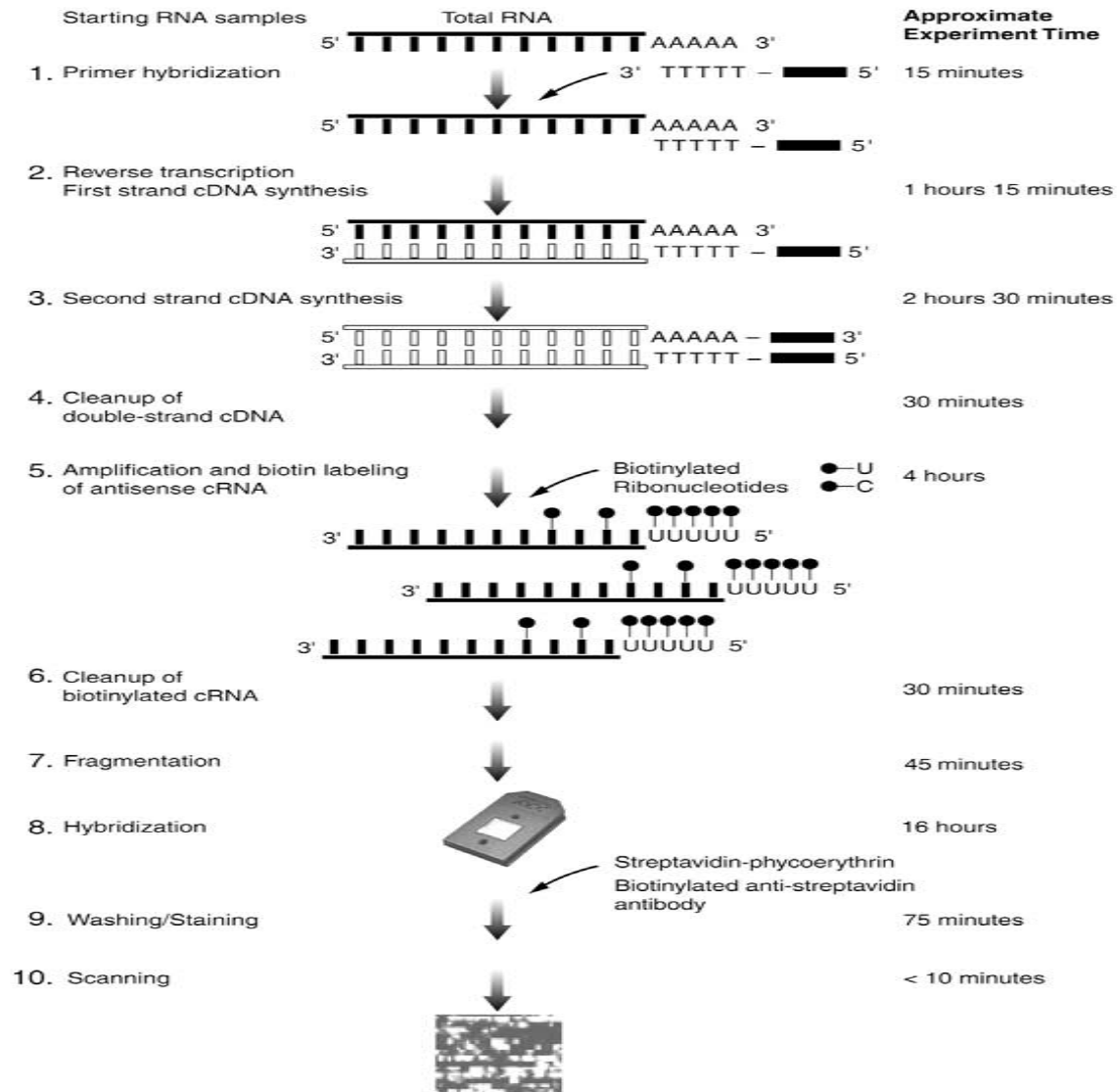
MM: the Mismatch differing from the Perfect Match only at the central base

PM: CAGACATAGTGT**C**TGTGTTTCTTCT

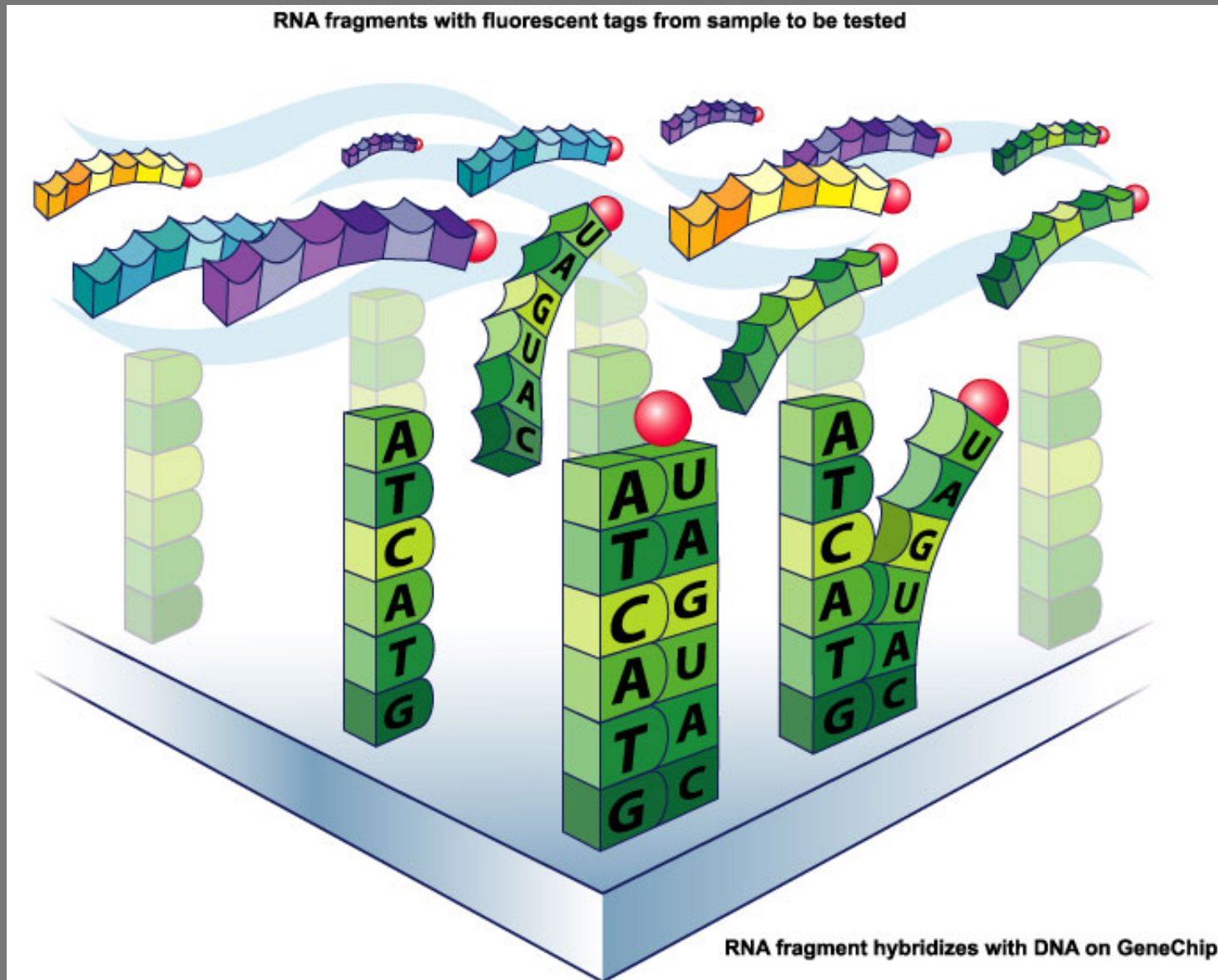
MM: CAGACATAGTGT**G**TGTGTTTCTTCT

# Sample Preparation

## Eukaryotic Target Labeling for GeneChip® Probe Arrays



# Hybridization to the Chip



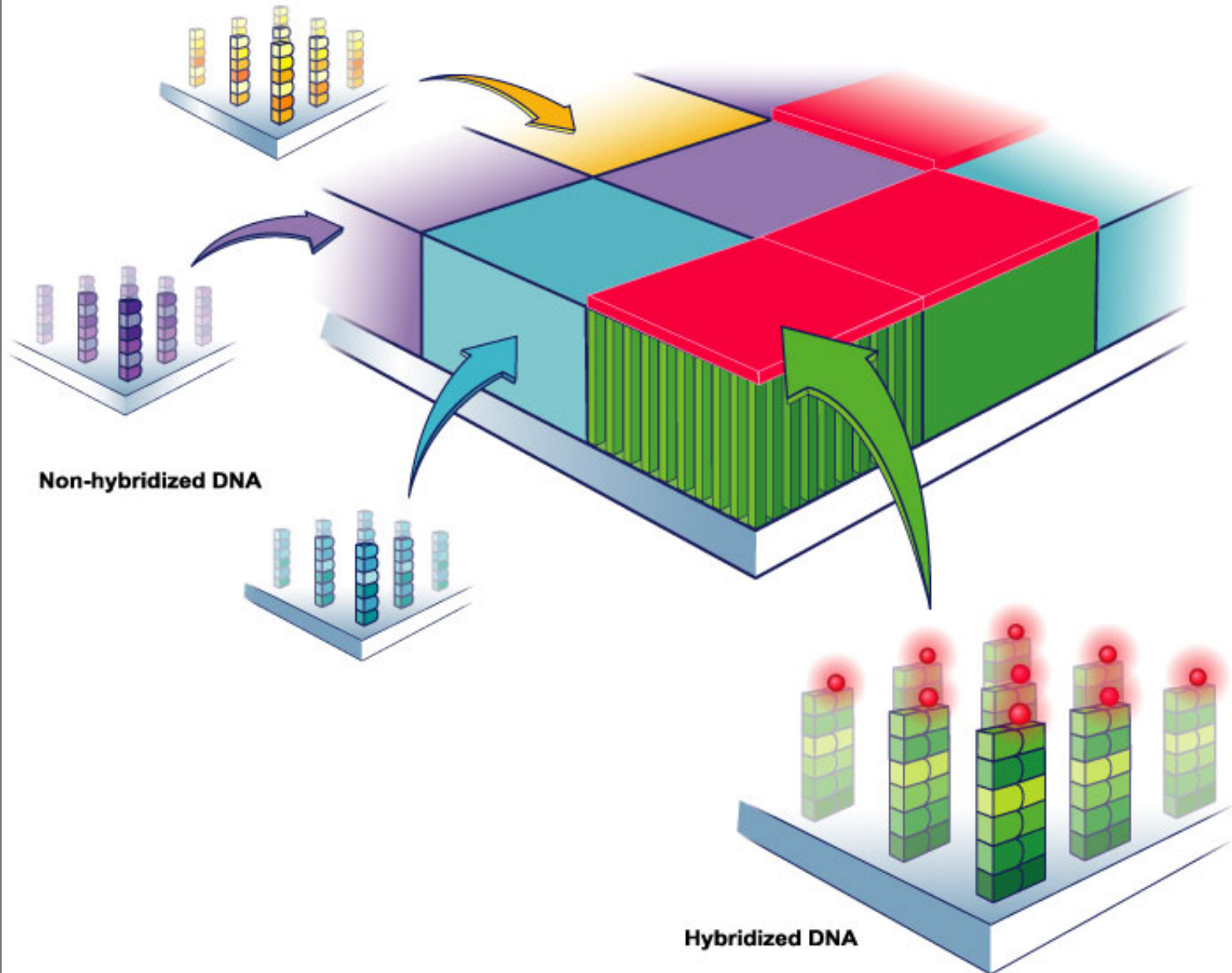
# The Chip is Scanned

- A scanner measures fluorescence

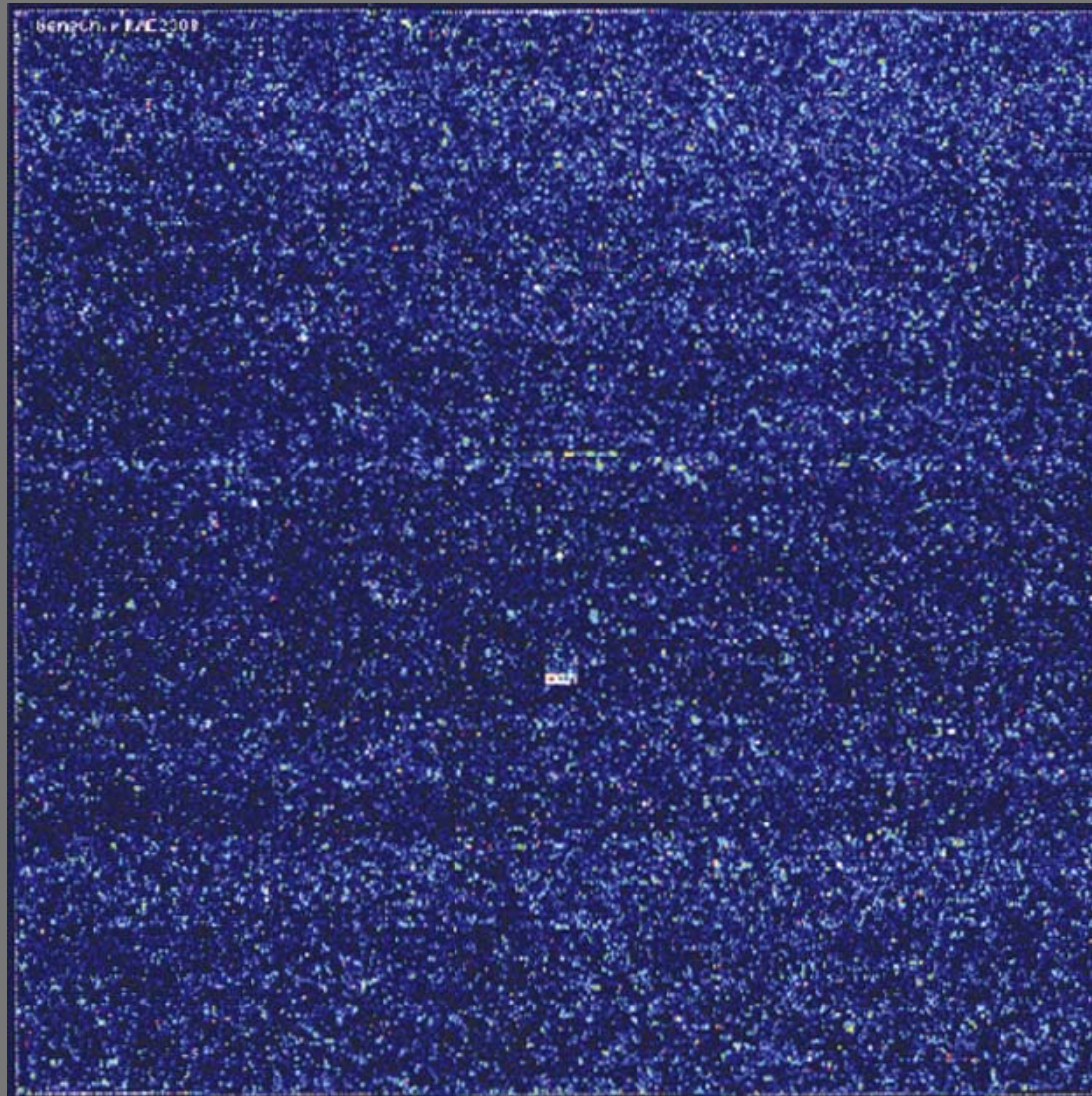


# In the Scanner

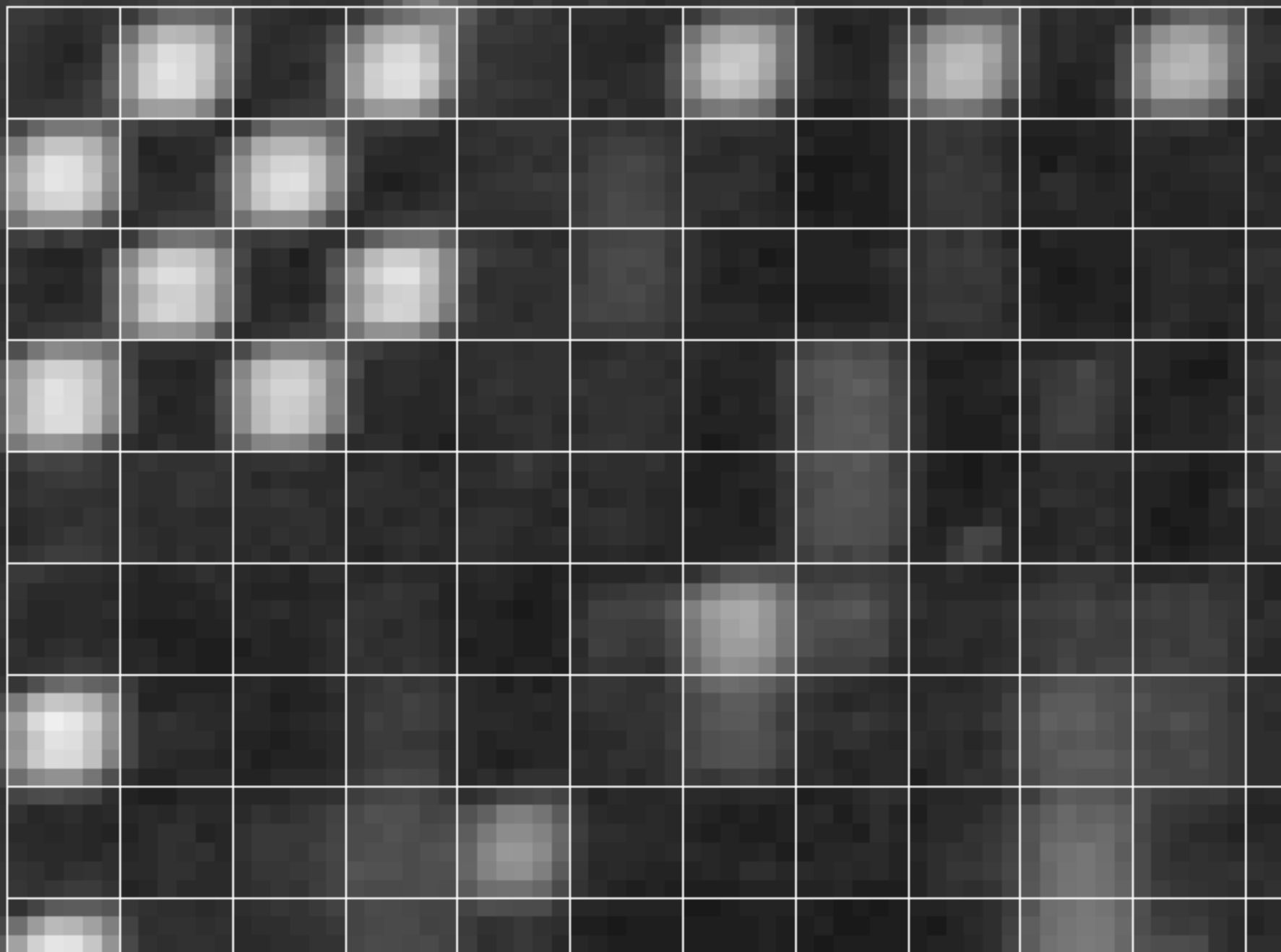
Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow

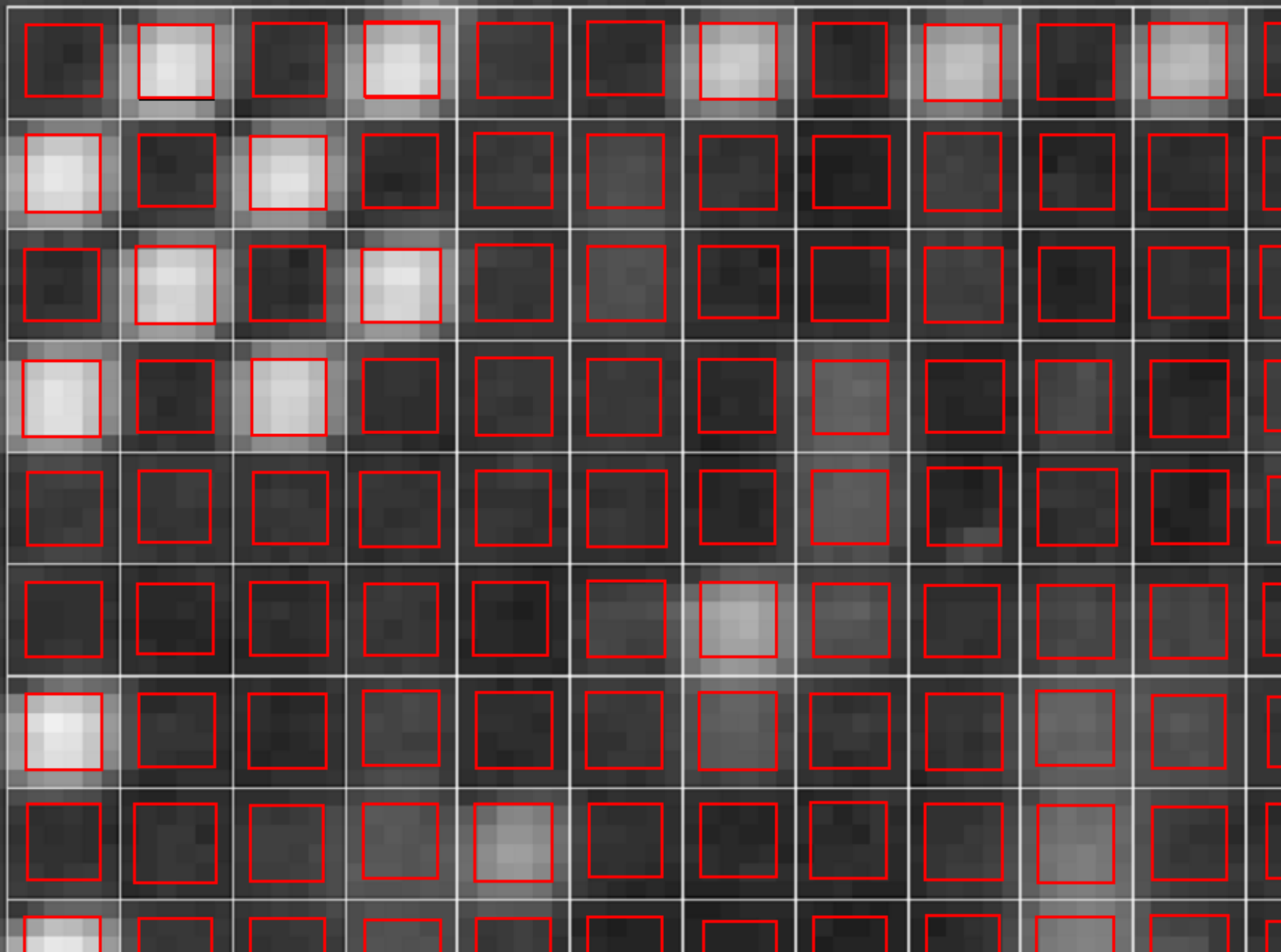


# Producing an Image











# Overview

- Introduction
- Brief Technology Overview
- **Preprocessing Steps**
  - Background correction/Signal adjustment
  - Normalization
  - Summarization
- Comparing the effect of different preprocessing methods on expression estimates
- Software
- Future/Ongoing work

# Computing Expression Measures as a Three Step Procedure

1. Background/Signal adjustment (B)
2. Normalization (N)
3. Summarization (S)

Let  $X$  be cel file data from multiple arrays then

Expression values =  $S(N(B(X)))$

# Background/Signal Adjustment

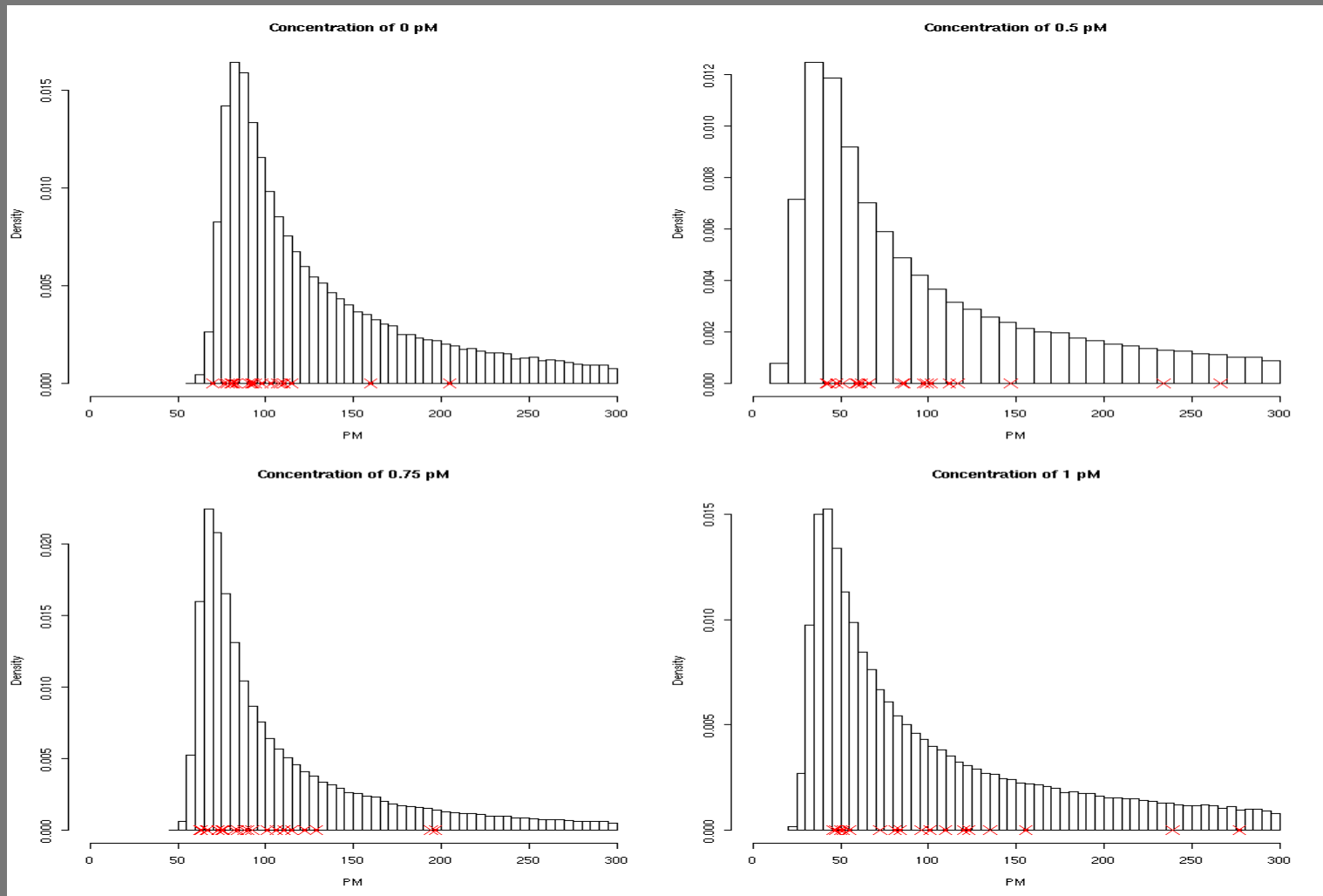
- A method which does some or all of the following
  - Corrects for background noise, processing effects
  - Adjusts for cross hybridization
  - Adjust estimated expression values to fall on proper scale
- Probe intensities are used in background adjustment to compute correction (unlike cDNA arrays where area surrounding spot might be used)

# Background Signal Methods

- Affymetrix
  - Location dependent background based on grids
    - I will refer to this as the MAS 5 background
  - Originally proposed subtracting MM from PM but this is problematic because as many as a third of MM's are greater than the respective PM
    - No longer used
  - Now uses what they refer to as the Ideal Mismatch which is MM when possible and something else when not possible (designed so that there is now no negatives)
    - Call this IMM

# Original RMA Background

- Convolution model is suggested by looking at density of observed empirical distributions



# Convolution Model

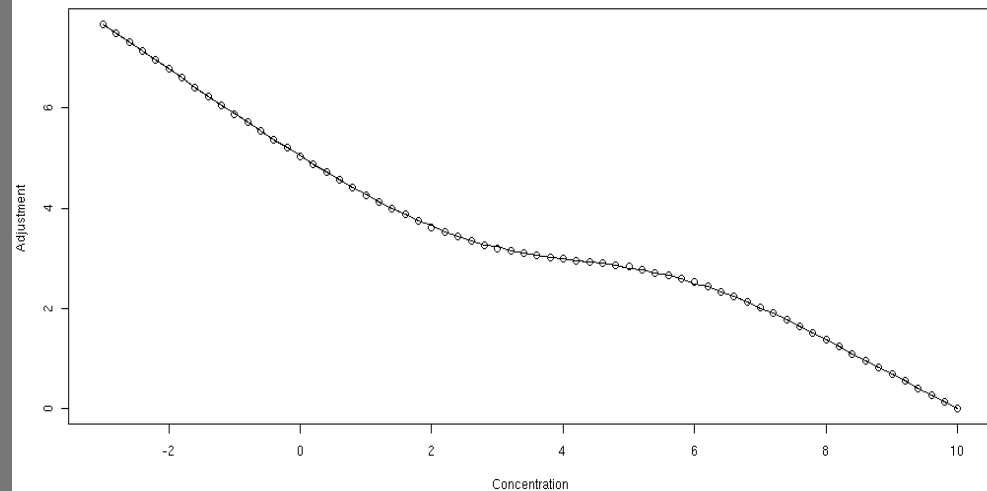
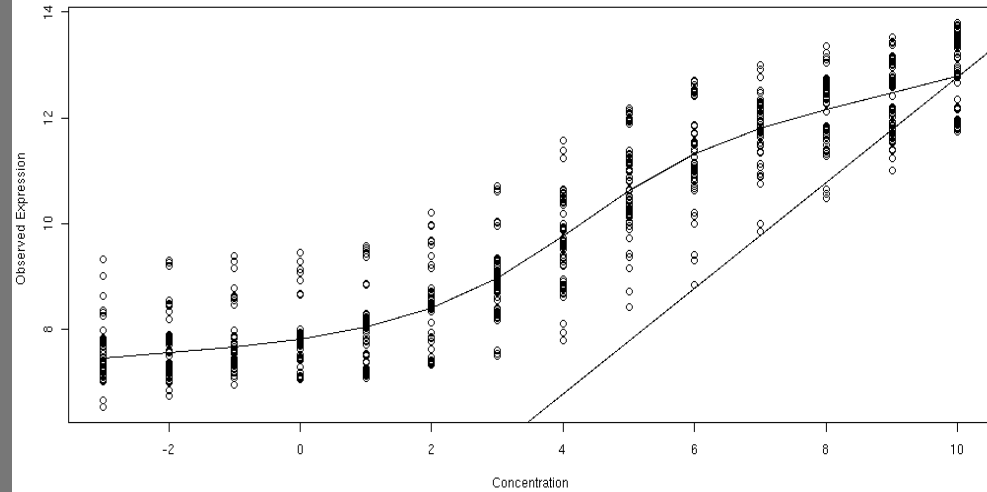
- $O = S + N$ 
  - $O$  is observed PM,  $S$  is signal (assumed exponential),  $N$  is noise (assumed normal, truncated at zero)
- Correction is then

$$E(S | O = o) = a + b \frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{o-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) - \Phi\left(\frac{o-a}{b}\right) - 1}$$

$$a = o - \mu - \sigma^2 \alpha, b = \sigma$$

# A Standard Curve Adjustment Based on Spike-in Information

- Observes that there is a curve that relates observed expression and spike-in concentration. The ideal would be to have a linear relationship between concentration and computed expression. The curve gives us a concentration dependent adjustment



# What About Non Spike-ins?

- We don't know a concentration for most probesets. If we did, or if we had a variable that related to concentration, the adjustment would be easy to perform
- Fit the following model

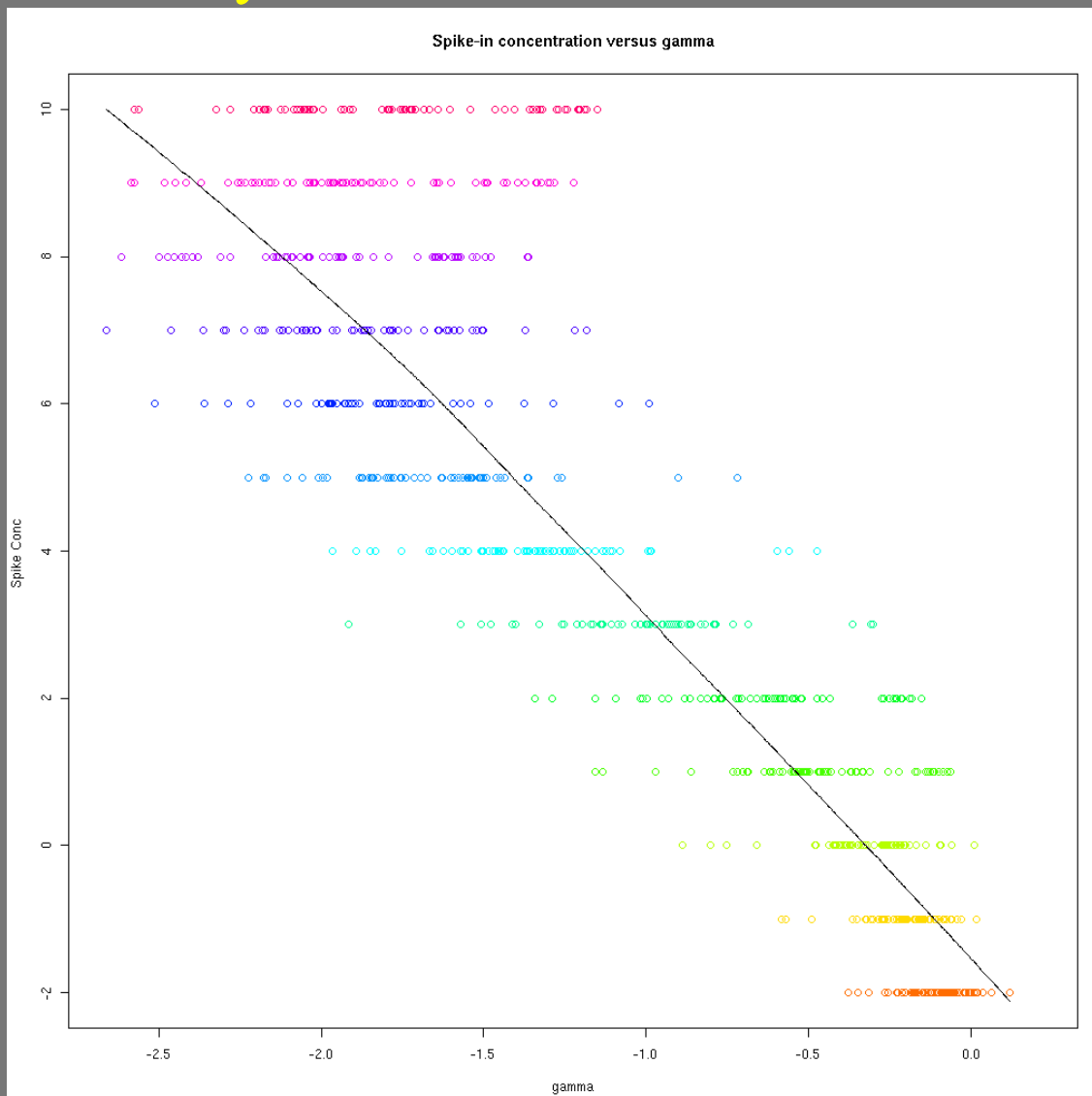
$$y_{1i}^{(k)} = \alpha_i^{(k)} + \varepsilon_i^{(k)}$$

$$y_{2i}^{(k)} = \alpha_i^{(k)} + \gamma^{(k)} + \varepsilon_i^{(k)}$$

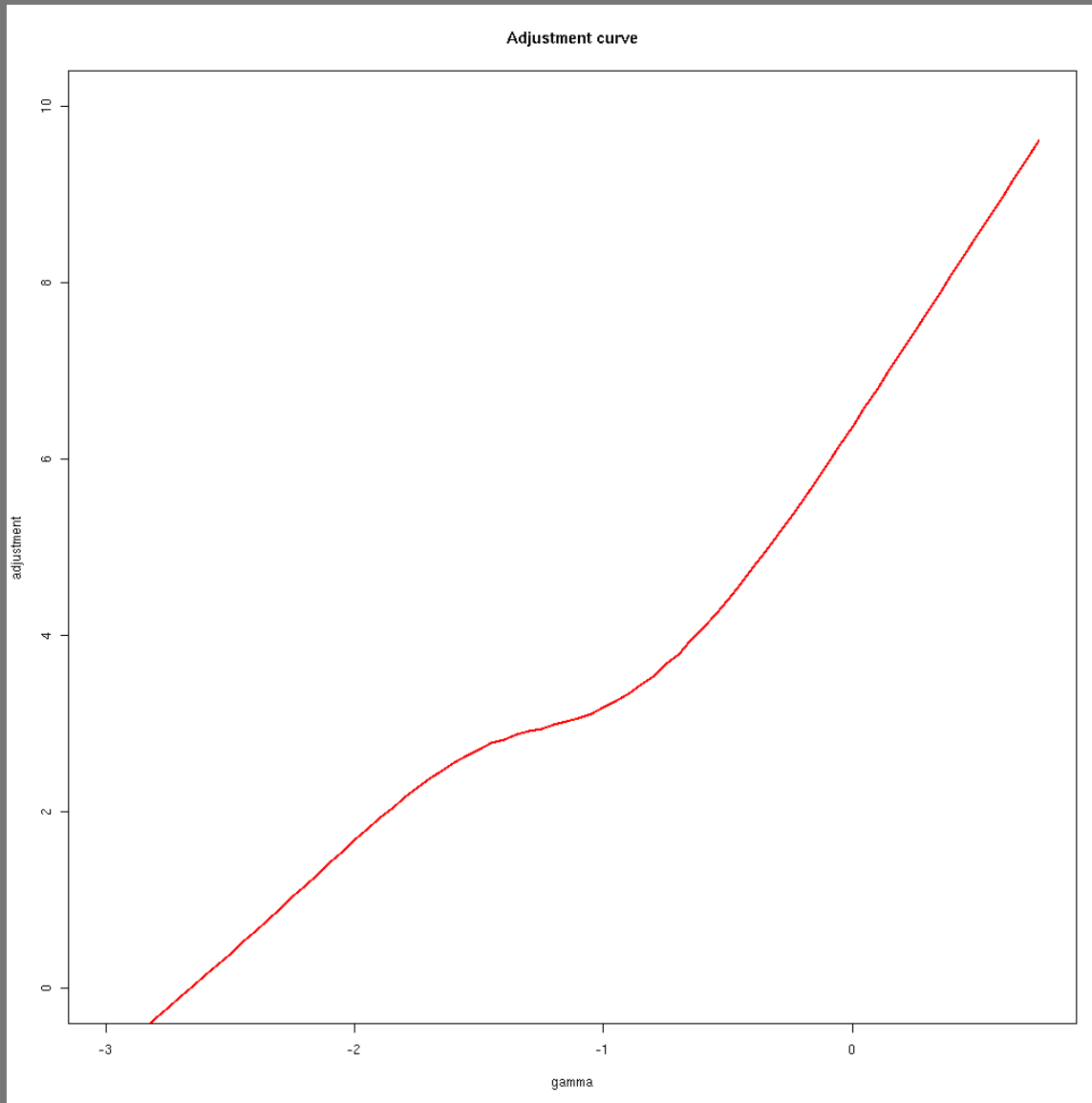
- Where  $y_{1i}^{(k)} = \log_2 PM_i^{(k)}$   
 $y_{2i}^{(k)} = \log_2 MM_i^{(k)}$



# Establishing a Relationship Between $\gamma$ and Concentration



# The Two Curves Yield an Adjustment Curve



# Normalization

*“Non-biological factors can contribute to the variability of data ... In order to reliably compare data from multiple probe arrays, differences of non-biological origin must be minimized.”*

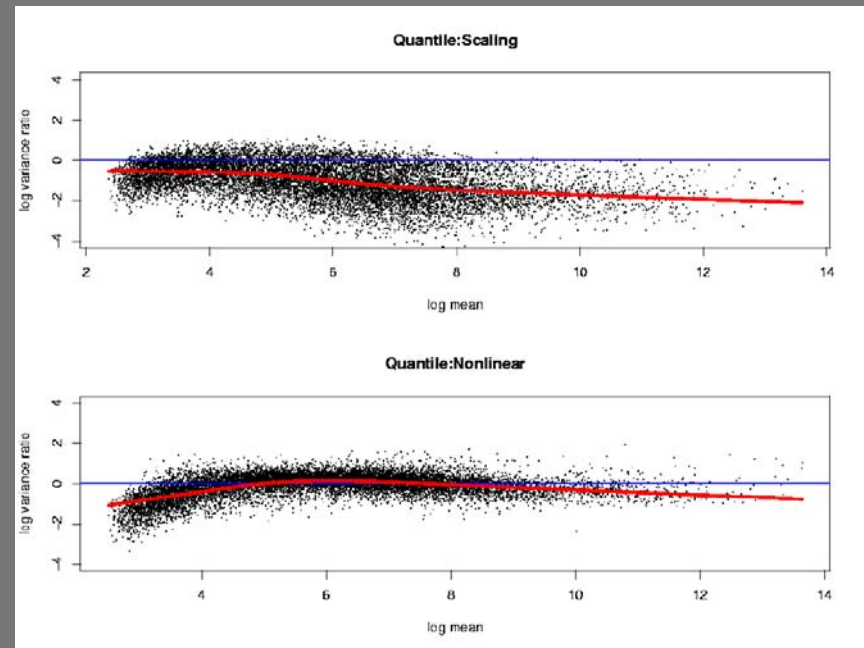
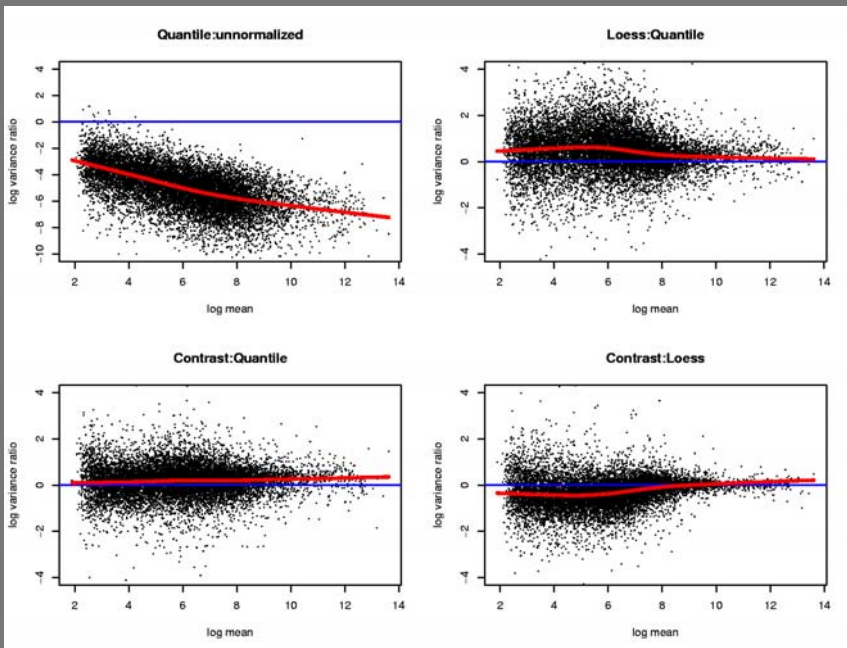
- Normalization is a process of reducing unwanted variation across chips. It may use information from multiple chips

# Normalization Methods

- Complete data (no reference chip, information from all arrays used)
  - Quantile normalization (Bolstad et al 2003)
  - Contrast (Åstrand)
  - Cyclic Loess
- Baseline (normalized using reference chip)
  - Scaling (Affymetrix)
  - Non linear (Li-Wong)
- Methods already compared in Bolstad et al (2003)

# Why Quantile Normalization?

- Quantile normalization found to perform acceptably in reducing variance without drastic bias effects
- Quantile normalization is fast



# Summarization

- Reduce the 11-20 probe intensities on each array to a single number for gene expression
- Main Approaches
  - Single chip
    - AvDiff (Affymetrix) – no longer recommended for use due to many flaws
    - Mas 5.0 (Affymetrix) – use a 1 step Tukey biweight to combine the probe intensities in log scale
  - Multiple Chip
    - MBEI (Li-Wong dChip) – a multiplicative model
    - RMA – a robust multi-chip linear model fit on the log scale

# RMA Model

- To each probeset (k), with i being number of probes and j being number of chips, fit the model:

$$y_{ij}^{(k)} = \alpha_i^{(k)} + \beta_j^{(k)} + \varepsilon_{ij}^{(k)}$$

where  $\alpha_i^{(k)}$  is a probe effect and  $\beta_j^{(k)}$  is the log gene expression.  $y_{ij}^{(k)}$  is the log2 background adjusted and normalized PM intensity

- Different ways to fit this model
  - Median polish – quick
  - Robust linear model – yields good quality diagnostic tools

# Overview

- Introduction
- Brief Technology Overview
- Preprocessing Steps
  - Background correction/Signal adjustment
  - Normalization
  - Summarization
- **Comparing the effect of different preprocessing methods on expression estimates**
- Software
- Future/Ongoing work

# Affymetrix Spike-in Data

- 59 chips. All but 1 of the rows are done as triplicates

	37777	684	1597	38734	39058	36311	36889	1024	36202	36085	40322	407	1091	1708
A	0	0.25	0.5	1	2	4	8	16	32	64	128	0	512	1024
B	0.25	0.5	1	2	4	8	16	32	64	128	256	0.25	1024	0
C	0.5	1	2	4	8	16	32	64	128	256	512	0.5	0	0.25
D	1	2	4	8	16	32	64	128	256	512	1024	1	0.25	0.5
E	2	4	8	16	32	64	128	256	512	1024	0	2	0.5	1
F	4	8	16	32	64	128	256	512	1024	0	0.25	4	1	2
G	8	16	32	64	128	256	512	1024	0	0.25	0.5	8	2	4
H	16	32	64	128	256	512	1024	0	0.25	0.5	1	16	4	8
I	32	64	128	256	512	1024	0	0.25	0.5	1	2	32	8	16
J	64	128	256	512	1024	0	0.25	0.5	1	2	4	64	16	32
K	128	256	512	1024	0	0.25	0.5	1	2	4	8	128	32	64
L	256	512	1024	0	0.25	0.5	1	2	4	8	16	256	64	128
M	512	1024	0	0.25	0.5	1	2	4	8	16	32	512	128	256
N	512	1024	0	0.25	0.5	1	2	4	8	16	32	512	128	256
O	512	1024	0	0.25	0.5	1	2	4	8	16	32	512	128	256
P	512	1024	0	0.25	0.5	1	2	4	8	16	32	512	128	256
Q	1024	0	0.25	0.5	1	2	4	8	16	32	64	1024	256	512
R	1024	0	0.25	0.5	1	2	4	8	16	32	64	1024	256	512
S	1024	0	0.25	0.5	1	2	4	8	16	32	64	1024	256	512
T	1024	0	0.25	0.5	1	2	4	8	16	32	64	1024	256	512

# **We will focus on assessing the impact of background adjustment methods on expression values**

- Impact of normalization has been previously addressed in Bolstad et al (2003)
- We will compare the impact of different background methods on expression values by
  - Signal adjusting using the chosen method
  - Normalizing using quantile normalization
  - Summarization using RMA: median polish
- Then we will compare the results

# Background Methods to be Compared

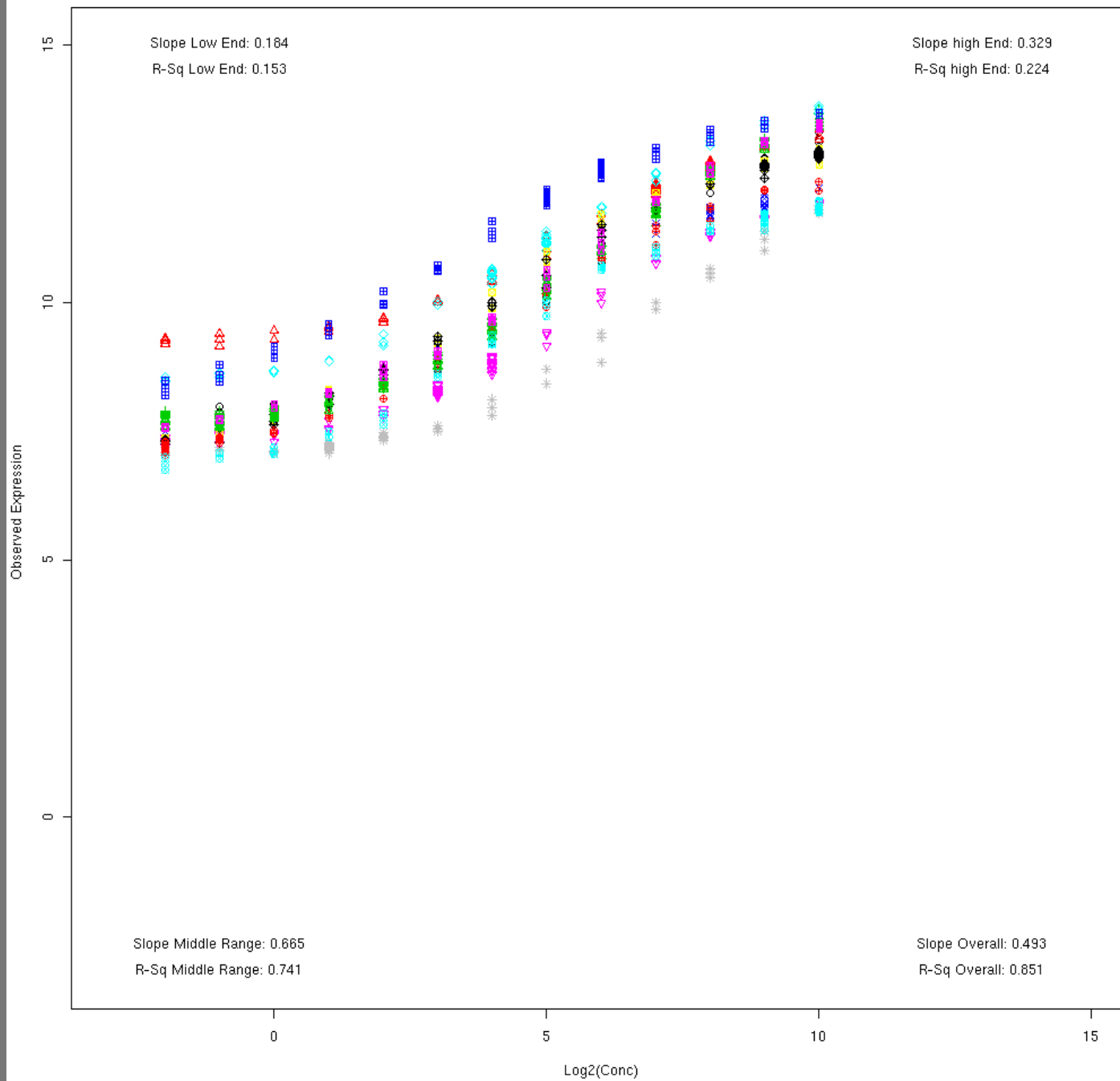
- None
- MAS 5.0 location specific background
- Ideal Mismatch
- MAS 5.0 and Ideal Mismatch
- RMA convolution model
- Using standard curve based on spike-in information to adjust signal

# Computing Relative Expression

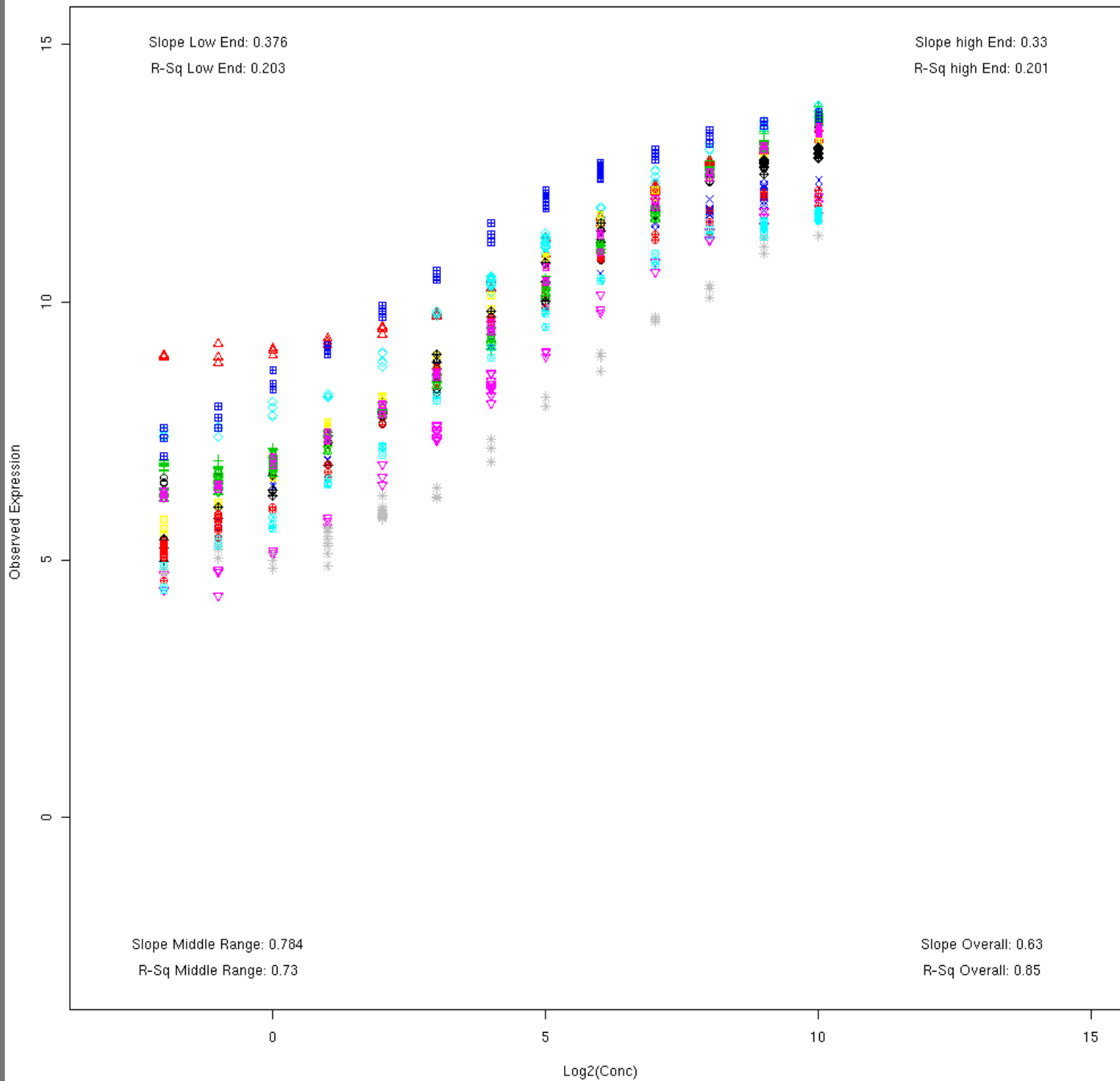
- We will average in log scale across spike-in concentration replicates
- If  $E_{i,j}$  is expression of probeset  $i$  in group  $j$ , then expression difference between group 1 and 2 is
  - $M_i = E_{i,1} - E_{i,2}$
- There are 14 dilution groups so there are  $14 \cdot 13 / 2 = 91$  different comparisons for each probeset

# **Observed expression versus spike-in concentration**

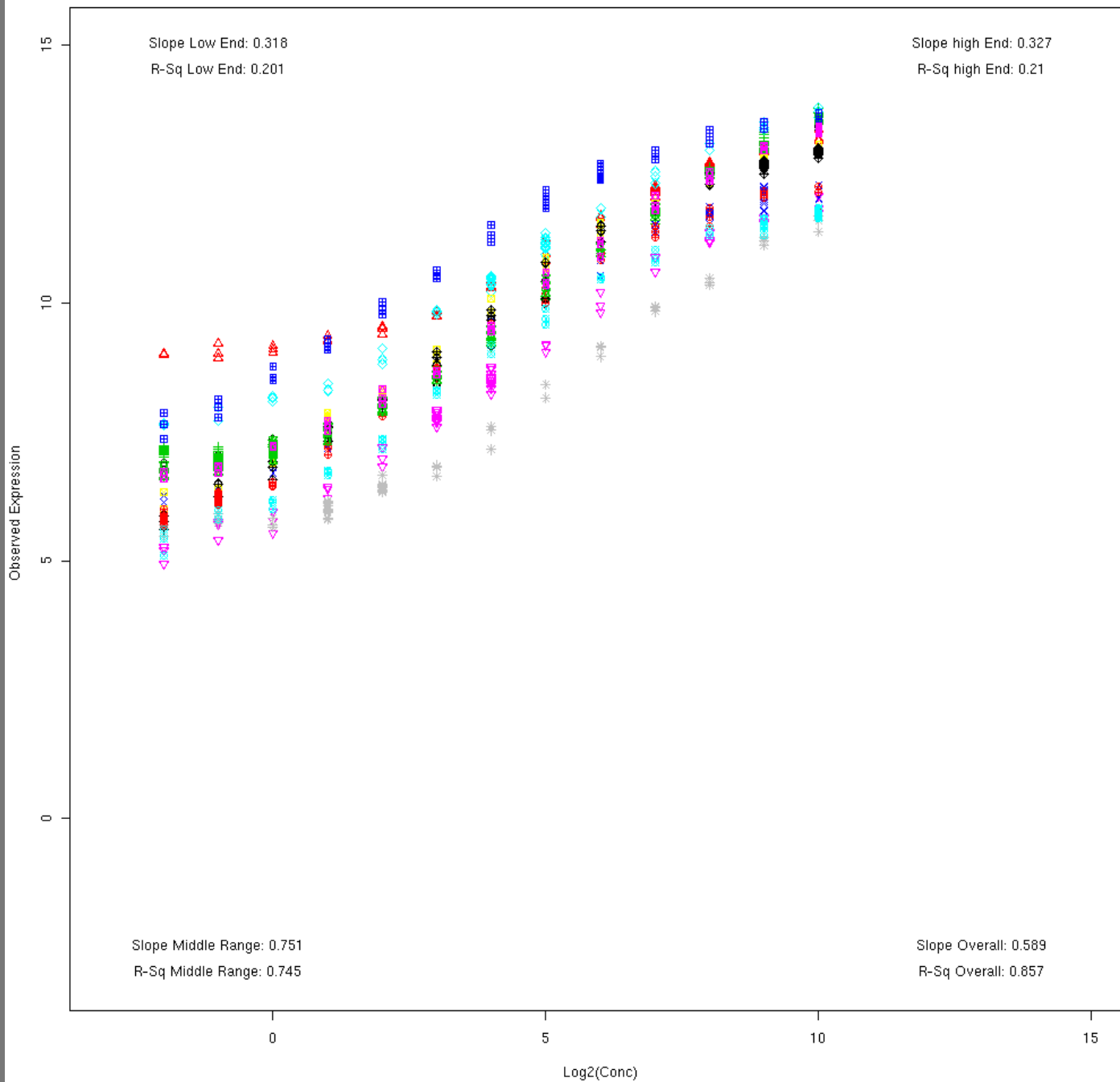
# No Background



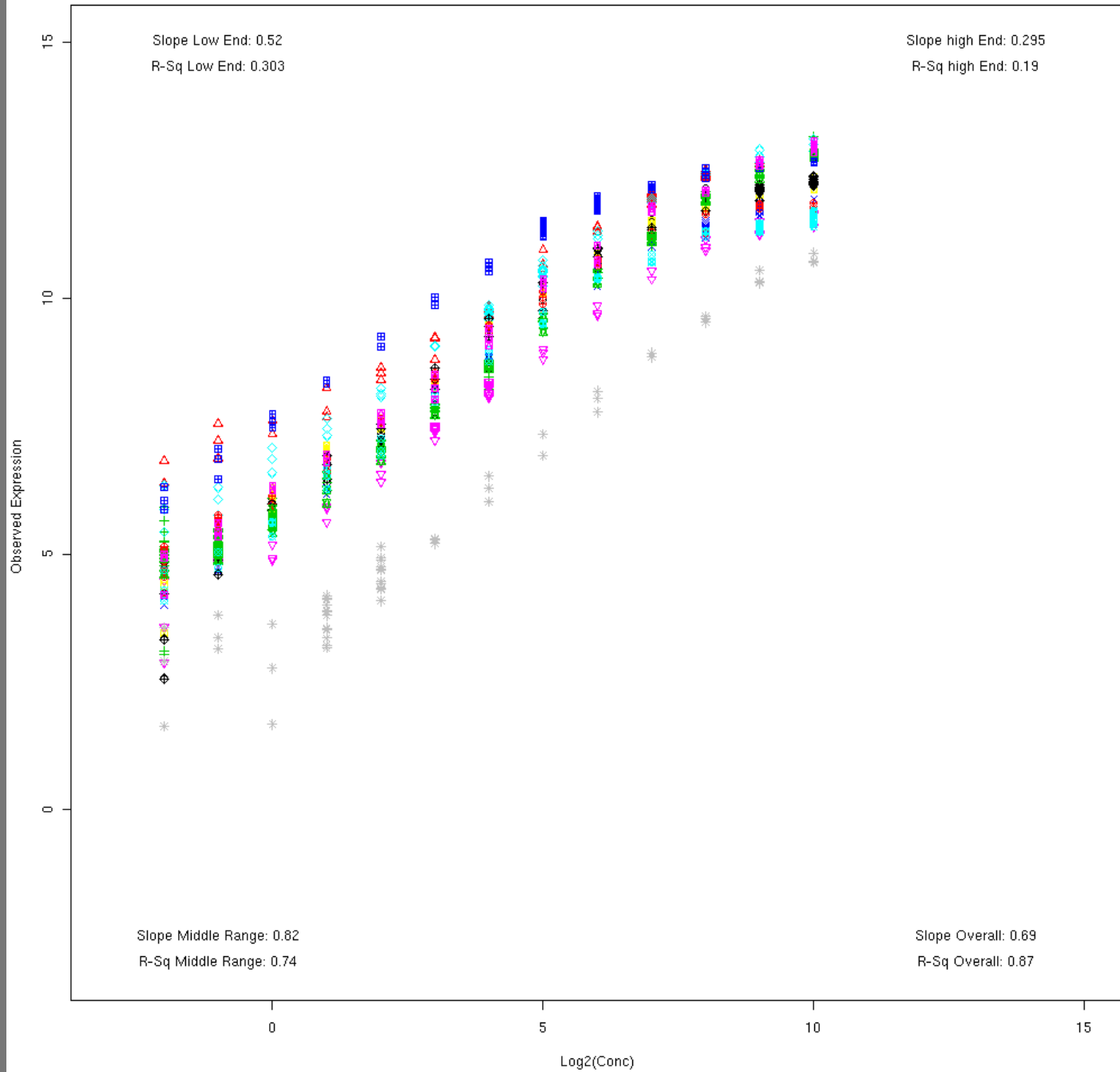
# Convolution background



# MAS 5.0 background

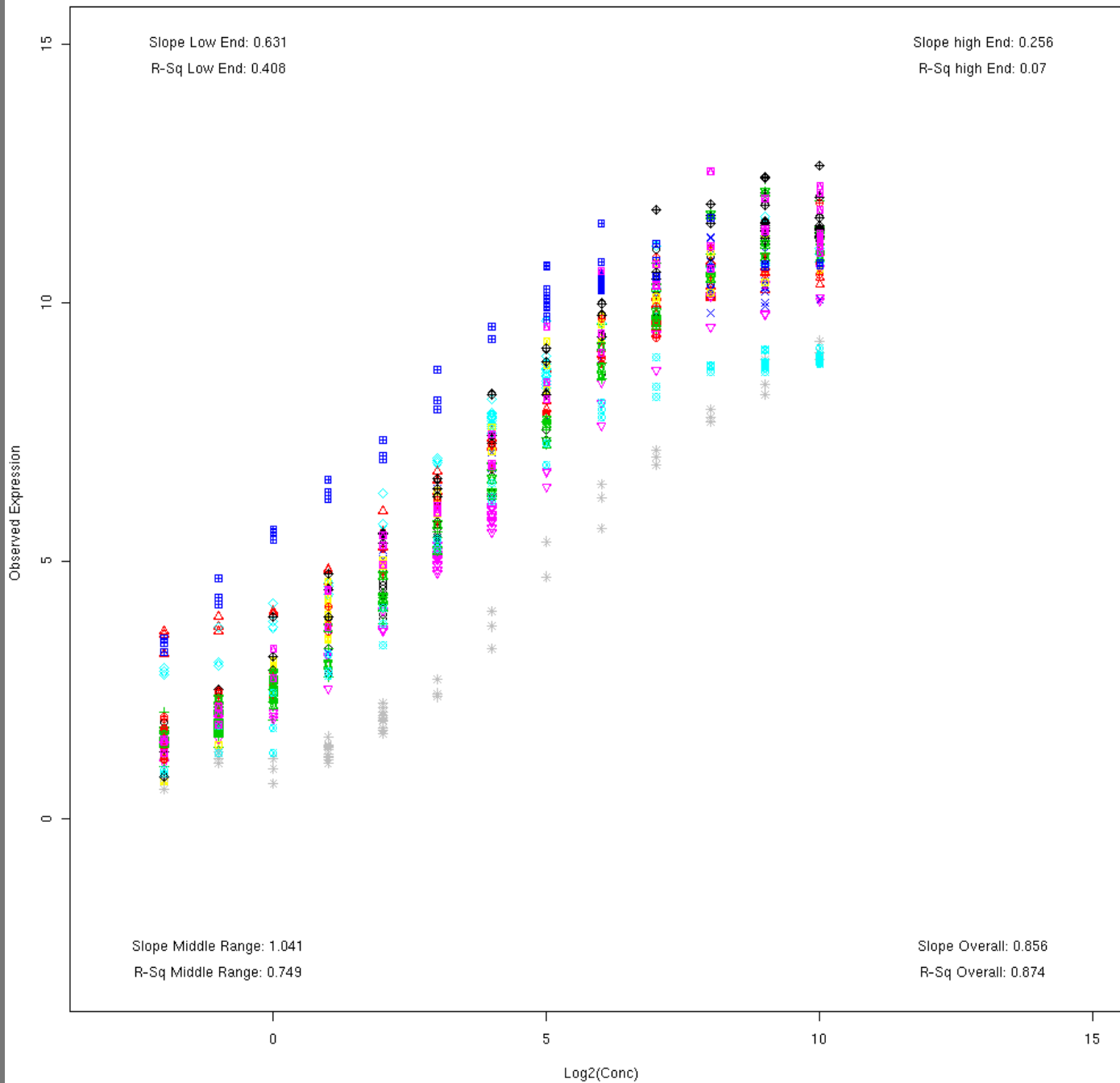


# IdealIMM



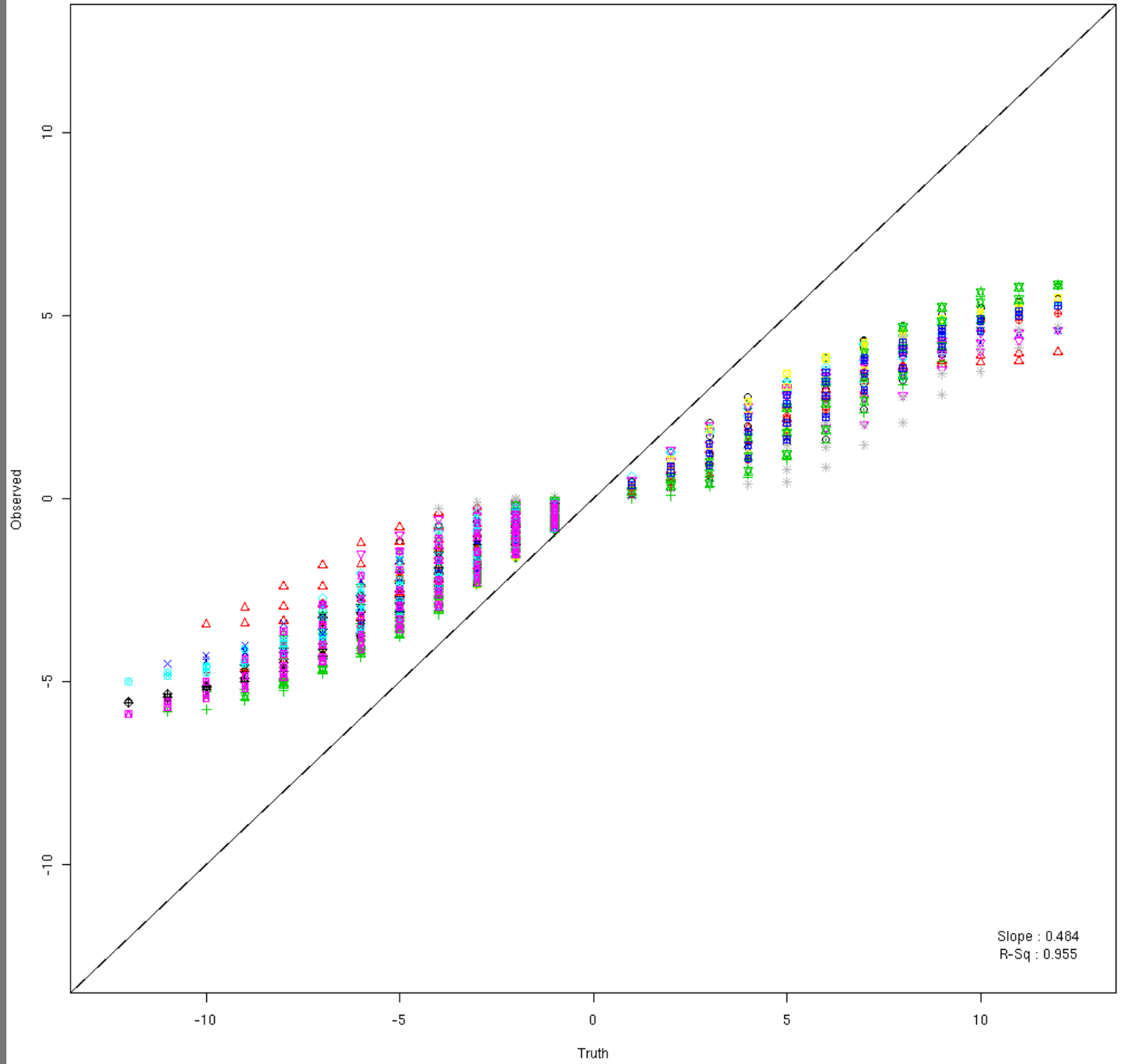


# Standard Curve Adjustment

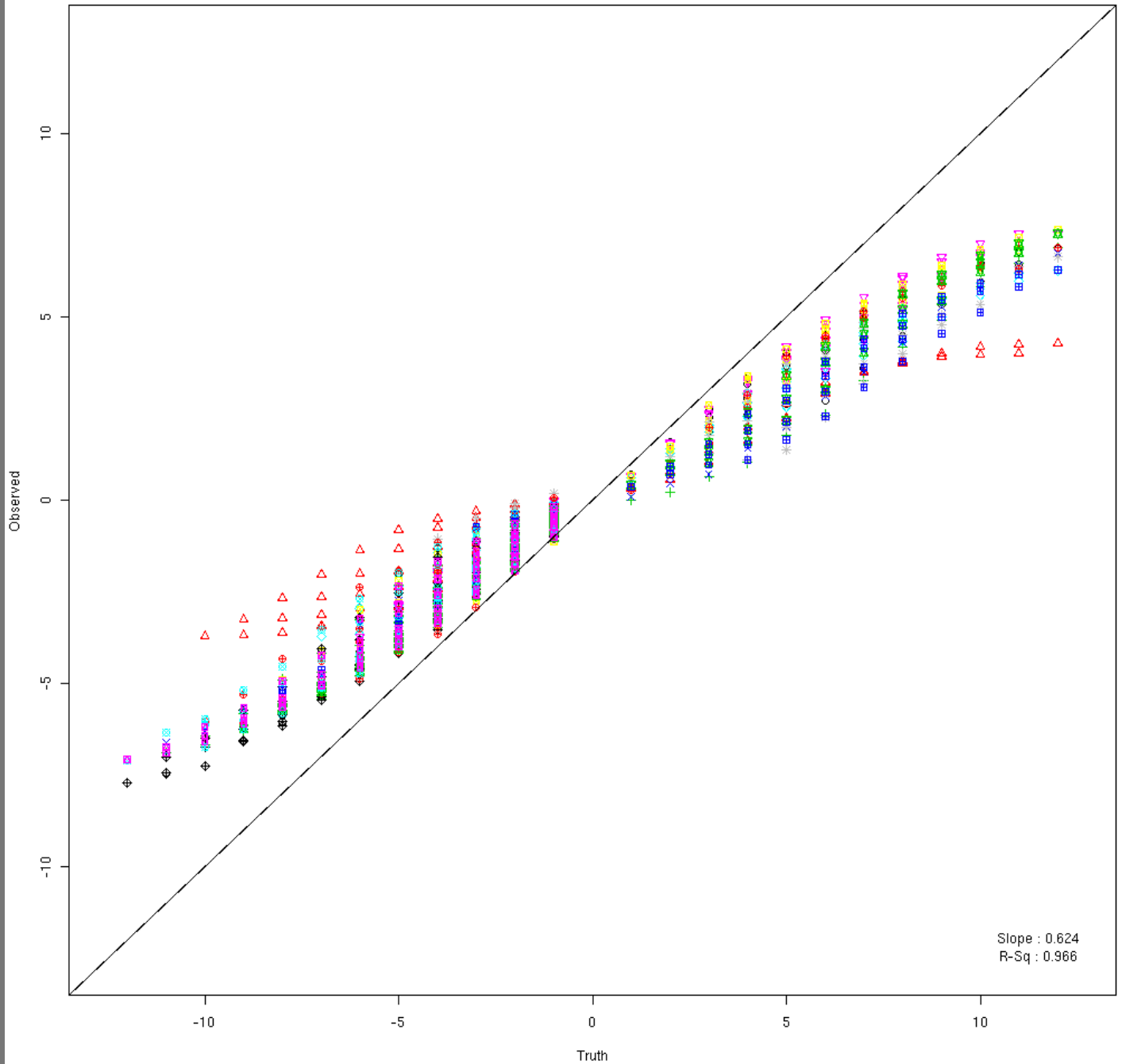


**Observed fold change  
versus expected fold  
change**

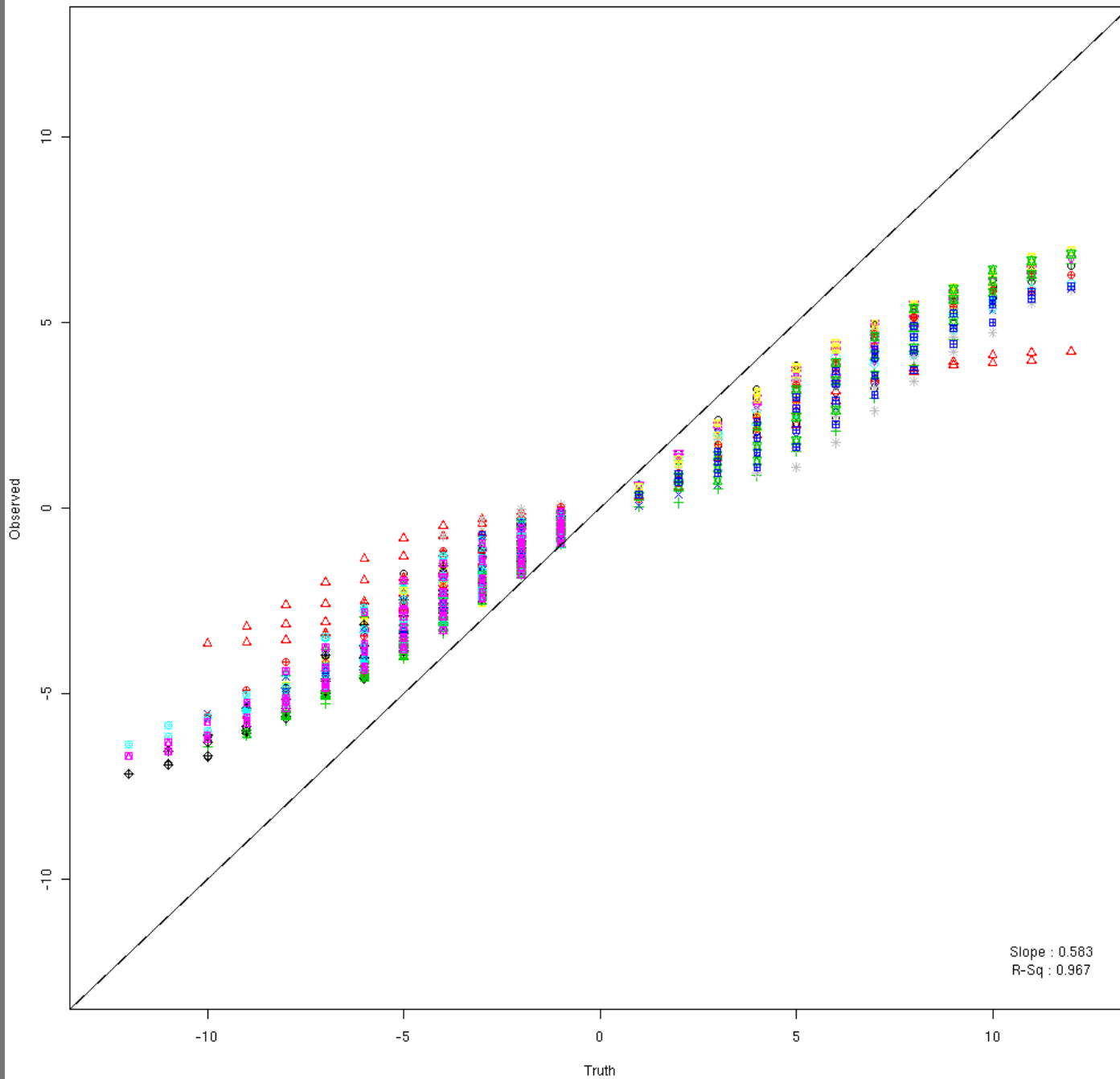
# No background



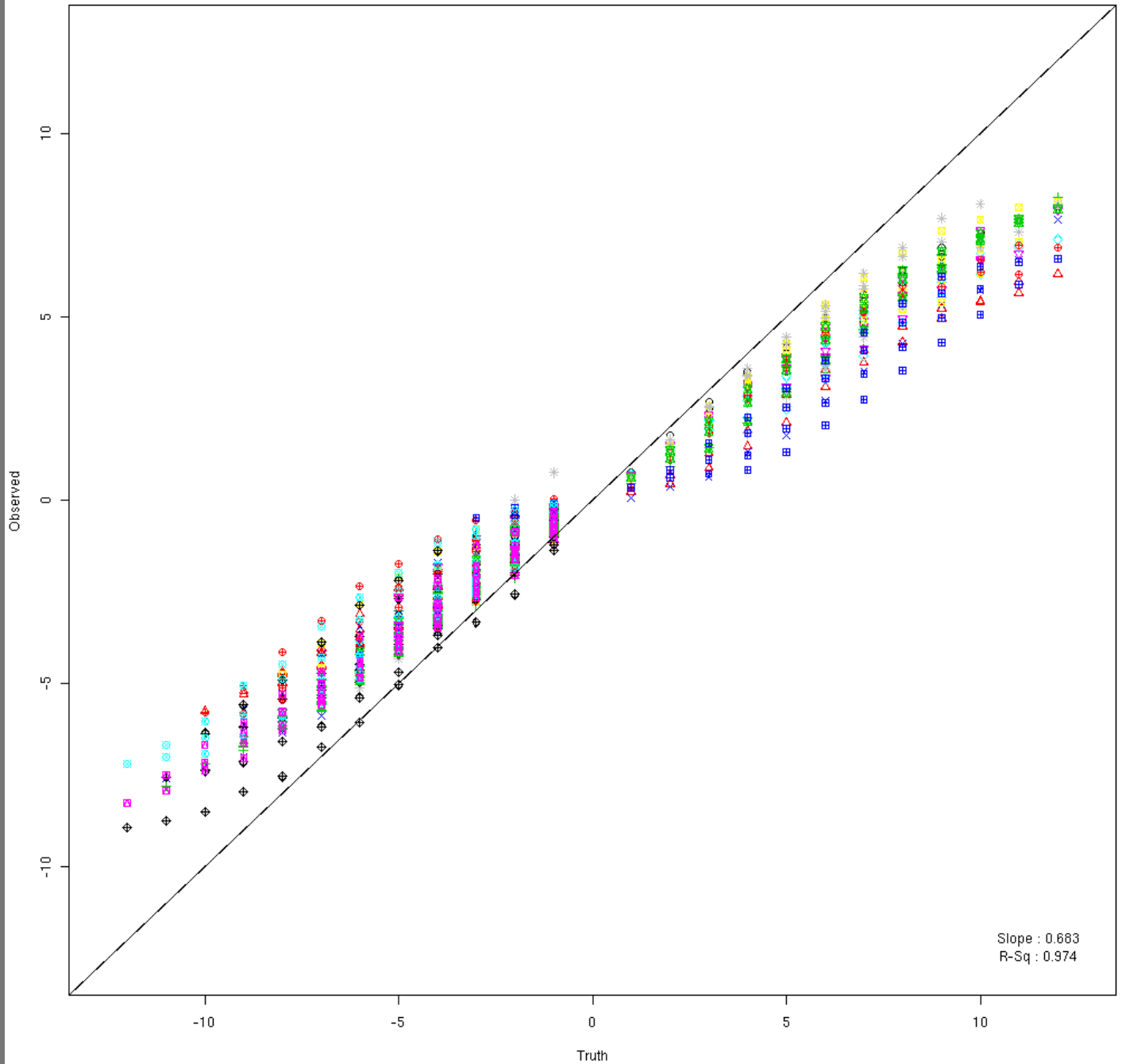
# Convolution background



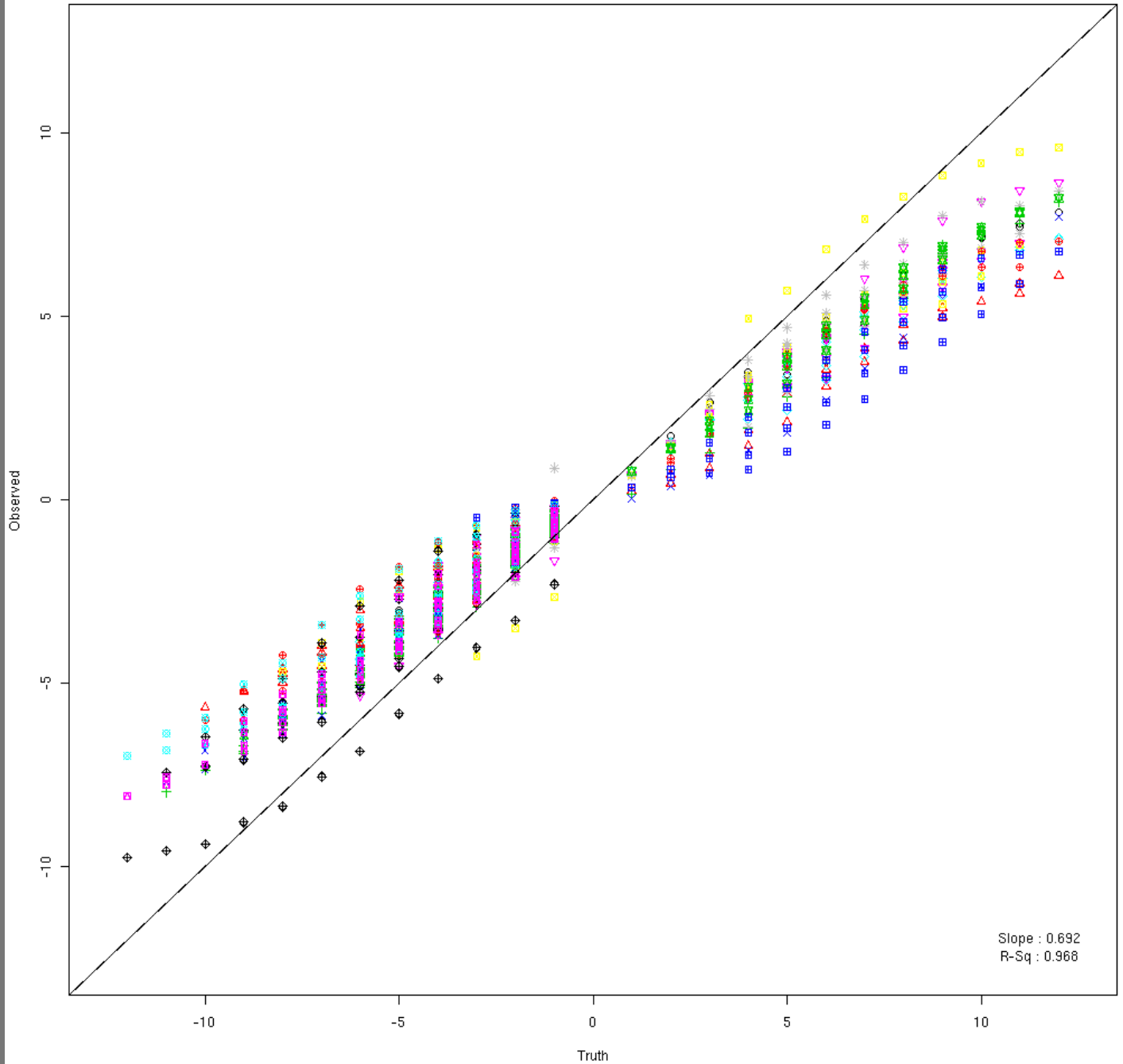
MAS 5.0 background



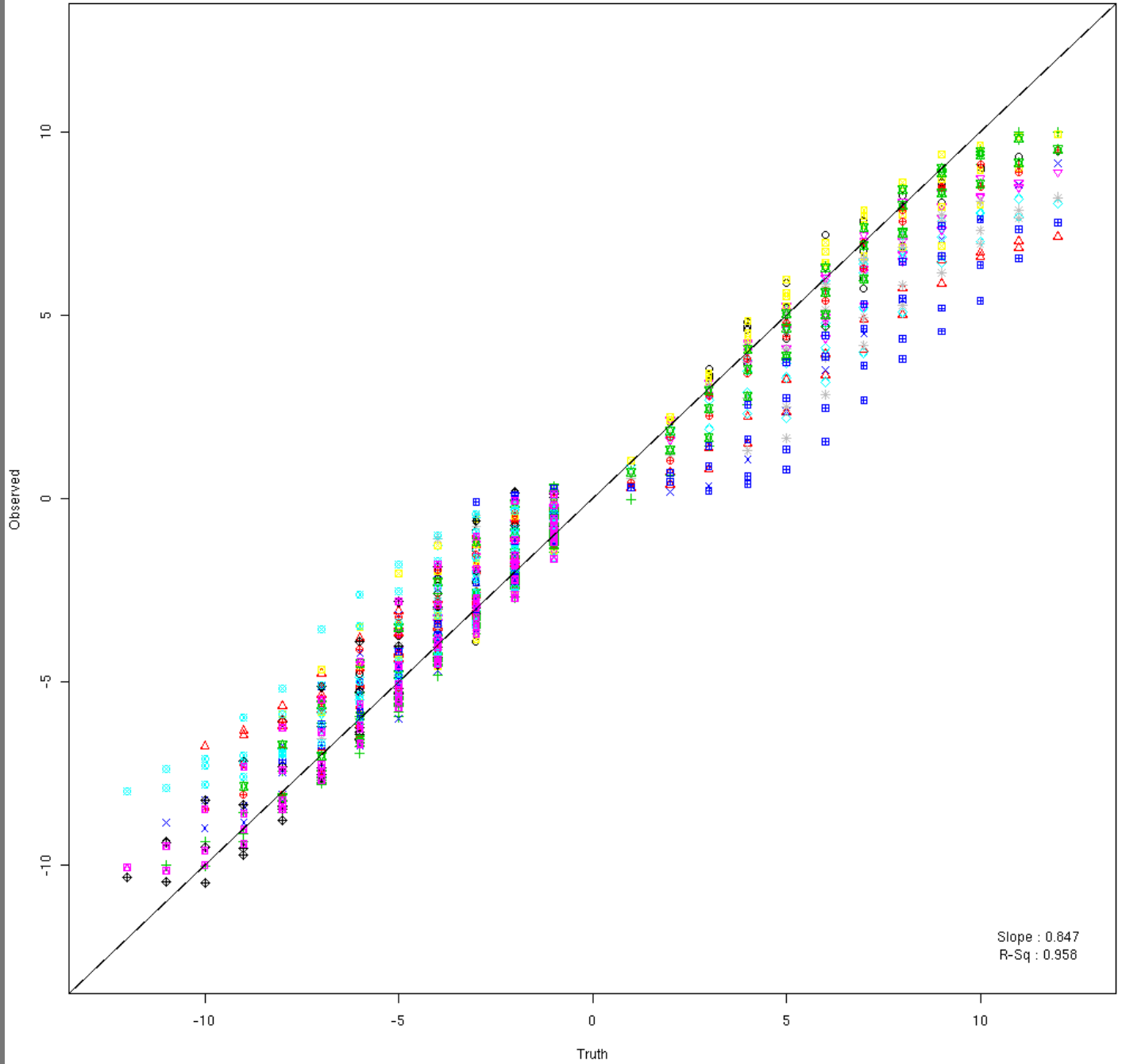
# Ideal Mismatch



MAS 5.0 bg then IdeallMM

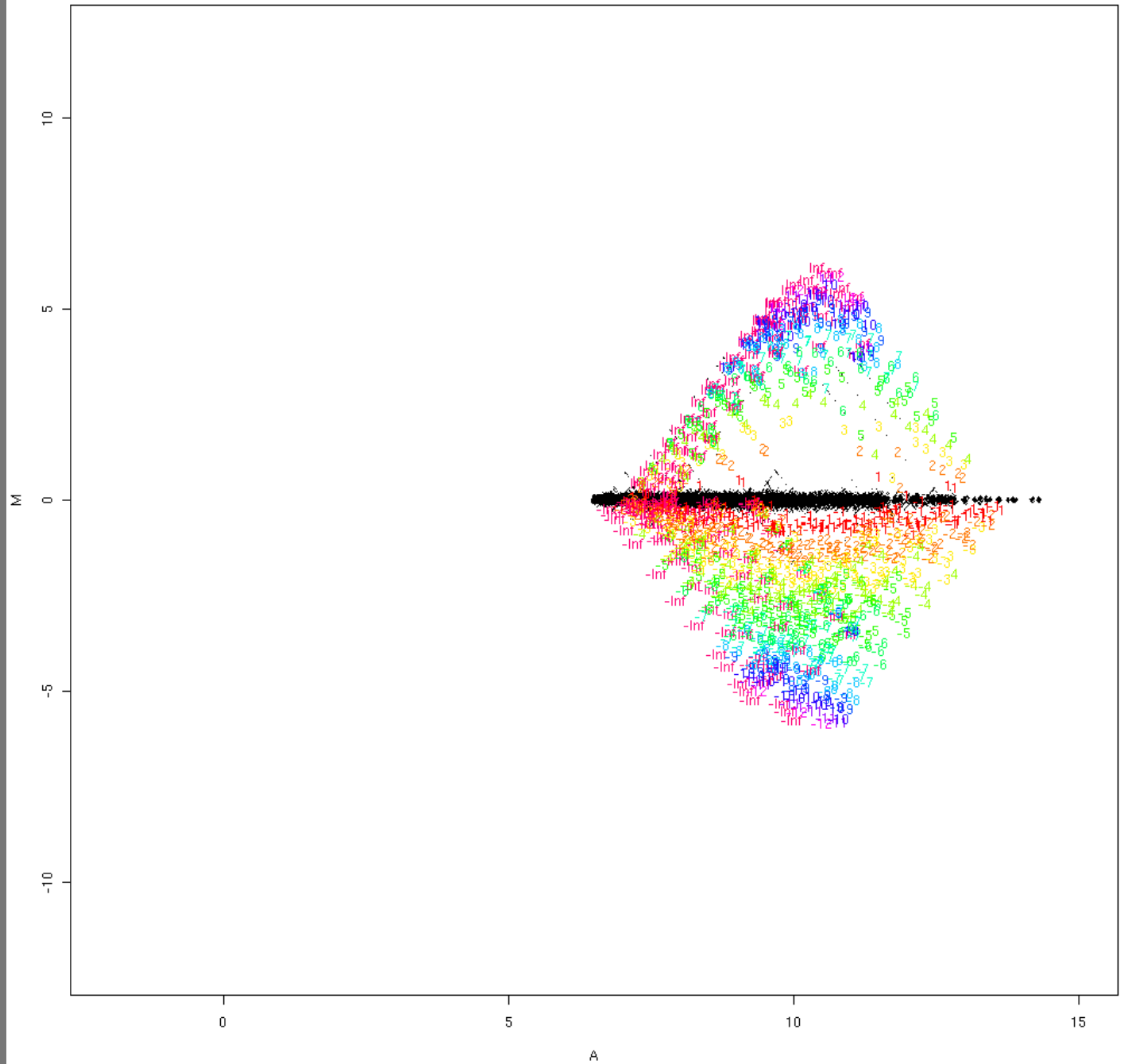


# Standard Curve Adjustment

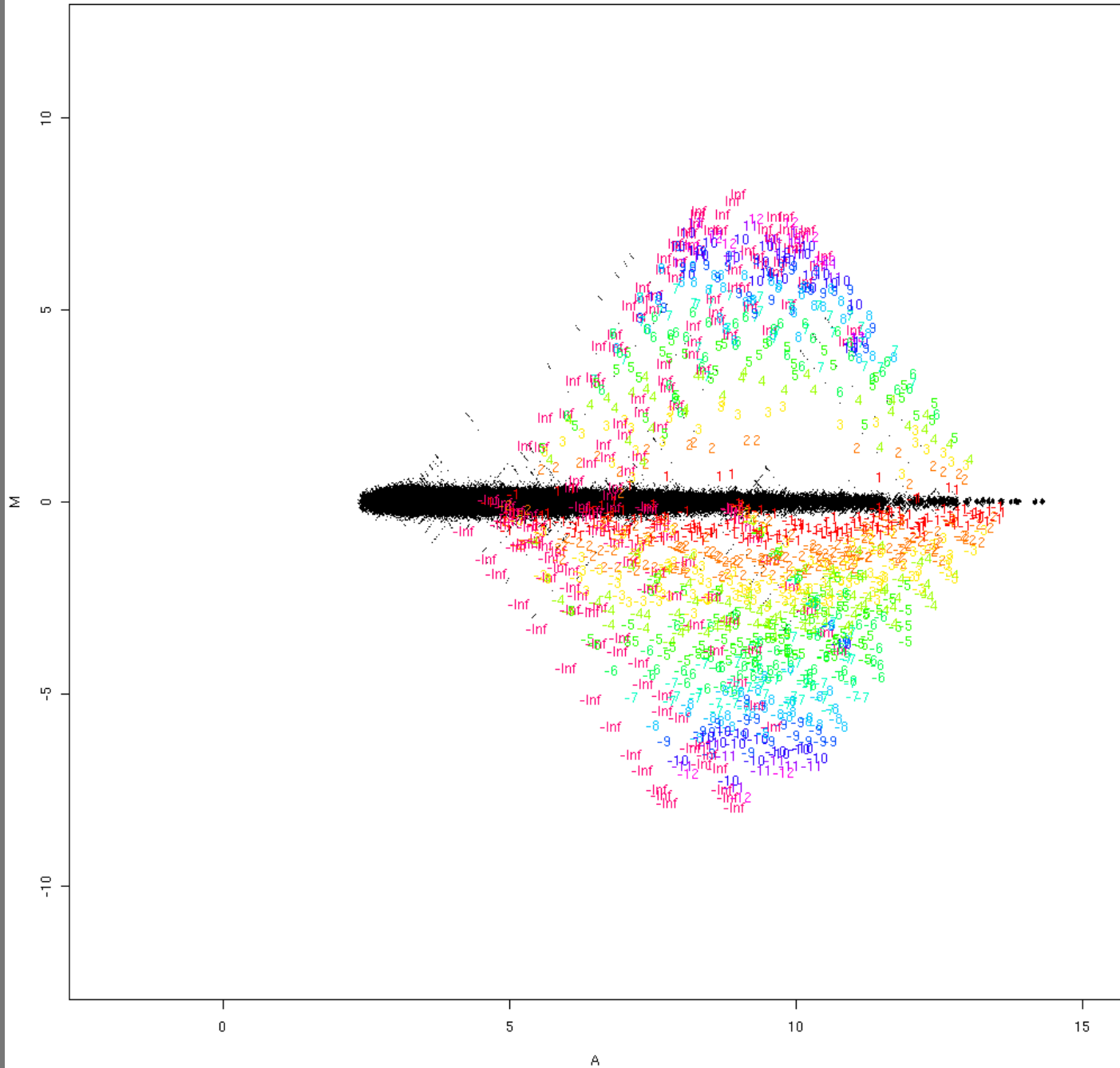


# Composite M vs A Plots

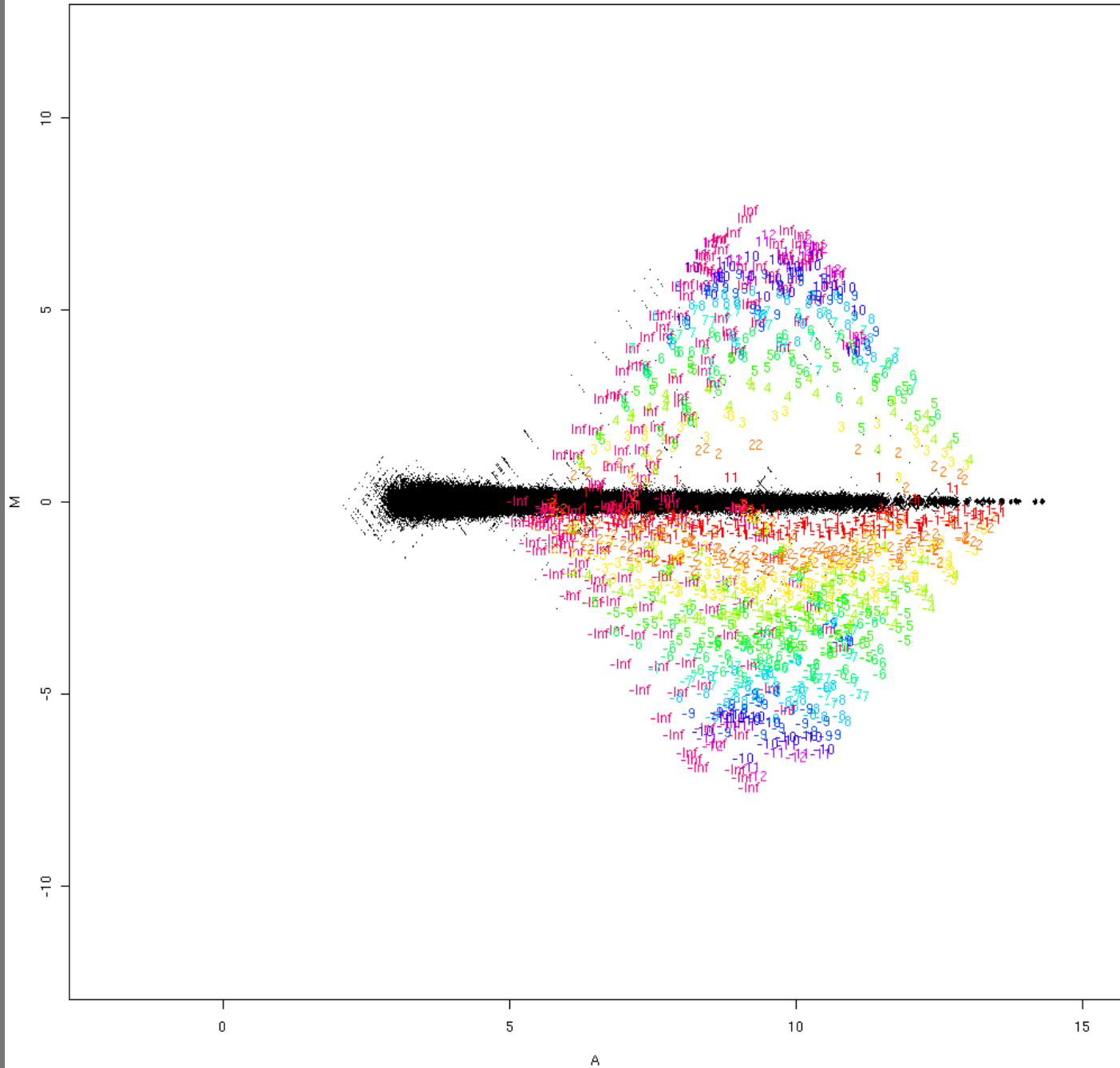
No background



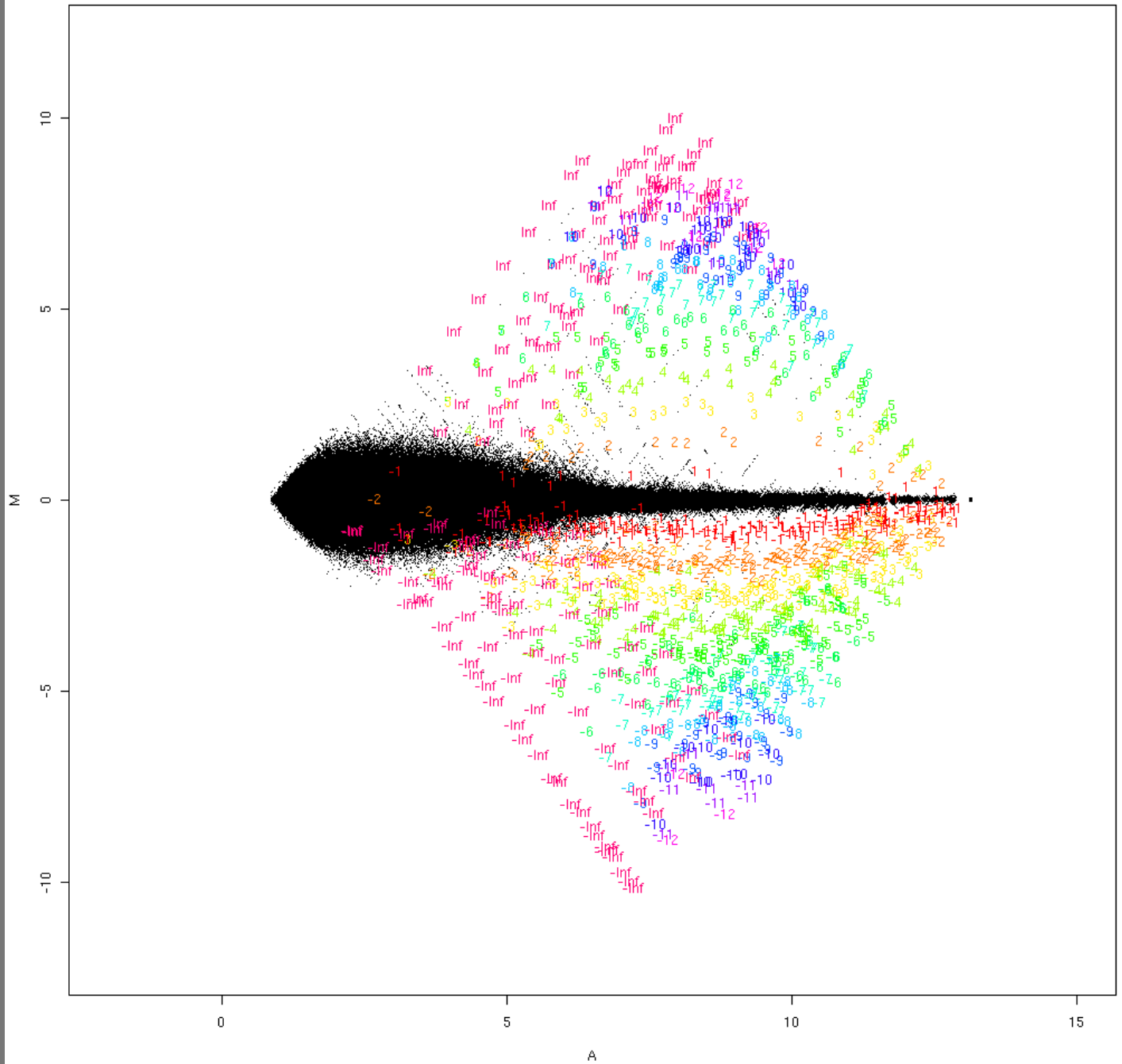
# Convolution model



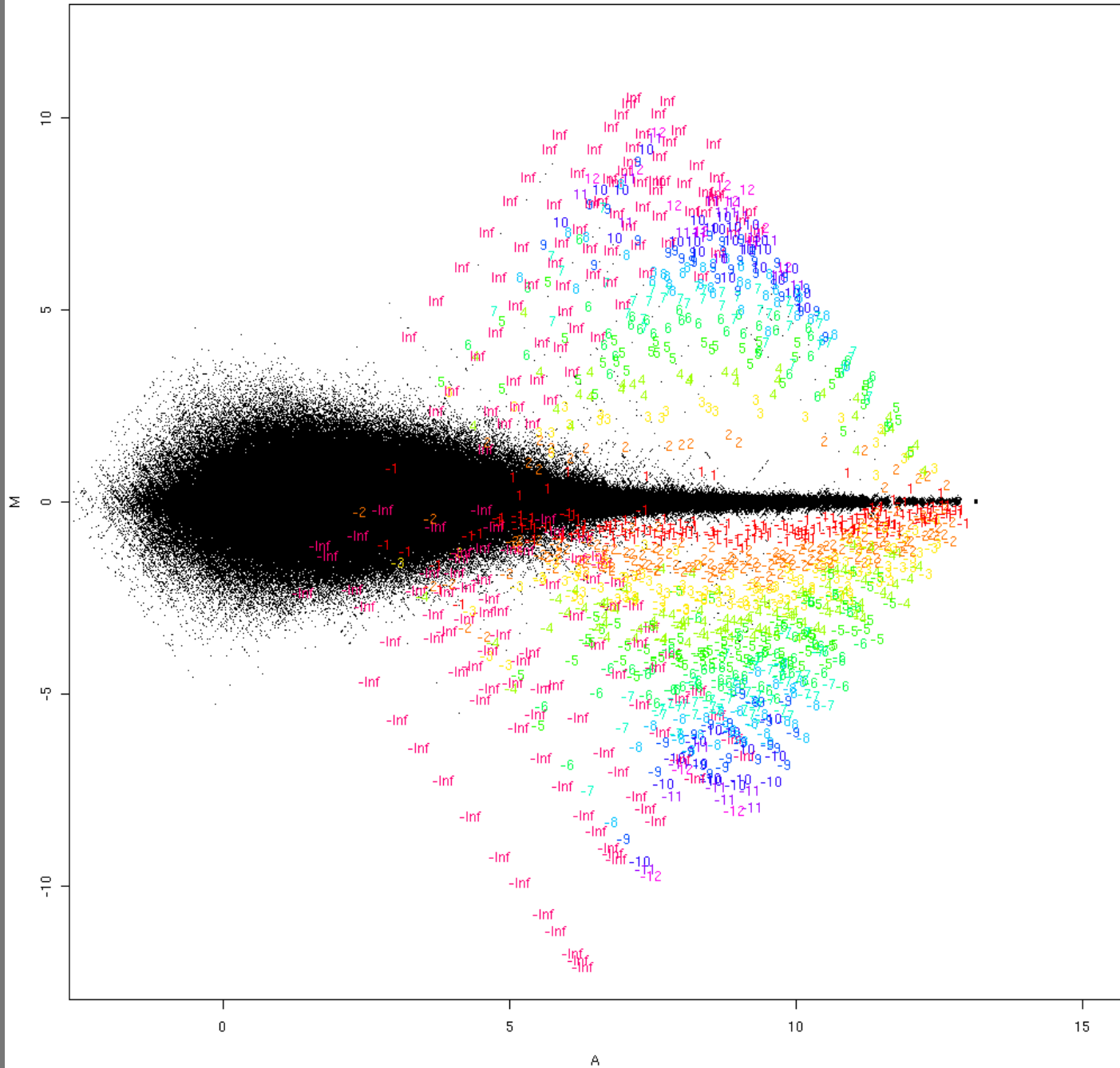
# Mas 5.0 Background



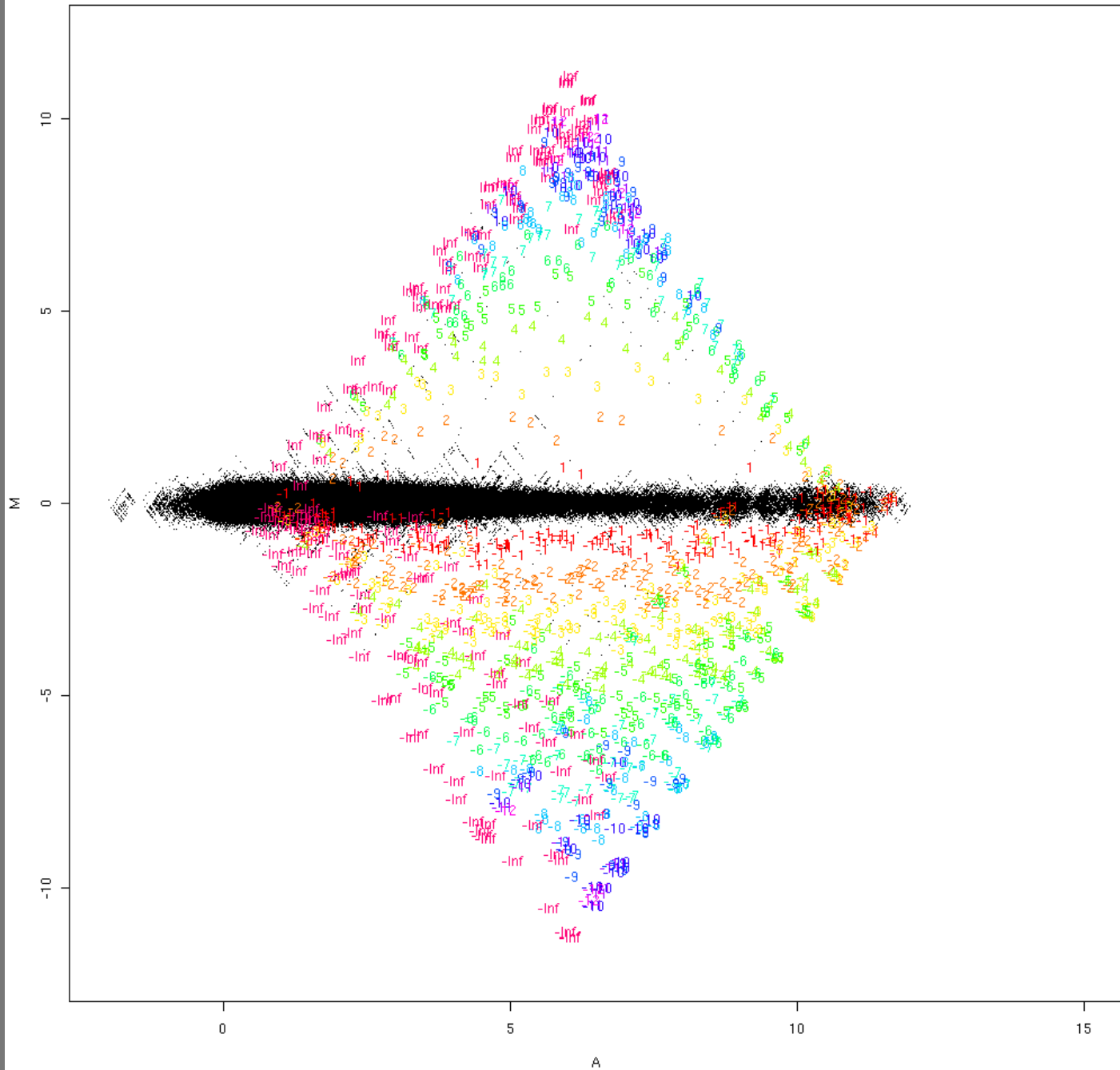
# Ideal Mismatch



Mas 5.0 background then Ideal Mismatch

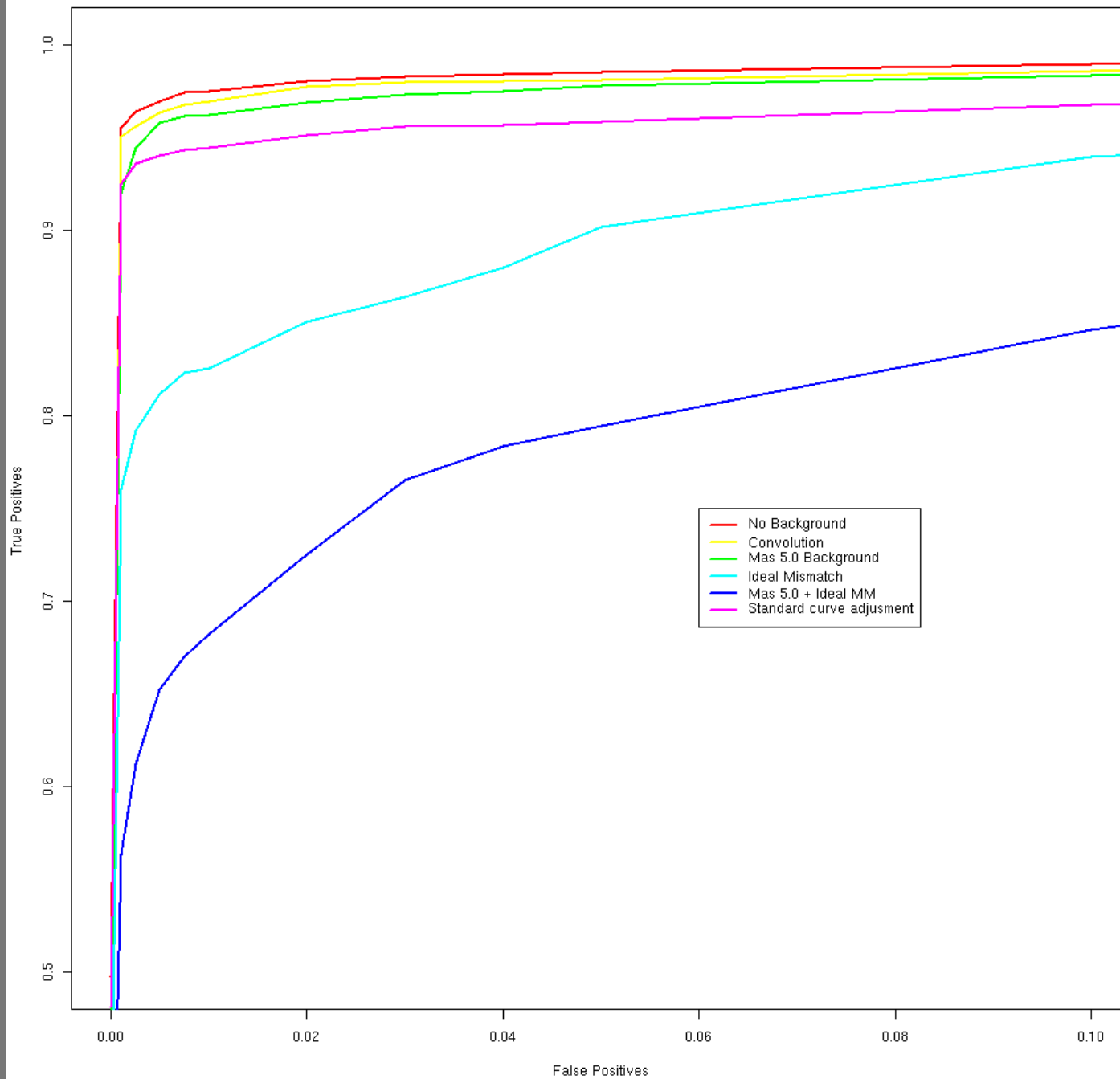


# Standard Curve Adjustment



# ROC Curves

ROC curves based upon Fold Change



# The Background Methods Have Different Tradeoffs

		Detect Differential Genes	
		Poor	Good
Change estimate Accurately Fold	Poor		<ul style="list-style-type: none"><li>•No Background</li><li>•Convolution</li><li>•MAS 5.0</li></ul>
	Good	<ul style="list-style-type: none"><li>•MAS 5 + IdealMM</li><li>•Ideal-Mismatch</li></ul>	<ul style="list-style-type: none"><li>•Standard Curve Adjustment</li></ul>

# These results are not limited to just this dataset

- Similar results have been observed with other spike-in experiments: Genelogic's spike-in datasets
- Datasets where we have QRT-PCR measurements for certain genes and array data can also be used in this sort of comparison

# Overview

- Introduction
- Brief Technology Overview
- Preprocessing Steps
  - Background correction/Signal adjustment
  - Normalization
  - Summarization
- Comparing the effect of different preprocessing methods on expression estimates
- **Software**
- Future/Ongoing work

# Software

- R packages
  - Contributions to *affy* which is part of Bioconductor <http://www.bioconductor.org>

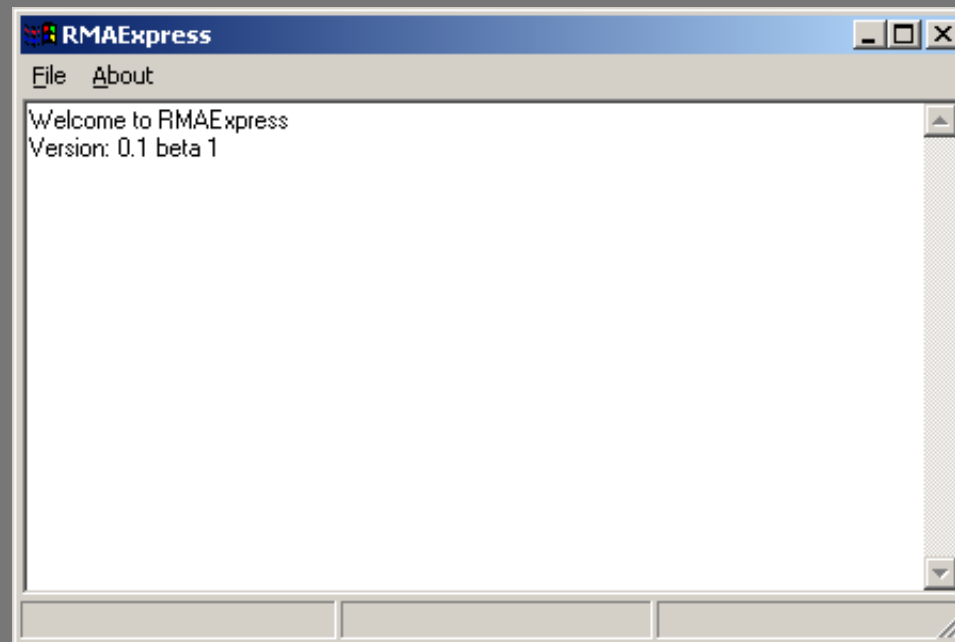
*rma()*, *normalize.quantiles()*,  
*bg.correct.rma()*, ...

- *AffyExtensions* A package for fitting more general probe level models  
<http://www.stat.berkeley.edu/users/bolstad/AffyExtensions/AffyExtensions.html>

*fitPLM()*, *threestep()*, ...

# Software

- Other
  - *RMAExpress*: a simple standalone GUI program for Windows for computing the RMA expression measure



# Overview

- Introduction
- Brief Technology Overview
- Preprocessing Steps
  - Background correction/Signal adjustment
  - Normalization
  - Summarization
- Comparing the effect of different preprocessing methods on expression estimates
- Software
- **Future/Ongoing work**

# Future/Ongoing work

- Extending the standard curve method to data without spike-ins
- Using robust probe-level models to develop moderated t-statistics

# Thesis Outline

- Introduction
- Background/Signal Adjustment
- Normalization
- Summarization (with a focus on the development of t-statistics from probe-level models)
- 1-2 chapters: each an application to real data

# References

1. Bolstad BM, Irizarry RA, Astrand M and Speed TP . (2003), A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003 Jan 22;19(2):185-193.
2. Irizarry R, Bolstad B, Collin F, Cope L, Hobbs B and Speed T (2003) Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research*, 2003, Vol. 31, No. 4 e15
3. Irizarry, R. et. al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, in press.
4. Affymetrix (2002) Statistical Algorithms Description Document [http://www.affymetrix.com/support/technical/whitepapers/sadd\\_whitepaper.pdf](http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf)
5. Bioconductor <http://www.bioconductor.org>
6. Affymetrix Spike-in experiment [http://www.affymetrix.com/analysis/download\\_center2.affx](http://www.affymetrix.com/analysis/download_center2.affx)
7. Affymetrix Website <http://www.affymetrix.com>

**ADDITIONAL MATERIAL**

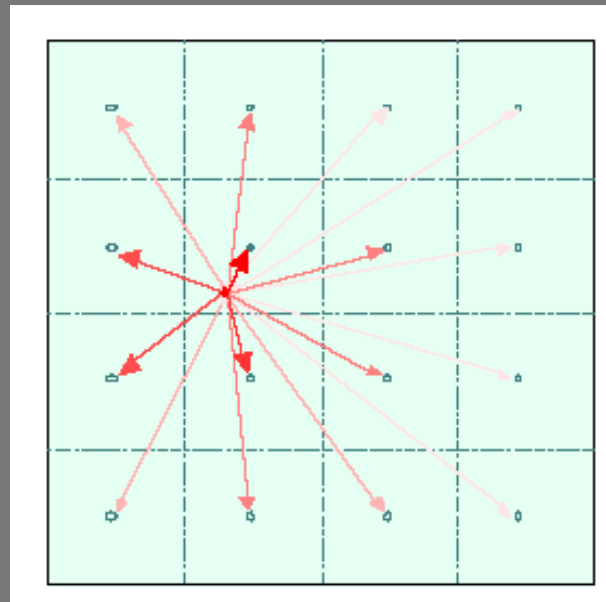
**Additional Material:  
Background/Signal  
Adjustment**

# MAS 5.0 Background

- Use the background correction method as described in Affymetrix “Statistical Algorithm Description Document”
- Break chip into  $k$  ( $k=1..16$ ) rectangular regions
  - lowest 2% is chosen as background for that region  $B_k$
  - Standard deviation for lowest 2% is chosen as noise for that zone  $N_k$

# MAS 5.0 Background (cont)

- The background adjustment to be used for cell at  $(x,y)$  is weighted average of the  $B_k$ , where the weights depend on the distance between  $(x,y)$  and the centroids of the regions:  $b(x,y)$



# MAS 5.0 Background (cont)

- A noise adjustment is computed in a similar way using  $N_k$  rather than  $B_k$  :  $n(x,y)$
- The background adjusted intensity is given by
  - $A(x,y) = \max(I(x,y) - b(x,y), \text{NoiseFrac} * n(x,y))$ 
    - Where  $\text{NoiseFrac} = 0.5$

# MAS 5.0 Mismatch Correction

- The way Affymetrix makes use of MM in MAS5.0
- Define biweight specific background (SB) for probe pair  $j$  in probeset  $i$  as
  - $SB_i = T_{bi}(\log_2(PM_{i,j}) - \log_2(MM_{i,j})):j = 1, \dots, n_i)$
  - $T_{bi}$  is Tukey biweight function

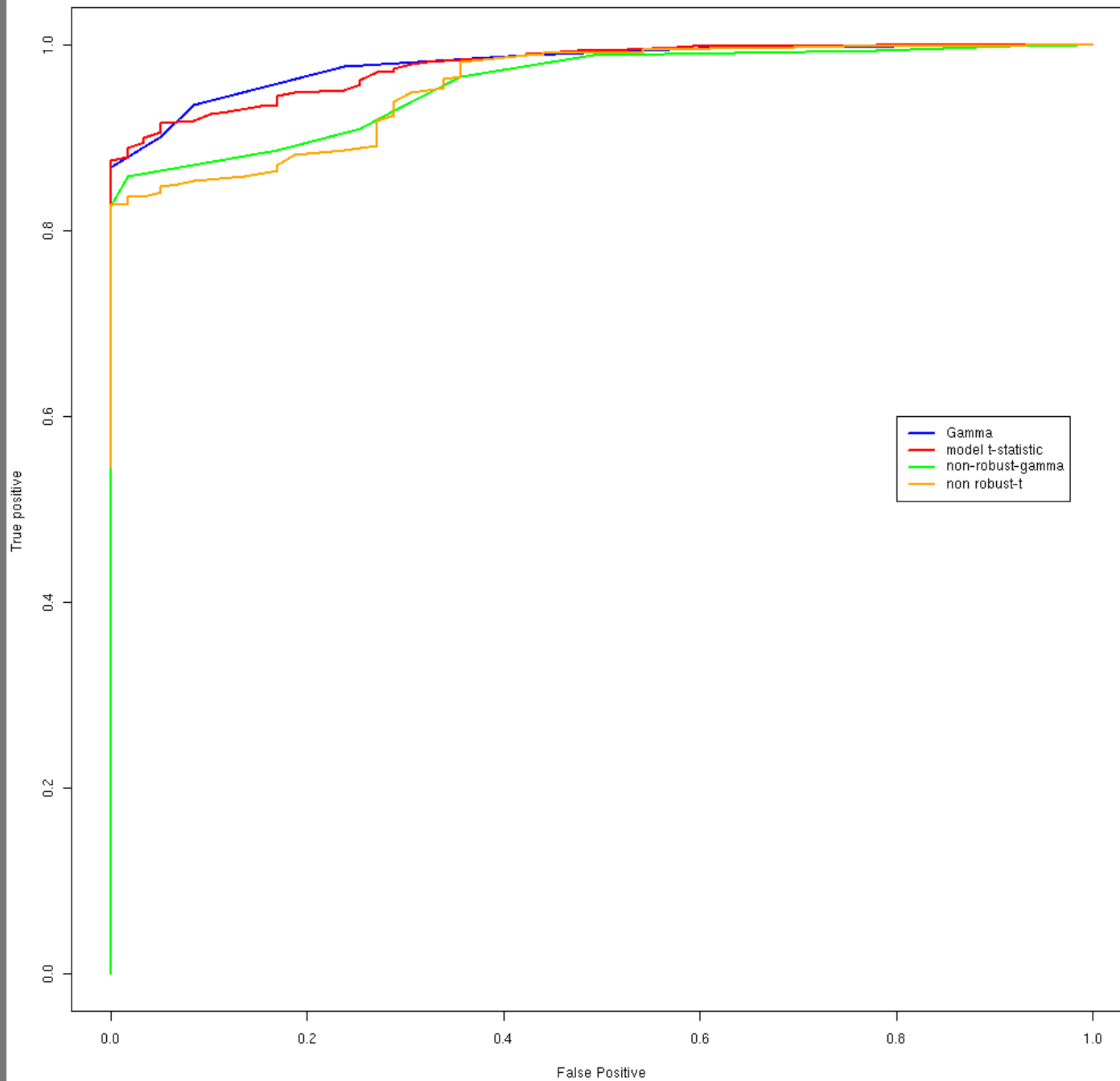
# MAS 5.0 Mismatch (cont)

- $IM_{i,j}$  is the ideal mismatch
- If  $MM_{i,j} < PM_{i,j}$ 
  - $IM_{i,j} = MM_{i,j}$
- If  $MM_{i,j} \geq PM_{i,j}$  and  $SB_i > \text{contrasttau}$ 
  - $IM_{i,j} = PM_{i,j} / 2^{(SB_i)}$
- If  $MM_{i,j} \geq PM_{i,j}$  and  $SB_i \leq \text{contrasttau}$ 
  - $IM_{i,j} = PM_{i,j} / 2^{(\text{contrasttau} / (1 + (\text{contrasttau} - SB_i) / \text{scaletau}))}$
- $\text{contrasttau} = 0.03$ ,  $\text{scaletau} = 10$
- Corrected  $PM_{i,j}$  is  $PM_{i,j} - MM_{i,j}$

# Detection: Using the standard curve background adjustment method to make P/A calls

- Threshold on Gamma or a t-statistic
- Consider fitting the models robustly
- Use Affymetrix spike-in data to judge how well we are doing at detection
- The following ROC plot shows we do quite well in detection and that robustness helps

ROC curve for P/A using thresholds on gamma (or t)



# **Additional Material: Normalization Methods**

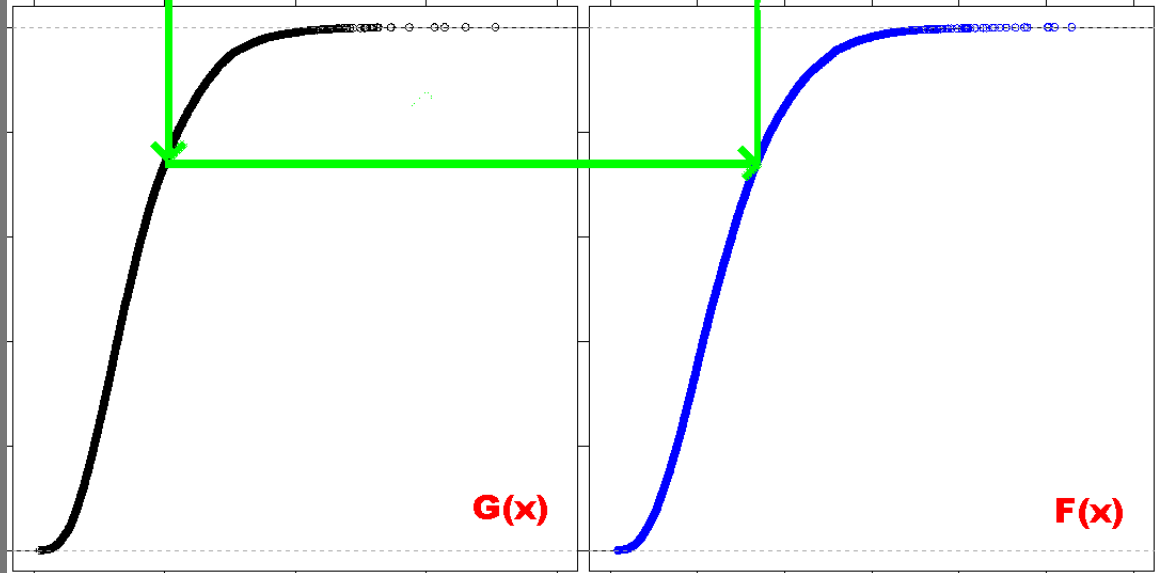
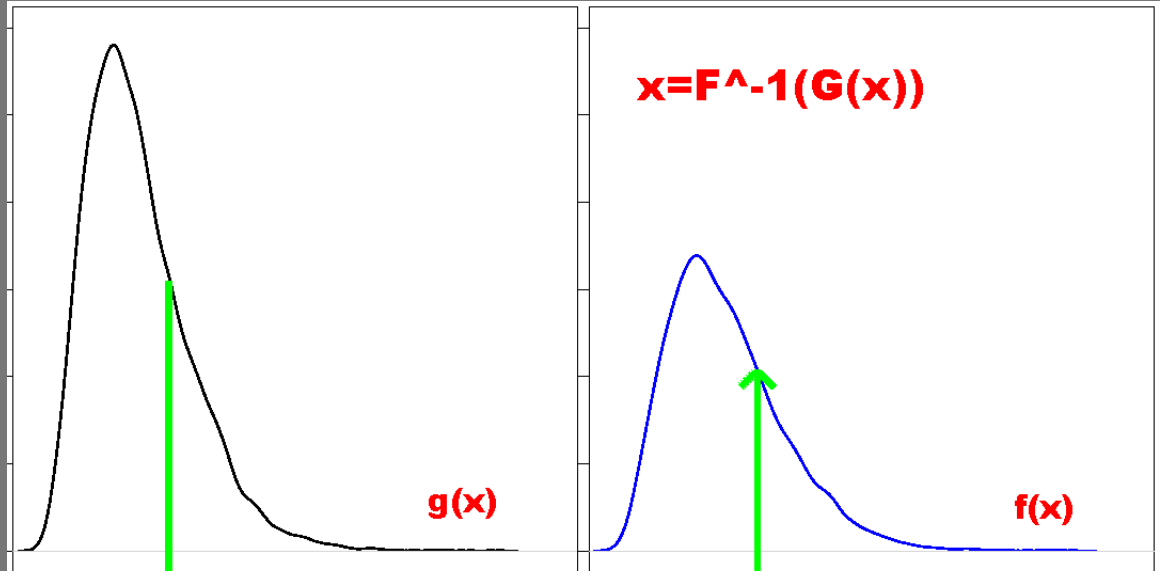
# Quantile Normalization

- See Bolstad et al (2003)
- Quantile normalization is a method to make the distribution of probe intensities the same for every chip
- The normalization distribution is chosen by averaging each quantile across chips
- The diagram that follows illustrates the transformation

Raw Data

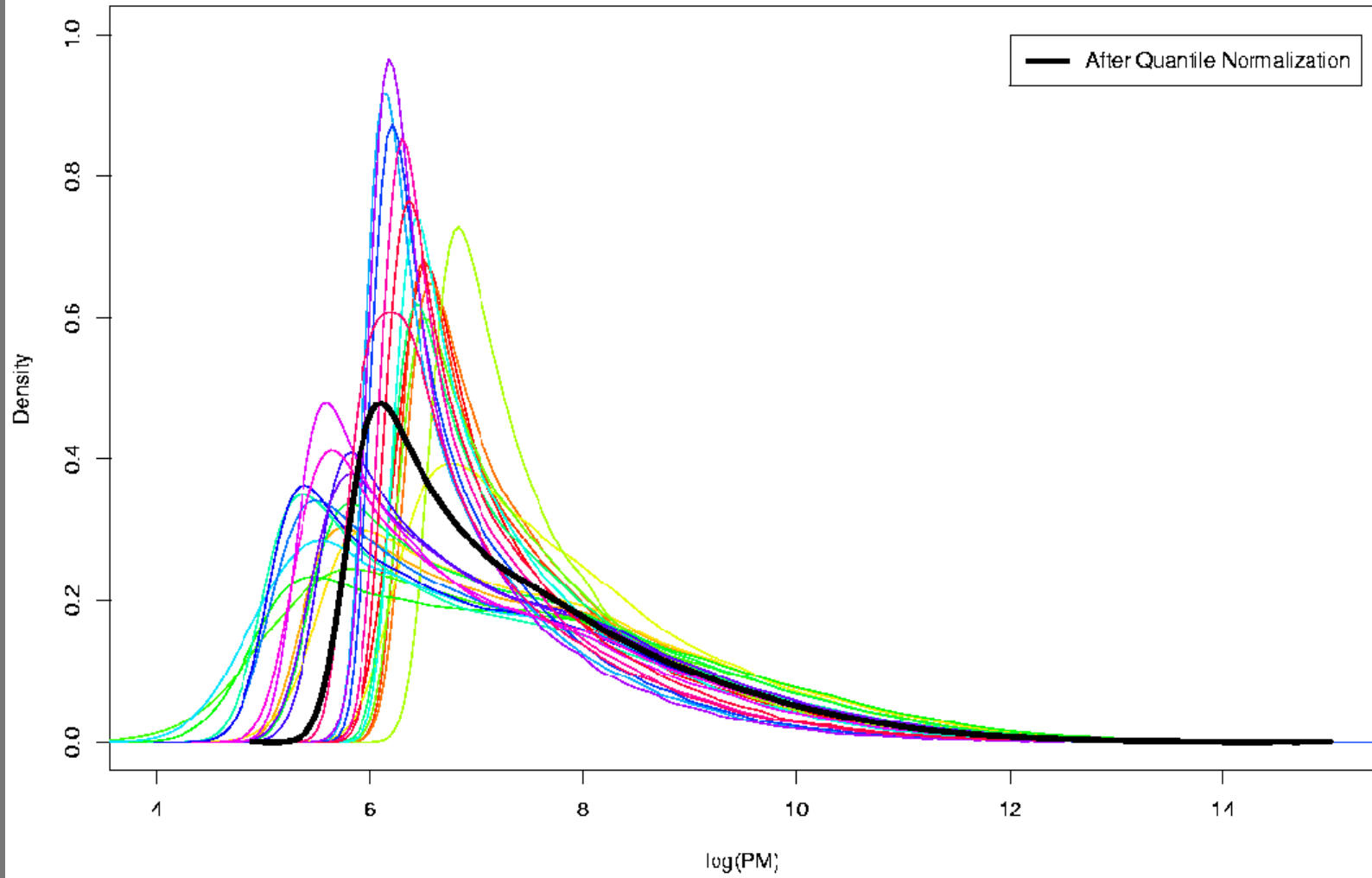
Normalization  
Distribution

Density Function



Distribution Function

Density of PM probe intensities for Spike-In chips



# Quantile Normalization (cont)

- The two distribution functions are effectively estimated by the sample quantiles
- Quantile normalization is fast
- After normalization, variability of expression measures across chips is reduced

# Normalization in MAS

- Compares a collection of experimental array with a baseline array, and normalizes the average intensity of the experimental array to the average intensity of the baseline array during normalization (sometime use a trimmed mean)
- We refer to this method as scaling
- MAS documentation applies normalization after summarization, we will use it before

# Other Prominent Normalization Methods

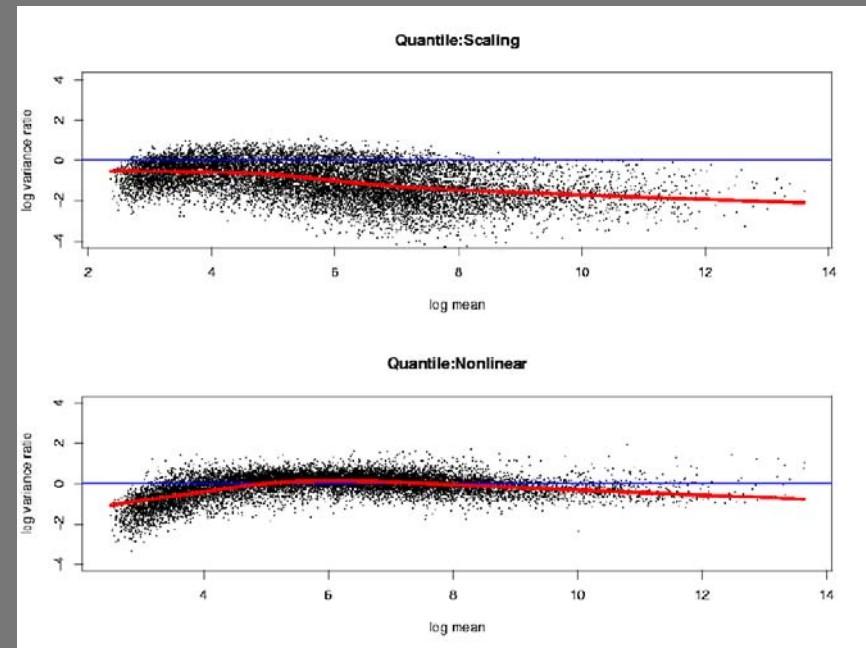
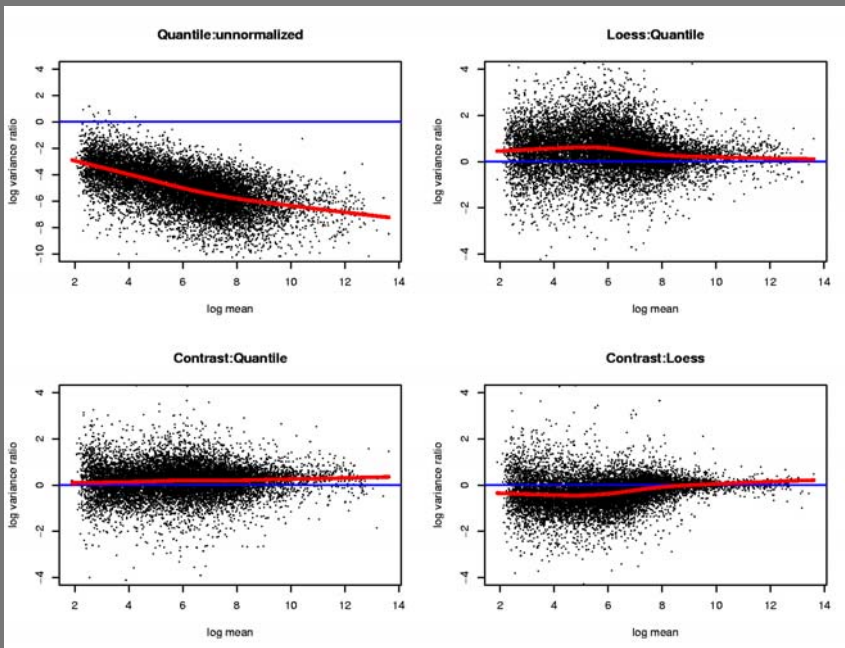
- Nonlinear – method used in dChip
  - pick a baseline chip then fit non linear relations (smoothing spines, running medians) between baseline chips and other chips
- Contrast, Cyclic loess
  - generalized M vs A loess normalization methods

# Bolstad et al (2003)

- Compares normalization methods in context of RMA measure
- Classifies normalization methods into two classes:
  - Complete Data Methods
    - Quantile
    - Contrast
    - Cyclic Loess
  - Baseline methods
    - Scaling
    - Non-linear

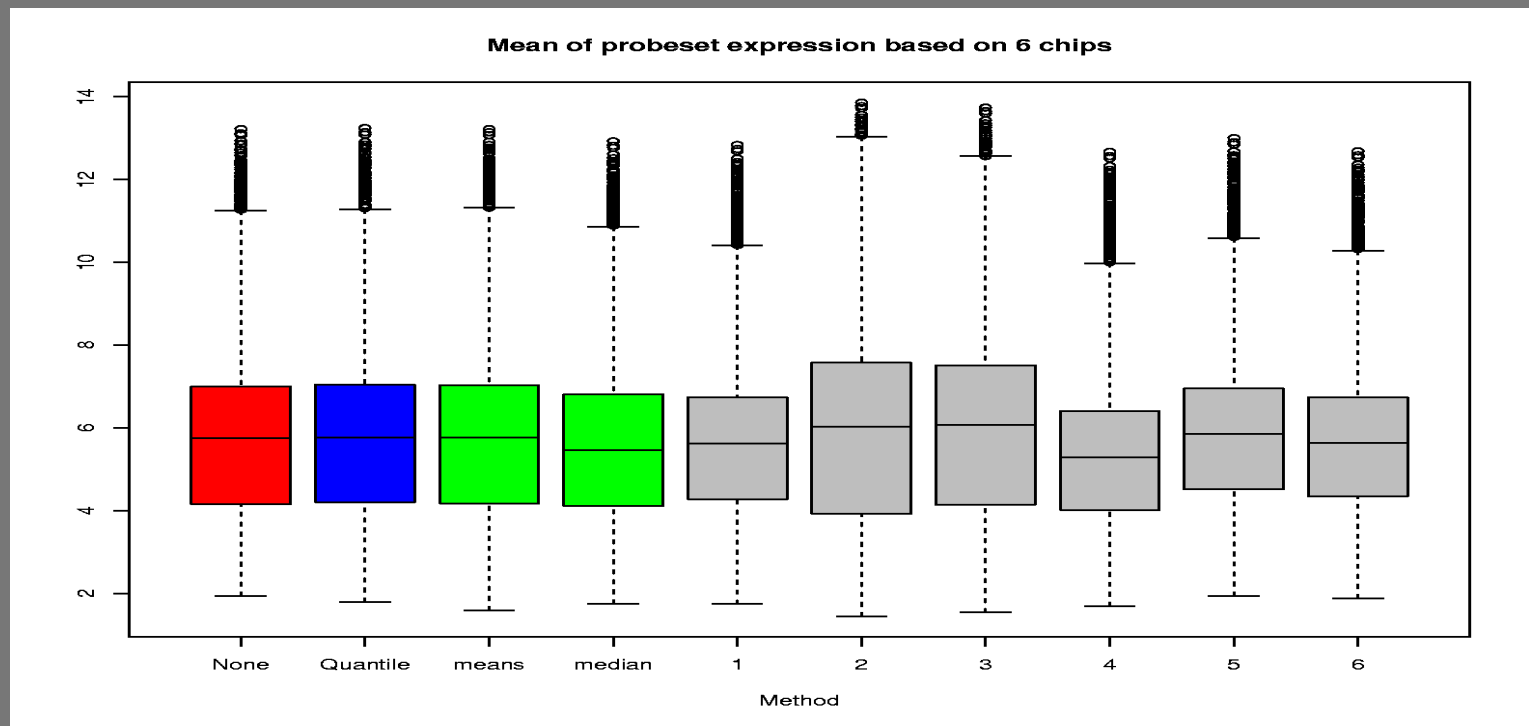
# Bolstad et al (2003) (cont)

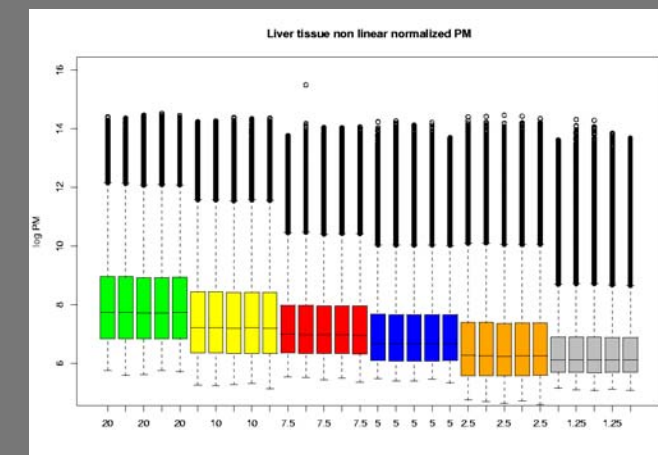
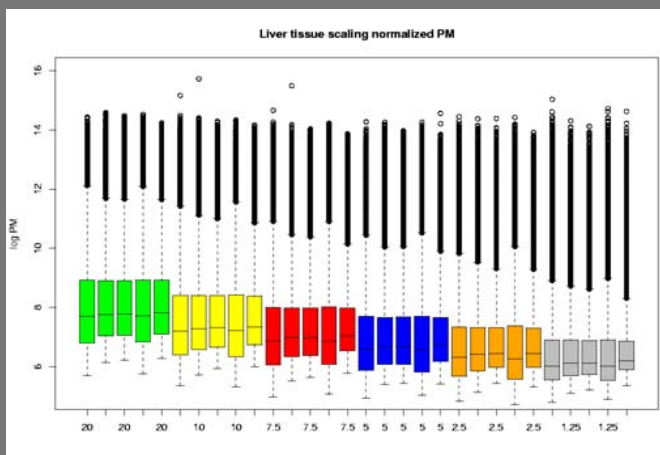
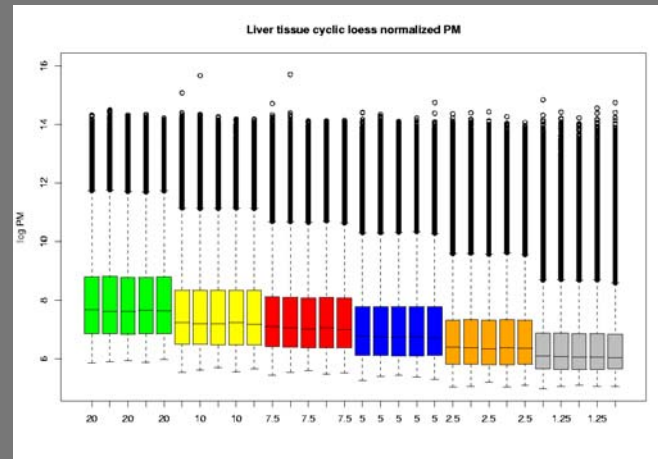
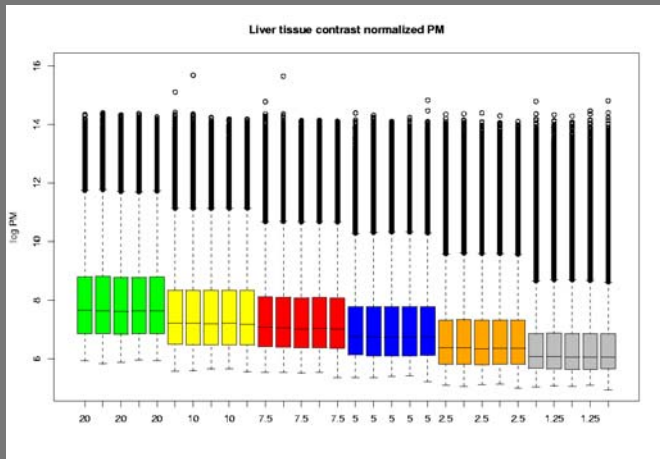
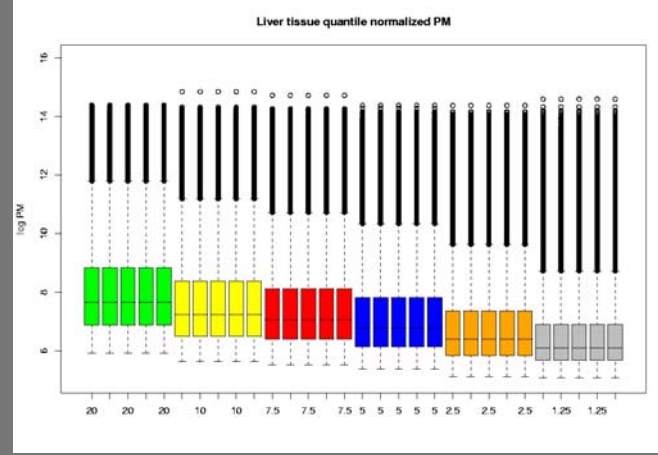
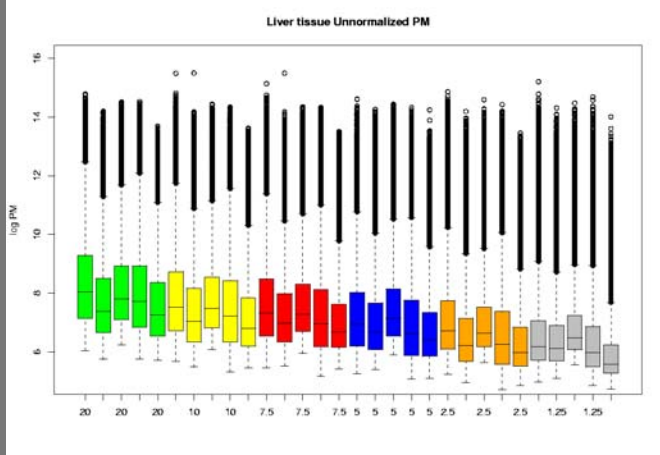
- Quantile normalization reduces between chip variability favorably when compared to other methods



# Bolstad et al (2003) (cont)

- Quantile method also found to perform well on the issue of bias (this was measured by using spike-in data)
- Complete data methods recommended over using a baseline





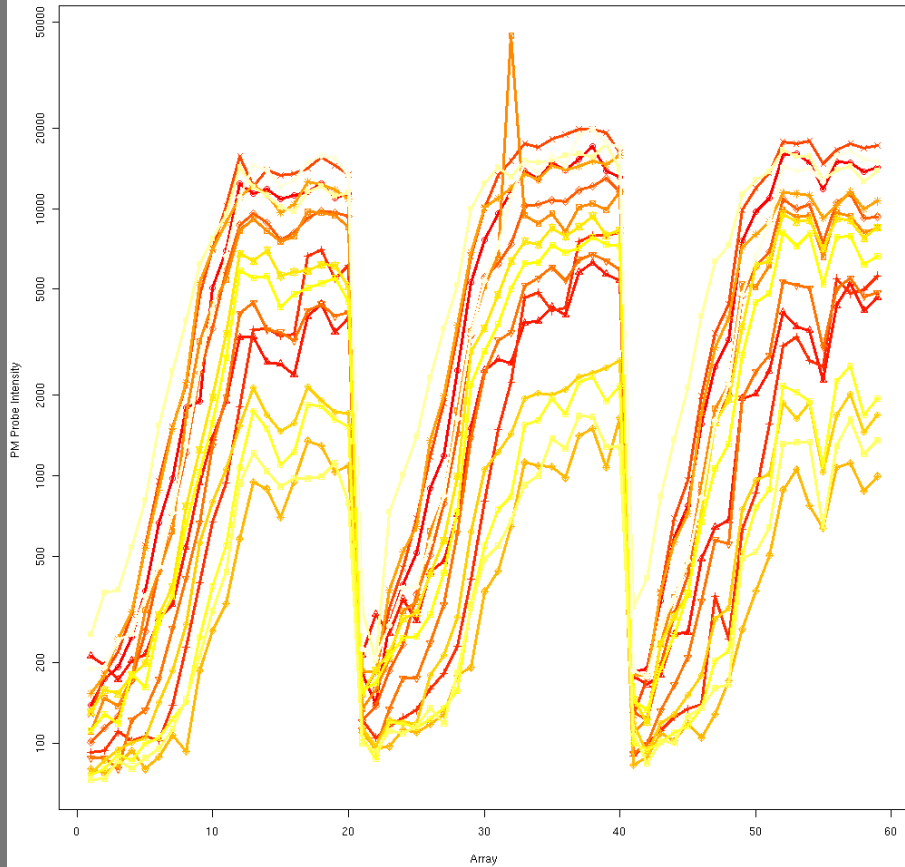
# **Additional Material: Summarization Methods**

# Motivation for RMA Model

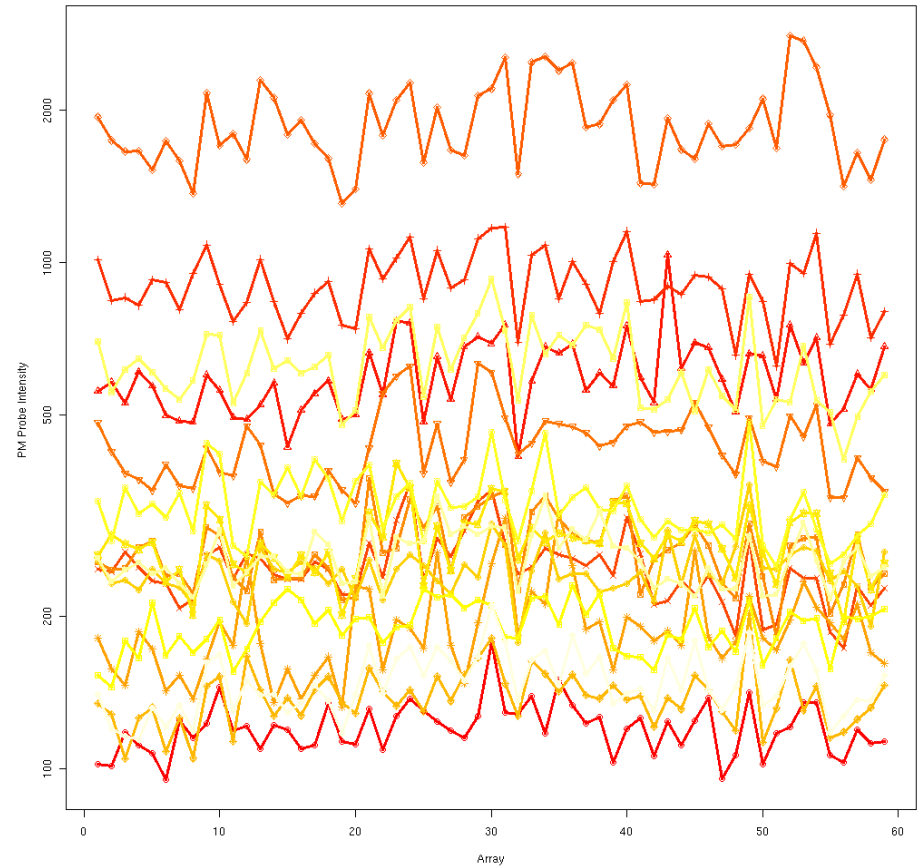
- Log scale stabilizes variance
- Parallel behavior in probe response observed across chips
- Robust because of outliers
- Following plots of probe responses across arrays show these features

# Parallel Behaviour for both a spike-in and a non spike-in

Probe behavior across chips (for a spike-in probset)

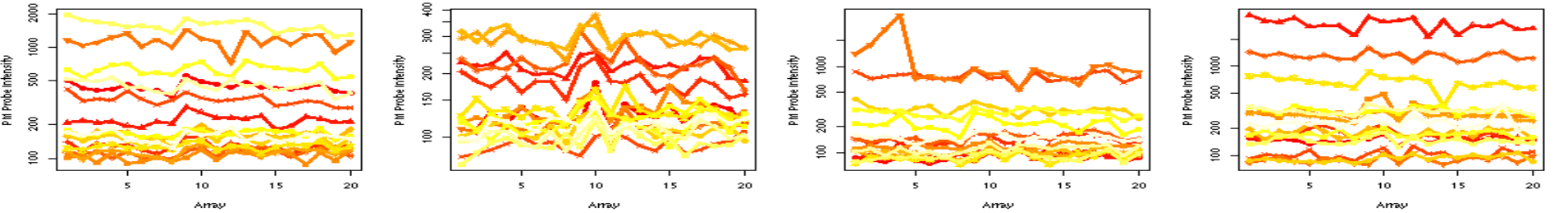


Probe behavior across chips (for a non spike-in probset)

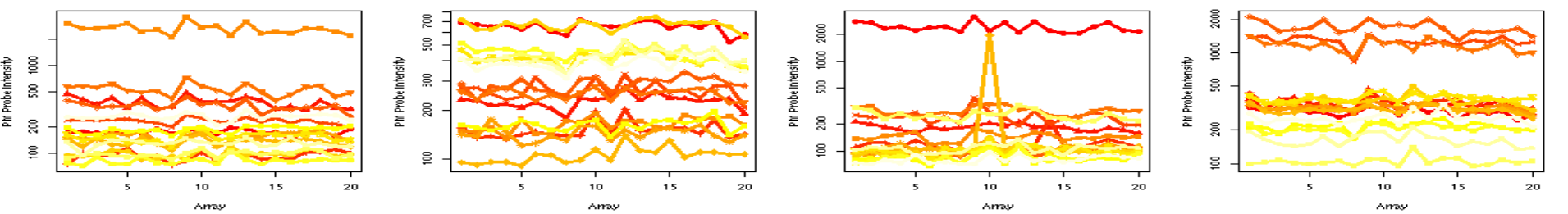


# A random set of non-spikeeins

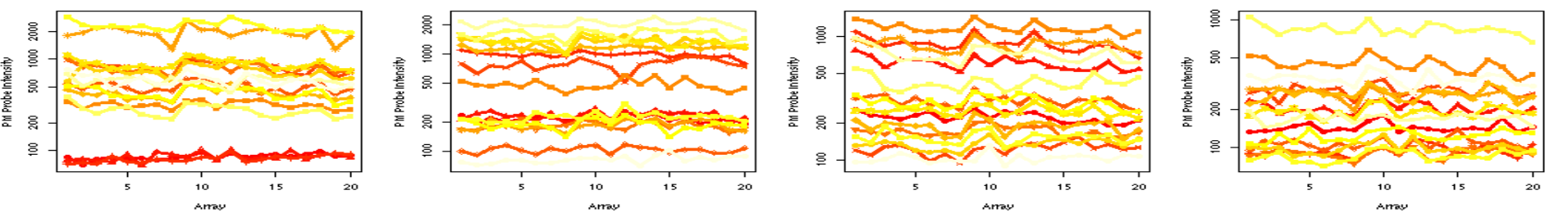
Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe)



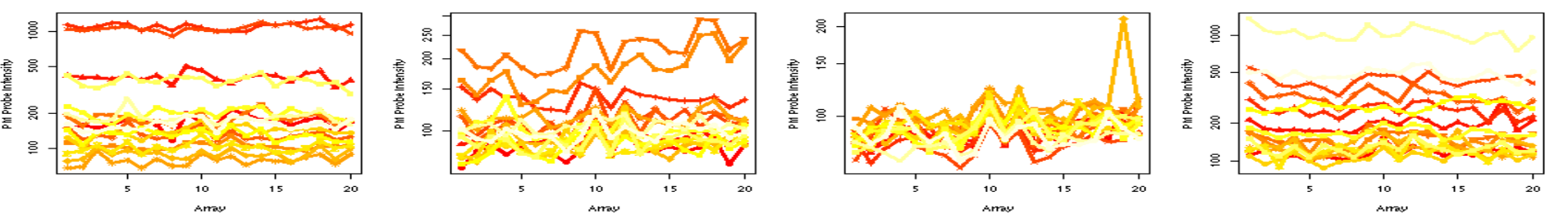
Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe)



Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe)



Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe) Probe behavior across chips (for a non spike-in probe)



# RMA: Robust Multichip Average

- Suppose we have  $j=1, \dots, J$  arrays and  $i=1, \dots, I$  probes for a given probeset
- Fit a robust linear model with probe and chip effects to log transformed data

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

Where  $\alpha_i$  is probe-effect and  $\beta_j$  chip effect

- Expression is on a log scale and given by

$$\mu + \beta_j$$

# RMA (cont)

- Method is compared with MAS 5.0 and Li-Wong MBEI in Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed (2002) Summaries of Affymetrix GeneChip Probe Level Data. Nucleic Acids Research
- Found to outperform other methods in most comparisons
- Current implementations use median-polish to fit the linear model, other robust linear model fitting procedures are being explored

# MAS 5.0: “The Statistical Algorithm”

- Using log-scale data for the probes related to the probeset on single chip
- Suppose  $P_i$  for  $i=1, \dots, I$  are preprocessed probe values
- Then expression is given by

$$E = T_{bi}(P_1, \dots, P_I)$$

Where  $T_{bi}()$  is the 1 step Tukey Biweight

# Other Expression Measures

- Not explored here
- AvDiff – the old Affymetrix method. Found lacking for a number of reasons
- Li-Wong MBEI (Model Based Expression Index) – implemented in the dChip software