

# Explorations in Probe Level Analysis of GeneChips

Ben Bolstad

June 8, 2005

# A Little Background

- PhD (2004) Biostatistics UC Berkeley.  
*Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization.*
- RMA (with Speed, Irizarry and Colin)
- BioConductor Core Developer
  - Packages dealing with Affymetrix Data

# What is RMA?

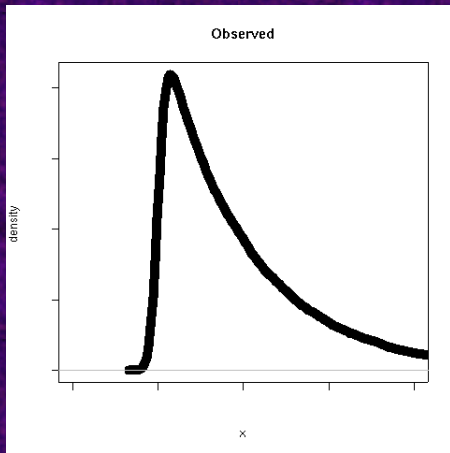
1. Convolution Background
  2. Quantile Normalization
  3. Probe Level Linear model on the log<sub>2</sub> scale fit robustly.
- Software
    - Bioconductor *affy* package [www.bioconductor.org](http://www.bioconductor.org)
    - RMAExpress  
[www.stat.berkeley.edu/~bolstad/RMAExpress](http://www.stat.berkeley.edu/~bolstad/RMAExpress)
    - Also available in some commercial software including S+ ArrayAnalzyer, GeneTraffic, GeneSpring, Genesifter.net, ArrayAssist, .....

# 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

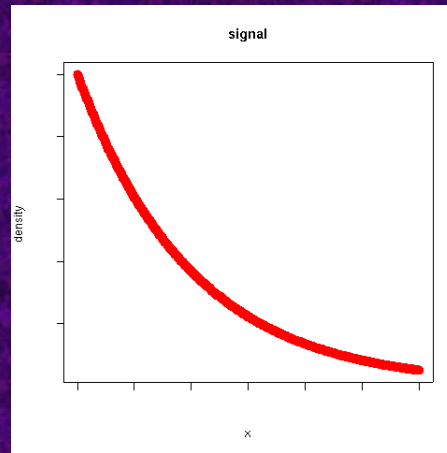
# RMA Background Approach

- Convolution Model



Observed  
O

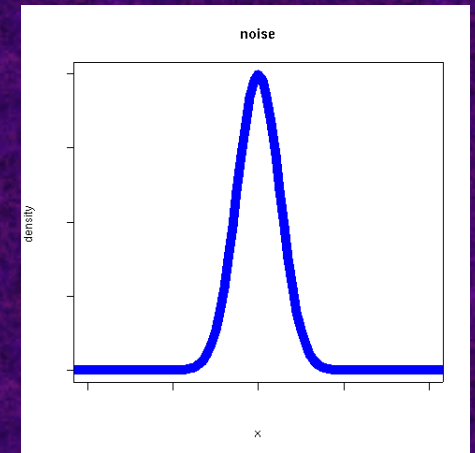
=



Signal  
S

$Exp(\alpha)$

+



Noise  
N

$N(\mu, \sigma^2)$

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

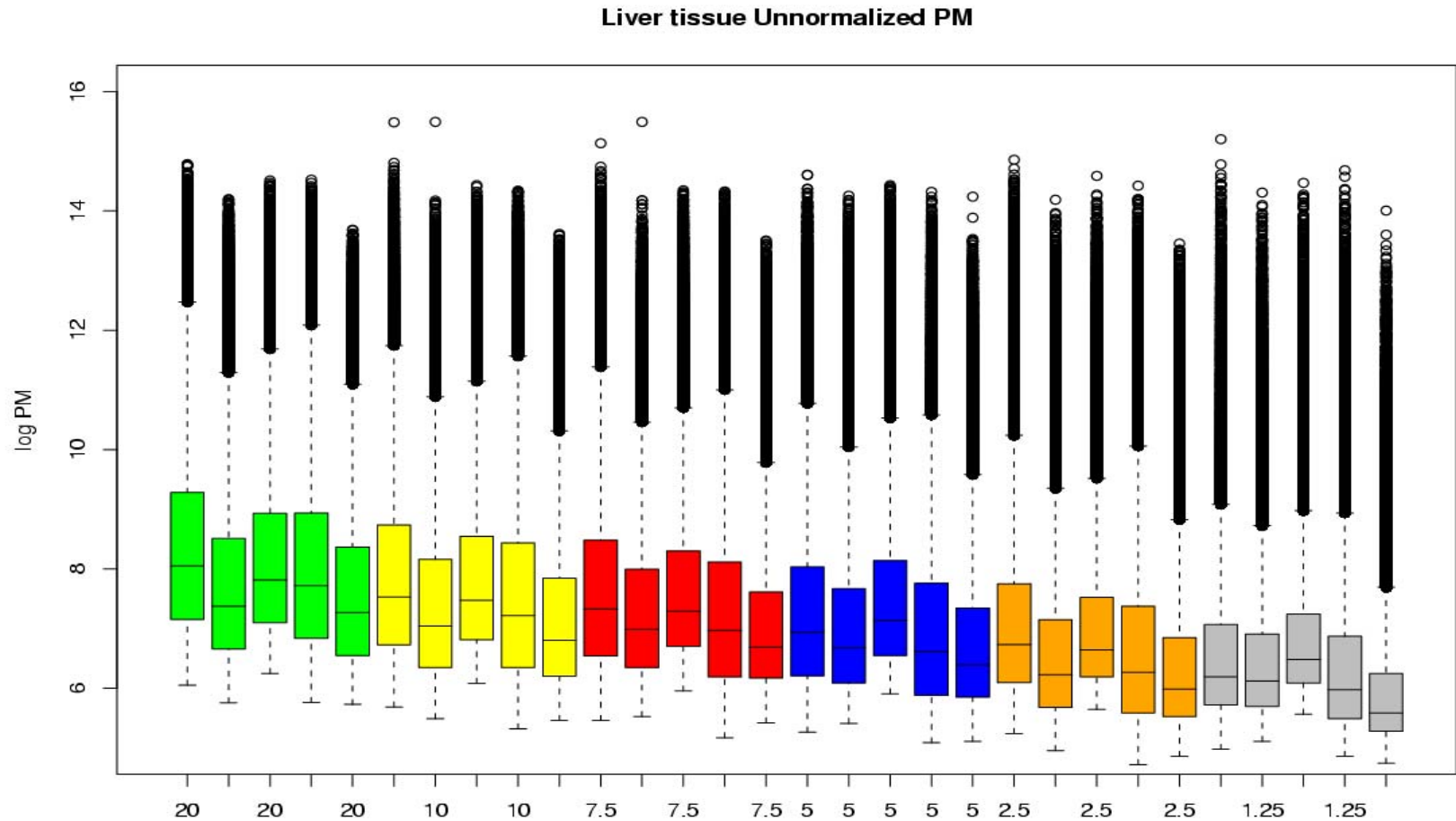
$$a = o - \mu - \sigma^2 \alpha, b = \sigma$$

# 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.”<sup>1</sup>*

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

# Non-Biological Variability

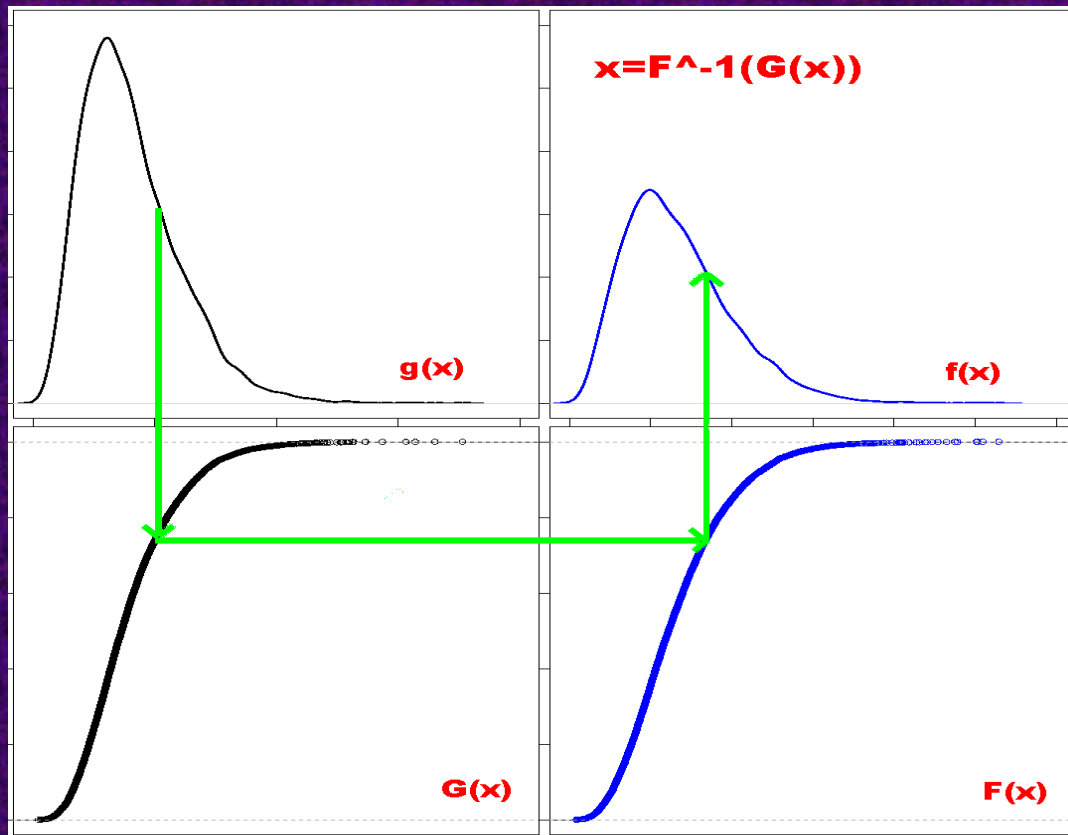


5 scanners for 6 dilution groups

# Quantile Normalization

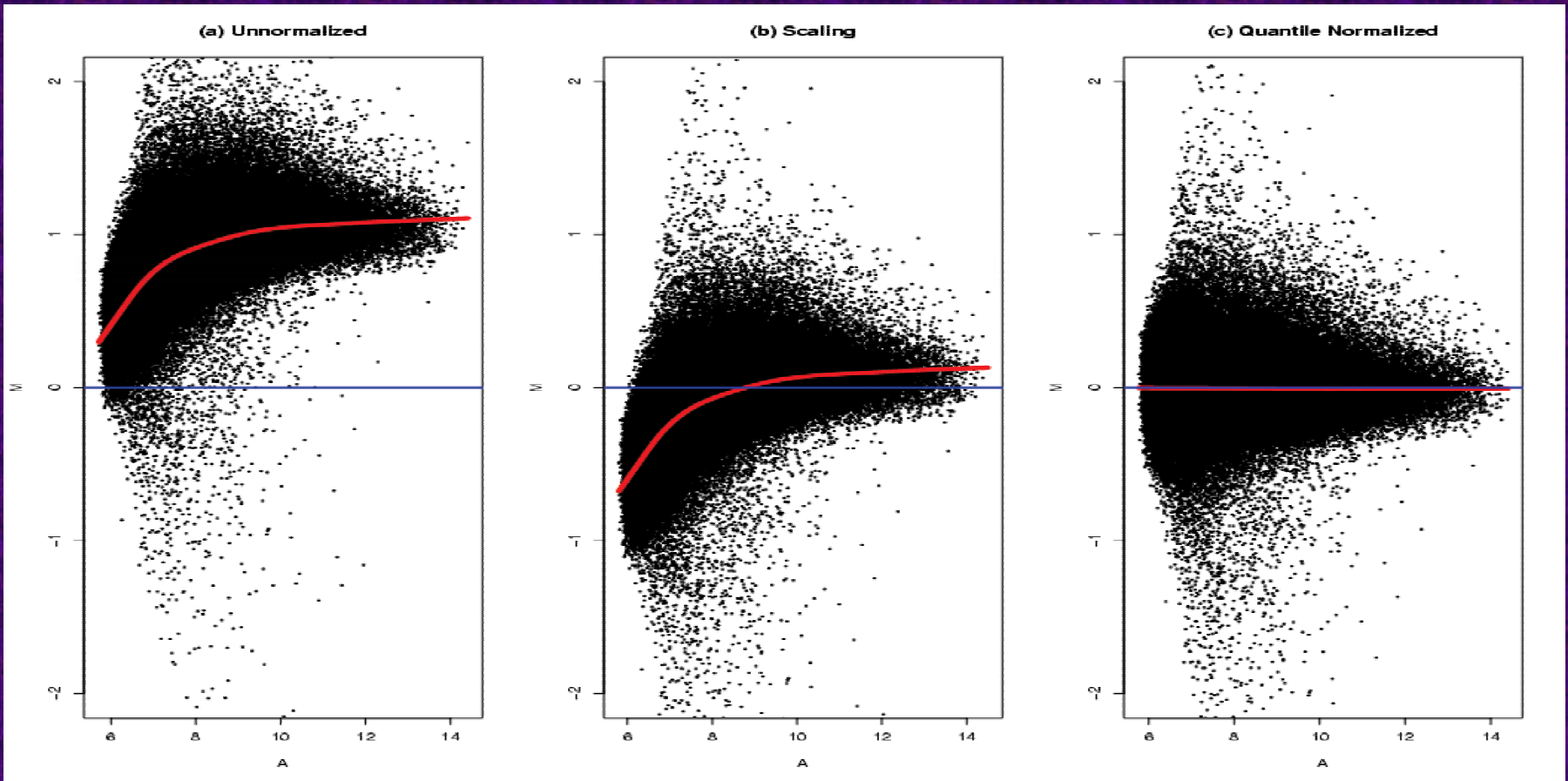
- Normalize so that the quantiles of each chip are equal. Simple and fast algorithm. Goal is to give same distribution to each chip.

Original  
Distribution



Target  
Distribution

# It works!!



Unnormalized

Scaled

Quantile  
Normalization

# Summarization

- Problem: Calculating gene expression values.
- How do we reduce the 11-20 probe intensities for each probeset on to a gene expression value?
- Our Approach
  - RMA – a robust multi-chip linear model fit on the log scale
- Some Other 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 on natural scale
    - PLIER, ....

# The RMA model

$$y_{kij} = m_k + \alpha_{ki} + \beta_{kj} + \varepsilon_{kij}$$

where  $y_{kij} = \log_2 N(B(PM_{kij}))$

$\alpha_{ki}$  is a probe-effect  $i=1, \dots, I$

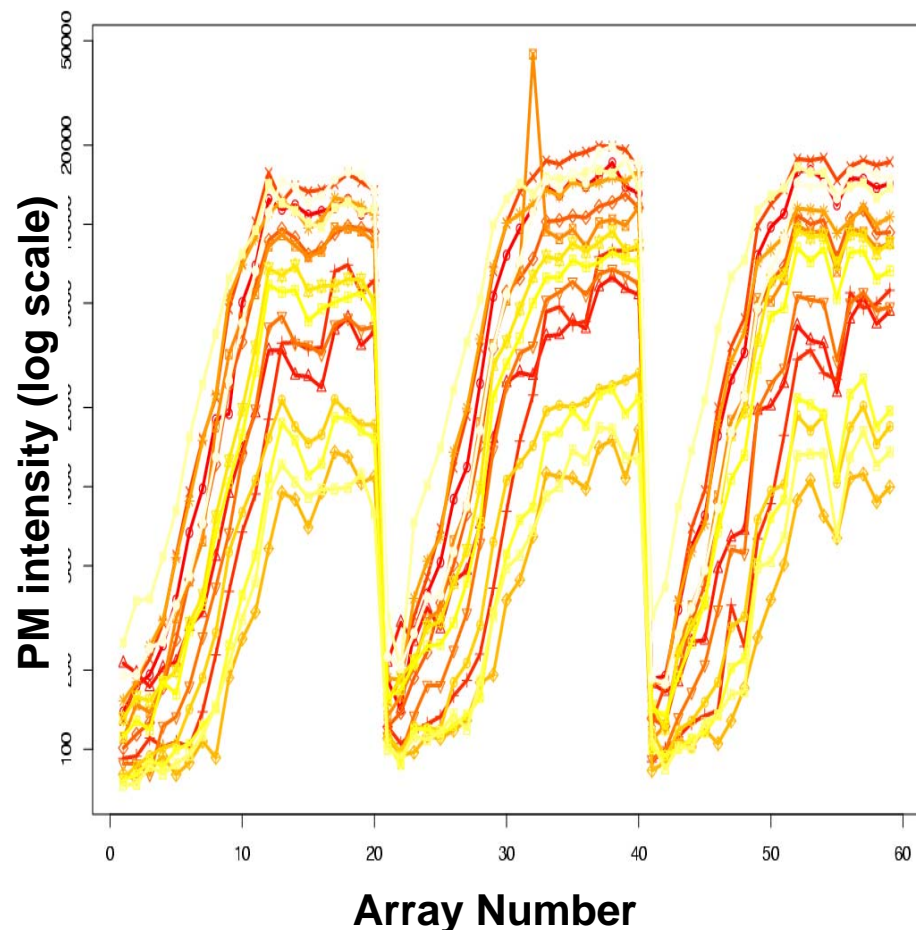
$\beta_{kj}$  is chip-effect ( $m_k + \beta_{kj}$  is log2 gene expression on array  $j$ )  $j=1, \dots, J$

$k=1, \dots, K$  is the number of probesets

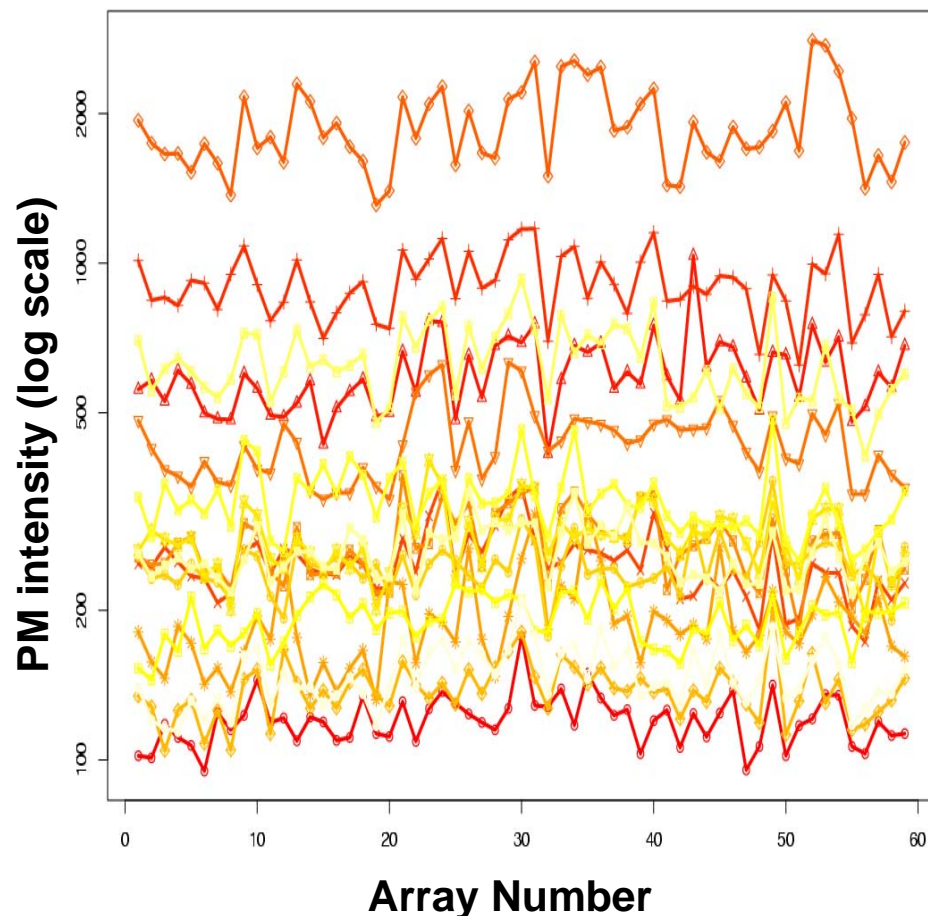
Why this model ? .....

# Parallel Behavior Suggests Multi-chip Model

Spike-in probeset

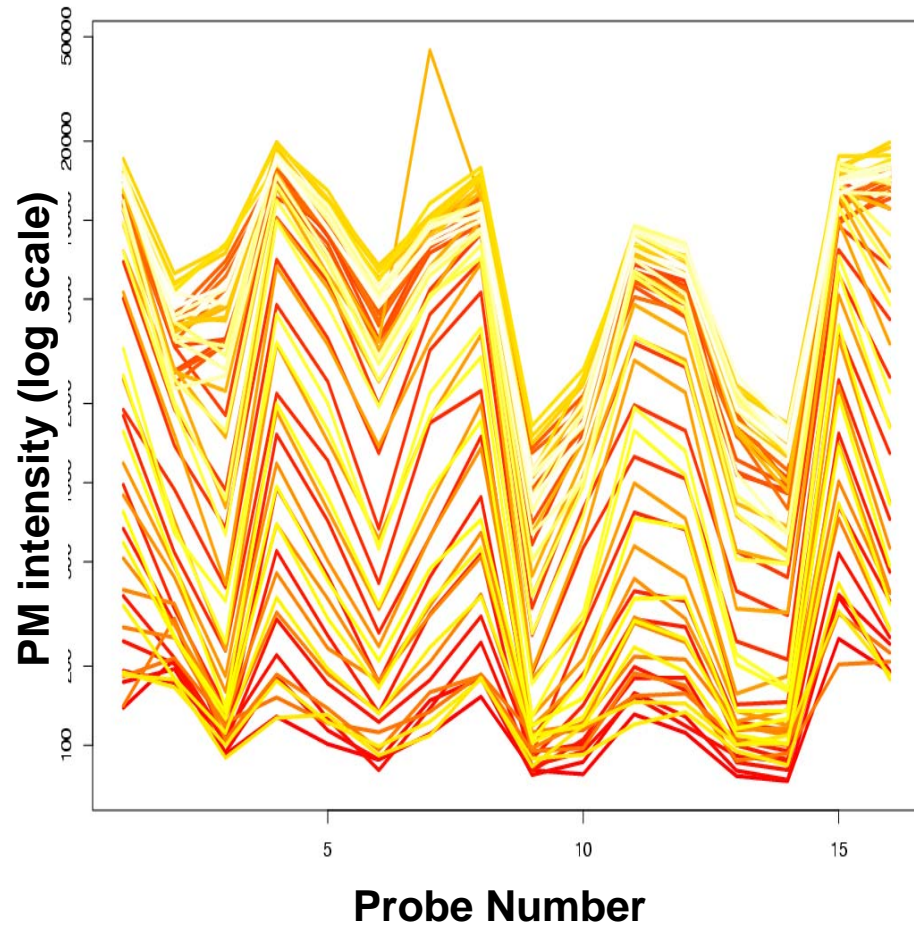


Non spike-in probeset

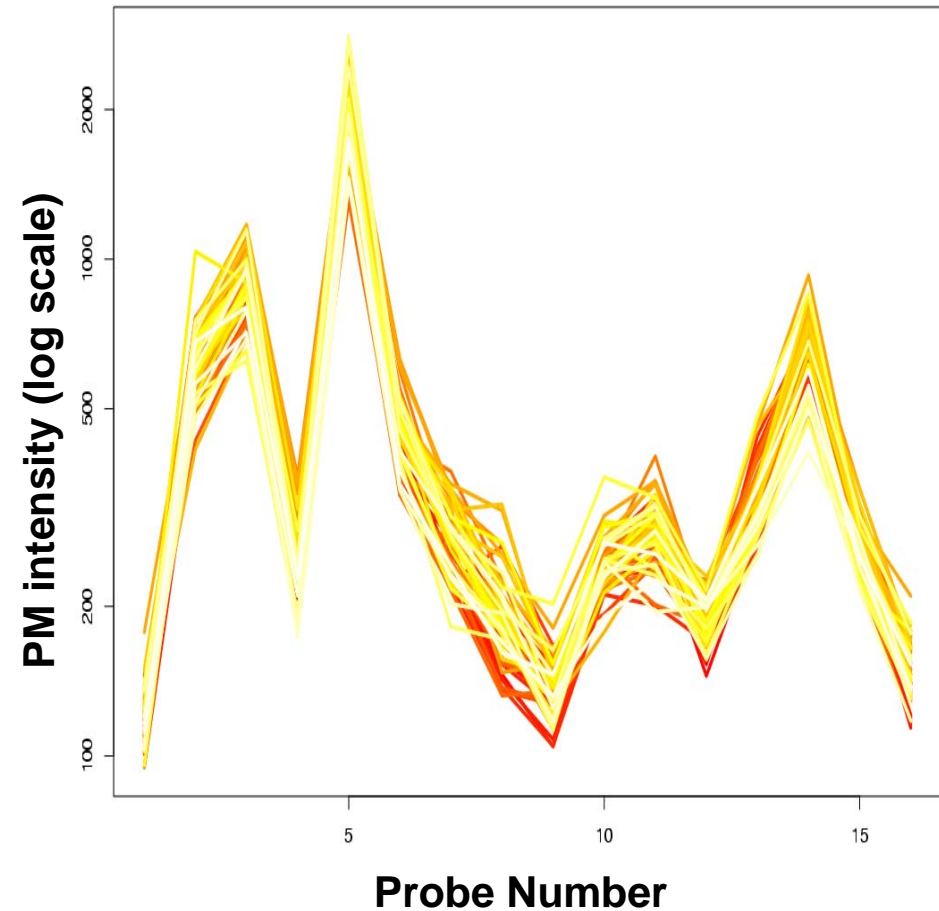


# Probe Pattern Suggests Including Probe-Effect

**Spike-in probeset**

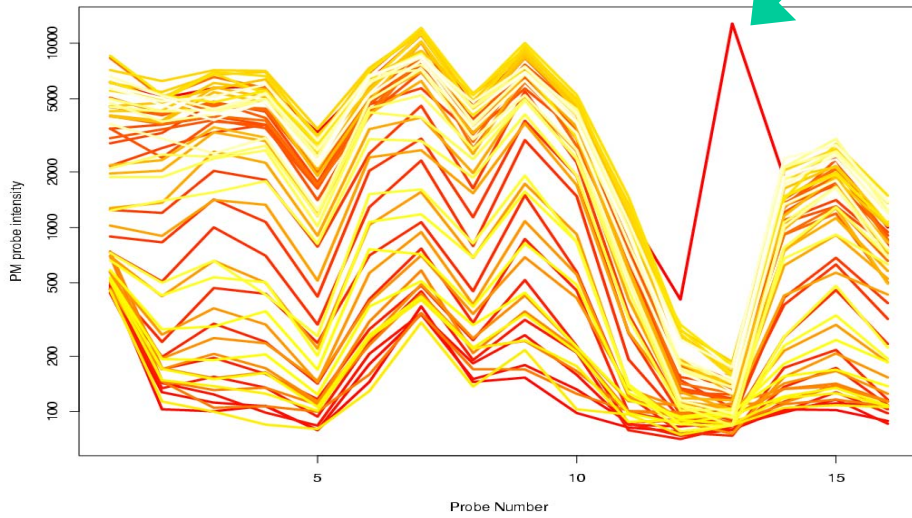


**Non spike-in probeset**

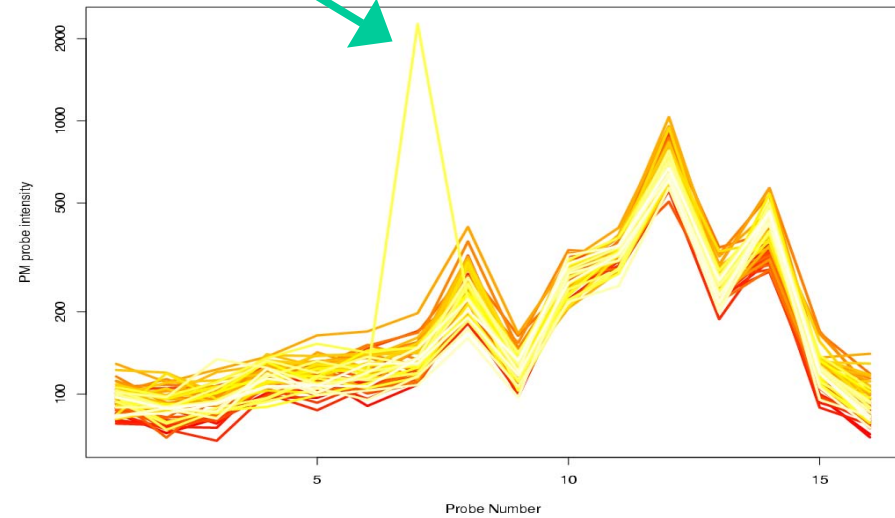


# Also Want Robustness

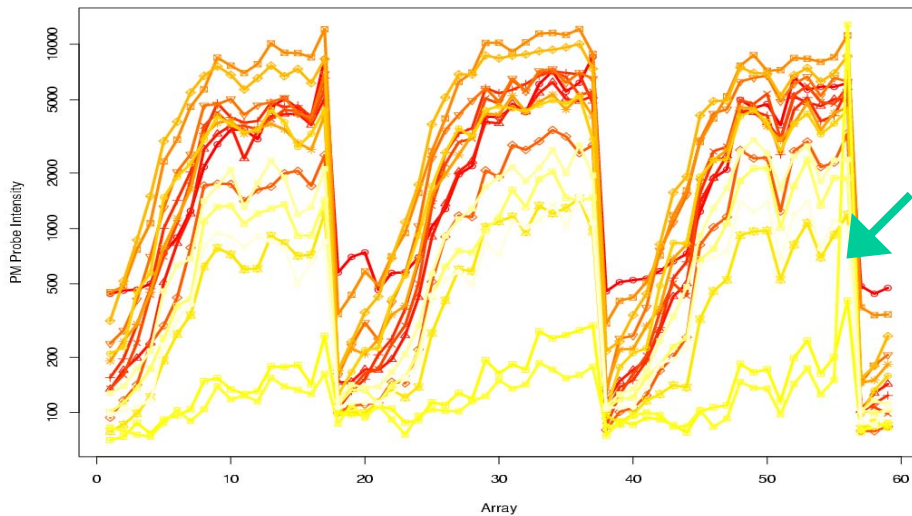
Probe behaviour (for spike-in)



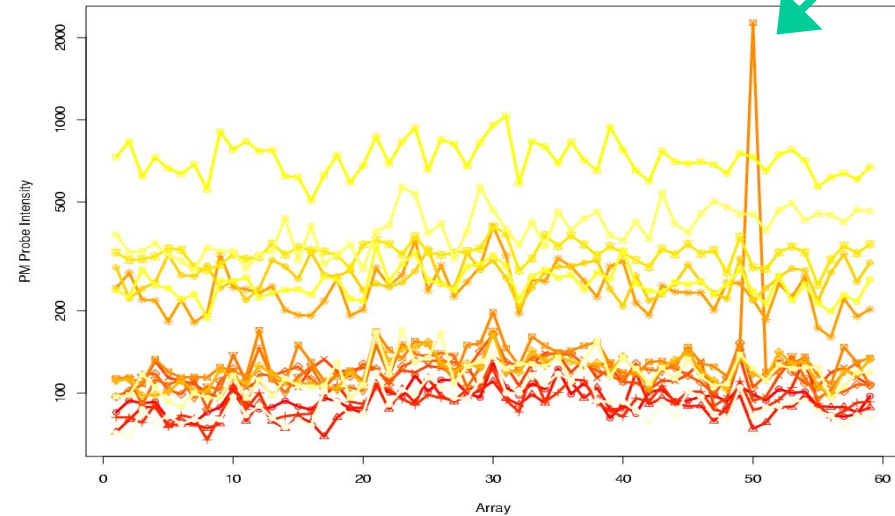
Probe behaviour (for non spike-in)



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



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



# Median Polish Algorithm

Pre-processed log2  
PM probe intensities  
for single probeset

$y_{11}$	$\cdots$	$y_{1J}$	$0$
$\vdots$	$\ddots$	$\vdots$	$\vdots$
$y_{I1}$	$\cdots$	$y_{IJ}$	$0$
$0$	$\cdots$	$0$	$0$

Sweep Rows

Sweep Columns

Iterate

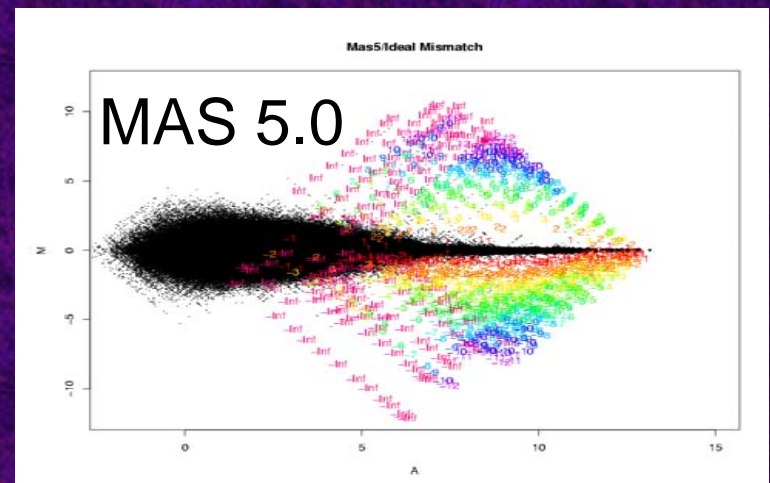
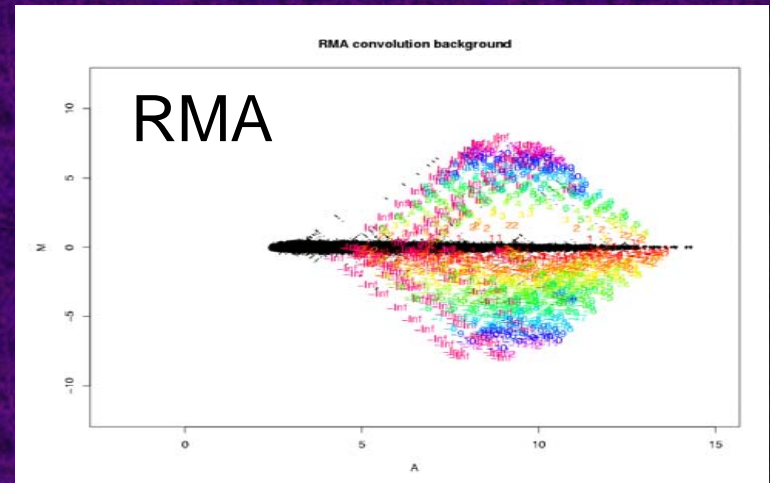
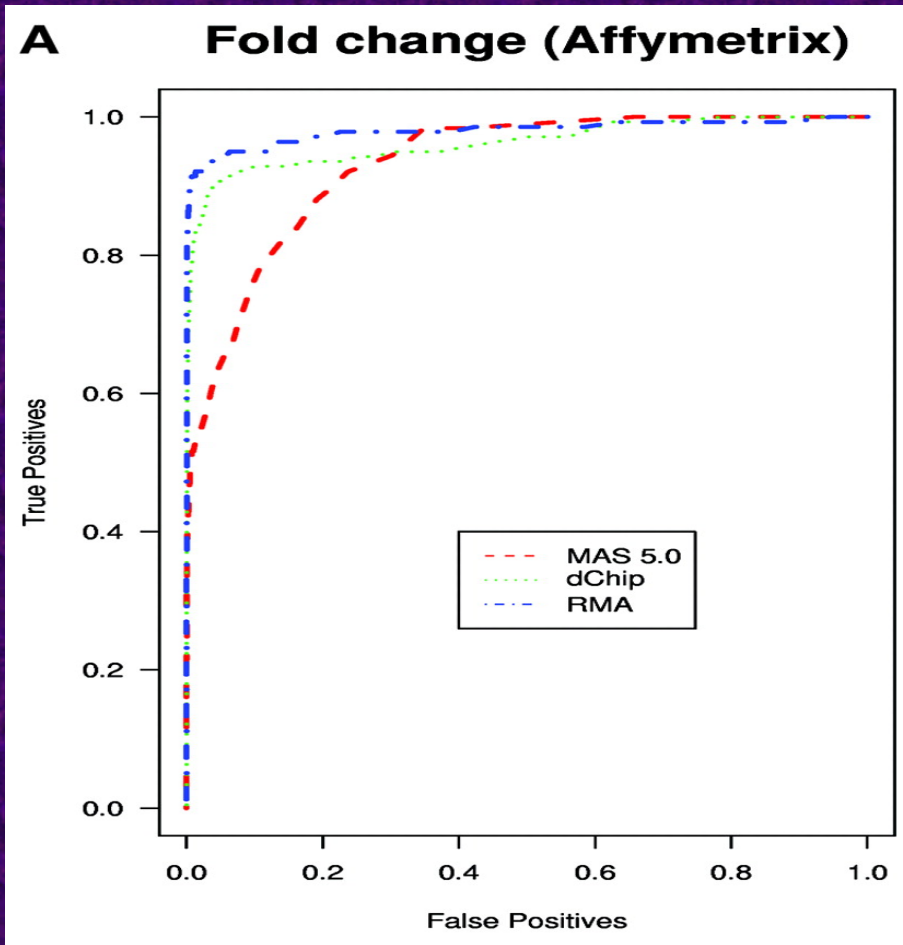
$\hat{\varepsilon}_{11}$	$\cdots$	$\hat{\varepsilon}_{1J}$	$\hat{\alpha}_1$
$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\hat{\varepsilon}_{I1}$	$\cdots$	$\hat{\varepsilon}_{IJ}$	$\hat{\alpha}_I$
$\hat{\beta}_1$	$\cdots$	$\hat{\beta}_J$	$\hat{m}$

Imposes  
Constraints

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

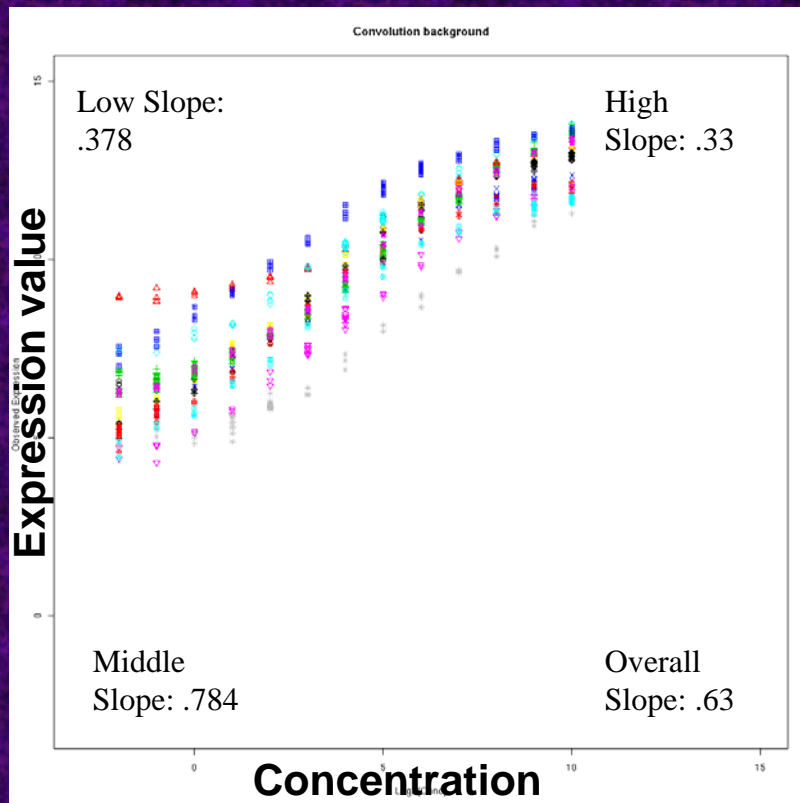
# RMA Mostly Does Well in Practice

Detecting Differential Expression    Not noisy in low intensities

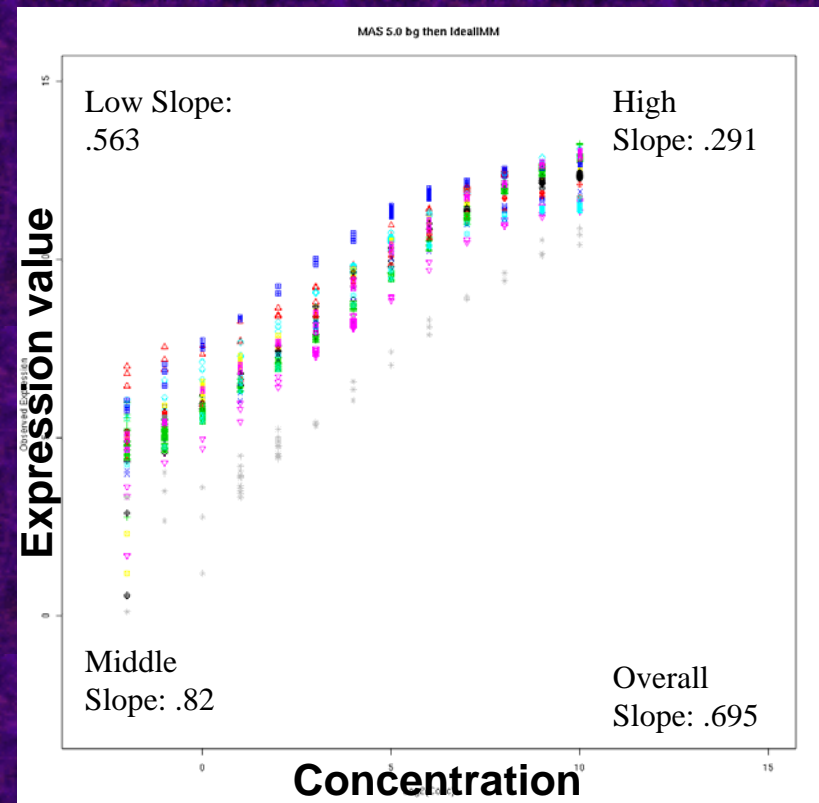


# One Drawback

RMA



MAS 5.0



GCRMA (Irizarry and Wu, JHU) which is RMA with a different background step fixes this to some extent

# More on how expression measures compared here

Competition results: Table 1 - Microsoft Internet Explorer

Address: <http://affycomp.biostat.jhsph.edu/AFFY2/TABLE5.hgu/1.html#table95a>

(^: #genes subnormal)

N	Method / Submitter	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	(perfection)	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
39	<a href="#">RMA_NBG/holstad</a>	0.01	0.02	0.06	0.90	0.09	0.02	0.09	0.10	0.09	0.04	0.54	0.90	0.93	0.63
31	<a href="#">cor523/cope</a>	0.02	0.03	0.12	0.88	0.12	0.06	0.13	0.10	0.12	0.08	0.54	0.77	0.61	0.60
38	<a href="#">W237/dario.greco</a>	0.02	0.04	0.17	0.87	0.12	0.05	0.13	0.10	0.12	0.07	0.35	0.54	0.39	0.39
8	<a href="#">RMAVSN/thomas.cappola</a>	0.02	0.04	0.15	0.89	0.12	0.06	0.13	0.10	0.12	0.08	0.46	0.59	0.43	0.49
40	<a href="#">RMAVSN/thomas.cappola</a>	0.02	0.04	0.15	0.89	0.12	0.06	0.13	0.10	0.12	0.08	0.46	0.59	0.43	0.49
36	<a href="#">gcrma113/zwu</a>	0.06	0.04	0.61	0.91	1.00	0.25	1.13	0.97	1.00	0.48	0.45	0.91	0.92	0.57
25	<a href="#">rsvd.pm/jack.liu</a>	0.06	0.11	0.34	0.89	0.53	0.12	0.53	0.77	0.53	0.16	0.42	0.90	0.96	0.54
34	<a href="#">GS_RMA/thon</a>	0.07	0.13	0.40	0.90	0.68	0.20	0.71	0.80	0.68	0.30	0.56	0.91	0.96	0.65
2	<a href="#">RMA/rafa</a>	0.07	0.13	0.40	0.90	0.68	0.20	0.71	0.80	0.68	0.31	0.57	0.91	0.96	0.65
26	<a href="#">rma-tog/dgreco</a>	0.07	0.13	0.40	0.90	0.68	0.20	0.71	0.80	0.68	0.31	0.57	0.91	0.96	0.65
35	<a href="#">GS_GCRMA/thon</a>	0.07	0.09	0.65	0.93	0.93	0.37	0.96	0.96	0.93	0.55	0.59	0.87	0.90	0.66
30	<a href="#">rsvd.bgc/jack.liu</a>	0.08	0.14	0.52	0.89	0.58	0.16	0.59	0.79	0.58	0.22	0.38	0.80	0.90	0.49
28	<a href="#">LW1/dgreco</a>	0.08	0.14	1.18	0.91	0.59	0.19	0.62	0.74	0.59	0.25	0.23	0.47	0.55	0.29
23	<a href="#">rsvd/jack.liu</a>	0.14	0.12	0.73	0.94	0.74	0.31	0.78	0.73	0.74	0.43	0.53	0.73	0.71	0.58
29	<a href="#">LW2/dgreco</a>	0.14	0.25	13.88	0.56	1.08	1.50	0.80	0.68	1.08	1.45	0.19	0.00	0.00	0.14
33	<a href="#">UM-Tr-Mu/jmacdon</a>	0.15	0.25	1.86	0.93	0.70	0.36	0.72	0.70	0.70	0.44	0.18	0.10	0.10	0.16
37	<a href="#">rsvd2/jack.liu</a>	0.17	0.28	1.74	0.91	0.75	0.46	0.74	0.81	0.75	0.52	0.29	0.16	0.21	0.26
27	<a href="#">rma-sep/dgreco</a>	0.18	0.28	0.96	0.90	0.71	0.27	0.72	0.84	0.71	0.39	0.38	0.53	0.63	0.42
1	<a href="#">MAS_5.0/rafa</a>	0.29	0.47	4.01	0.91	0.77	0.58	0.73	0.77	0.77	0.64	0.09	0.00	0.00	0.06
0	(perfection)	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Done Internet

# RMA Model is a Special Case of the General Probe Level Model

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

- Where  $f(\mathbf{X})$  is function of factor (and possibly covariate) variables (our interest will be in linear functions)
- $y_{kij}$  is a pre-processed probe intensity (usually log scale)
- Assume that  $E[\varepsilon_{kij}] = 0$

$$\text{Var}[\varepsilon_{kij}] = \sigma_k^2$$

# Limitations of Median Polish

- No algorithmic flexibility to fit alternative models
- No standard error estimates
- But it does have some advantages which is why it has been used
  - Fast
  - Very robust

# An Alternative Method for Fitting a Probe-level Model (PLM)

- Robust regression using M-estimation
- In this talk, we will use Huber's influence function  $\psi$ . The software handles many more.
- Fitting algorithm is IRLS with weights dependent on current residuals  $\frac{\psi(r_{kij})}{r_{kij}}$
- Software for fitting such models is part of BioConductor package *affyPLM*

# 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

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

# Summarization PLM

- Array effect model

$$y_{kij} = \alpha_{ki} + \beta_{kj} + \varepsilon_{kij}$$

Pre-processed  
Log PM intensity

Probe Effect

Array Effect

With constraint  $\sum_{i=1}^I \alpha_{ki} = 0$

ie basically same model as before, just different fitting mechanism

# Detecting Differential Expression

- Problem: Given an experiment with two treatment groups correctly identify the differential genes without incorrectly choosing non differential genes.
- Question 1: How do different methods for DDE compare?
- Question 2: Can we do better using PLM's?

# How Do We Know Which Genes are Differential?

- Spike-in datasets.
  - Transcripts at known concentrations differing across arrays. Common background cRNA.
  - Typically, Latin Square design

# 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

# Testing for Differential Expression

- On a probeset by probeset basis compute a test statistic
- Should include something related to observed FC and some variability estimate
- A “t” statistic: something of the form  $t = \frac{\bar{X}}{SE}$

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

ie Average expression  
measures across arrays in  
group

# Expression Values Based Statistics

Fold Change	$\bar{X}_l - \bar{X}_m$	FC
Two Sample T-statistic	$(\bar{X}_l - \bar{X}_m) / \sqrt{\frac{s_l^2}{n_l} + \frac{s_m^2}{n_m}}$	T.std
Simple Moderation	$(\bar{X}_l - \bar{X}_m) / \left( \sqrt{\frac{s_l^2}{n_l} + \frac{s_m^2}{n_m}} + s_{\text{med}} \right)$	T.mod
Limma “Ebayes”		T.ebayes
“Robust”	$(\tilde{X}_l - \tilde{X}_m) / \sqrt{\frac{\tilde{s}_l^2}{n_l} + \frac{\tilde{s}_m^2}{n_m}}$	T.robust

# Probe Level Model Test Statistics

- Suppose that  $\Sigma$  is component of the variance-covariance matrix related to  $\beta$
- Let  $\mathbf{c}$  be the contrast vector defined such that the  $j$ th element of  $\mathbf{c}$  is

$$\begin{aligned} & \frac{1}{n_l} \text{ if array } j \text{ is in group } l \\ & -\frac{1}{n_m} \text{ if array } j \text{ is in group } m \end{aligned}$$

0 otherwise

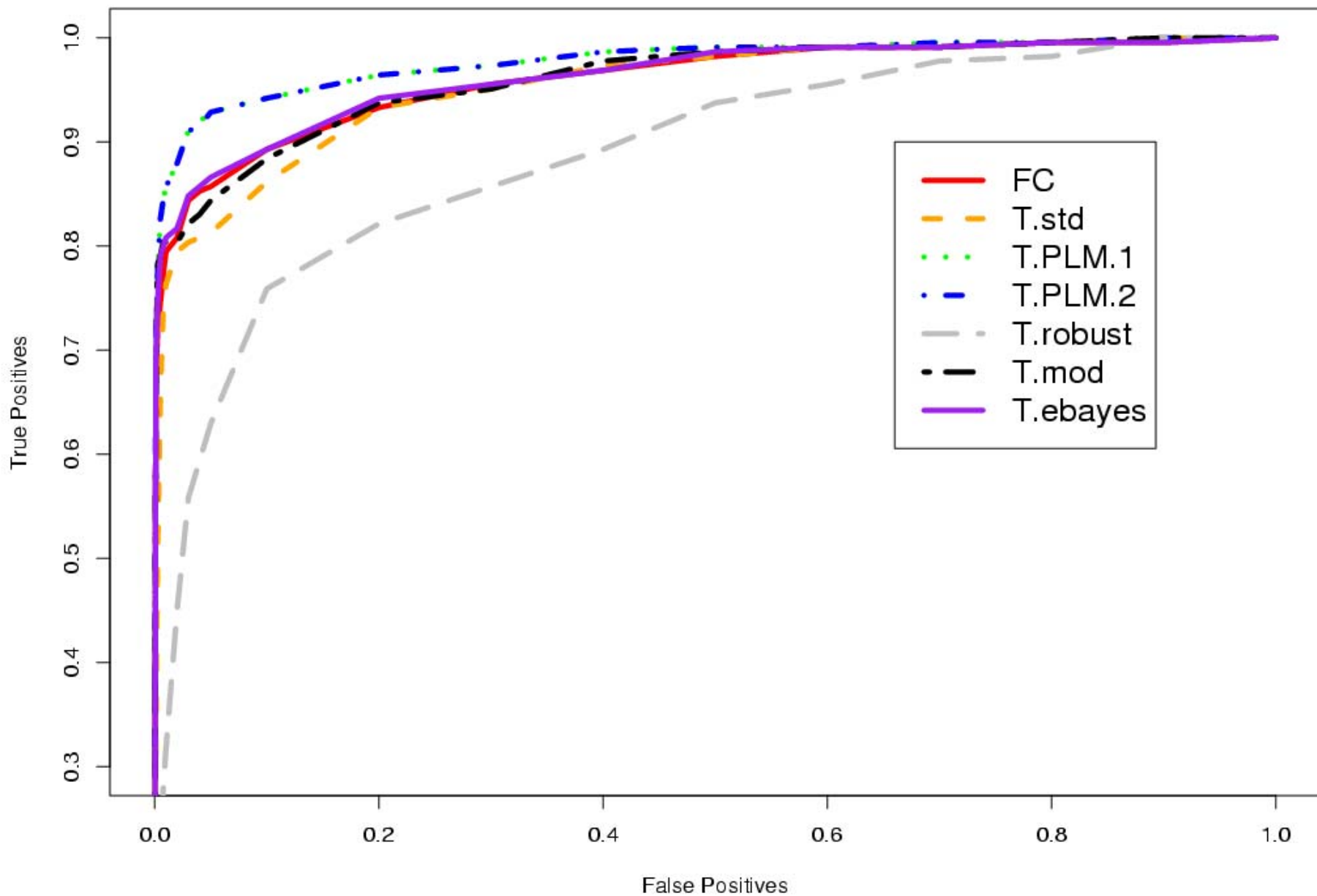
$$t_{\text{PLM.1}} = \frac{\mathbf{c}^T \boldsymbol{\beta}}{\sqrt{\sum_{j=1}^J c_j^2 \Sigma_{jj}}}$$

$$t_{\text{PLM.2}} = \frac{\mathbf{c}^T \boldsymbol{\beta}}{\sqrt{\mathbf{c}^T \boldsymbol{\Sigma} \mathbf{c}}}$$

# 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 3 vs 3 comparisons

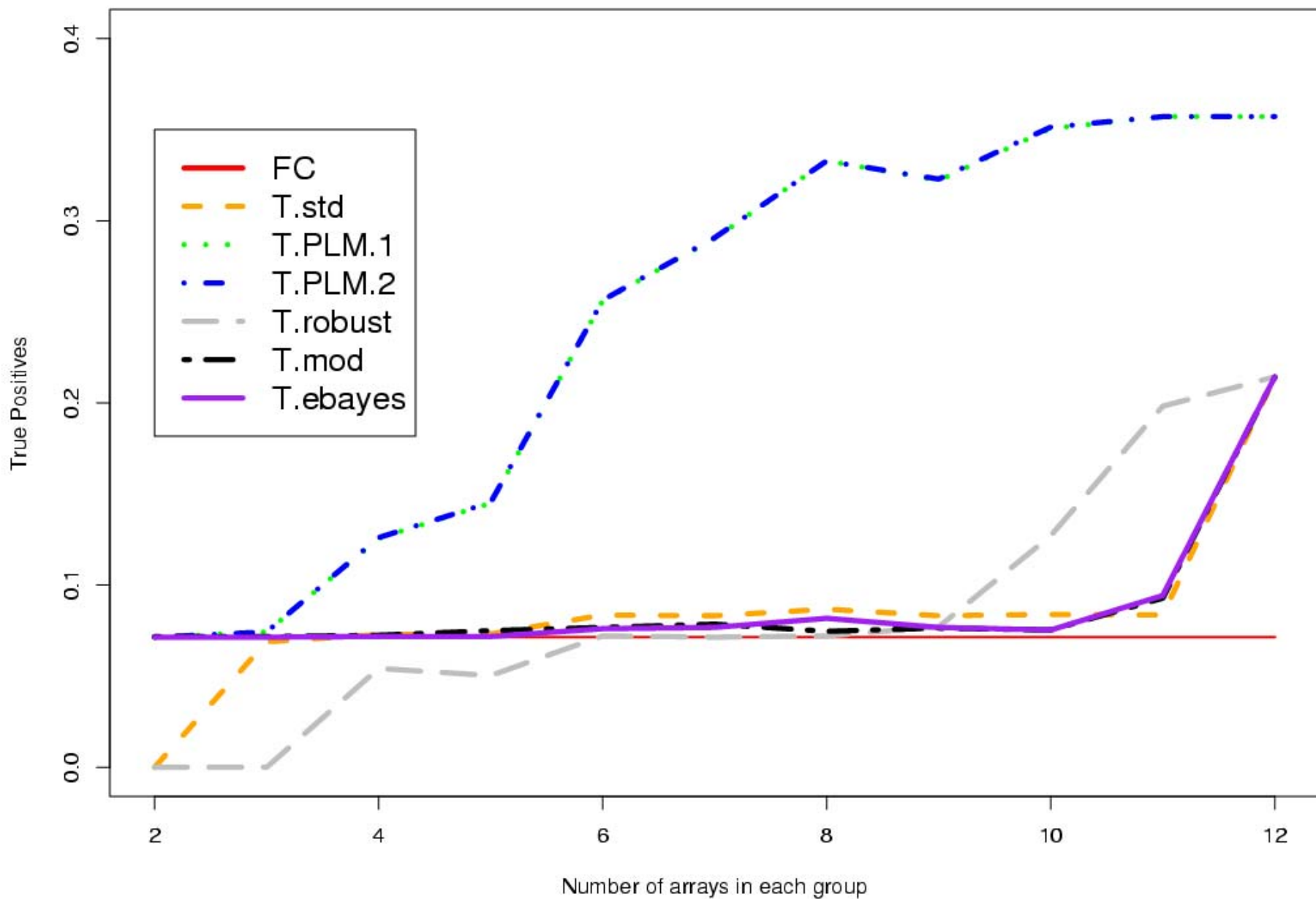
### Affy Spikein: 3 on 3



# 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



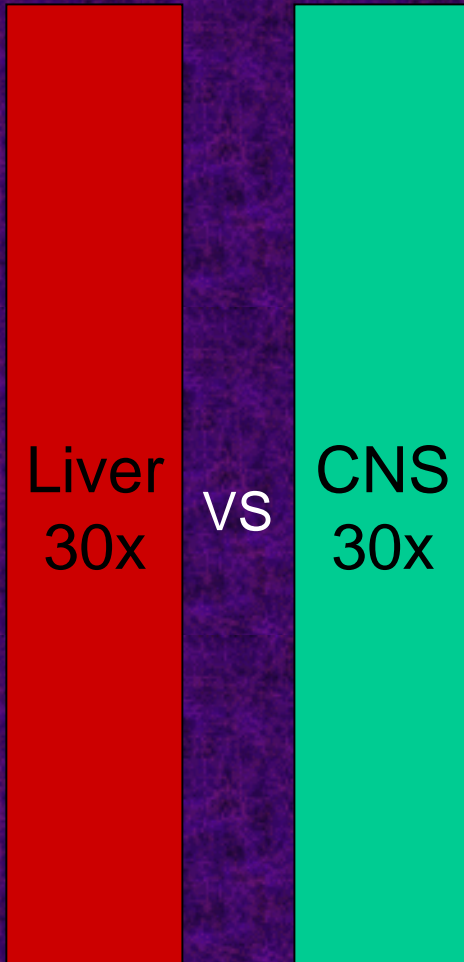




# How About With More “Real” Data?

GeneLogic Dilution/Mixture Dataset

Learning Set



400 top genes  
“truth”



Test Set

Mixture Data

Dilution Series Data

# Mixture Data 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

# Moderation for the PLM test statistic

Method	5 vs 5		
	0% FP	5% FP	AUC
PLM.2	0.058	0.95	0.761
Ebeyes	0	0.943	0.776
PLM Moderated	0.053	0.963	0.795

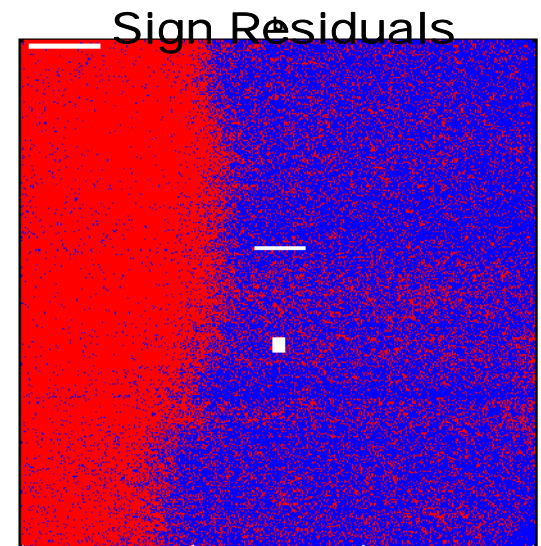
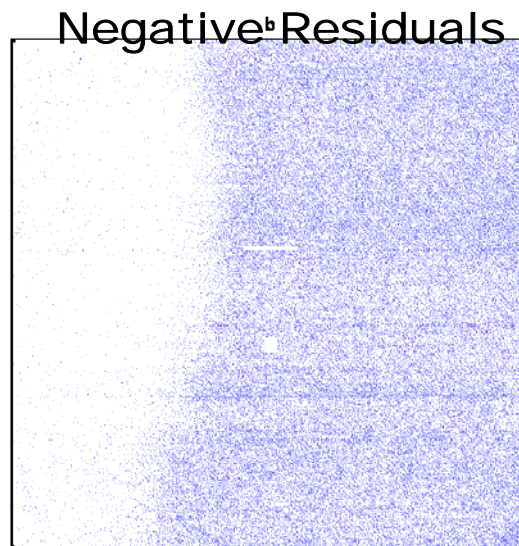
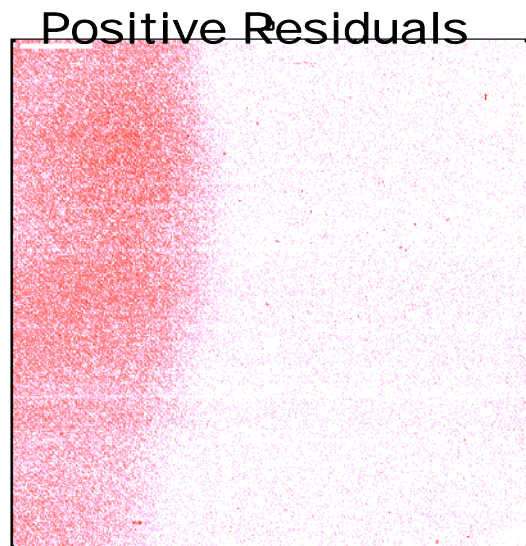
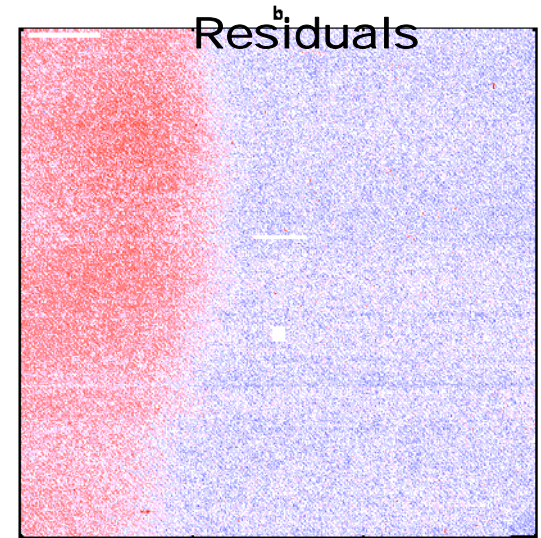
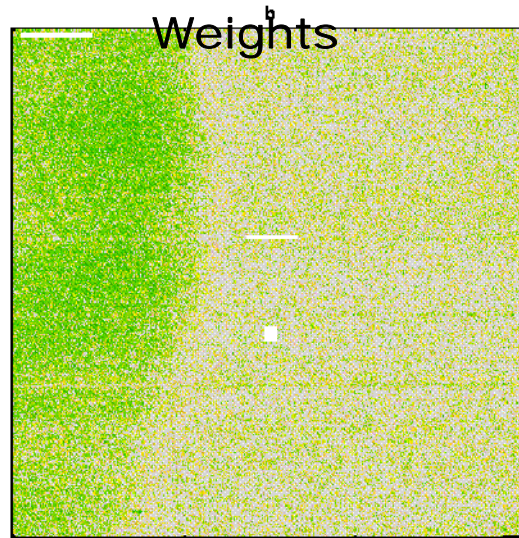
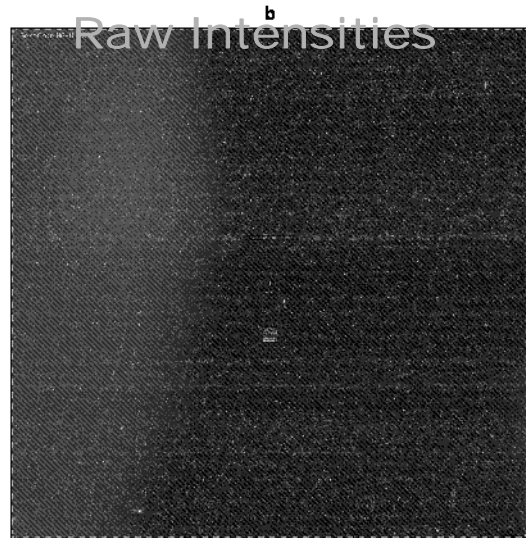
$$\Sigma_{\text{mod}} = (1 - p)\Sigma + p\lambda$$

$\lambda$  is  $\Sigma$  averaged over all probesets

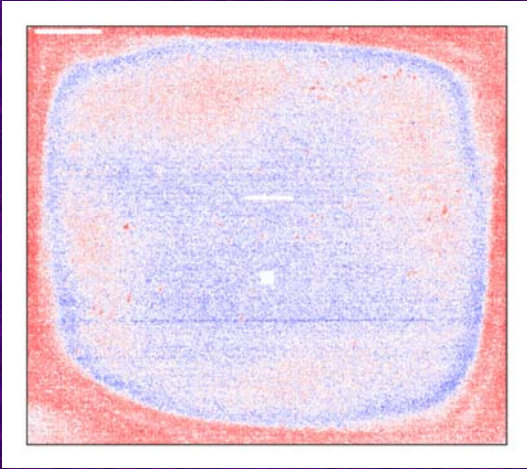
# Quality Assessment

- Problem: Judge quality of chip data
- Question: Can we do this with the output of the Probe Level Modeling procedures?

# Chip Pseudo-images

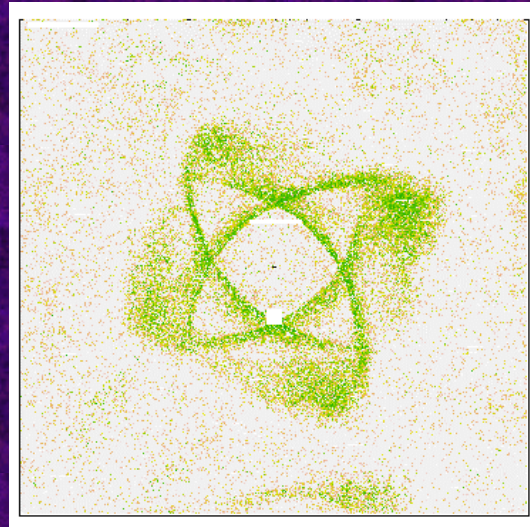


# An Image Gallery

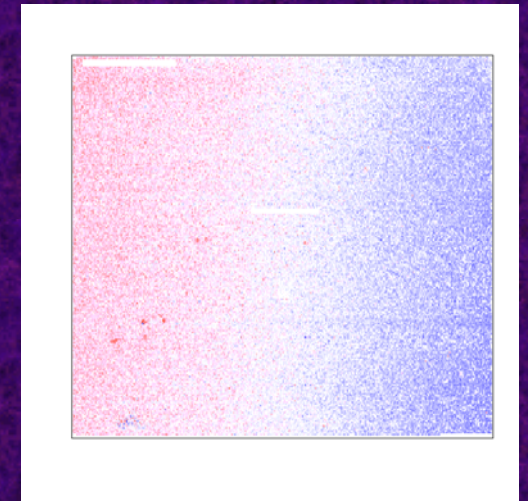


“Ring of Fire”

“Crop Circles”

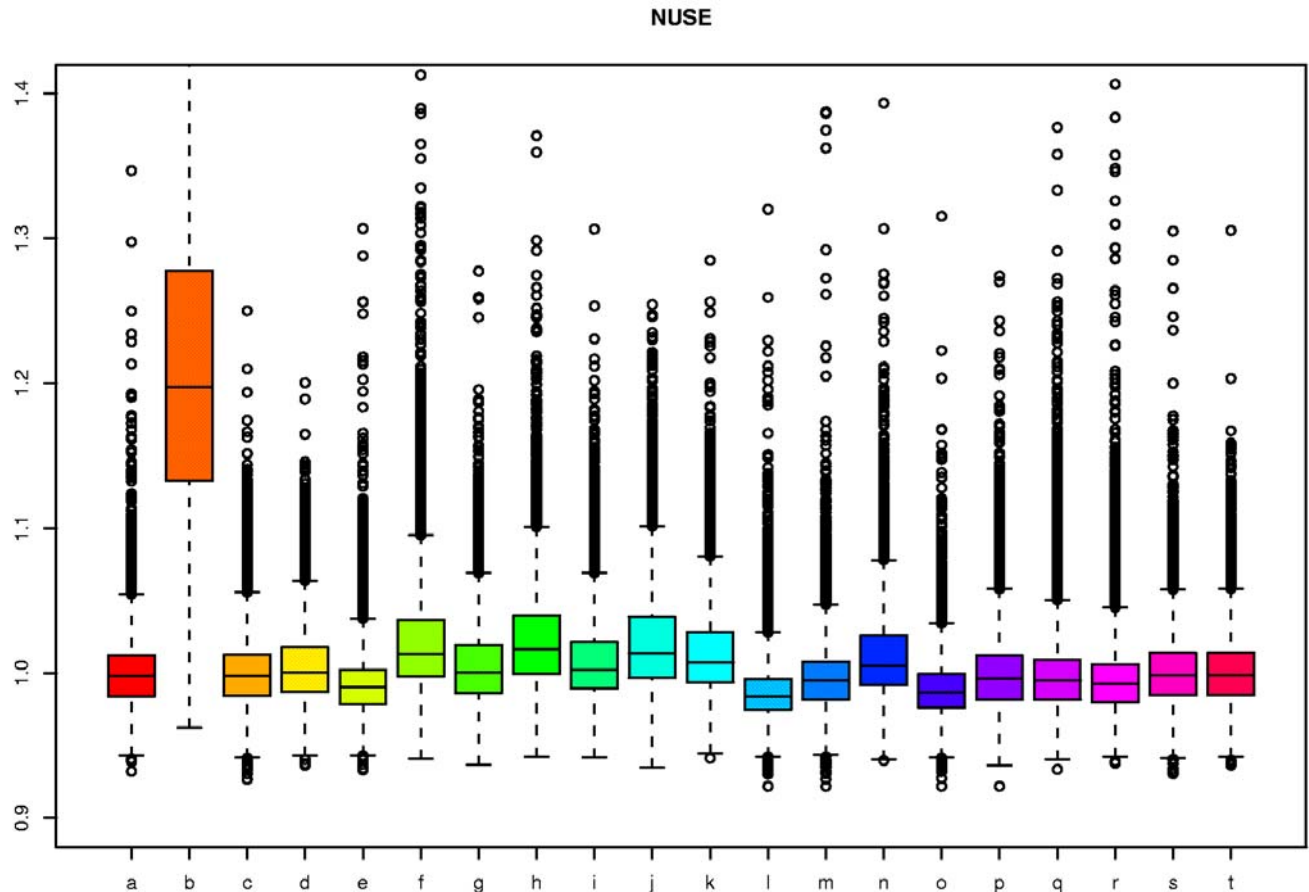


“Tricolor”



# NUSE Plots

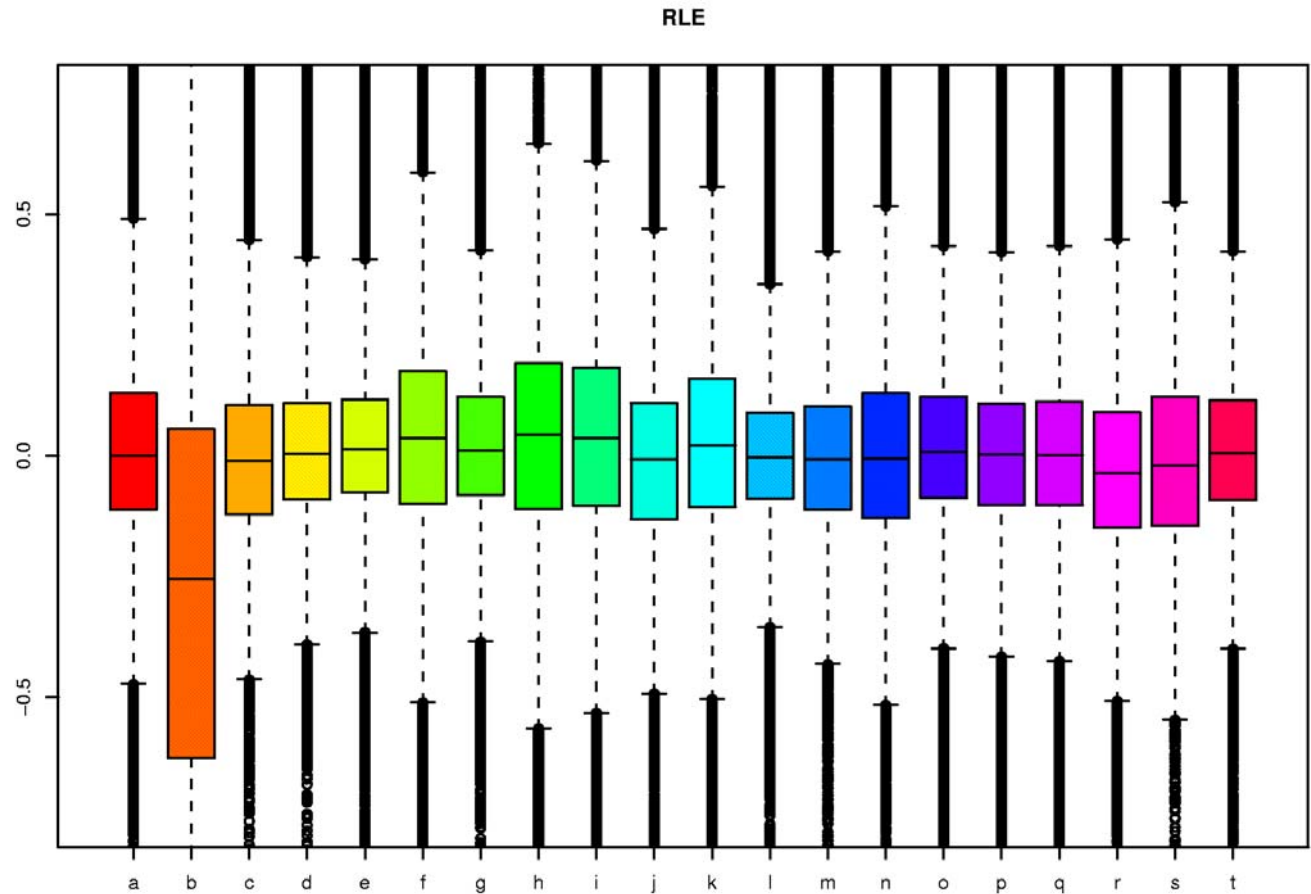
Normalized  
Unscaled  
Standard  
Errors



$$NUSE(\beta_{kj}) = \frac{SE(\beta_{kj})}{\text{Median}_j [SE(\beta_{kj})]}$$

# RLE Plots

Relative  
Log  
Expression

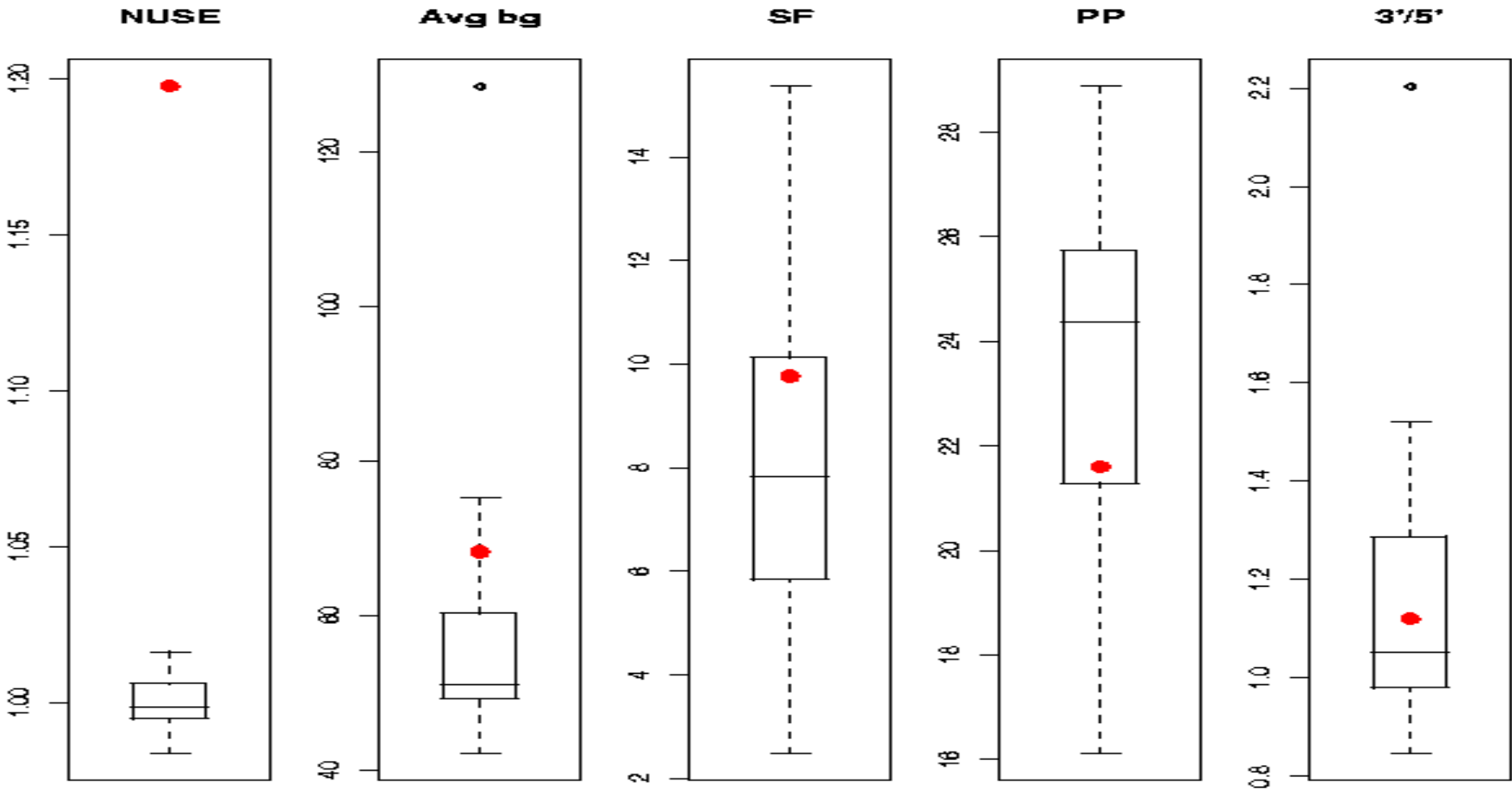


$$RLE(\beta_{kj}) = \beta_{kj} - \text{median}_j(\beta_{kj})$$

# Numeric Quality Summaries for Each Array

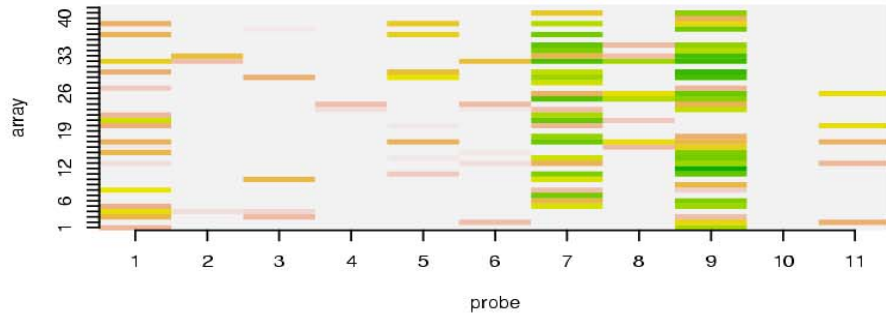
- Median NUSE
- IQR NUSE
- Median RLE
- IQR RLE
  
- Aberrant values indicate quality problem with array and more specifically possible problems with expression values from that array

# Comparing with Affymetrix Quality Measures

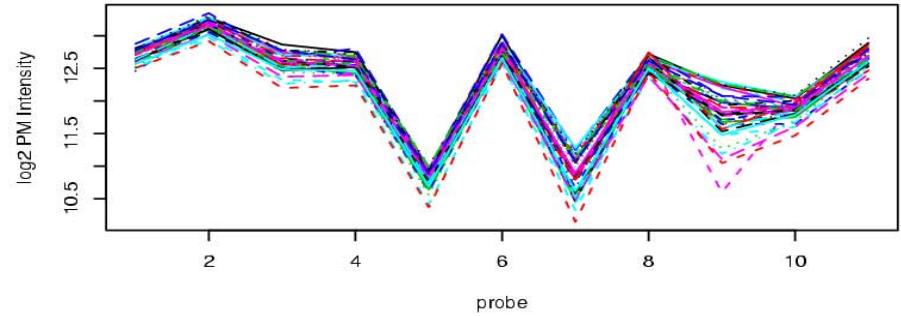


# Discordant Probes

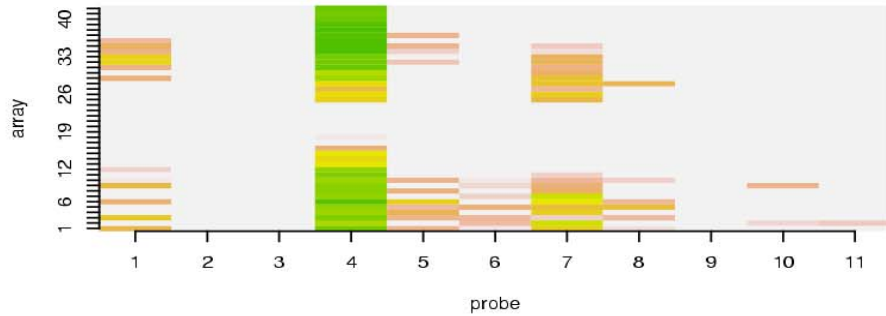
200061\_s\_at



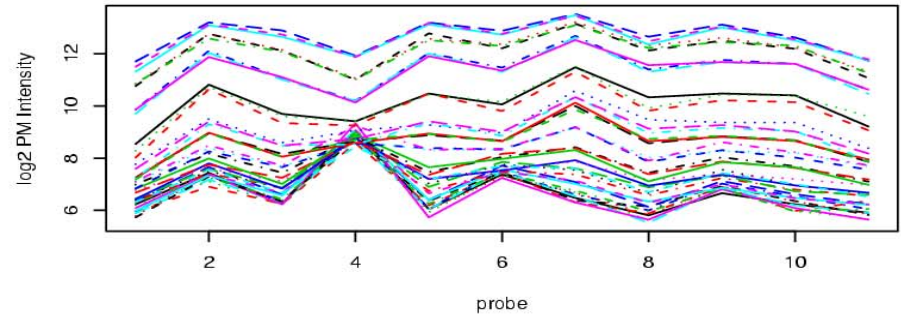
200061\_s\_at



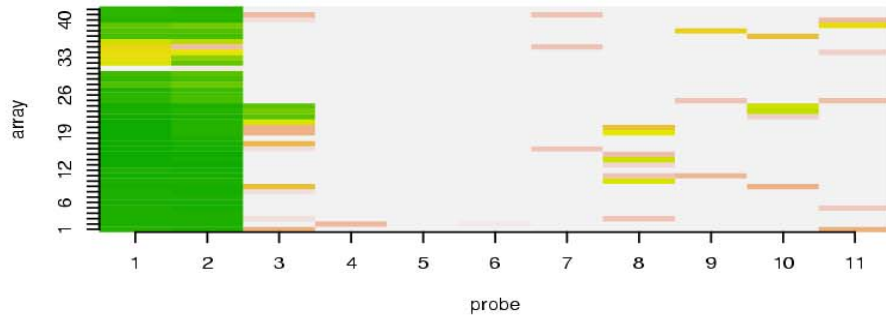
204513\_s\_at



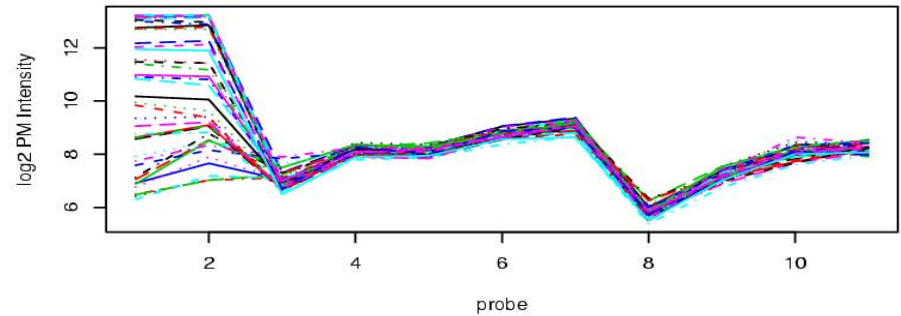
204513\_s\_at



213441\_x\_at

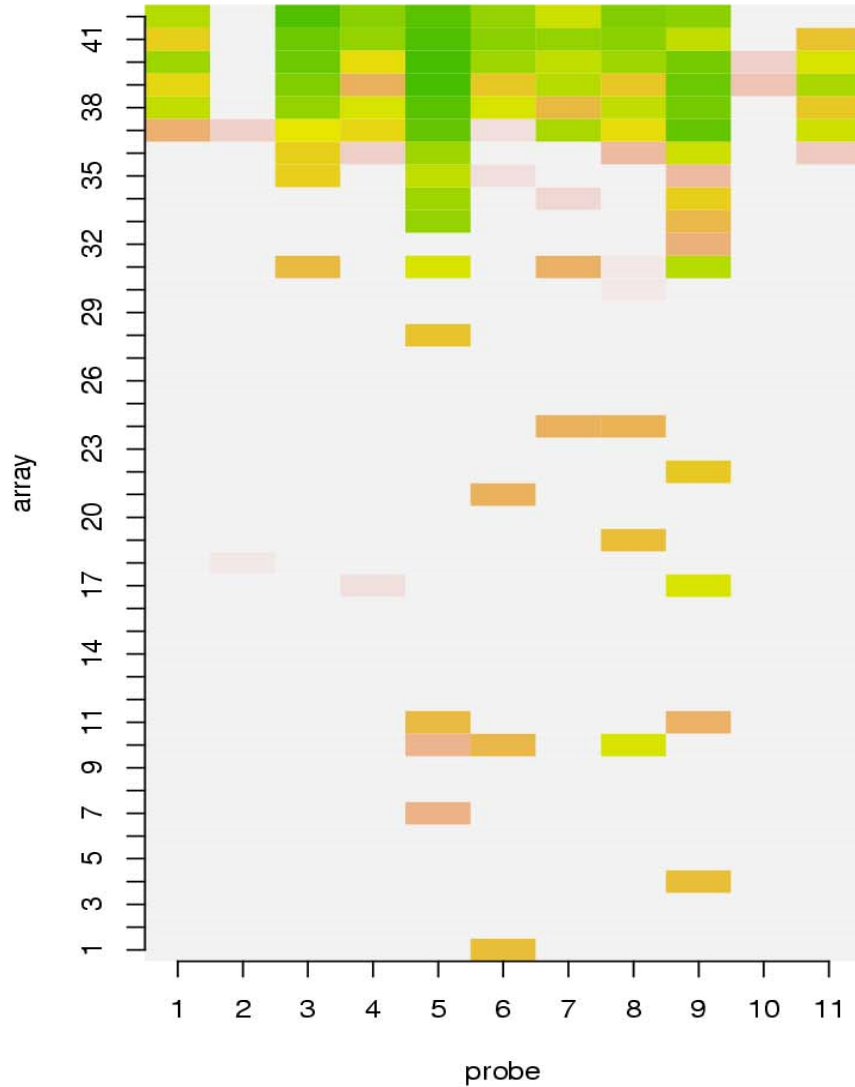


213441\_x\_at

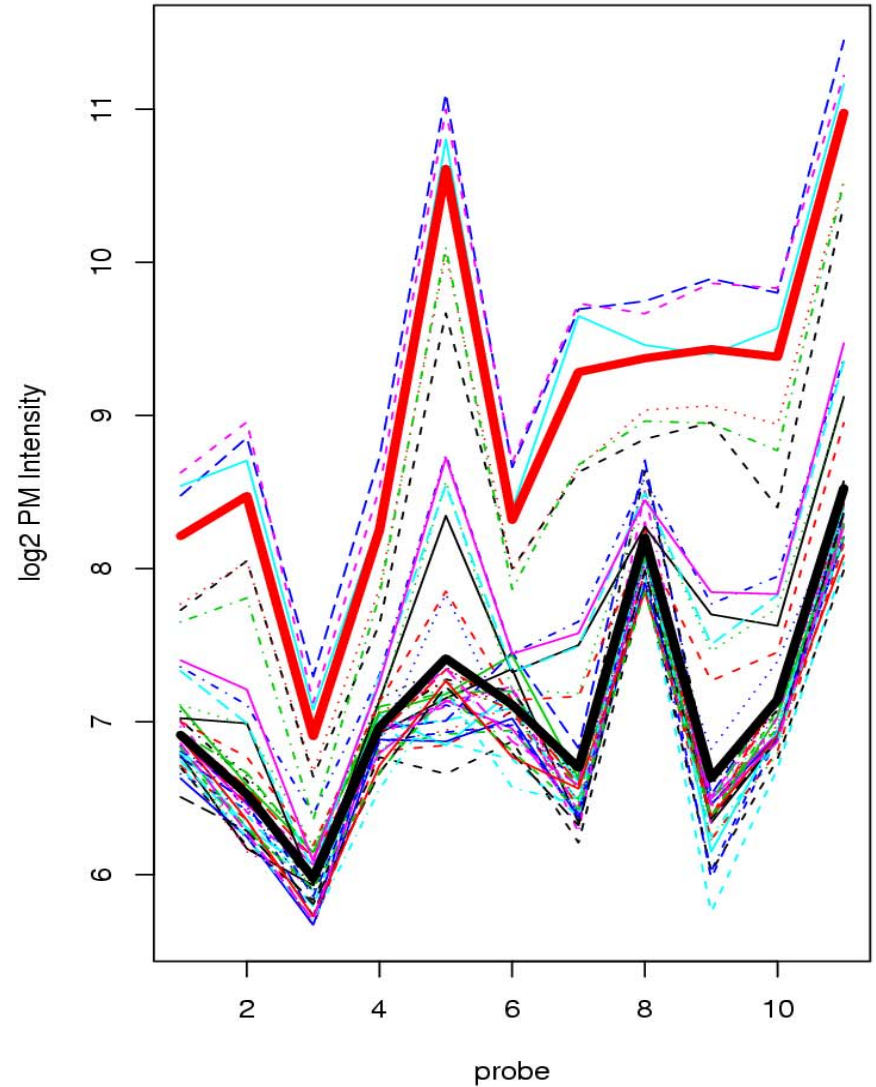


# Discordant Arrays

203173\_s\_at



203173\_s\_at



# Acknowledgements

- Terry Speed (UC Berkeley)
- Julia Brettschneider (UC Berkeley)
- Francois Colin
- Rafael Irizarry (Johns Hopkins)
  
- Bioconductor Core  
<http://www.bioconductor.org>