Probe-Level Data Analysis of Affymetrix GeneChip Expression Data using Open-source Software

Ben Bolstad

bmb@bmbolstad.com

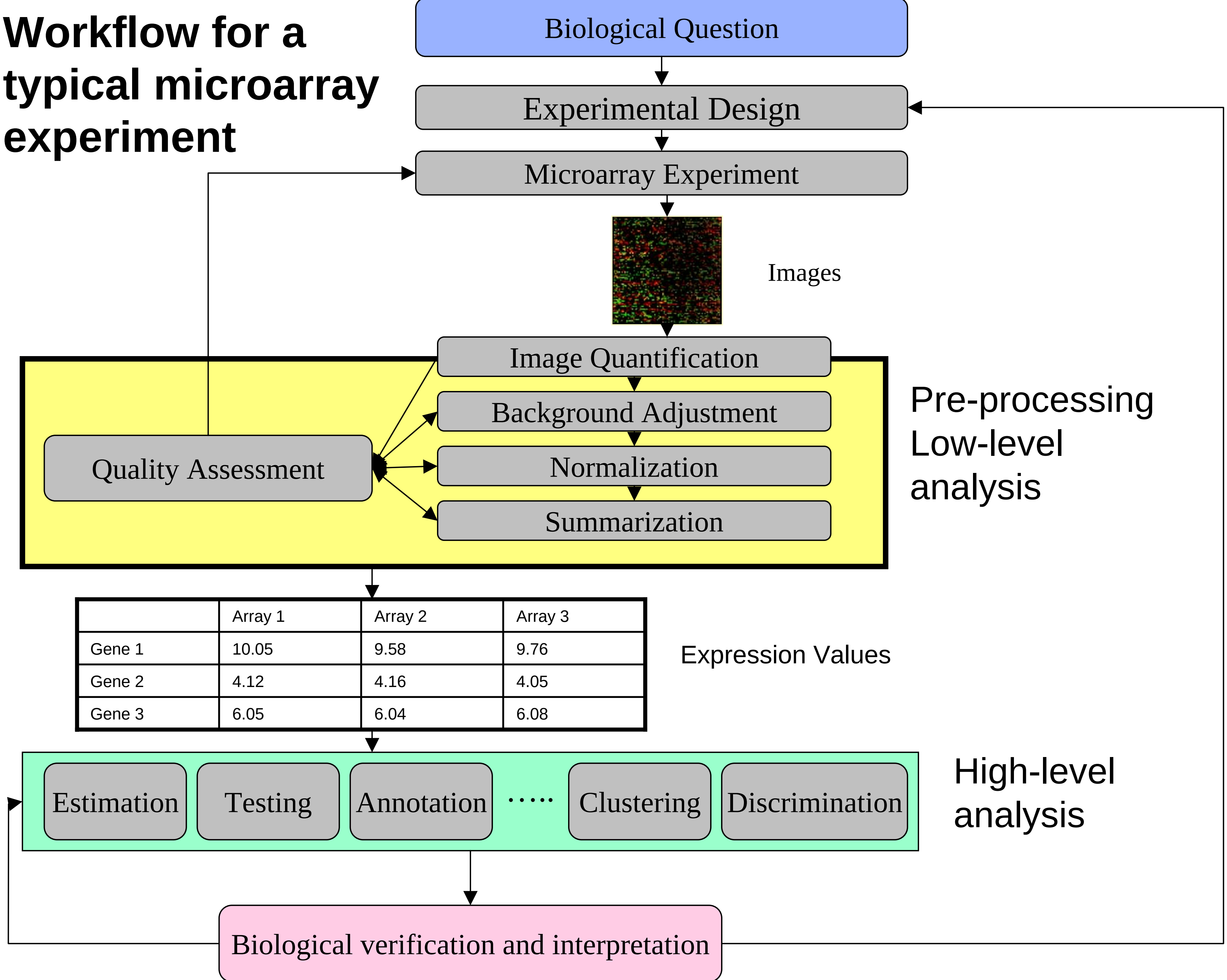
http://bmbolstad.com

August 7, 2006

Outline

- Introduction to probe-level data analysis
- Probe-level analysis using the RMA framework and extensions
- Example analysis using BioConductor tools

Workflow for a typical microarray experiment



Biological Question

Experimental Design

Microarray Experiment

Images

Image Quantification

Background Adjustment

Normalization

Summarization

Quality Assessment

Pre-processing
Low-level
analysis

	Array 1	Array 2	Array 3
Gene 1	10.05	9.58	9.76
Gene 2	4.12	4.16	4.05
Gene 3	6.05	6.04	6.08

Expression Values

Estimation

Testing

Annotation

…..

Clustering

Discrimination

High-level
analysis

Biological verification and interpretation

Introduction to Probe-Level Analysis

- Also known as “Pre-processing” or “low-level analysis”
- Pre-processing typically constitutes the initial (and possibly most important) step in the analysis of data from any microarray experiment
- Often ignored or treated like a black box (but it shouldn’t be)
- Consists of:
  - Data exploration
  - Background correction, normalization, summarization
  - Quality Assessment
- These are interlinked steps
- Probe intensities rather than expression values are the data used.

Background Correction/Signal Adjustment

- A method which does some or all of the following:
  - Corrects for background noise, processing effects on the array
  - Adjusts for cross hybridization (non-specific binding)
  - Adjust estimated expression values to fall across an appropriate range

Normalization

- Normalization is the process of reducing unwanted variation (variation due to technical effects) either within or between arrays. It may use information from multiple chips.
- Typical assumptions of most major normalization methods are (one or both of the following):
  - Only a minority of genes are expected to be differentially expressed between conditions
  - Any differential expression is as likely to be up-regulation as down-regulation (ie about as many genes going up in expression as are going down between conditions)

Summarization

- Reducing multiple measurements on the same gene down to a single measurement by combining in some manner. ie take each of the multiple probe intensities for a probeset and derive a single number representing probeset expression value.

Quality Assessment

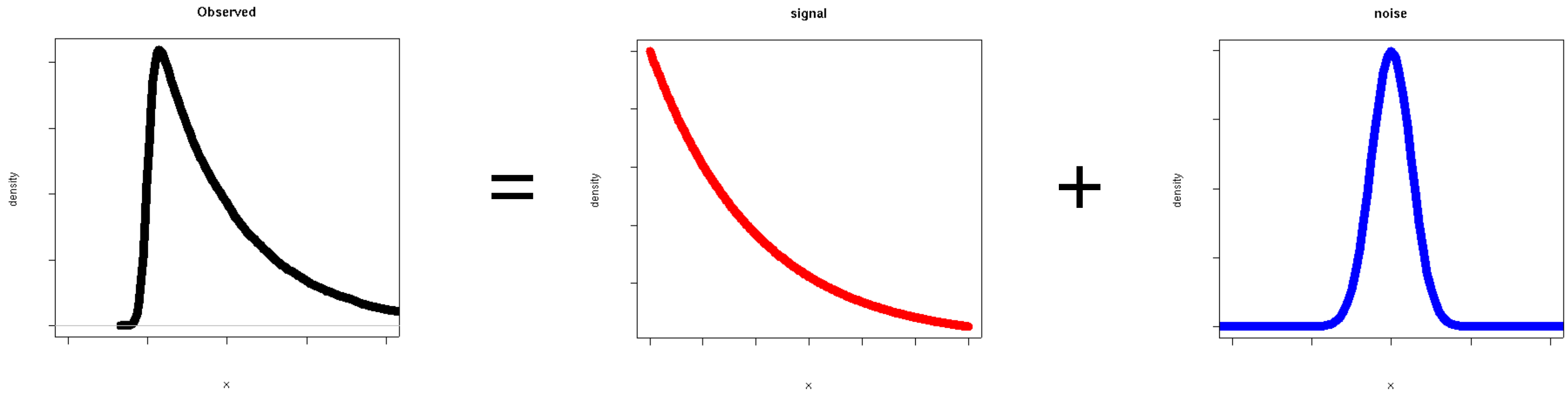
- Need to be able to differentiate between good and bad data.
- Bad data could be caused by poor hybridization, artifacts on the arrays, inconsistent sample handling, …..
- An admirable goal would be to reduce systematic differences with data analysis techniques.
- Sometimes there is no option but to completely discard an array from further analysis. How to decide …..

Whats RMA?

- **R**obust **M**ulti-array **A**nalysis
  - Background correction using a convolution model (GCRMA modifies this stage)
  - Quantile Normalization across arrays
  - Multi-array probe-level model fit to each probeset
  - Quality assessment

RMA Background Approach

- Convolution Model



Observed PM

Signal S $exp(\alpha)$

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

$$\mathrm{E}\left(S \middle| PM = pm\right) = a + b\frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{pm-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{pm-a}{b}\right) - 1}$$

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

GCRMA Background Approach

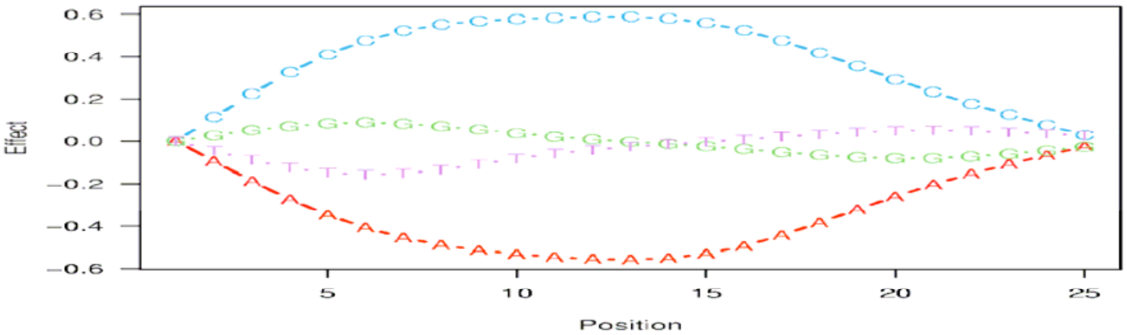
- $PM=O_{pm}+N_{pm}+S$
- $MM=O_{mm}+N_{mm}$

- O – Optical noise
- N – non-specific binding
- S – Signal

- Assume O is distributed Normal
- $\log(N_{pm})$and $\log(N_{mm})$ are assumed bi-variate normal with correlation 0.7
- log(S) assumed exponential(1)

GCRMA Background cont

- An experiment was carried out where yeast RNA was hybridized to human chips, so all binding expected to be non specific.
- Fitted a model to predict log intensity from sequence composition gives base and position effects



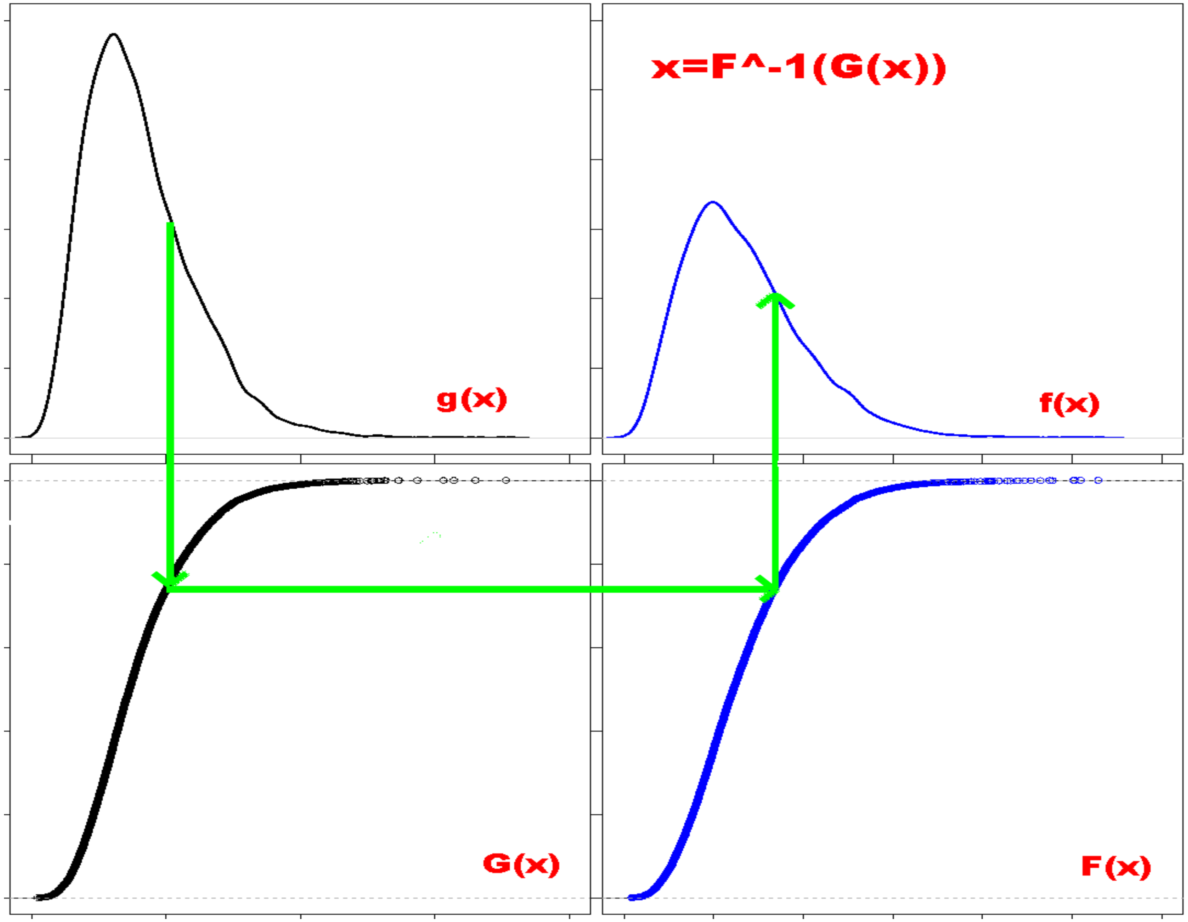
- Uses these effects to predict an affinity for any gıven sequence call this A. The means of the distributions for the $N_{pm,}$ N_{mm} terms are functions of the affinities.

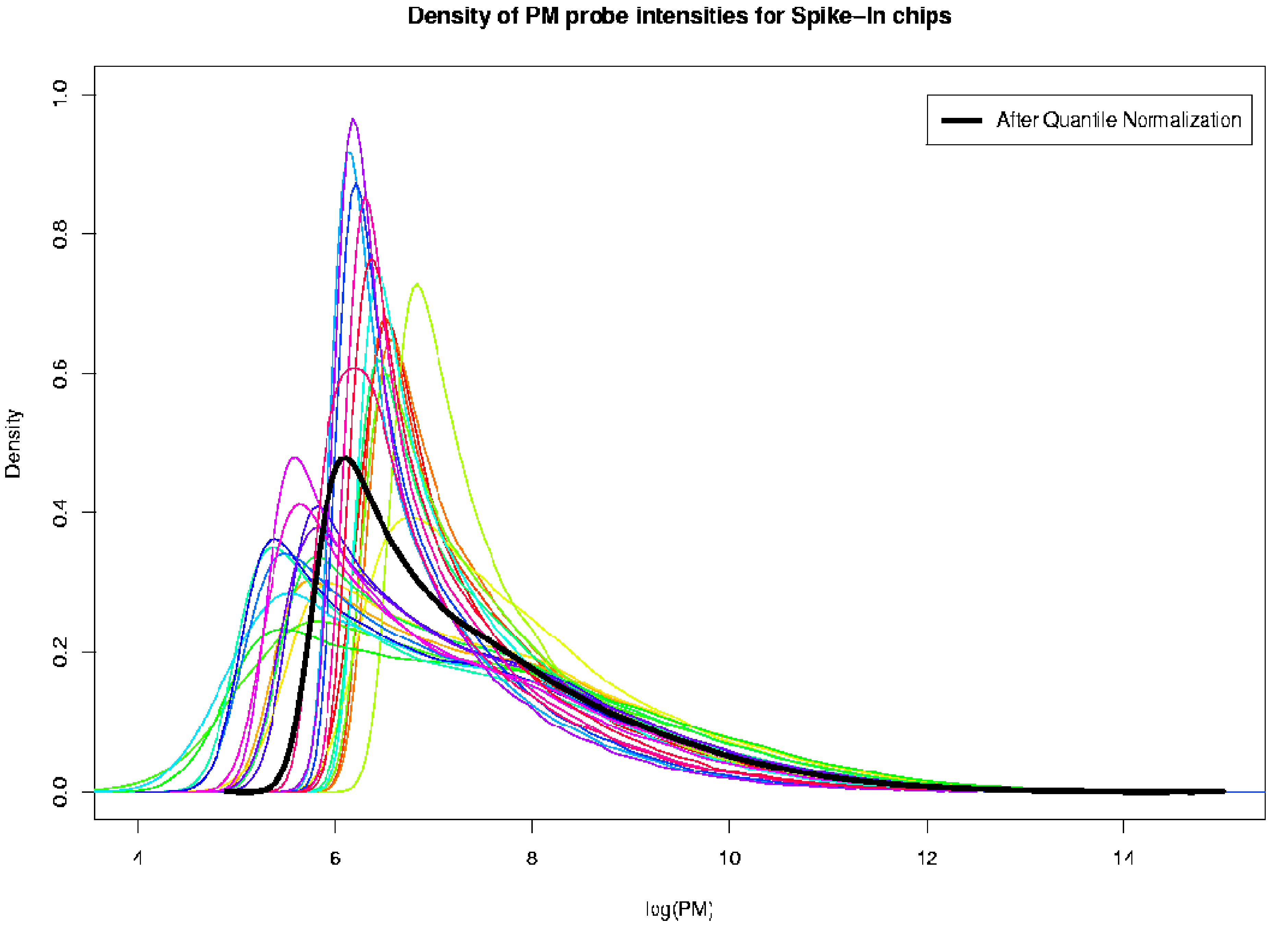
Normalization

- In case of single channel microarray data this is carried out only across arrays.
- Could generalize methods we applied to two color arrays, but several problems:
  - Typically several orders of magnitude more probes on an Affymetrix array then spots on a two channel array
  - With single channel arrays we are dealing with absolute intensities rather than relative intensities.
- Need something fast

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.



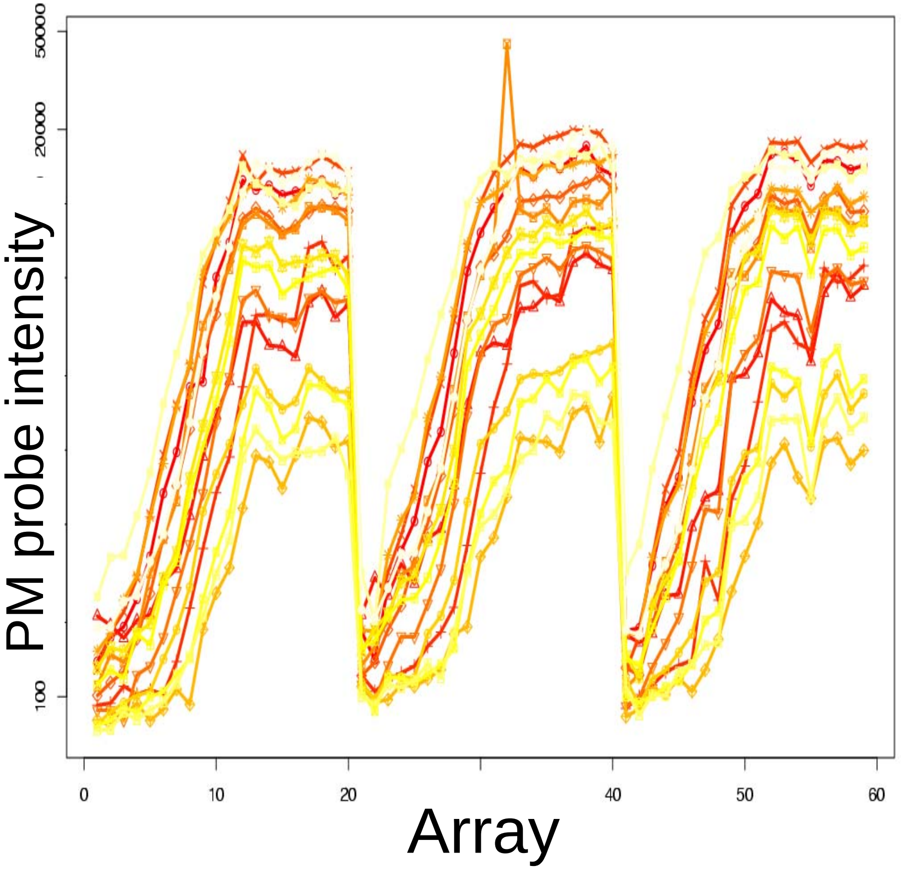

Density of PM probe intensities for Spike−In chips
After Quantile Normalization
Density
0.0
0.2
0.4
0.6
0.8
1.0
4
6
8
10
12
14
log(PM)

Summarization

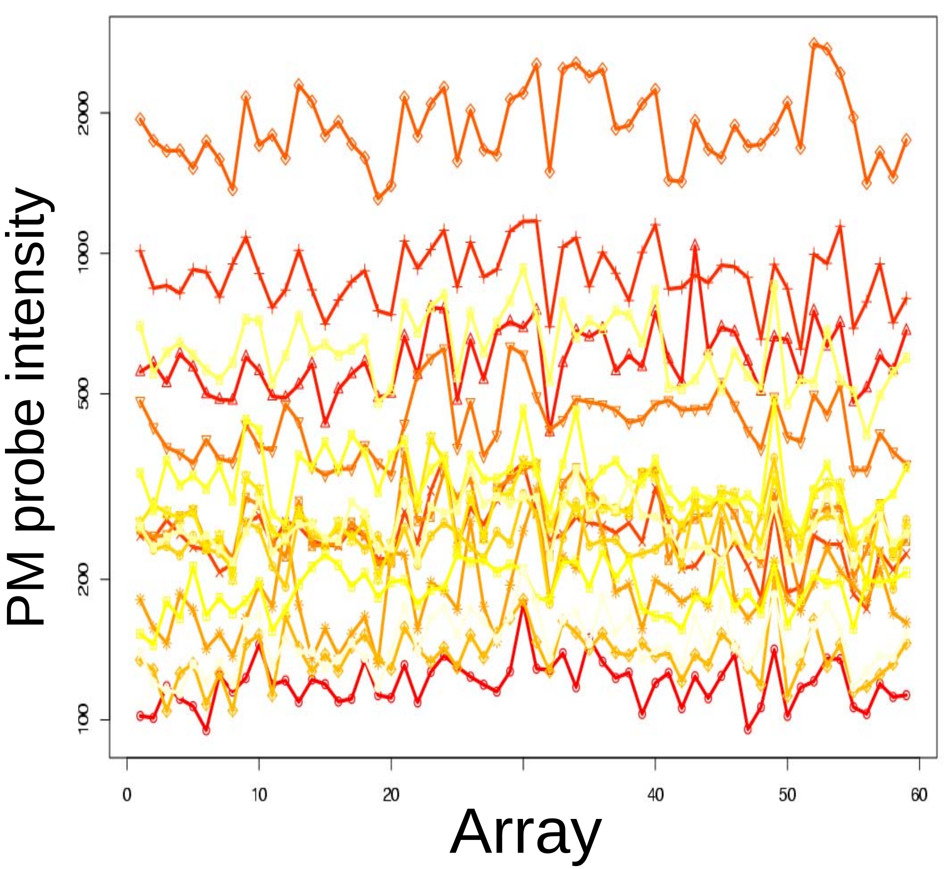
- Need to take the normalized background corrected probe intensities and reduce to sensible gene expression measures.
- RMA uses a multi-array model fit to logarithmic scale data.

Parallel Behavior Suggests Multi-chip Model

Differentially expressing

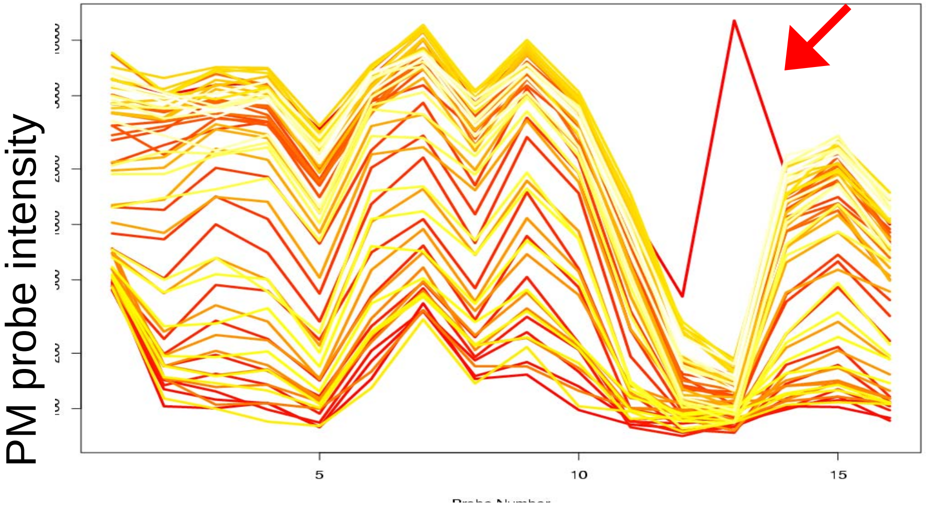


Non Differential

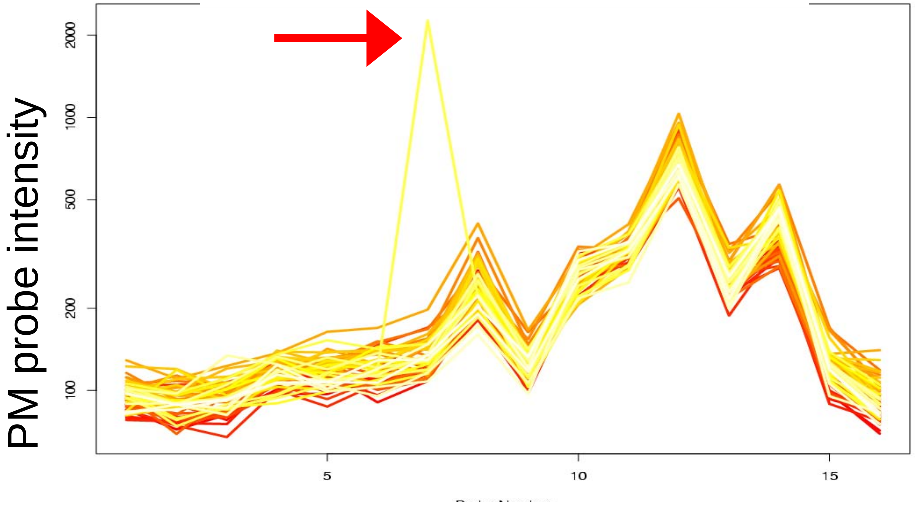


Also Want Robustness

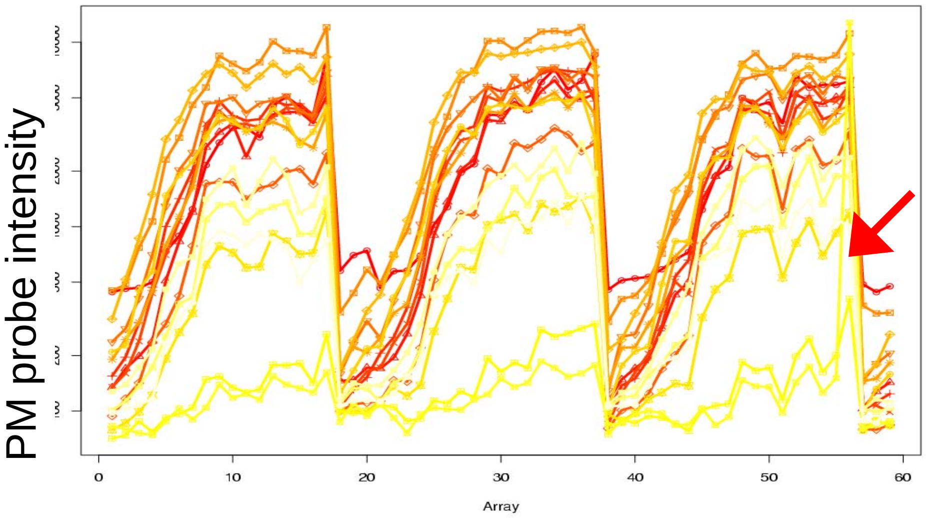
Differentially expressing



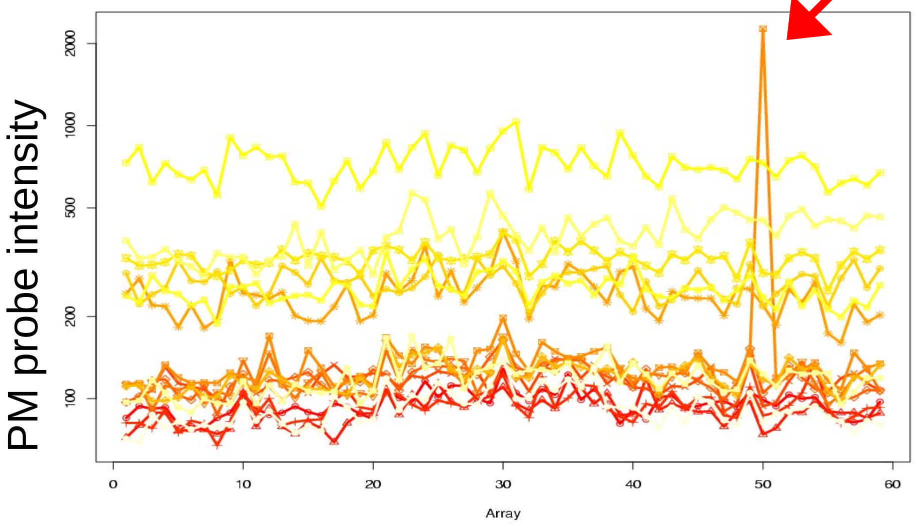
Non Differential



Differentially expressing



Non Differential



The RMA model

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

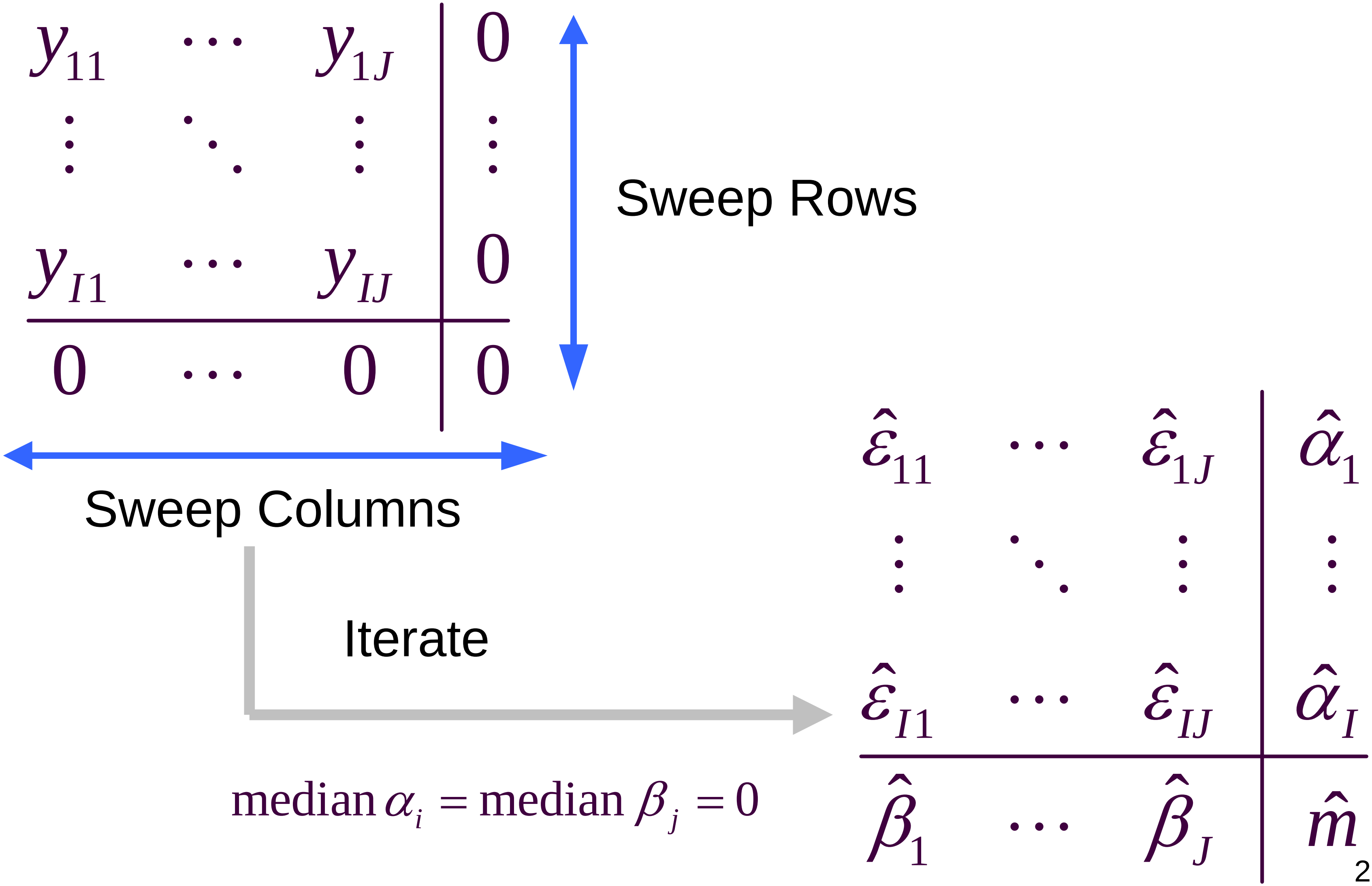
where $y_{kij} = \log_2 \mathrm{N}\left(\mathrm{B}\left(PM_{kij}\right)\right)$

α_{ki} is a probe-effect i= 1,…,I

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

k=1,…,K is the number of probesets

Median Polish Algorithm



$$\begin{array}{ccc|c} y_{11} & \cdots & y_{1J} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ y_{I1} & \cdots & y_{IJ} & 0 \\ \hline 0 & \cdots & 0 & 0 \end{array}$$

Sweep Rows

Sweep Columns

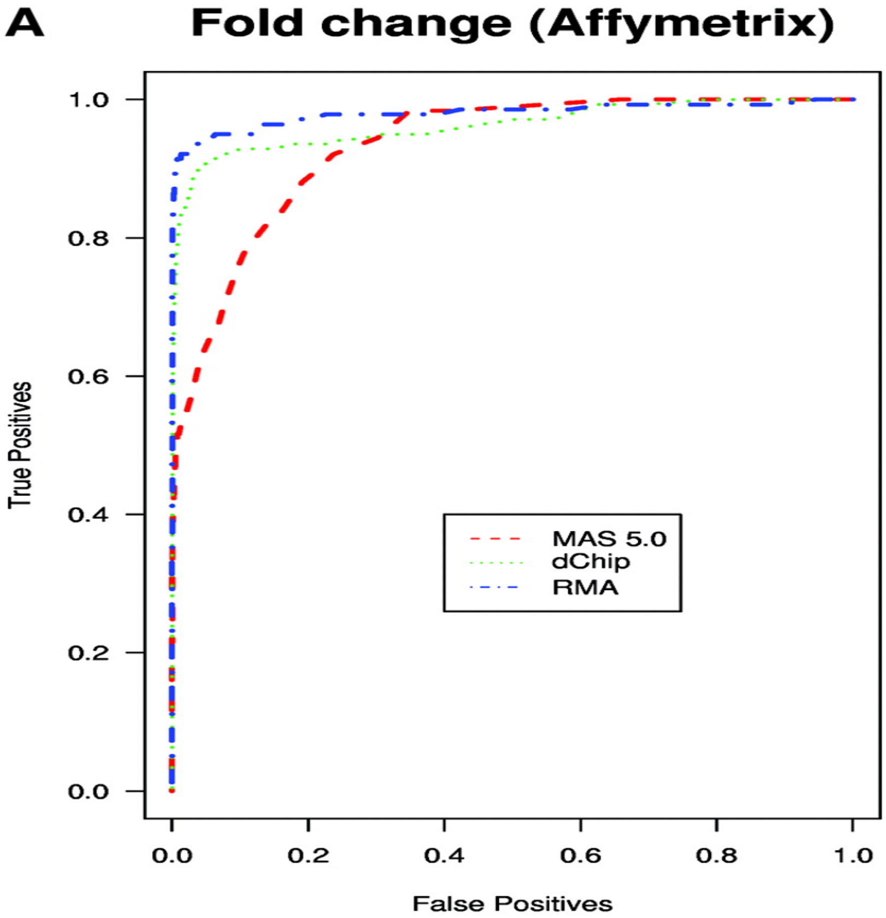
Iterate

$$\begin{array}{ccc|c} \hat{\varepsilon}_{11} & \cdots & \hat{\varepsilon}_{1J} & \hat{\alpha}_1 \\ \vdots & \ddots & \vdots & \vdots \\ \hat{\varepsilon}_{I1} & \cdots & \hat{\varepsilon}_{IJ} & \hat{\alpha}_I \\ \hline \hat{\beta}_1 & \cdots & \hat{\beta}_J & \hat{m} \end{array}$$

$$\text{median}\,\alpha_i = \text{median}\,\beta_j = 0$$

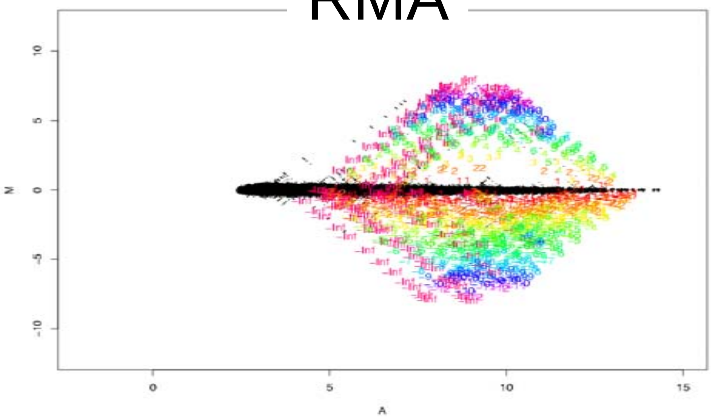
RMA mostly does well in practice

Detecting Differential Expression

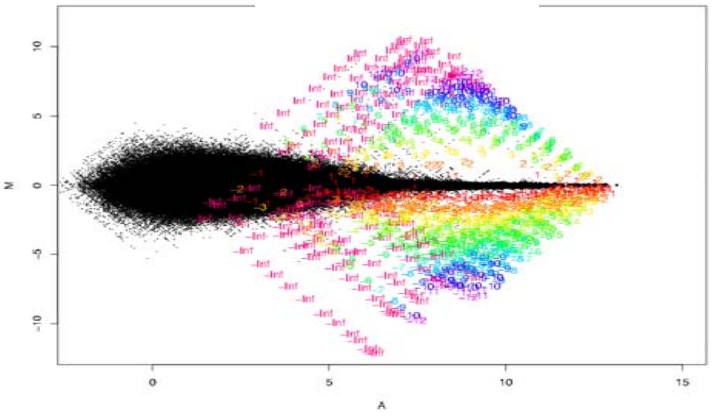


Not noisy in low intensities

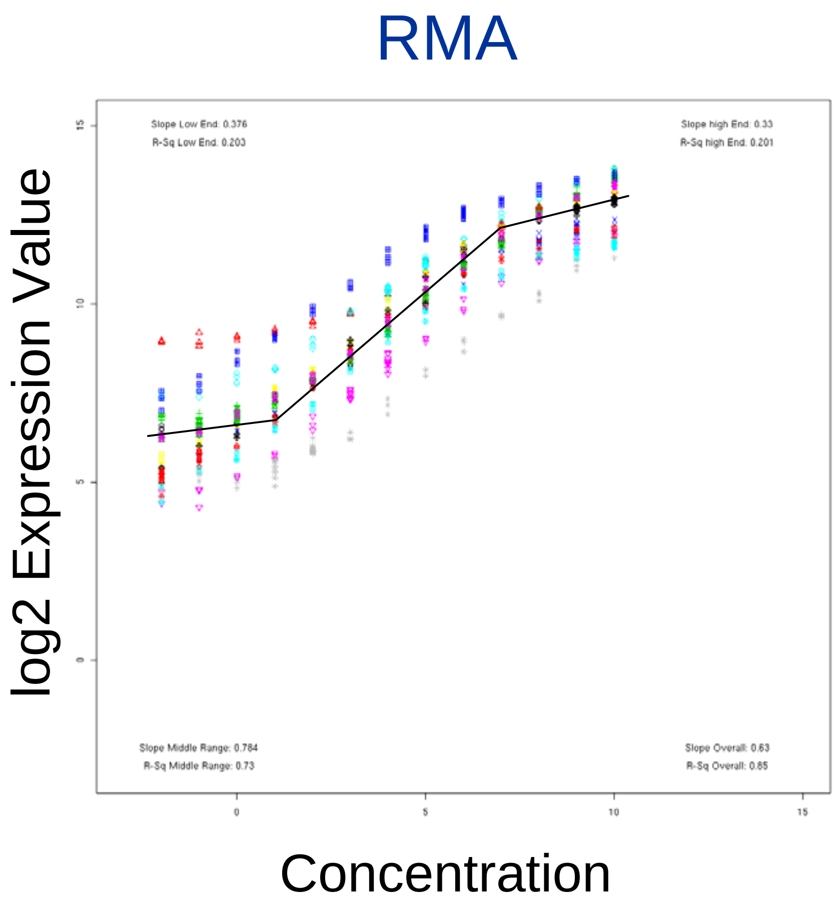
RMA

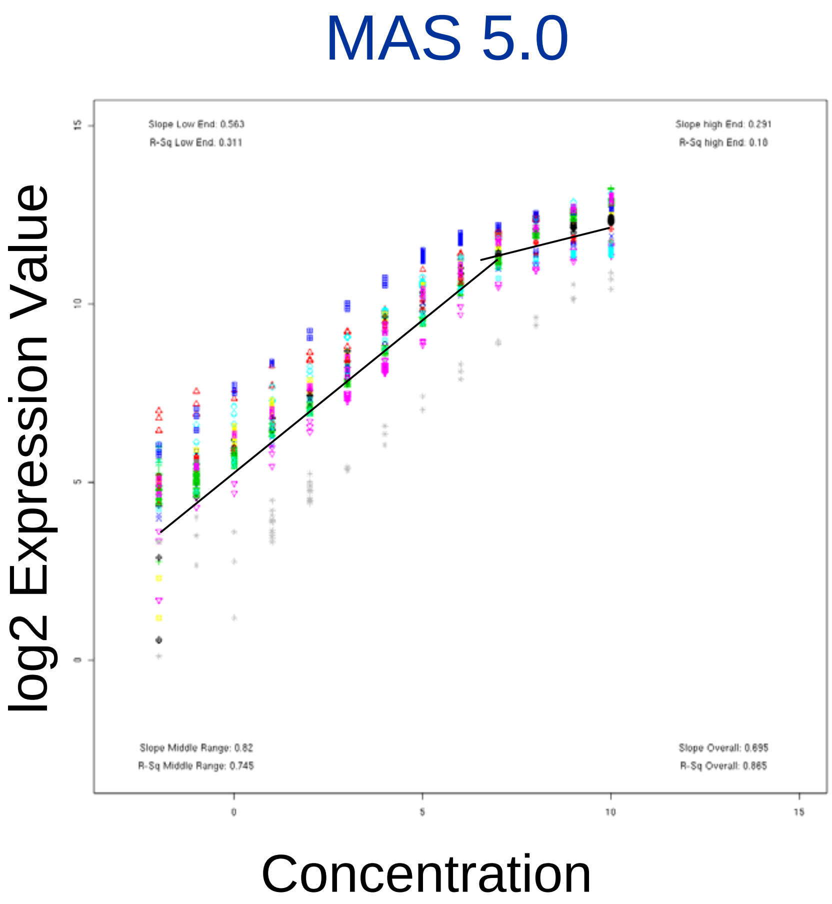


MAS 5.0



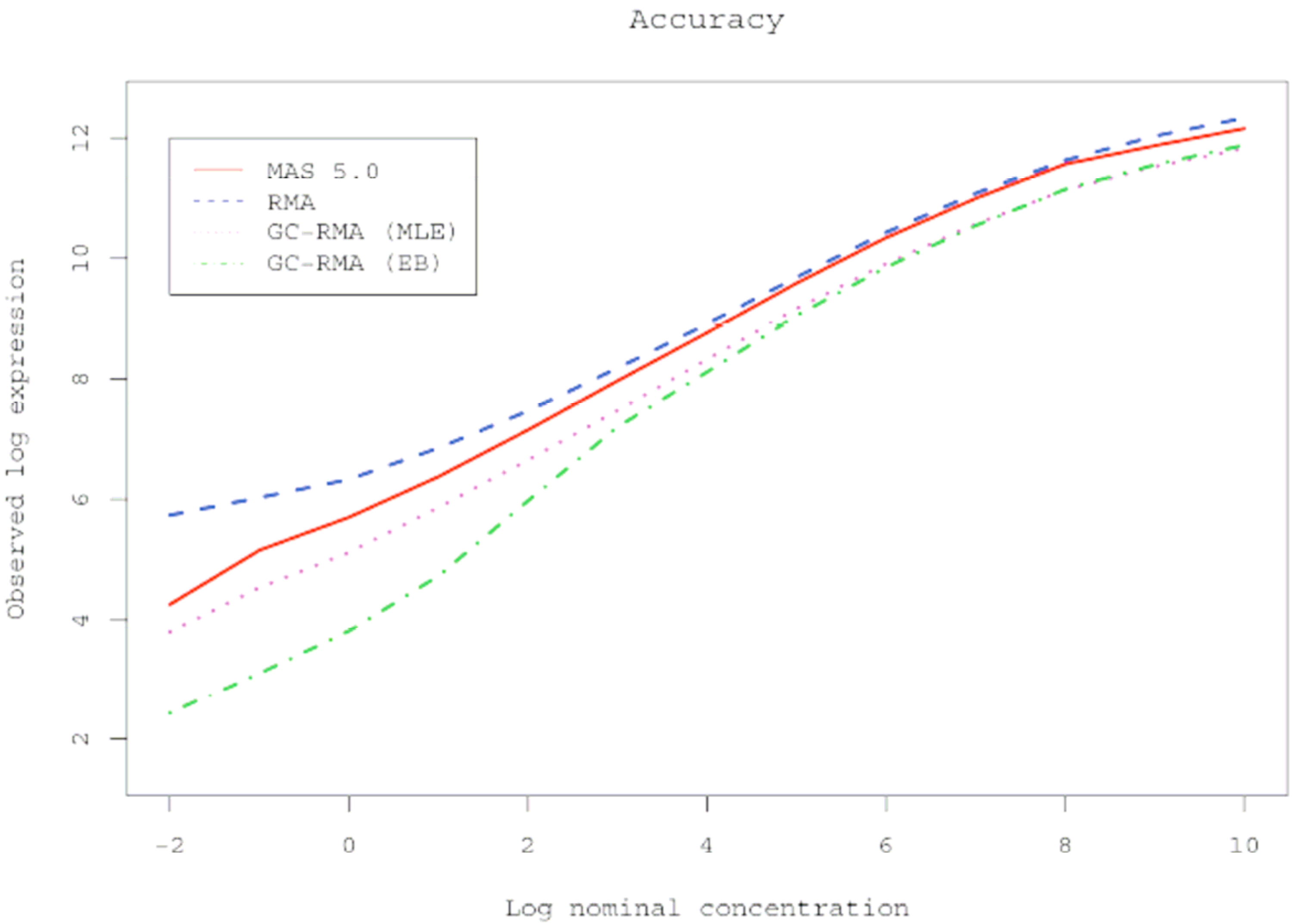
One Drawback





Linearity across concentration. GCRMA fixes this problem

GCRMA improve linearity



- See affycomp for more comparisons between RMA, GCRMA, MAS5 and many other expression measures.
- http://affycomp.biostat.jhsph.edu/

- Assessments shown in this talk are based on Affymetrix U95A Spike-in dataset

An Alternative Method for Fitting a PLM

- Robust regression using M-estimation
- In this talk, we will use Huber's influence function. The software handles many more.
- Fitting algorithm is Iteratively Re-weighted Least Squares with weights dependent on current residuals

$$\frac{\psi\left(r_{kij}\right)}{r_{kij}}$$

We Will Focus on the Summarization PLM

Array Effect
(Expression value)

- Array effect model

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

Pre-processed
Log PM intensity

Probe Effect

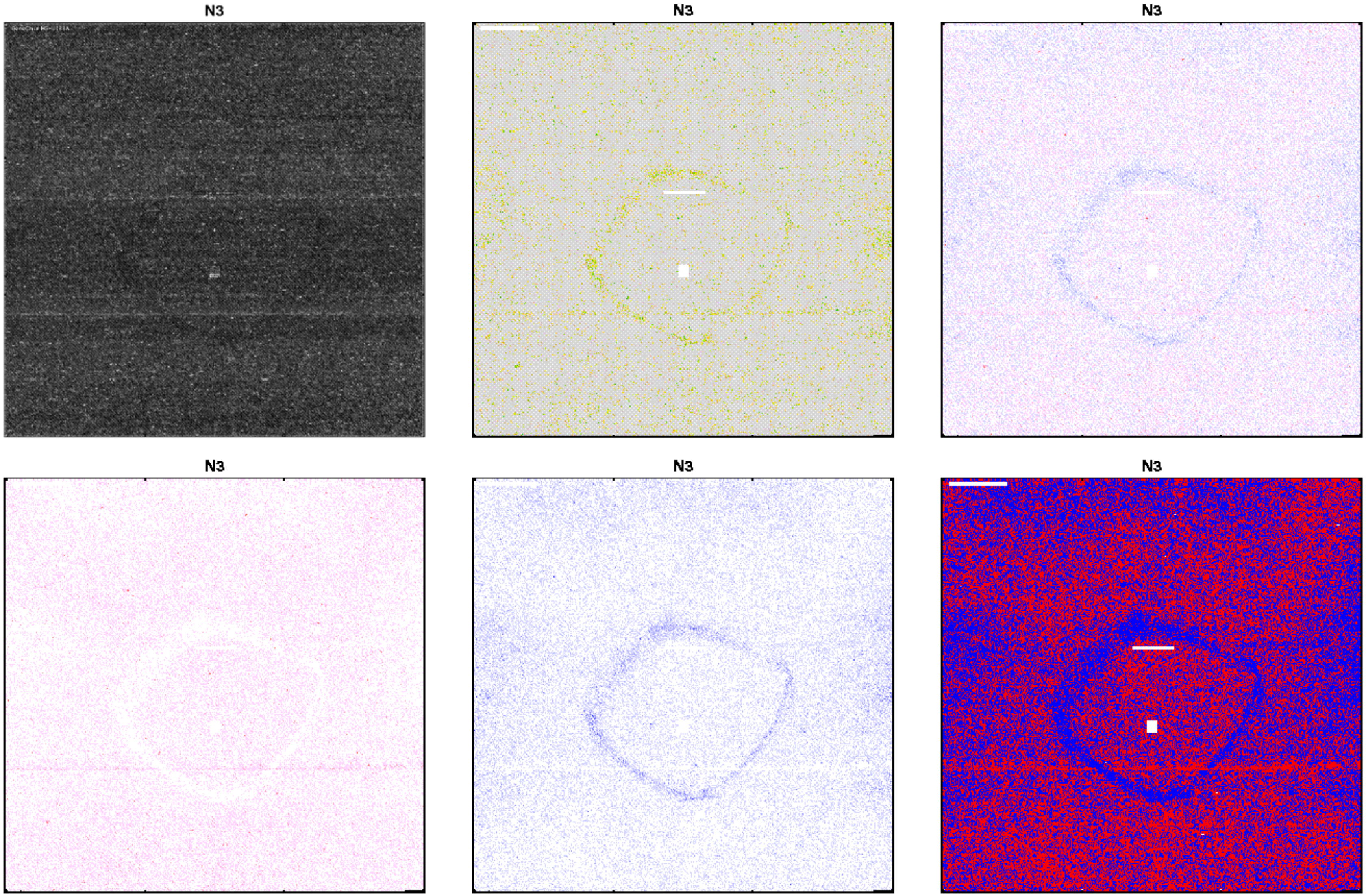
With constraint

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

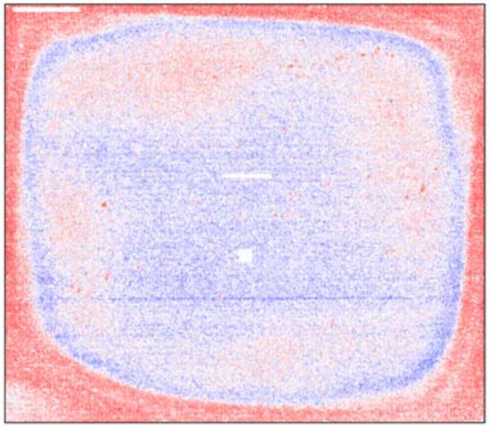
Quality Assessment

- Problem: Judge quality of chip data
- Question: Can we do this with the output of the Probe Level Modeling procedures?
- Answer: **Yes**. Use weights, residuals, standard errors and expression values.

Chip pseudo-images

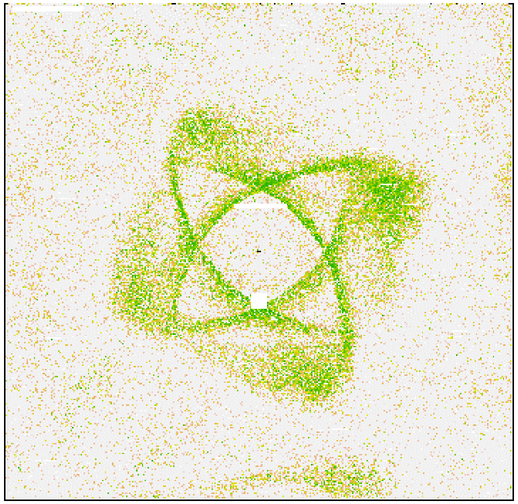


An Image Gallery

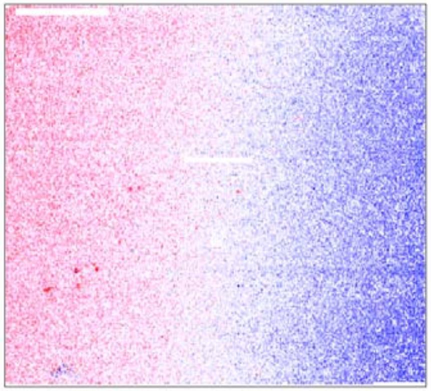


“Ring of Fire”

“Crop Circles”



“Tricolor”



http://PLMImageGallery.bmbolstad.com

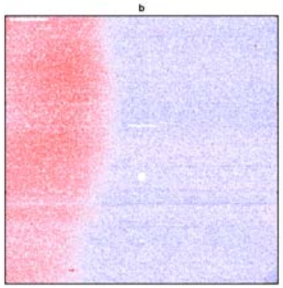
NUSE Plots

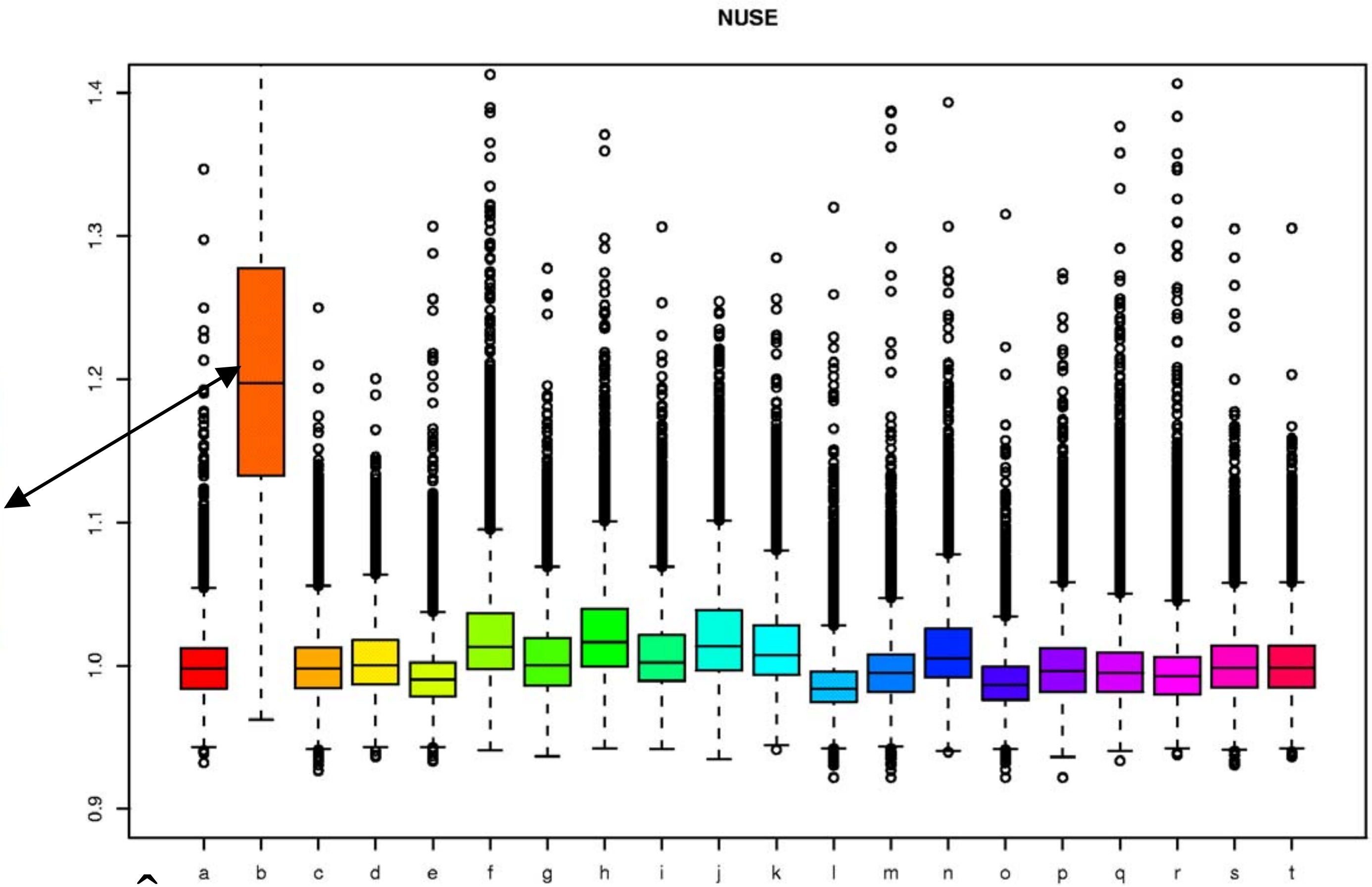
Normalized

Unscaled

Standard

Errors





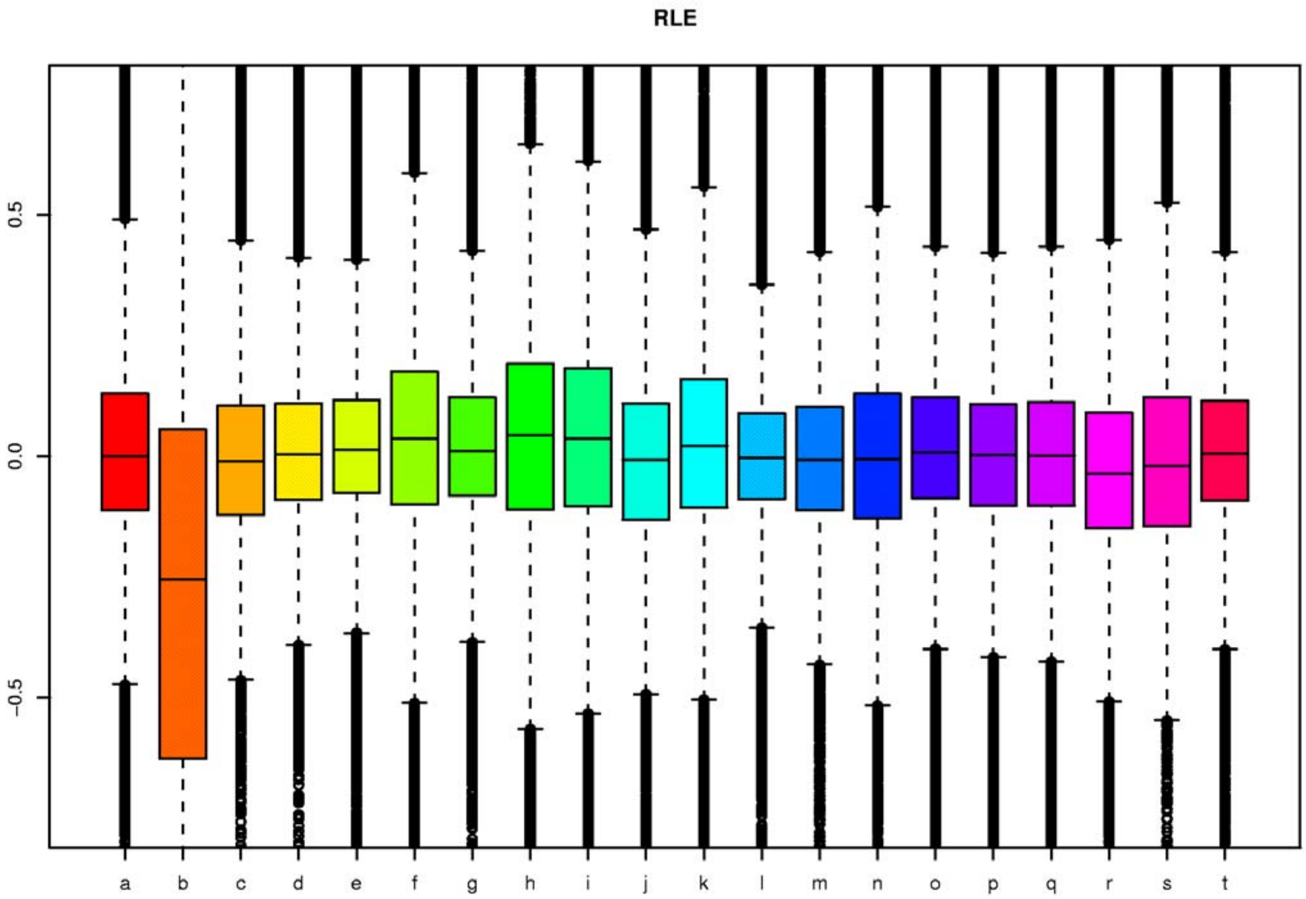
$$NUSE(\hat{\beta}_{kj}) = \frac{SE(\hat{\beta}_{kj})}{med_{j}(SE(\hat{\beta}_{kj}))}$$

RLE Plots

Relative

Log

Expression



$$RLE(\hat{\beta}_{kj}) = \hat{\beta}_{kj} - med_j(\hat{\beta}_{kj})$$

BIOCONDUCTOR

open source software for bioinformatics

- Based on the R language
- Approx 160 packages (at 1.8 Release Apr 2006)
- All source code is available
- Microarray data is a major focus, but also currently some software for dealing with Mass Spec data, Cell Based Assays (Flow Cytometry), with others application areas planned and expected.

- http://www.bioconductor.org

Installing BioConductor

```
source("http://www.bioconductor.org/biocLite.R")
biocLite()
```

- Installs a small (approx 20) subset of the packages
- Additional packages can be installed

```
biocLite(c(“simpleaffy”,”makecdfenv”))
```

- This handles all the (inter) dependencies between the different packages

Dealing with Affymetrix Data

- **affy** – Data structures for storing probe intensity data. Supplies RMA, general functionality for combining different background, normalization, summarization schemes. Basic methods for examining probe intensity data.
- **affyPLM** – Methods for fitting probe level models. QC tools.
- **gcrma** – provides the GCRMA expression measure and background correction
- **simpleaffy** – provides Affymetrix standard QC

Affymetrix Meta-data Packages

- *cdfenv packages* – contain processed CDF information
- *Probe packages* – contain probe sequence information
- *Annotation packages* – contain annotation information created using public data repositories
- eg for u133A chips these would be
  - **hgu133acdf**
  - **hgu133aprobe**
  - **hgu133a**

- Automatically downloaded and installed on first use.

Case Study

- Data retrieved from a public repository, GEO
- Data Series GSE2603
- Minn et al (2005) Genes that mediate breast cancer metastasis to lung. Nature. 2005 Jul 28;436(7050):518-24
- 121 HG-U133A microarrays

Starting up

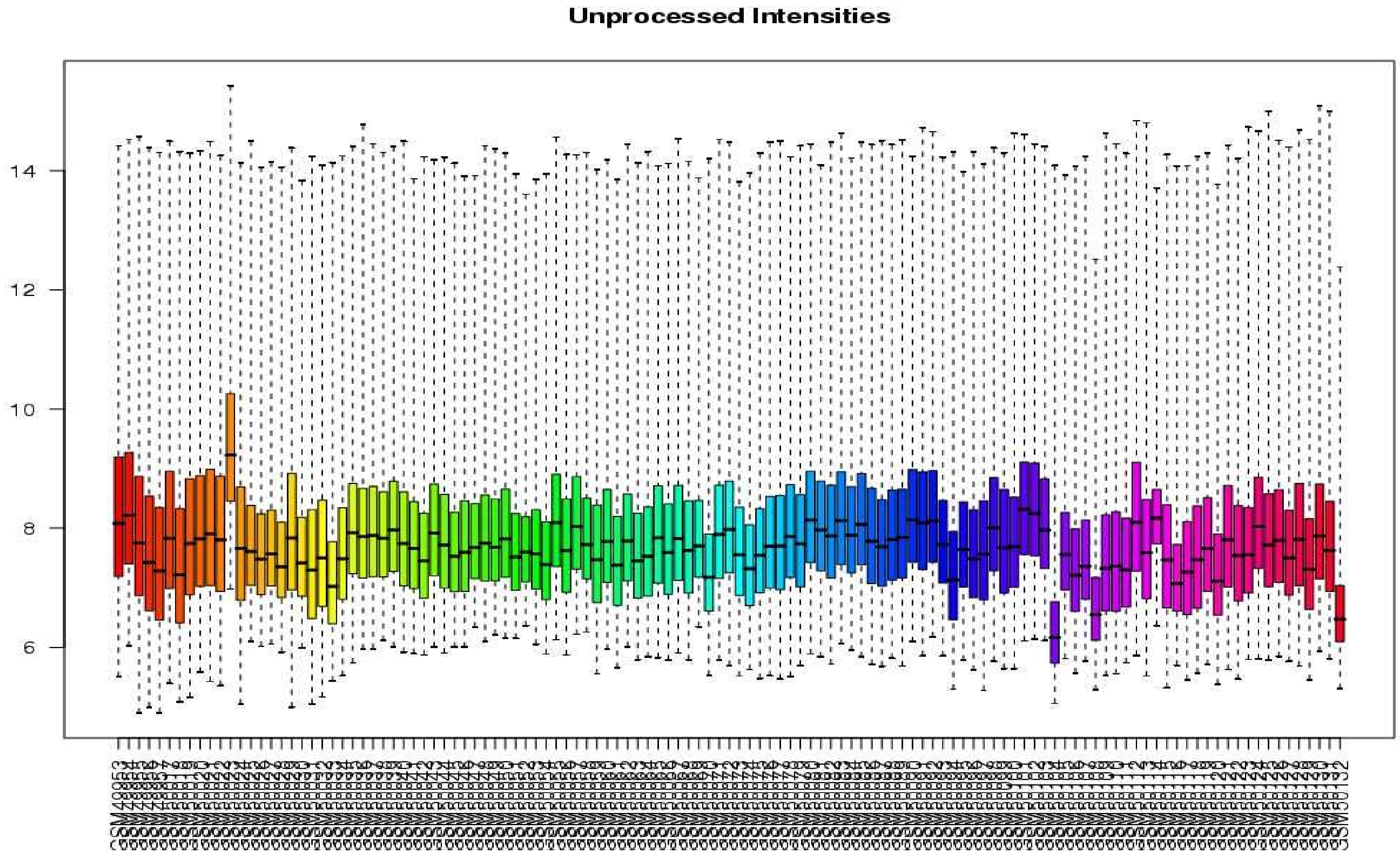
```
library(affyPLM)
### loads requisite packages including
### affy, Biobase, gcrma etc
```

Reading in the data

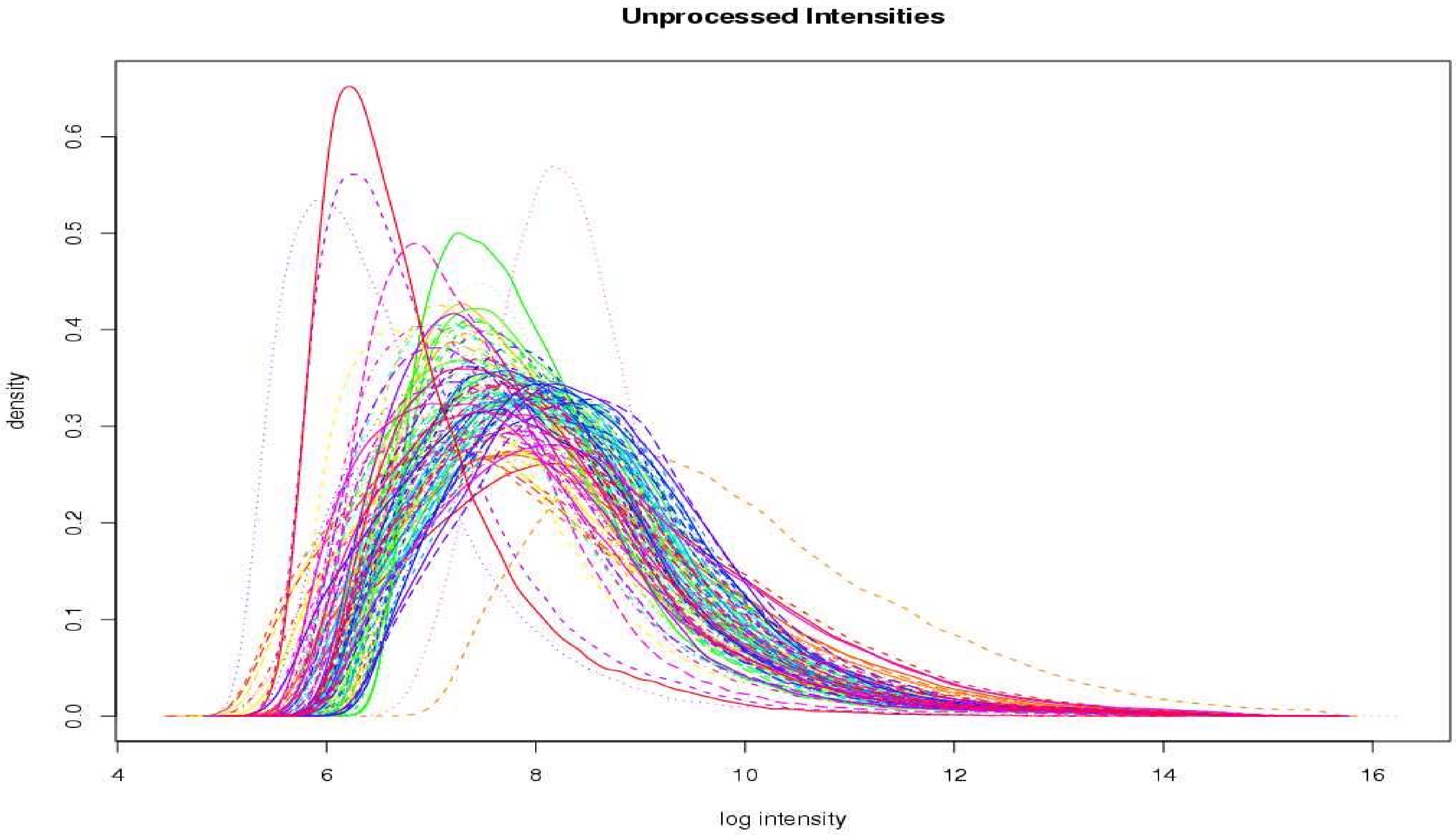
```
abatch.raw <- ReadAffy()
```

- Reads the all the CEL files in current directory into an R S4 object known as an `AffyBatch`
- Note we don’t need to supply the CDF file. Instead a processed version of it will get automatically downloaded if needed on the first use of that chip type.
- An `AffyBatch` is an object which can store probe-intensities, along with meta-data such as phenotypic data, for a set of arrays.
- Accessor functions like `pm()`, `mm()` allow access to the PM or MM probe intensities.
- Other functions can be used to visually examine the data

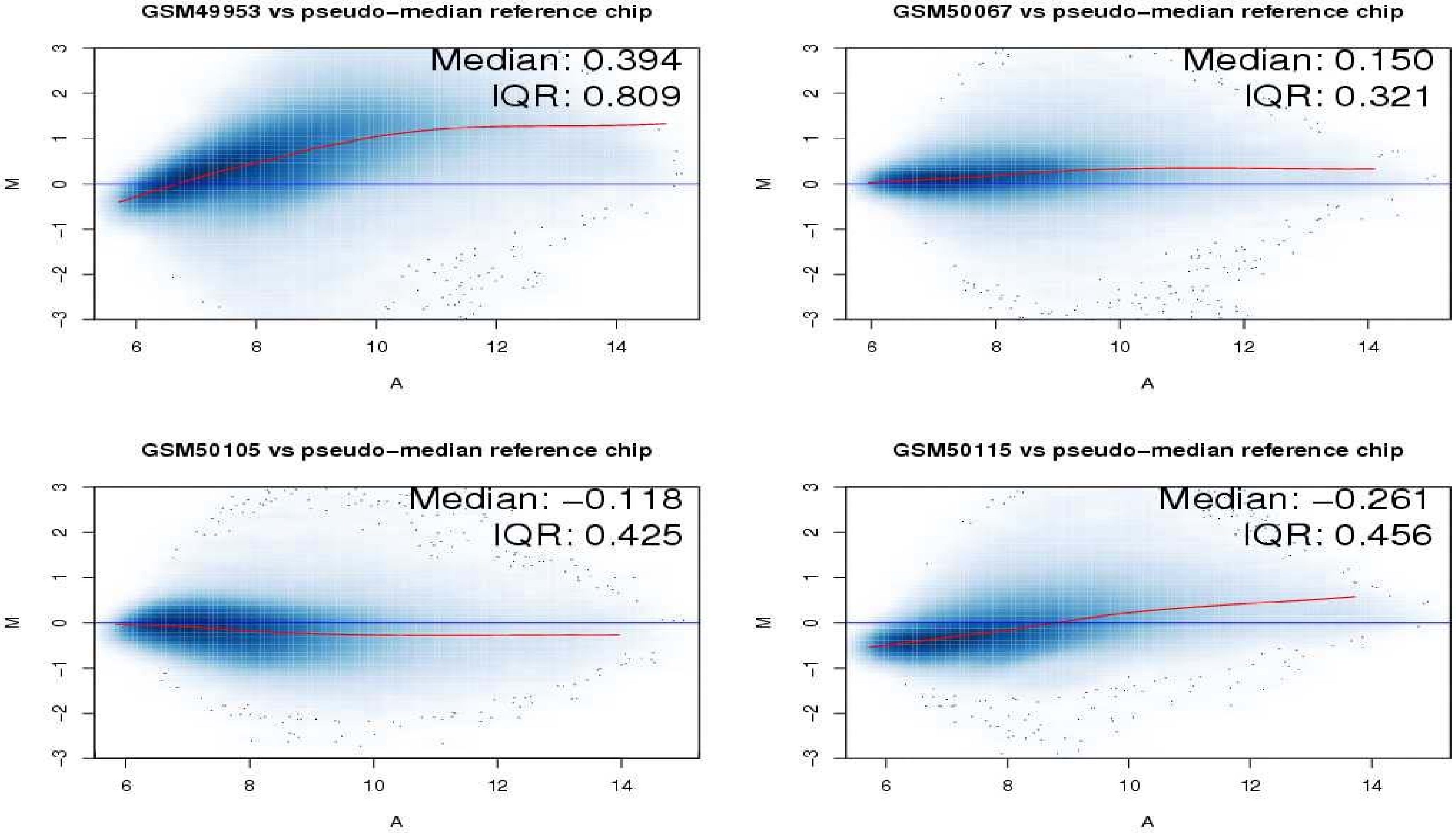
…..



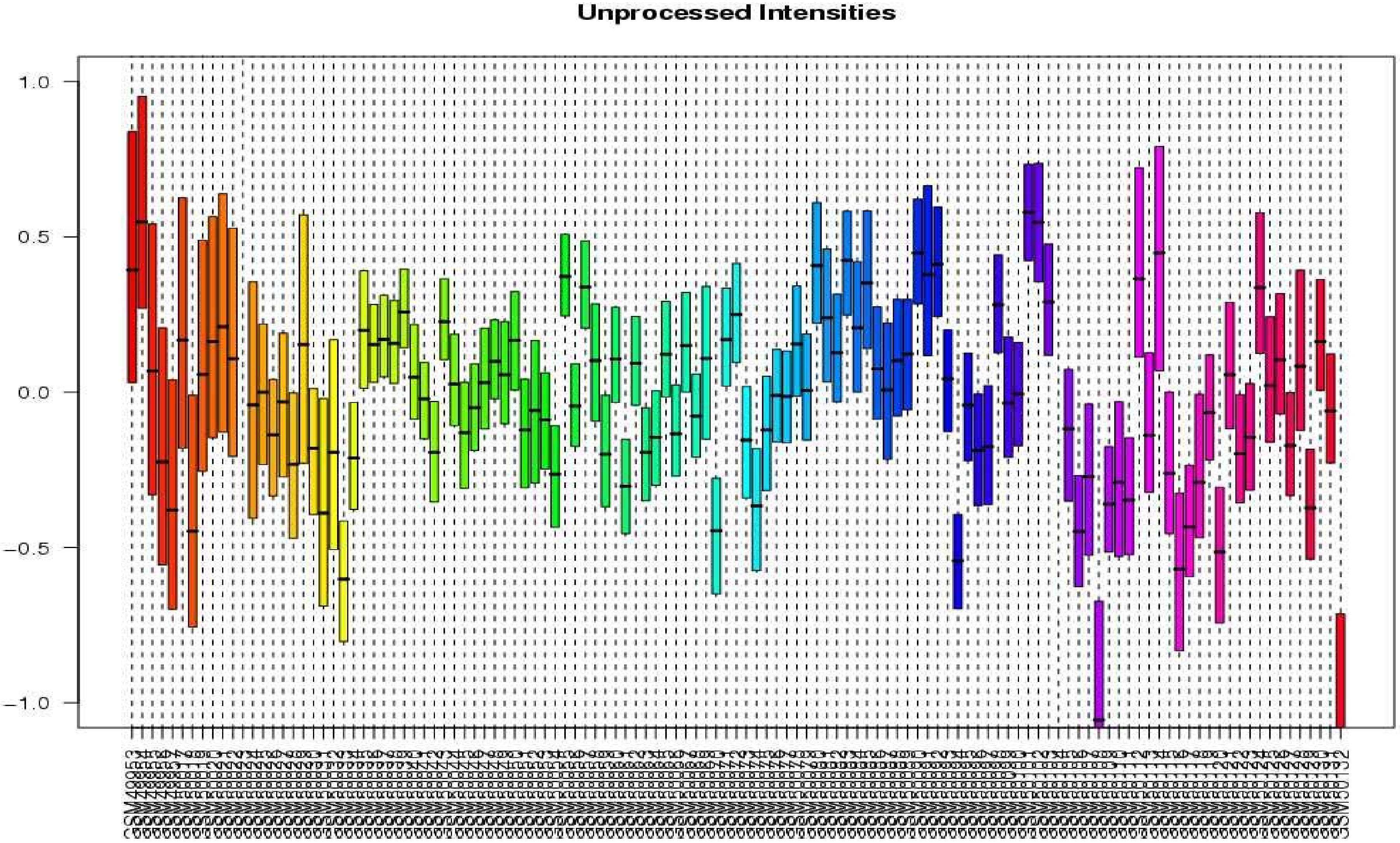
```
boxplot(abatch.raw)
```



```
hist(abatch.raw)
```

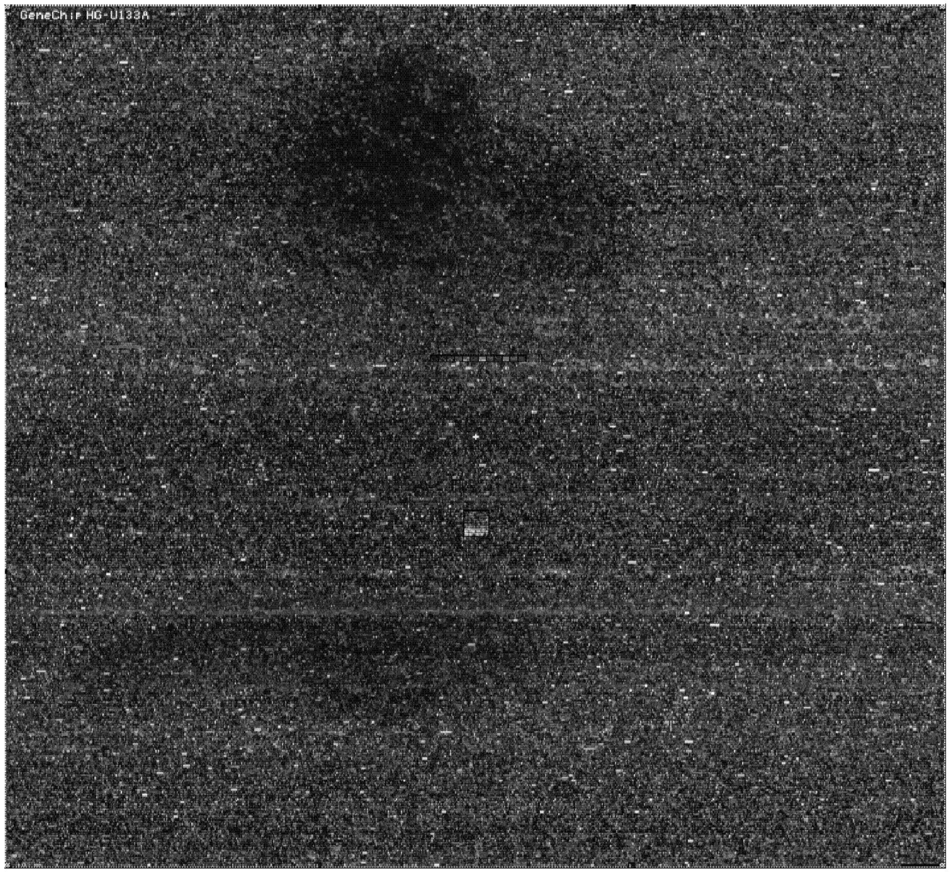


```
MAplot(abatch.raw,plot.method=“smoothScatter”,
        which=c(1,56,94,104))
```



```
Mbox(abatch.raw)
```

GSM50110.CEL



```
image(abatch.raw[,99])
```

Manually Preprocessing

- Background correction

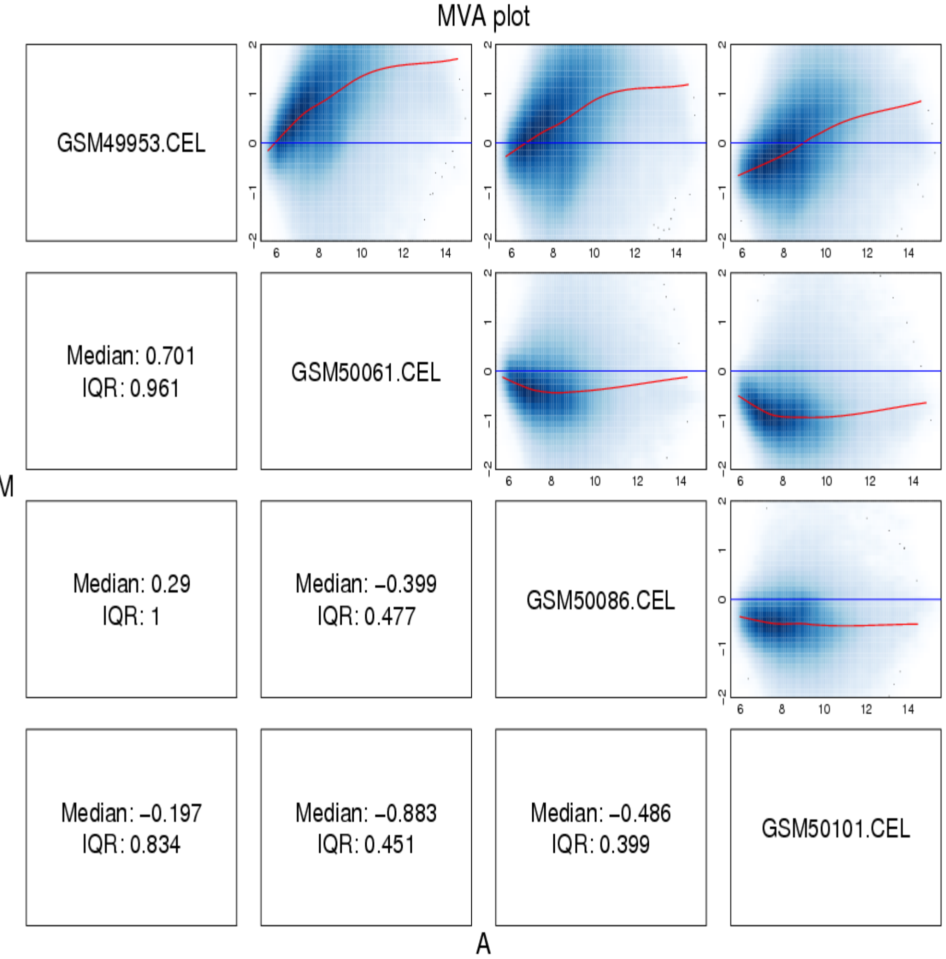
```
abatch.rmabg<-bg.correct.rma(abatch.raw)
abatch.gcrmabg <- bg.correct.gcrma(abatch.raw)
```

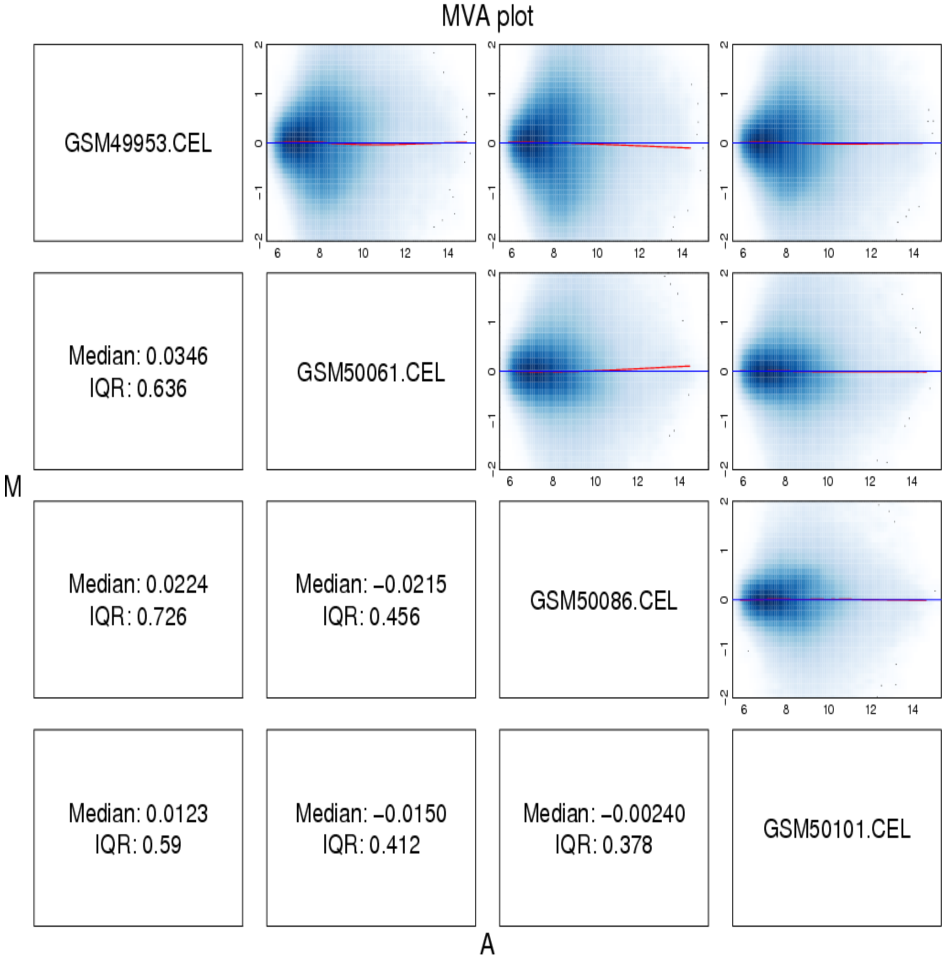
- Normalization

```
abatch.norm <- normalize(abatch.raw)
```

↑

Defaults to quantile normalization, but an
Optional argument can be used to select
an alternative method.



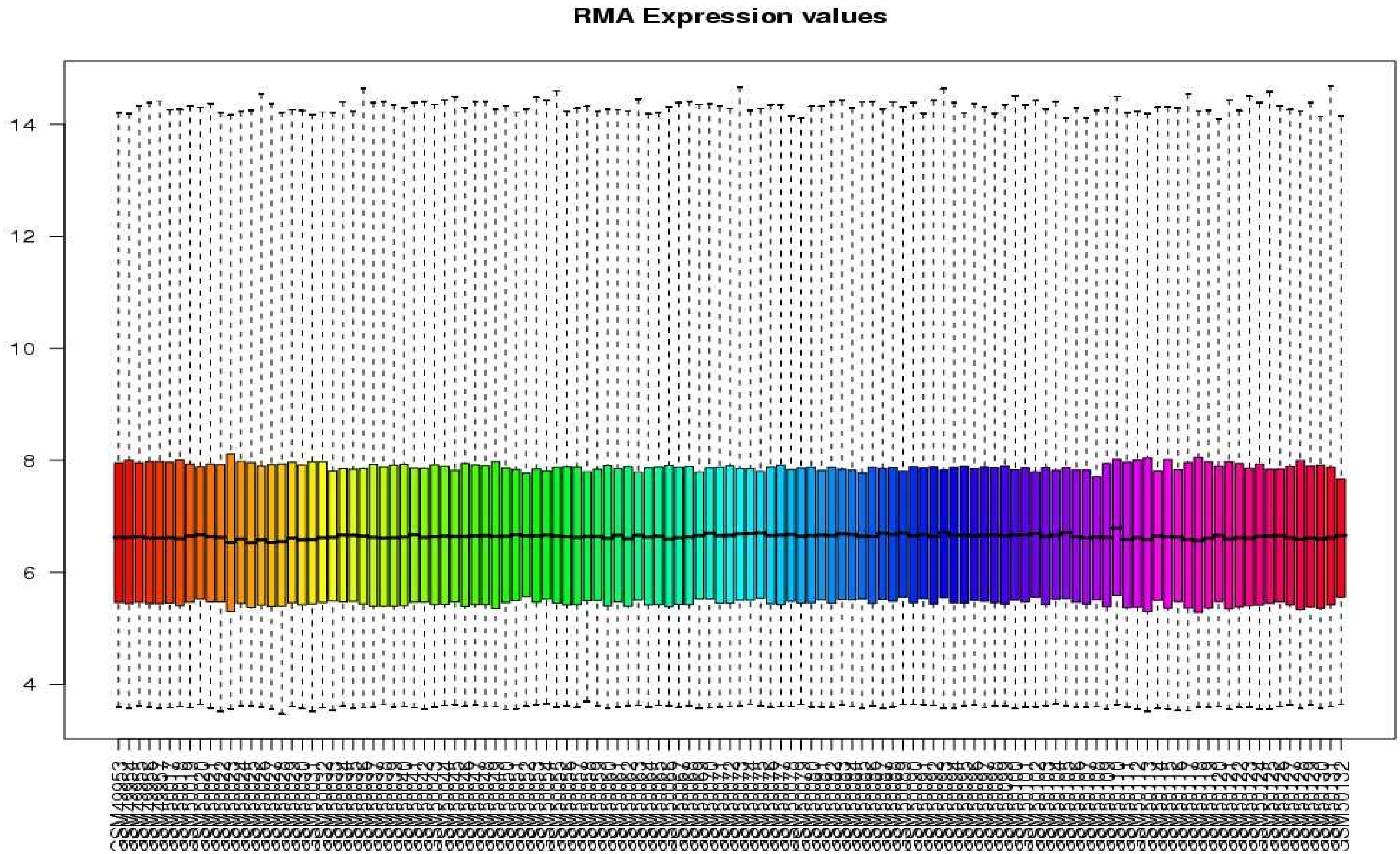


```
MAplot(abatch.rawdata, which=c(1,50,75,90), pairs=TRUE,
  ylim=c(-2,2), plot.method="smoothScatter")
MAplot(abatch.norm, which=c(1,50,75,90), pairs=TRUE,
  ylim=c(-2,2), plot.method="smoothScatter")
```

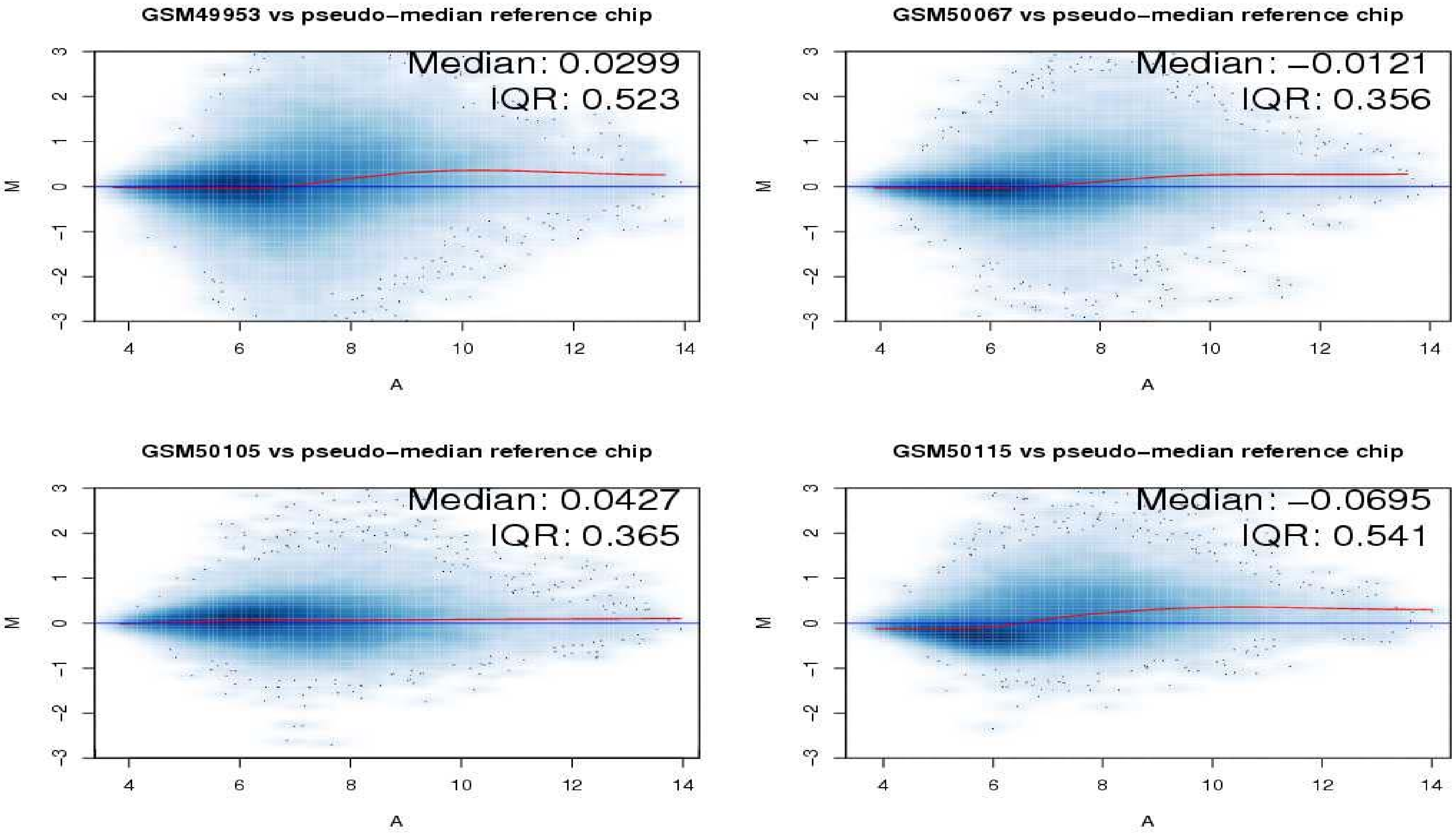
Computing RMA

```
eset.rma <- rma(abatch.raw)
```

- The function `rma()` returns an `exprSet` (in the future this likely to be replaced by the `eSet`) containing RMA values.
- An `exprSet` stores expression values and related meta-data. Many BioConductor functions for high-level analysis accept these as input.
- `gcrma()` can be used to get GCRMA values



```
boxplot(eset.rma)
```



```
MAplot(eset.rma,plot.method=“smoothScatter”,
          which=c(1,56,94,104))
```

Other ways to get expression measures

- General methods give user control over which pre-processing steps occur:

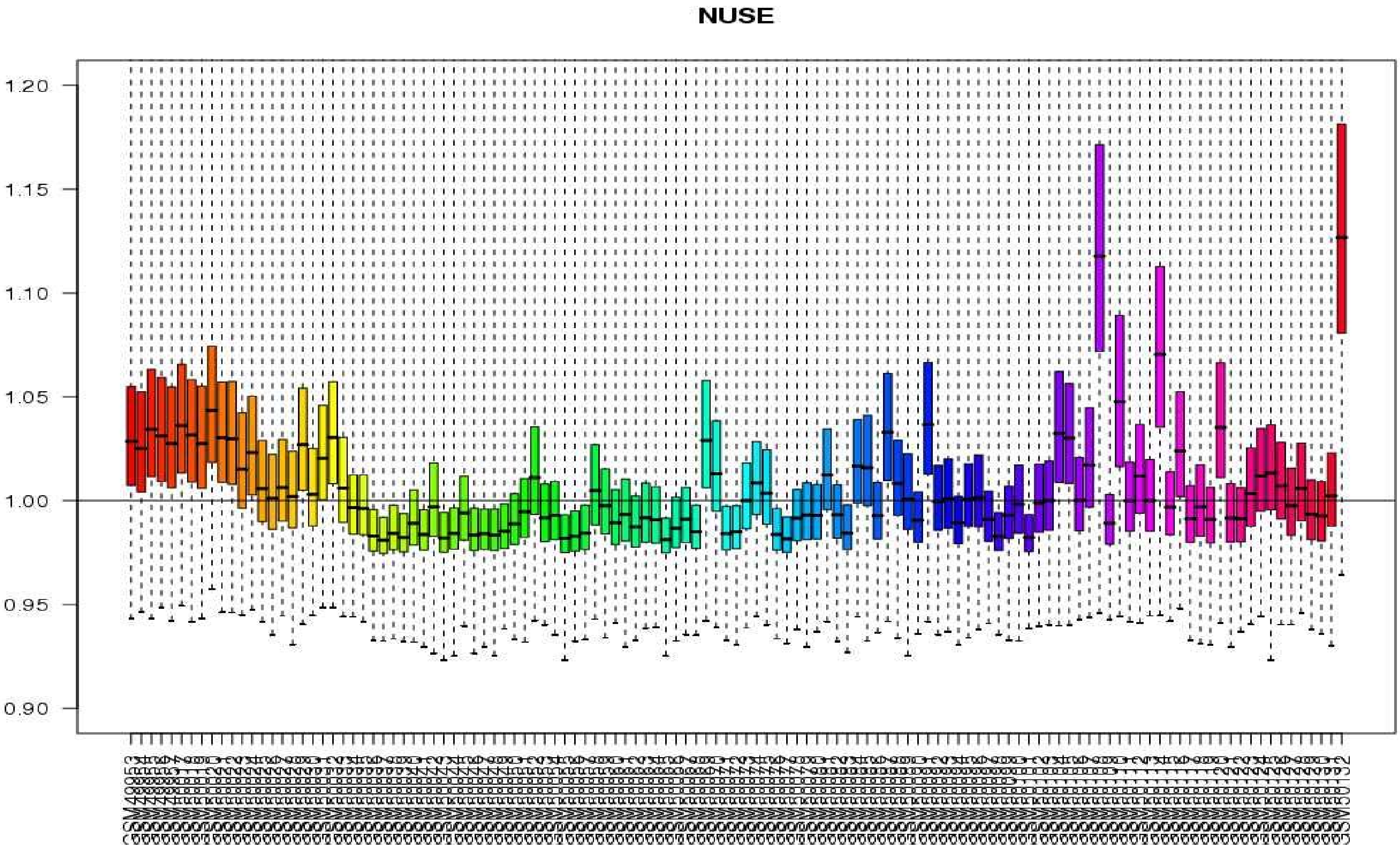
`threestep()`- memory and run time efficient

`expresso()` – easily extensible by outside users but slower and generally consumes much more memory

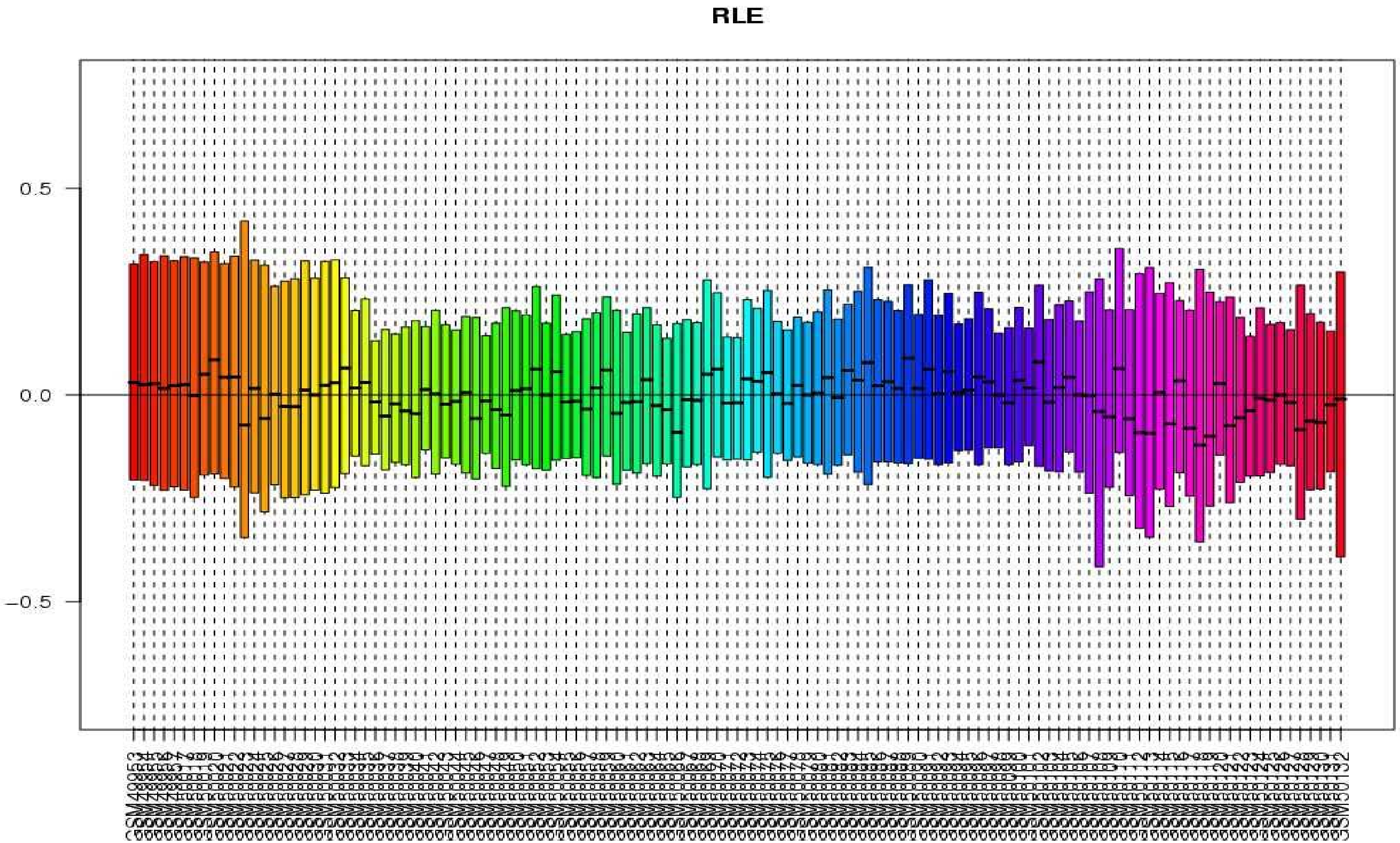
Carrying out QC Assessment

```
Pset <- fitPLM(abatch.raw)
```

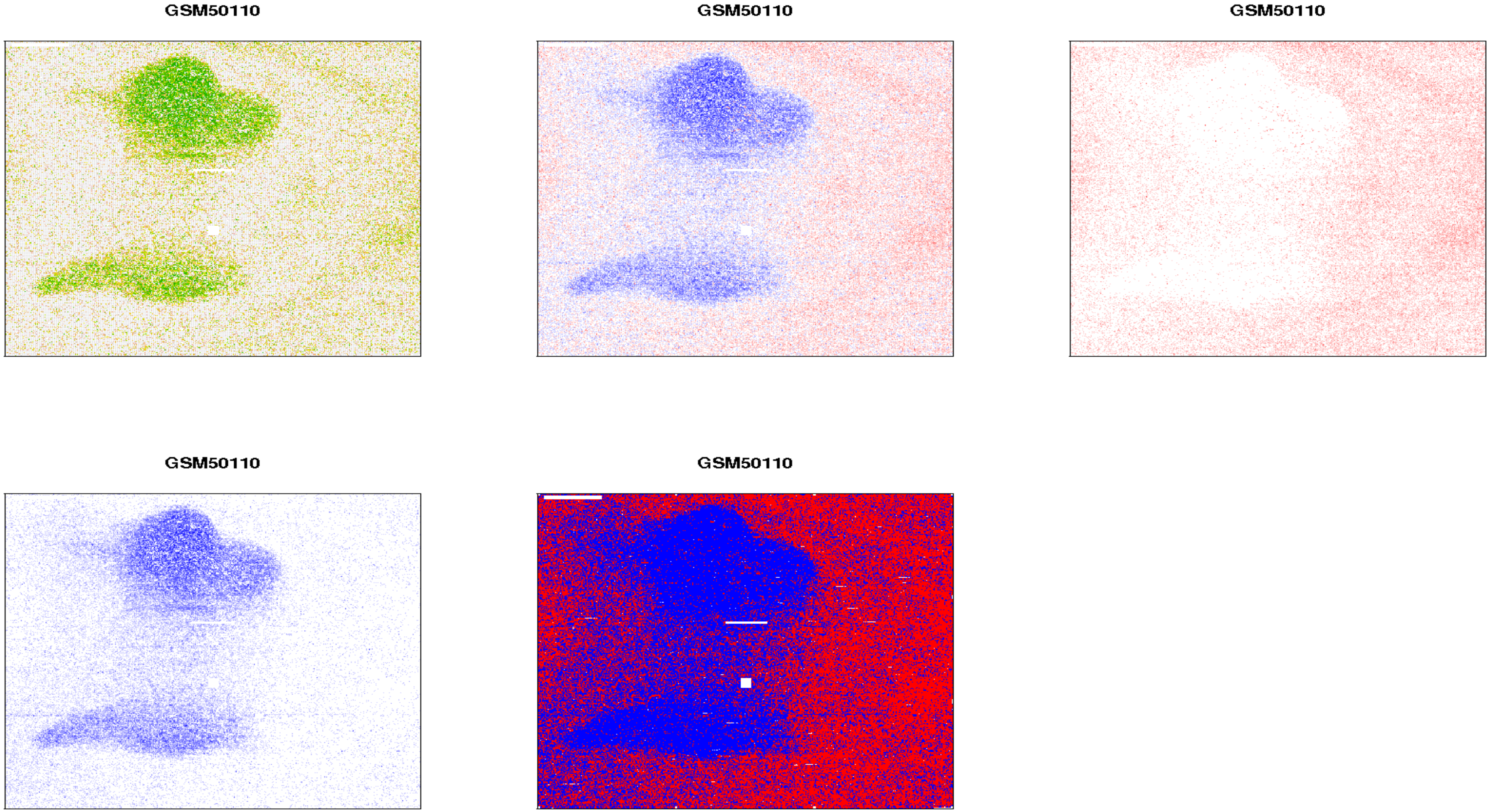
- A `PLMset` object is the return value of fitPLM(). It stores parameter estimates and their standard errors. Also residuals and weights from the IRLS procedure.



```
NUSE(Pset)
NUSE(Pset,type=“stats”) # get median/IQR
```



```
RLE(Pset)
RLE(Pset,type="stats") # get median/IQR
```



```
image(Pset,which=99)
image(Pset,which=99,type=“resids”)
image(Pset,which=99,type=“pos.resids”)
image(Pset,which=99,type=“neg.resids”)
image(Pset,which=99,type=“sign.resids”)
```

Future Developments

- oligo – a package supporting low-level analysis of SNP, tiling and expression arrays
- BufferedMatrix – R tools for dealing with extremely large data objects outside main memory

Acknowledgements

- Terry Speed
- Rafael Irizarry
- Julia Brettschneider
- Francois Colin
- Zhijin (Jean) Wu
- Robert Gentleman
- Wolfgang Huber

- Any one else I happened to forget …

References

- Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics, 19, 185 (2003).
- Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, and Speed TP. Summaries of Affymetrix GeneChip Probe Level Data. Nucleic Acids Research, 31(4):e15, 2003.
- Bolstad BM, Collin F, Brettschneider J, Simpson K, Cope L, Irizarry RA, and Speed TP. (2005) Quality Assessment of Affymetrix GeneChip Data in Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Gentleman R, Carey V, Huber W, Irizarry R, and Dudoit S. (Eds.), Springer
- Wu, Z., Irizarry, R., Gentleman, R., Martinez Murillo, F. Spencer, F. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of American Statistical Association 99, 909-917 (2004)
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, and Zhang J. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5(10):R80

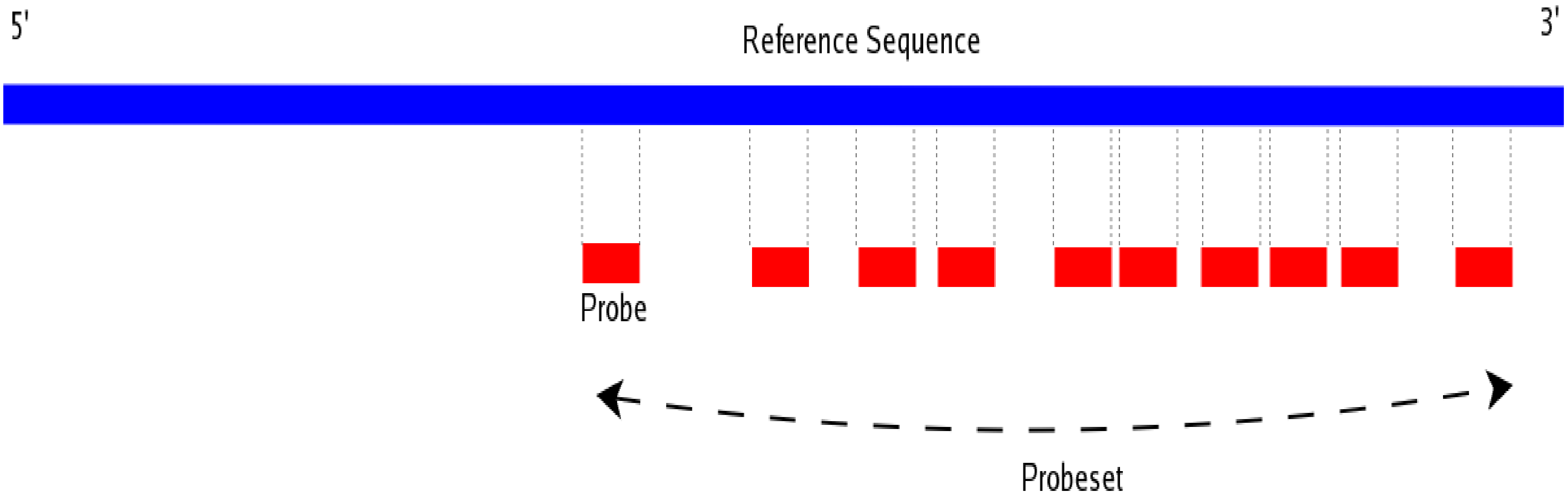
Supplemental Material

Affymetrix GeneChip

- Commericial mass produced high density oligonucleotide array technology developed by Affymetrix http://www.affymetrix.com
- Single channel microarray
- Todays talk relates to arrays designed for expression analysis

Image courtesy of Affymetrix Press Website.

Probes and Probesets



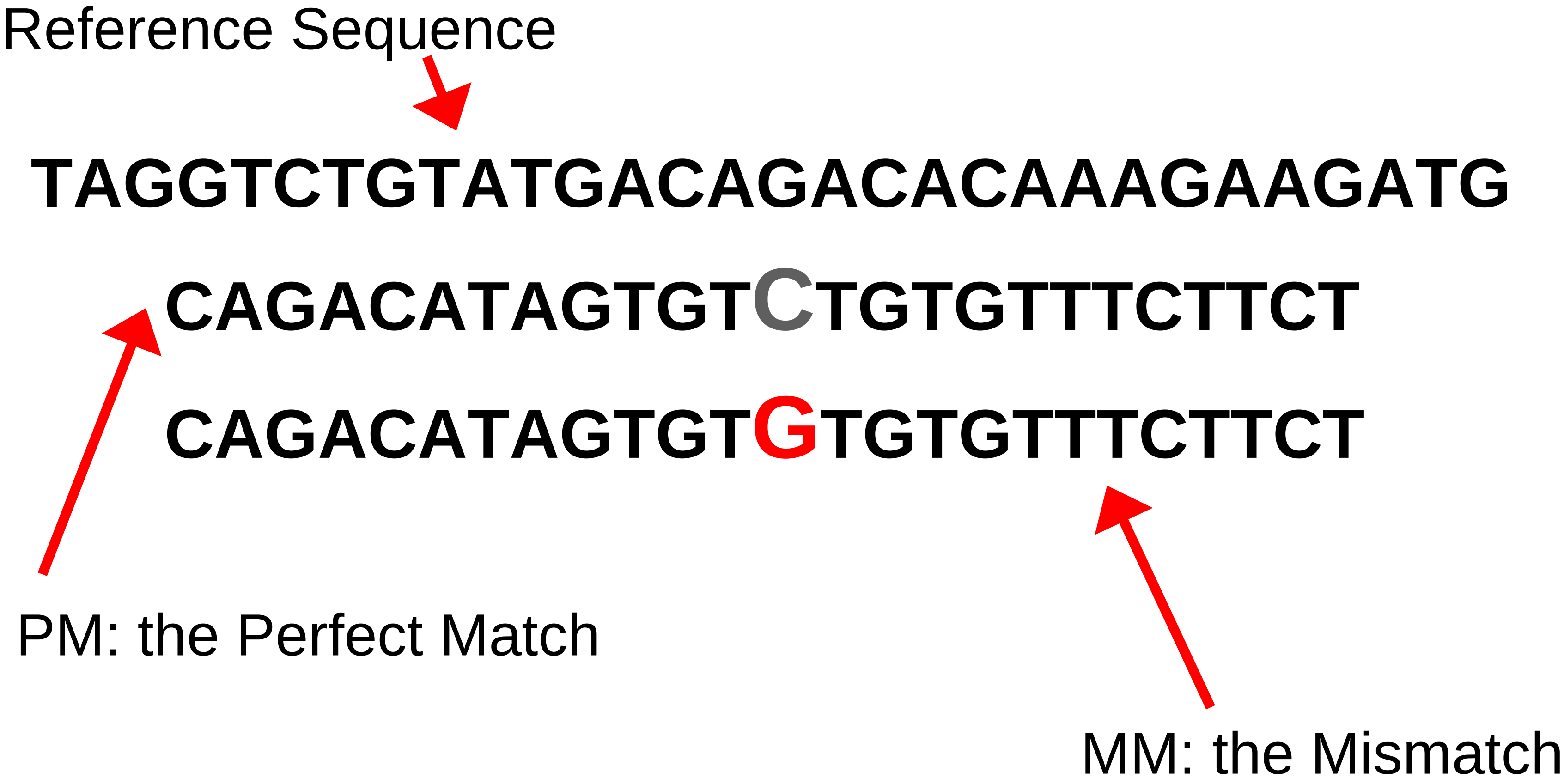
Typically 11 probe(pairs) in a probeset

Latest GeneChips have as many as:

54,000 probesets

1.3 Million probes

Two Probe Types



Reference Sequence

TAGGTCTGTATGACAGACACAAAGAAGATG

CAGACATAGTGTCTGTGTTTCTTCT

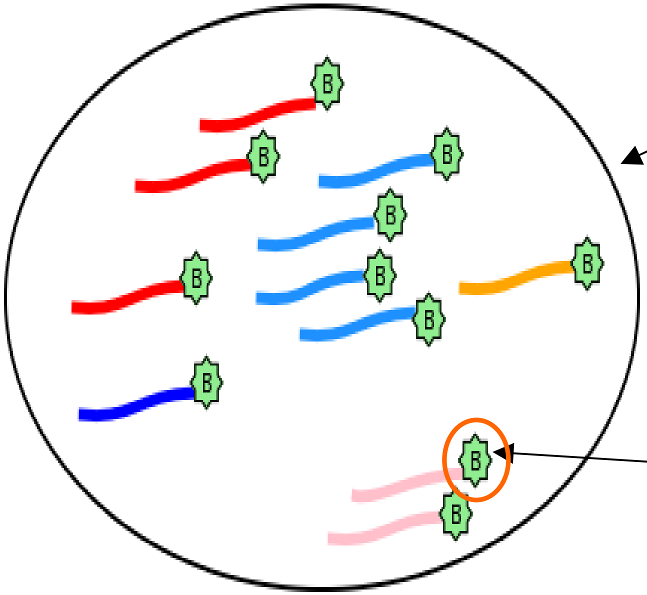
CAGACATAGTGTGTGTGTTTCTTCT

PM: the Perfect Match

MM: the Mismatch

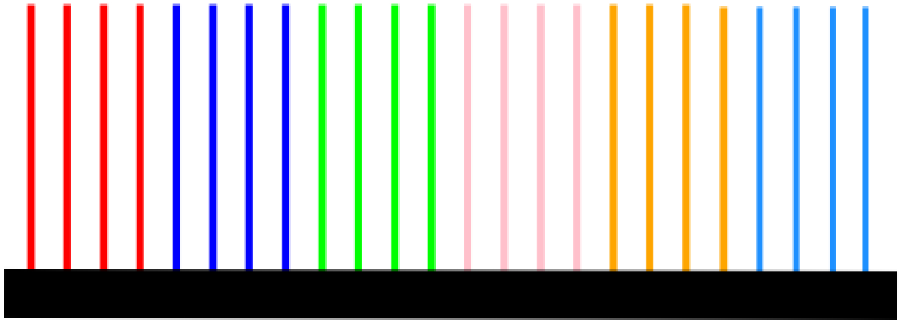
Note that about 30% of MM probe intensities are brighter than corresponding PM probe intensities.

Hybridization to the Chip

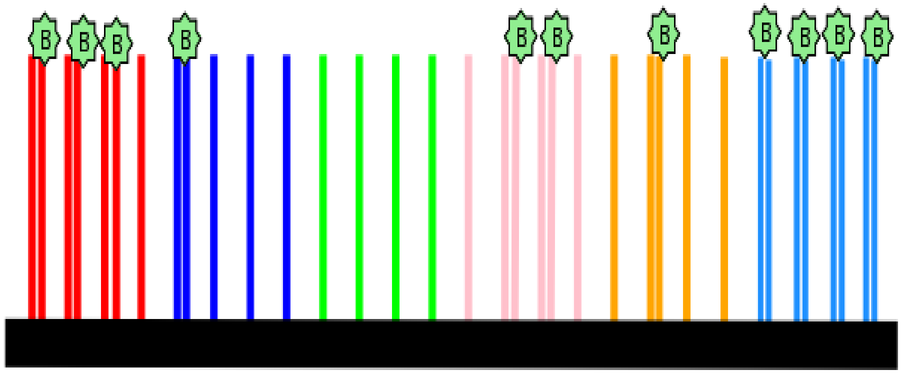


Sample of Fragmented Labeled RNA

Labeling molecule that fluoresces

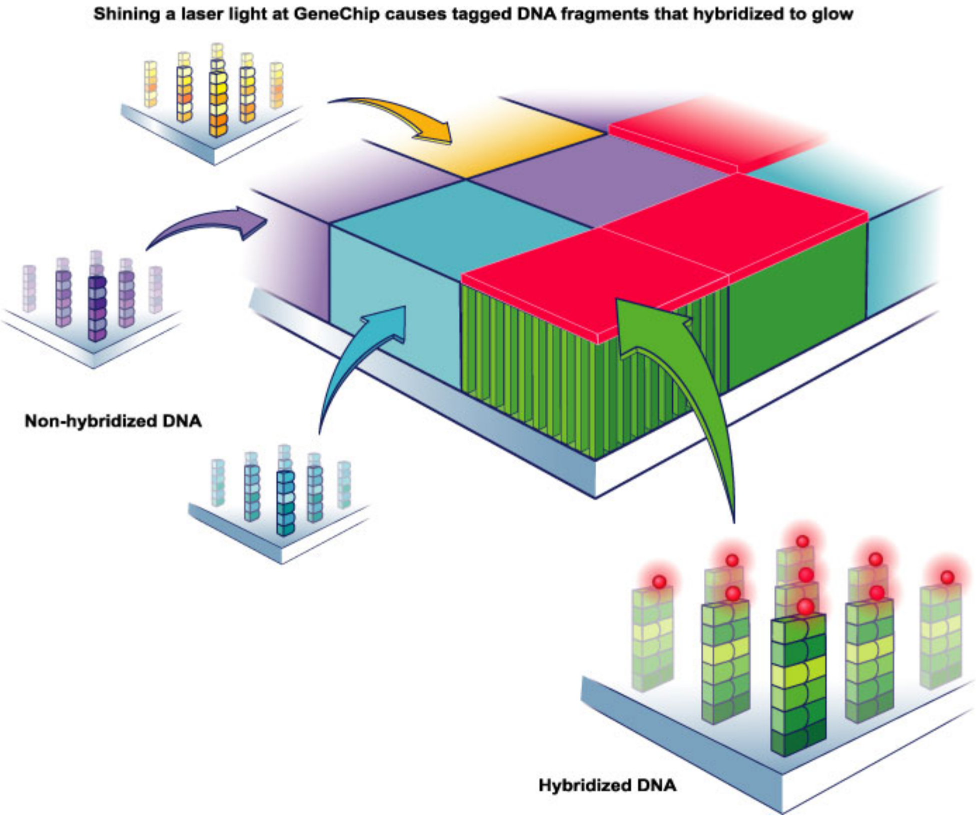


Before Hybridization



After Hybridization

The Chip is Scanned


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

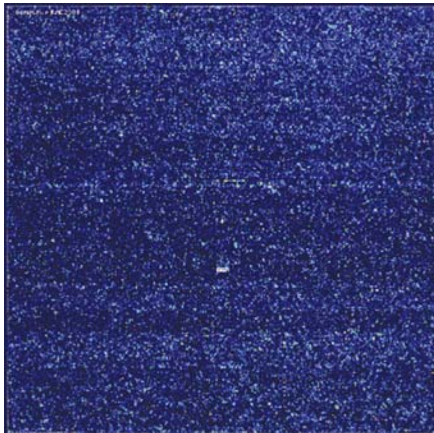
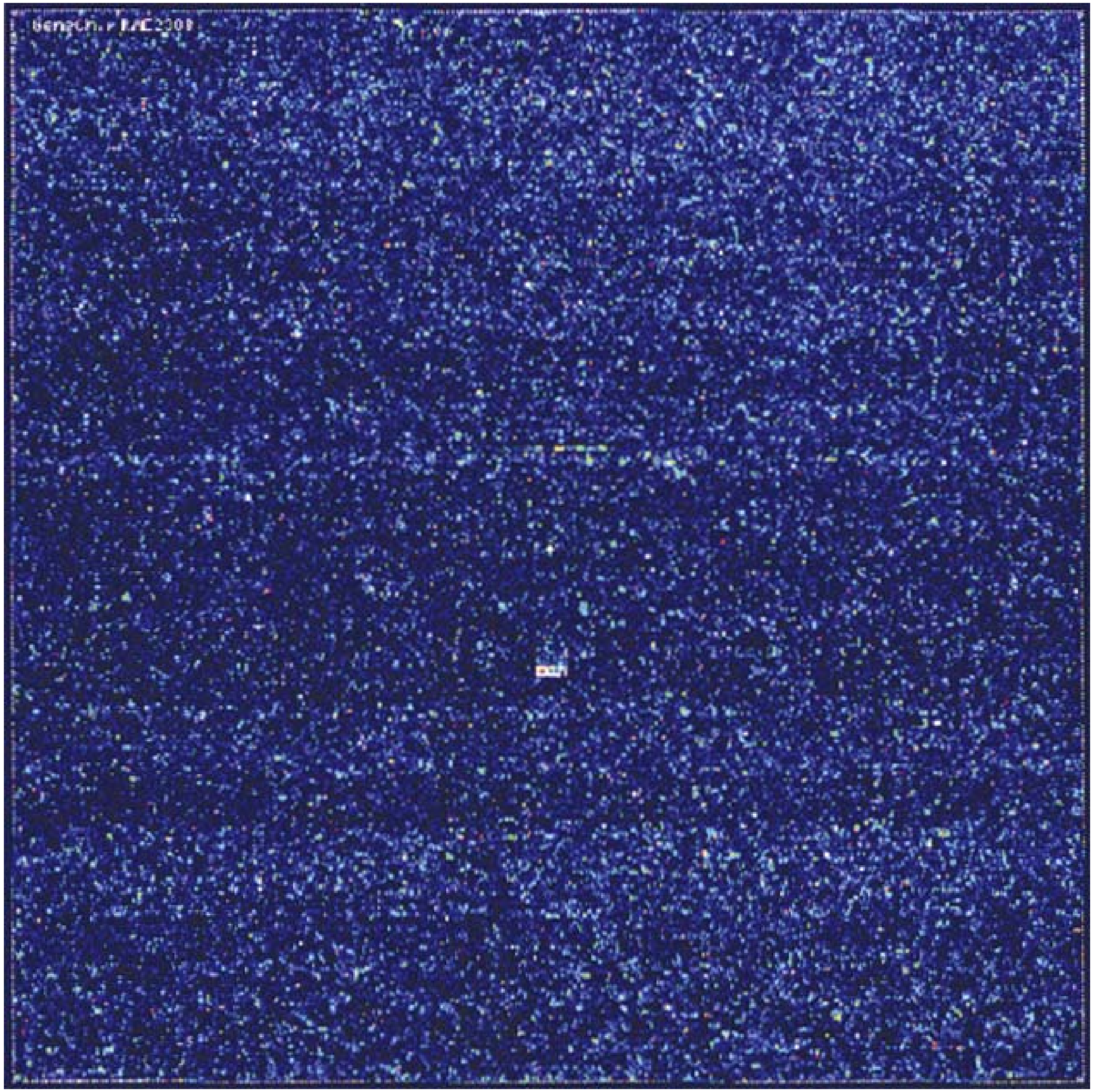
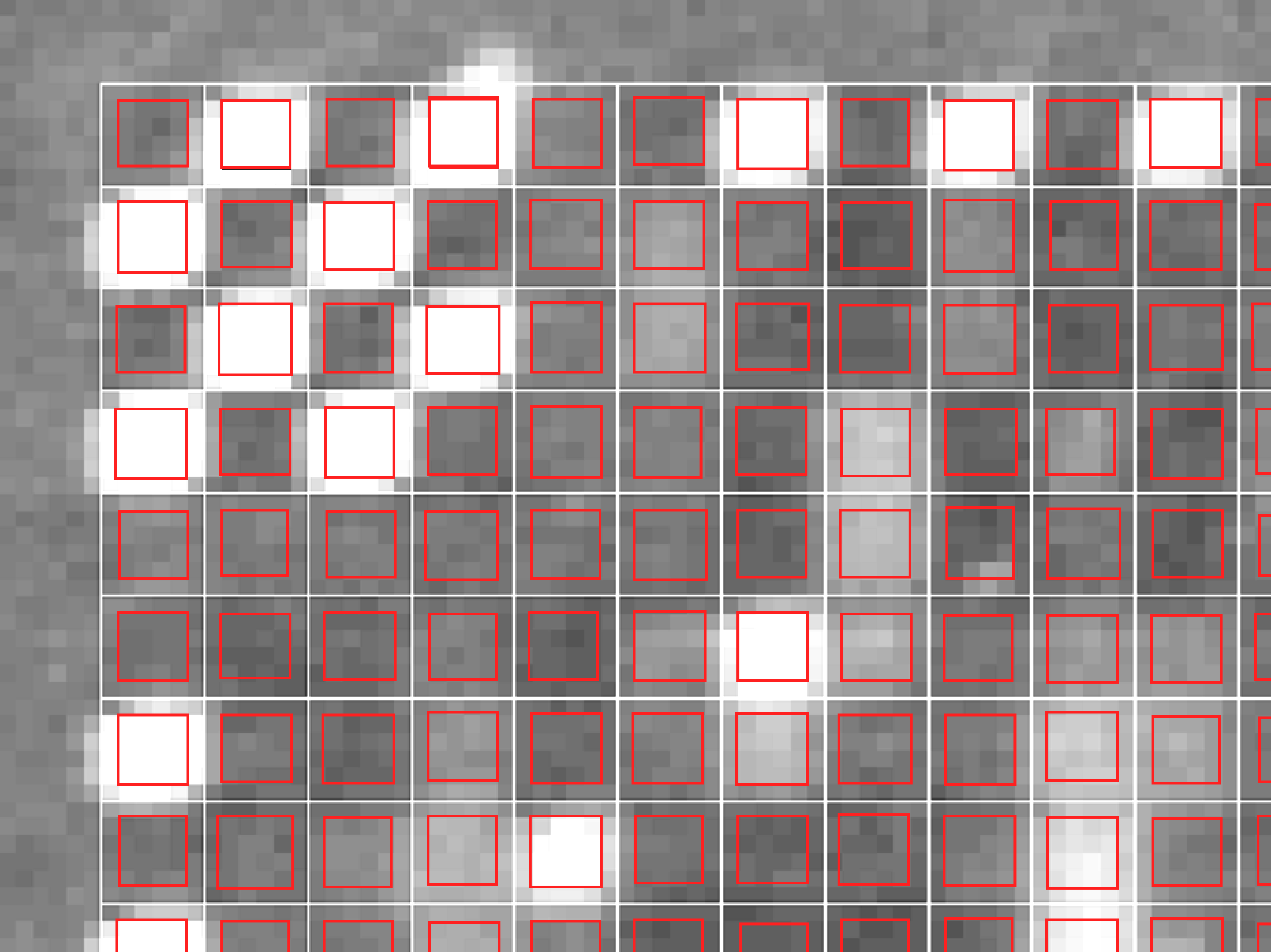


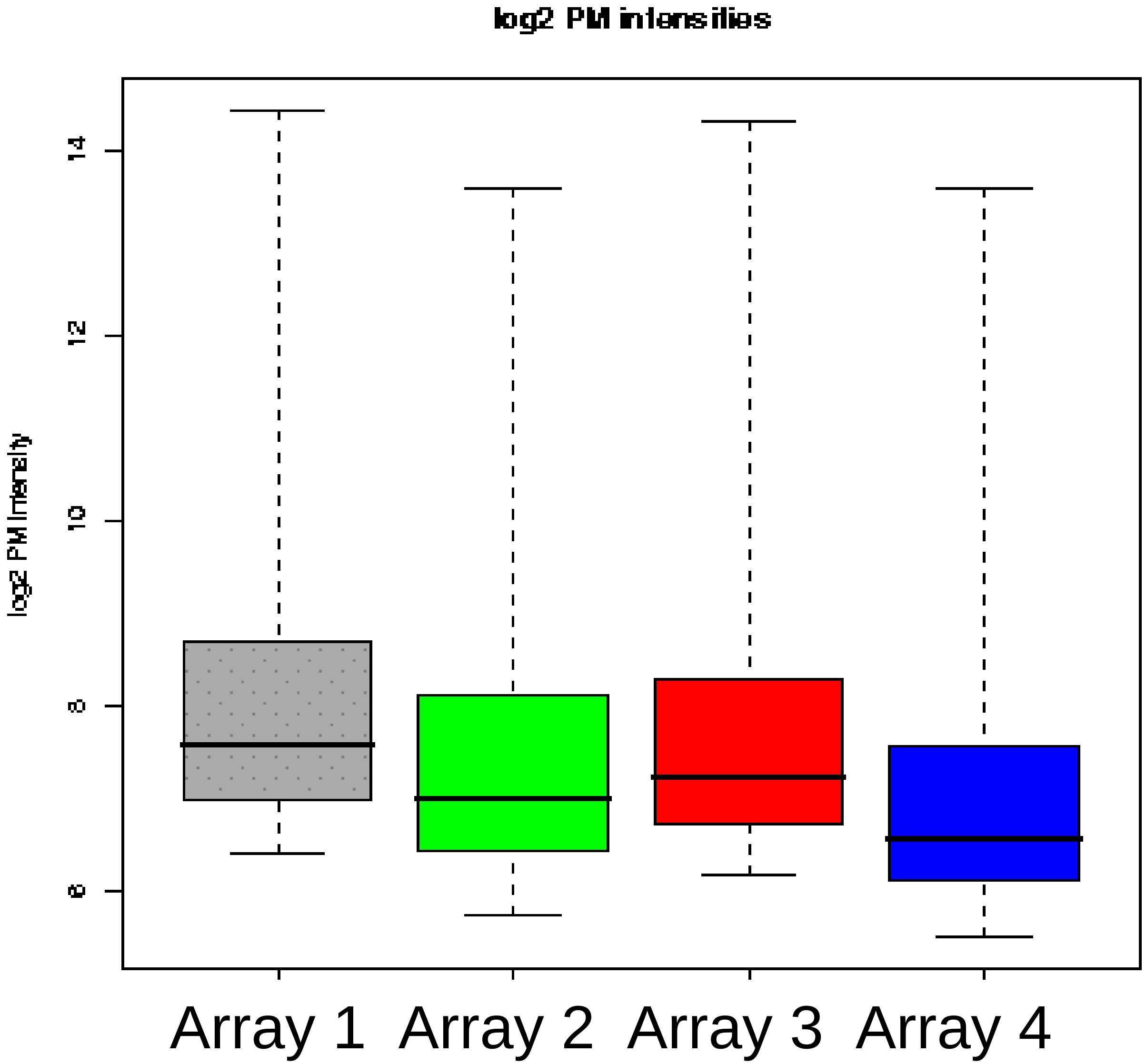
Image Analysis



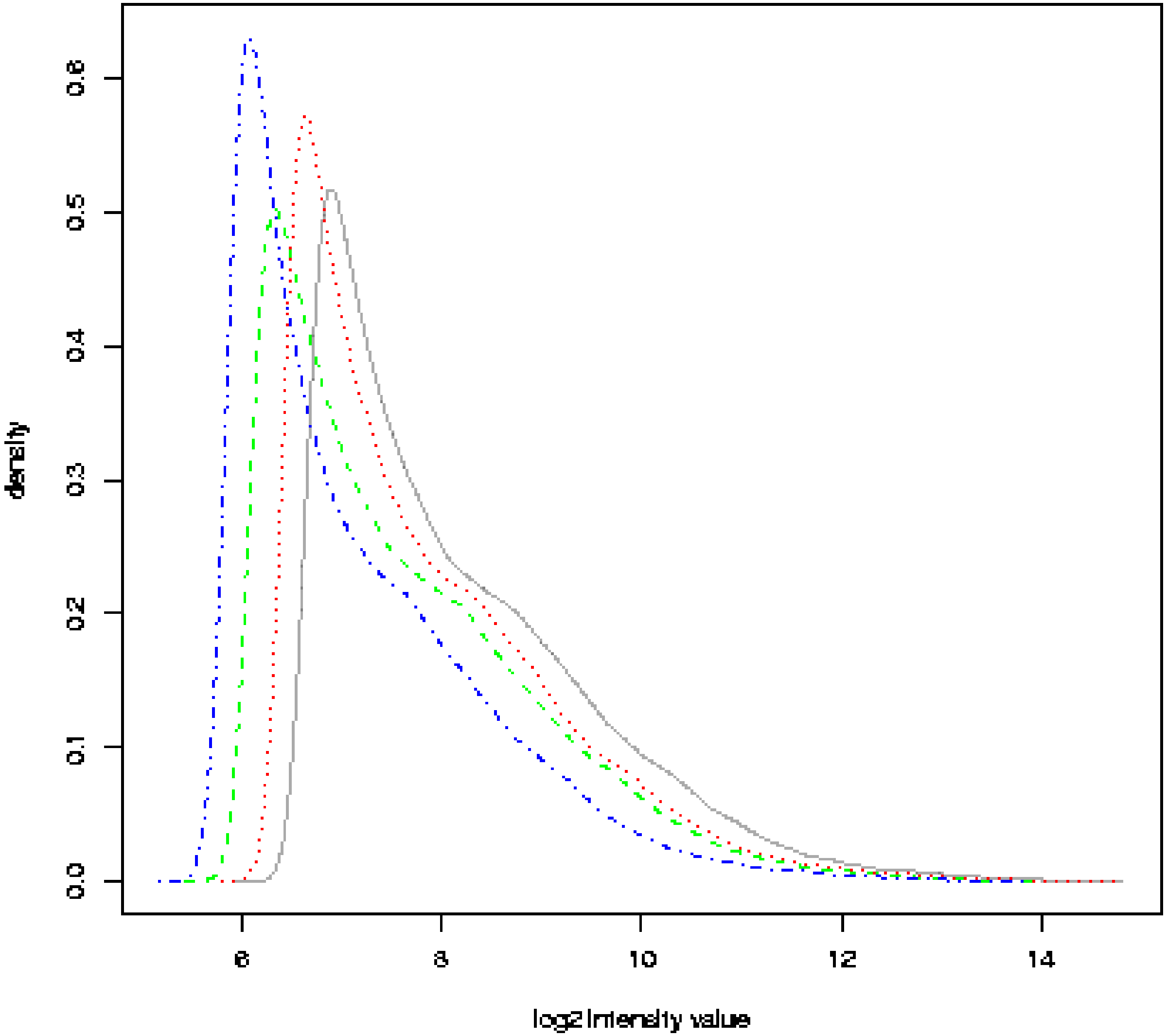


Courtesy: F. Colin

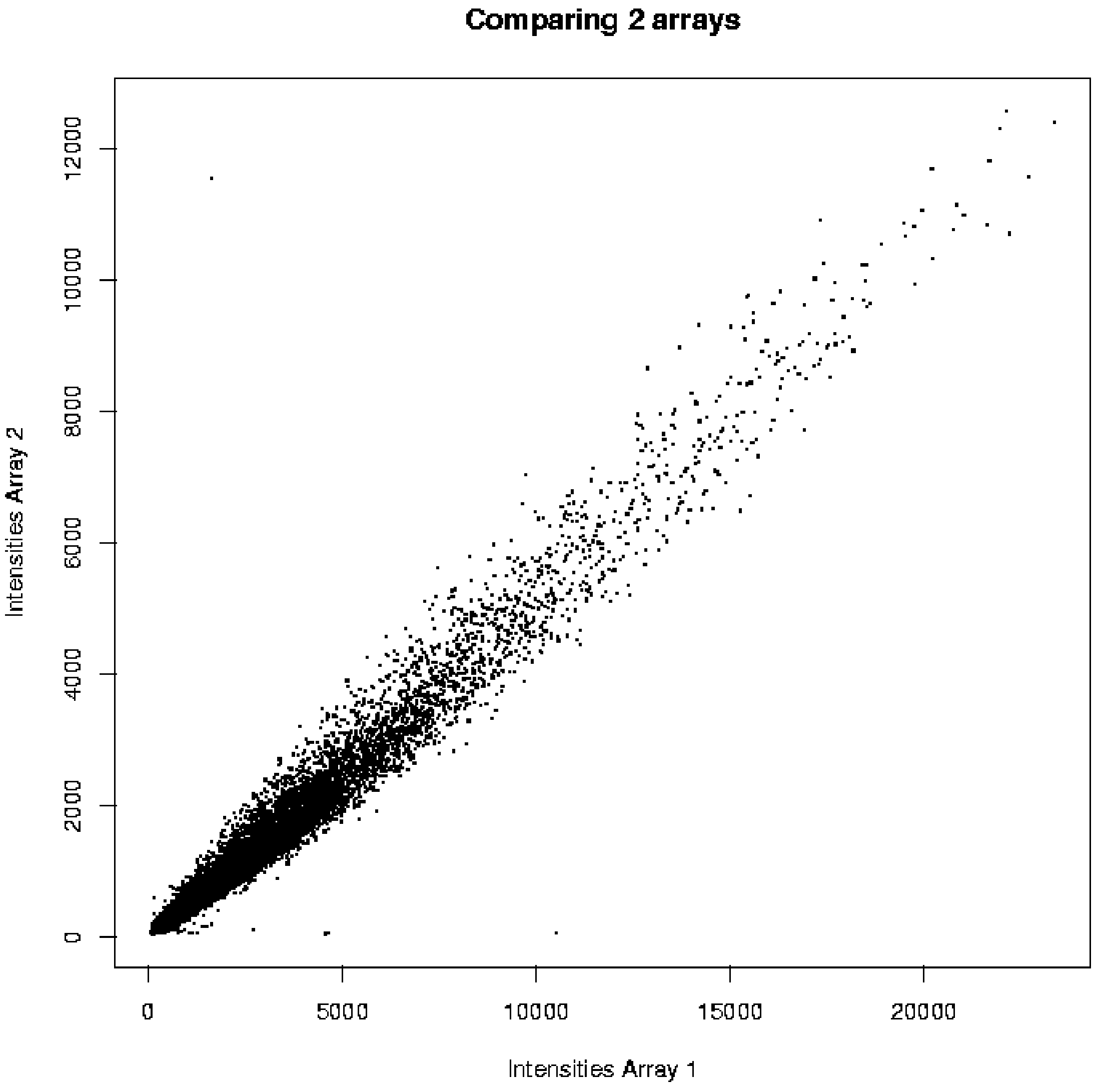
Boxplot raw intensities



Density plots



Comparing arrays



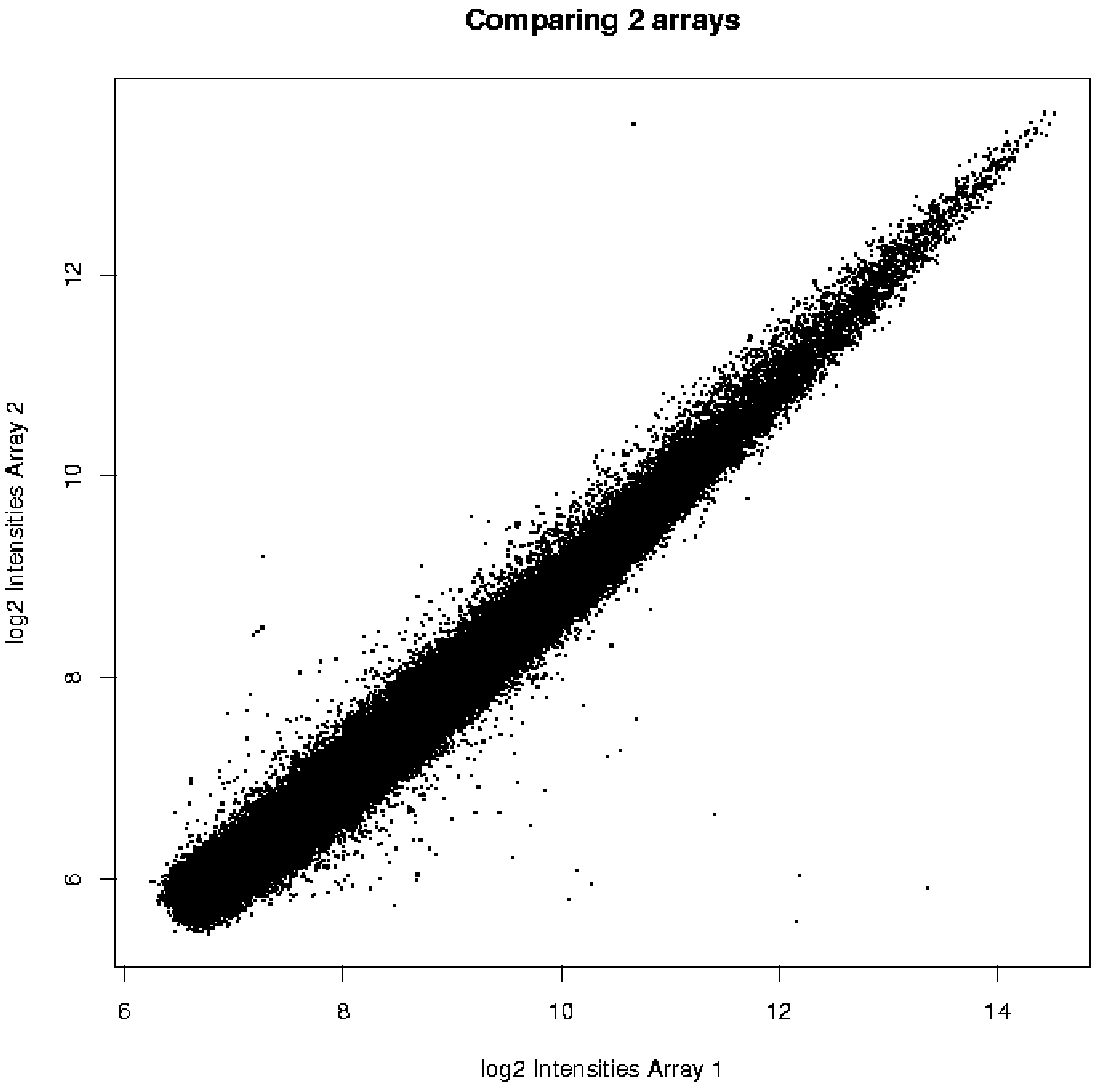
Array2

vs

Array 1

Bad

Comparing arrays



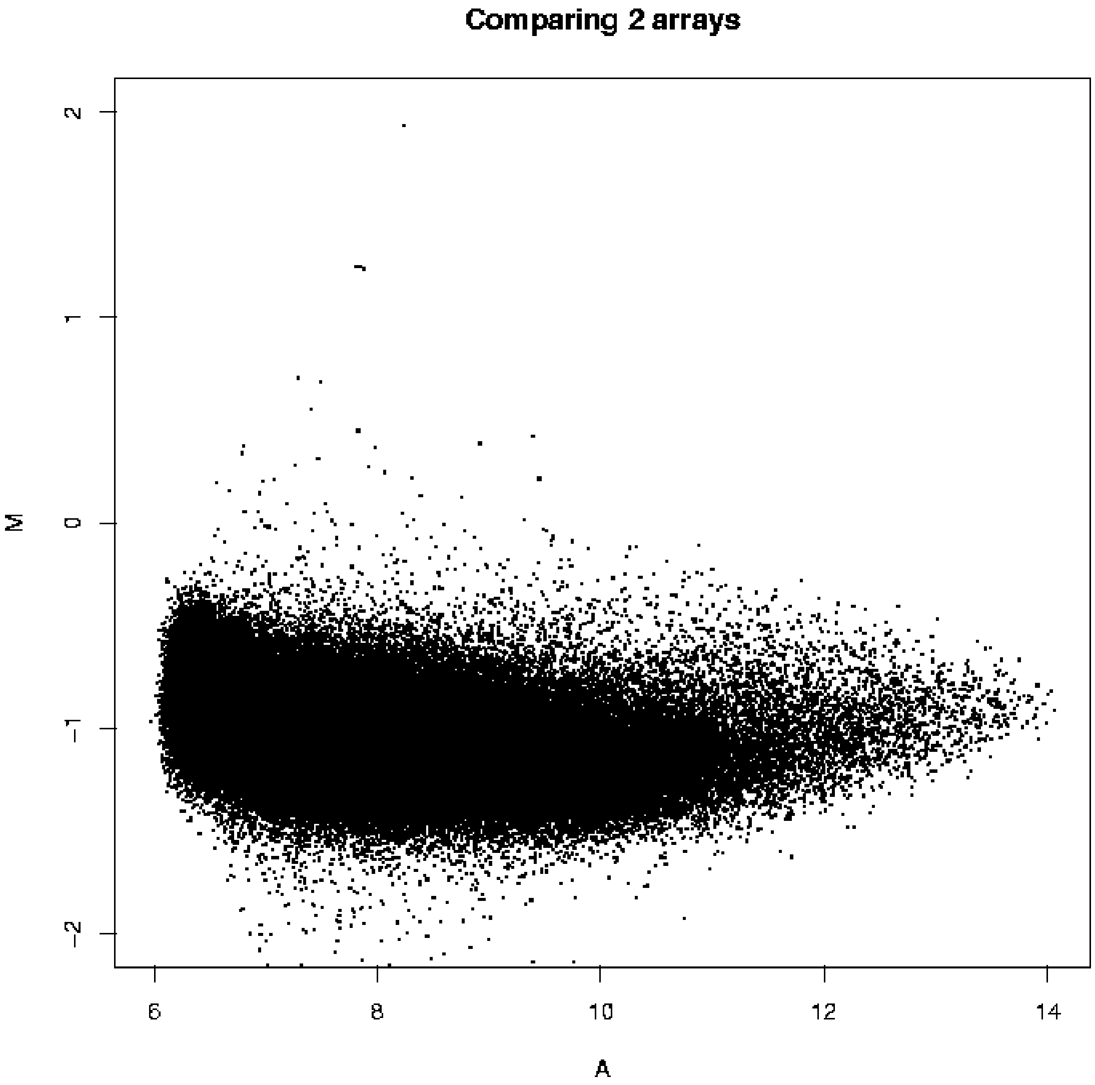
Log2 Array2

vs

Log2 Array 1

Better

Comparing arrays



M = log2(Array2/Array1)

Vs

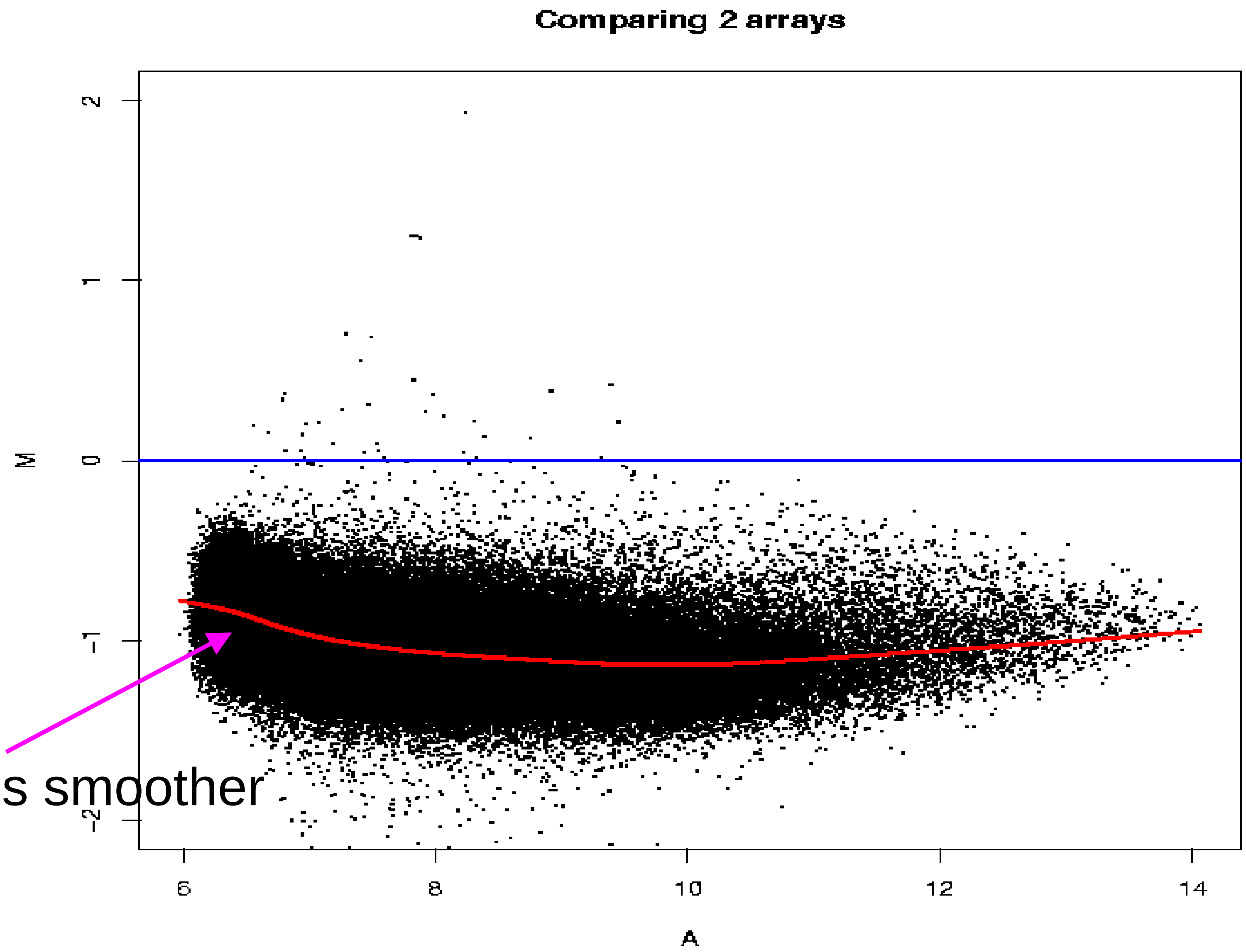
A = ½ log2(Array2*Array1)

Best

M= Minus

A=Average

Typical MA-plot

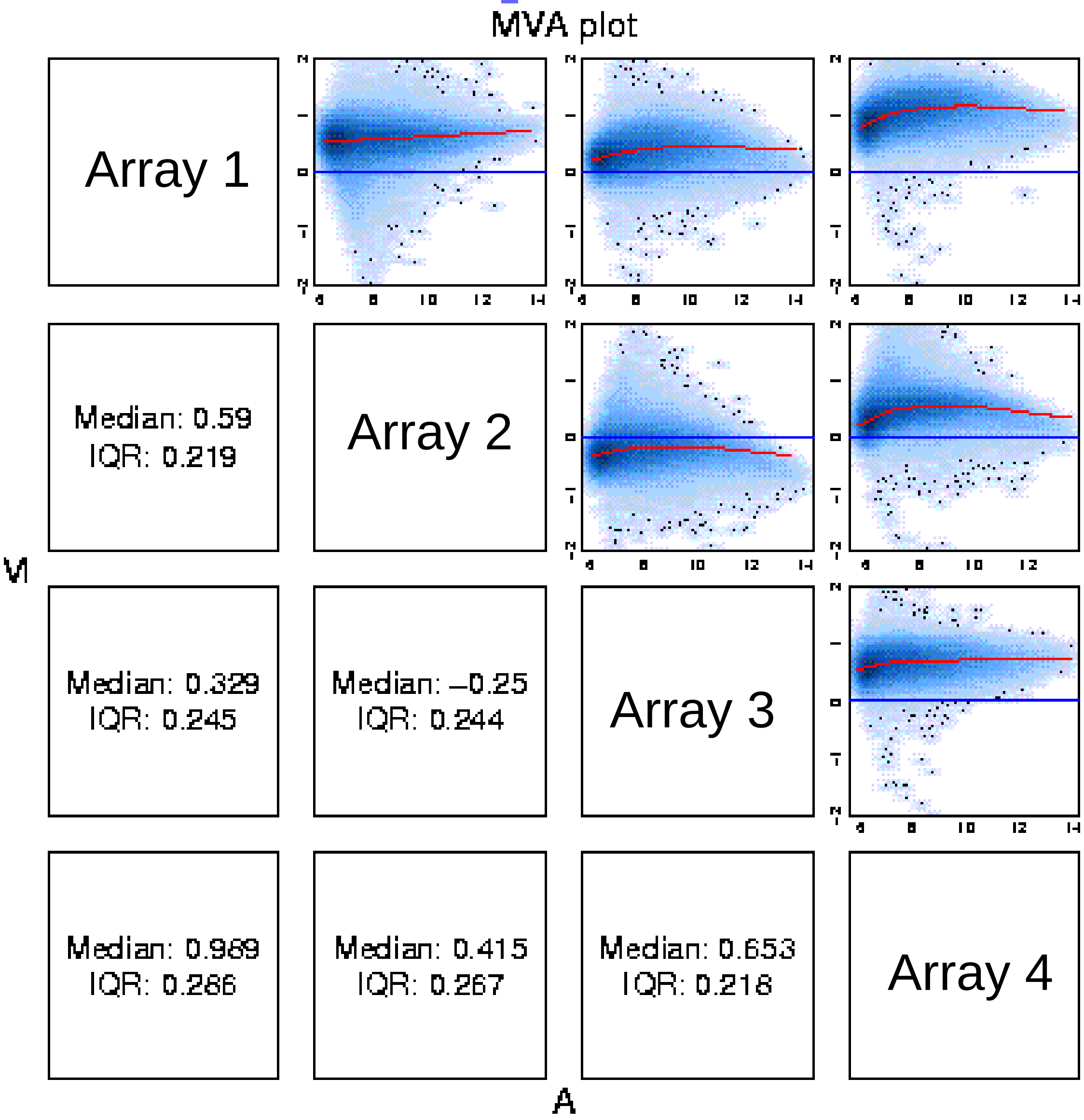


Loess smoother

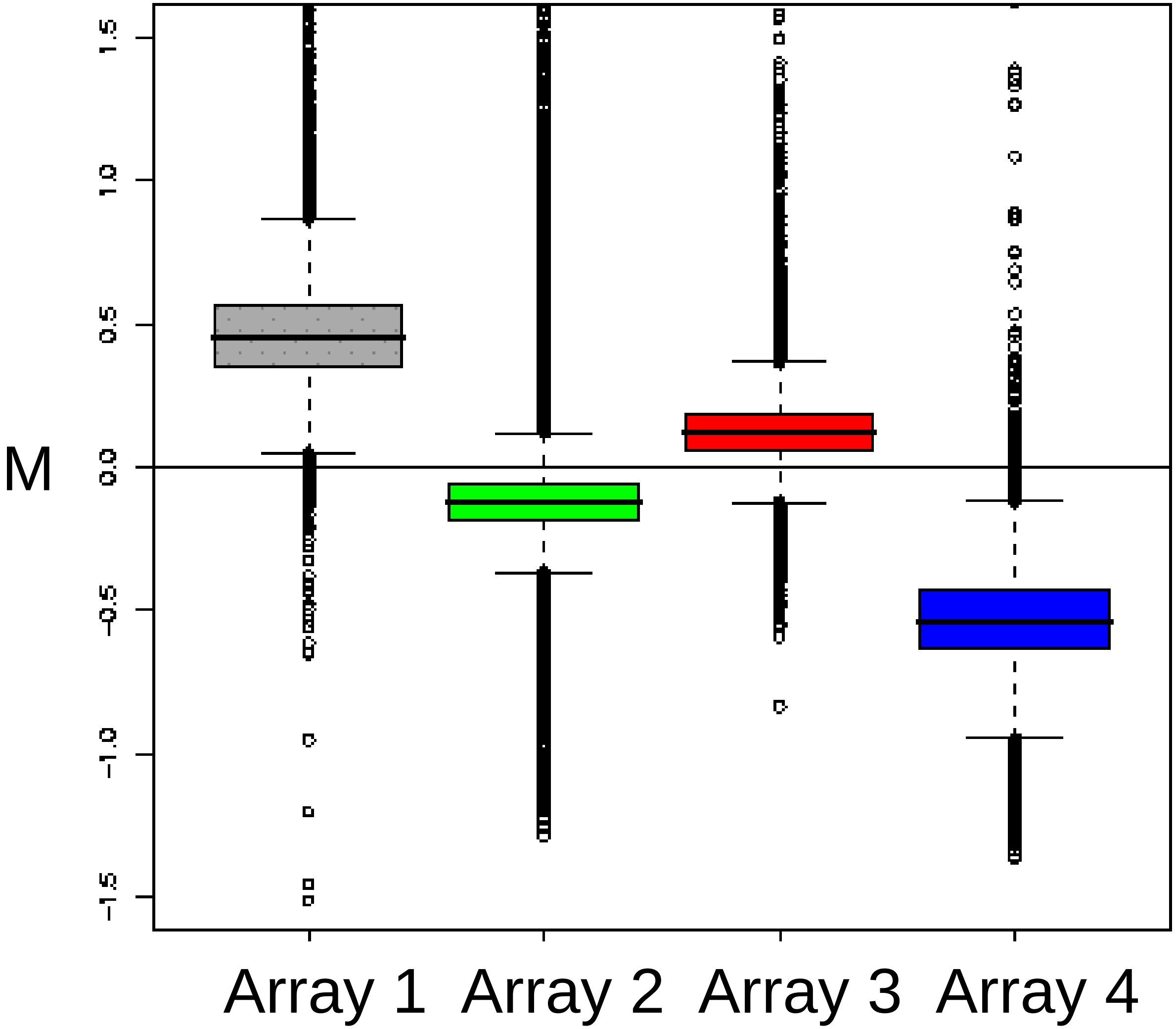
Pairwise MA plots

$M = \log_2 array_i / array_j$

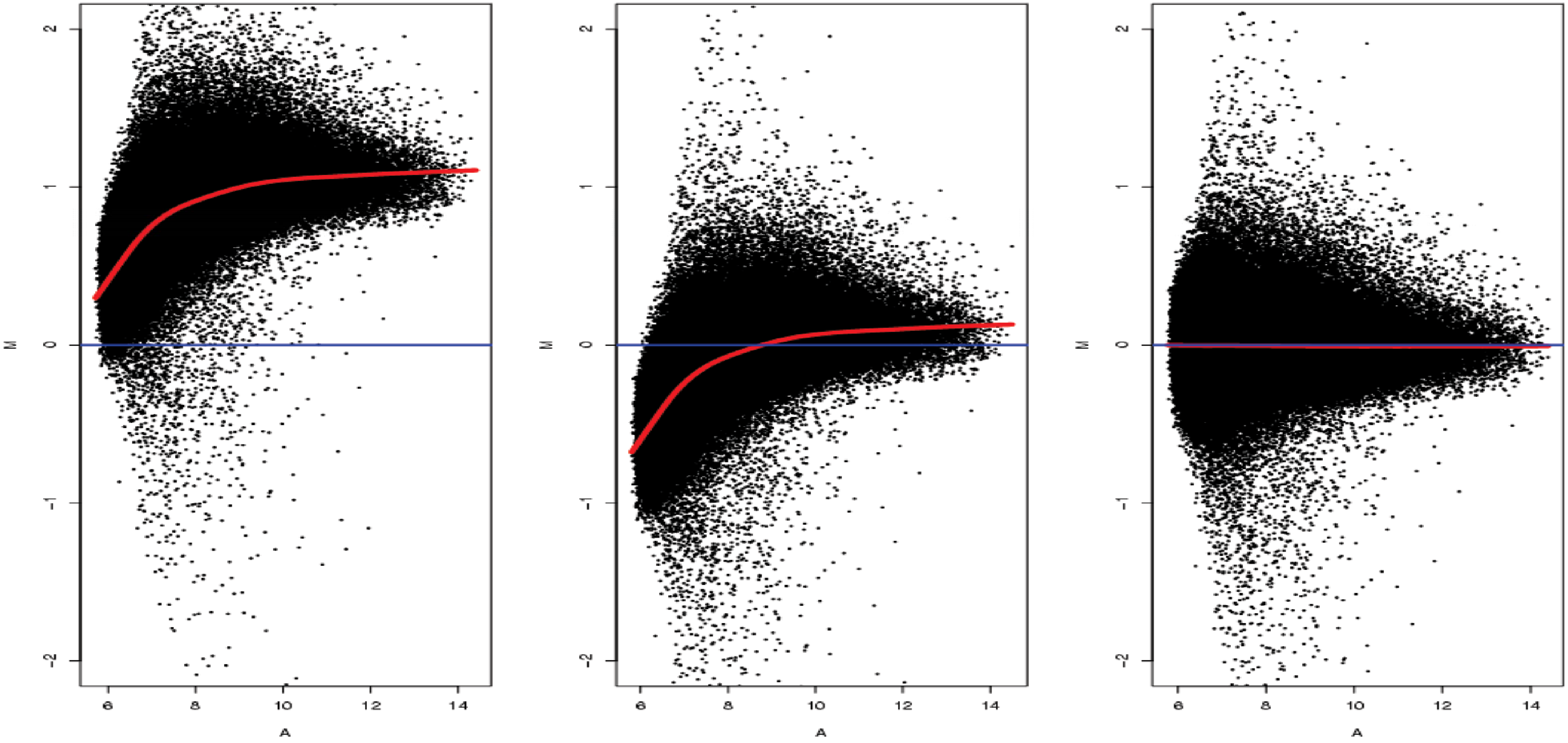
$A = 1/2 * \log_2(array_i * array_j)$



Boxplots comparing M



It works!!



Unnormalized

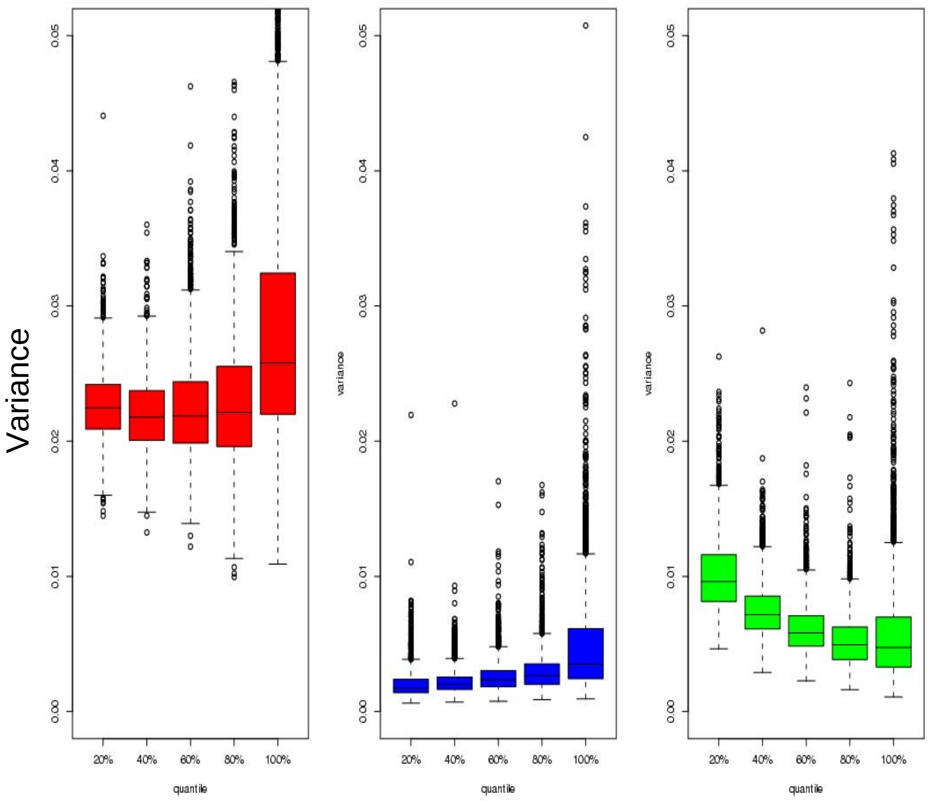
Scaling

Quantile Normalization

This is probe intensity data for two chips hybridized using same sample pool but scan on different scanners.

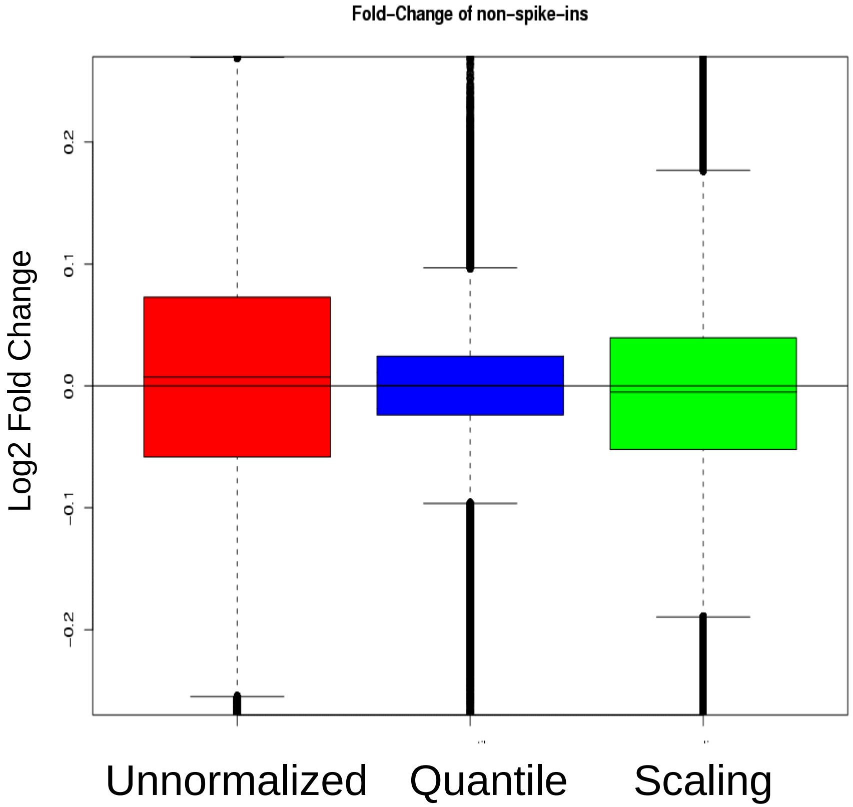
It Reduces Variability

Expression Values



Unnormalized Quantile Scaling

Fold change for
Non differential genes



Also no serious bias effects. For more see Bolstad et al (2003)

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 Popular 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 (Affymetrix)

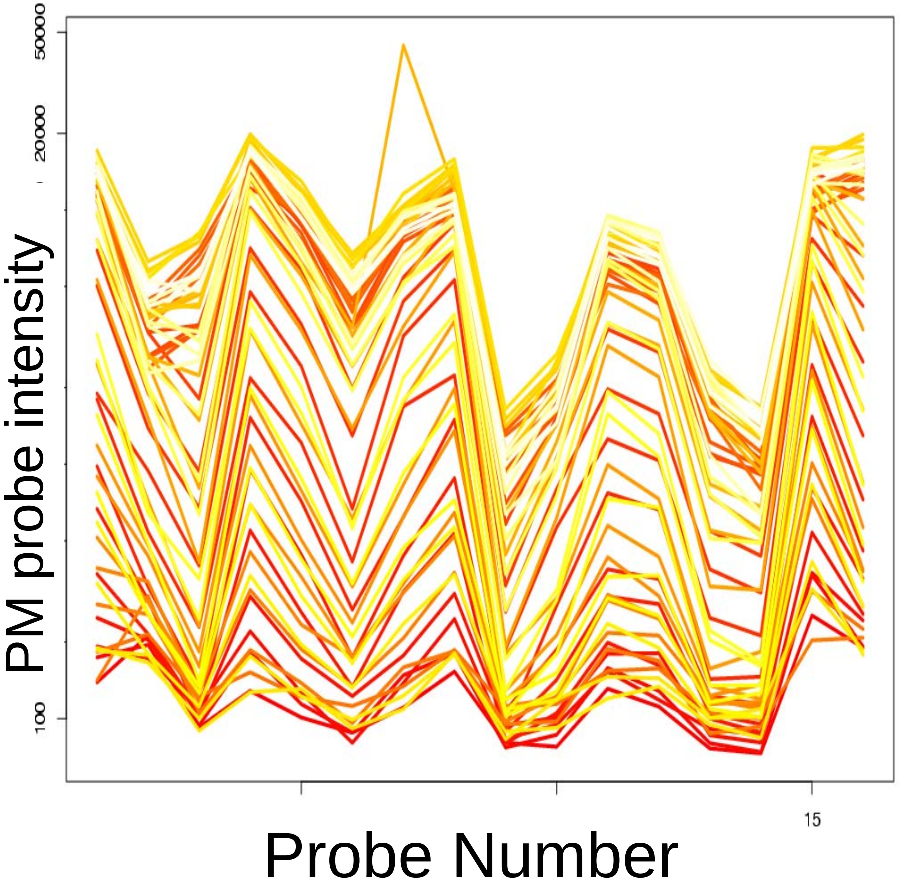
General Probe Level Model

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

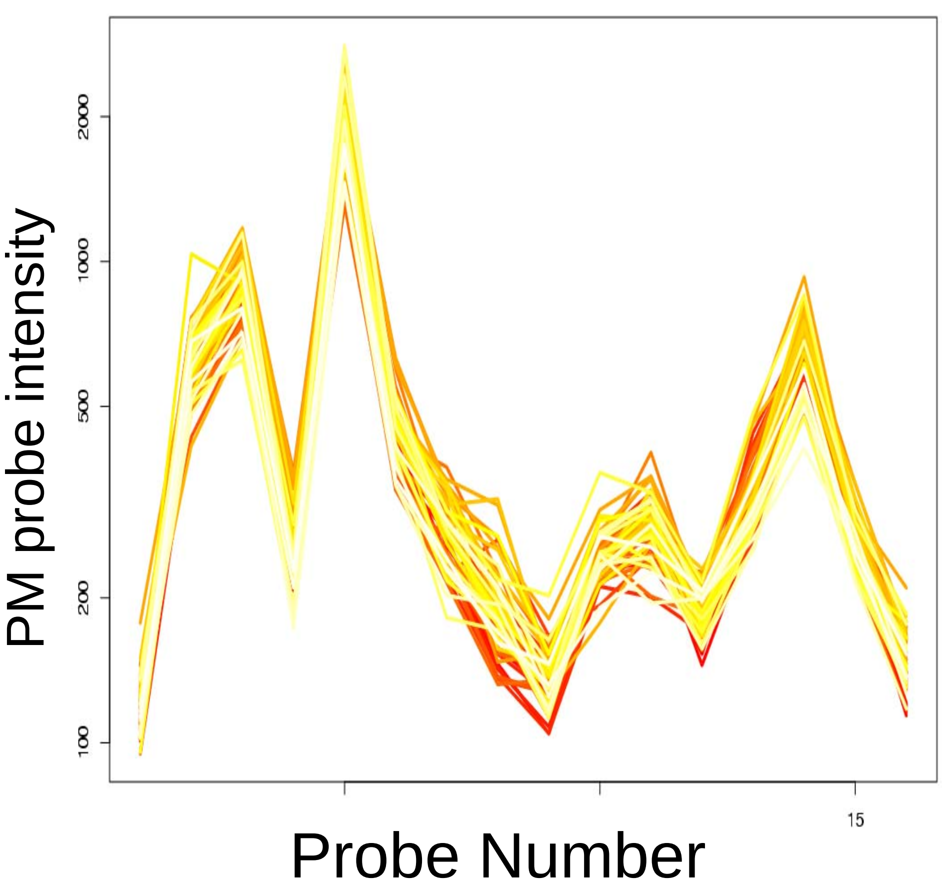
- Where f(X) is function of factor (and possibly covariate) variables (our interest will be in linear functions)
- y_{kij} is a pre-processed probe intensity (usually log scale)
- Assume that $\mathrm{Var}\left[\varepsilon_{kij}\right] = \sigma_k{}^2$

Probe Pattern Suggests Including Probe-Effect

Differentially expressing



Non Differential



Variance Covariance Estimates

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

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

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

We will use this form

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

Normalization

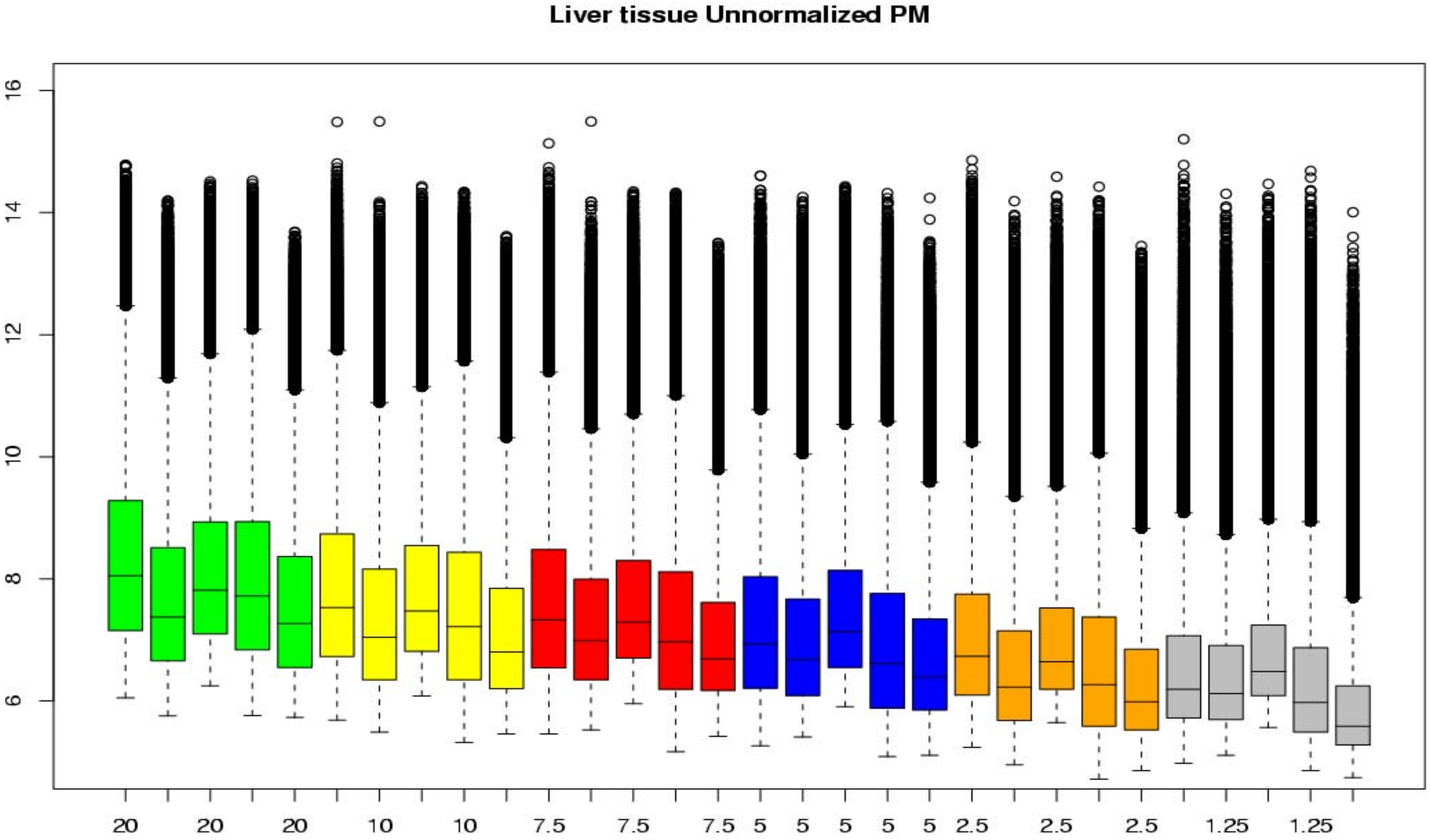
“Non-biological factors can contribute to the variability of data ... In order to reliably compare data from multiple probe arrays, differences of non-biological origin must be minimized.“[1]

- Normalization is the process of reducing unwanted variation either within or between arrays. It may use information from multiple chips.
- Typical assumptions of most major normalization methods are (one or both of the following):
  - Only a minority of genes are expected to be differentially expressed between conditions
  - Any differential expression is as likely to be up-regulation as down-regulation (ie about as many genes going up in expression as are going down between conditions)

[1] GeneChip 3.1 Expression Analysis Algorithm Tutorial, Affymetrix technical support

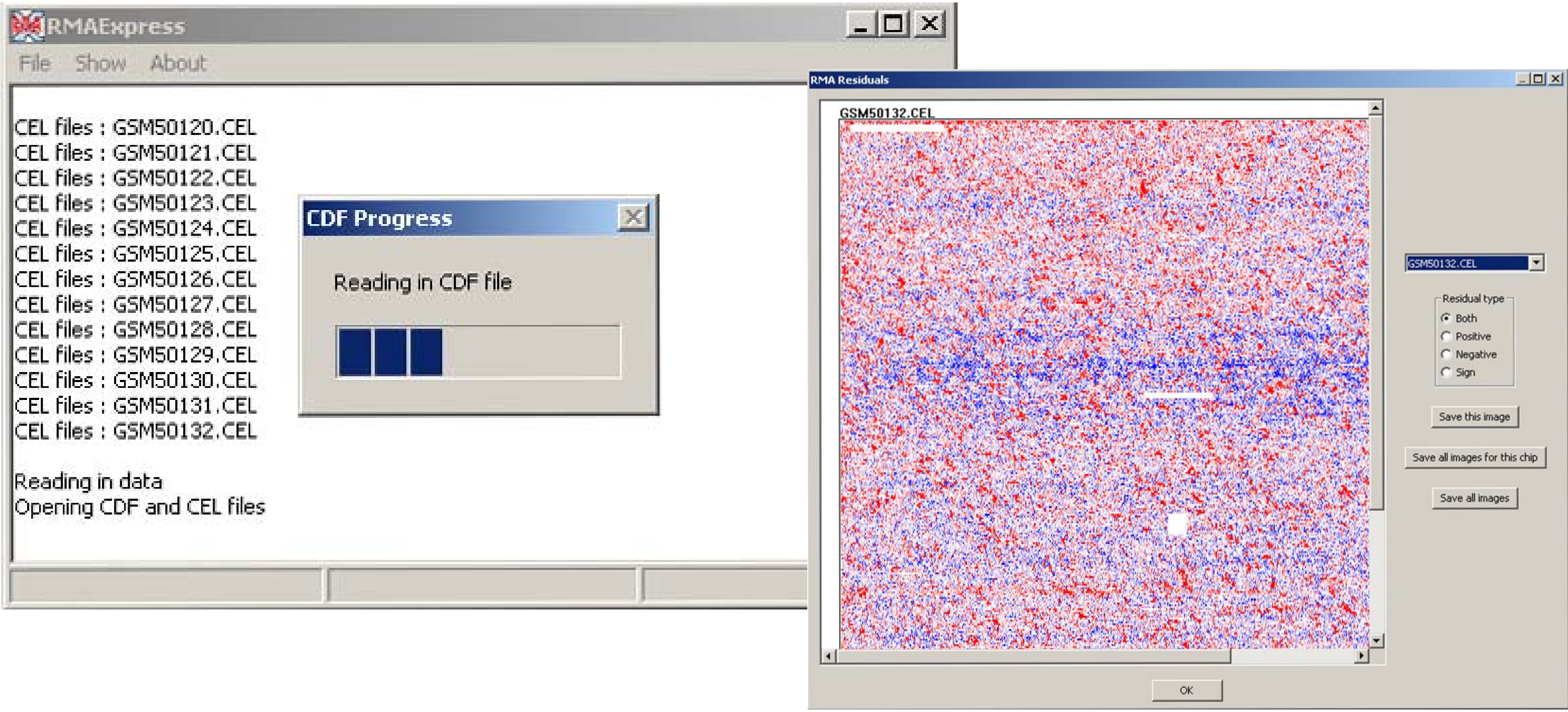
Non-Biological variability is a problem

Log2 PM intensity



5 scanners for 6 dilution groups

RMAExpress



```
CEL files : GSM50120.CEL
CEL files : GSM50121.CEL
CEL files : GSM50122.CEL
CEL files : GSM50123.CEL
CEL files : GSM50124.CEL
CEL files : GSM50125.CEL
CEL files : GSM50126.CEL
CEL files : GSM50127.CEL
CEL files : GSM50128.CEL
CEL files : GSM50129.CEL
CEL files : GSM50130.CEL
CEL files : GSM50131.CEL
CEL files : GSM50132.CEL

Reading in data
Opening CDF and CEL files
```

CDF Progress

Reading in CDF file

Residual type: ● Both ○ Positive ○ Negative ○ Sign

- http://rmaexpress.bmbolstad.com
- Implemented in C++. Open source.
- Compiled builds supplied for Windows users. Source code for Unix users. Cross-platform