

Low-Level Analysis of High-Density Oligonucleotide Microarray Data

Ben Bolstad

<http://www.stat.berkeley.edu/~bolstad>

Biostatistics, University of California, Berkeley

University of Florida

Feb 17, 2004

Outline

- What is Low-Level Analysis?
- Affymetrix GeneChip Technology
- Two topics in Low-level analysis
 - Constructing a gene expression measure
 - Probe level models for detecting differential expression

Low-Level Analysis

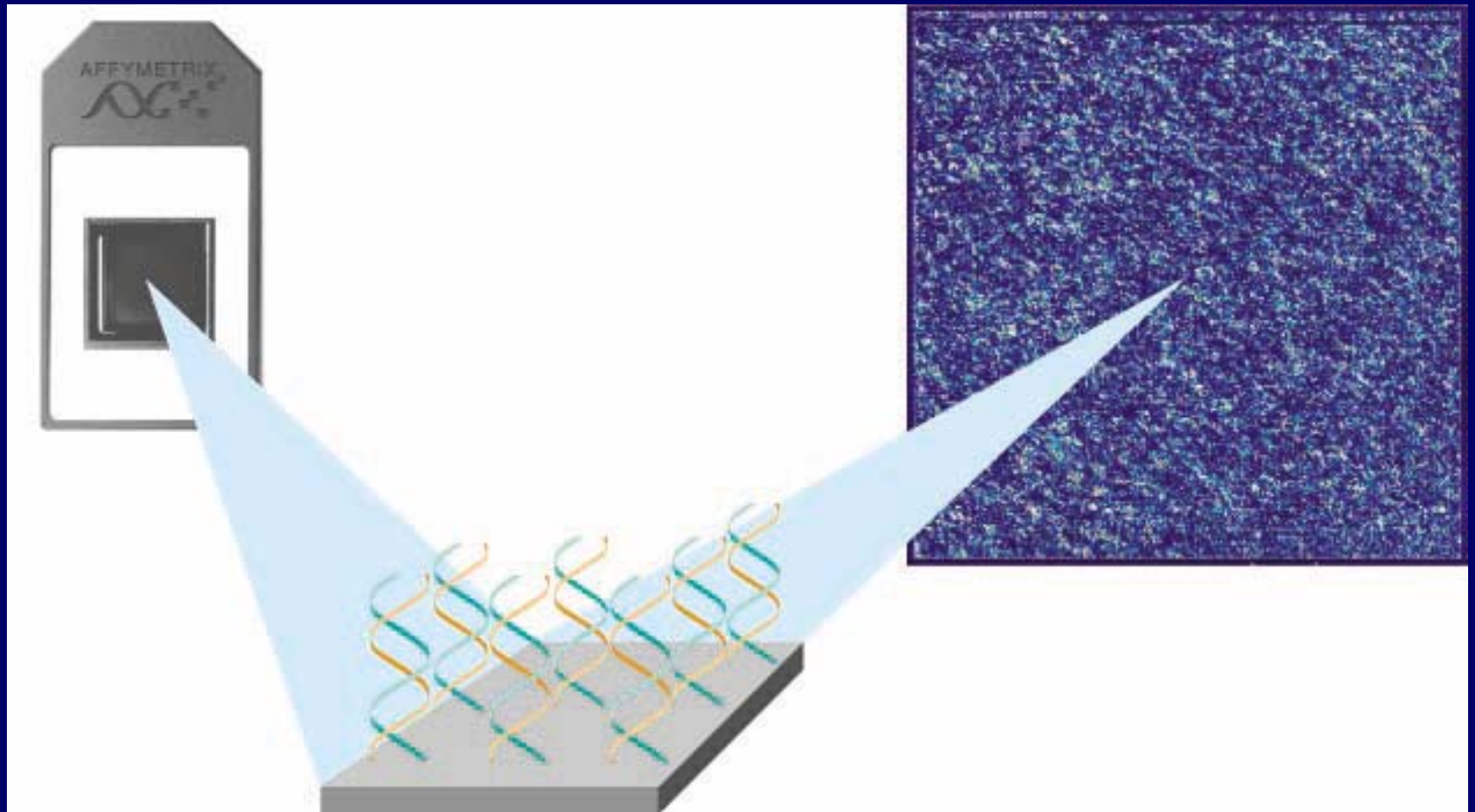
- What is low-level analysis?
 - Analysis and manipulation of probe intensity data
 - Expression calculation: Background, Normalization, Summarization
 - Determining presence/absence
 - Quality control diagnostics
- Why do we do it?
 - Hopefully it will allow us to produce better, more biologically meaningful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

High-Level Analysis

- Clustering/Classification
- Pathway Analysis
- Cell Cycle
- Gene function
- Anything where a more biological interpretation is desired

Such matters will not be discussed further in today's talk.

From Chip To Data

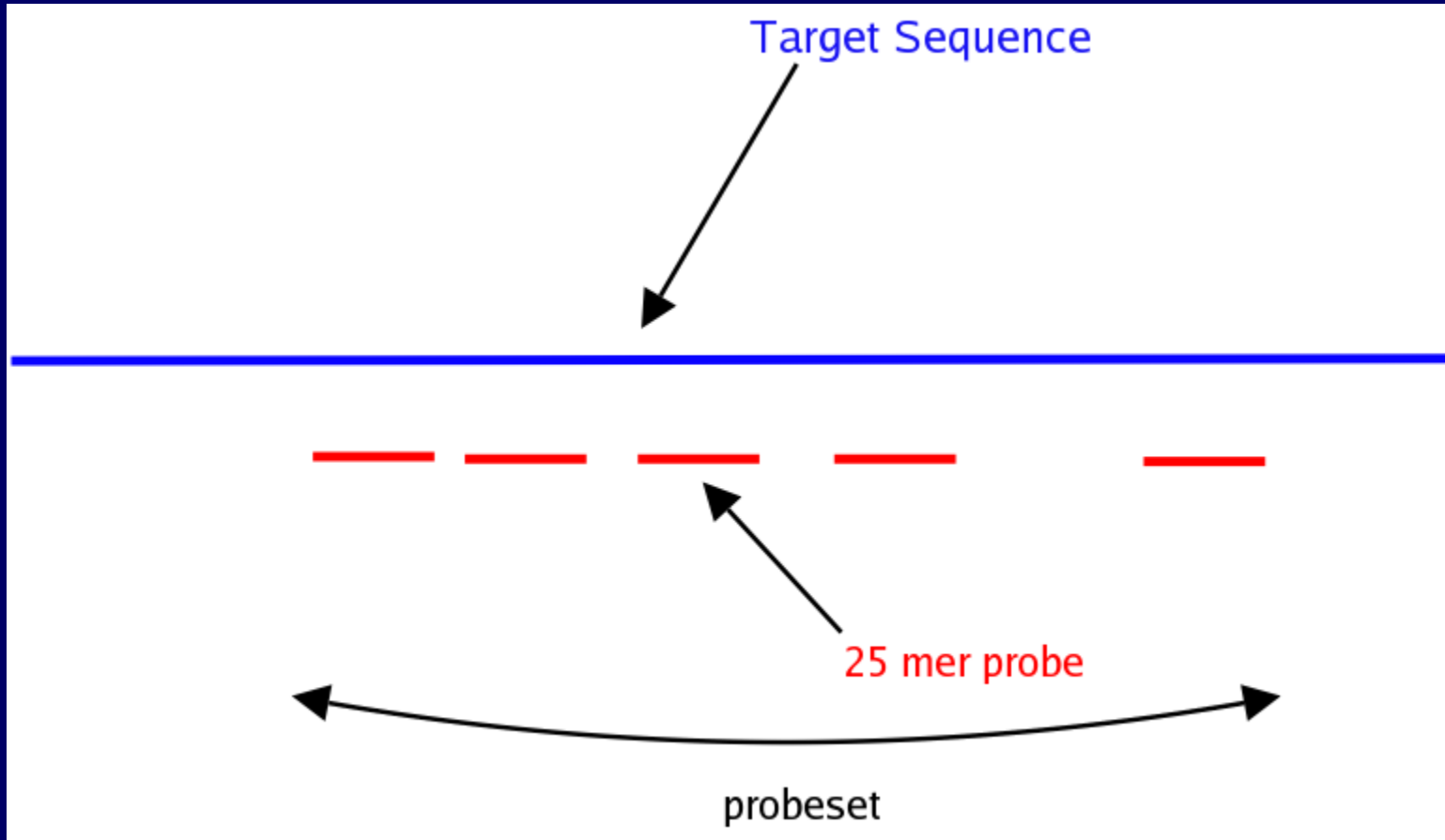


Brief Technology Overview

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



Probes and Probesets



Two Probe Types

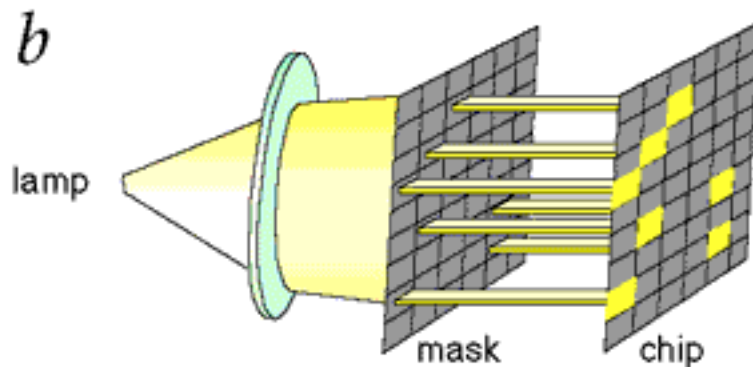
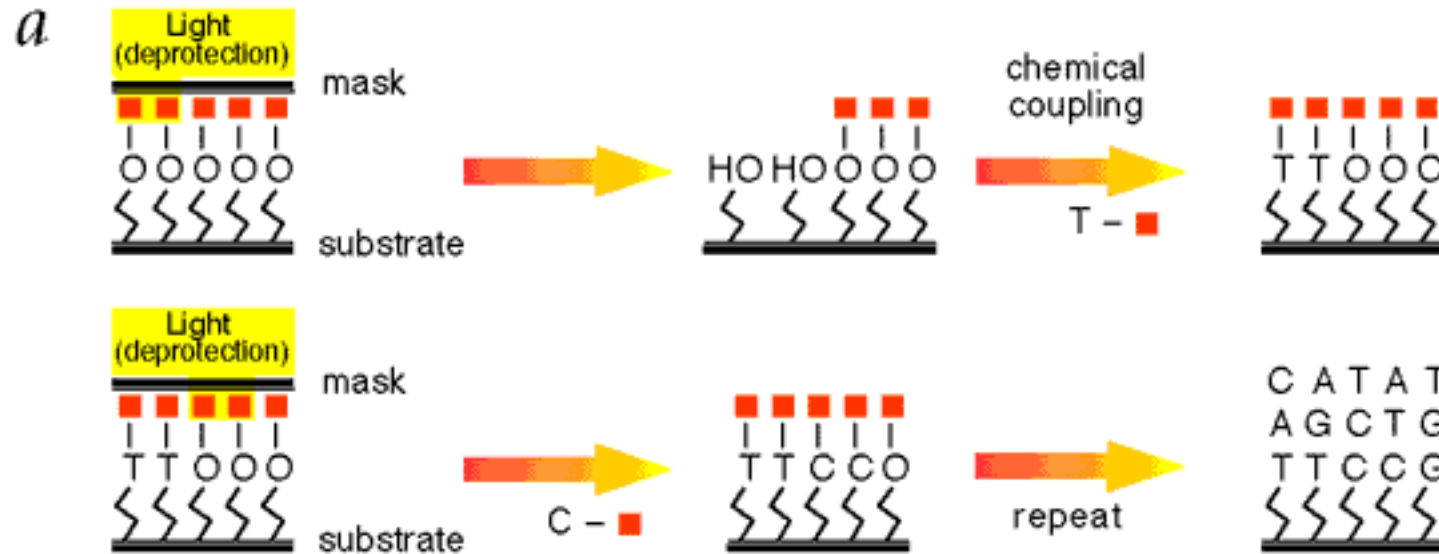
PM: the Perfect Match

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

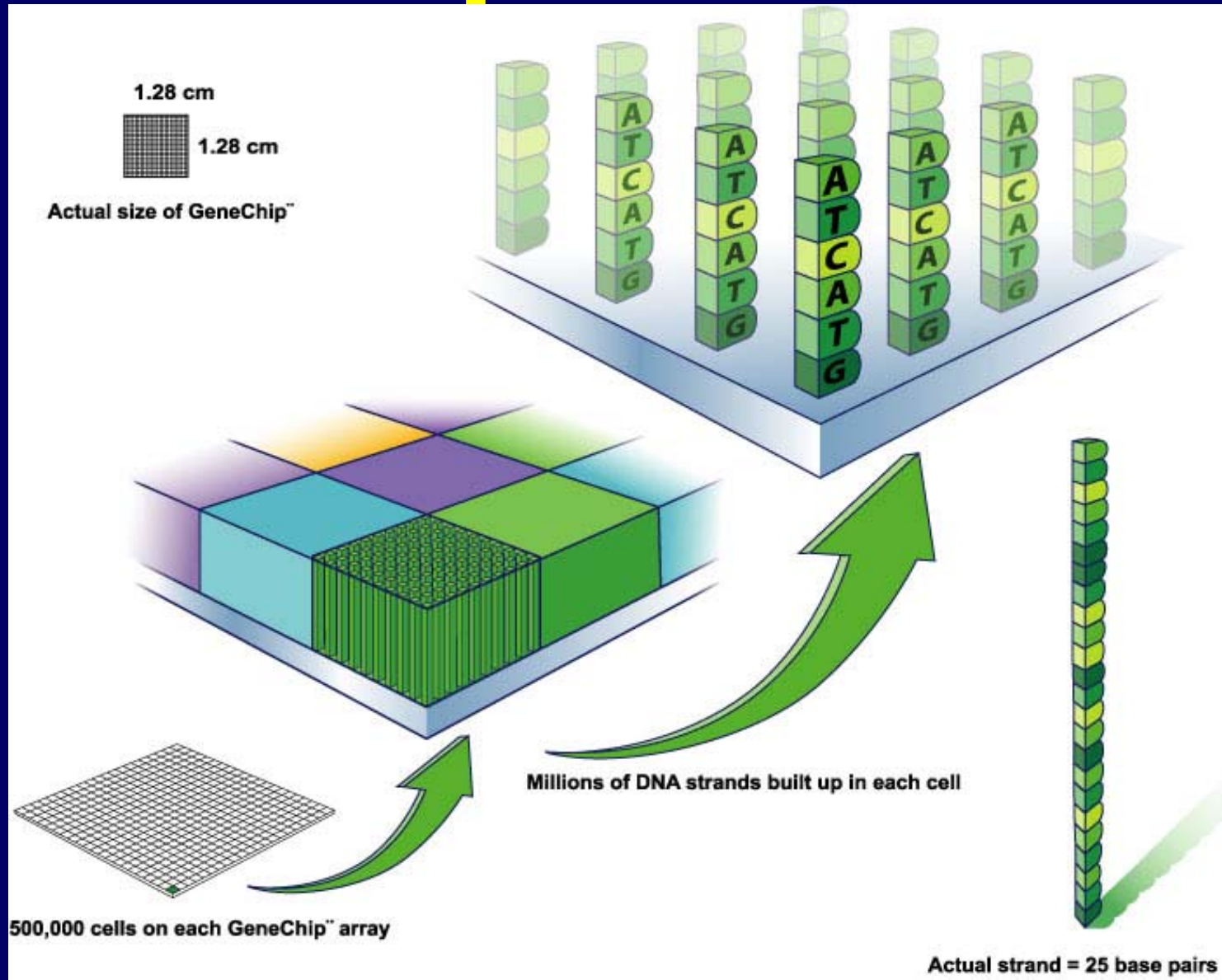
PM: CAGACATAGTGTCTGTGTTTCTTCT

MM: CAGACATAGTGTGTGTGTTTCTTCT

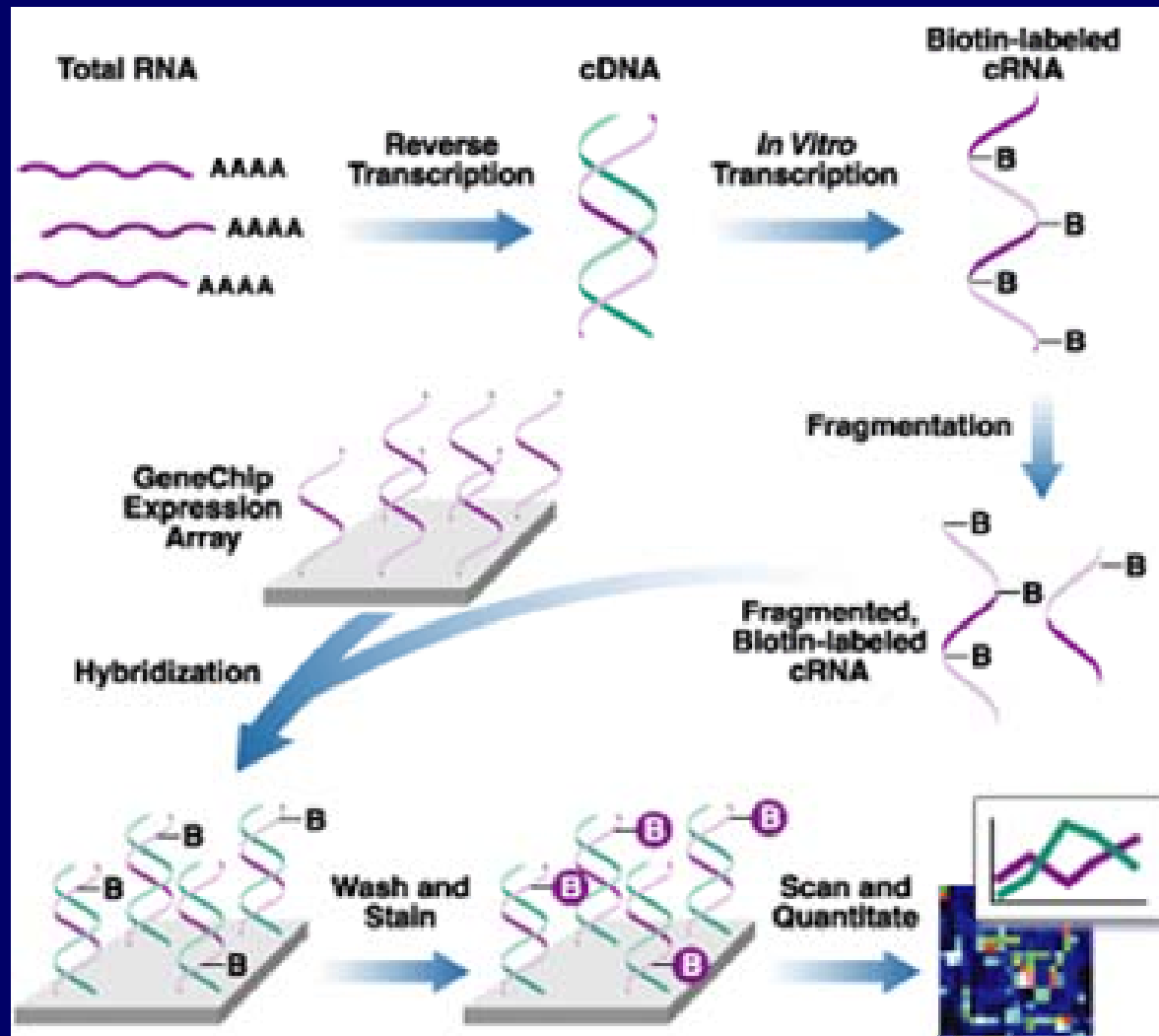
Constructing the Chip



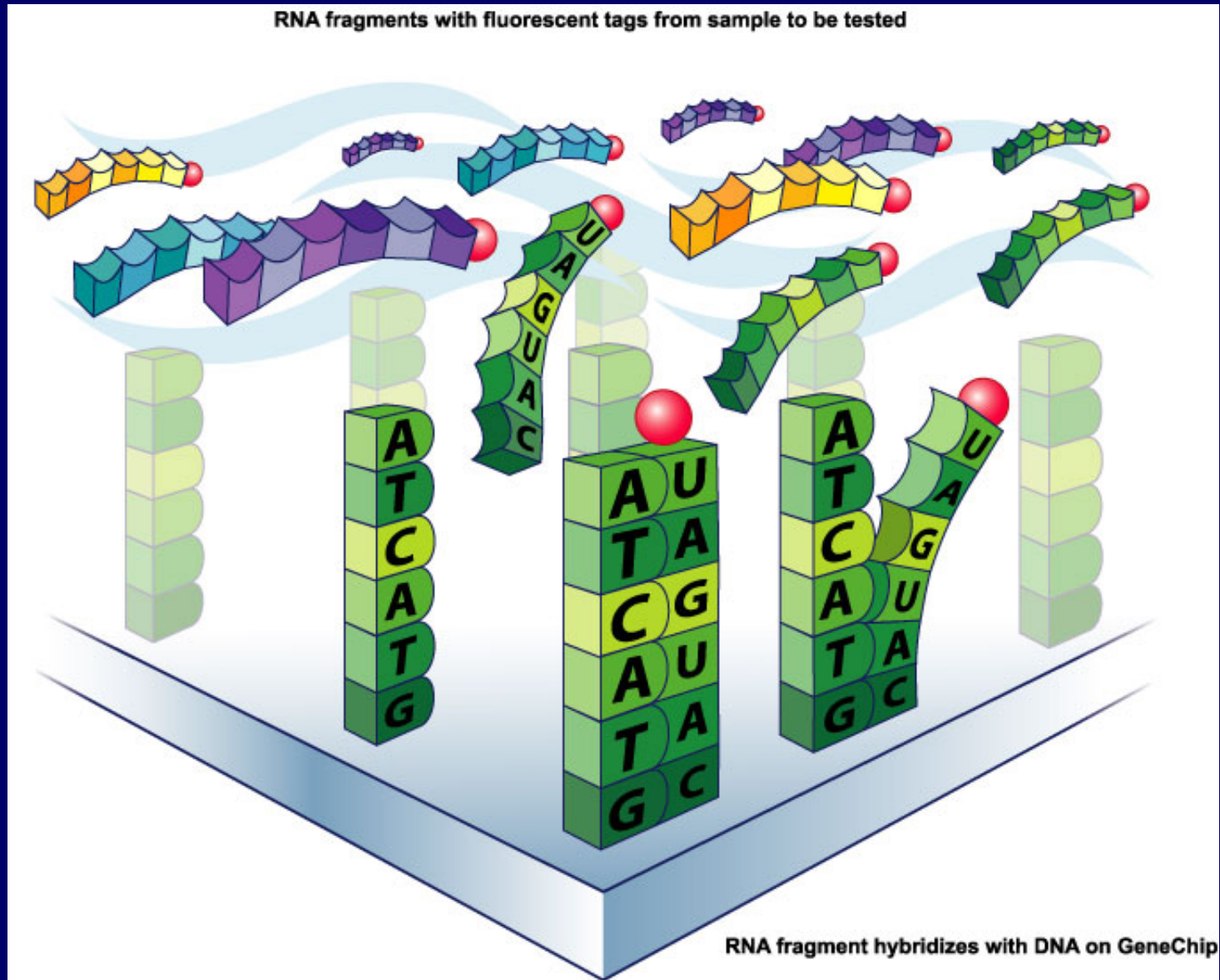
Focusing on a Single GeneChip Cell Location



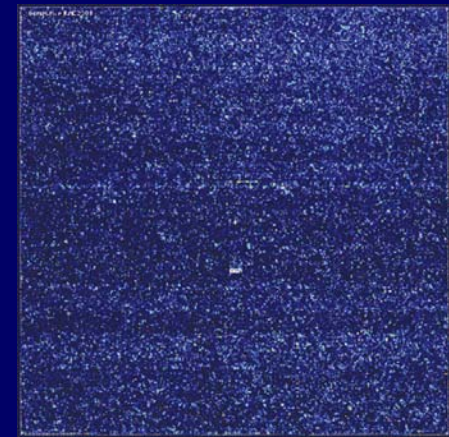
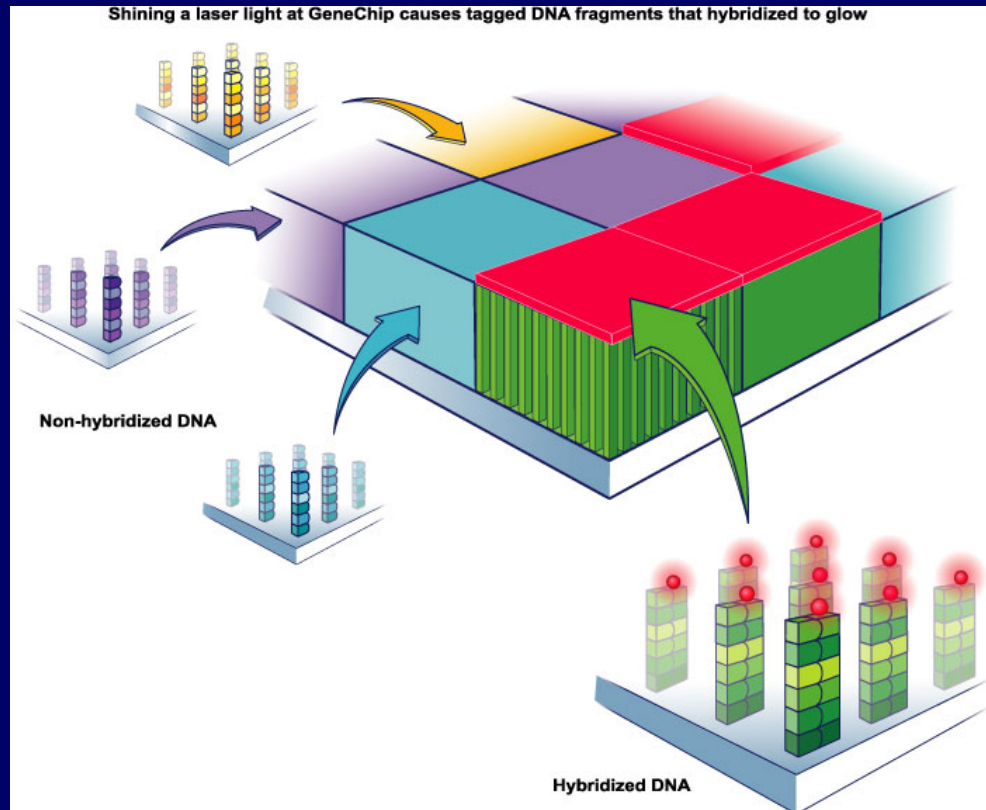
Sample Preparation



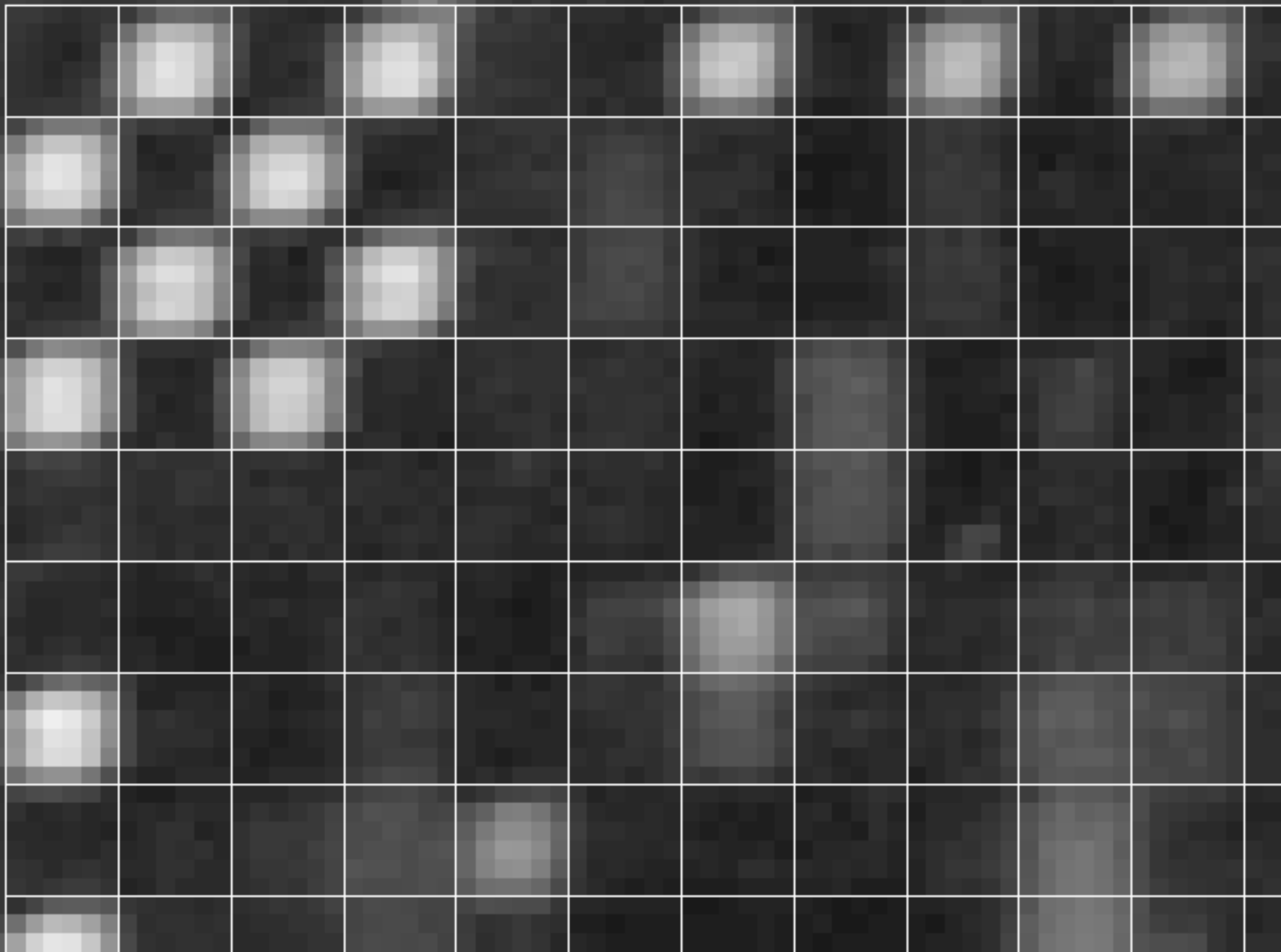
Hybridization to the Chip

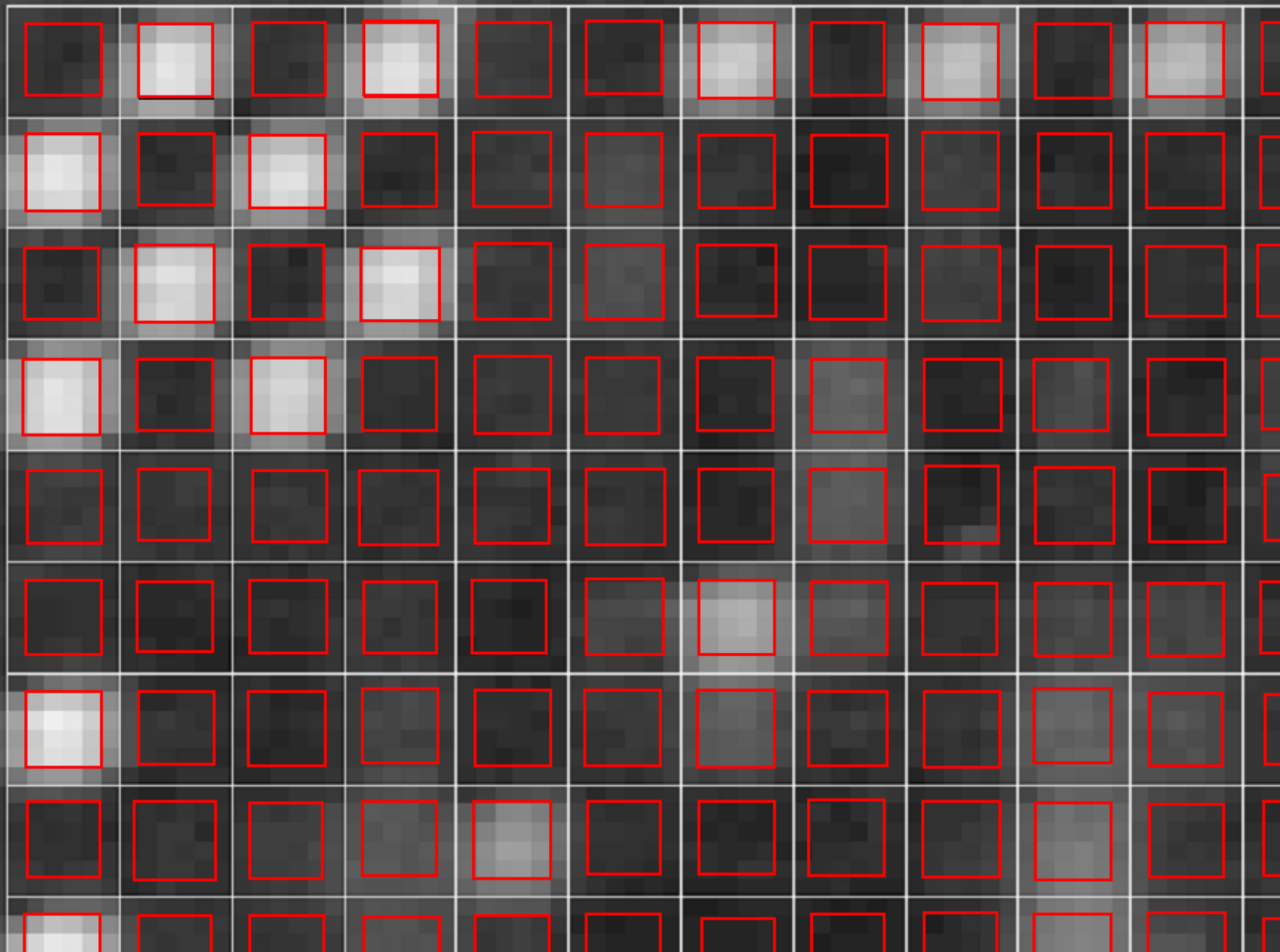


The Chip is Scanned











Constructing a gene expression measure

Computing Expression Measures: A Three Step Procedure

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

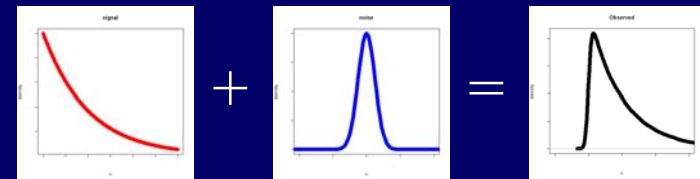
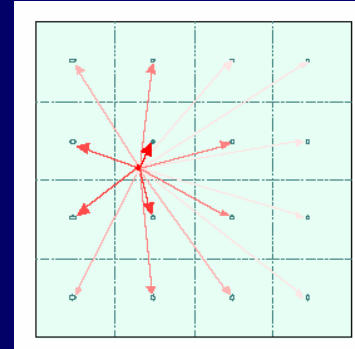
Let X be cel file data from multiple arrays then
Expression values = $S(N(B(X)))$

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 Methods

- Affymetrix
 - Location dependent
 - Ideal mismatch
- RMA
 - Convolution model
- Other
 - Standard curve adjustment
 - GCRMA (Wu et al 2003)



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.”¹

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

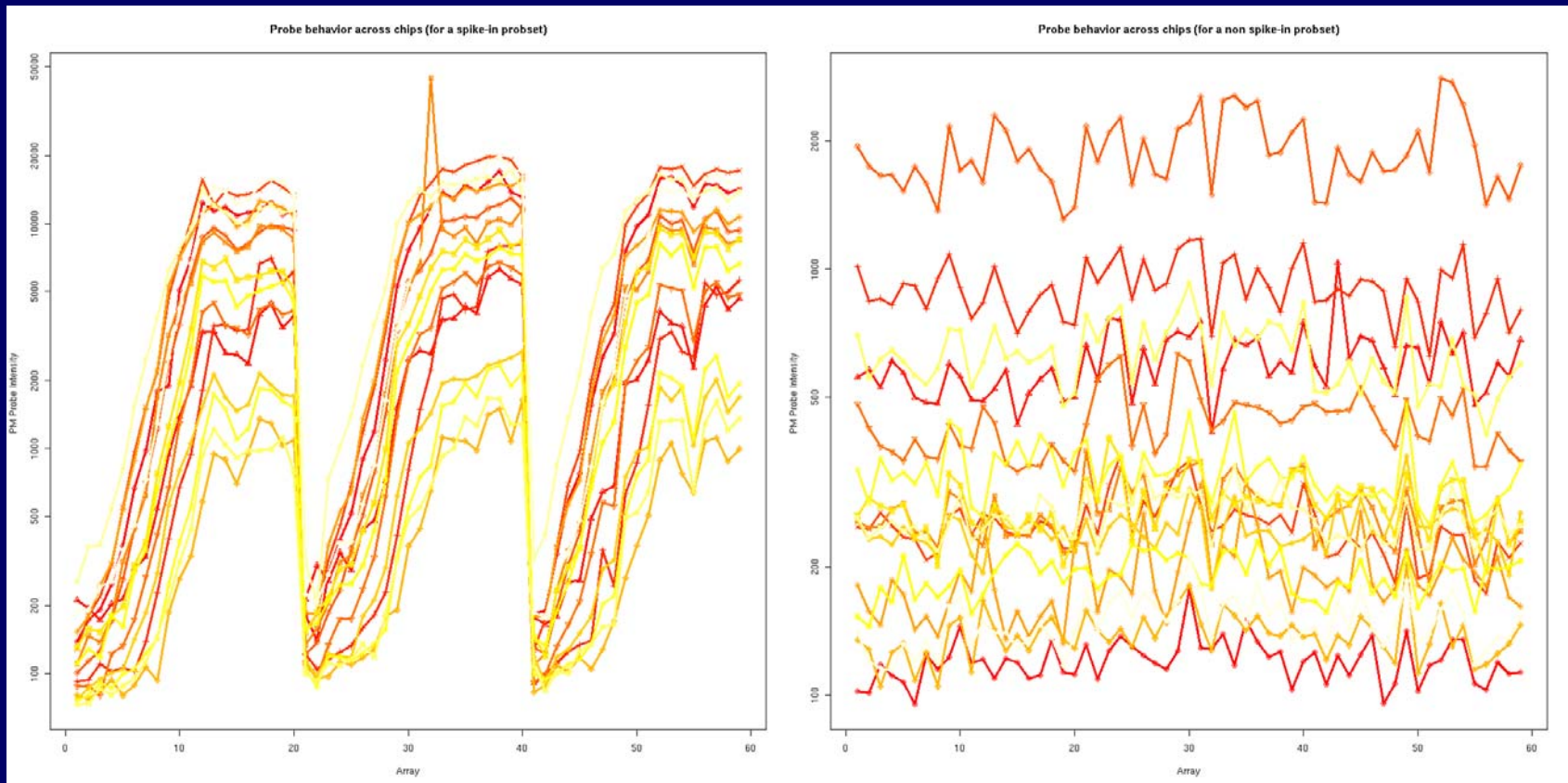
Normalization Methods

- Methods already compared in Bolstad et al (2003)
- 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)

Summarization

- Reduce the 11-20 probe intensities for each probeset 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

Parallel Behaviour Suggests Multi-chip Model



In this context what is RMA?

- Convolution Background
- Quantile Normalization
- Linear model on the log₂ scale fit robustly.

- Software for implementing RMA is in the Bioconductor *affy* package

Affymetrix Spike-in Data

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

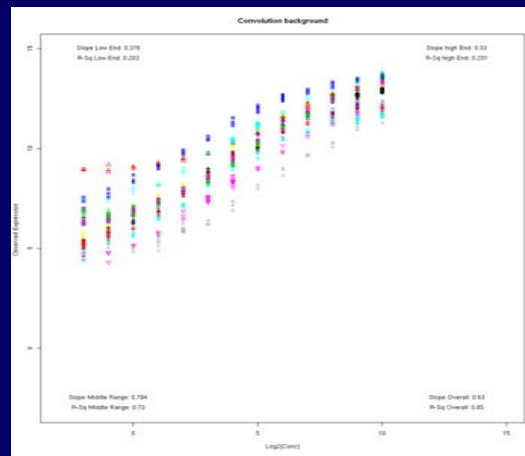
	37777	684	1597	38734	39058	36311	36889	1024	36202	36085	40322	407	1091	1708
A	0	0.25	0.5	1	2	4	8	16	32	64	128	0	512	1024
B	0.25	0.5	1	2	4	8	16	32	64	128	256	0.25	1024	0
C	0.5	1	2	4	8	16	32	64	128	256	512	0.5	0	0.25
D	1	2	4	8	16	32	64	128	256	512	1024	1	0.25	0.5
E	2	4	8	16	32	64	128	256	512	1024	0	2	0.5	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

Comparing the background methods

- Using an Affymetrix spike-in experiment we shall examine
 - Observed vs spike-in concentration
 - Observed vs expected fold change
 - Composite M vs A plots
 - ROC curves
- In each case we will compute expression values use standard RMA methodology. (ie quantile normalization, median polish summarization)

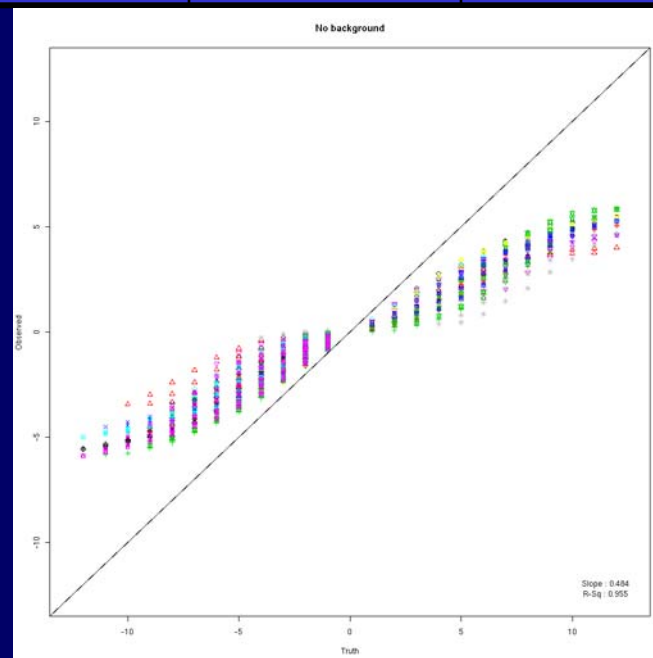
Assessing Bias: Observed Expression vs Spike-in Concentration

Slope	None	RMA	MAS 5	IMM	MAS5/ IMM	S.C.A.
All	0.493	0.63	0.589	0.69	0.695	0.856
Mid	0.665	0.784	0.751	0.82	0.82	1.041
Low	0.184	0.376	0.318	0.52	0.563	0.631
High	0.329	0.33	0.327	0.295	0.291	0.256



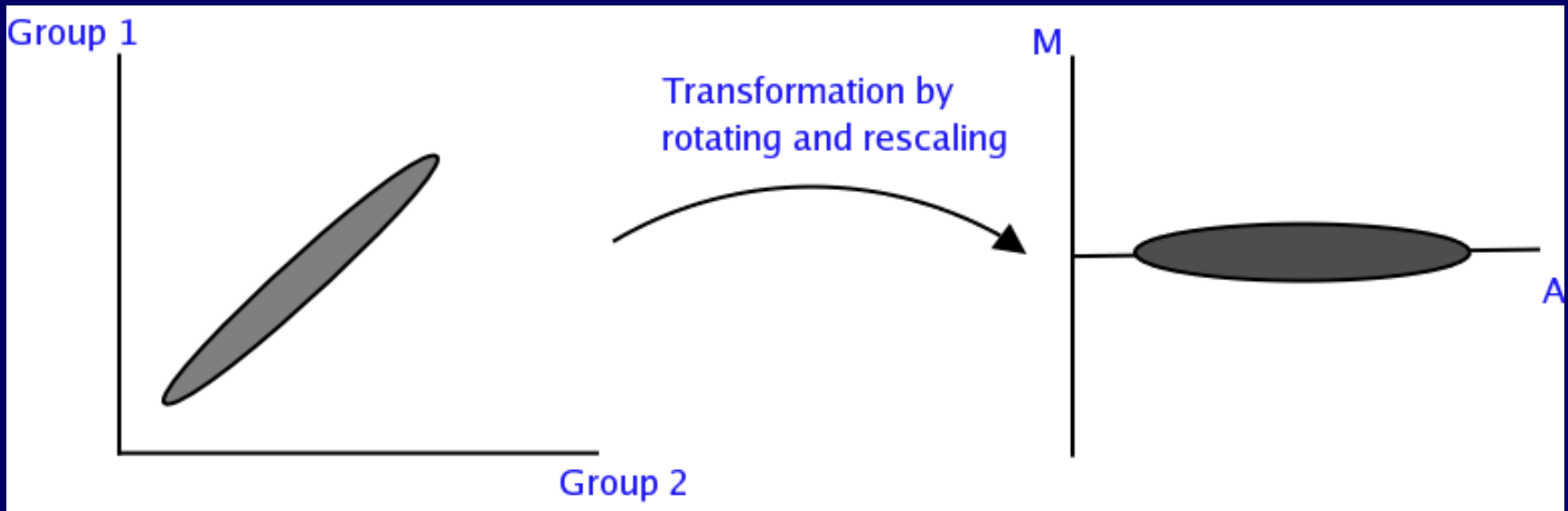
Assessing Bias: Observed Fold-change versus Expected Fold-change

Slope	None	RMA	MAS 5	IMM	MAS5/ IMM	S.C.A.
All	0.484	0.624	0.583	0.683	0.692	0.847

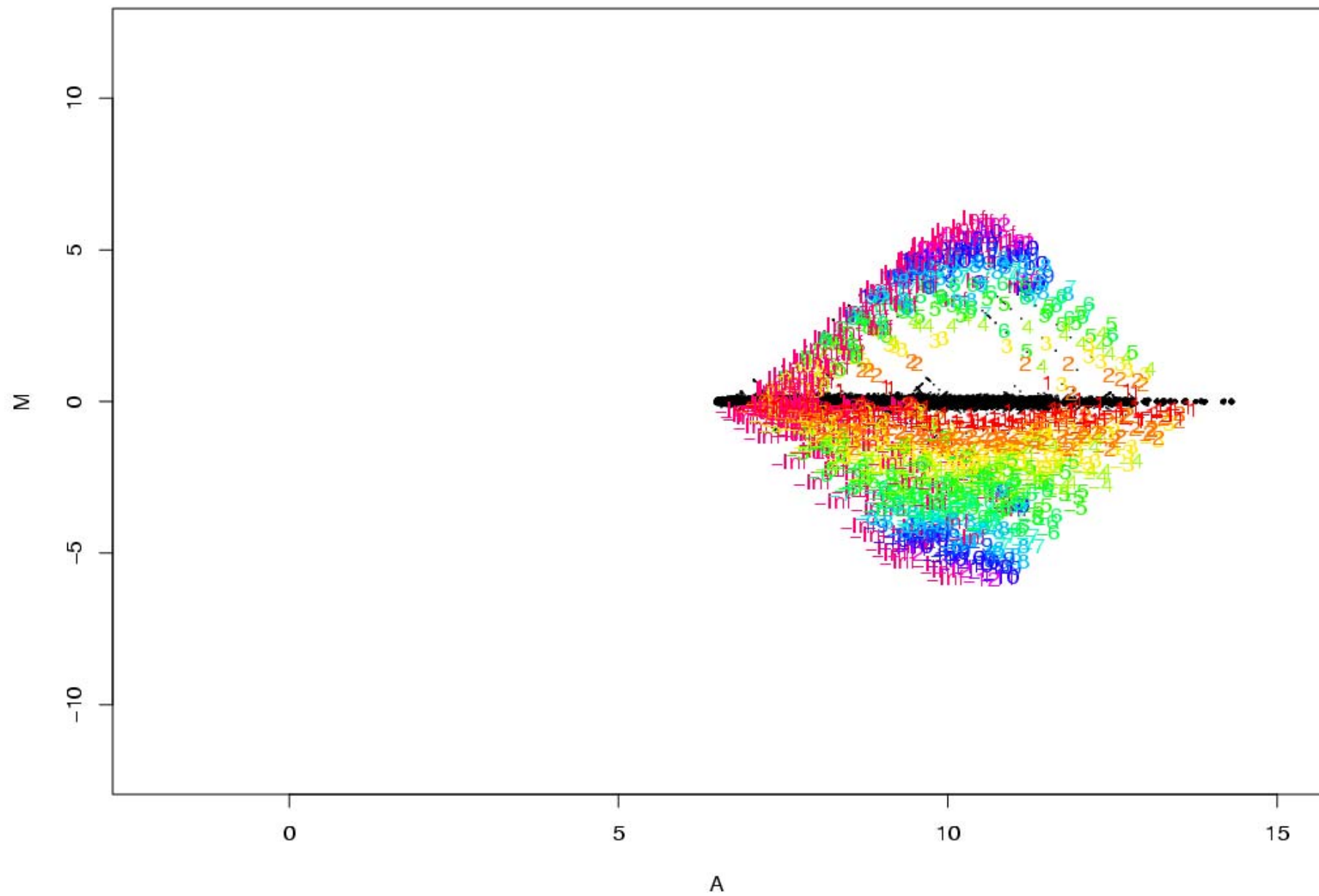


Assessing Variability: M vs A plots

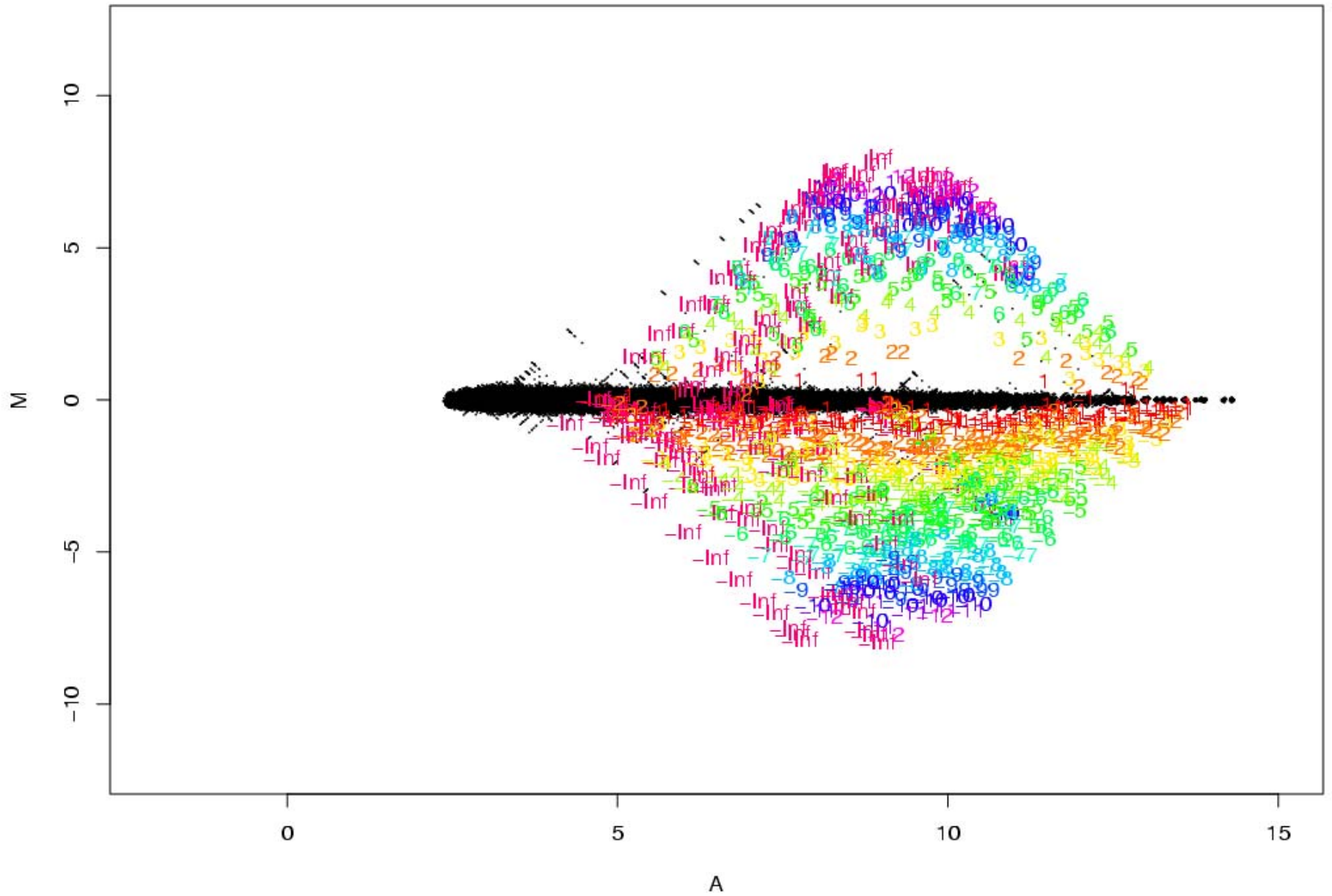
- Vertical Axis is M a log₂ fold-change.
- Horizontal Axis is A an average absolute expression value.
- Ideally non differential genes tight about M=0



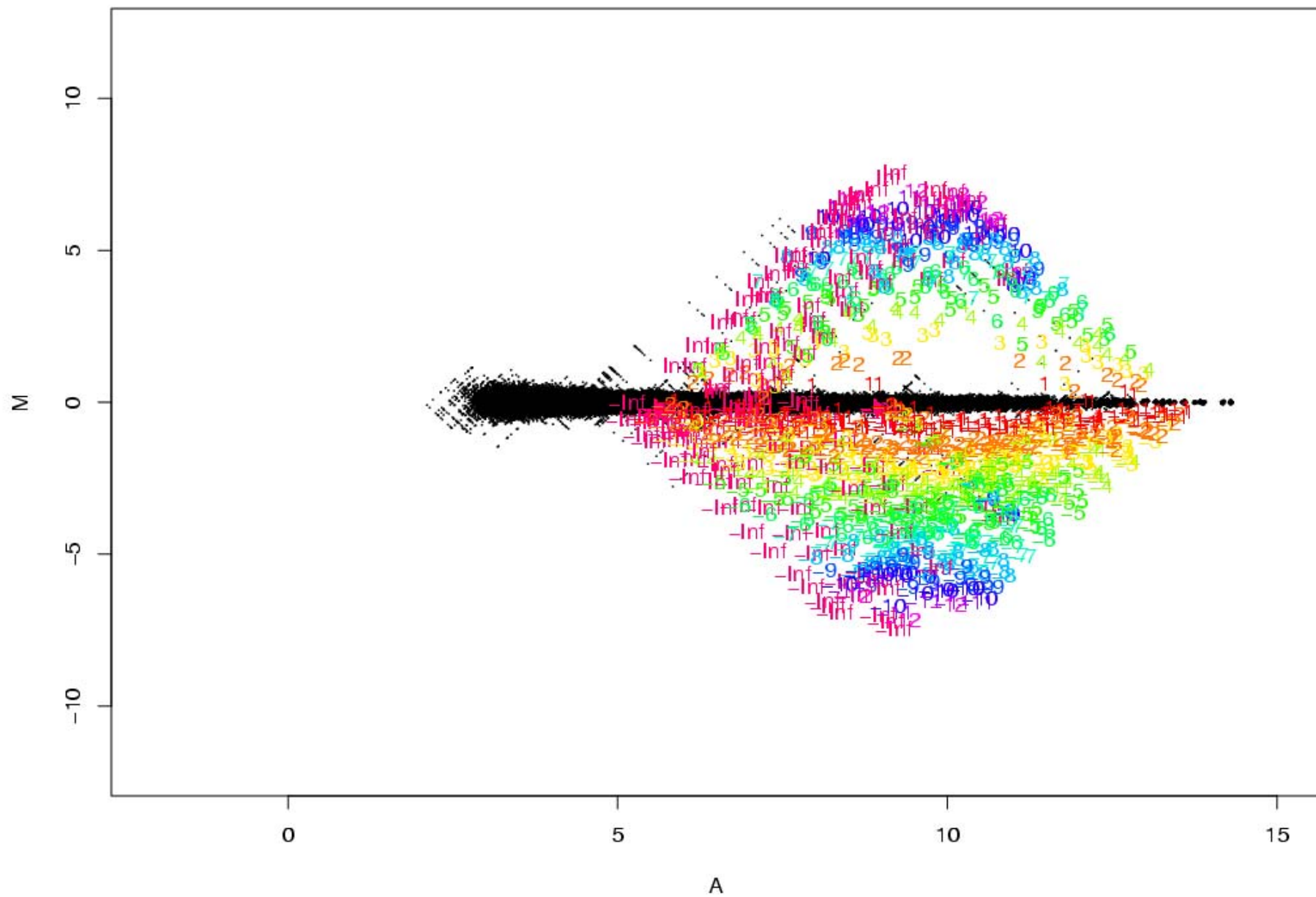
No background



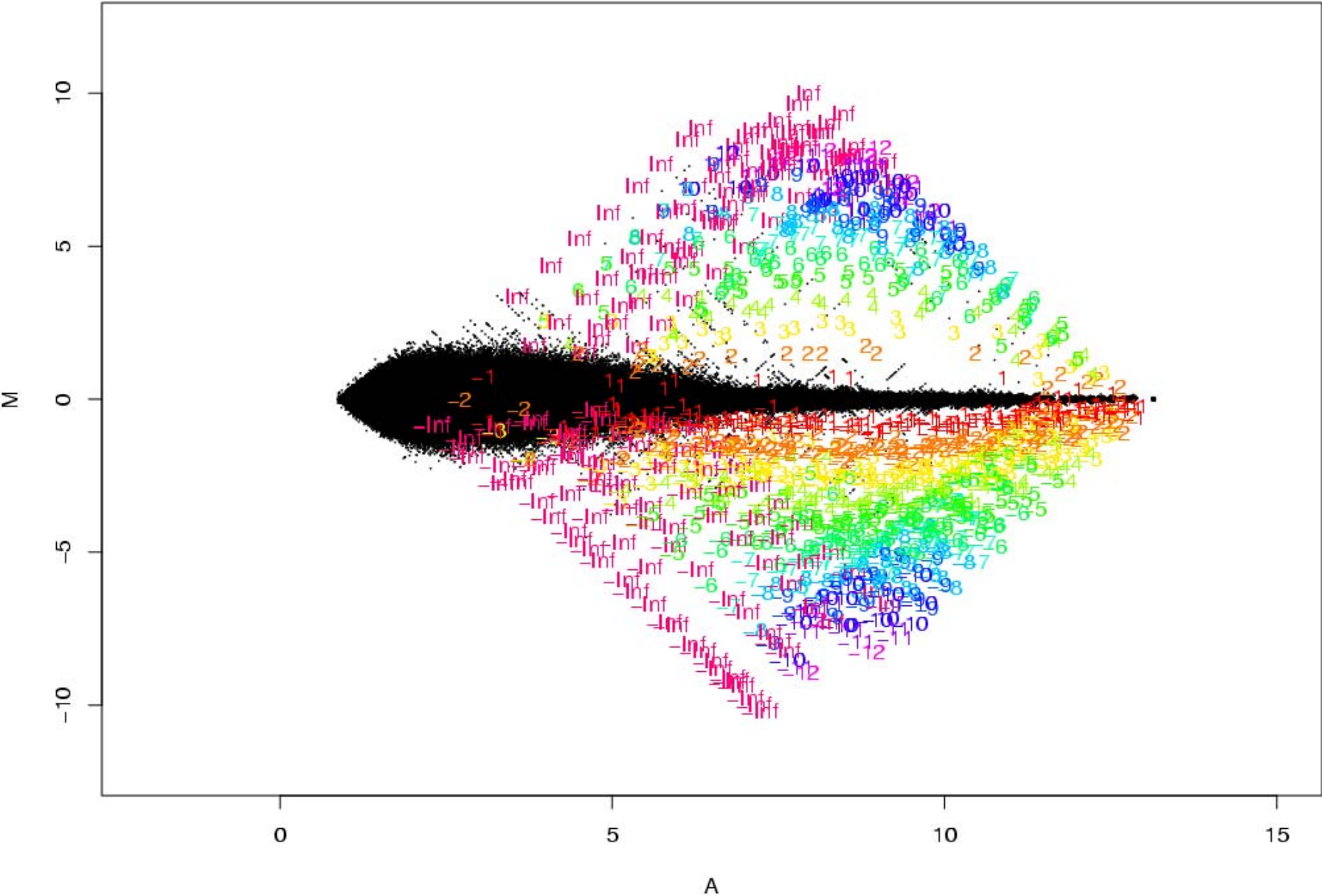
RMA convolution background



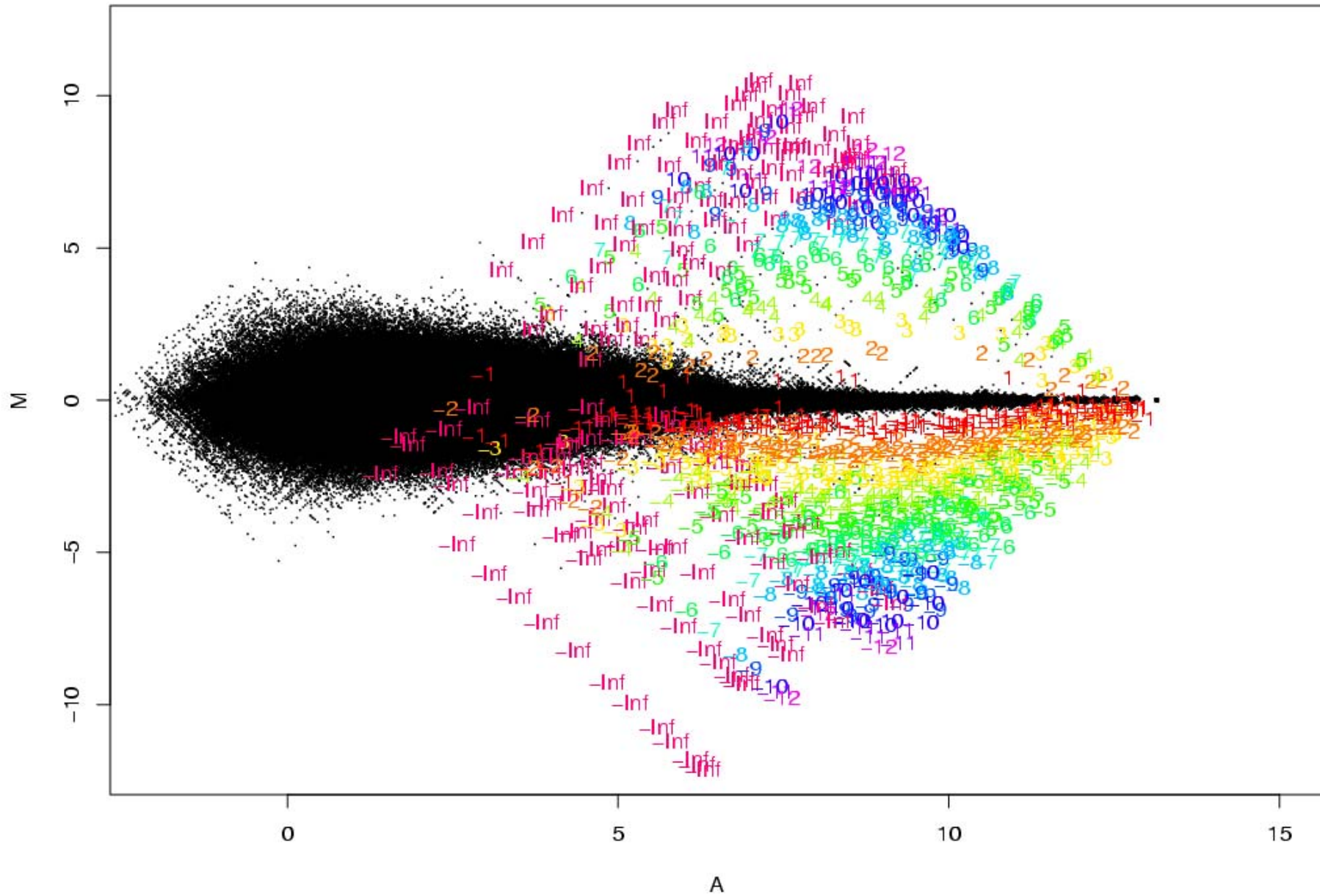
MAS 5 background



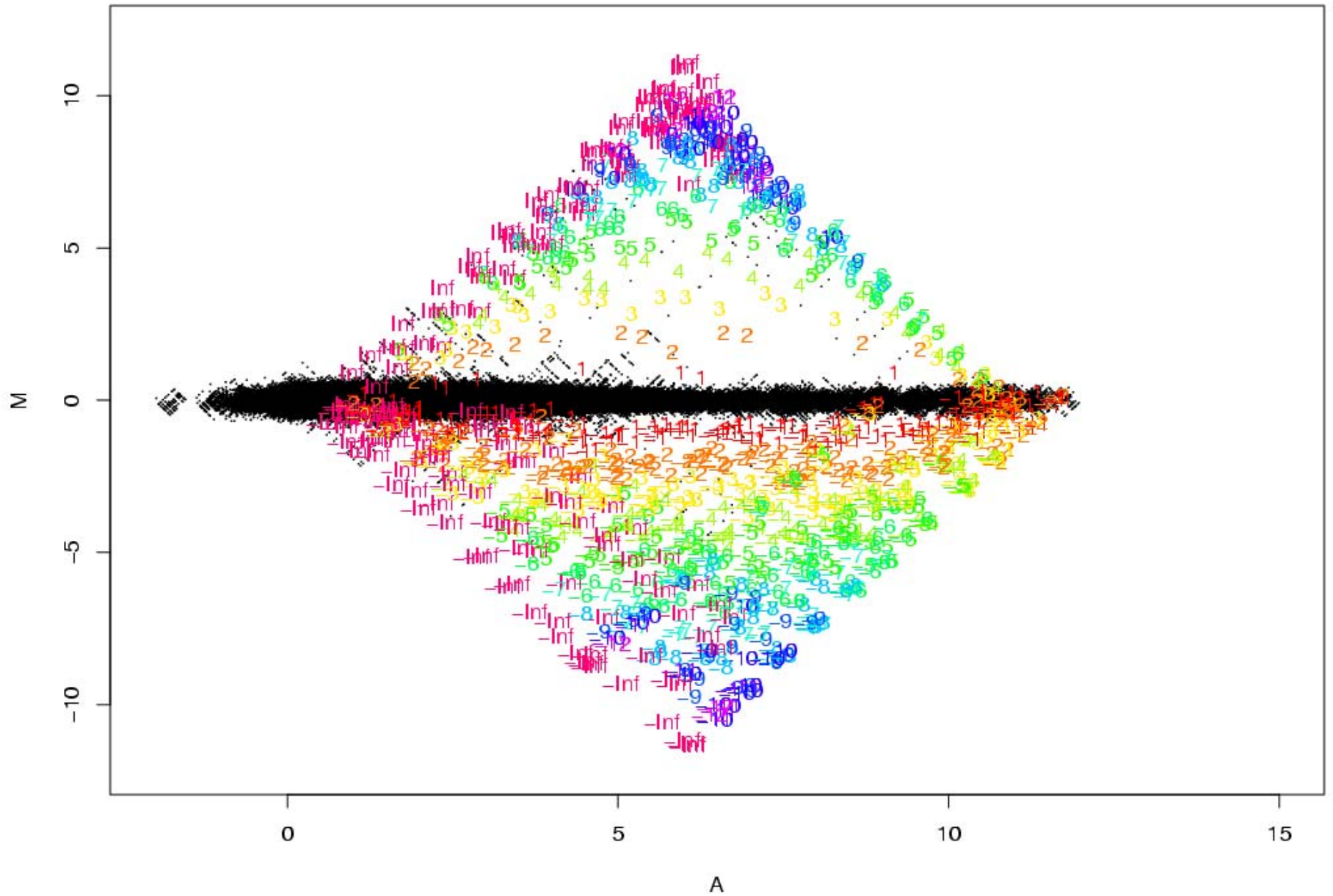
Ideal Mismatch



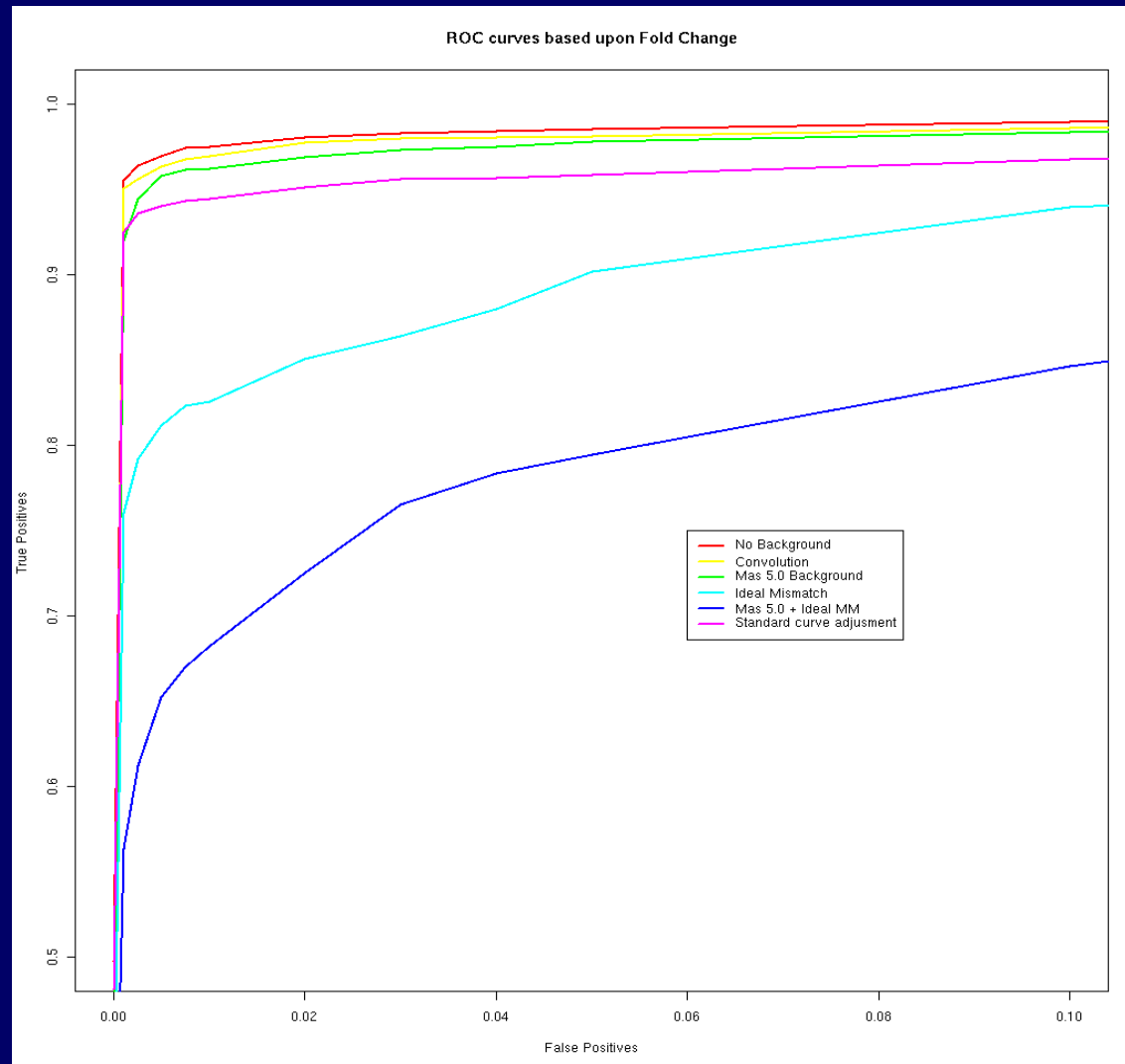
Mas5/Ideal Mismatch



Standard Curve Adjustment



Detecting Differential Expression: ROC Curves



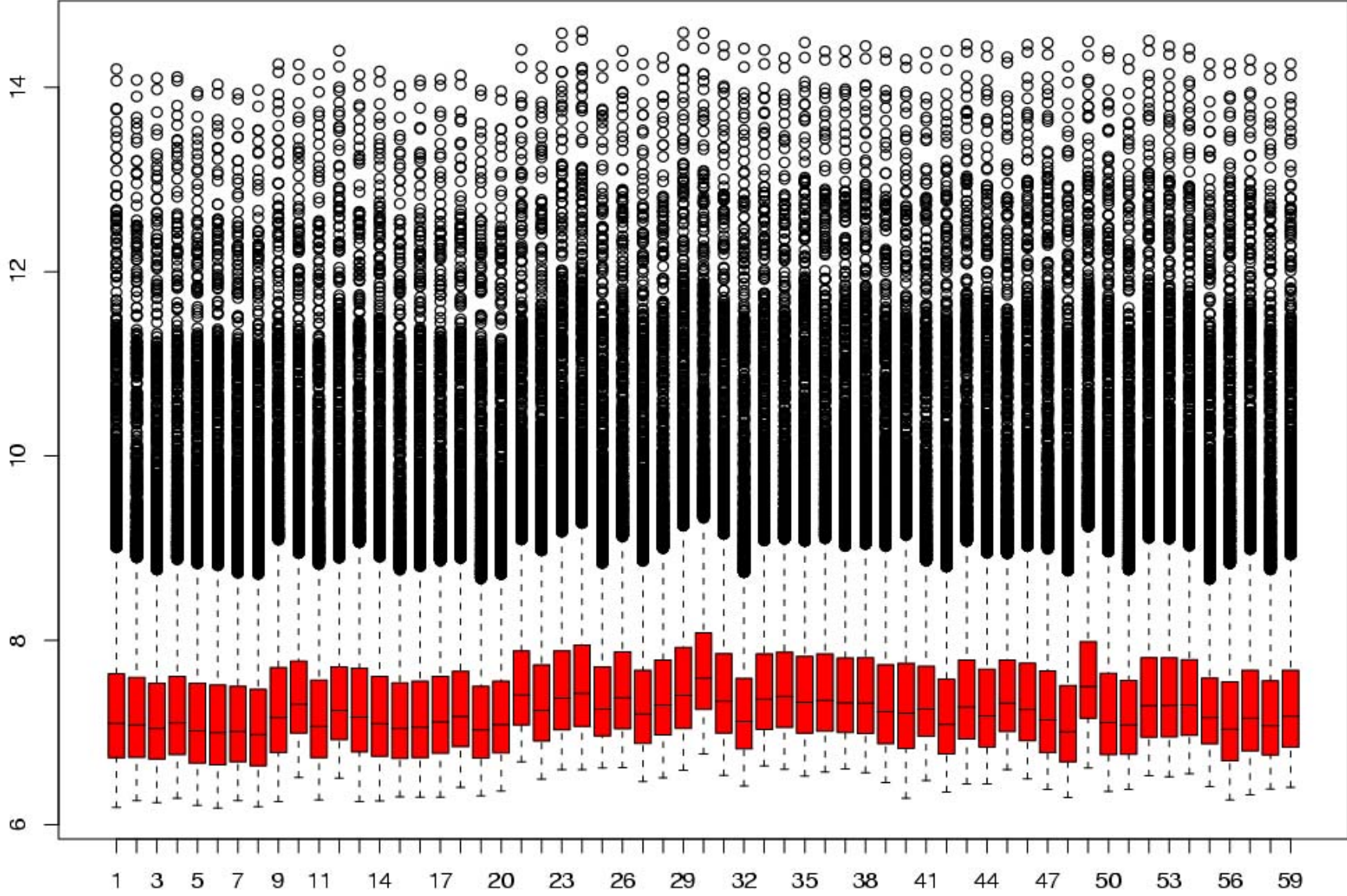
Summary of Trade-offs

Background Method	Detect Differential Genes	Accurate estimates of Fold change
No Background	Good	Poor
RMA	Good	Poor
MAS 5.0	Good	Poor
Ideal Mismatch	Poor	Good
MAS5.0/IdealIMM	Poor	Good
Standard Curve Adjustment	Good	Good

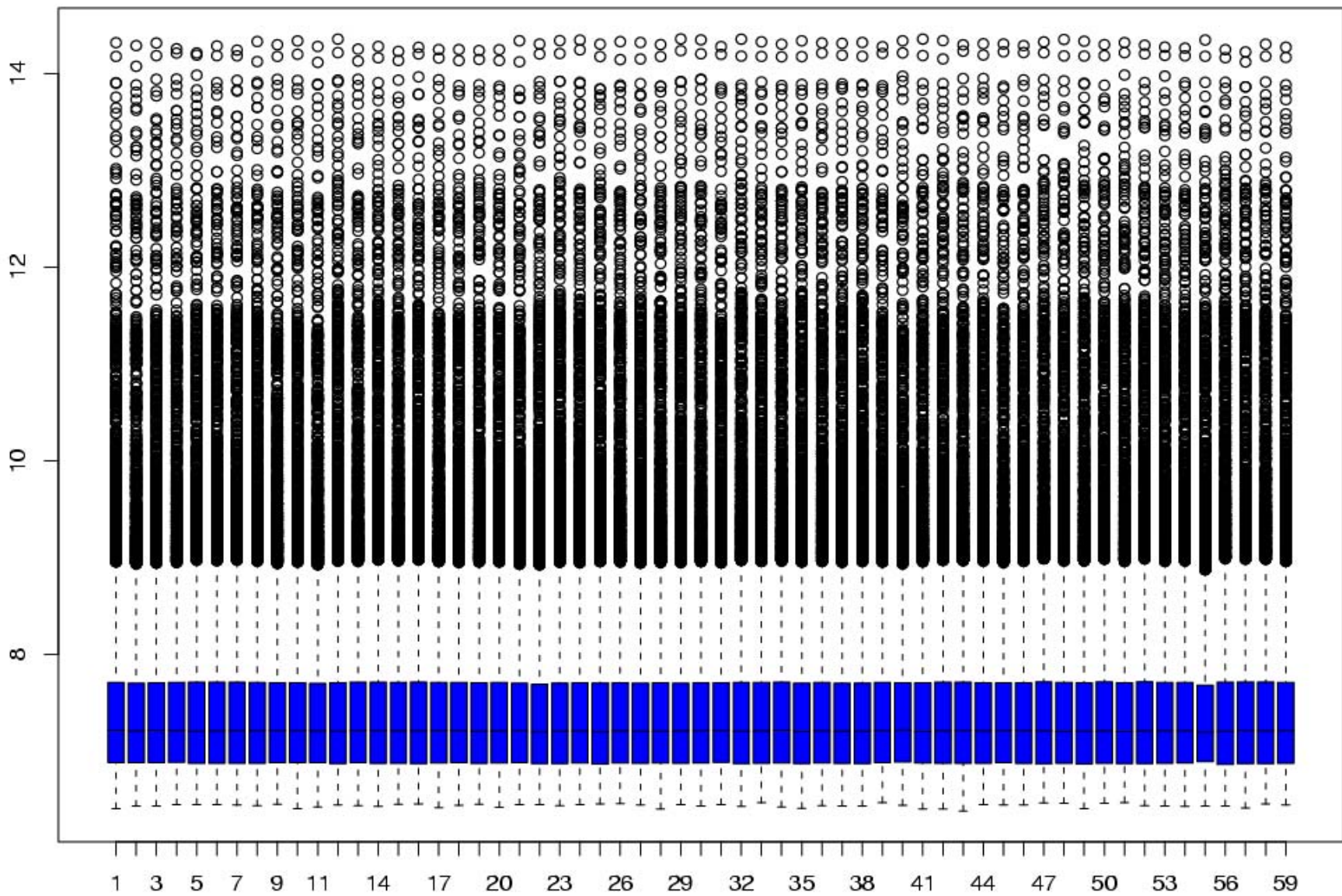
Comparing the Normalization Methods

- Want to reduce variation but at the same time we do not want to introduce any bias
- First a quick examination of the expression values by array
- Using same spike-in experiment as before, this time no background correction, only normalization and median polish summarization.

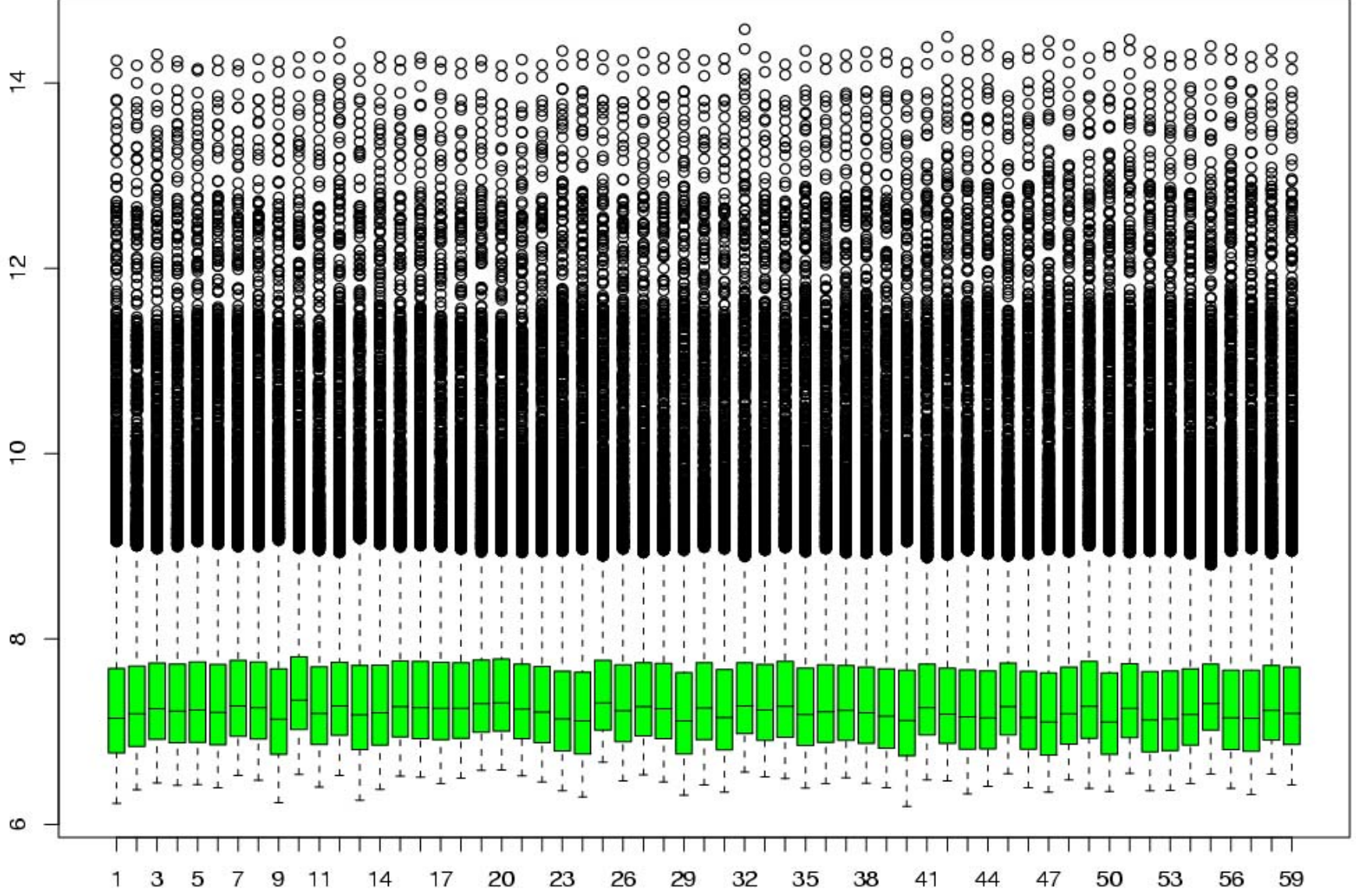
No Normalization



Quantiles

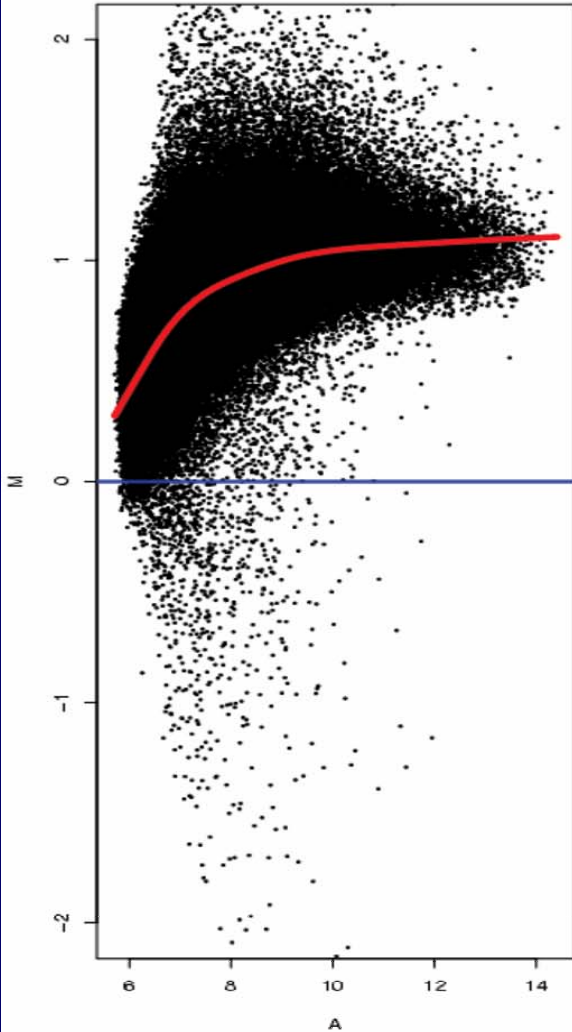


Scaling

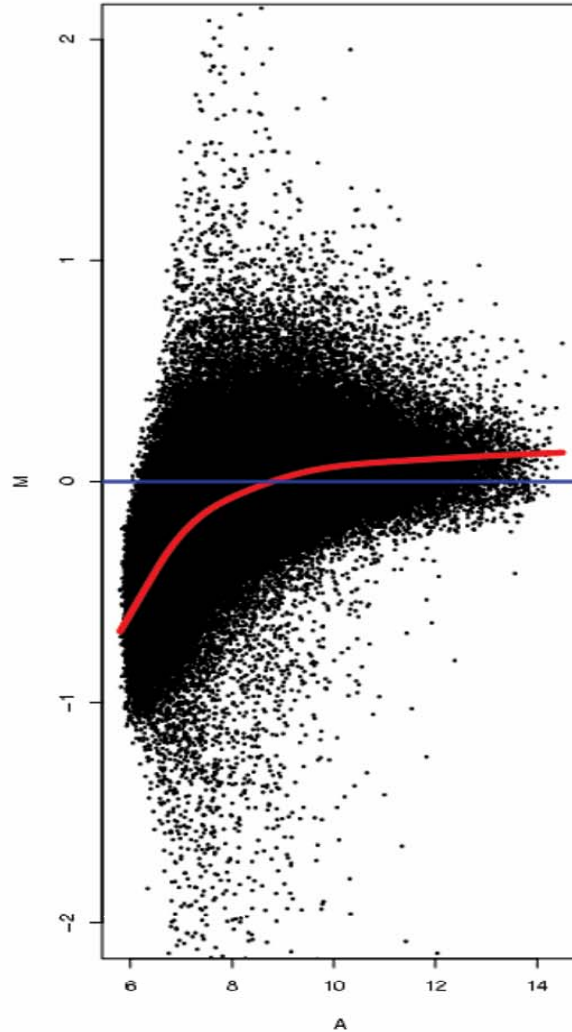


Scaling is Not Sufficient

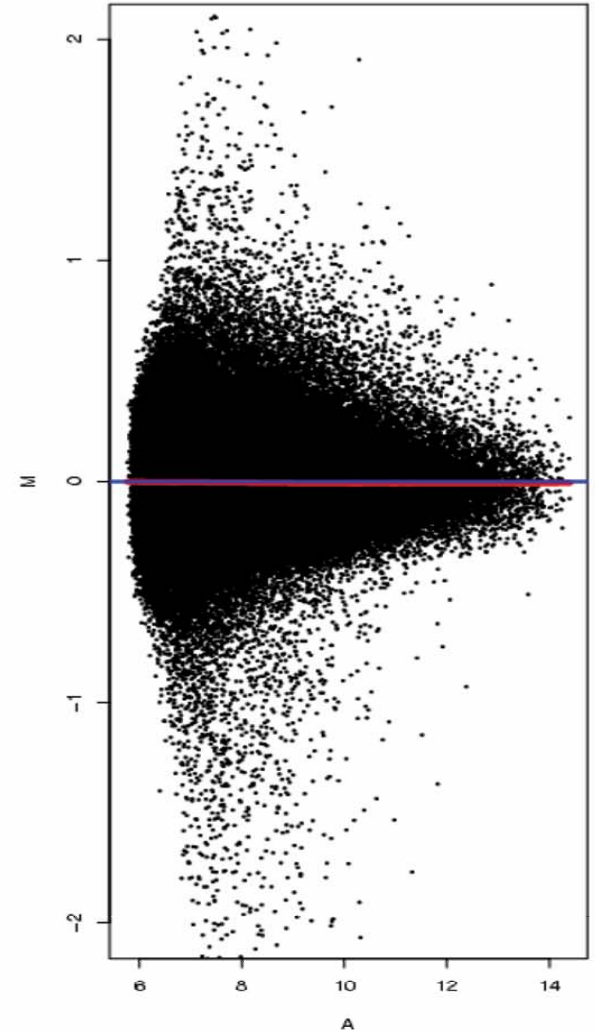
(a) Unnormalized



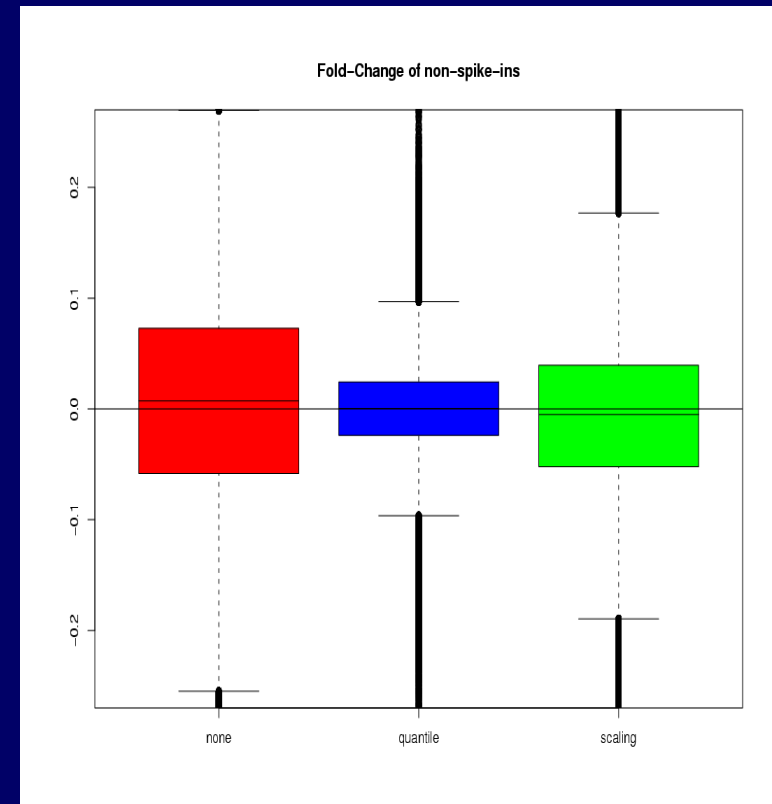
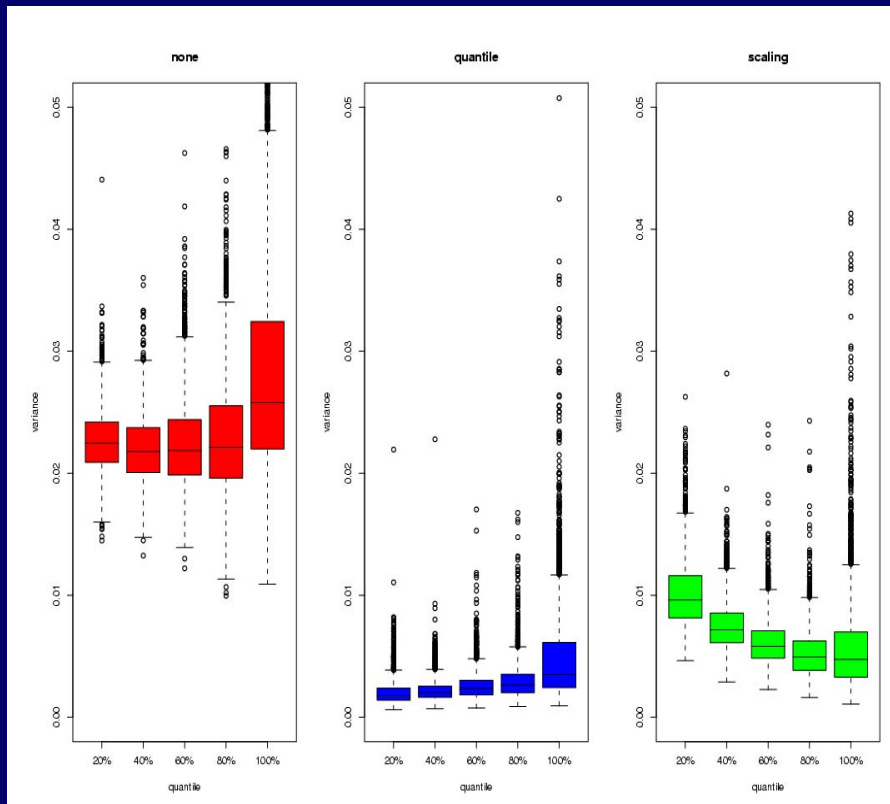
(b) Scaling



(c) Quantile Normalized



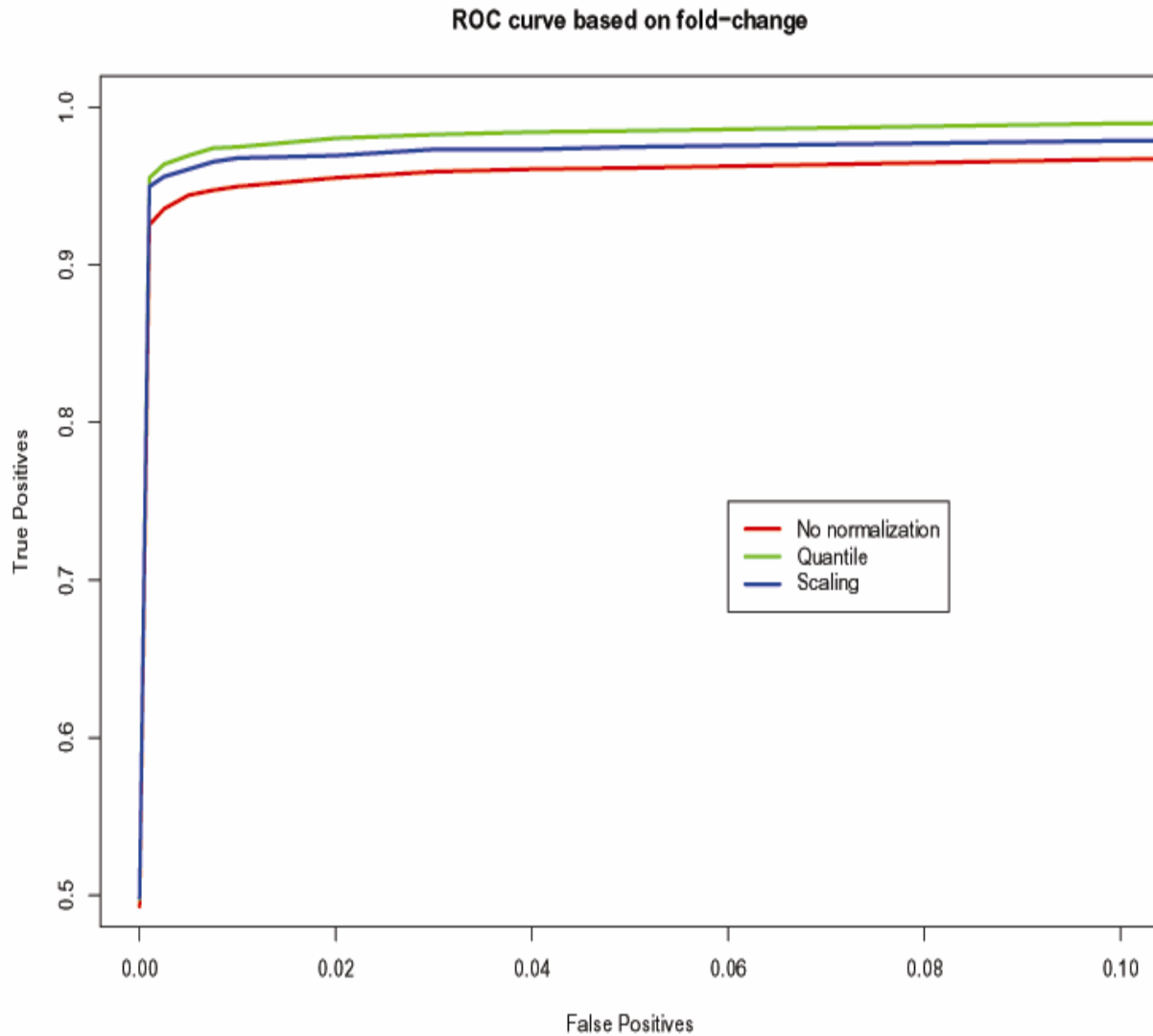
Variability of Non-Differential Genes is Reduced



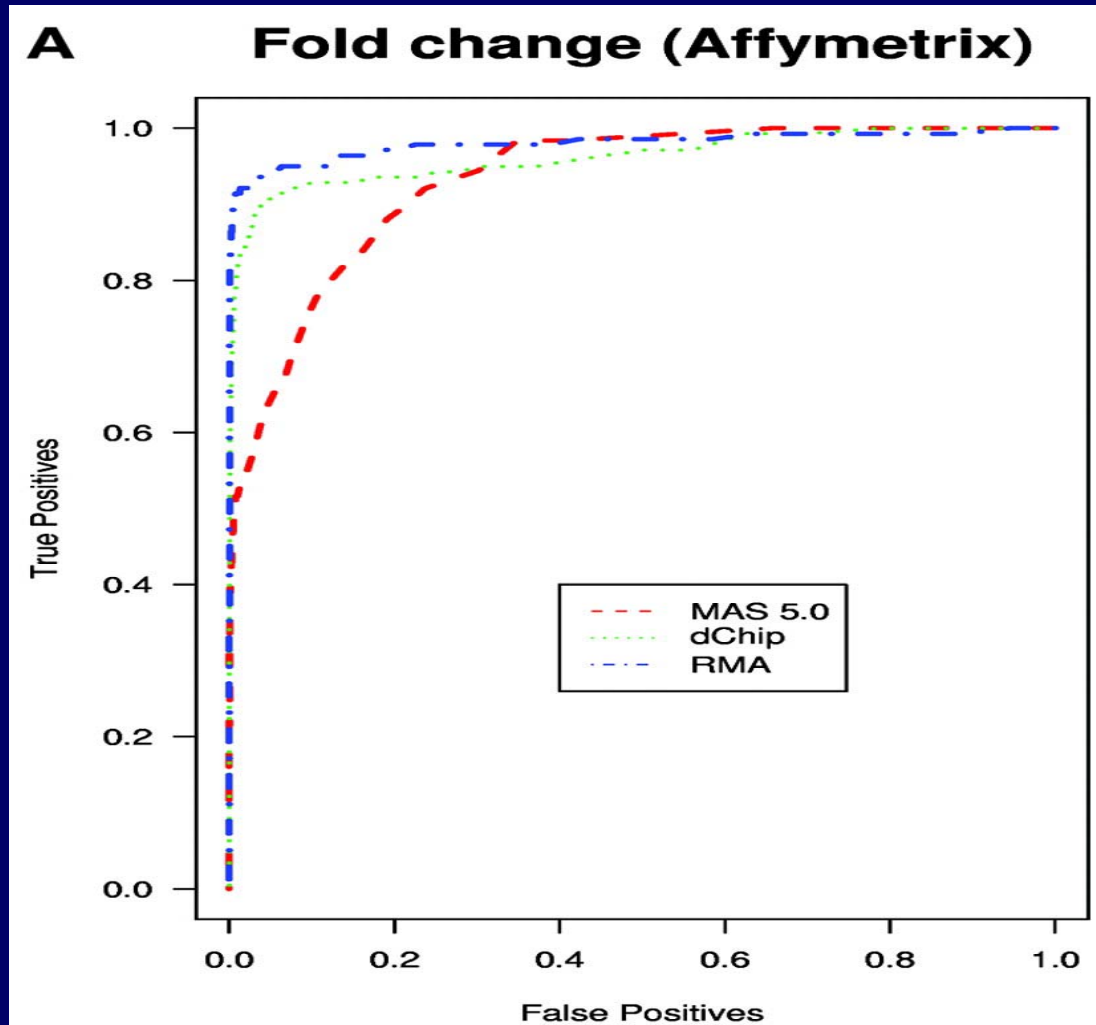
Little effect on Spike-ins

Method	All	Low	Mid	High	FC
No Normalization	0.493 (0.845)	0.185 (0.148)	0.664 (0.733)	0.328 (0.207)	0.484 (0.952)
Quantile	0.493 (0.851)	0.184 (0.153)	0.665 (0.741)	0.329 (0.224)	0.484 (0.955)
Scaling	0.493 (0.852)	0.186 (0.156)	0.663 (0.742)	0.33 (0.225)	0.484 (0.954)

ROC Curves



Comparing Established Expression Measures



Probe Level Models for Detection of Differential Expression

General Probe Level Model

$$y_{ij} = f(\mathbf{X}) + \varepsilon_{ij}$$

- Where $f(\mathbf{X})$ is a linear function of factor (and possibly covariate) variables
- Assume that $E[\varepsilon_{ij}] = 0$

$$\text{Var}[\varepsilon_{ij}] = \sigma^2$$

$$y_{ij} = \log_2 \mathbf{N}\left(\mathbf{B}(PM_{ij})\right)$$

We Will Focus on Two Particular PLM

- Array effect model

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

- Treatment effect model

$$y_{ij} = \alpha_i + \tau_{l_j} + \varepsilon_{ij}$$

In both cases $\sum_{i=1}^I \alpha_i = 0$

Fitting the PLM

- Robust regression using M-estimation
- By default, we will use Huber's ψ
- Fitting algorithm is IRLS with weights $\frac{\psi(r)}{r}$
- Software for fitting such models is part of *affyPLM* package of Bioconductor

Variance Covariance Estimates

- Suppose model is $Y = X\beta + \varepsilon$
- Huber (1981) gives three forms for estimating variance covariance matrix

$$\kappa^2 \frac{1/(n-p) \sum_i \psi(r_i)^2}{\left[1/n \sum_i \psi'(r_i)\right]^2} (X^T X)^{-1}$$

$$\kappa \frac{1/(n-p) \sum_i \psi(r_i)^2}{1/n \sum_i \psi'(r_i)} W^{-1}$$

$$\frac{1}{\kappa} 1/(n-p) \sum_i \psi(r_i)^2 W^{-1} (X^T X) W^{-1}$$

We will use this form

$$W = X^T \Psi' X$$

Fold Change

$$FC = \bar{X}_l - \bar{X}_m$$

Where

$$\bar{X}_l = \frac{\sum \beta_j \text{Ind}(j \in \text{group } l)}{\sum \text{Ind}(j \in \text{group } l)}$$

Simple t-statistic

$$t = \frac{\bar{X}_l - \bar{X}_m}{\sqrt{\frac{s_l^2}{n_l} + \frac{s_m^2}{n_m}}}$$

“Robust” t-statistic

$$t = \frac{\tilde{X}_l - \tilde{X}_m}{\sqrt{\frac{\tilde{S}_l^2}{n_l} + \frac{\tilde{S}_m^2}{n_m}}}$$

- Use medians in place of means
- Use MAD in place of standard deviation

Simple Moderated t-Statistic

$$t = \frac{\bar{X}_l - \bar{X}_m}{\sqrt{\frac{s_l^2}{n_l} + \frac{s_m^2}{n_m} + S_{\text{med}}}}$$

- S_{med} is median $\sqrt{\frac{s_l^2}{n_l} + \frac{s_m^2}{n_m}}$ across all genes

Limma “ebayes” t-statistic

- Generalization of Bayesian method of Lonnstadt and Speed (2002) to the general linear model case
- An alternative and much more sophisticated moderated t-statistic

Probe Level Model test statistics

- Suppose that Σ is component of the variance-covariance matrix related to β
- Let c be the contrast vector defined such that the j th element of c is $1/n_l$ if array j is in group l and $-1/n_m$ if array j is in group m , 0 otherwise

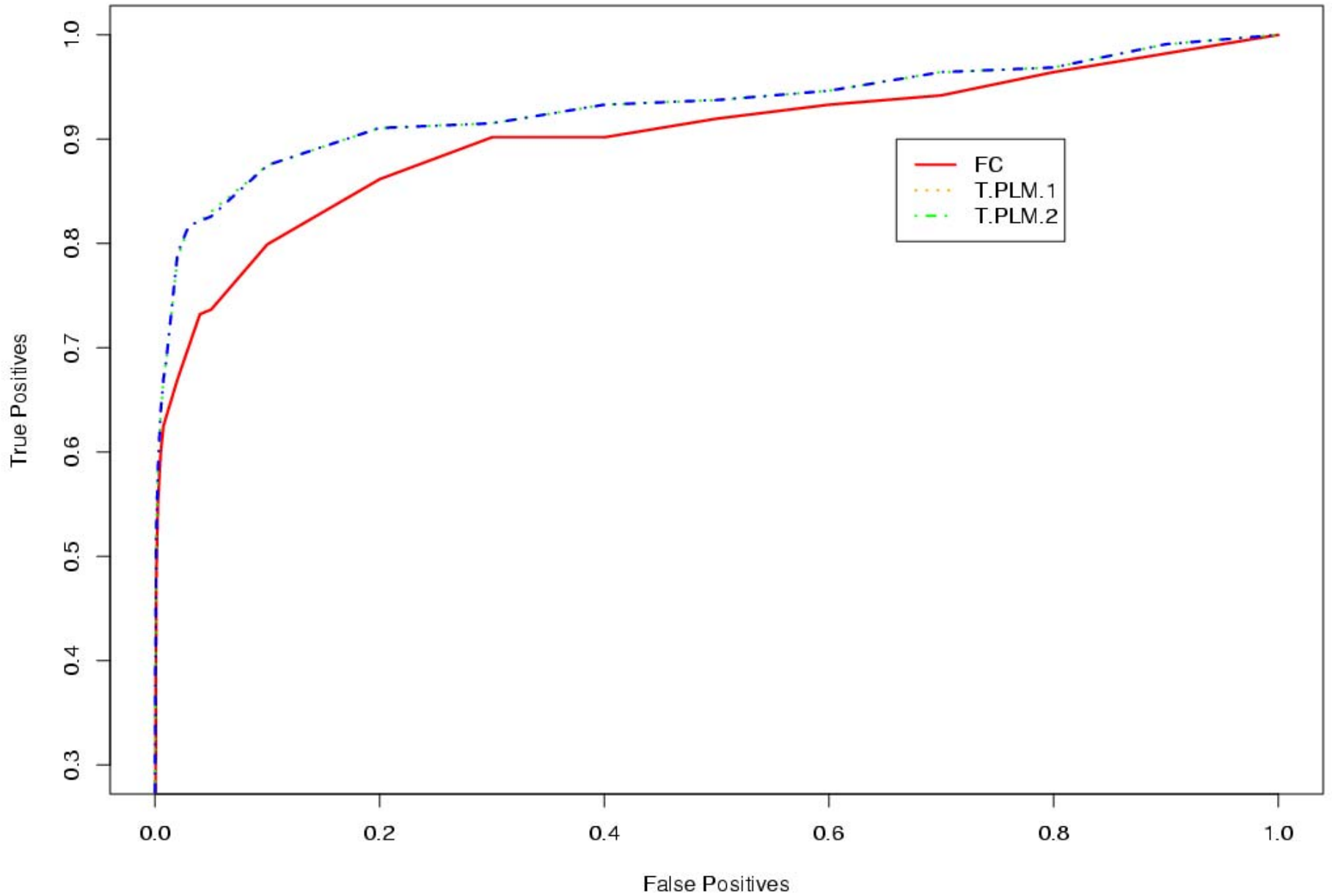
Probe Level Model test statistics

$$t_{\text{PLM.1}} = \frac{\mathbf{c}^T \boldsymbol{\beta}}{\sqrt{\sum_{j=1}^J c_j^2 \Sigma_{jj}}} \quad t_{\text{PLM.2}} = \frac{\mathbf{c}^T \boldsymbol{\beta}}{\sqrt{\mathbf{c}^T \boldsymbol{\Sigma} \mathbf{c}}}$$

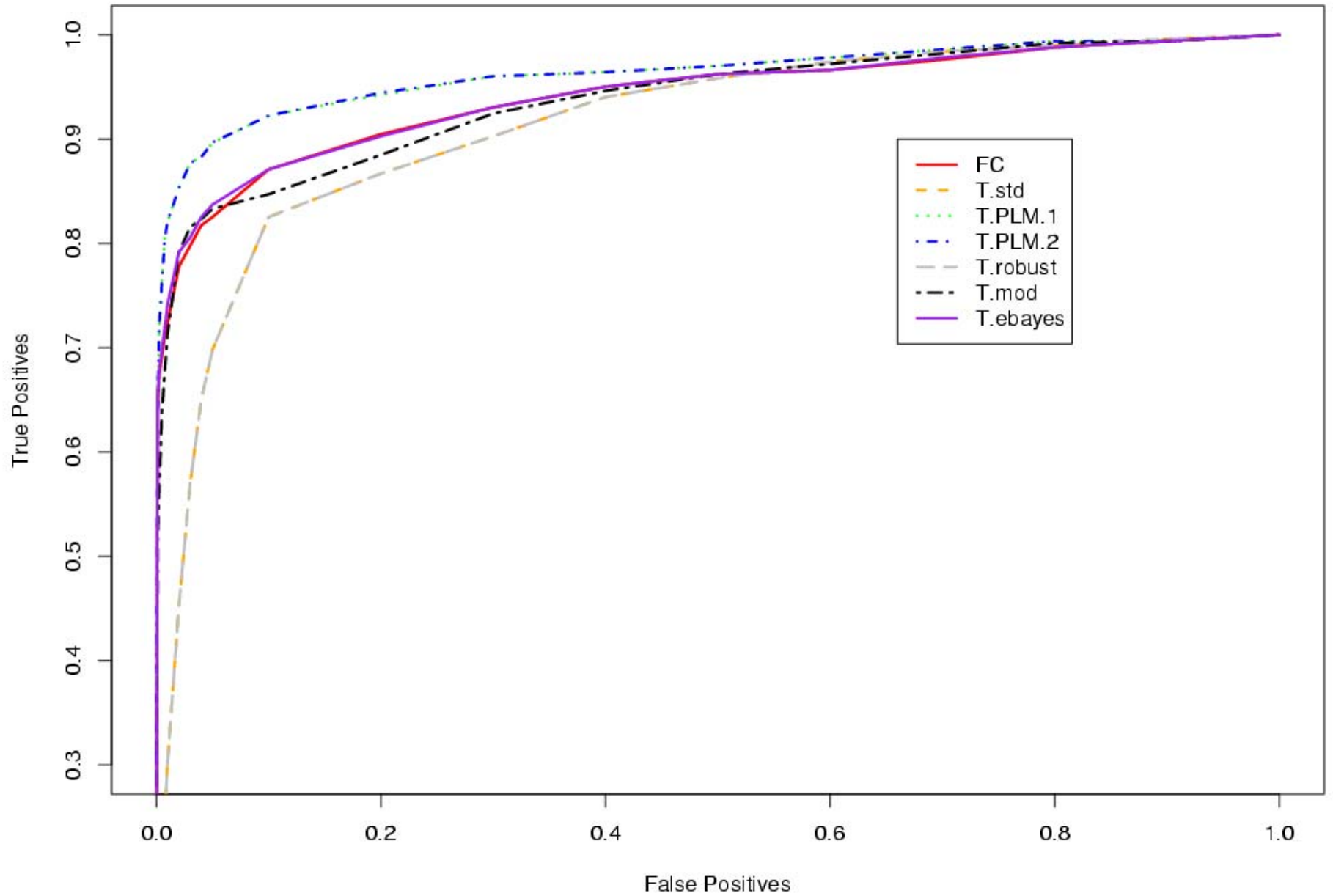
A First Comparison

- 8 chips from Affymetrix HG-U95A spike-in dataset
 - 4 arrays for each of two concentration profiles
- Fit an array effect model to all 8 chips
 - Compare the performance of the different methods by looking at all comparisons
 - 1 vs 1
 - 2 vs 2
 - 3 vs 3
 - 4 vs 4

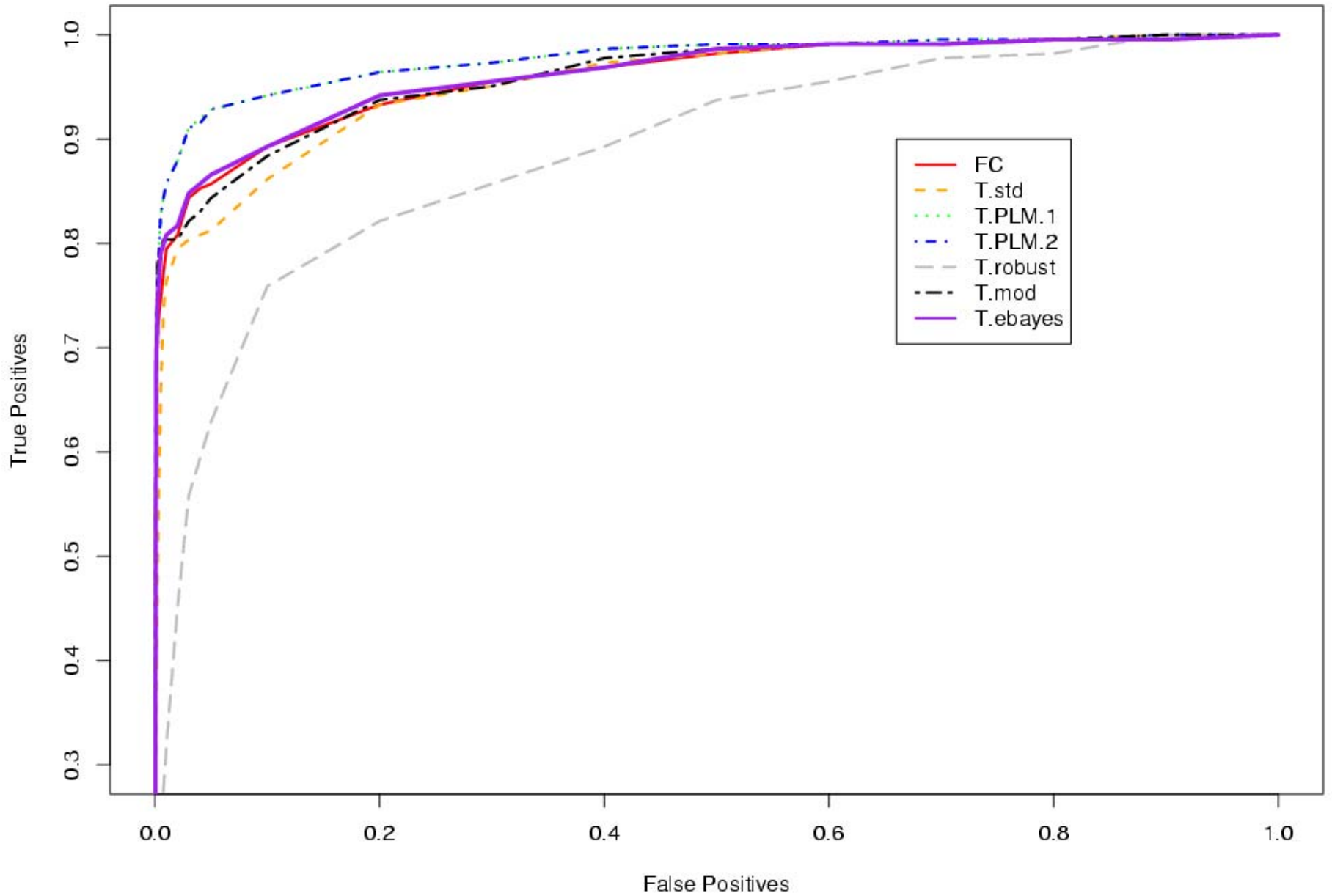
Affy Spikein (4 chips): 1 on 1



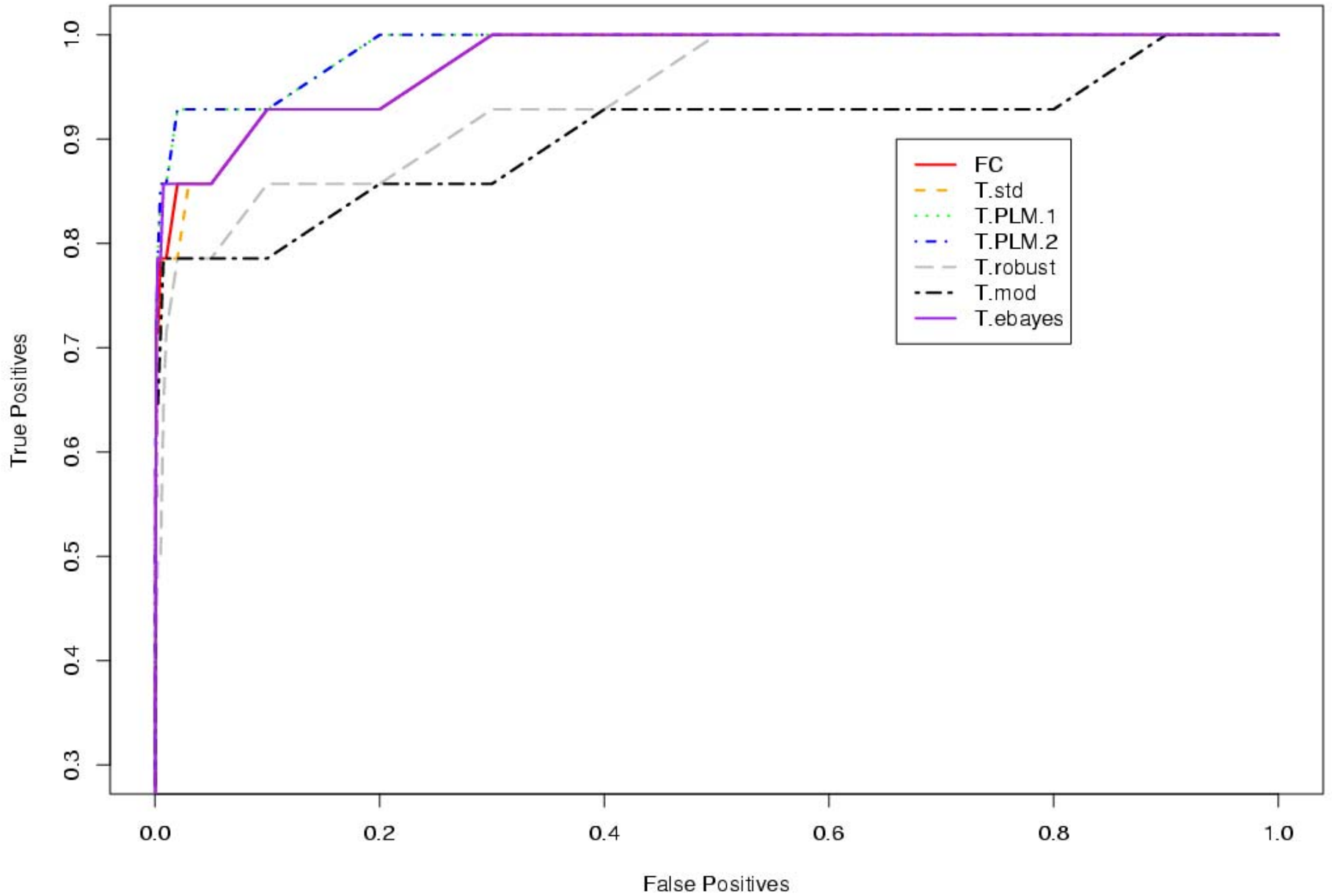
Affy Spikein: 2 on 2



Affy Spikein: 3 on 3



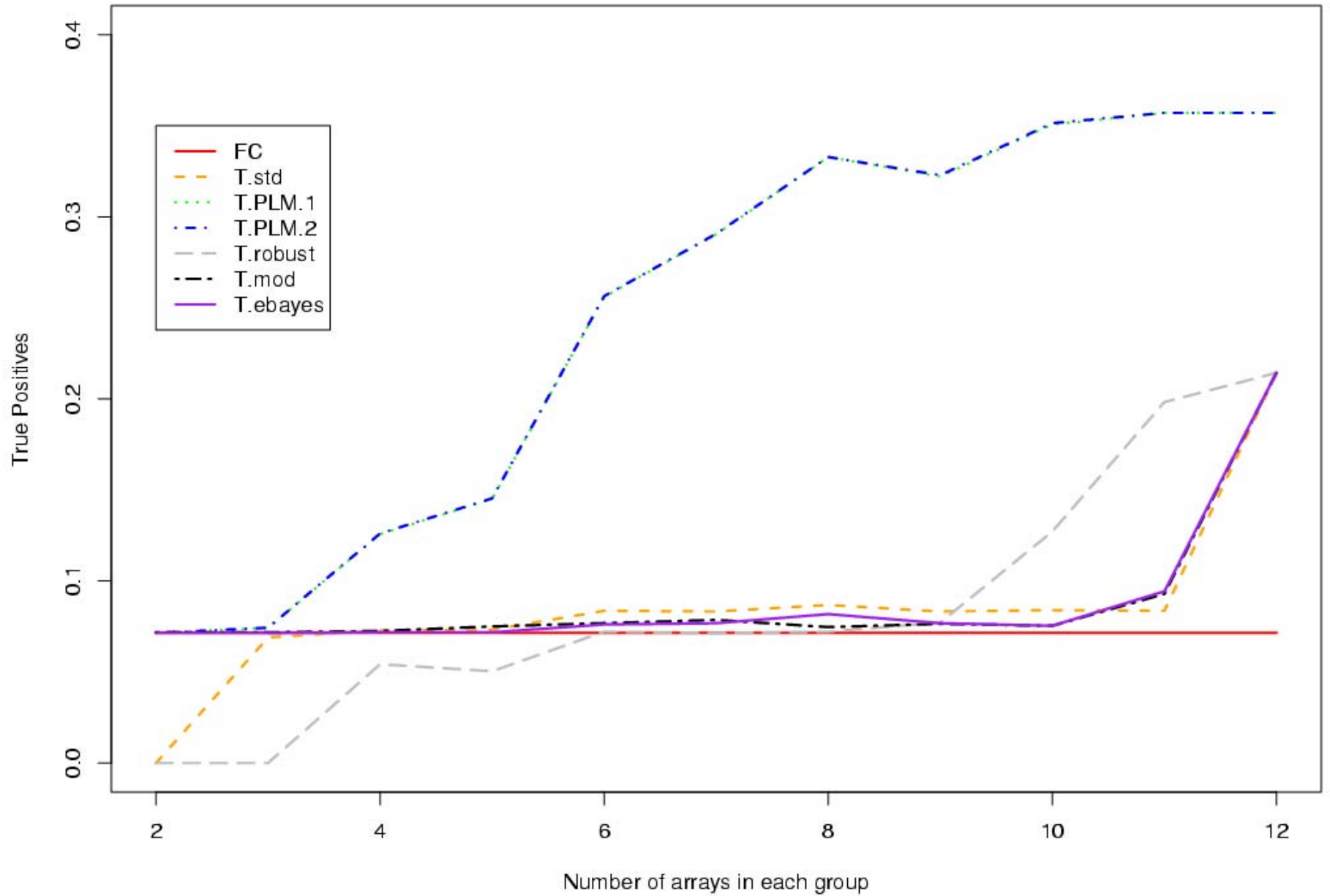
Affy Spikein: 4 on 4



What Happens as the Number of Arrays Increases?

- Expand comparison to all 24 Arrays with same concentration profiles from Affymetrix HG-U95A spike-in dataset
- Fit an array effect model to all 24 arrays
- Look at comparisons between equal number of chips

True positives when FP = 0



A Larger Comparison

- Look at the entire 59 chips for Affymetrix HG-U95A spike-in dataset
- Examine two cases. After standard preprocessing
 - Fit a model to all 59 chips
 - Fit models for each pairwise comparison
- There are 91 pairwise comparisons

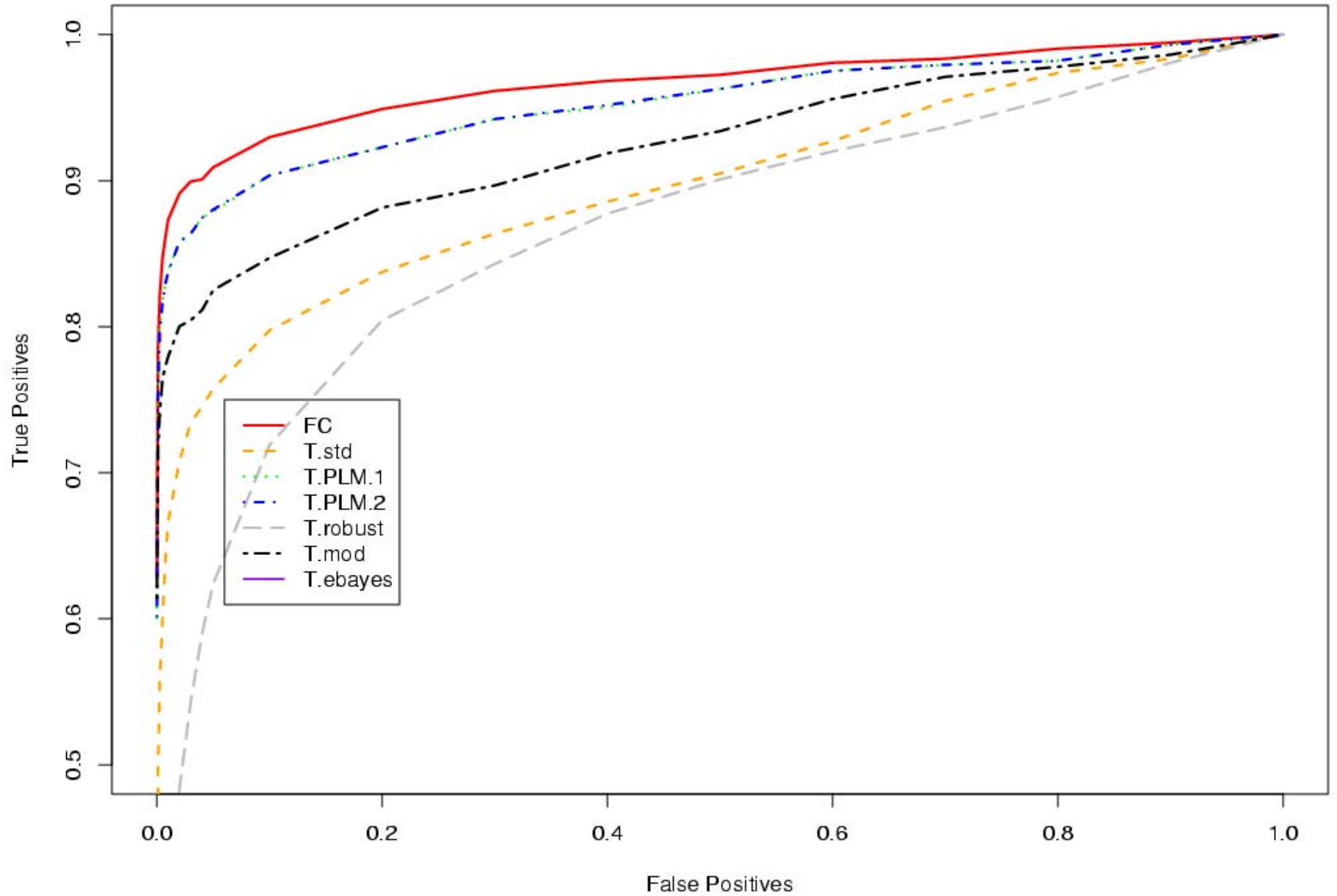
Results

Method	Individual Models			Single Model		
	0% FP	5% FP	AUC	0% FP	5% FP	AUC
FC	0.451	0.985	0.975	0.444	0.982	0.971
Std	0.323	0.982	0.956	0.301	0.975	0.952
Robust	0.16	0.939	0.857	0.144	0.935	0.852
Mod	0.437	0.987	0.975	0.413	0.98	0.97
PLM.1	0.653	0.991	0.979	0.54	0.951	0.93
PLM.2	0.657	0.991	0.979	0.539	0.951	0.93
Ebayes	0.514	0.988	0.978	0.45	0.986	0.974

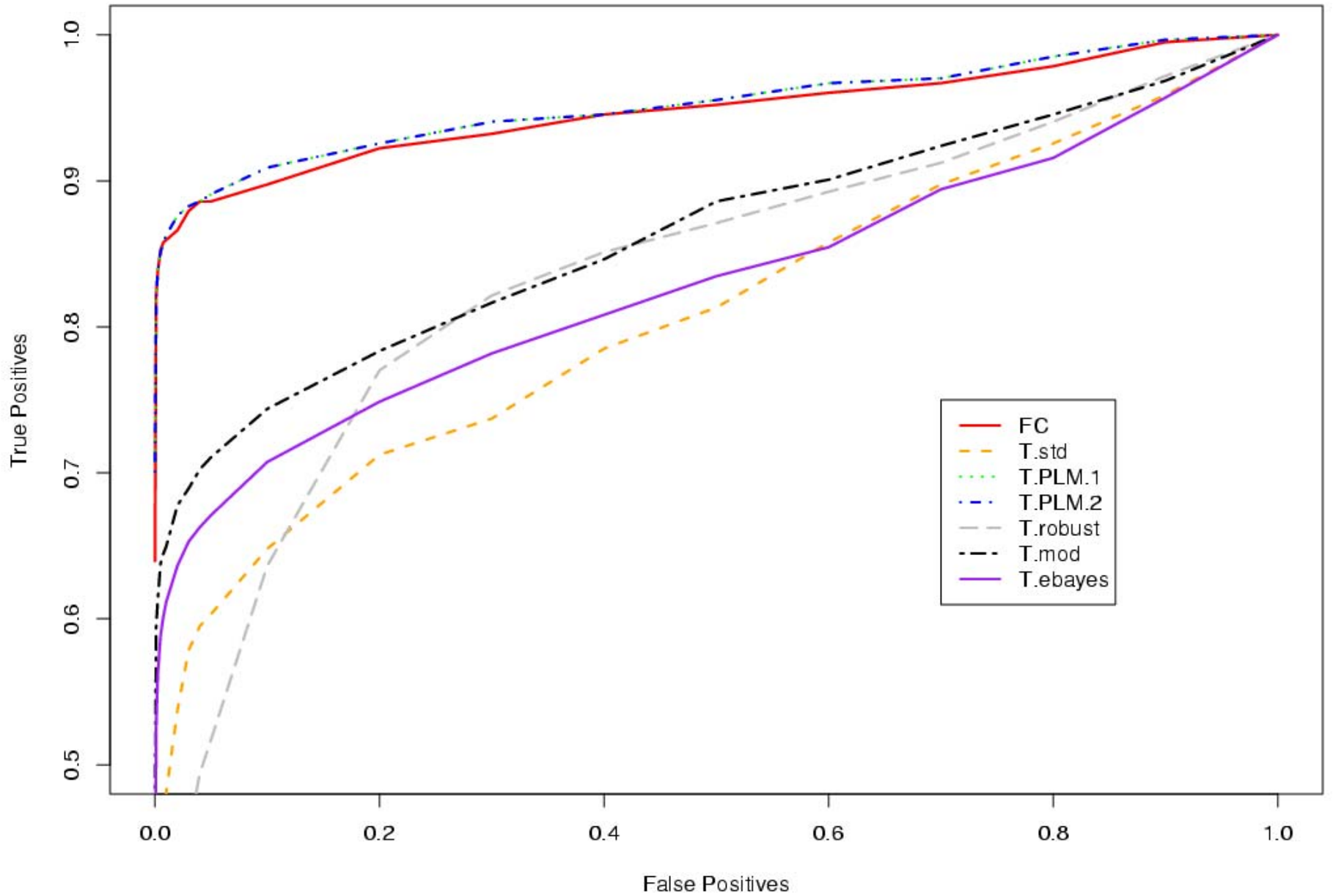
More Spike-in Datasets

- Two GeneLogic Spike-in datasets
 - AML dataset (34 arrays)
 - Tonsil dataset (36 arrays)
- In each case use single models fitted to all arrays

Genelogic Tonsil data



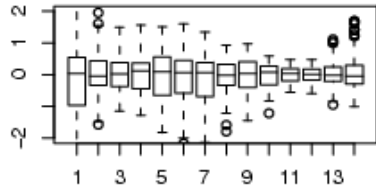
GL AML data



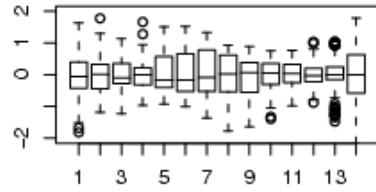
What is going on here?

- Examine residuals stratified by concentration group
 - Spike-ins
 - Randomly chosen non-differential probesets at low, medium and high average expression

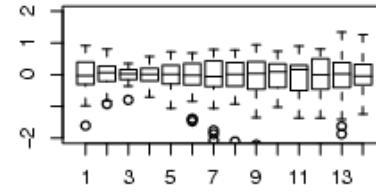
37777_at



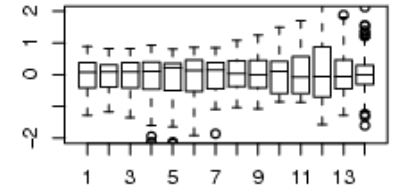
684_at



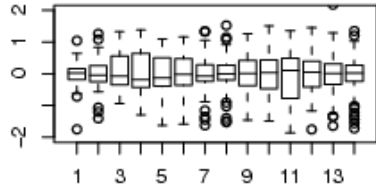
1597_at



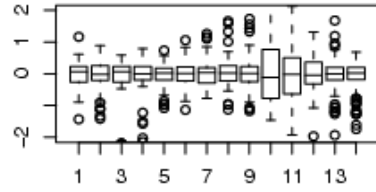
38734_at



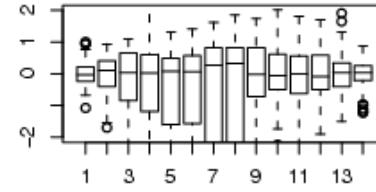
39058_at



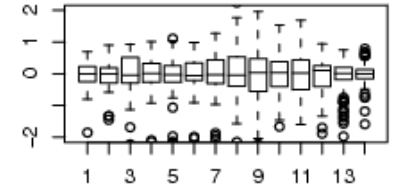
36311_at



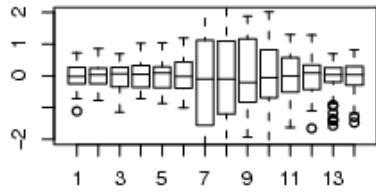
36889_at



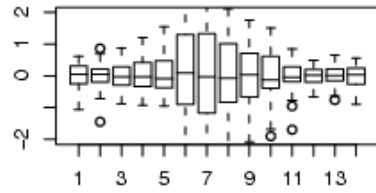
1024_at



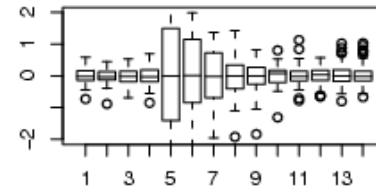
36202_at



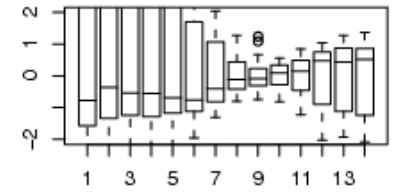
36085_at



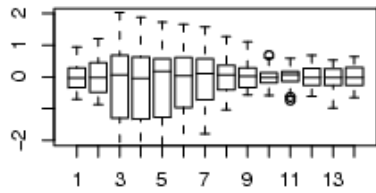
40322_at



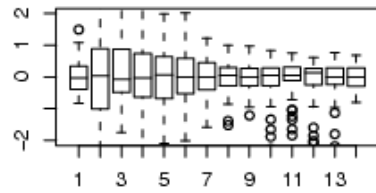
407_at



1091_at



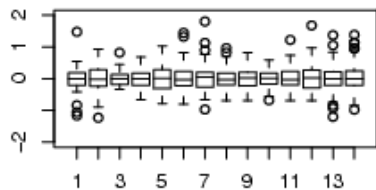
1708_at



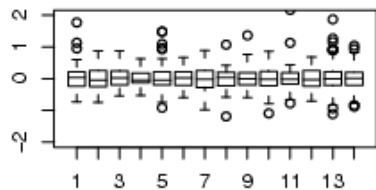
Affymetrix Spike-ins

Low Non-Differential

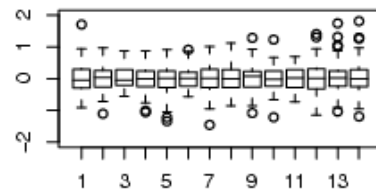
41128_at



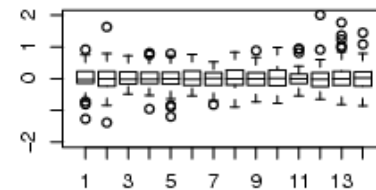
40430_at



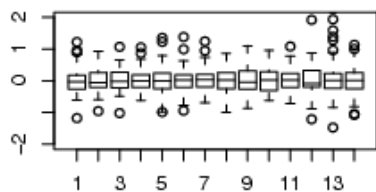
36087_at



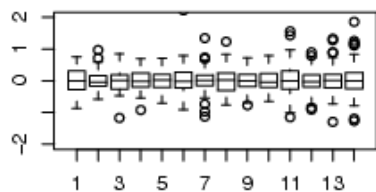
39040_at



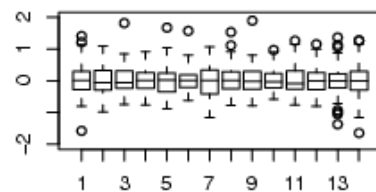
40686_at



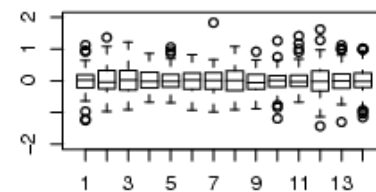
33147_at



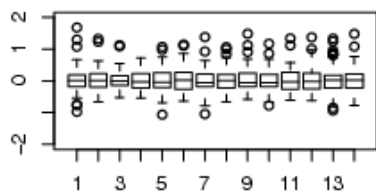
34517_at



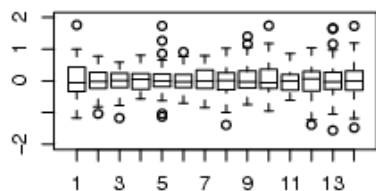
636_at



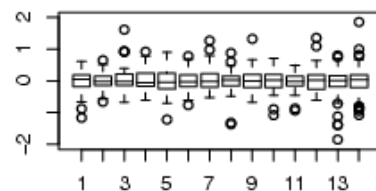
35681_r_at



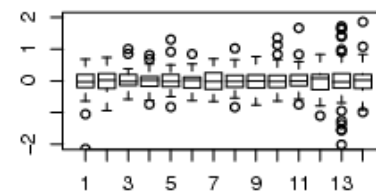
34483_at



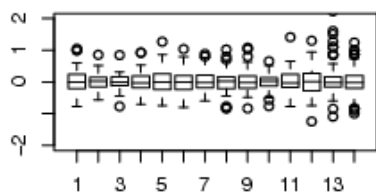
37373_at



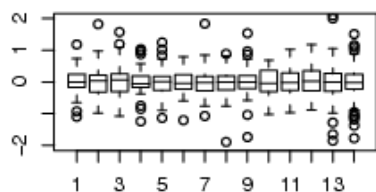
32005_at



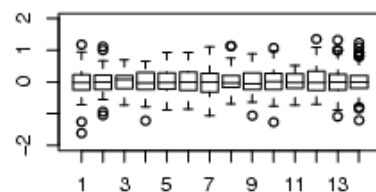
34471_at



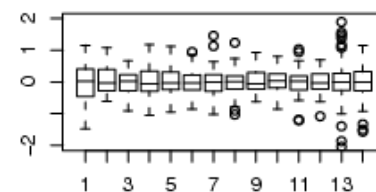
33351_at



40696_at

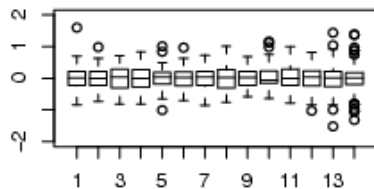


41716_at

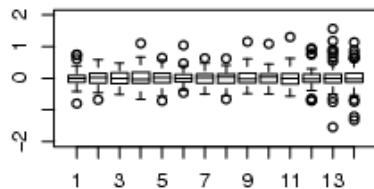


Middle Non-Differential

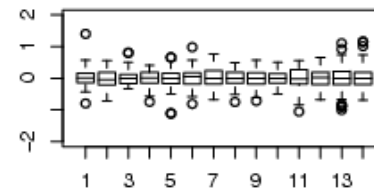
39638_at



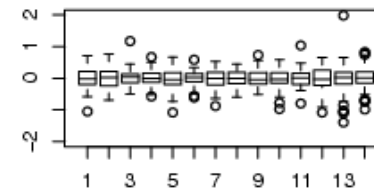
31406_at



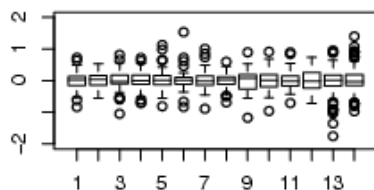
39196_i_at



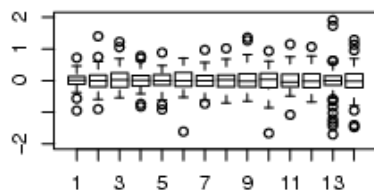
41233_at



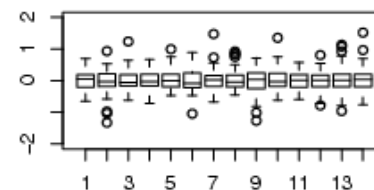
1248_at



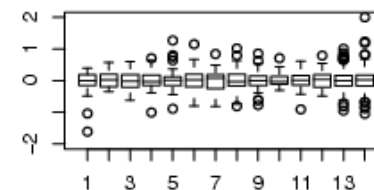
38968_at



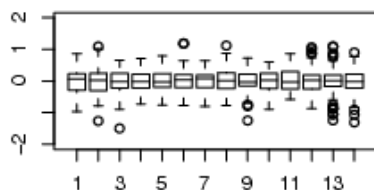
36264_at



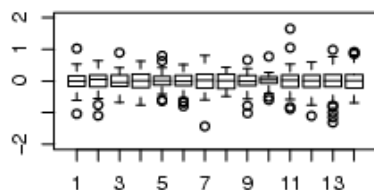
37417_at



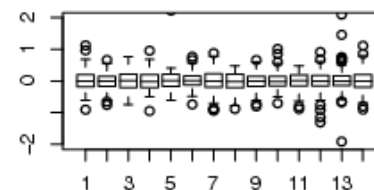
40110_at



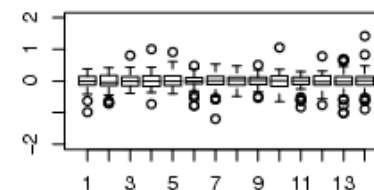
630_at



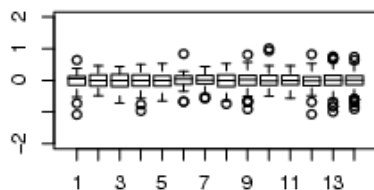
34147_g_at



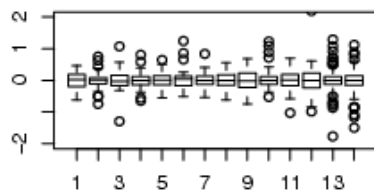
38592_s_at



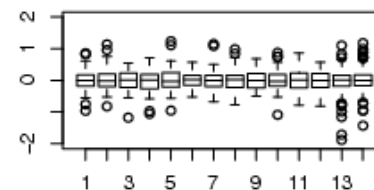
38950_r_at



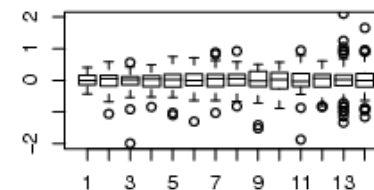
1650_g_at



40404_s_at

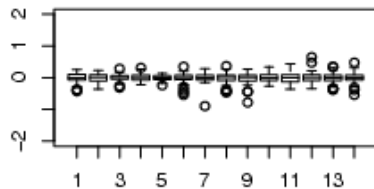


38609_at

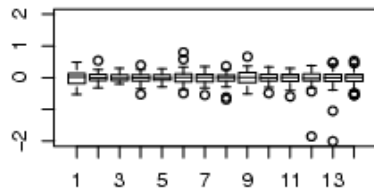


High Non-Differential

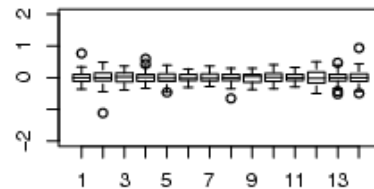
40887_g_at



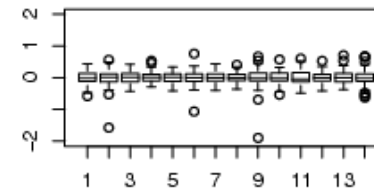
40886_at



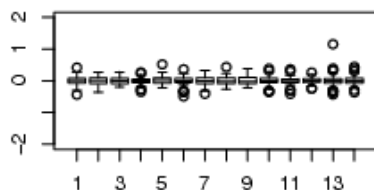
39473_r_at



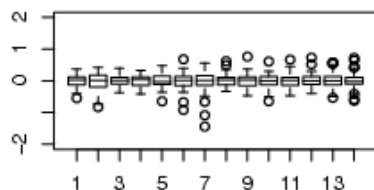
31538_at



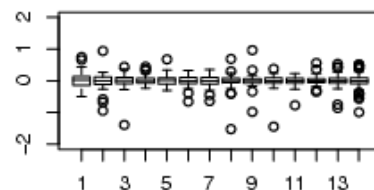
256_s_at



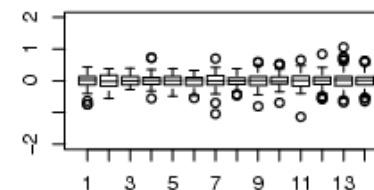
32438_at



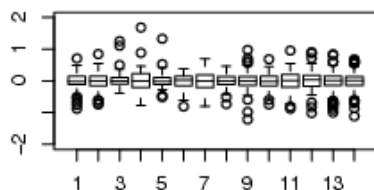
35905_s_at



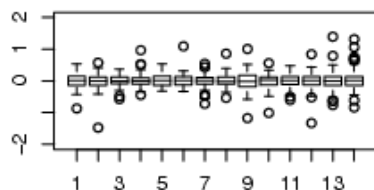
33660_at



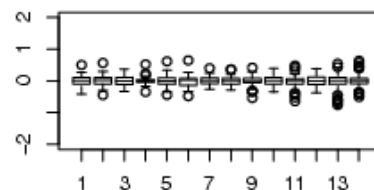
34085_at



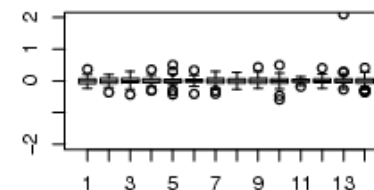
33677_at



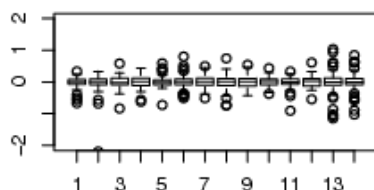
31623_f_at



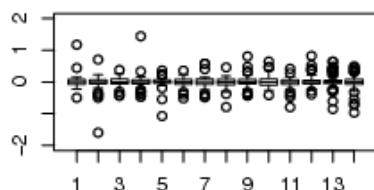
37746_r_at



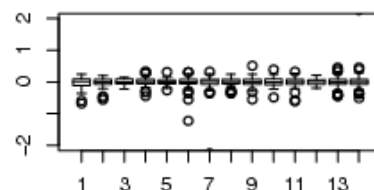
38061_at



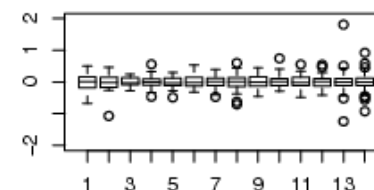
36546_r_at

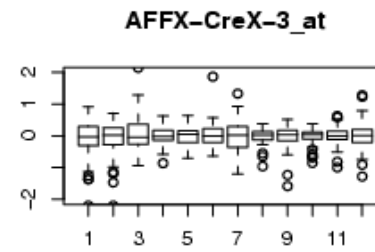
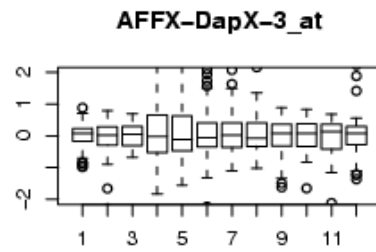
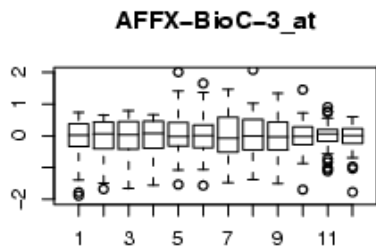
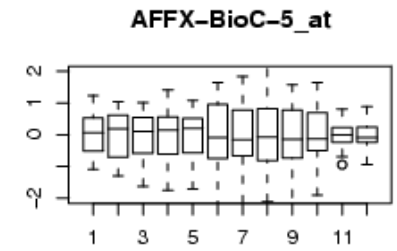
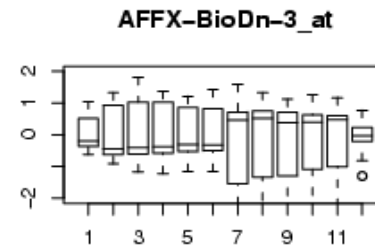
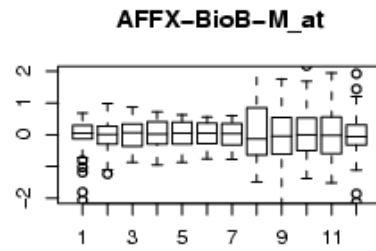
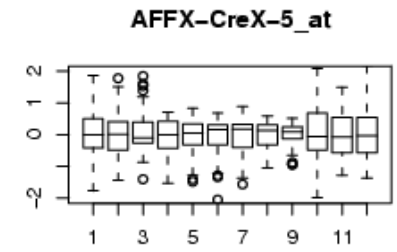
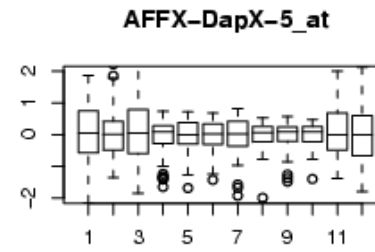
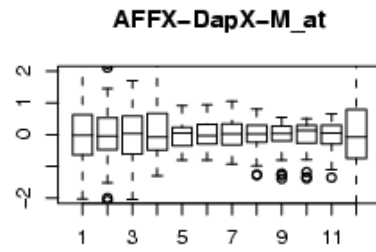
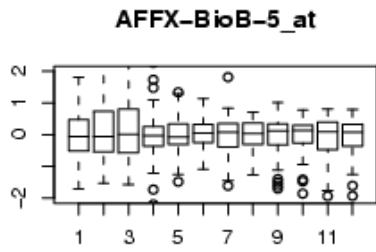


41210_at

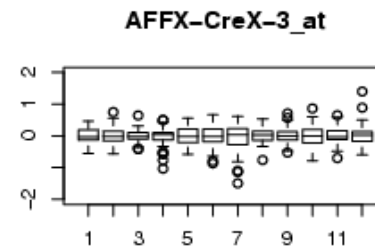
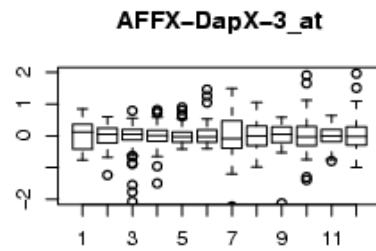
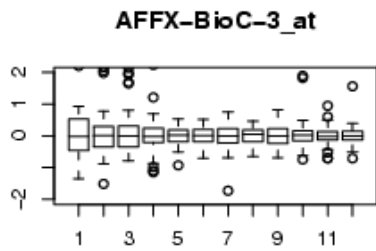
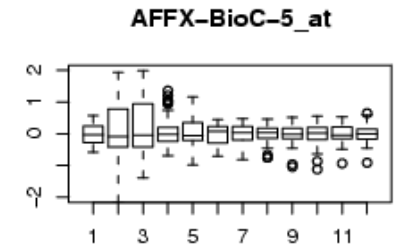
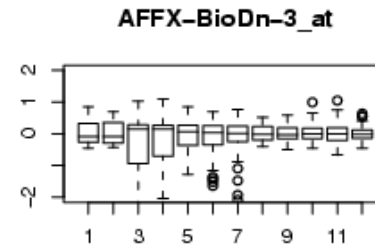
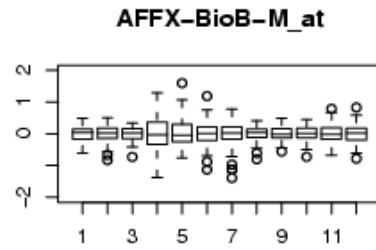
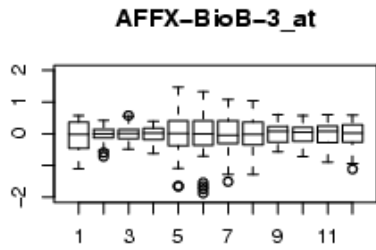
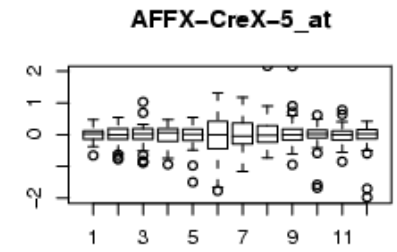
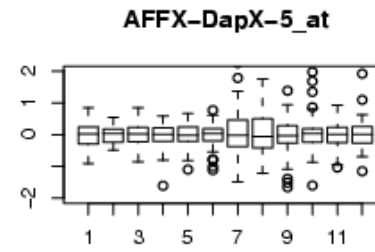
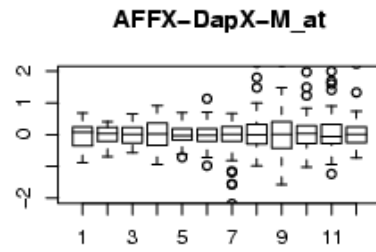
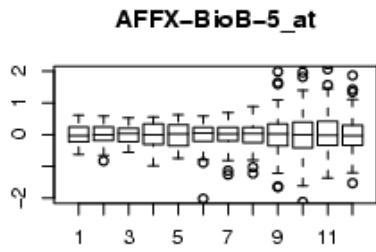


882_at





GeneLogic AML Spike-ins



GeneLogic Tonsil dataset

How About With More “Real” Data?

- Previous comparisons were for Spike-in data where only 11 or 14 probesets were expected to show any change between conditions. Consider GeneLogic Dilution/Mixture study. Using the 30 Liver and 30 CNS arrays to give a “truth”
- Use the 75:25 (5 arrays) and 25:75 (5 arrays) mixture arrays to test
- Choose 400 probesets with most extreme t-statistics from Dilution set to define “truth”

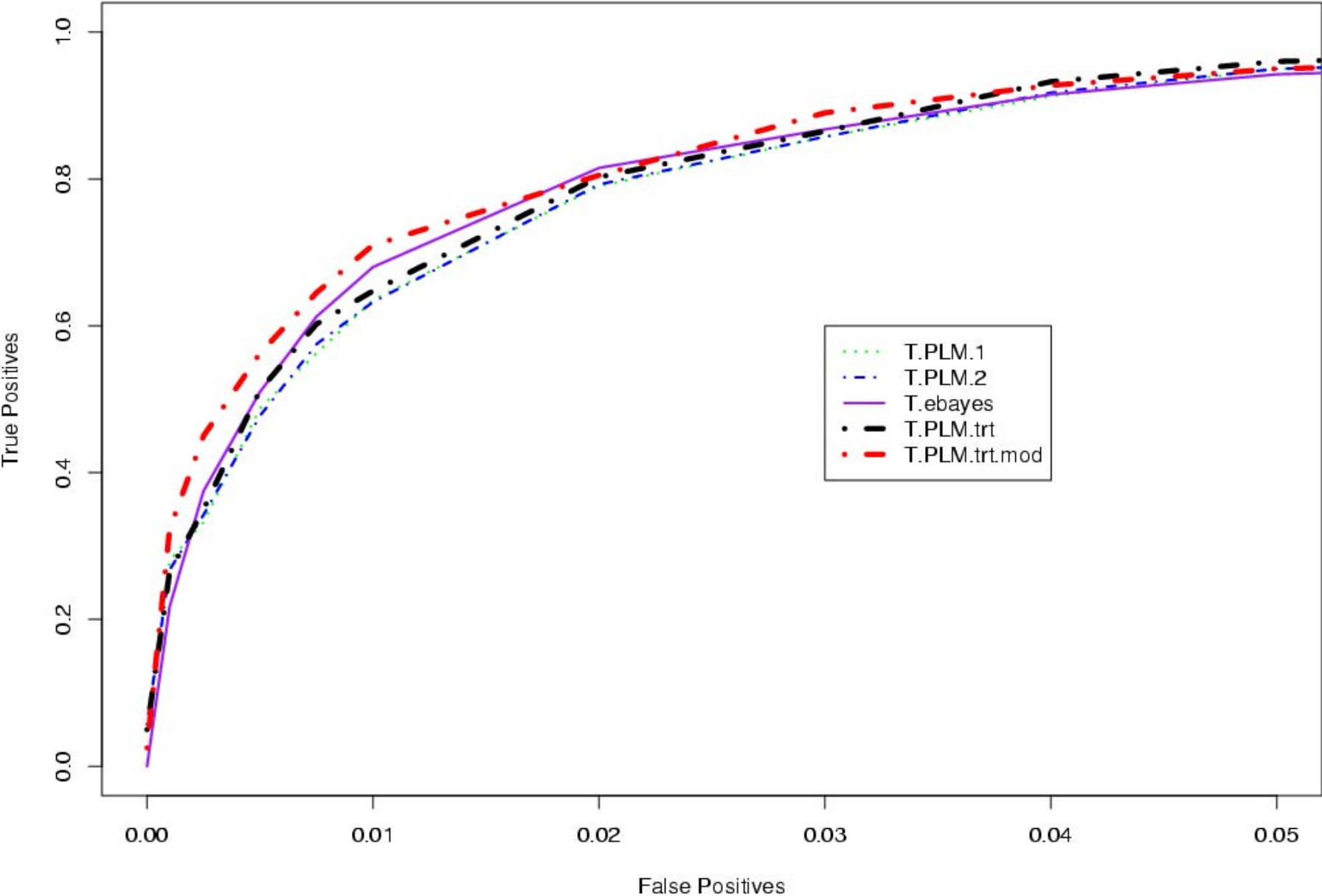
Results

Method	3 vs 3			4 vs 4			5 vs 5		
	0% FP	5% FP	AUC	0% FP	5% FP	AUC	0% FP	5% FP	AUC
FC	0.007	0.886	0.697	0.008	0.888	0.703	0.005	0.888	0.708
Std	0.004	0.793	0.53	0.008	0.872	0.626	0.018	0.902	0.675
Robust	0.002	0.485	0.271	0.005	0.747	0.49	0.01	0.743	0.488
Mod	0.007	0.908	0.697	0.002	0.932	0.735	0	0.948	0.76
PLM.1	0.056	0.943	0.751	0.057	0.947	0.756	0.056	0.95	0.76
PLM.2	0.057	0.943	0.752	0.057	0.948	0.758	0.058	0.95	0.761
Ebayes	0.001	0.918	0.744	0	0.933	0.761	0	0.943	0.776

What about the treatment effect model?

- The limma ebayes test statistic seems to be outperforming the PLM test statistics in the AUC statistic. Closer examination of ROC curve shows it exceeding all other methods between 0.25% and 2.5% false positives.
- Try the Treatment effect model with
 - PLM.2
 - PLM.2 with a simple moderation

Mixture data: 5 on 5



Ongoing work in this area

- Technology changes: what still works?
What doesn't?
- Better moderation for the PLM test statistic
- Other probe-level models

Acknowledgements

- Terry Speed (UC Berkeley)
- Francois Colin (UC Berkeley)
- Rafael Irizarry (Johns Hopkins)

- Bioconductor Core
<http://www.bioconductor.org>

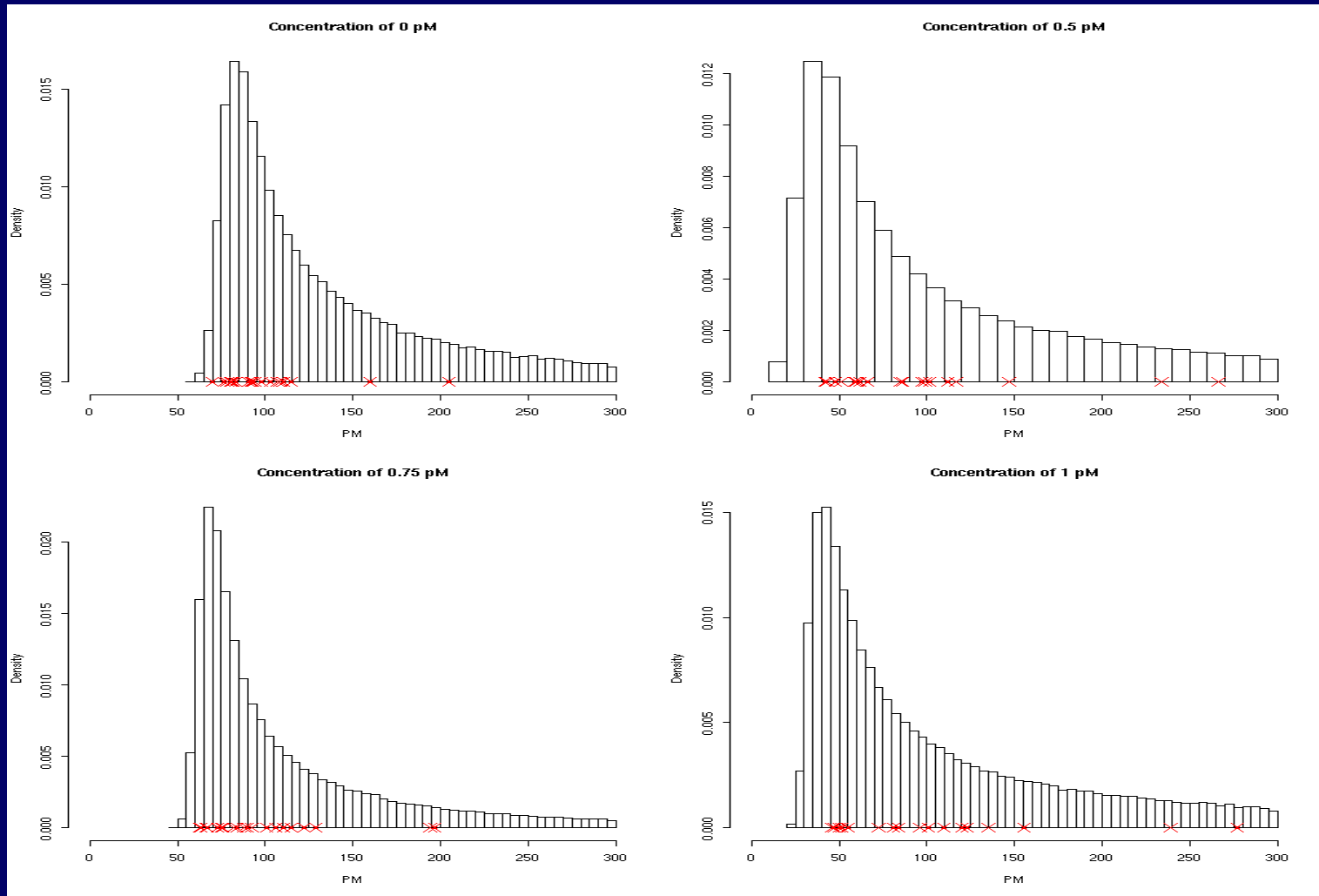
Additional Slides

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

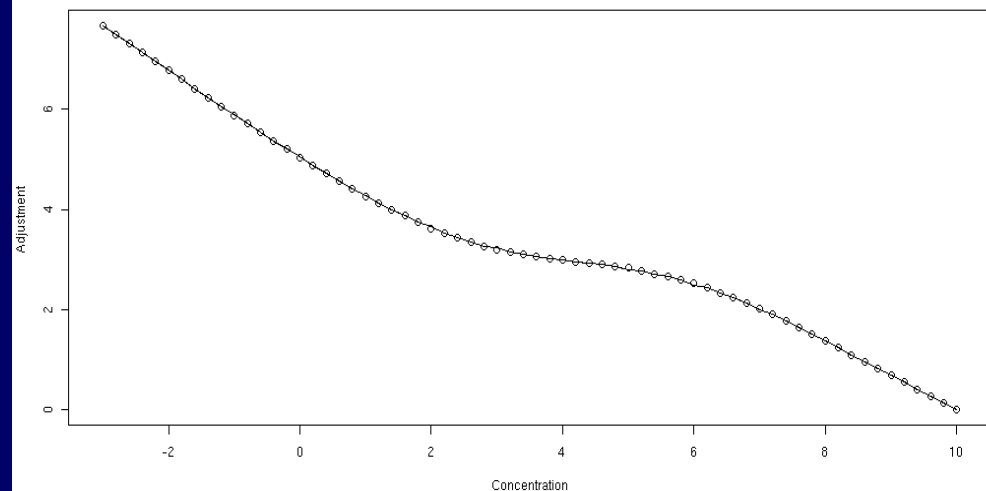
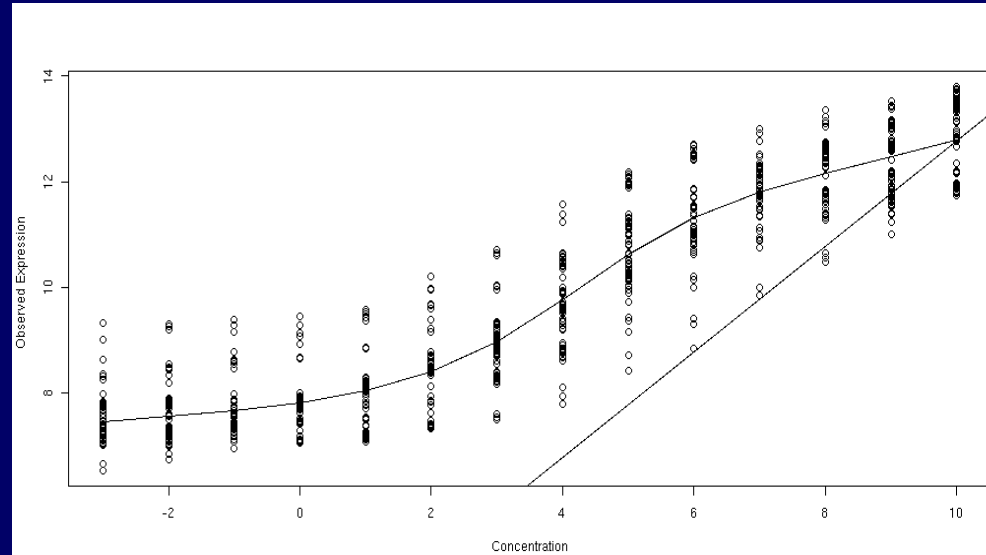
- $O = S + N$
 - O is observed PM, S is signal (assumed exponential), N is noise (assumed normal, truncated at zero)
- Correction is then

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

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

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?

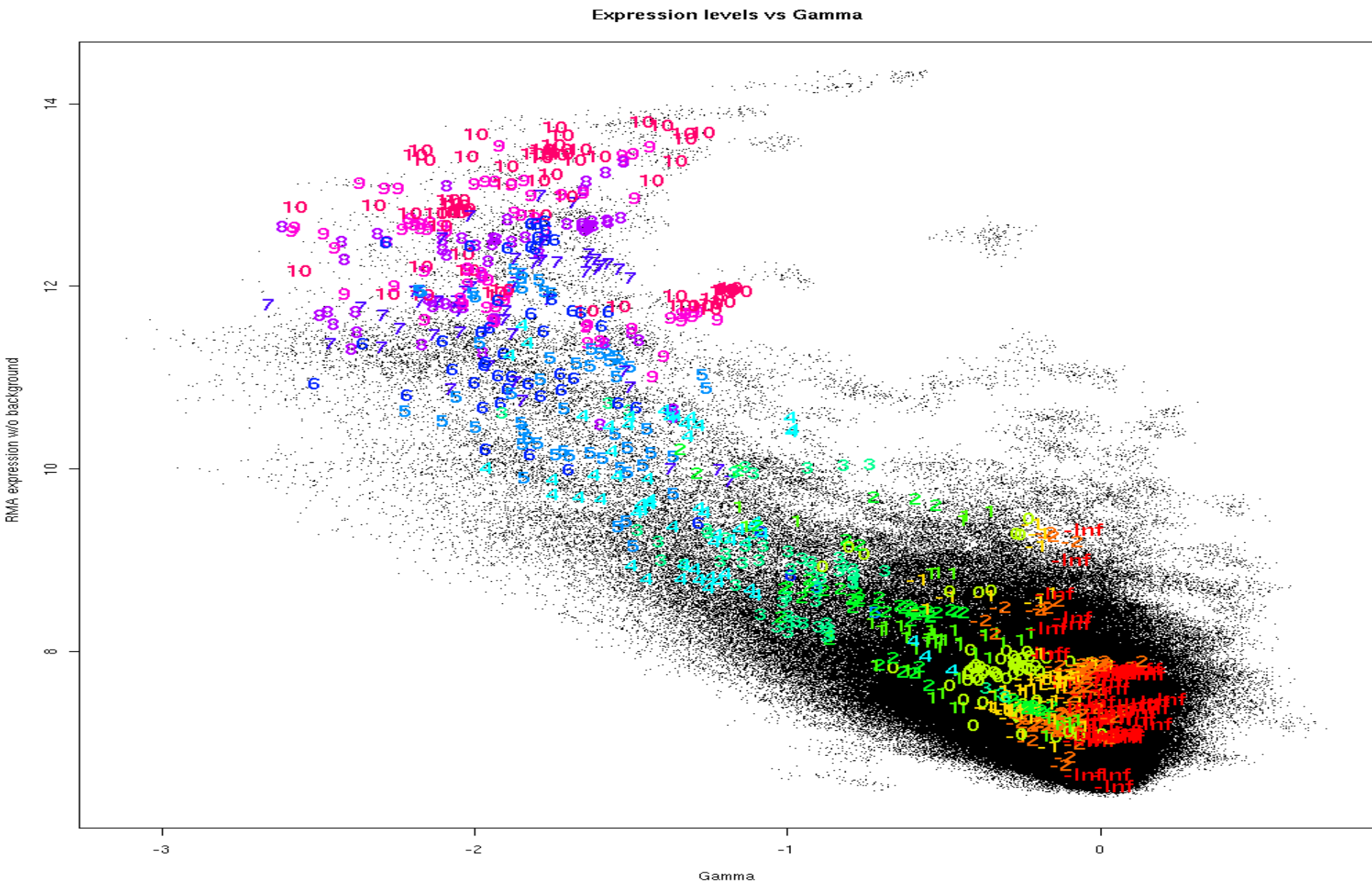
- We don't know a concentration for most probesets. If we did, or if we had a variable that related to concentration, the adjustment would be easy to perform
- Fit the following model

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

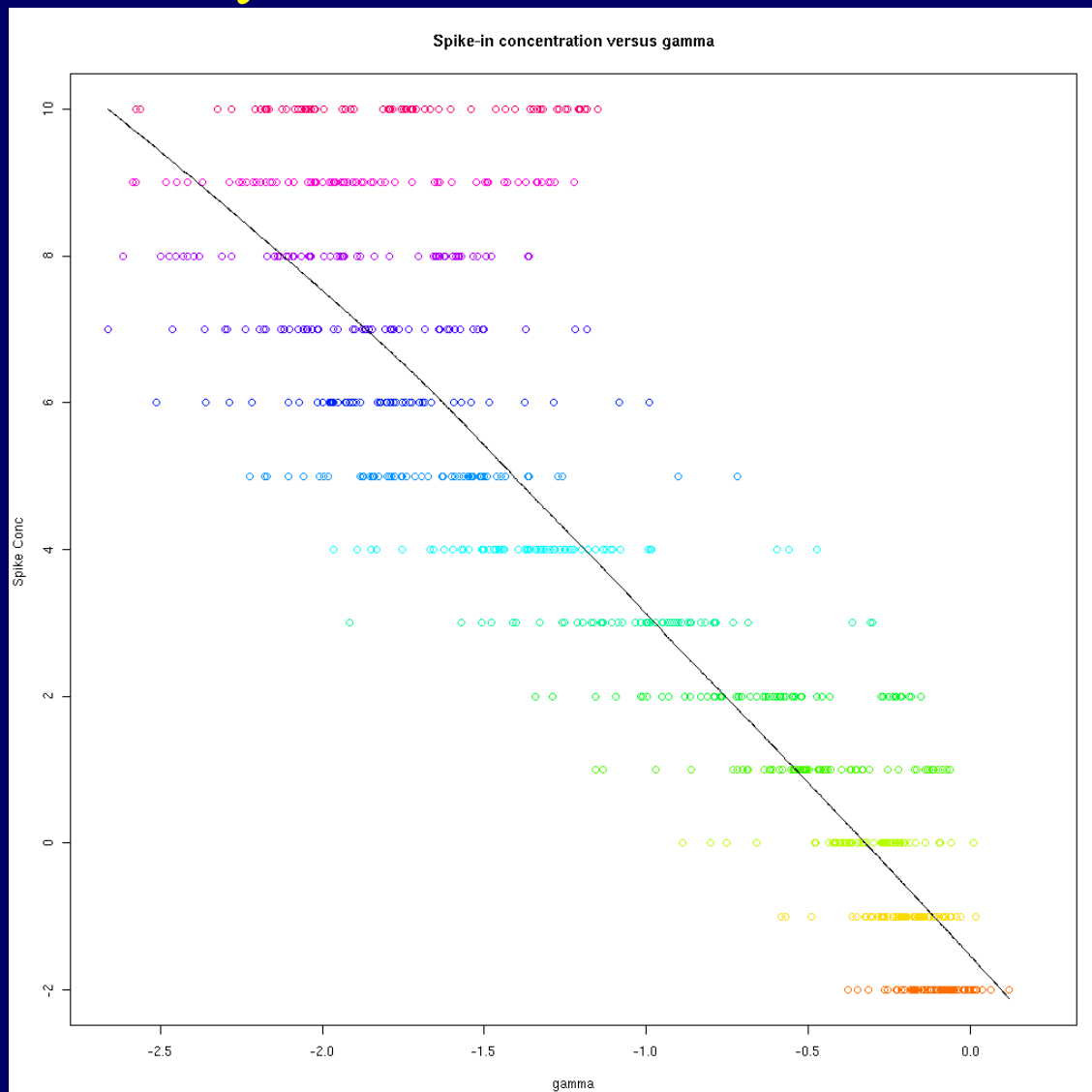
$$y_{2i}^{(k)} = \alpha_i^{(k)} + \gamma^{(k)} + \varepsilon_i^{(k)}$$

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

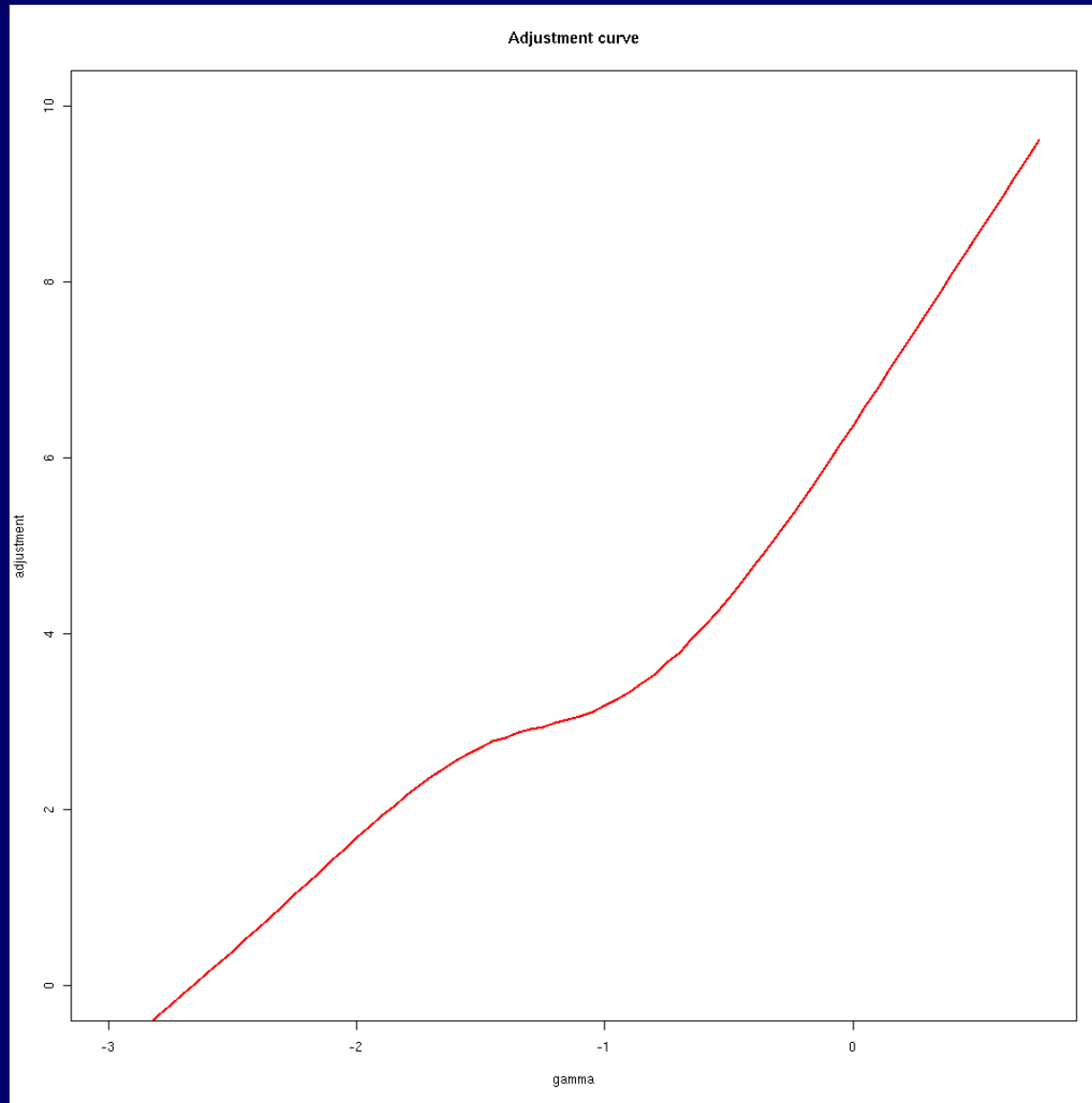
γ Relates to Concentration



Establishing a Relationship Between γ and Concentration

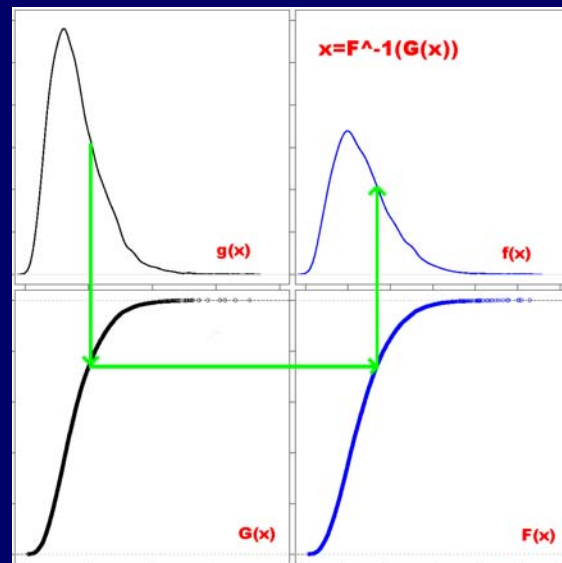


The Two Curves Yield an Adjustment Curve



Quantile Normalization

- Normalize so that the quantiles of each chip are equal. Simple and fast algorithm. Goal is to give same distribution to each chip.



- We will illustrate the algorithm with an example.

Sort columns of original matrix

$$\begin{bmatrix} 1 & 5 & 3 & 5 \\ 2 & 1 & 6 & 7 \\ 3 & 2 & 2 & 6 \\ 4 & 6 & 1 & 8 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & 1 & 5 \\ 2 & 2 & 2 & 6 \\ 3 & 5 & 3 & 7 \\ 4 & 6 & 6 & 8 \end{bmatrix}$$

Take averages across rows

$$\begin{bmatrix} 1 & 1 & 1 & 5 \\ 2 & 2 & 2 & 6 \\ 3 & 5 & 3 & 7 \\ 4 & 6 & 6 & 8 \end{bmatrix} \rightarrow \begin{bmatrix} 2 \\ 3 \\ 4.5 \\ 6 \end{bmatrix}$$

Set average as value for All elements in the row

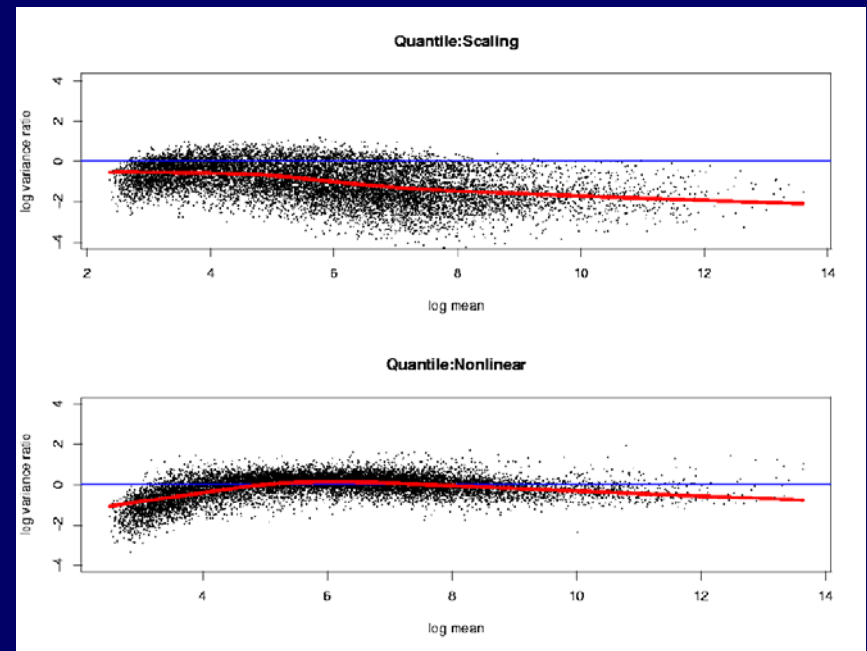
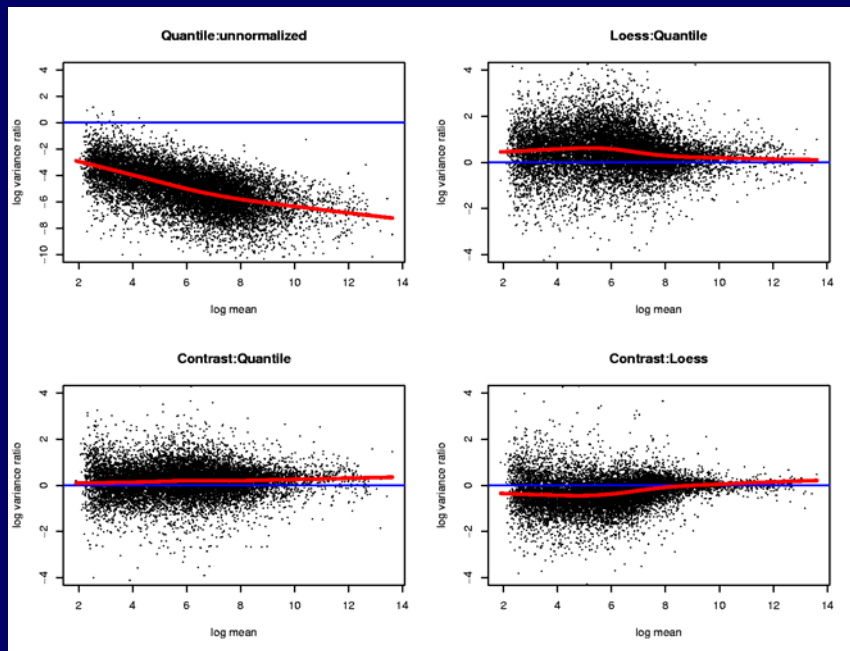
$$\begin{bmatrix} 2 \\ 3 \\ 4.5 \\ 6 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4.5 & 4.5 & 4.5 & 4.5 \\ 6 & 6 & 6 & 6 \end{bmatrix}$$

Unsort columns of matrix to original order

$$\begin{bmatrix} 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4.5 & 4.5 & 4.5 & 4.5 \\ 6 & 6 & 6 & 6 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 4.5 & 4.5 & 2 \\ 3 & 2 & 6 & 4.5 \\ 4.5 & 3 & 3 & 3 \\ 6 & 6 & 2 & 6 \end{bmatrix}$$

Why Quantile Normalization?

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



RMA Model

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

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

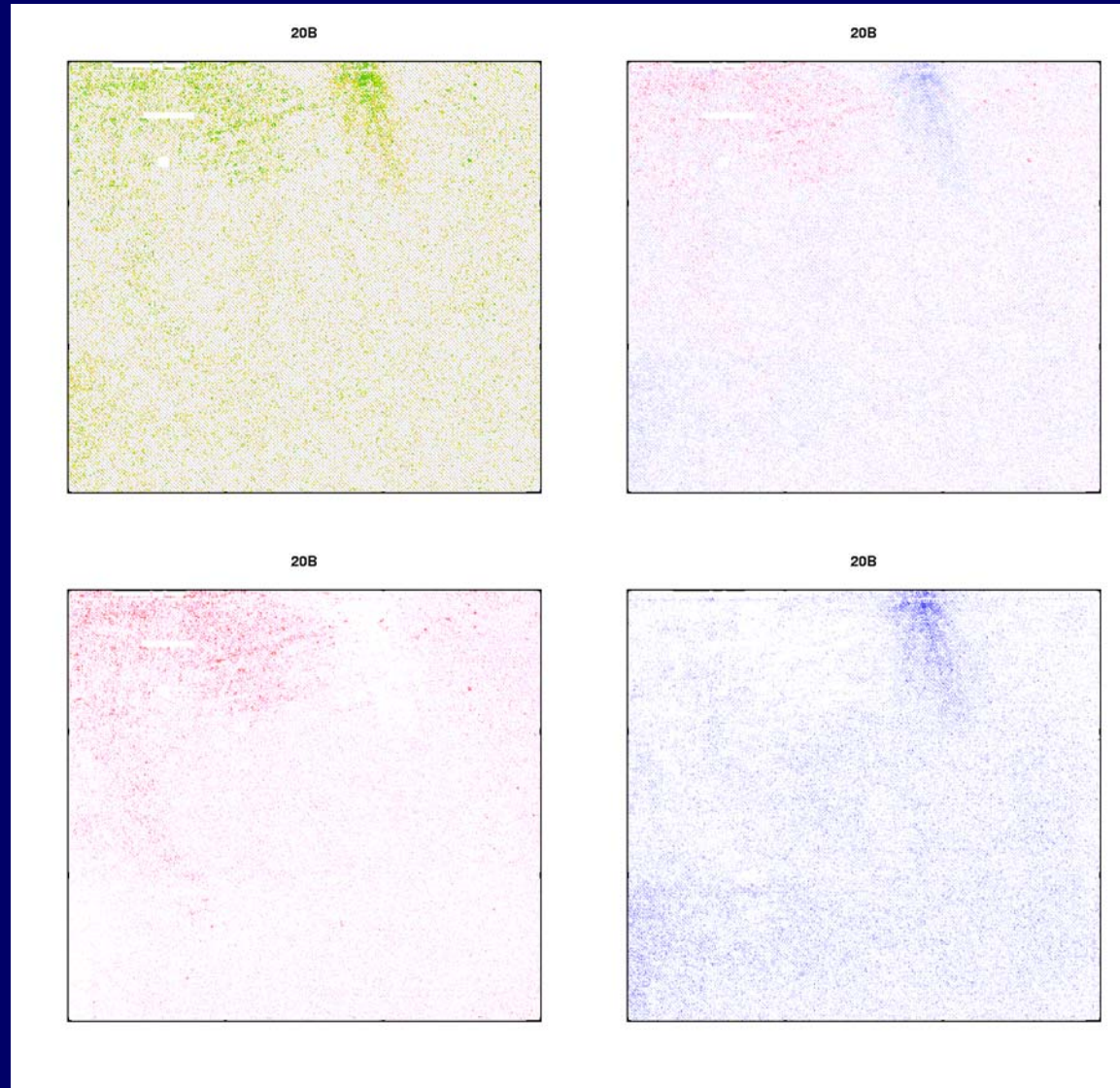
where $\alpha_i^{(k)}$ is a probe effect and $\beta_j^{(k)}$ is the log gene expression. $y_{ij}^{(k)}$ is the log2 background adjusted and normalized PM intensity

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

Probe Level Models are based on RMA

- RMA method
 - Convolution Model Background
 - Quantile Normalization
 - Summarization using a robust multi-chip model on the log scale. Model is fitted using the median polish algorithm on a probeset by probeset basis

Pseudo-chip images using PLM outputs aid QC



Basic RMA model

Let $y_{ij} = \log_2 N(B(PM_{ij}))$

then $y_{ij} = m + \alpha_i + \beta_j + \varepsilon_{ij}$

where α_i is probe-effect

β_j is chip-effect ($m + \beta_j$ is log2 gene expression on array j)

Median-polish imposes constraints

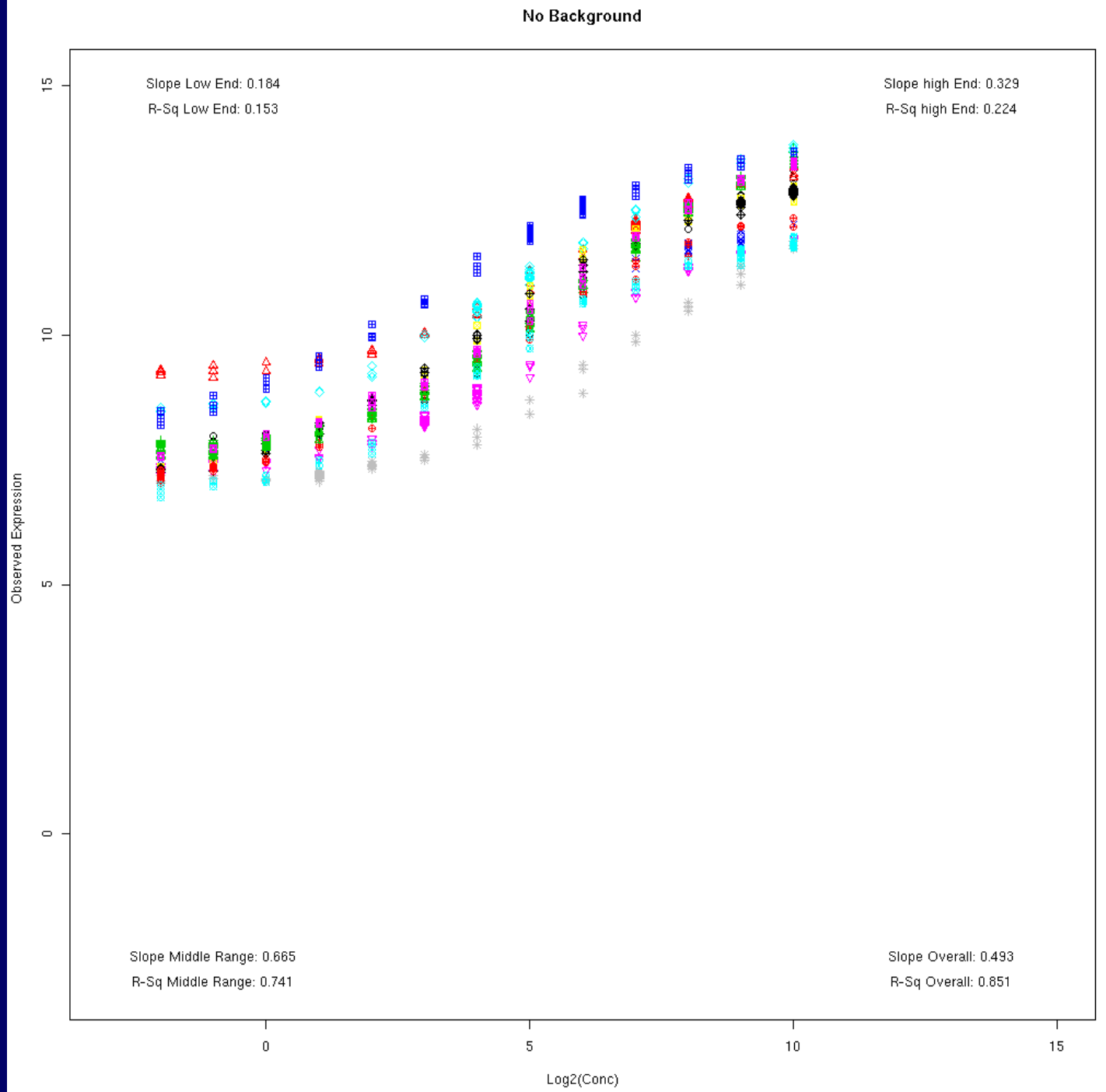
$$\text{median}_i \alpha_i = \text{median}_j \beta_j = 0$$

$$\text{median}_i \varepsilon_{ij} = \text{median}_j \varepsilon_{ij} = 0$$

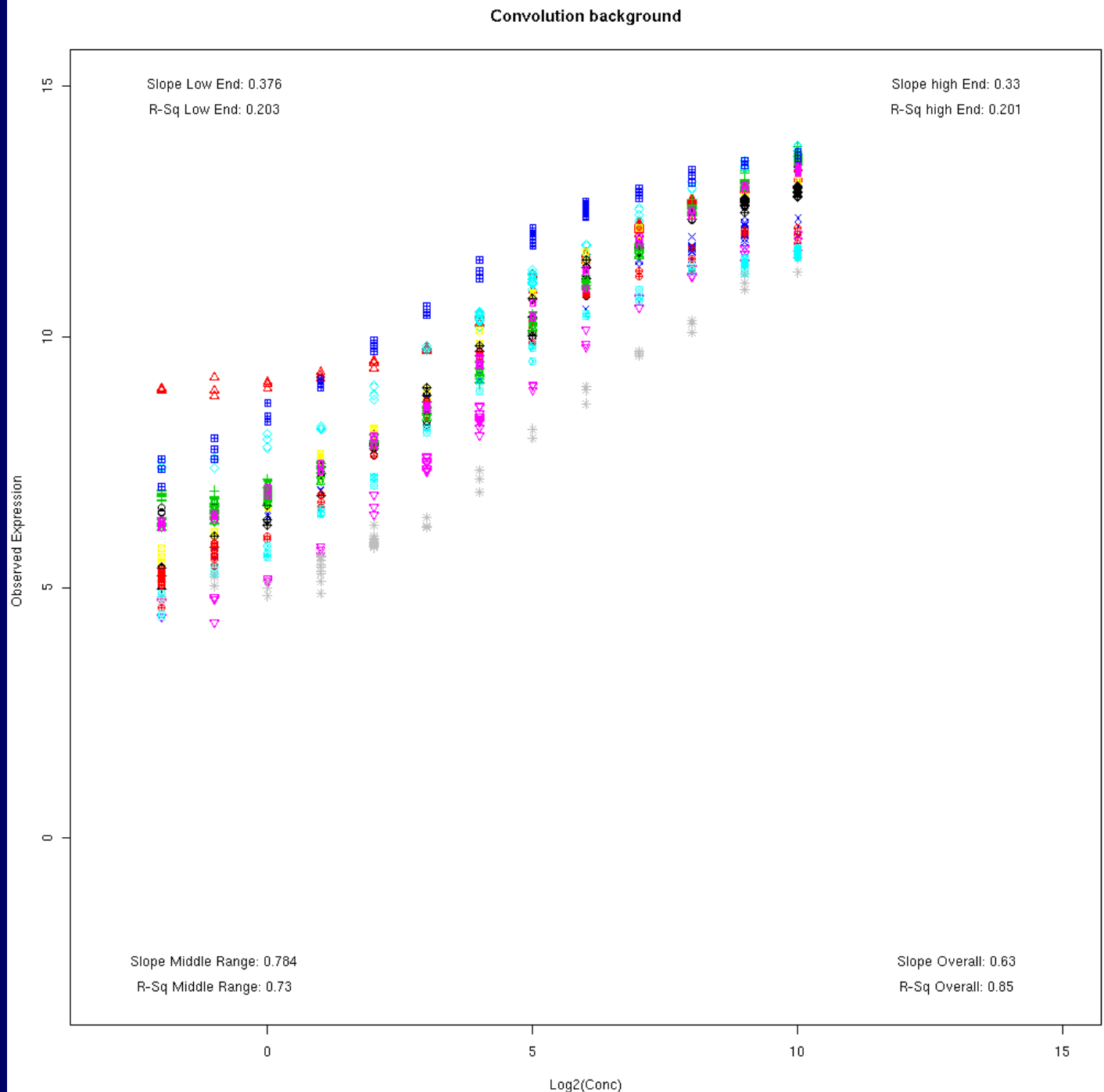
Advantages/Disadvantages of RMA/Median polish

- Advantages
 - Fast
 - Gene expression measures perform favorably when compared with MAS 5.0, Li-Wong MBEI
 - Robust against outliers
- Disadvantages
 - No standard error estimates
 - No algorithmic flexibility to fit alternative models

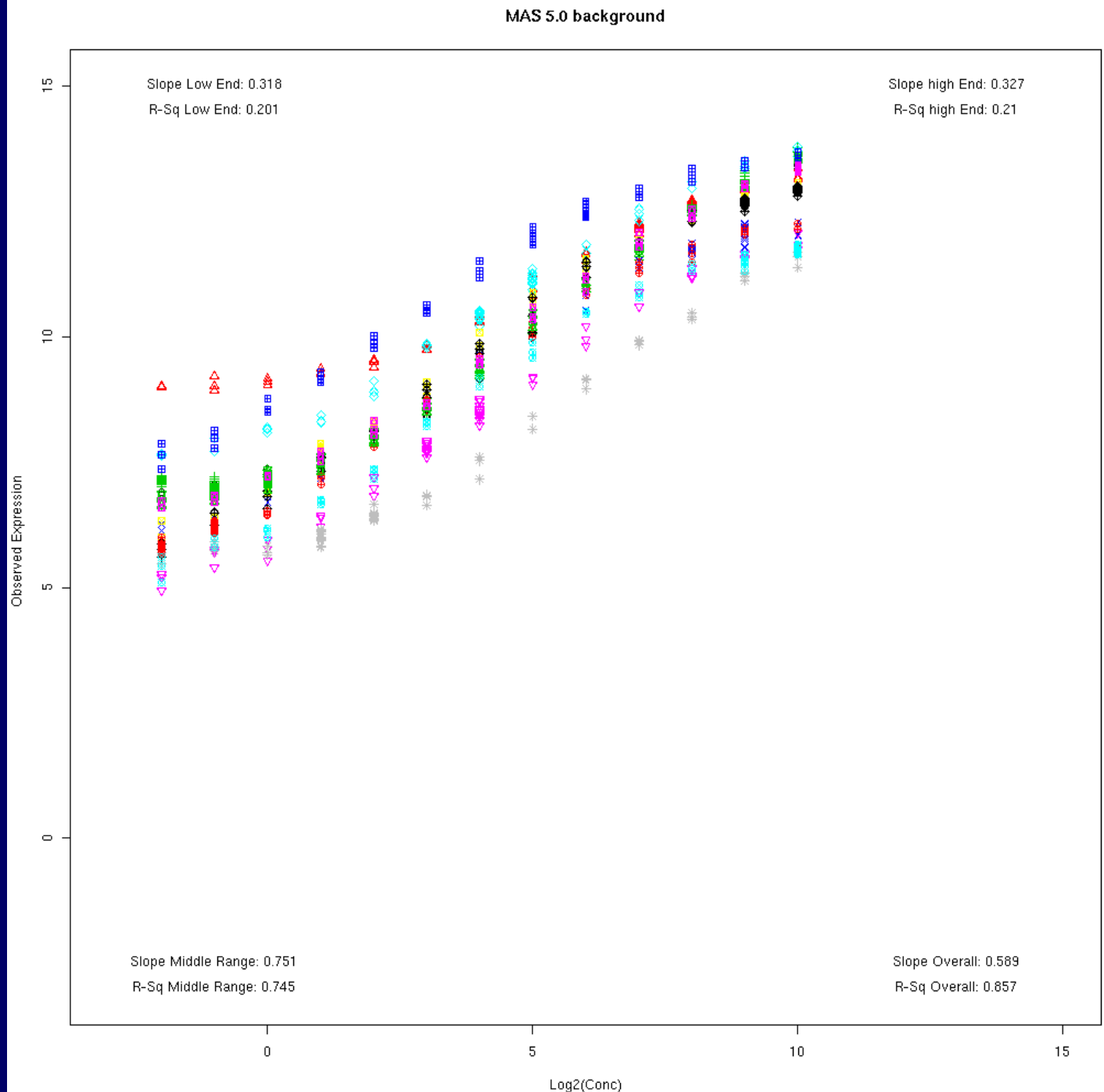
Slope	Value
All	0.493
Mid	0.665
Low	0.184
High	0.329



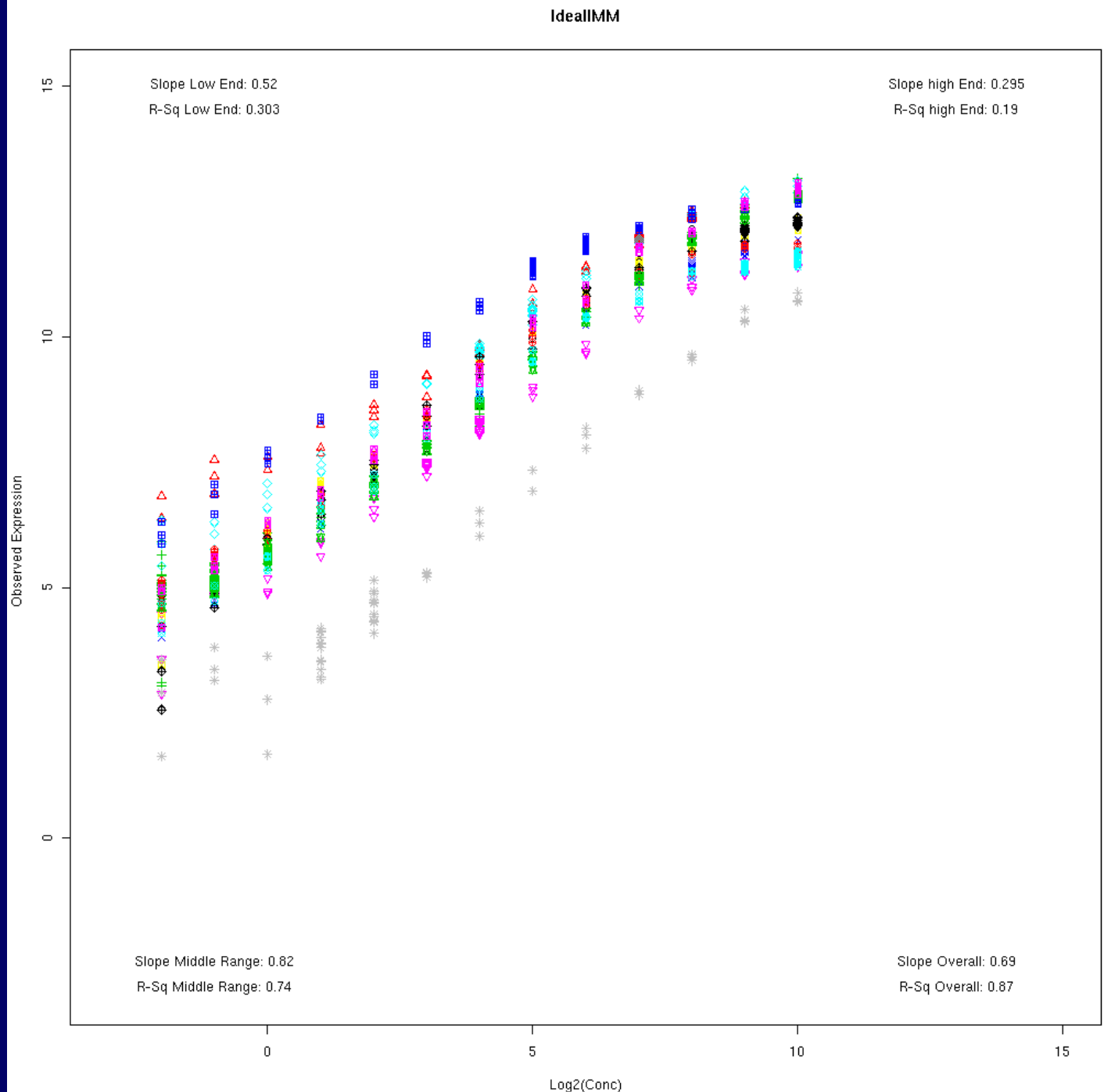
Slope	Value
All	0.63
Mid	0.784
Low	0.376
High	0.33



Slope	Value
All	0.589
Mid	0.751
Low	0.318
High	0.327

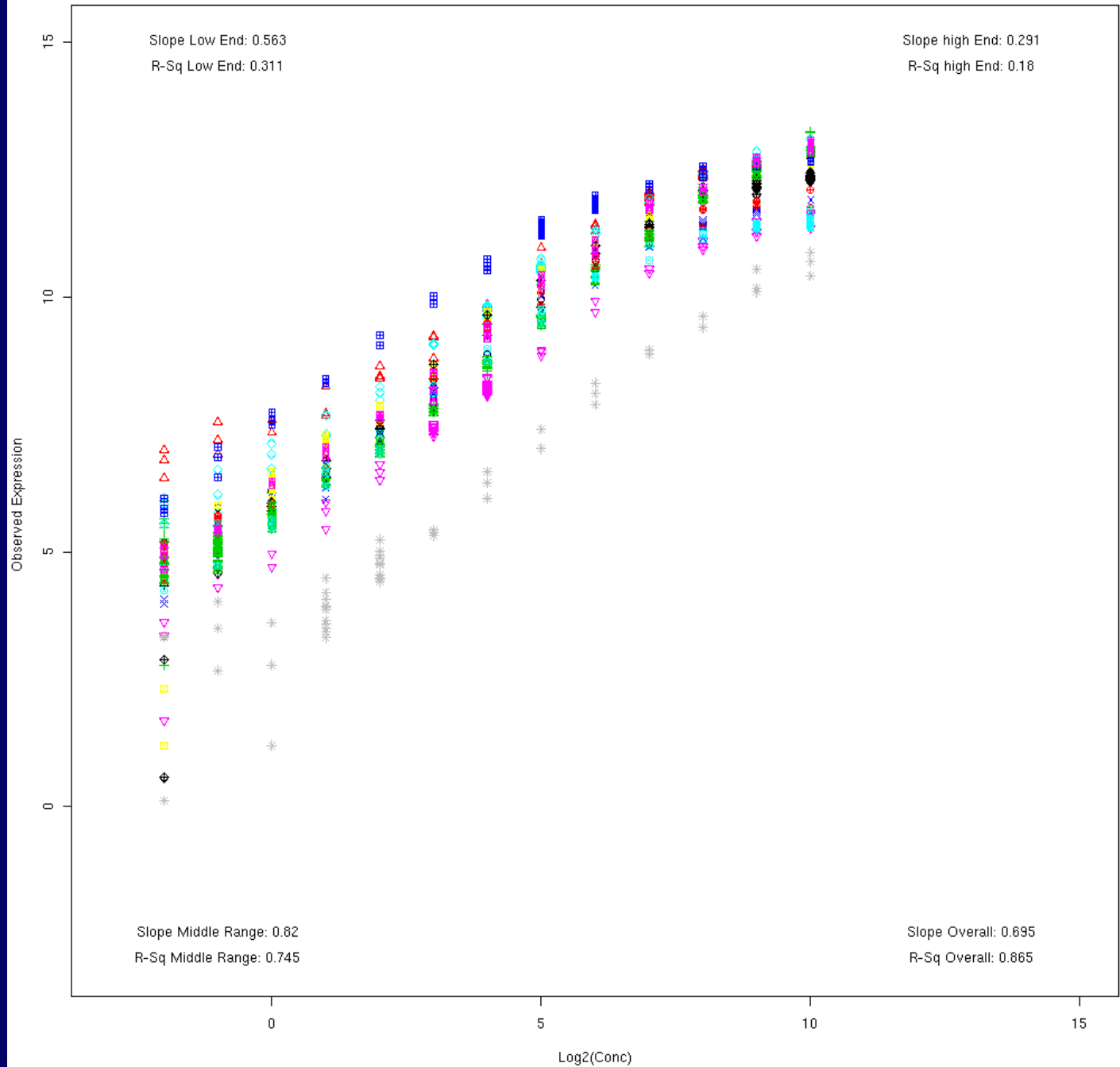


Slope	Value
All	0.69
Mid	0.82
Low	0.52
High	0.295



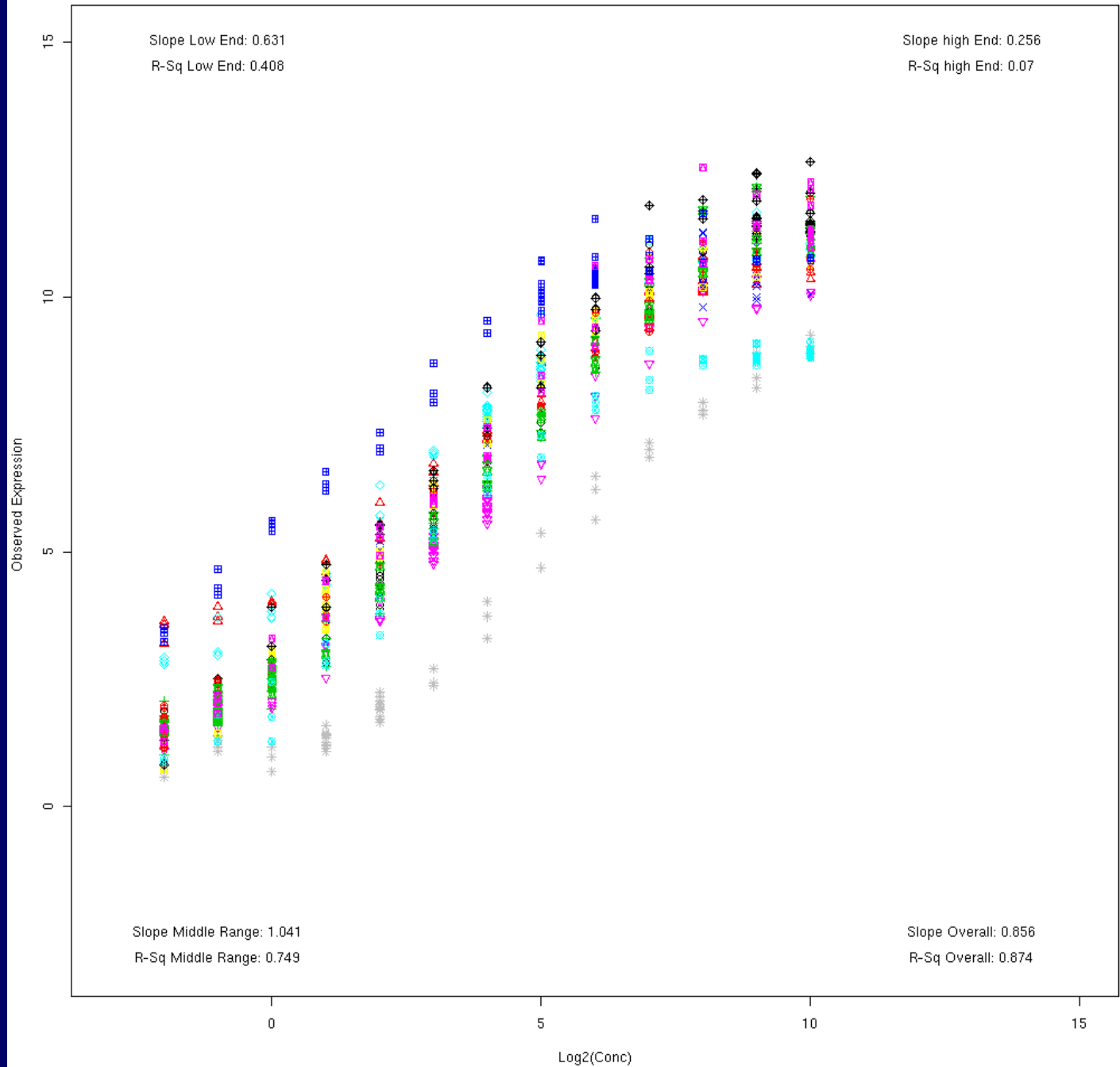
Slope	Value
All	0.695
Mid	0.82
Low	0.563
High	0.291

MAS 5.0 bg then IdealIMM

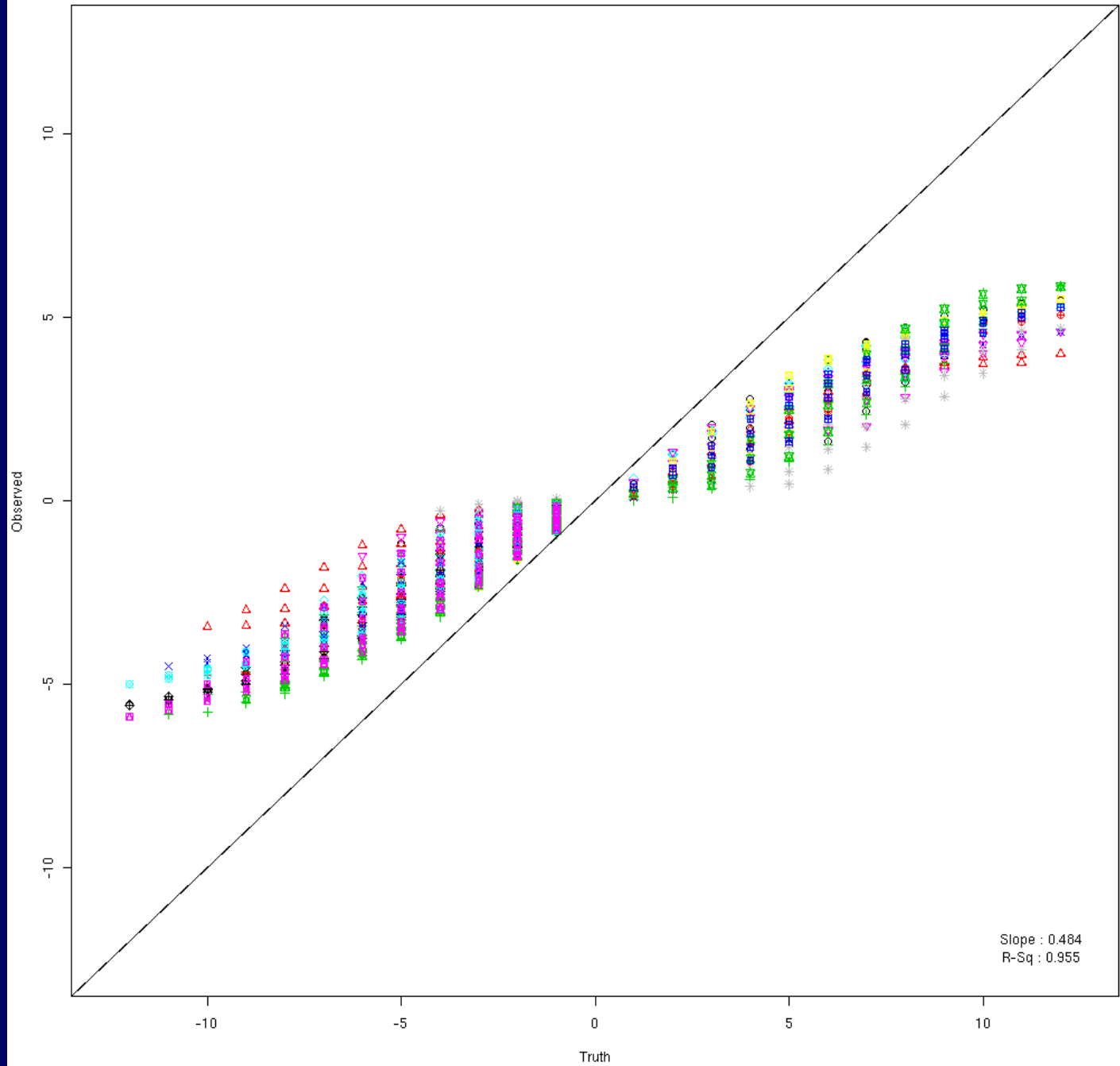


Slope	Value
All	0.856
Mid	1.041
Low	0.631
High	0.256

Standard Curve Adjustment



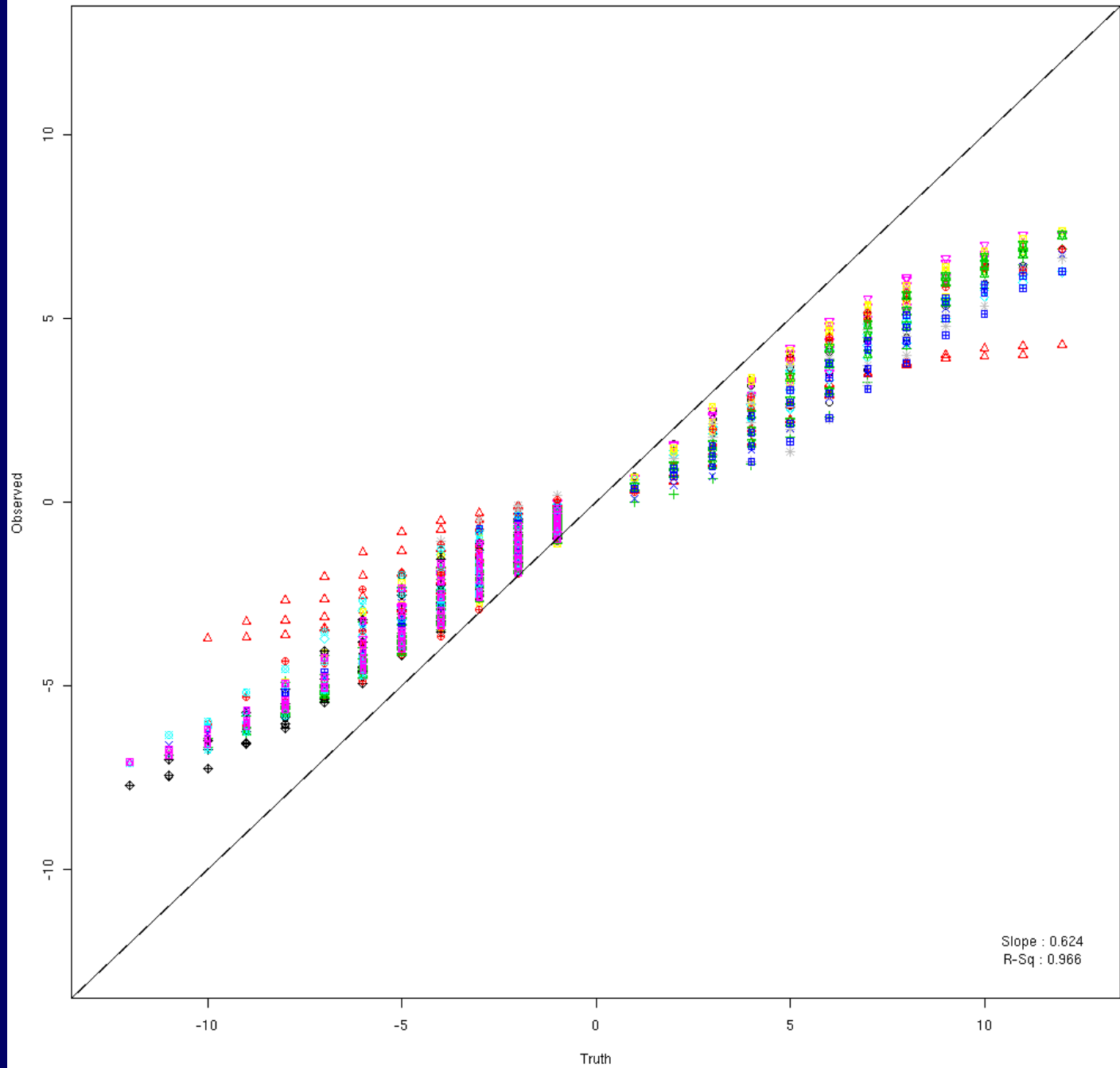
No background



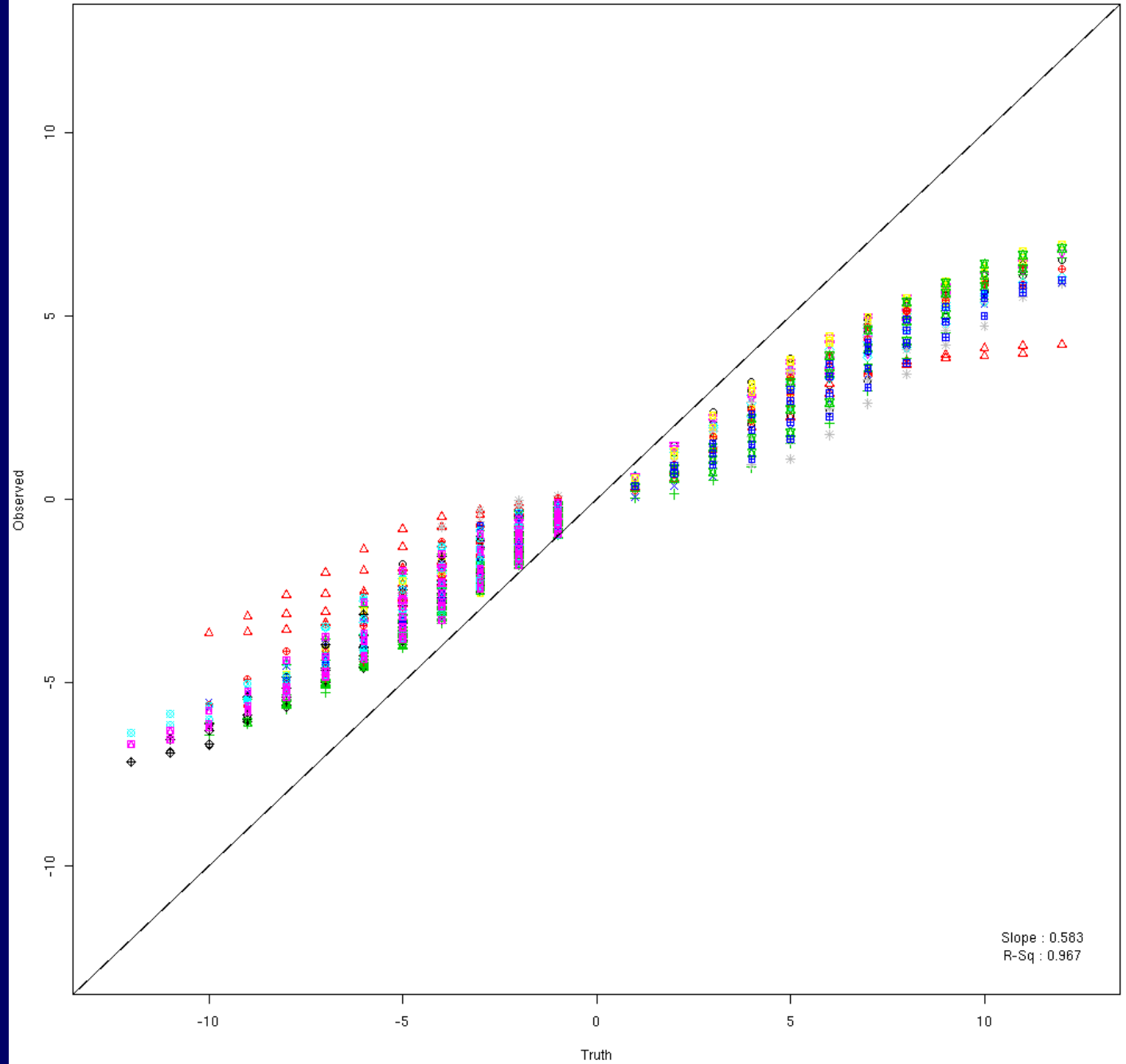
Slope: 0.484

Convolution background

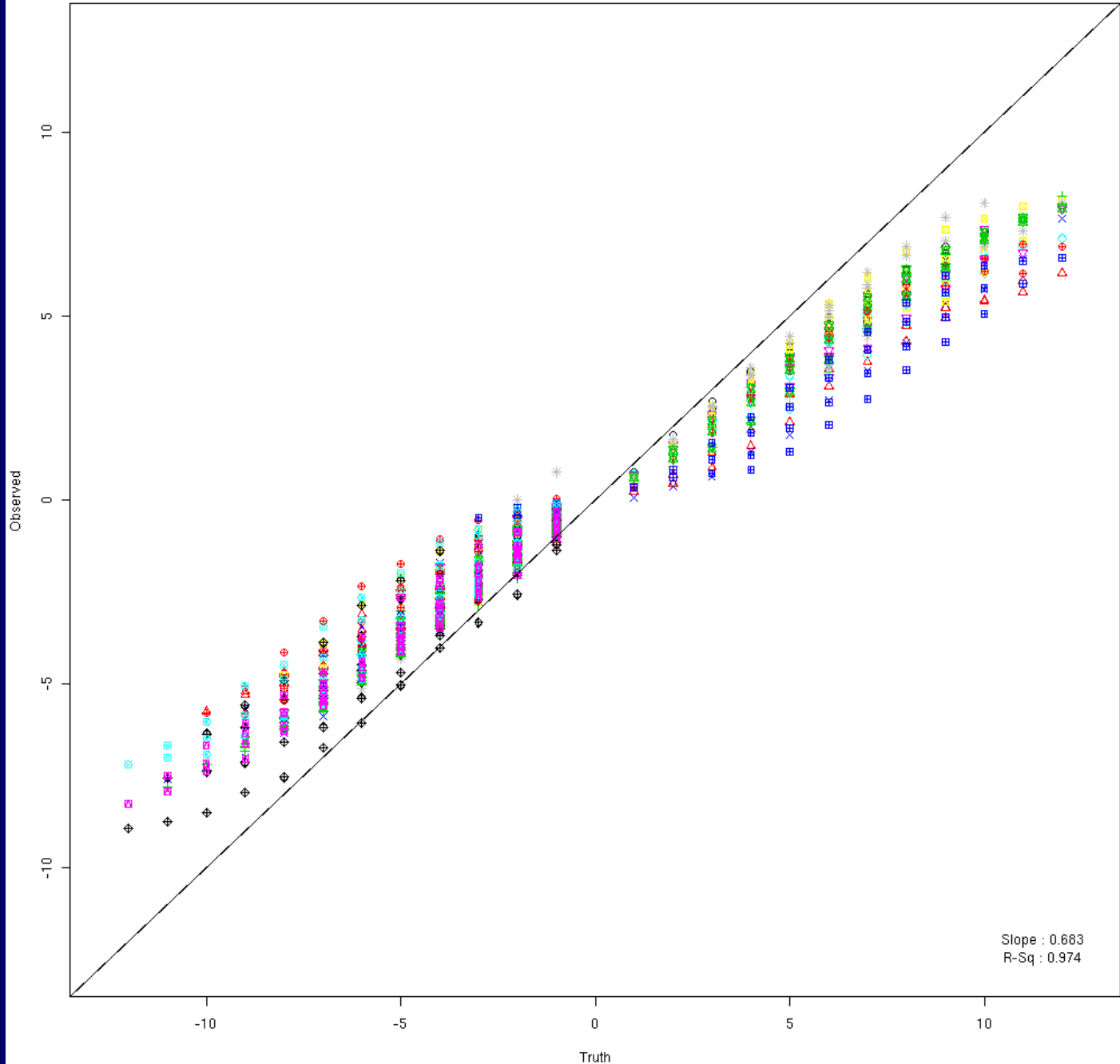
Slope: 0.624



Slope: 0.583

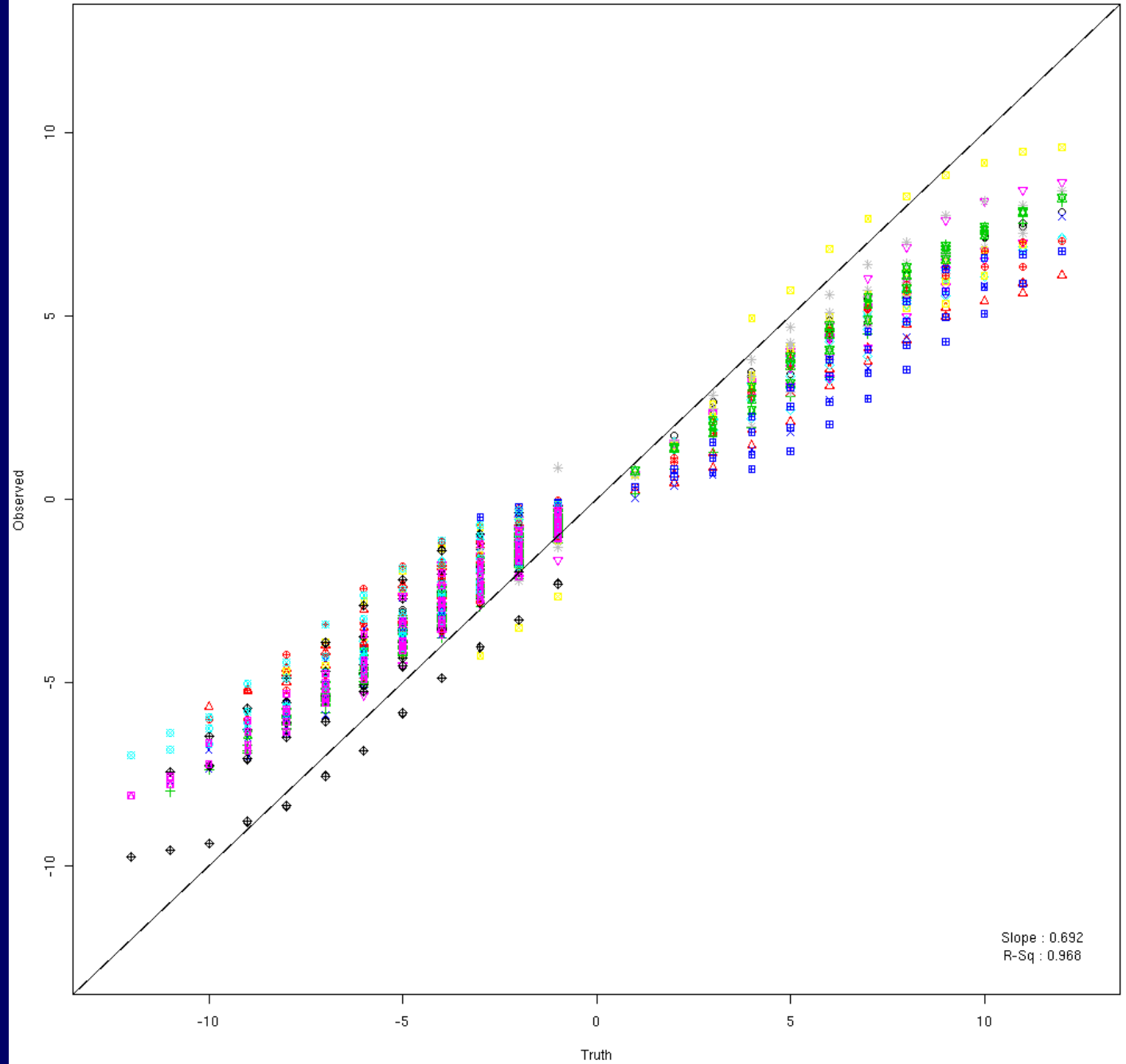


Ideal Mismatch



Slope: 0.683

Slope: 0.692



Standard Curve Adjustment

Slope: 0.847

