Statistical Analysis of Low-level High Density Oligonucleotide Array Data

smallTalk 2003
July 15, 2003 San Jose

Ben Bolstad <bolstad@stat.berkeley.edu>
Biostatistics
University of California, Berkeley
http://www.stat.berkeley.edu/~bolstad

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

Where do we start?

We skip image analysis.

We start with probe intensity data from CEL files. It is the probe intensity information that we will use for our low level analysis.

Computing expression summaries: A 3 step process

Background/Signal adjustment (B)

Normalization (N)

Summarization (S)

Let X be cel file data from multiple arrays then

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

Background/Signal Adjustment

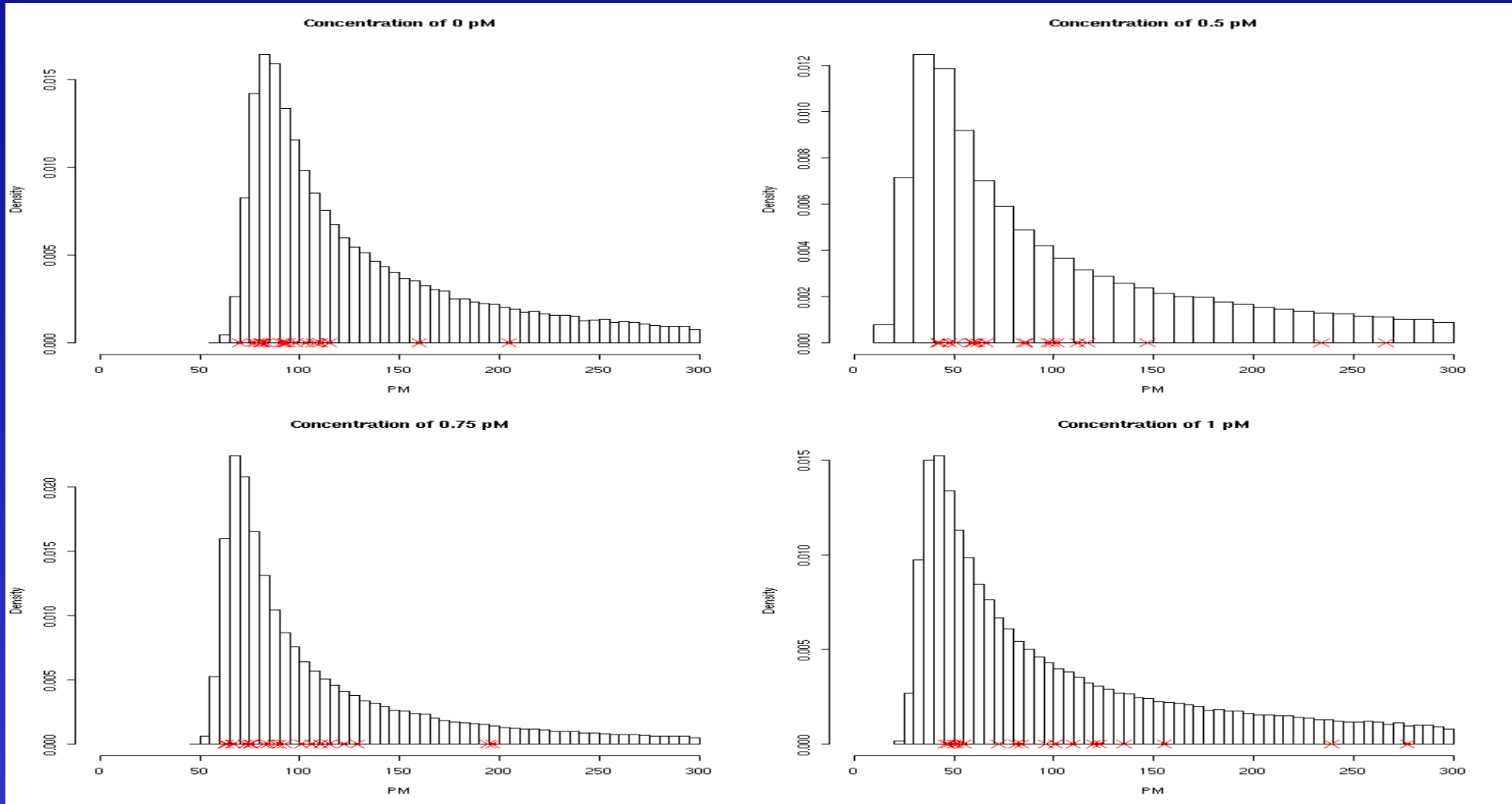
- A method which does some or all of the following
  - Corrects for background noise, processing effects
  - Adjusts for cross hybridization
  - Adjust estimated expression values to fall on proper scale
- Probe intensities are used in background adjustment to compute correction (unlike cDNA arrays where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

RMA convolution model

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



Convolution Model

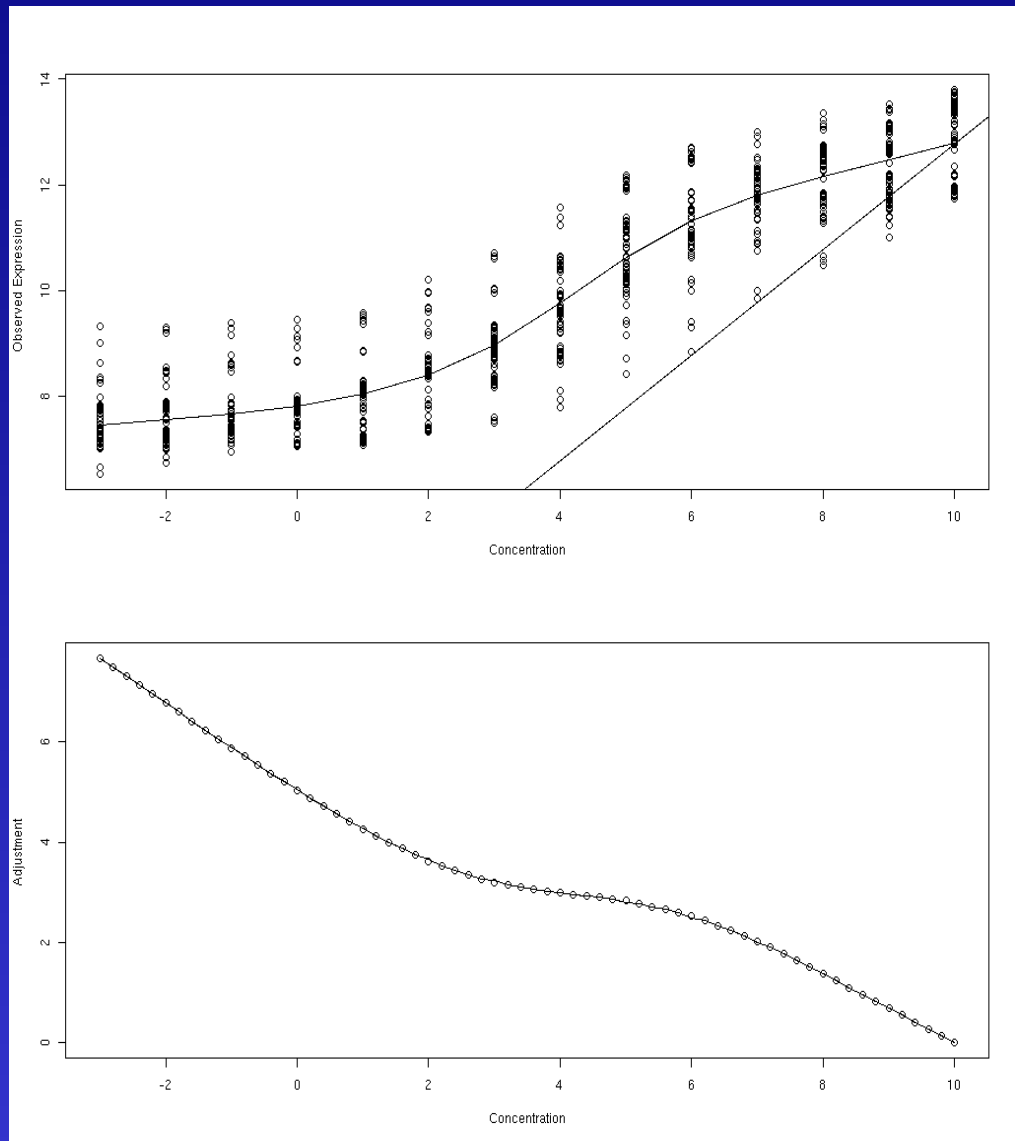
- O = S +N
  - O is observed PM, S is signal (assumed exponential), N is noise (assumed normal, truncated at zero)
- Correction is then

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

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

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-spikeins?

- 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

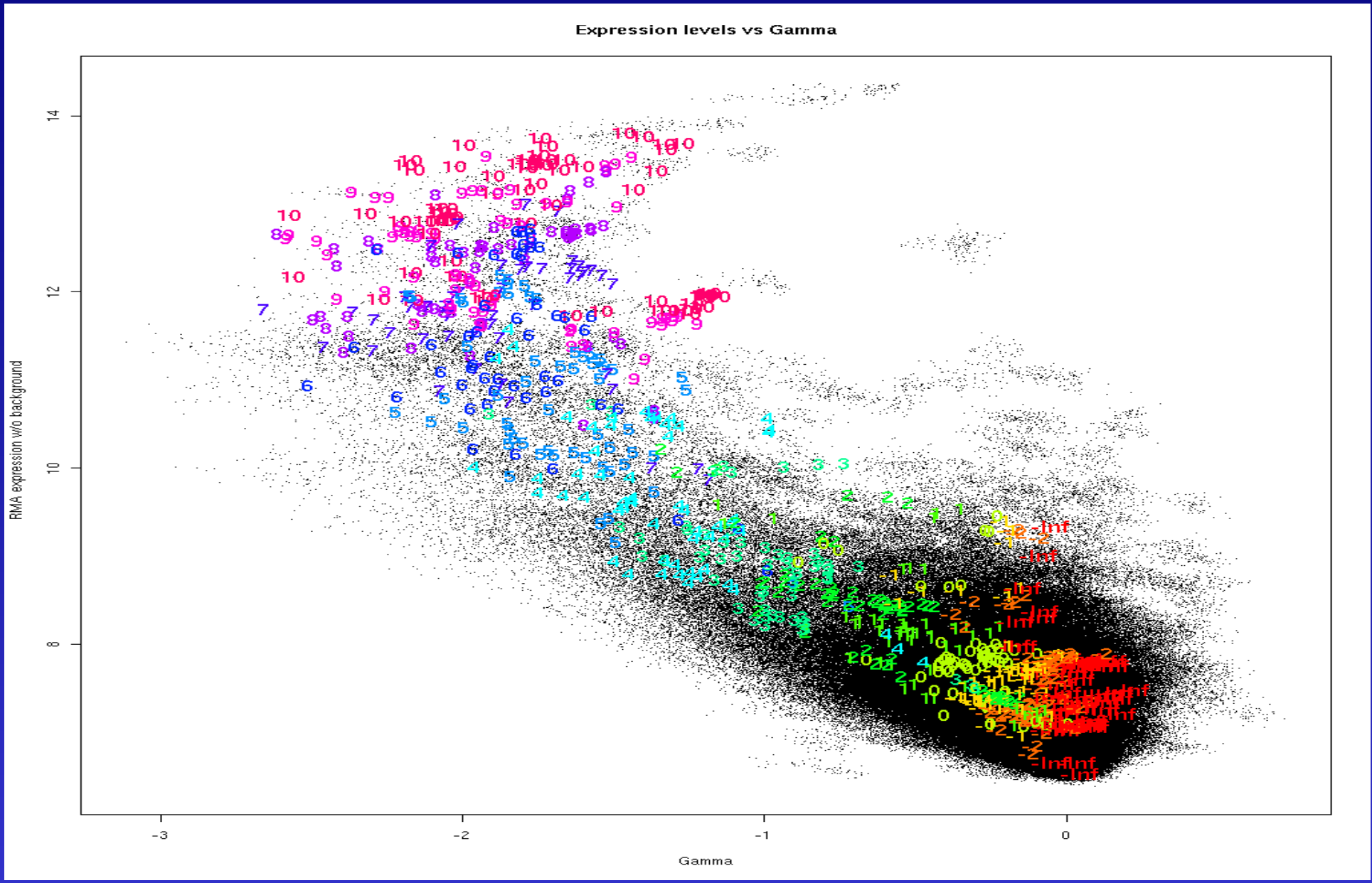
$$y_{1i}^{(k)} = \alpha_i^{(k)} + \varepsilon_i^{(k)}$$

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

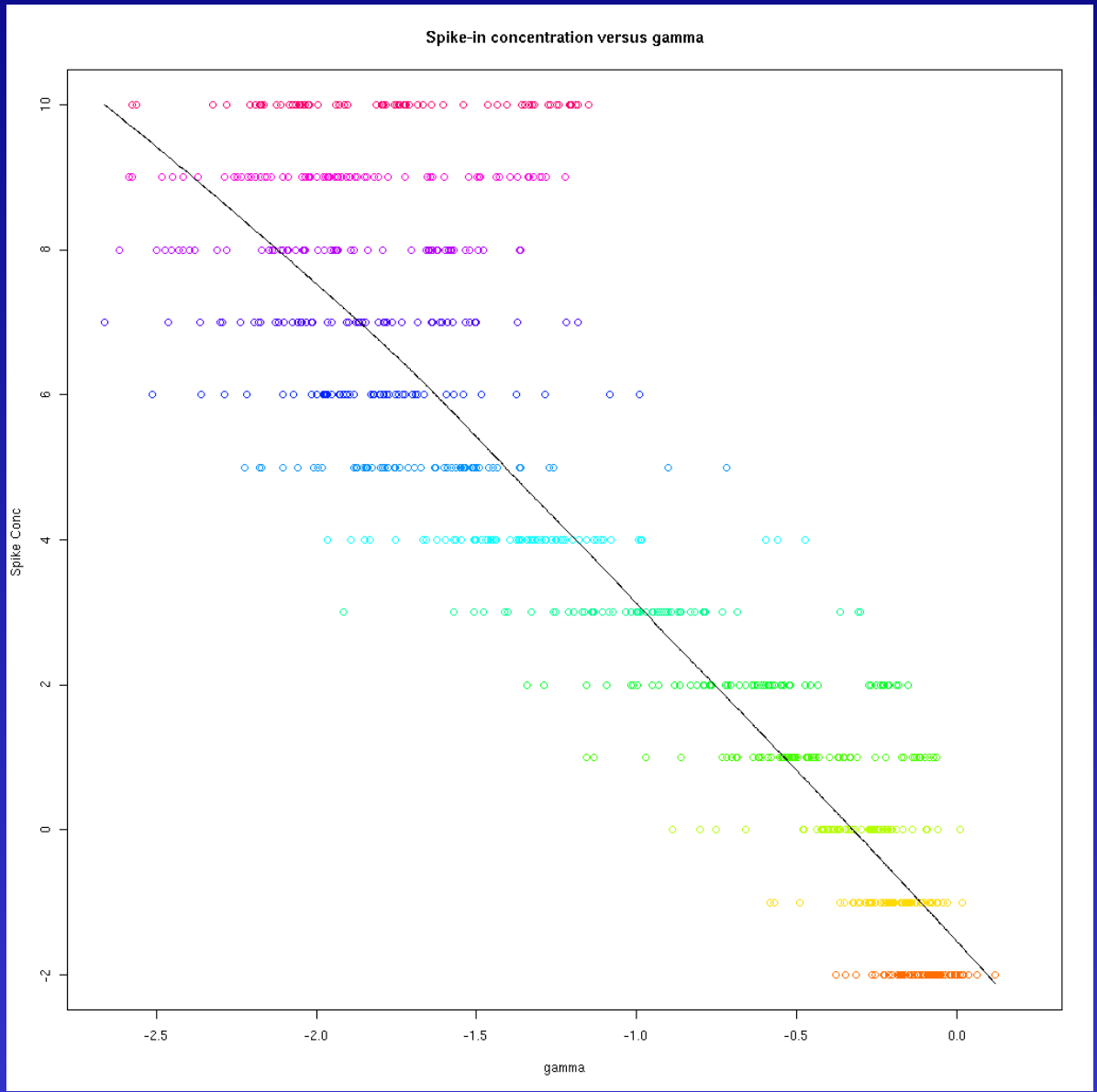
- Where $y_{1i}^{(k)} = \log_2 PM_i^{(k)}$

$$y_{2i}^{(k)} = \log_2 MM_i^{(k)}$$

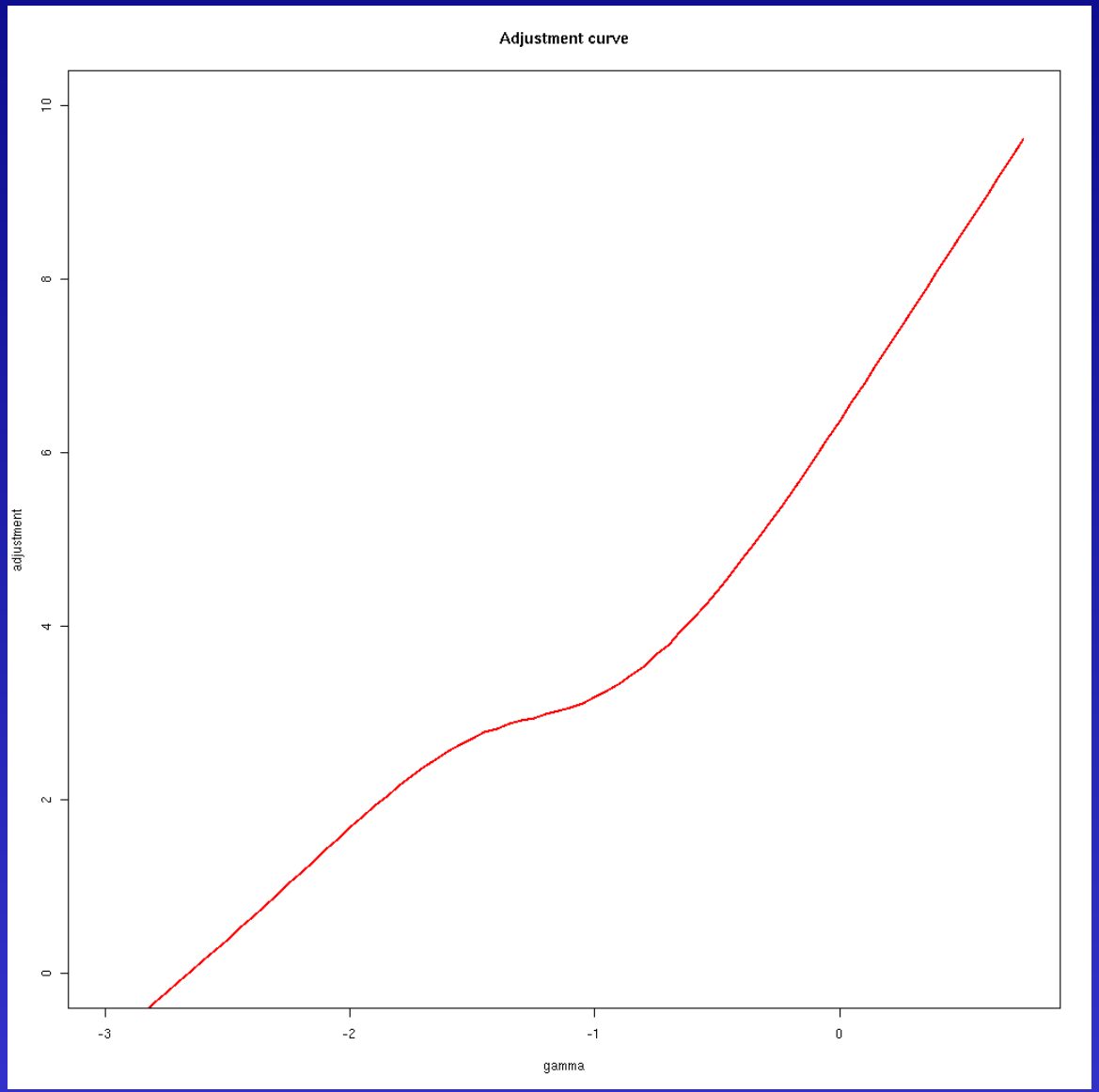
γ Relates to Concentration



Establishing a Relationship Between γ and Concentration



The Two Curves Yield an Adjustment Curve



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

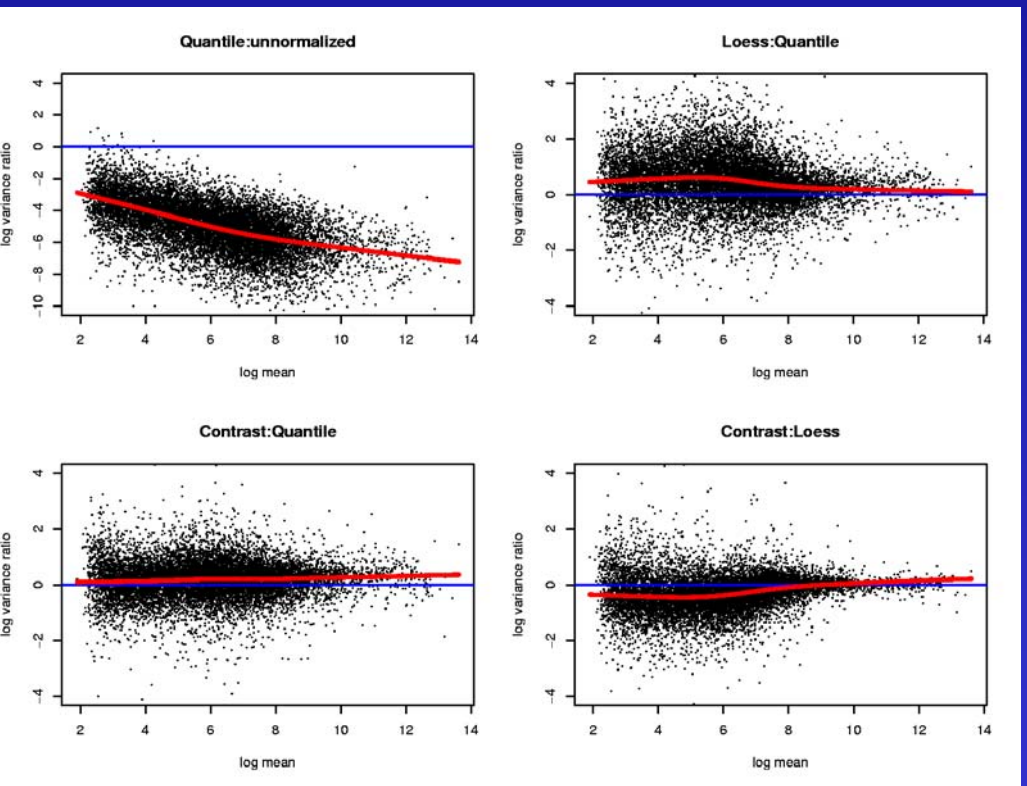
Baseline (normalized using reference chip)

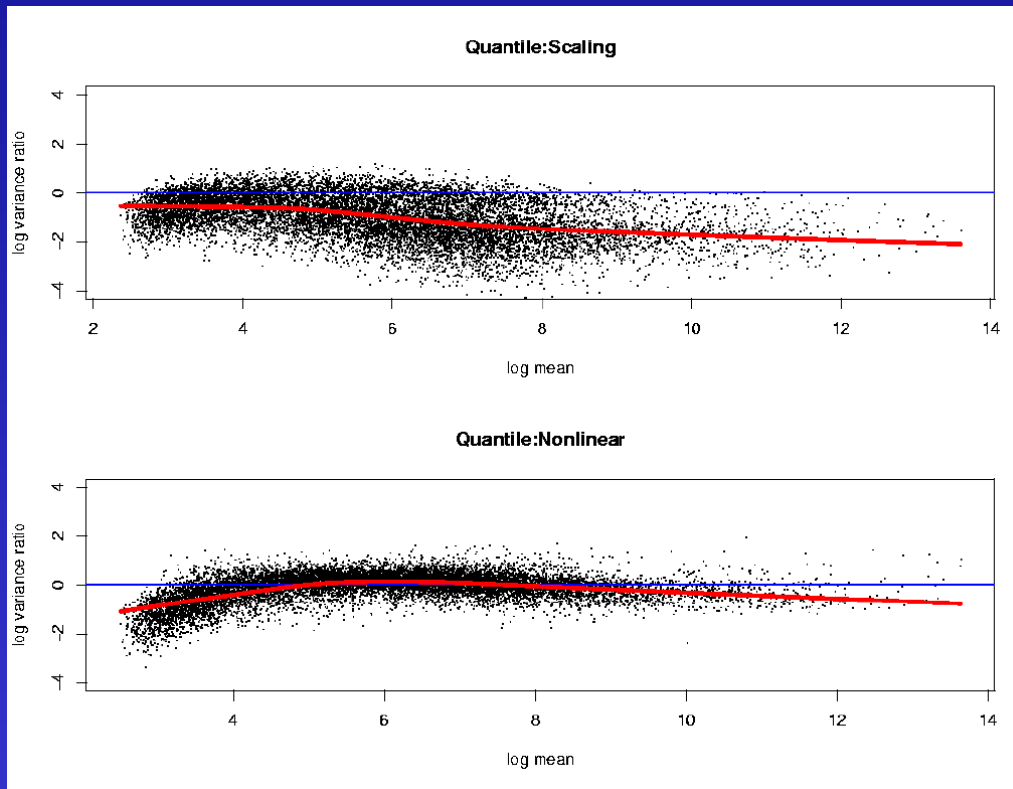
- Scaling (Affymetrix)
- Non linear (Li-Wong)

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

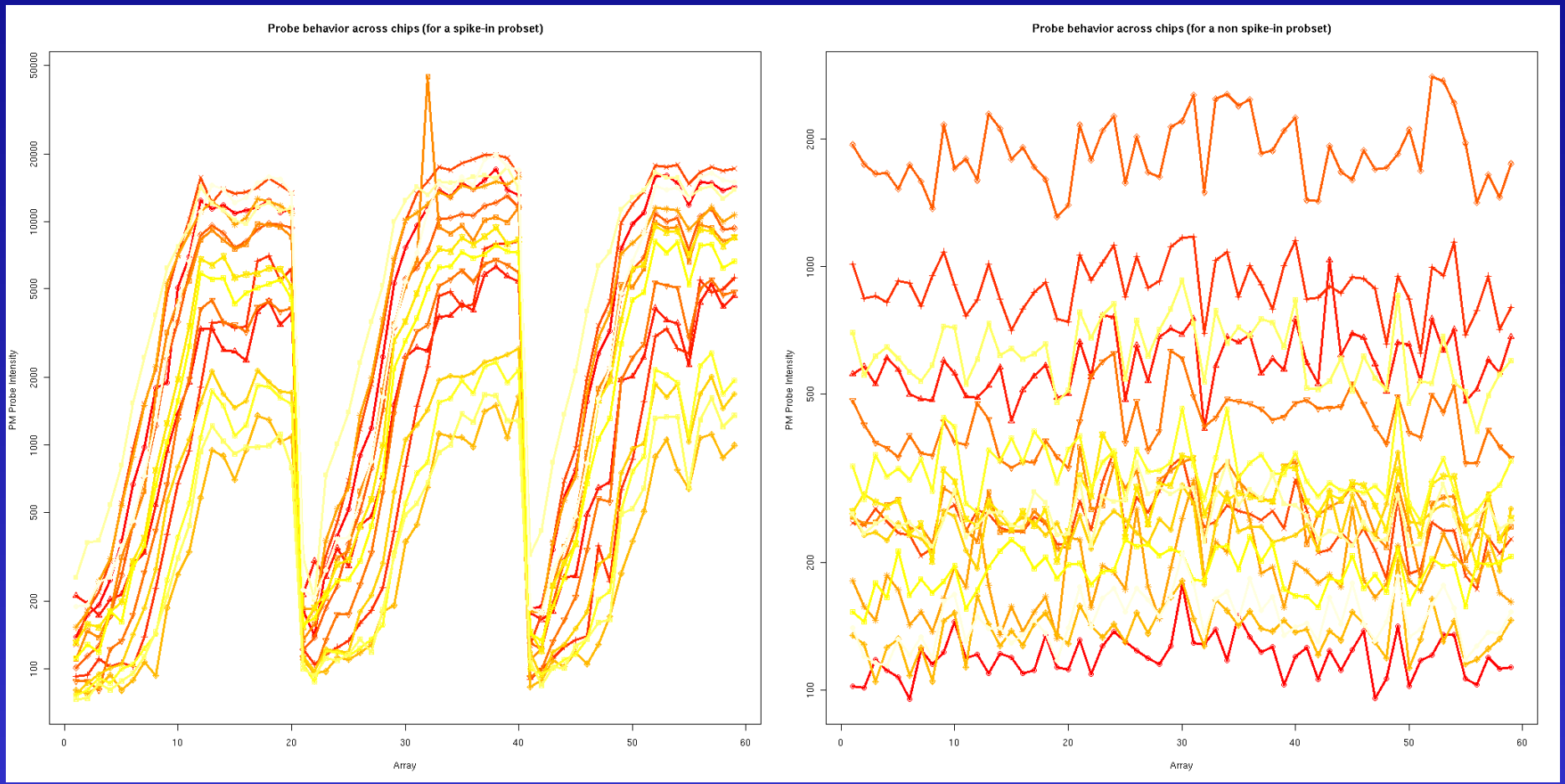




Summarization

- 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

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



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

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

Focus will be on assessing the impact of background adjustment methods

- 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

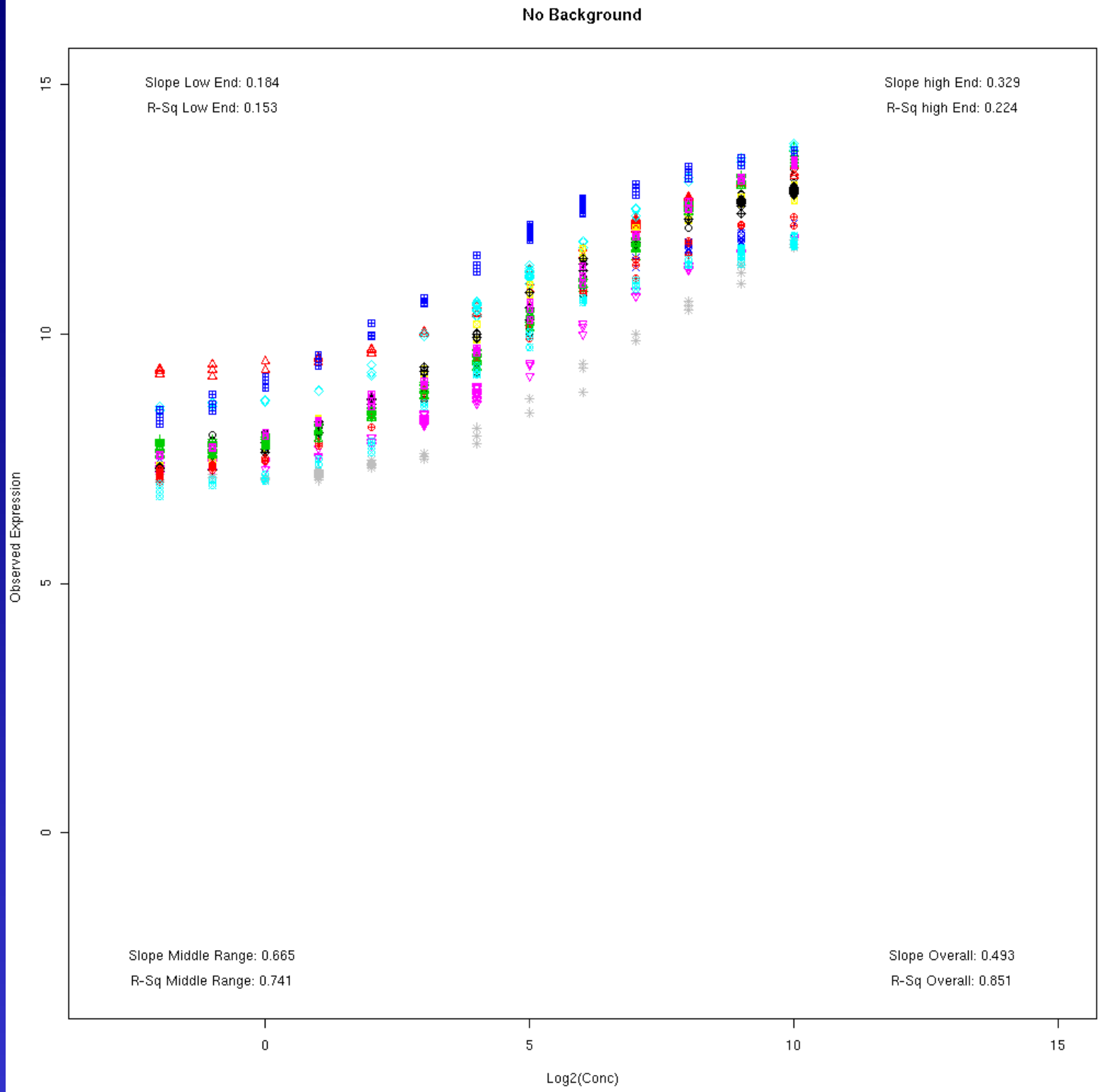
- None
- MAS 5.0 location specific background
- Ideal Mismatch
- MAS 5.0 and Ideal Mismatch
- RMA convolution model
- Using standard curve based on spike-in information to adjust signal

Computing Relative Expression

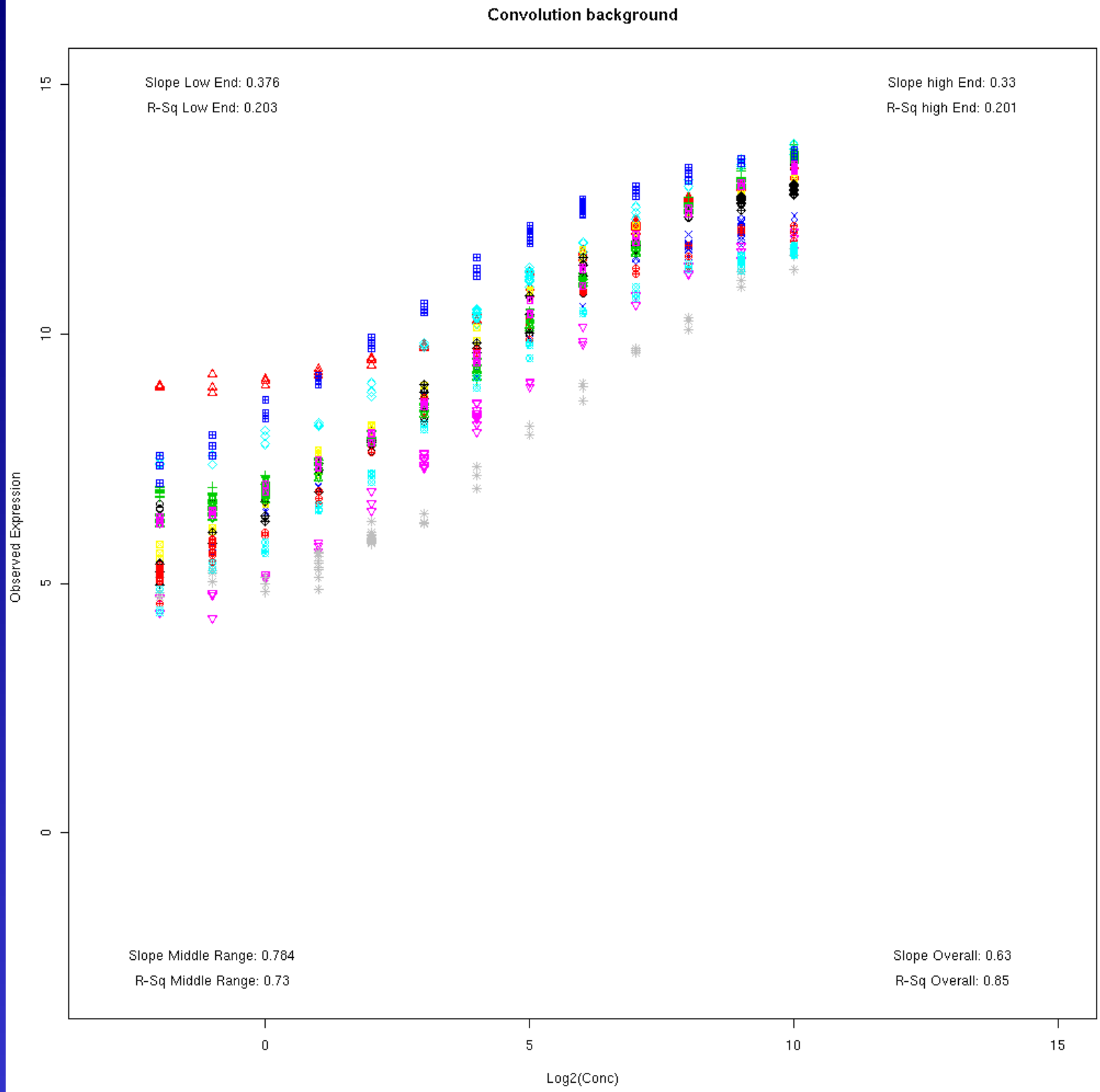
- We will average in log scale across spike-in concentration replicates
- If $E_{i,j}$ is expression of probeset i in group j, then expression difference between group 1 and 2 is
  - $M_i = E_{i,1} - E_{i,2}$
- There are 14 dilution groups so there are 14*13/2 = 91 different comparisons for each probeset

Observed expression versus spike-in concentration

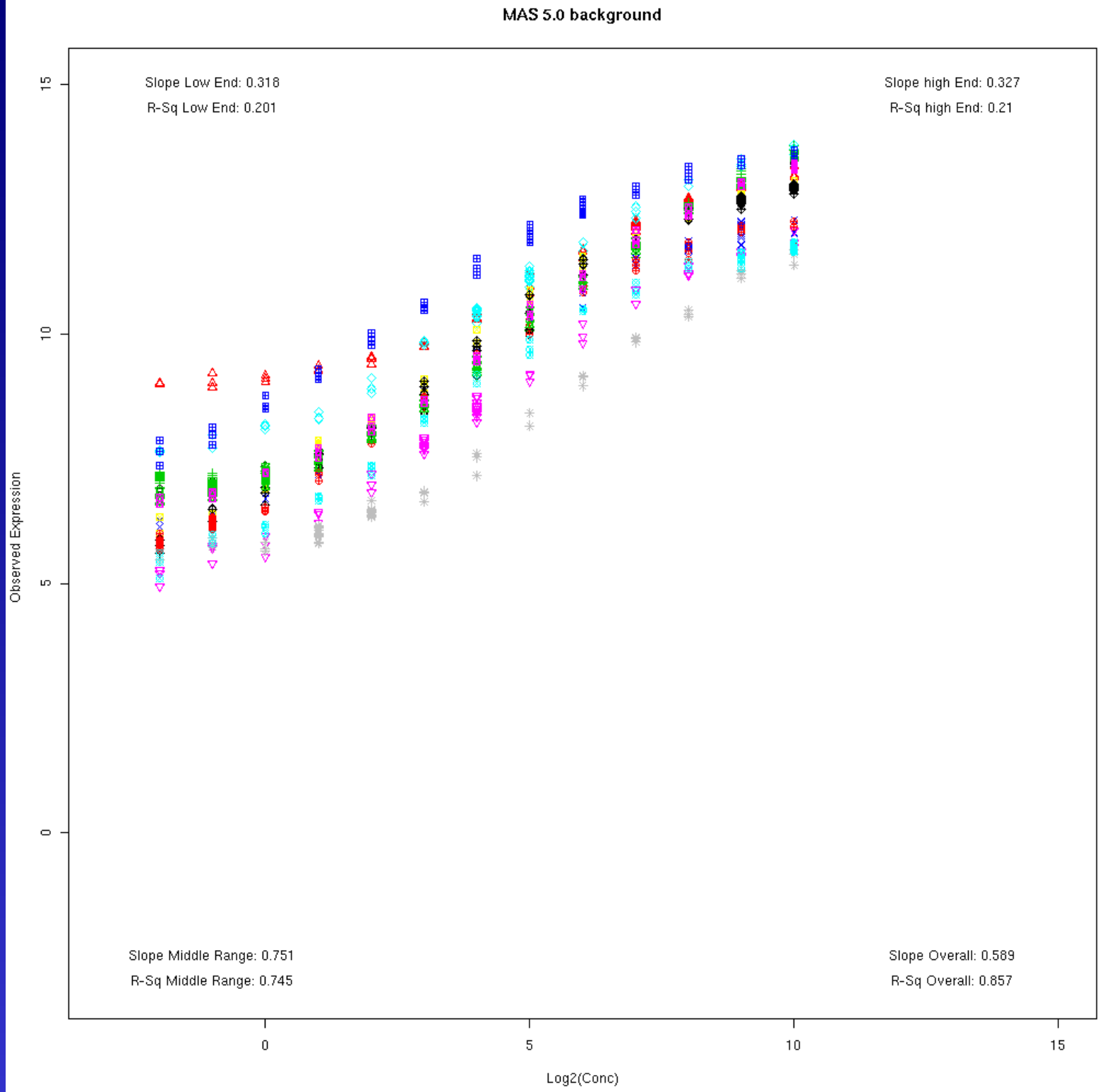
Slope	Value
All	0.493
Mid	0.665
Low	0.184
High	0.329



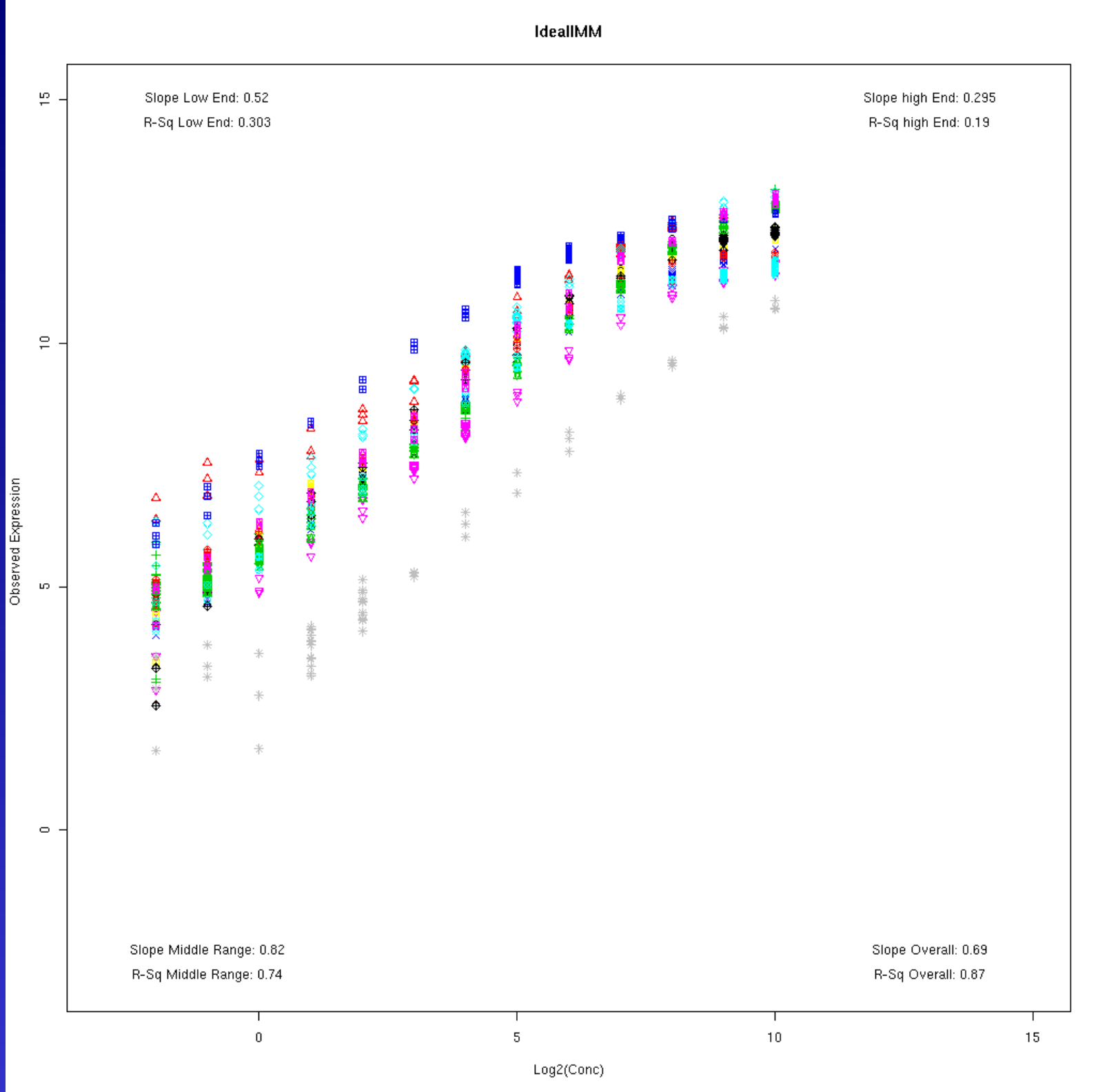
Slope	Value
All	0.63
Mid	0.784
Low	0.376
High	0.33



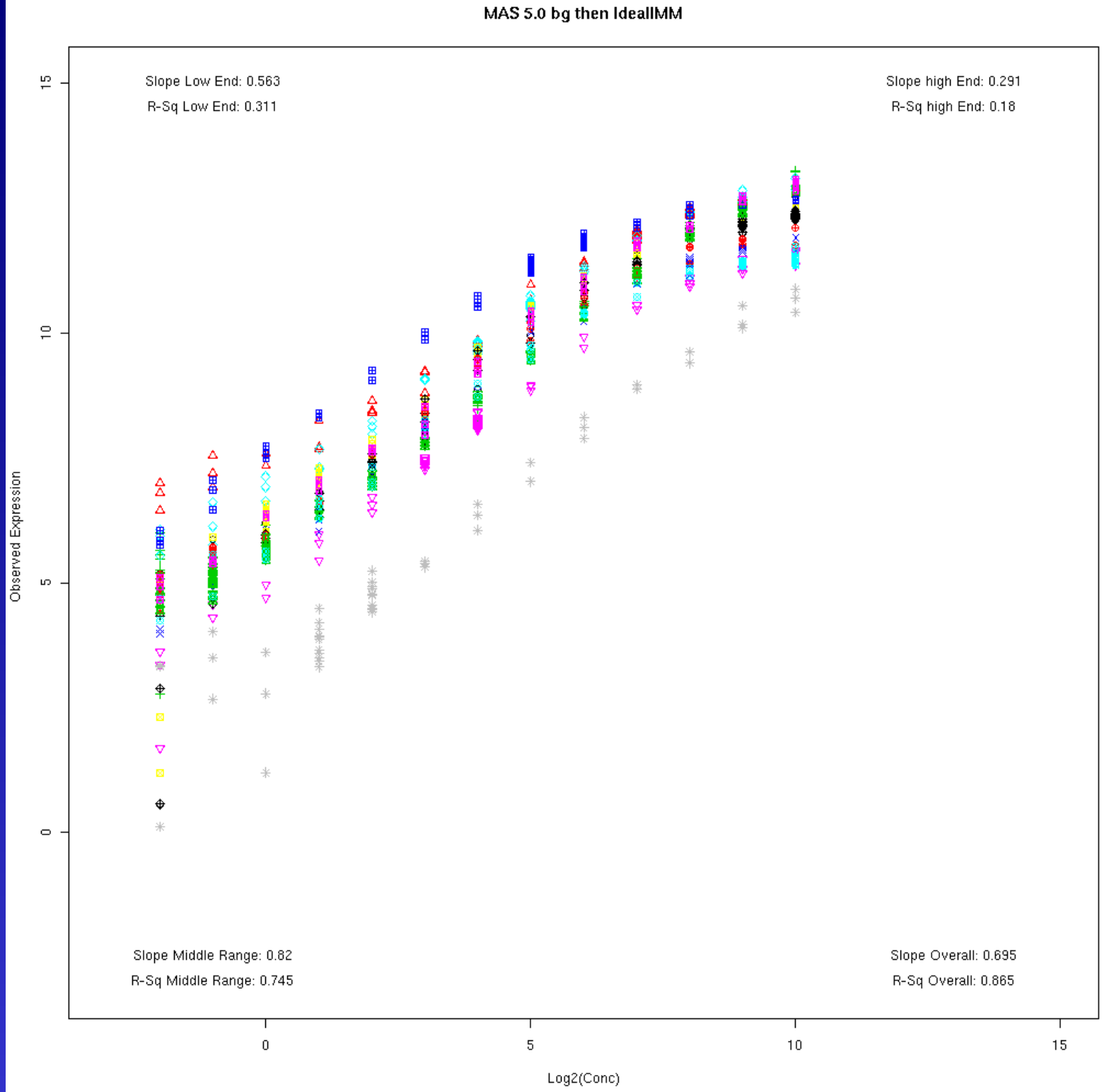
Slope	Value
All	0.589
Mid	0.751
Low	0.318
High	0.327



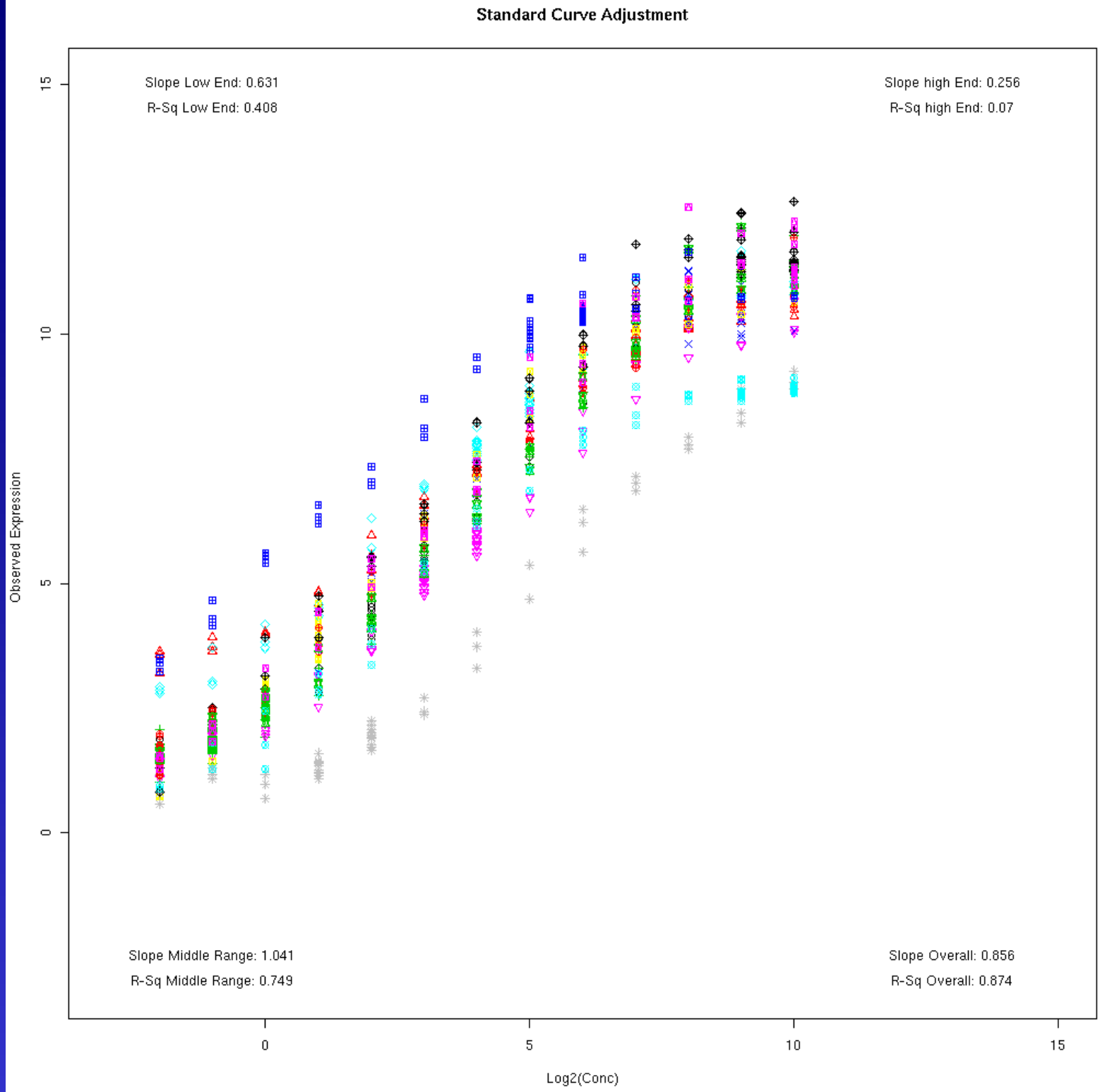
Slope	Value
All	0.69
Mid	0.82
Low	0.52
High	0.295



Slope	Value
All	0.695
Mid	0.82
Low	0.563
High	0.291

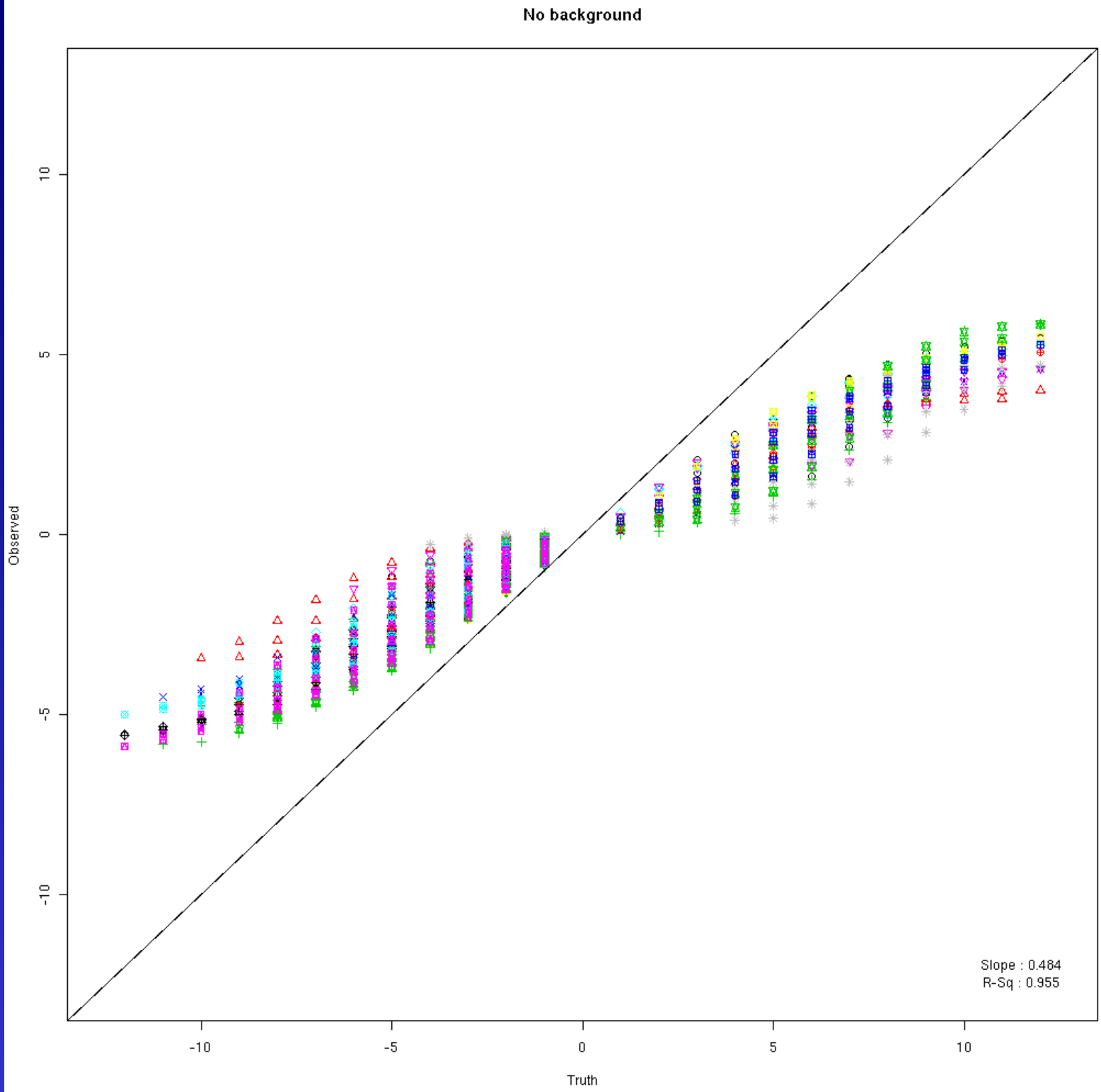


Slope	Value
All	0.856
Mid	1.041
Low	0.631
High	0.256

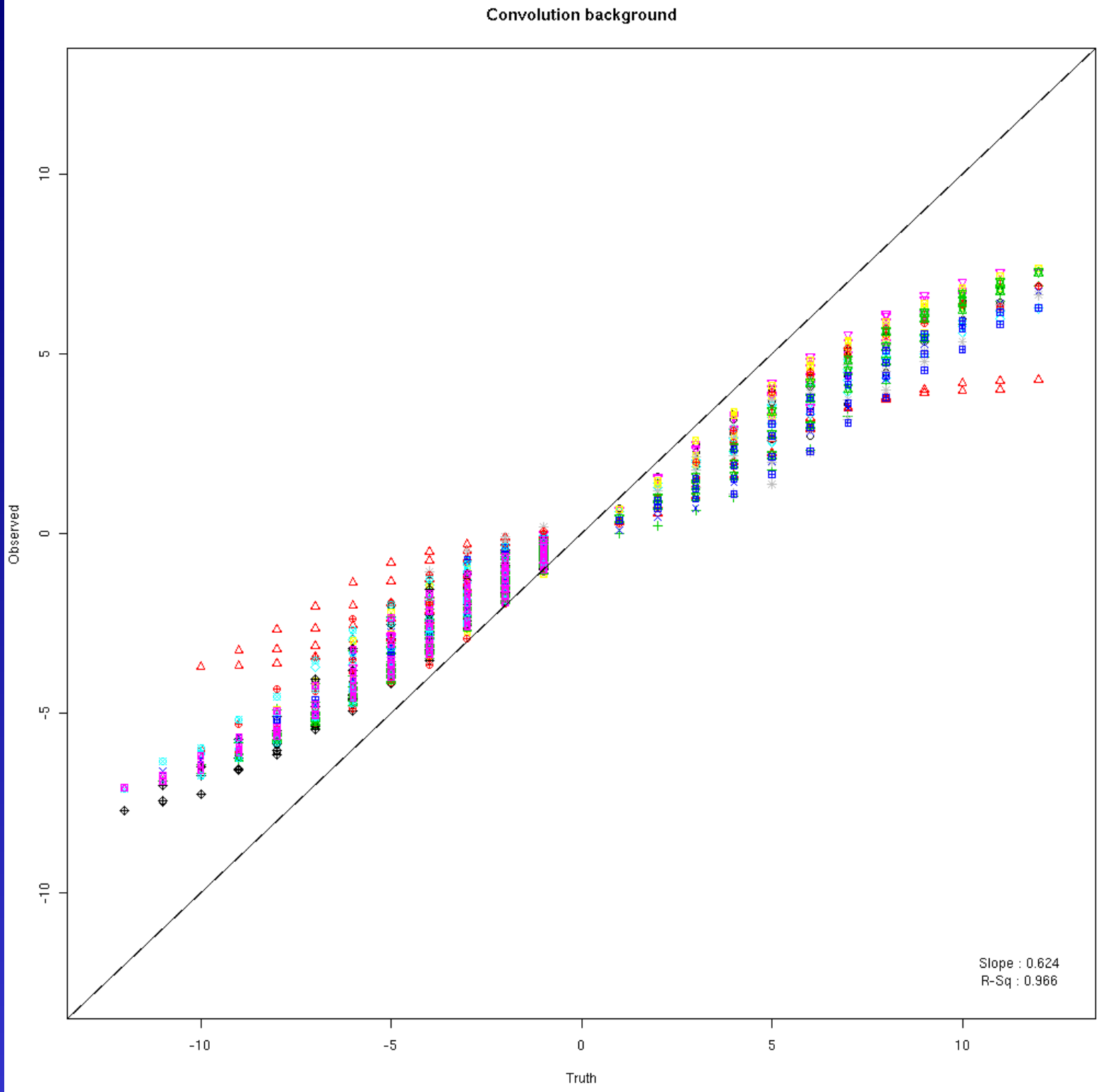


Observed fold change versus expected fold change

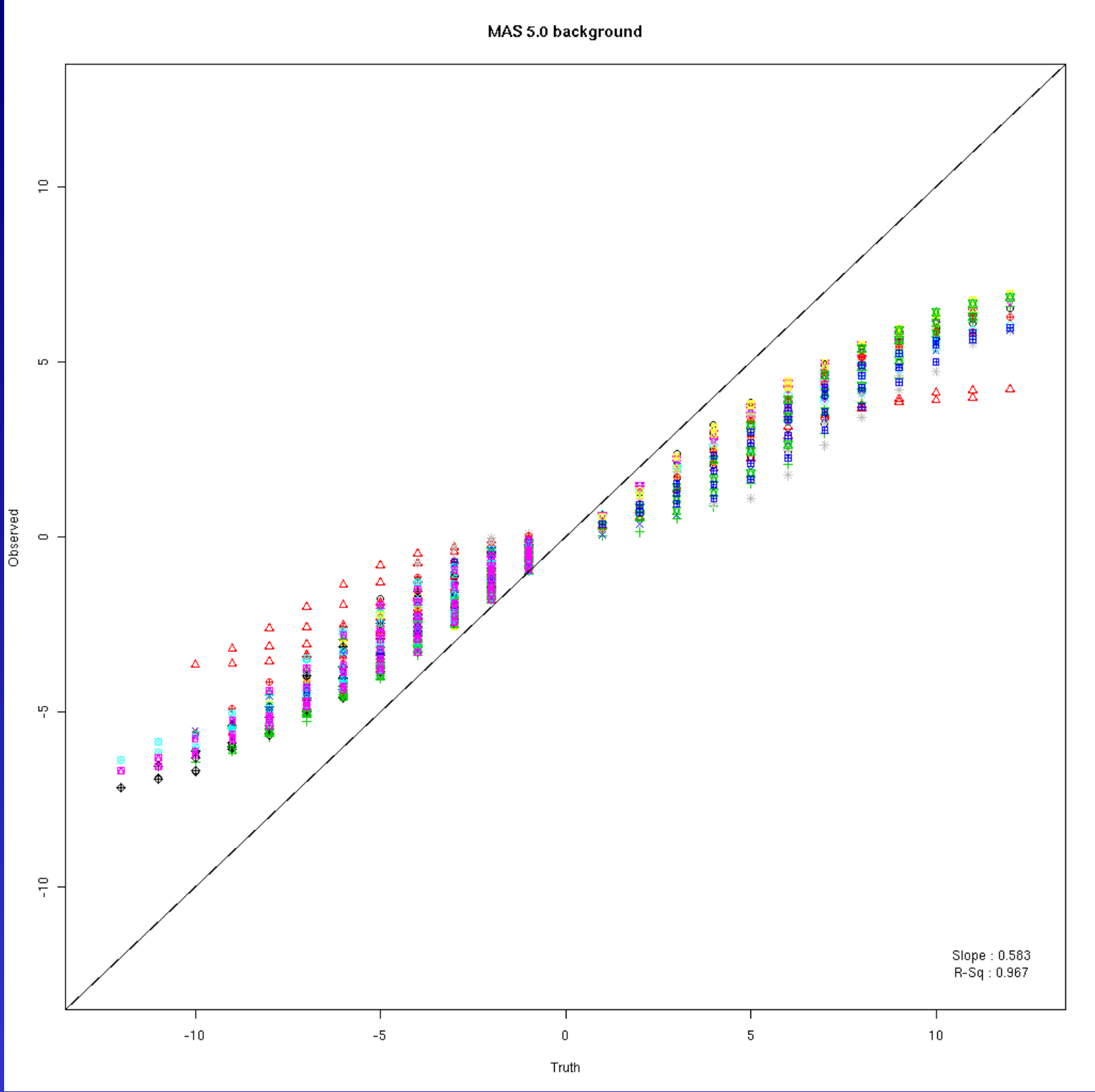
Slope: 0.484


No background
Observed
Truth
10
5
0
-5
-10
-10
-5
0
5
10
Slope : 0.484
R-Sq : 0.955

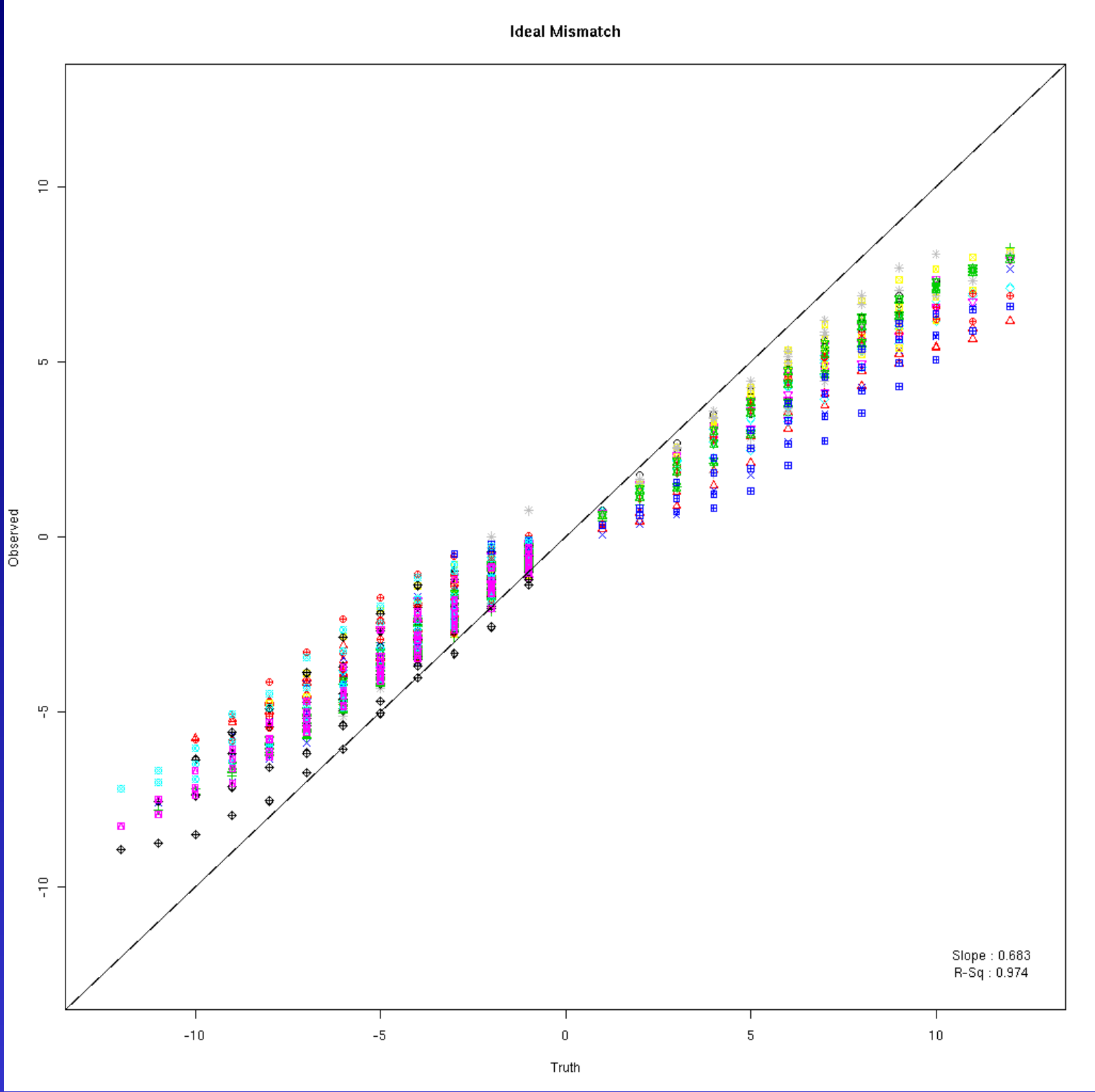
Slope: 0.624


Convolution background
Observed
Truth
10
5
0
-5
-10
-10
-5
0
5
10
Slope : 0.624
R-Sq : 0.966

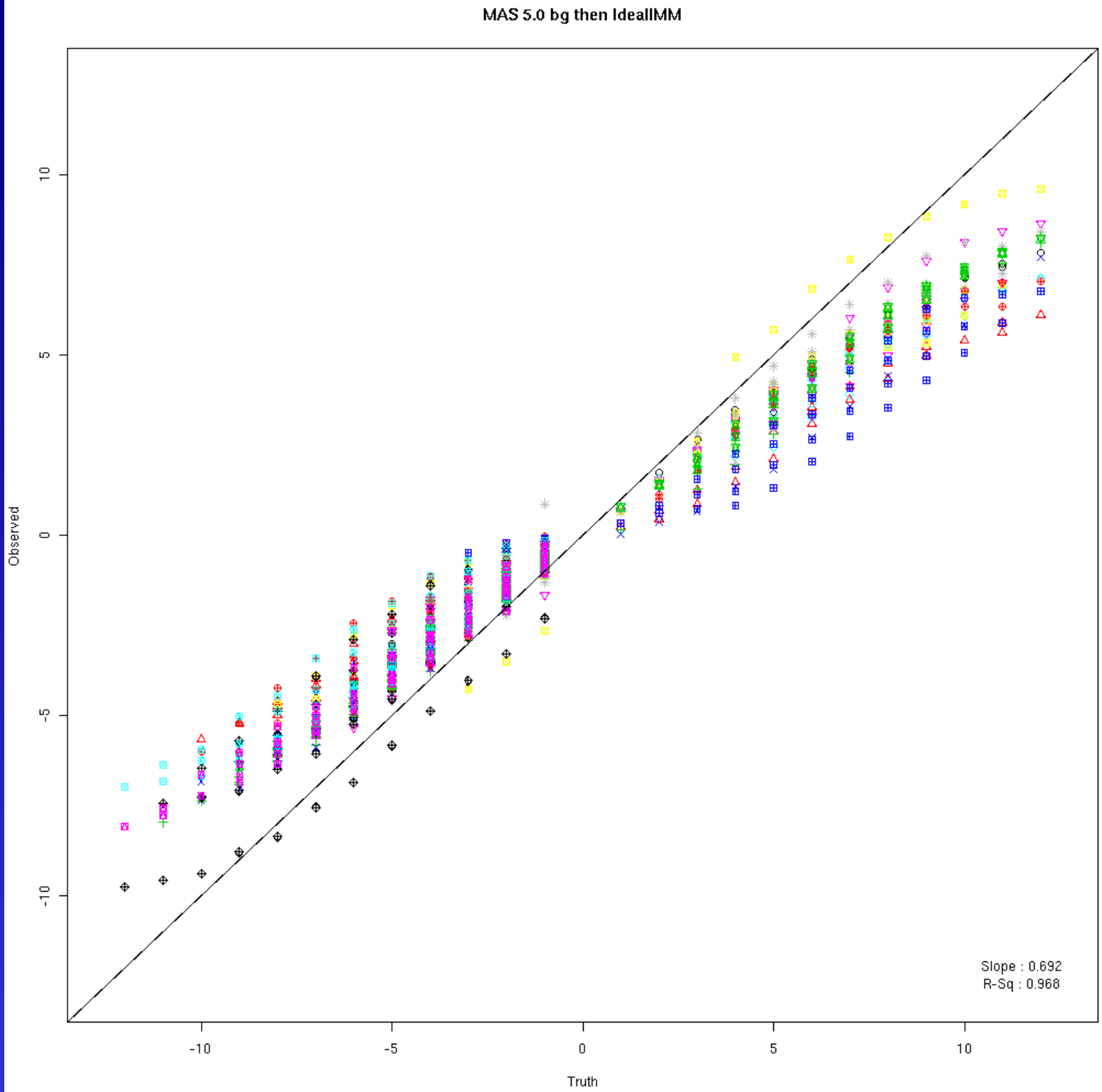
Slope: 0.583


MAS 5.0 background
Observed
Truth
10
5
0
-5
-10
-10
-5
0
5
10
Slope : 0.583
R-Sq : 0.967

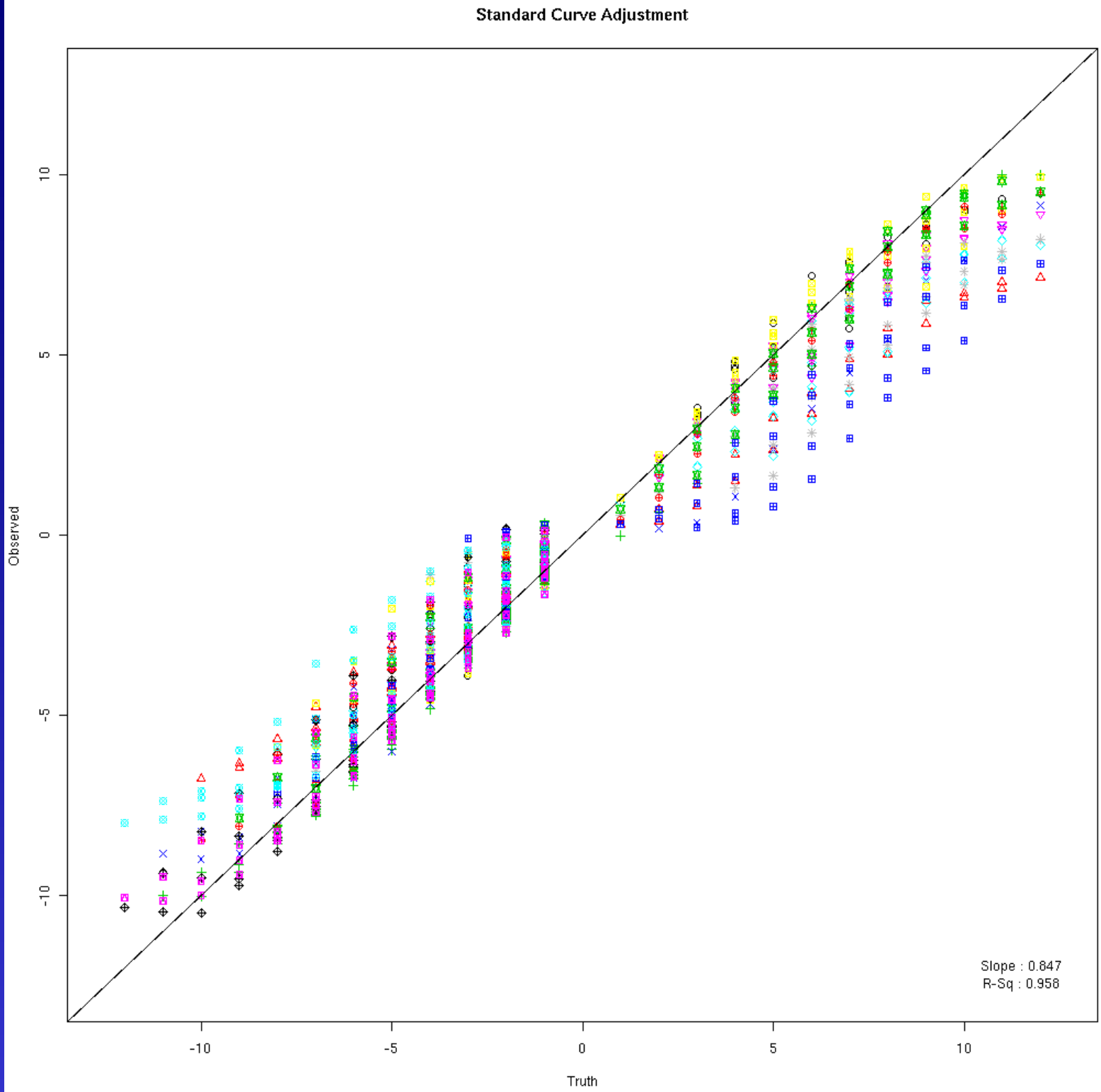
Slope: 0.683


Ideal Mismatch
Observed
Truth
10
5
0
-5
-10
-10
-5
0
5
10
Slope : 0.683
R-Sq : 0.974

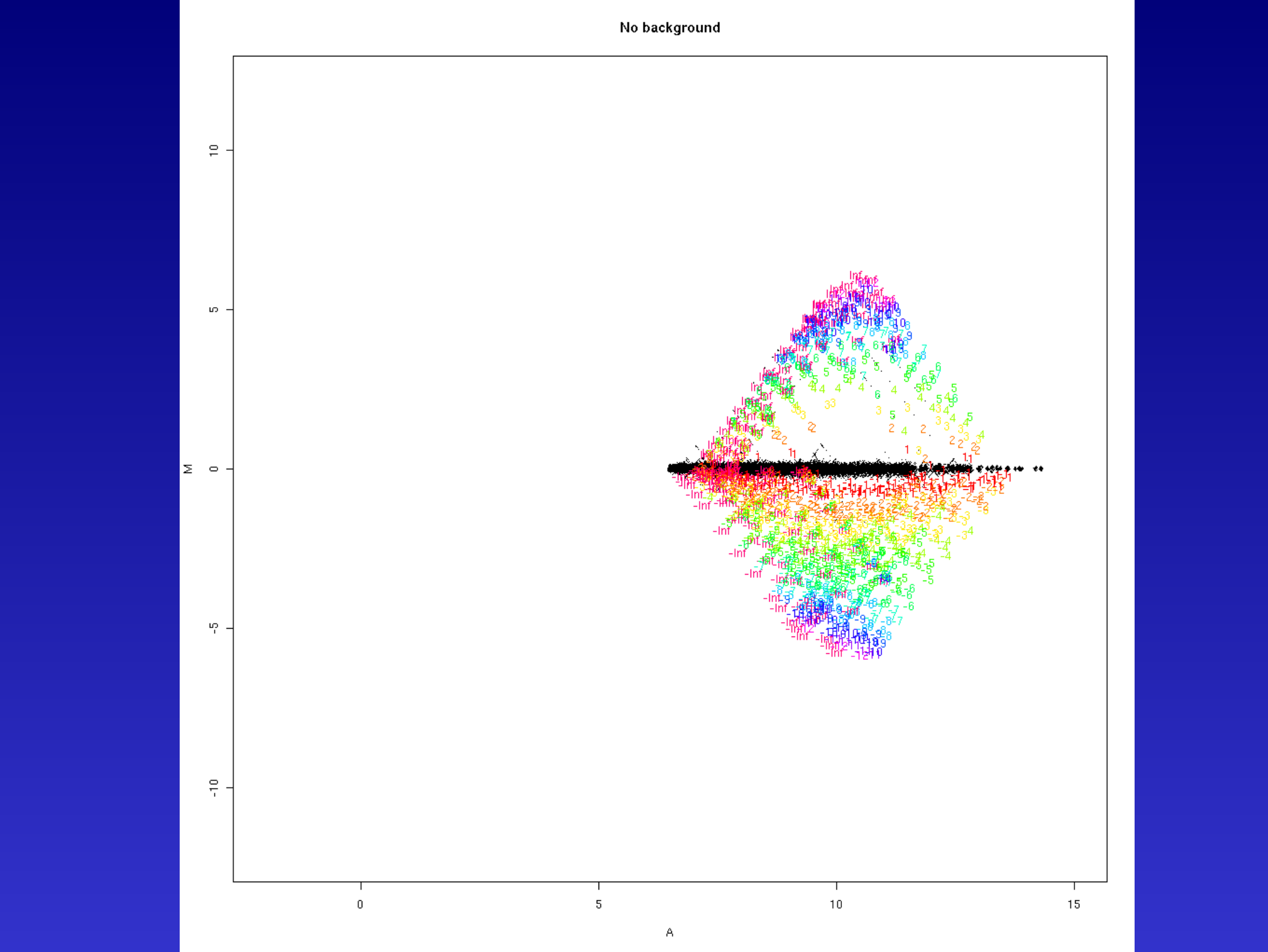
Slope: 0.692

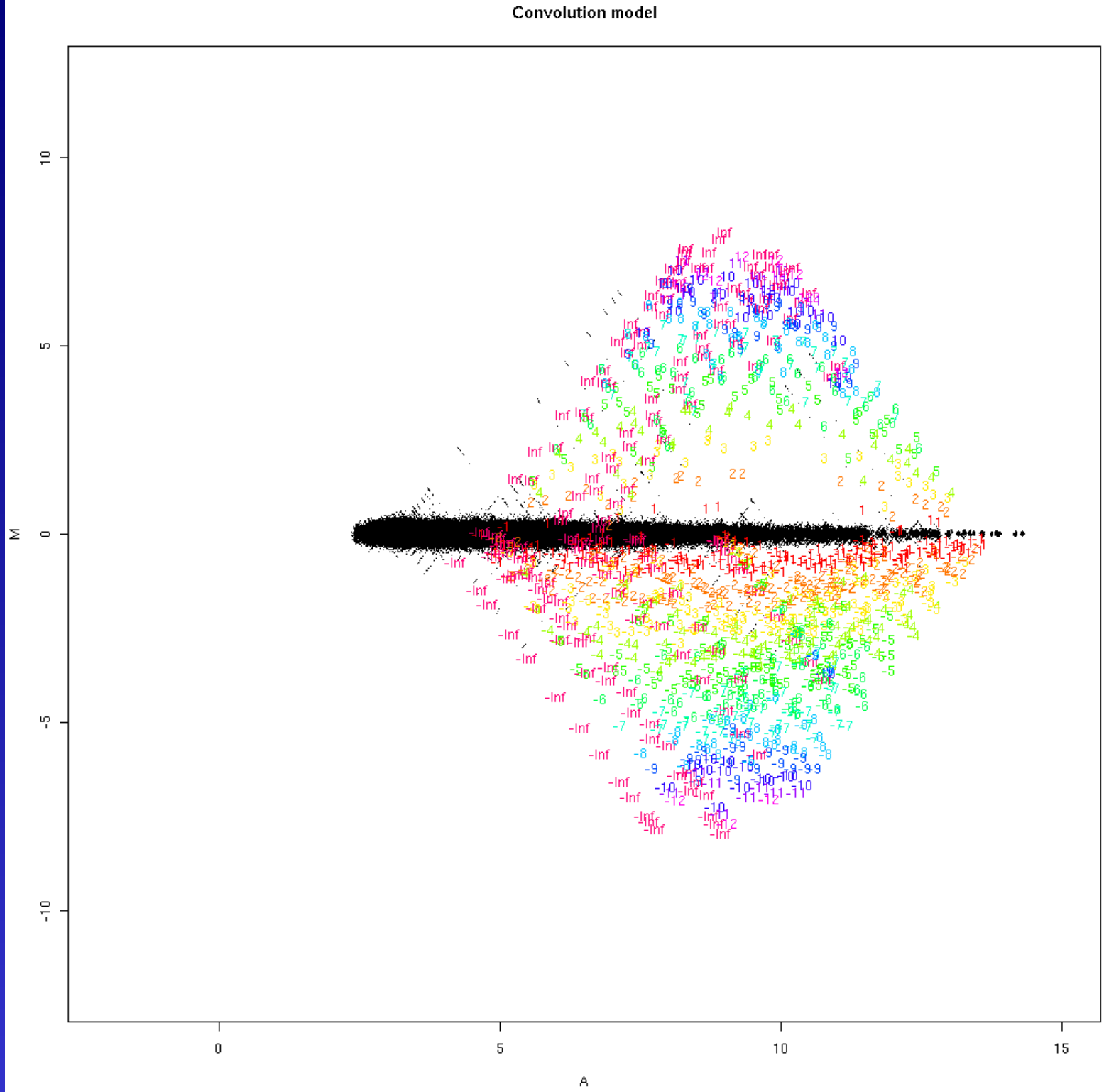

MAS 5.0 bg then IdeallMM
Observed
Truth
10
5
0
-5
-10
-10
-5
0
5
10
Slope : 0.692
R-Sq : 0.968

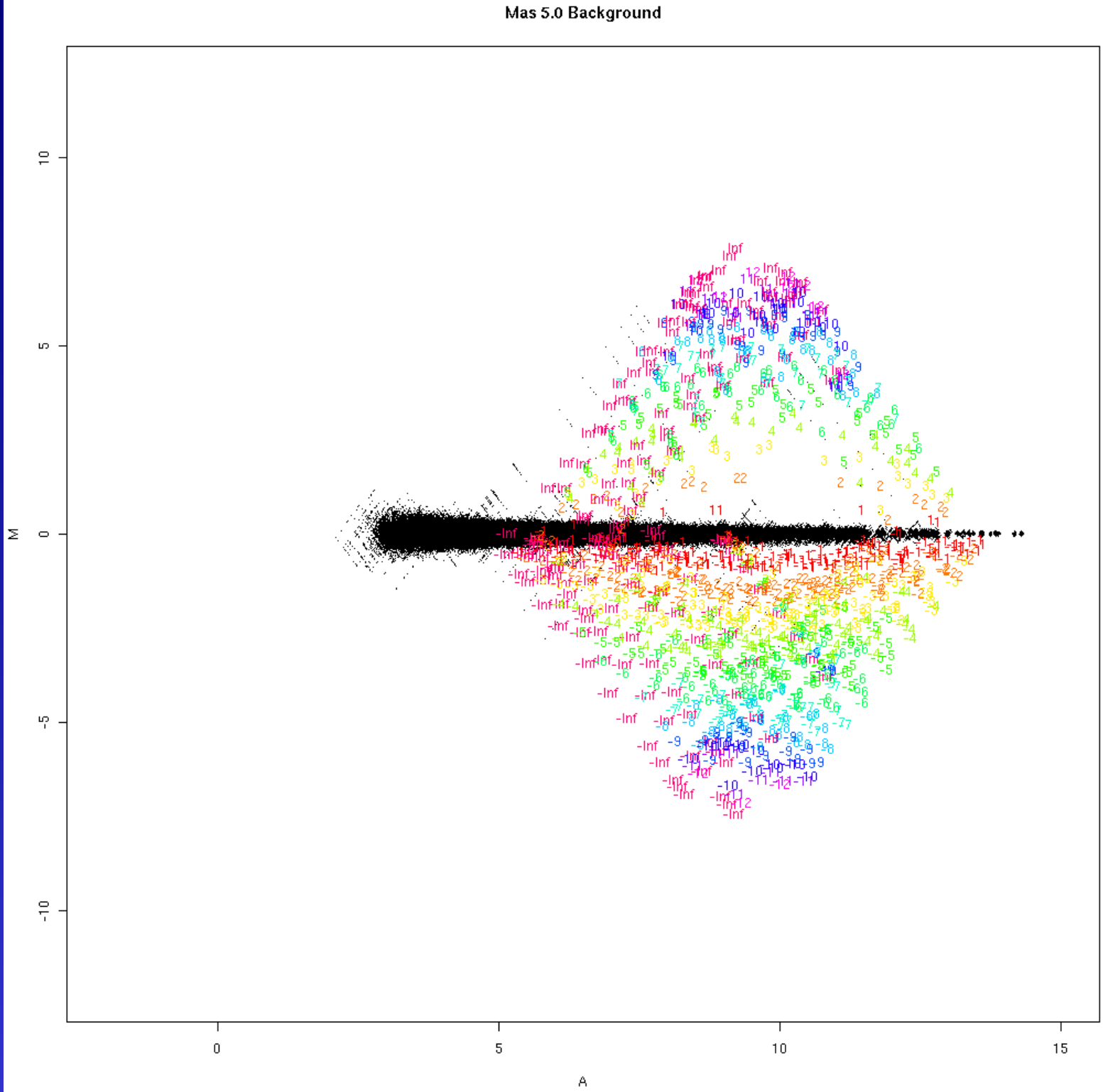
Slope: 0.847

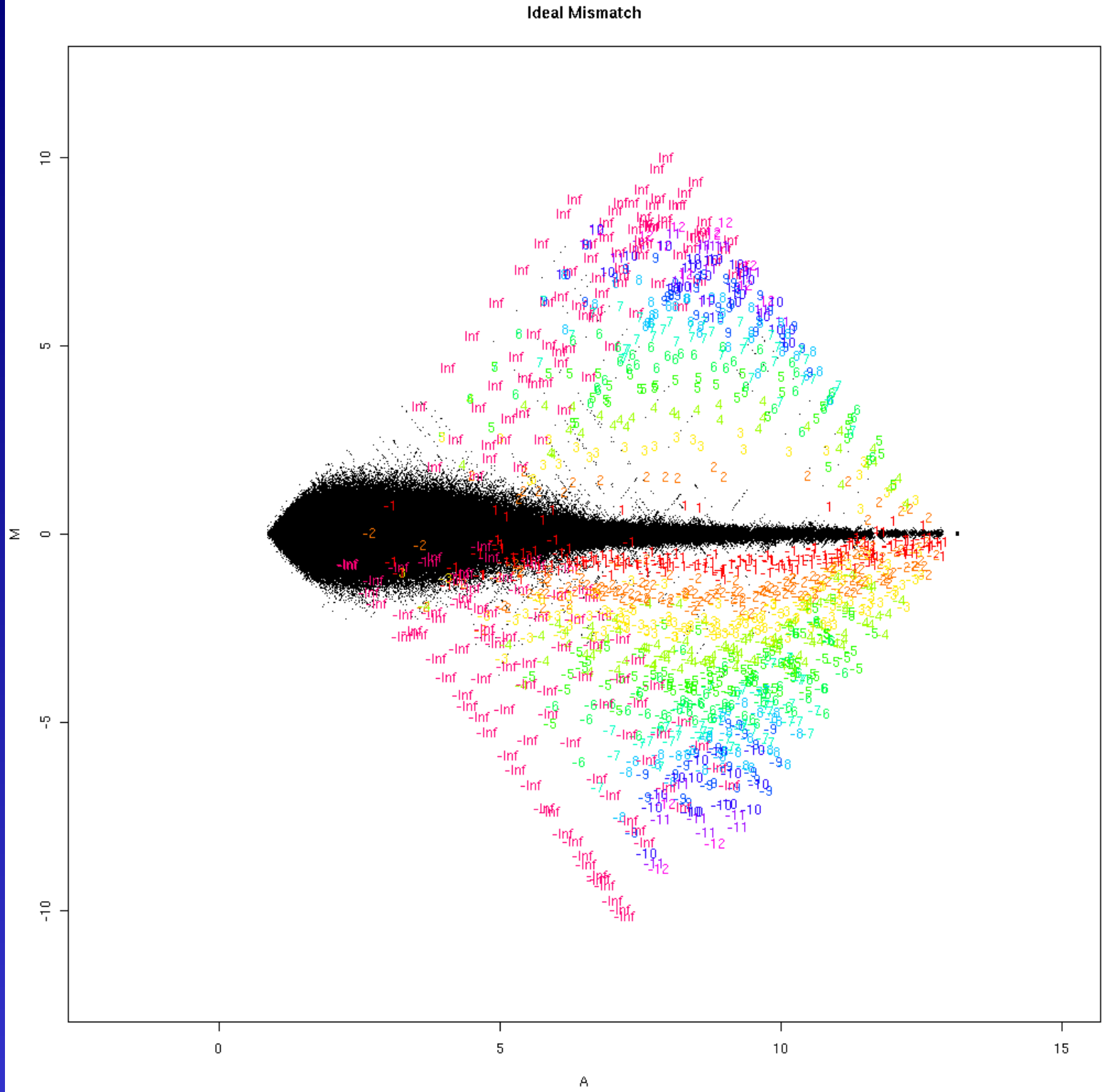

Standard Curve Adjustment
Observed
Truth
10
5
0
-5
-10
-10
-5
0
5
10
Slope : 0.847
R-Sq : 0.958

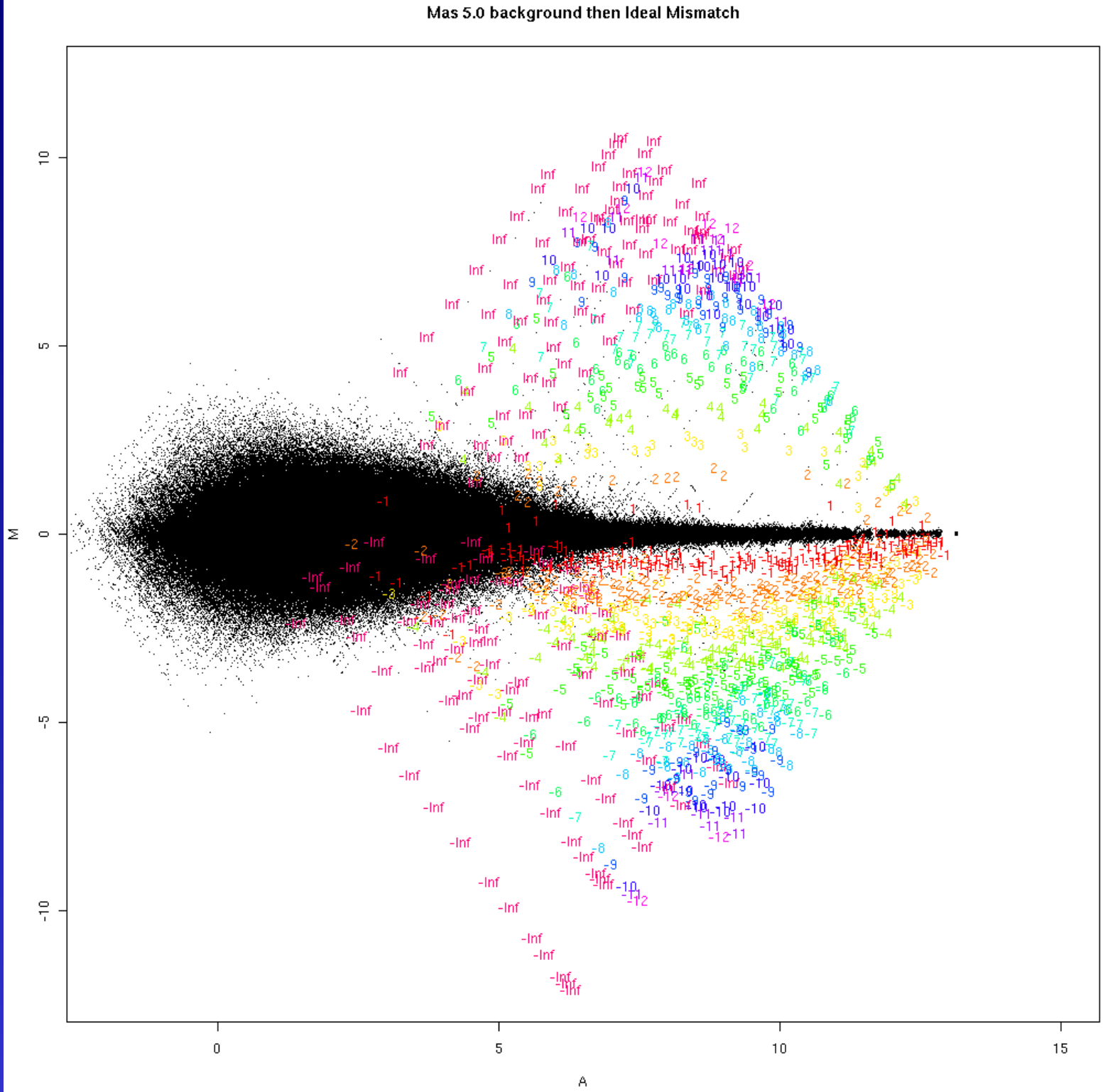
Composite M vs A Plots

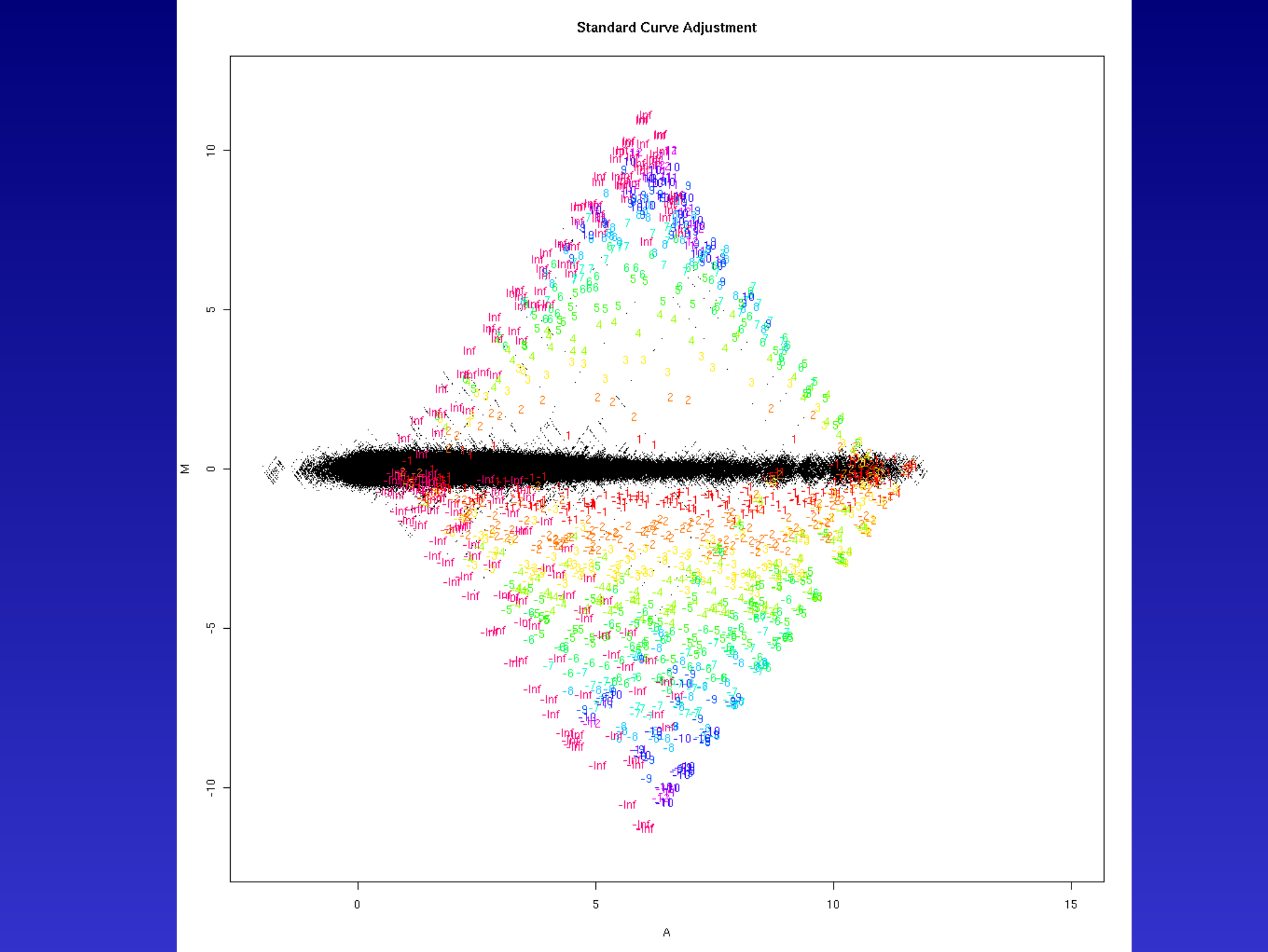

No background
M
A
10
5
0
-5
-10
0
5
10
15


Convolution model
M
A
10
5
0
-5
-10
0
5
10
15

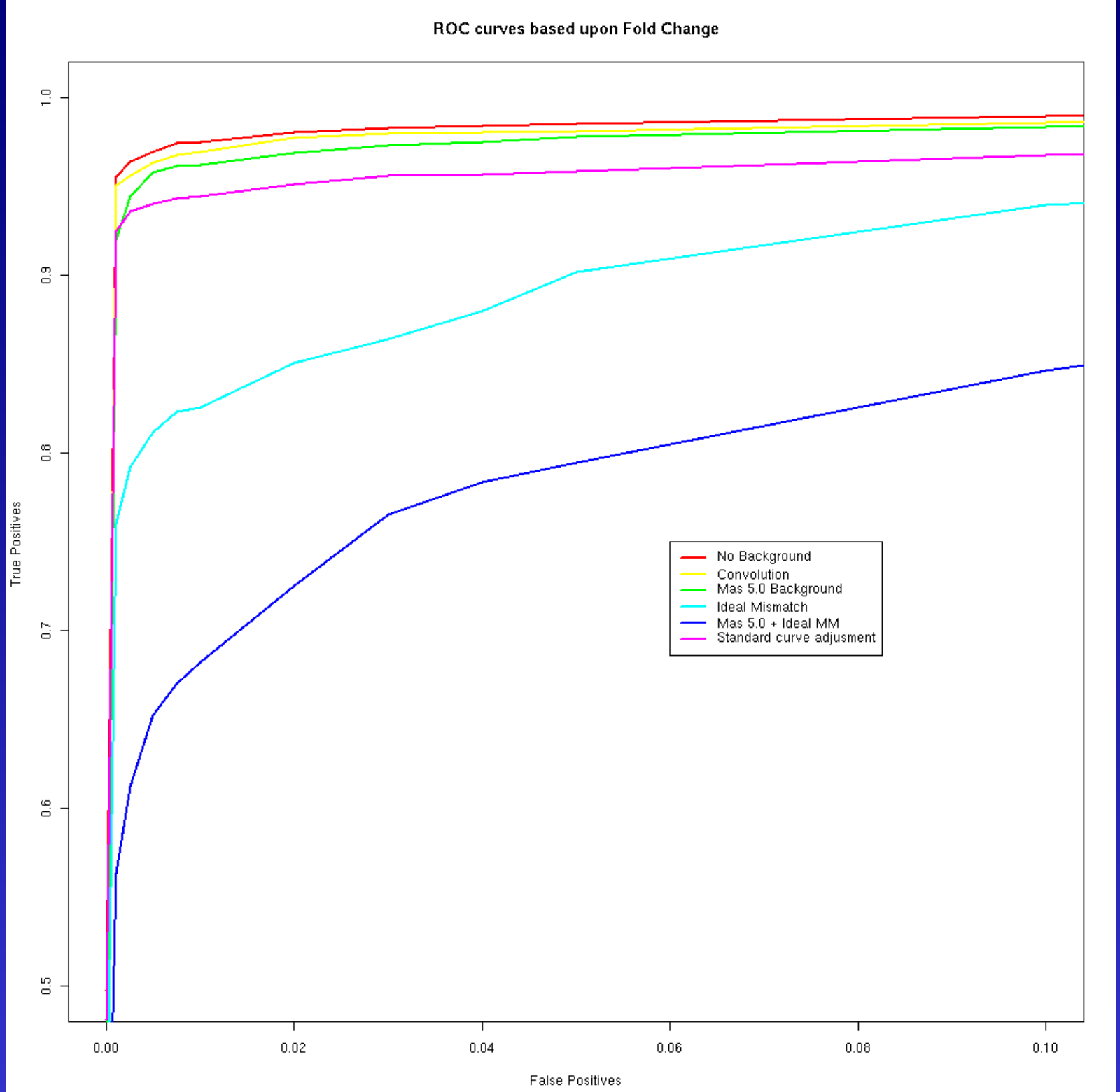

Mas 5.0 Background
M
A
10
5
0
-5
-10
0
5
10
15


Ideal Mismatch
M
A
10
5
0
-5
-10
0
5
10
15


Mas 5.0 background then Ideal Mismatch
M
A
10
5
0
-5
-10
0
5
10
15


Standard Curve Adjustment
M
A
10
5
0
-5
-10
0
5
10
15

ROC Curves


ROC curves based upon Fold Change
True Positives
False Positives
1.0
0.9
0.8
0.7
0.6
0.5
0.00
0.02
0.04
0.06
0.08
0.10
No Background
Convolution
Mas 5.0 Background
Ideal Mismatch
Mas 5.0 + Ideal MM
Standard curve adjusment

The Background Methods Have Different Tradeoffs

		Detect Differential Genes	
		Poor	Good
Accurate estimate of Fold Change	Poor		•No Background •Convolution •MAS 5.0
	Good	•MAS 5 + IdealMM •Ideal-Mismatch	•Standard Curve Adjustment

Results 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

Comparing RMA with MAS 5.0, dChip MBEI and others

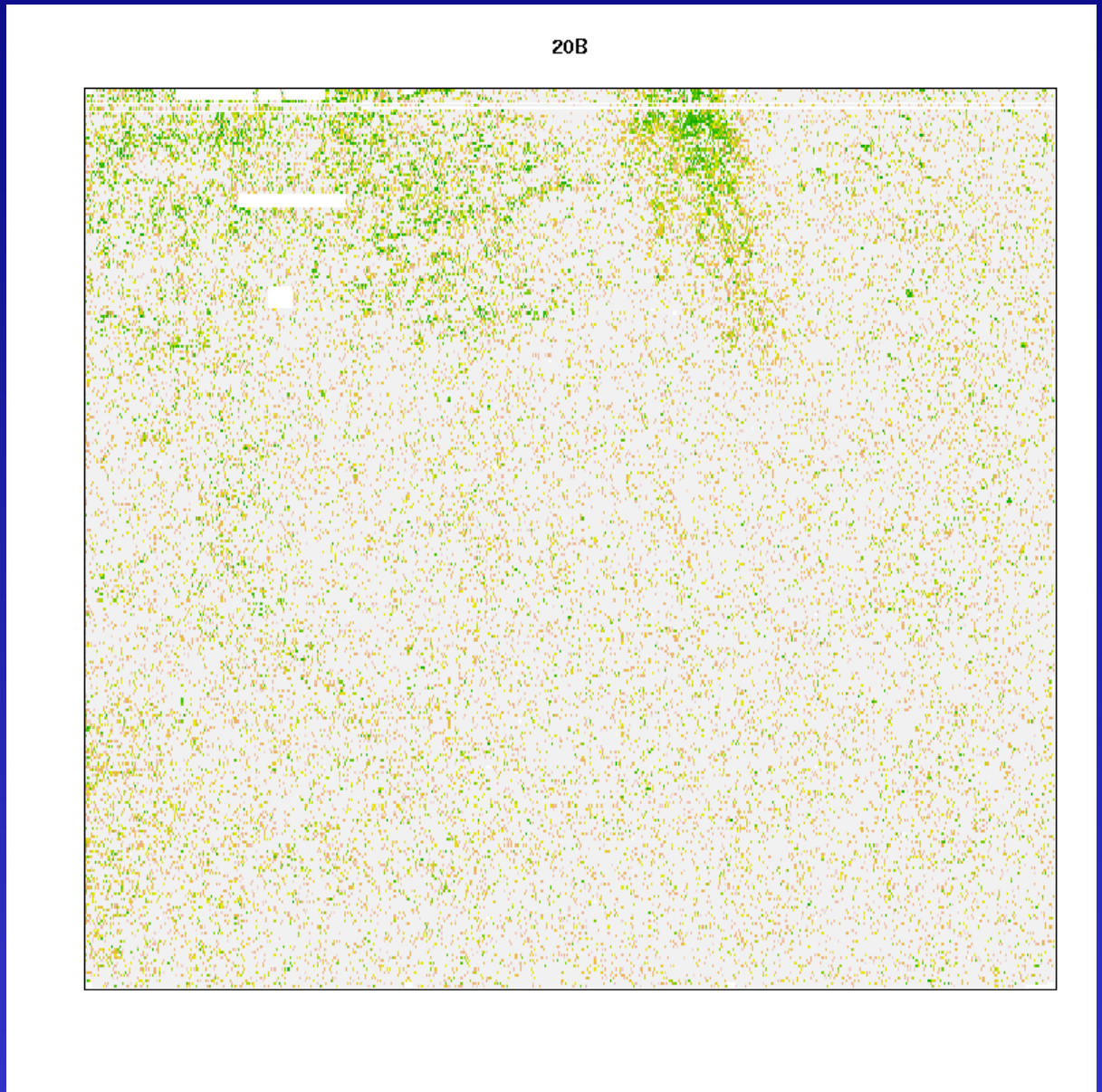
This article compares RMA with MAS 5.0 and dChip MBEI:

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**

A competition and comparison framework

http://affycomp.biostat.jhsph.edu/

Fitting using a robust linear model gives quality diagnostics



Software

- R packages
  - *affy* which is part of Bioconductor

http://www.bioconductor.org

`rma(), normalize.quantiles(), bg.correct.rma(),…`

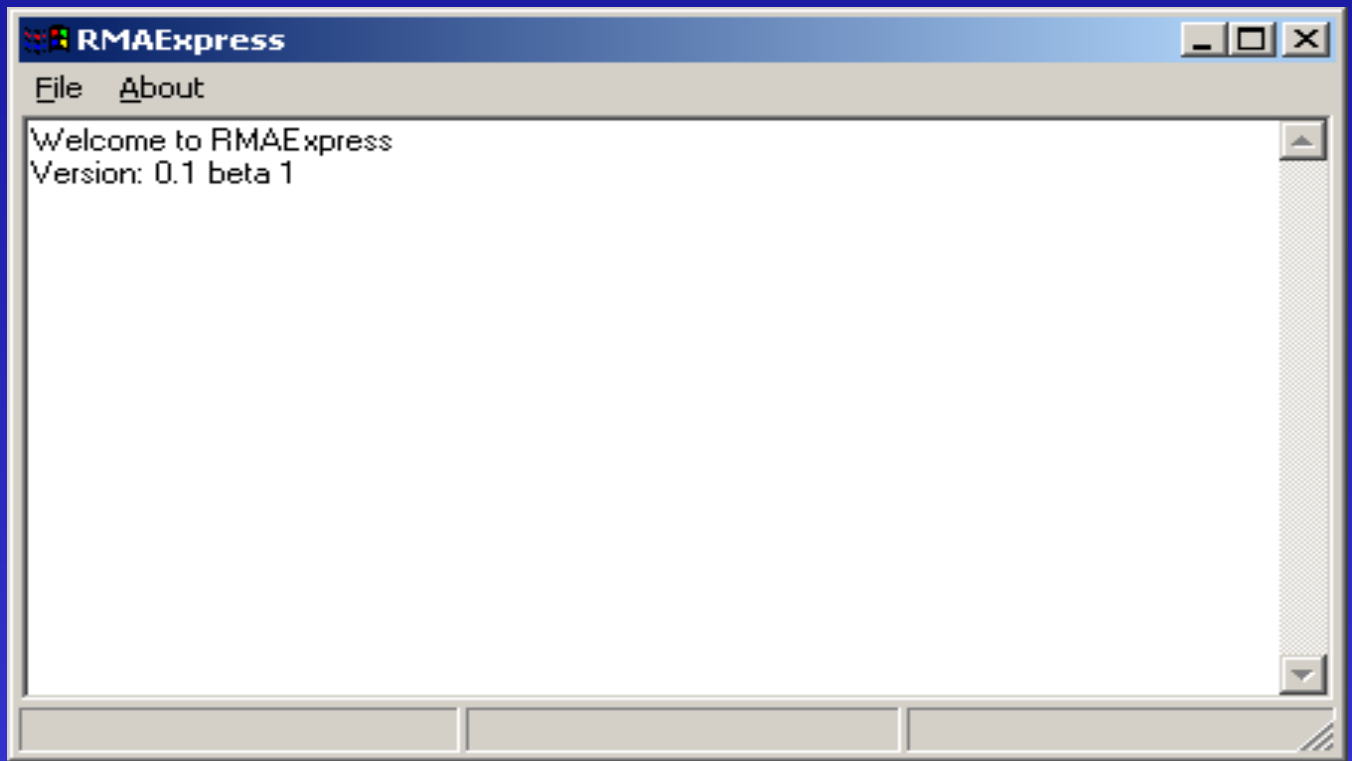
  - *AffyExtensions* A package for fitting more general probe level models

http://www.stat.berkeley.edu/~bolstad/AffyExtensions/AffyExtensions.html

`fitPLM(), threestep(),…`

Software

- *RMAExpress:* a simple standalone GUI program for Windows for computing the RMA expression measure
  http://www.stat.berkeley.edu/~bolstad/RMAExpress/RMAExpress.html



Some references

1. Bolstad BM, Irizarry RA, Astrand M and Speed TP . (2003), A comparison of normalization methods for high density oligonucleotide array data basedon 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
5. Bioconductor http://www.bioconductor.org
6. Affymetrix Spike-in experiment http://www.affymetrix.com/analysis/download_center2.affx
7. Affymetrix Website http://www.affymetrix.com

Acknowledgements

- Terry Speed (Statistics, UCB)
- Rafael Irizarry (Biostatistics, John Hopkins)
- Francois Collin (Genelogic)