

# Low level analysis of Affymetrix Genechip data

Ben Bolstad

Division of Biostatistics, University of California, Berkeley

Australasian Biometrics/NZSA 2001

December 2001

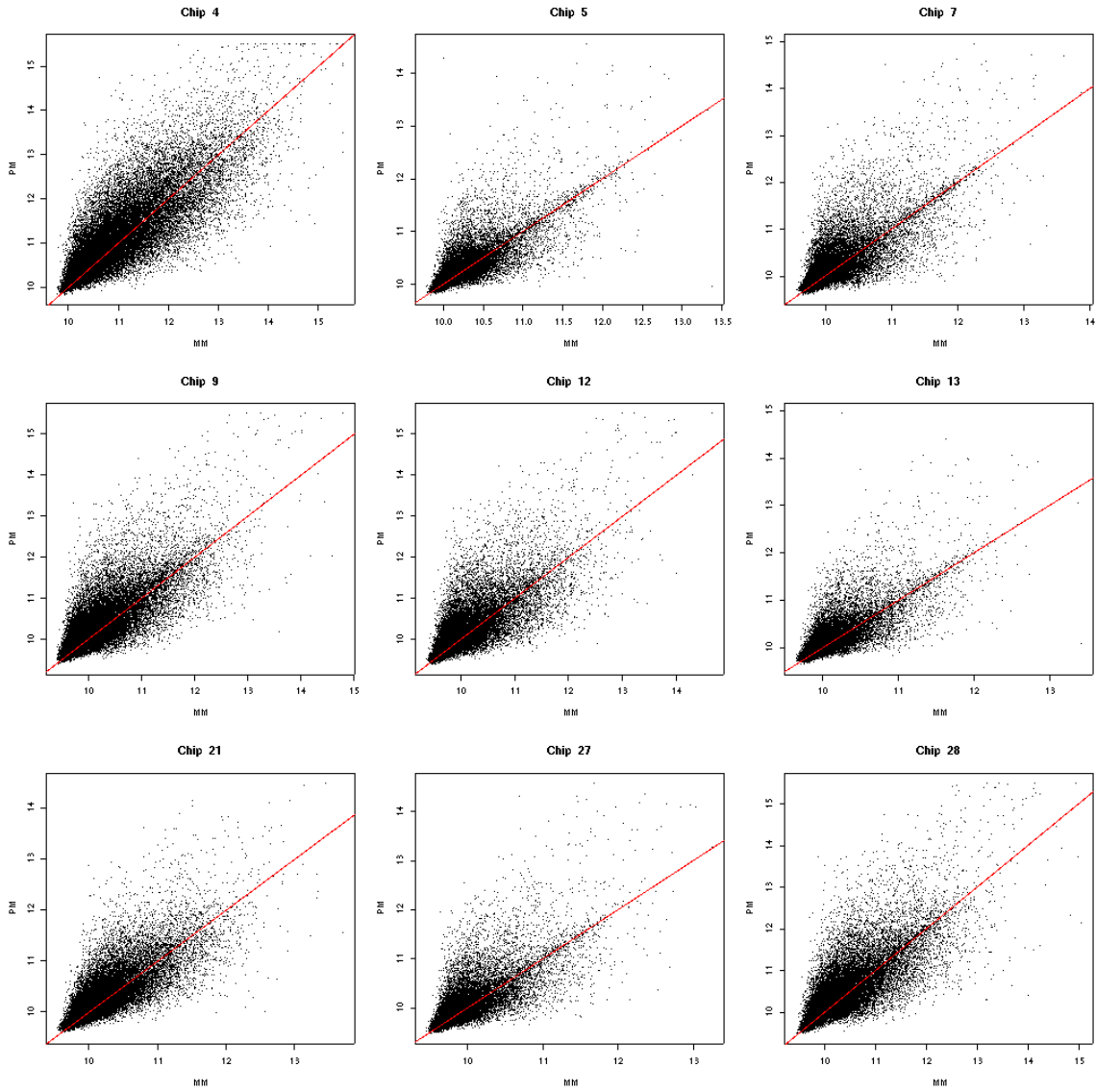
# Outline

1. Data exploration
2. Background
3. Normalization
4. Measures of expression
5. Quality issues

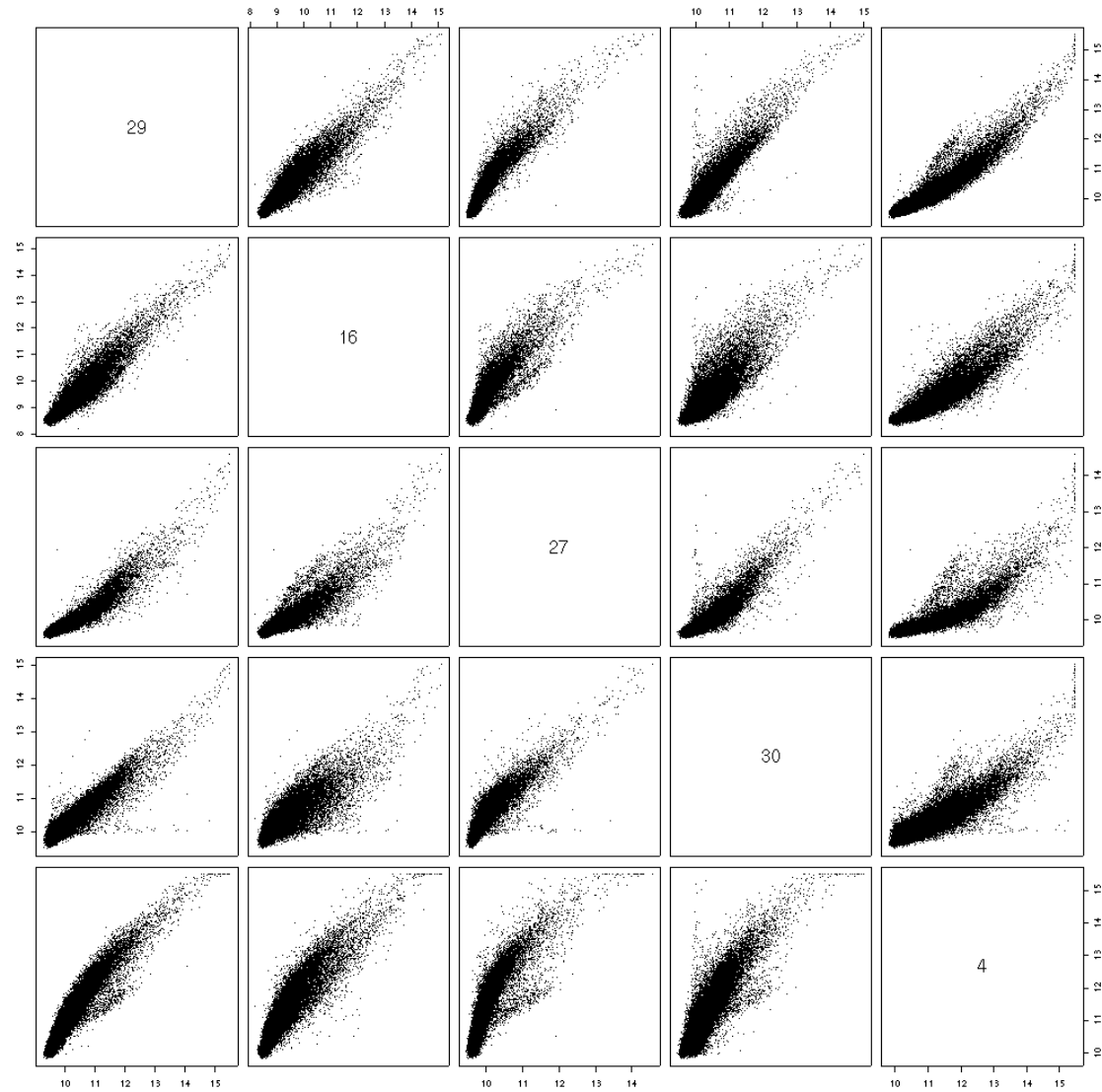
## Some data exploration

1. PM vs MM on individual chips.
2. PM vs PM, MM vs MM pairwise between chips
3. Distribution of intensities within/between chips
4. Some cel images

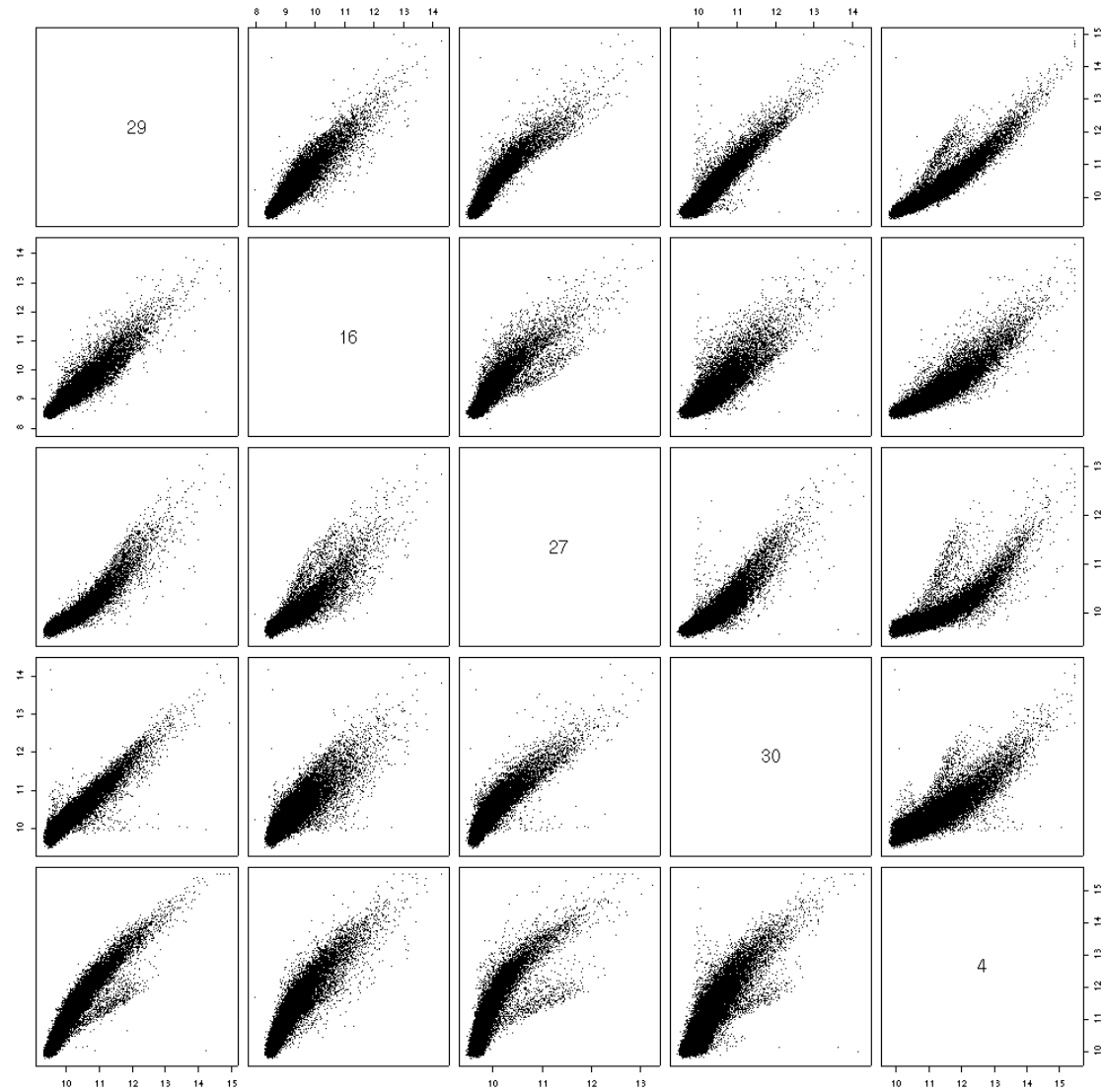
# PM vs MM



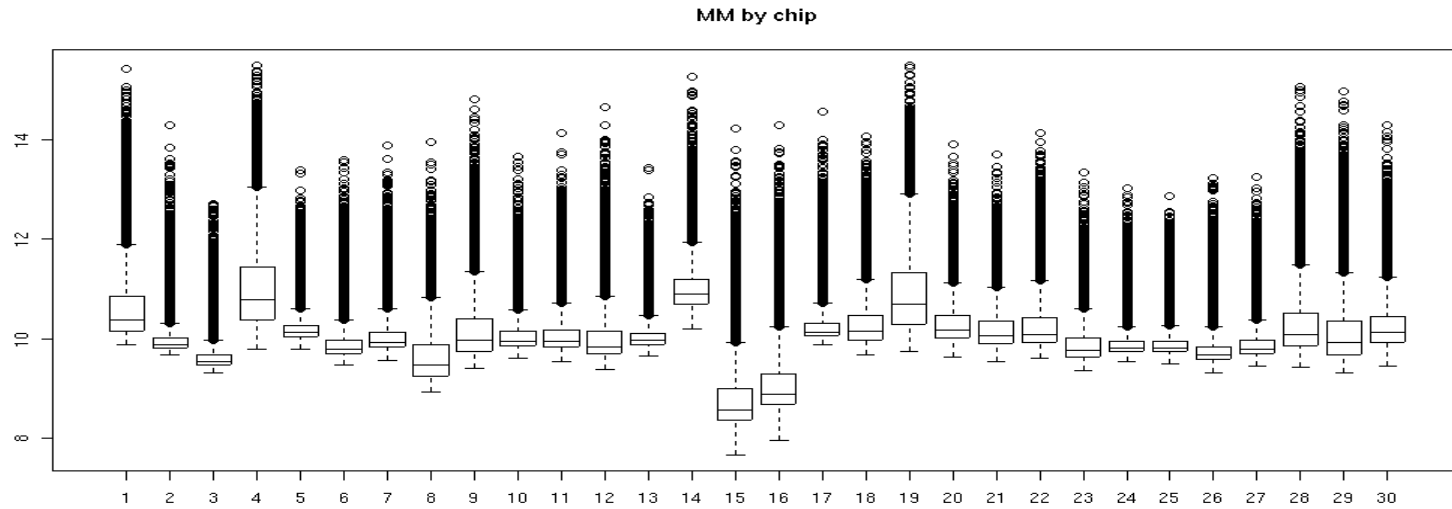
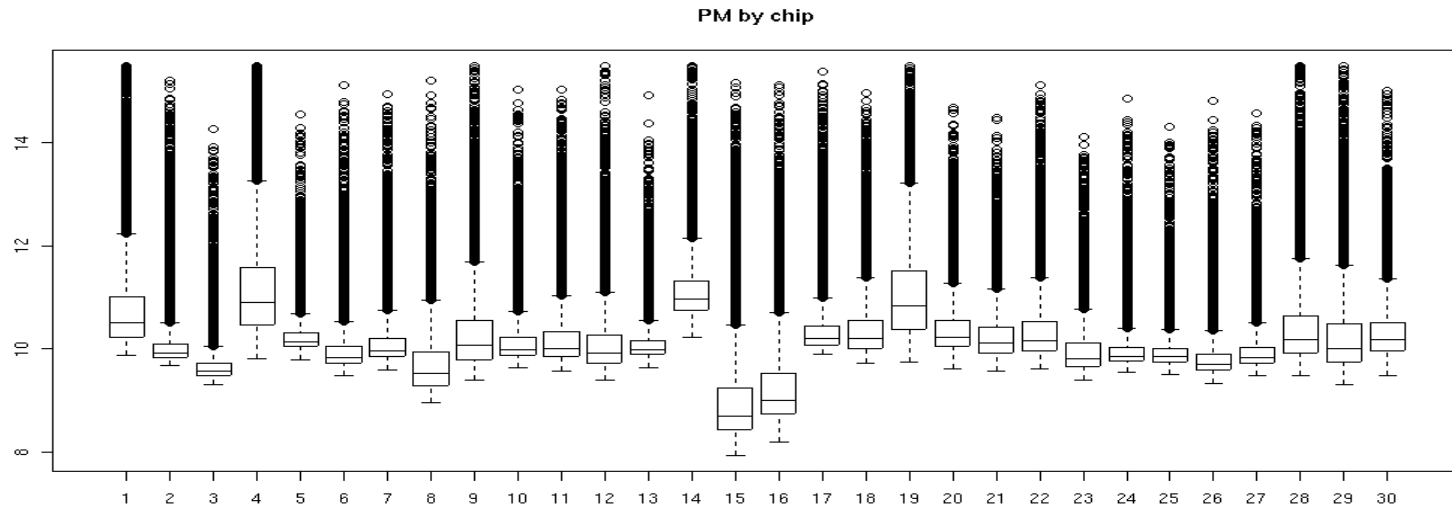
# PM vs PM



# MM vs MM

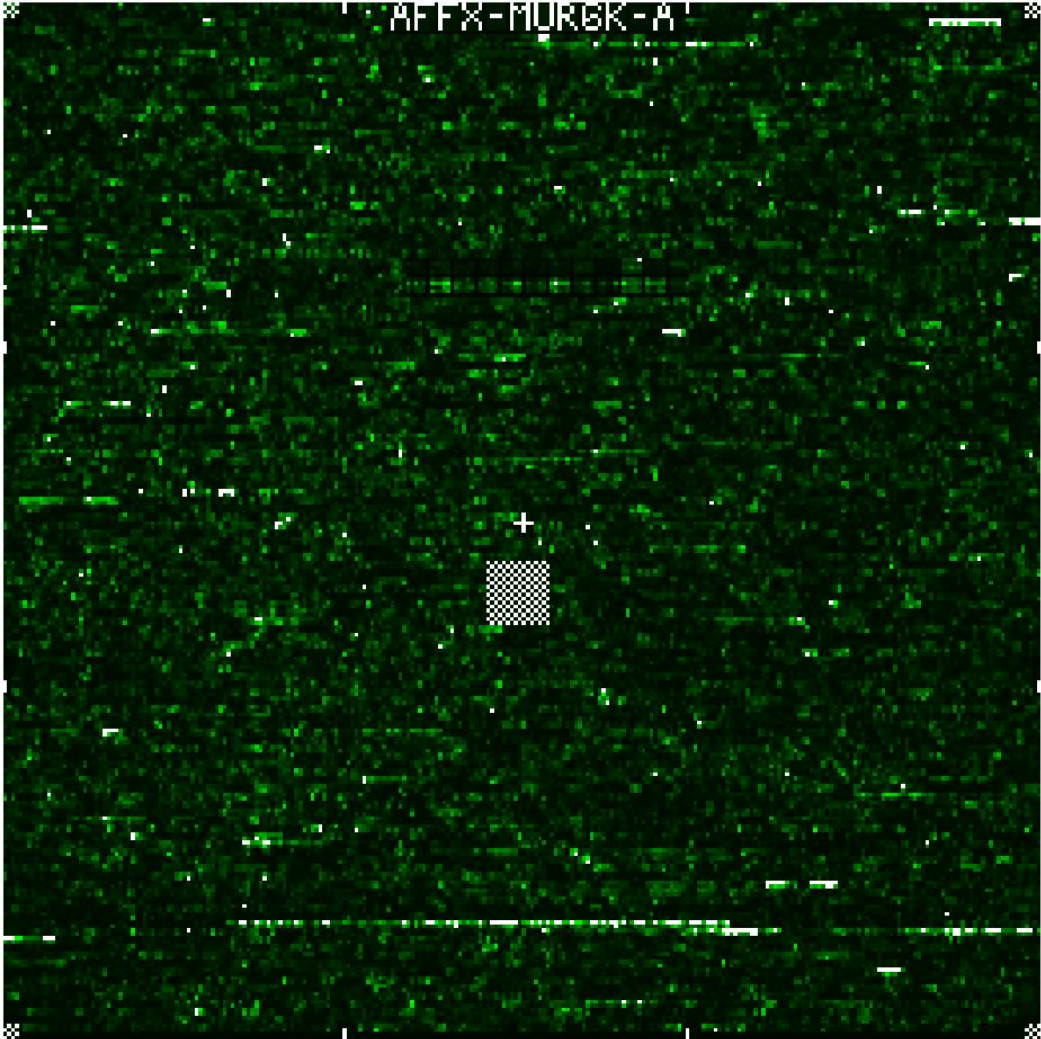


# Distribution of intensities across chips



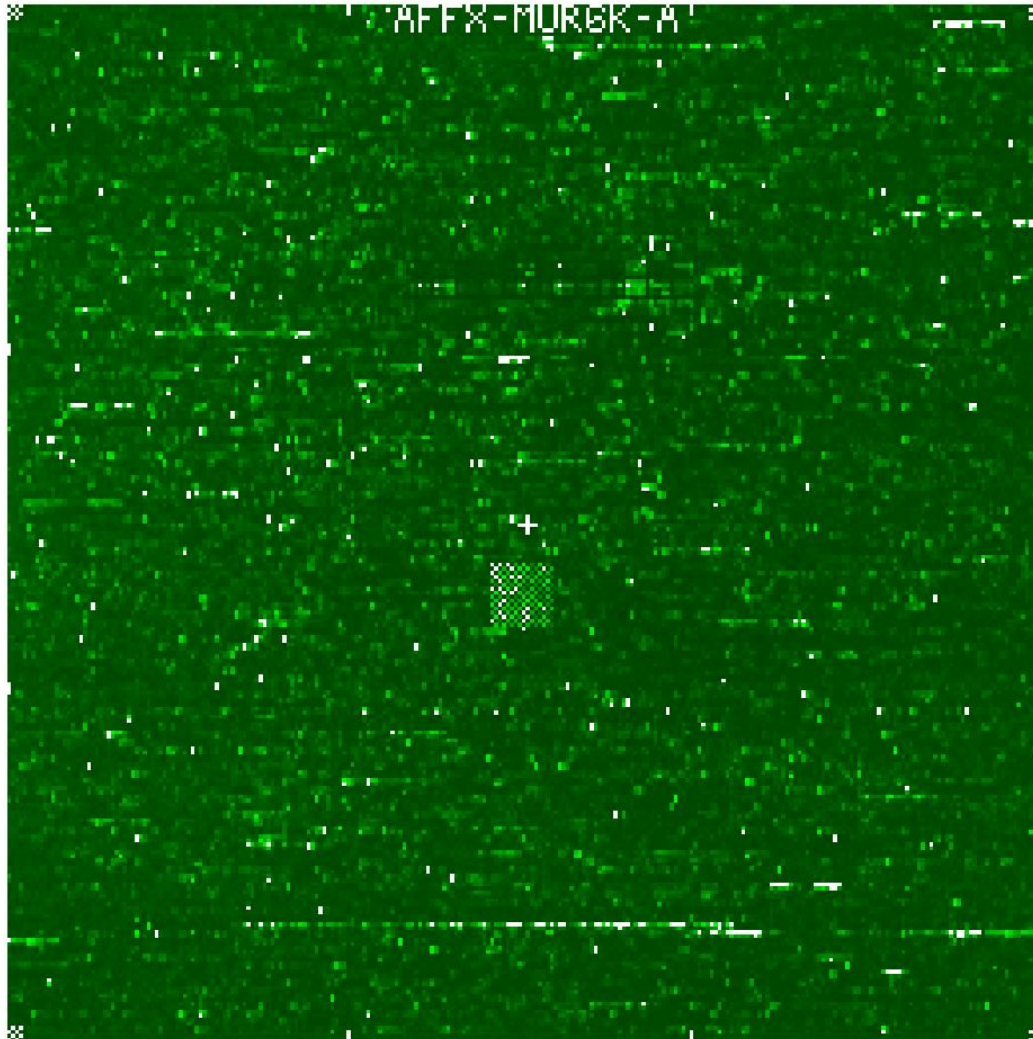
# CEL image

Plot of X04A.cel



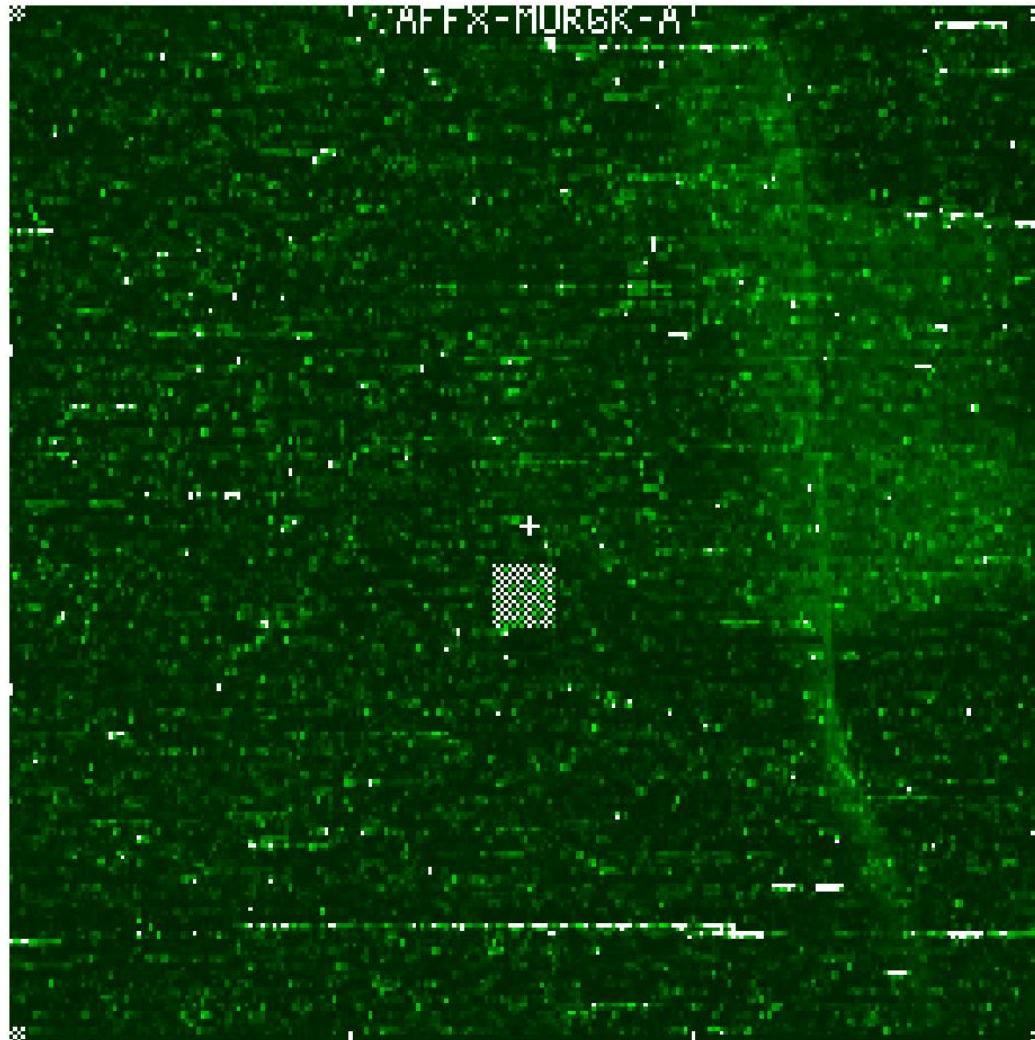
# Noiser CEL image

Plot of X05A.cel



# CEL image with some artifacts

Plot of X08A.cel



## Background

- A measurement of signal intensity caused by autofluorescence of the array surface and non specific binding.
- Since probes are so densely packed on chip must use probes themselves (rather than region adjacent to probes as in cDNA arrays) to calculate the background.
- In theory the MM should serve as a biological background correction for the PM.

## Background - Affymetrix

1. The array is divided into sectors (16 by default)
2. within each sector probes are ranked by intensity, the lowest 2% identified and average of these probes calculated. This value will be the background for the sector.
3. Background value is subtracted from all probes in that sector.

Source: GeneChip 3.1 Expression Analysis Algorithm Tutorial, Affymetrix technical support

## Background - Naef et al

In Naef et al they consider a subset of probes where the difference  $|PM - MM| < \epsilon$  as representative of the background. Naef et al suggest  $\epsilon = 50$  but state that using  $\epsilon = 100$  makes little difference in the background estimates, the claim is that background is visible mostly in the low intensities and that the distribution of these probes is formed by the conjunction of a normal and a “step” function. By fitting gaussians, they estimate a mean background (and variance). Naef et al use only PM in their calculation of an expression measure.

Felix Naef, Daniel A. Lim, Nila Patil, Marcelo O. Magnasco , "From features to expression: High-density oligonucleotide array analysis revisited", LANL e-print physics/0102010

## 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.”

Source: GeneChip 3.1 Expression Analysis Algorithm Tutorial, Affymetrix technical support

## Normalization - Affy

Use a global normalization (or scaling). Procedure is to choose a baseline array and use its average intensity or pick a target intensity. Then each chip is multiplied by a factor to give all chips same average intensity. Average intensity is calculated by averaging with exclusions on highest 2% and lowest 2% of values.

For example to normalize chip<sub>1</sub> against chip<sub>2</sub> the normalization factor is given by

$$\hat{\beta} = \frac{\sum_{chip2} (PM - MM)}{\sum_{chip1} (PM - MM)}$$

and thus the normalization for chip 1 is given by

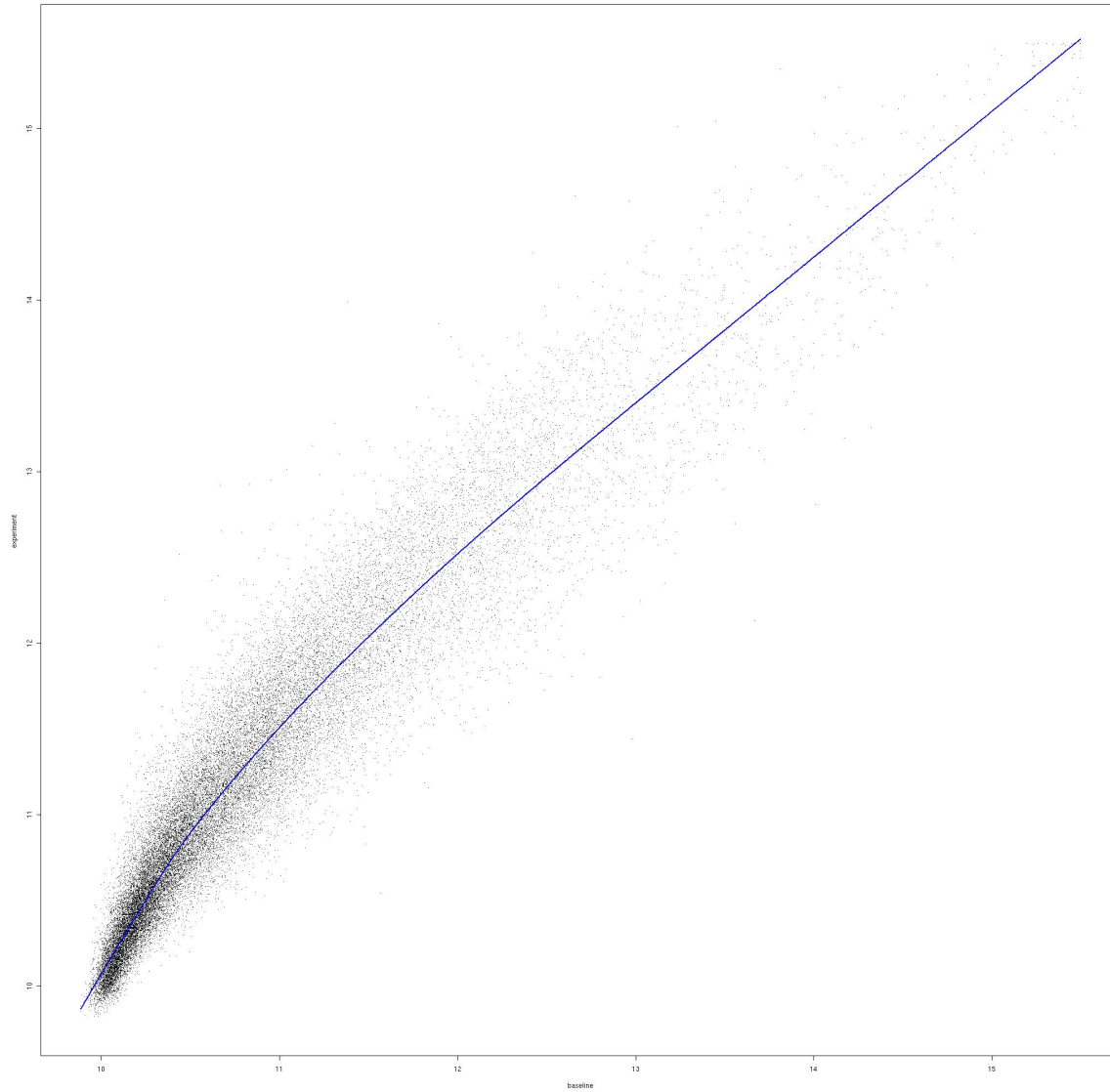
$$(PM - MM)_{new} = \hat{\beta} (PM - MM)_{old}$$

## Normalization - Schadt et al

Fit a non linear normalization relationship between a baseline and other chips using invariant difference selection algorithm. A set of probes is said to be invariant if ordering of probes in one chip is same in other set. Using their method they pick the invariant set of genes and then fit the non linear relation using cross validated smoothing splines. In a set of chips they choose the array having median intensity as the baseline and normalize all the other chips to this chip.

Feature Extraction and Normalization Algorithm for High Density Oligonucleotide Gene Expression Array Data. E. Schadt et al. UCLA preprint 304

# Schadt et al, cont



## Normalization - Irizarry and Speed

Method is based on  $M$  versus  $A$  plots. Where

$$M = \log_2 \left( PM_i / PM_j \right)$$

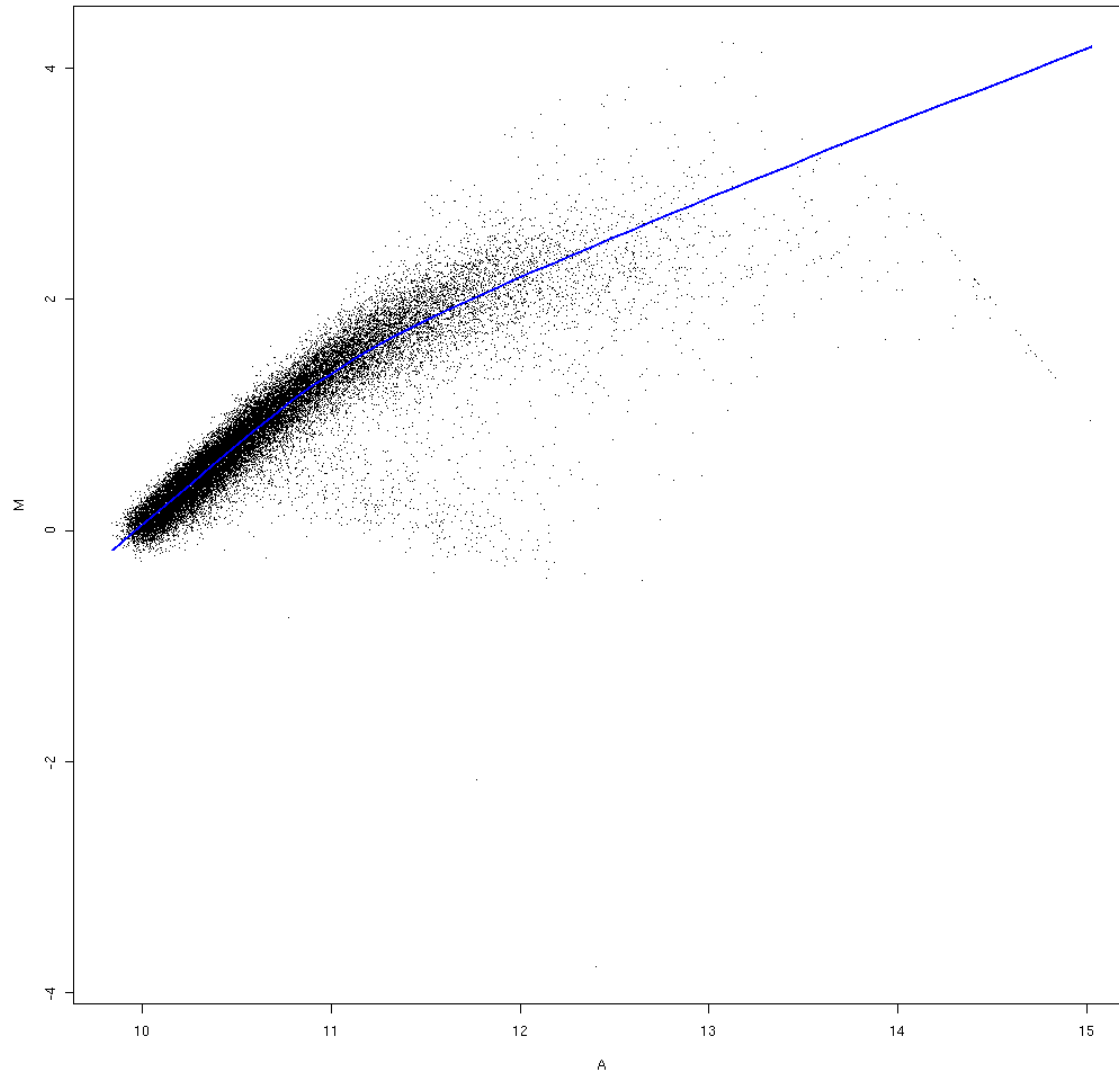
and

$$A = \log_2 \sqrt{PM_i \times PM_j}$$

Two chips are normalized by using a lowess smoother. For a collection of chips an iterative method based on all pairwise combinations is used.

Irizarry, RA, Hobbs, B, and Speed, T (2001) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Manuscript in preparation.

# Irizarry and Speed, cont

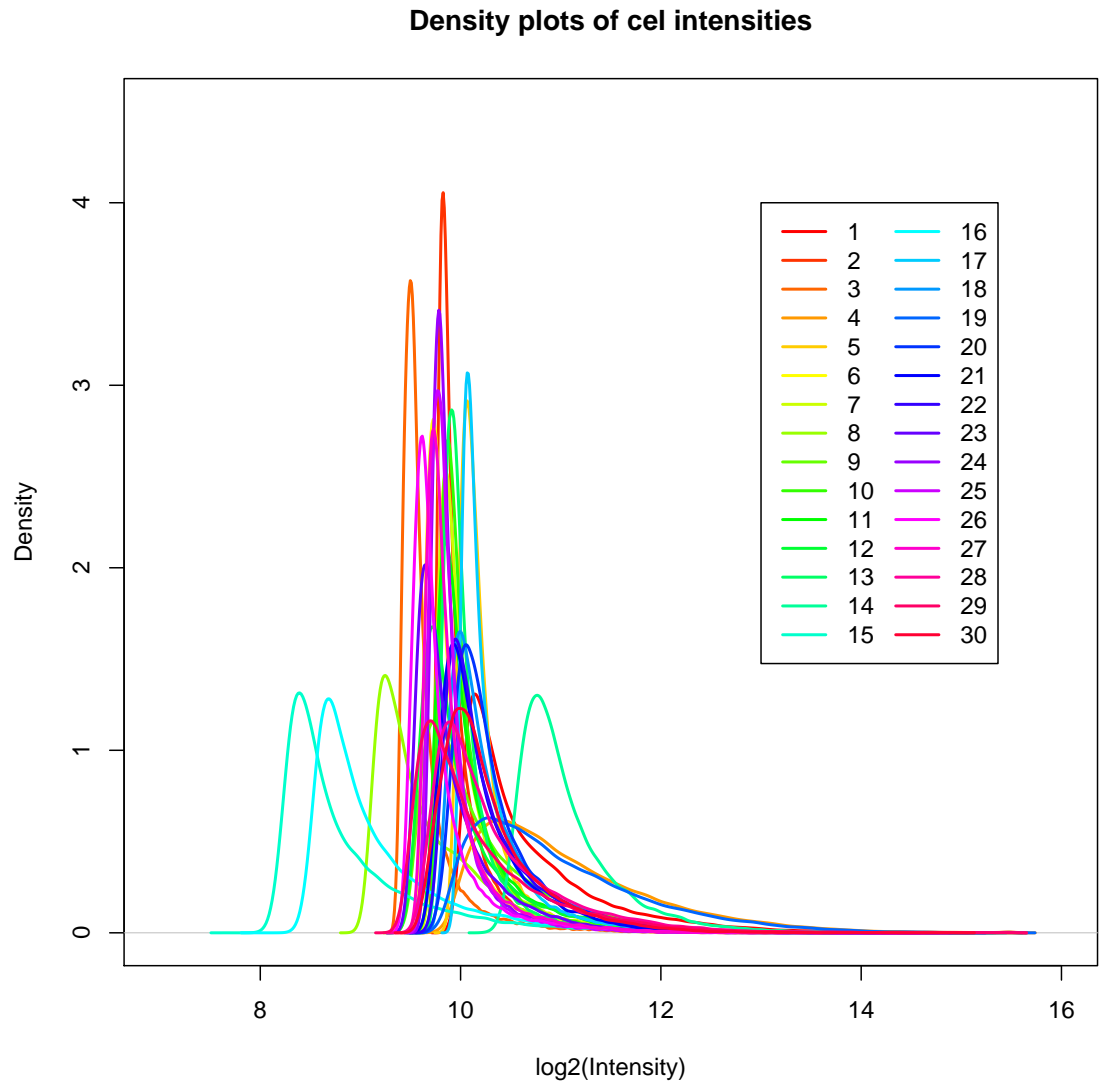


## Normalization - Quantile normalization

Based upon the assumption that the distribution of intensities for each chip should be the same. That is each chip is really the transformation of an underlying common distribution.

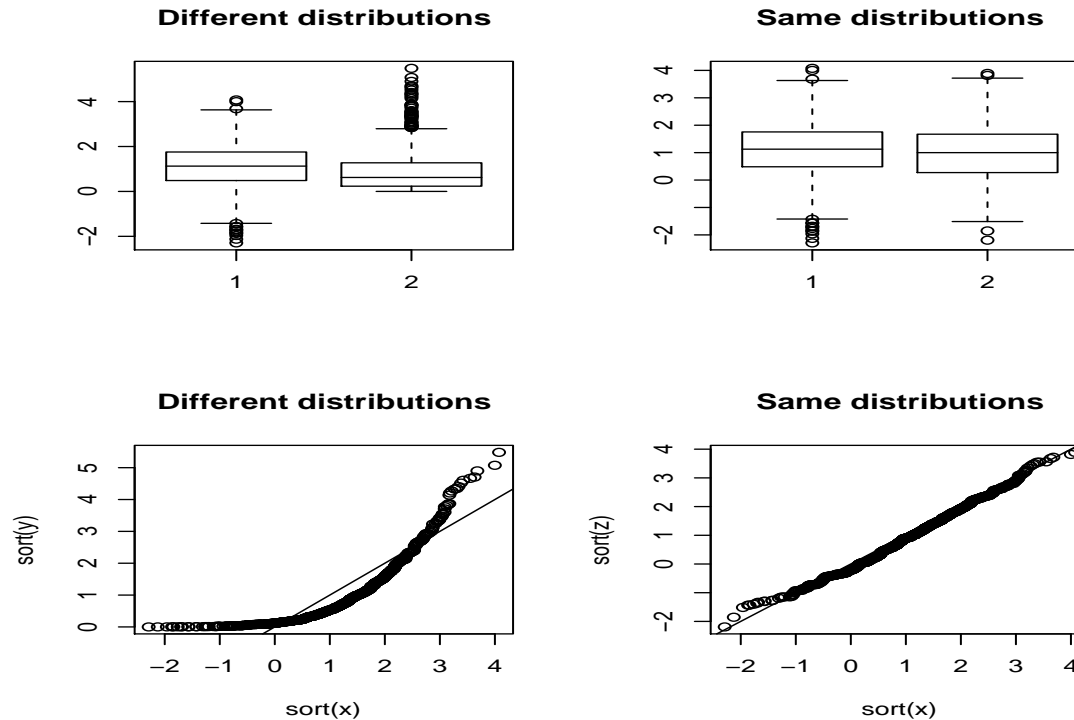
Bolstad (2001) Probe Level Quantile Normalization of High Density Oligonucleotide Array Data, Unpublished Manuscript

# Quantile normalization - Validity of assumption



# Quantile normalization - Method

Consider traditional quantile-quantile plot. Two data vectors from the same distribution will have a diagonal line quantile-quantile plot. Use this idea to motivate algorithm.



## Quantile normalization - Algorithm

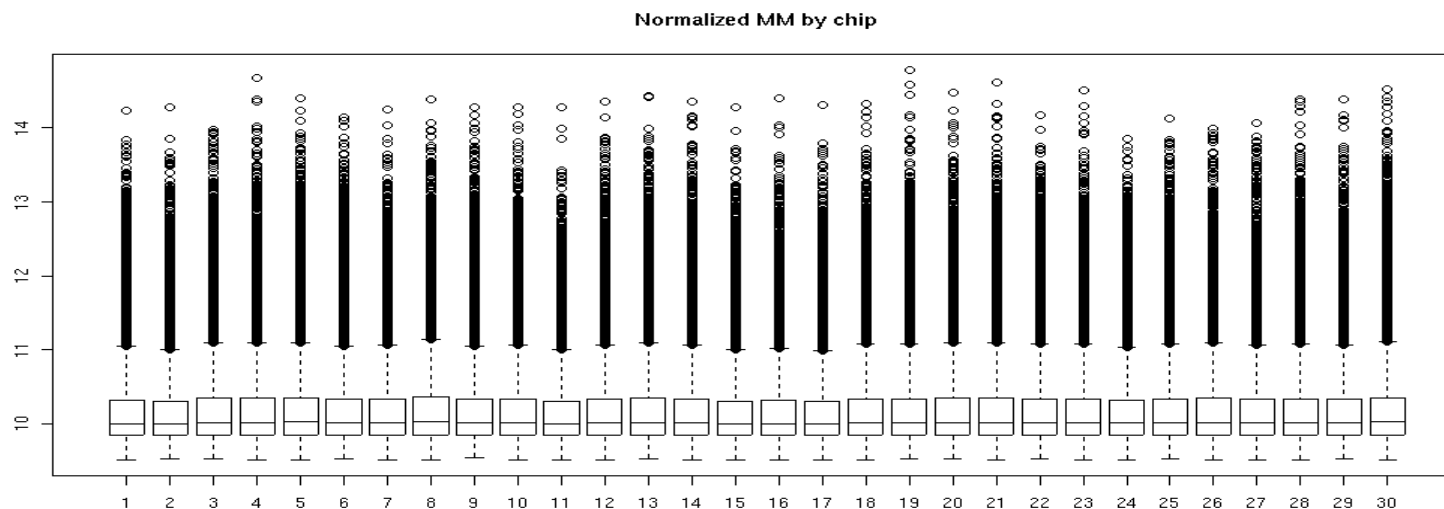
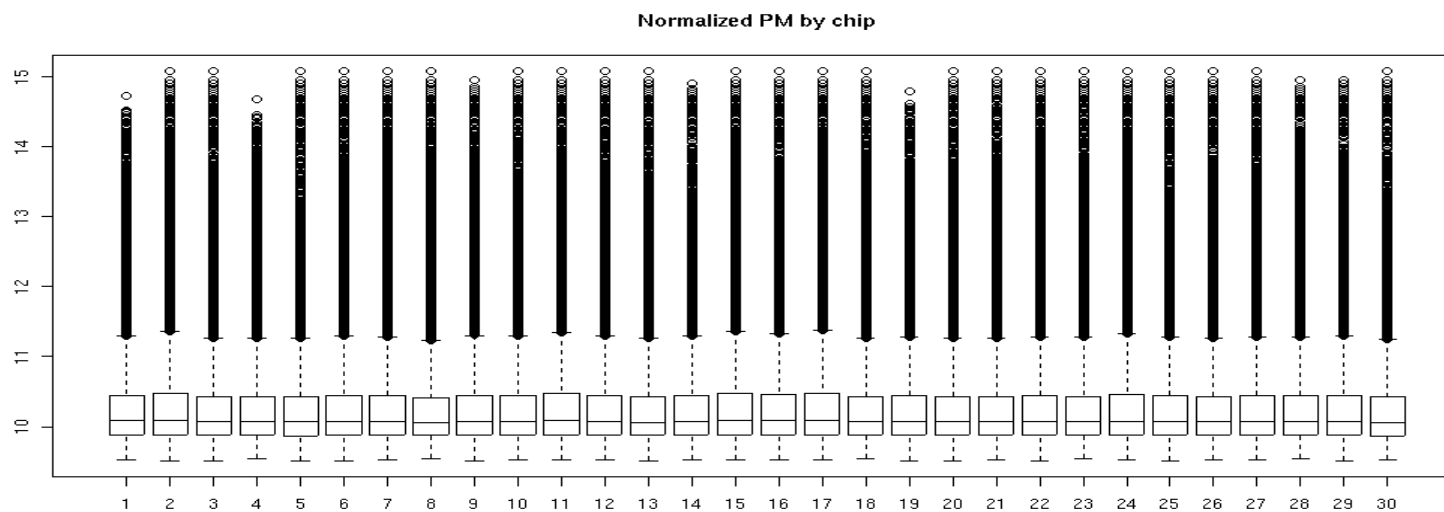
1. Given  $N$  datasets of length  $p$  form  $X$  of dimension  $p \times N$  where each dataset is a column
2. Set  $d = \left( \frac{1}{\sqrt{N}}, \dots, \frac{1}{\sqrt{N}} \right)$
3. Sort each column of  $X$  to give  $X_{\text{sort}}$
4. Project each row of  $X_{\text{sort}}$  onto  $d$  to get  $X'_{\text{sort}}$
5. Get  $X_{\text{norm}}$  by rearranging each column of  $X'_{\text{sort}}$  to have the same ordering as original  $X$

## Quantile normalization - Algorithm (a few notes)

1. If  $q_i = (q_{i1}, \dots, q_{iN})$  is a row in  $X_{\text{sort}}$  then the corresponding row in  $X'_{\text{sort}}$  is given by  $q'_i = \text{proj}_{\mathbf{d}} q_i$
2. The projection is equivalent to taking the average of the quantile in a particular row and substituting this value for each of the individual elements in that row

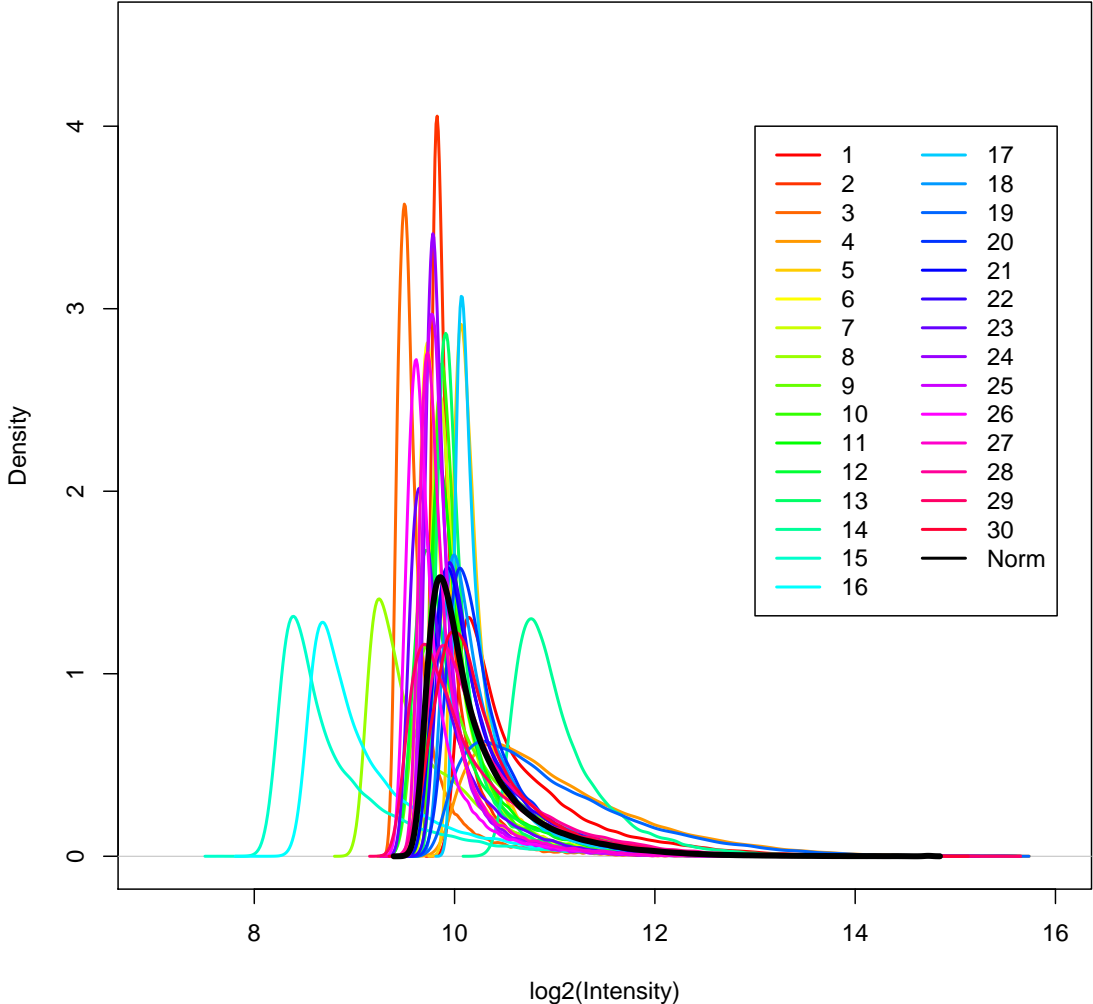
$$\text{proj}_{\mathbf{d}} q_i = \frac{\mathbf{q}_i \cdot \mathbf{d}}{\mathbf{d} \cdot \mathbf{d}} \mathbf{d} = \frac{1}{\sqrt{N}} \sum_{j=1}^N q_{ij} \mathbf{d} = \left( \frac{1}{N} \sum_{j=1}^N q_{ij}, \dots, \frac{1}{N} \sum_{j=1}^N q_{ij} \right)$$

# Quantile normalization - Results

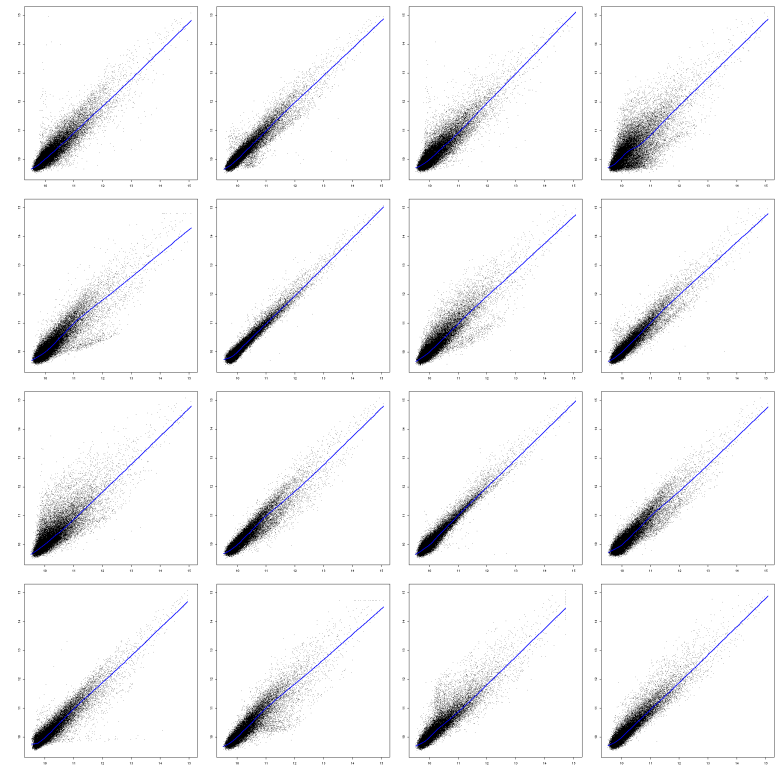
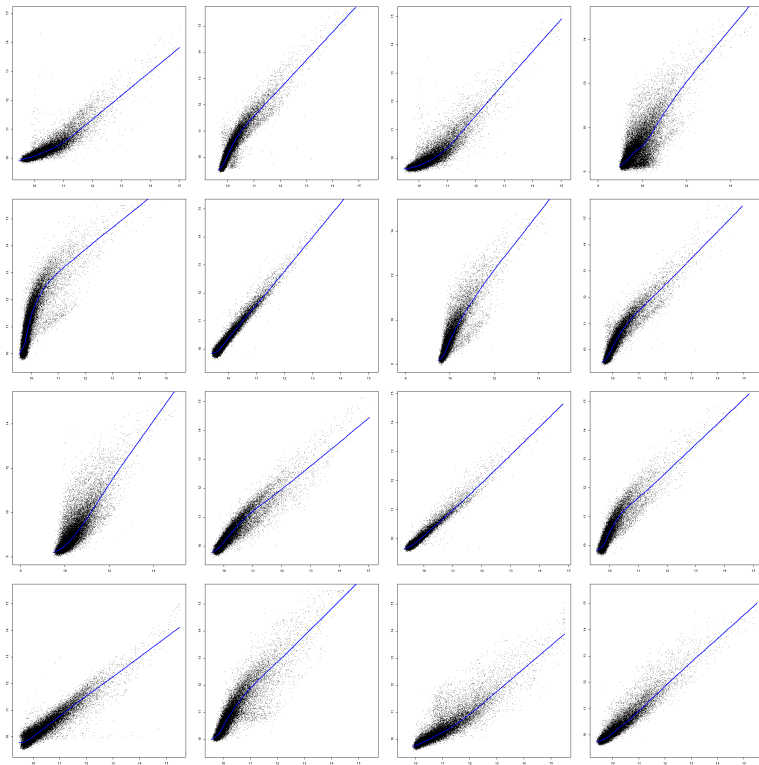


# Quantile normalization - Results cont

Density plots of cel intensities with normalized distribution

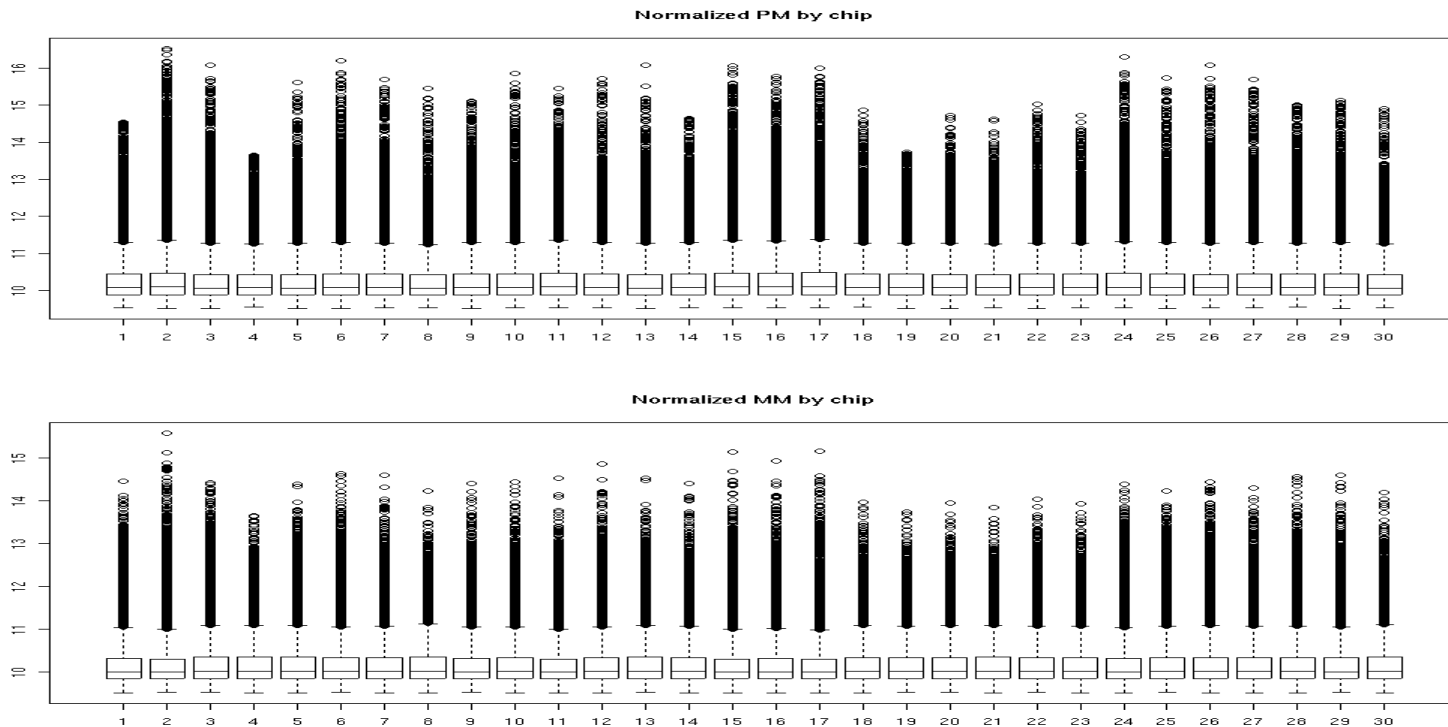


# Quantile normalization - Results PMvsPM



# Quantile normalization - Future direction

1. A tail adjustment to allow the tails to differentiate a little more.



2. Extending methodology (or finding new method) to normalize hundreds of chips

## Measures of expression

The goal is to produce a measure that will serve as an indicator of the level of expression of a transcript using the PM (and possibly MM values). The values of the PM and MM probes for a probeset will be combined to produce this measure.

## Measures of expression - Avg Diff

Used by Affymetrix in their software. Average difference is

$$\text{Avg Diff} = \frac{\sum(PM - MM)}{\#\text{probe pairs}}$$

with the following provisions made to robustify. The standard deviation of the pm-mm is computed (after removing the biggest and smallest). Any pm-mm that deviates from the mean by more than 3 standard deviations is discarded and then the average is computed.

## Measures of expression - Li-Wong

The Li-Wong method provides a Model Based Expression Index (MBEI).  
For a gene  $n$

$$y_{ij}^{(n)} = \theta_i^{(n)} \phi_j^{(n)} + \epsilon_{ij}^{(n)}$$

with  $\sum_j \phi_j^2 = J$  and  $\epsilon_{ij} \sim N(0, \sigma^2)$  where  $\theta_i^{(n)}$  is the expression index,  $\phi_j^{(n)}$  is the probe pattern and  $y_{ij} = PM_{ij} - MM_{ij}$ . Note that  $I = 1, \dots, I$  the number of chips and  $j = 1, \dots, J$  number of probe pairs.

Li and Wong (2001), Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection PNAS 98 pp31-36

Li and Wong (2001), Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application Genome Biology 2(8) pp 1-11

## Measures of expression - Irizarry-Speed

For a probeset expression measure is average  $\log_2(PM - bg)$ .  $bg$  is based upon a model (similar to Naef, but using a log normal  $bg$ ) and is a chip wide estimate. Using bias, variance and goodness of fit criteria this measure seems to be performing better than AvDiff and Li-Wong.

Irizarry, RA, Hobbs, B, and Speed, T (2001) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Manuscript in preparation.

## Measures of expression - New Affymetrix method

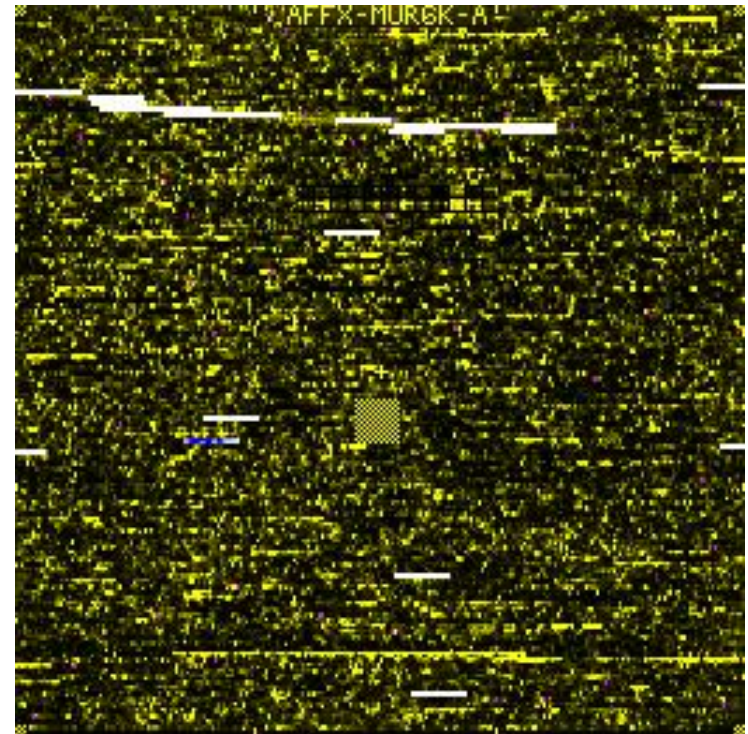
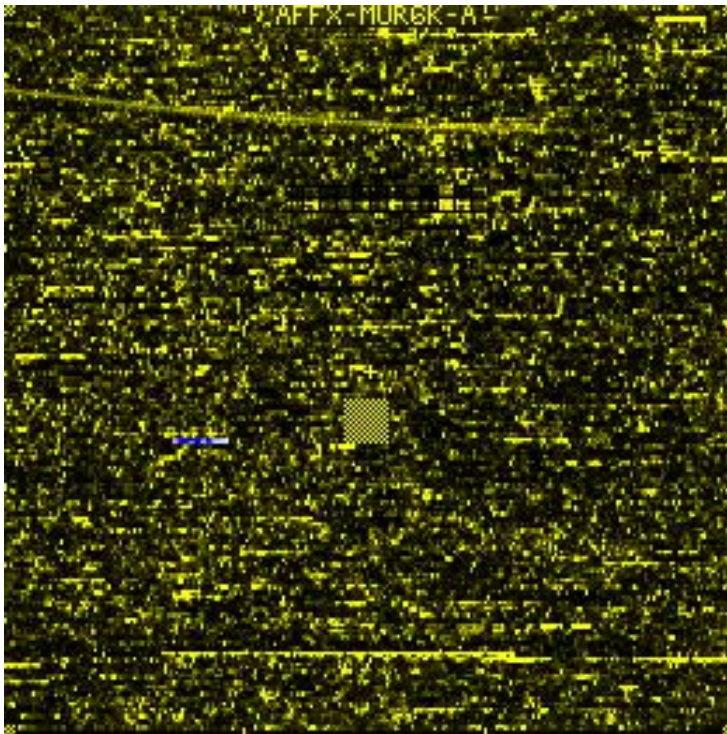
Recognizing problems with AvDiff Affymetrix have announced a new algorithm. Expression measure for a probeset is calculated using Tukey Biweight on  $\log(PM - CT)$  where  $CT$  is stray signal. Stray signal is a probe specific correction using  $MM$  (when physically possible ie  $PM - MM$  is positive) or based upon “Stray proportion” when it is not.

[http://www.affymetrix.com/products/algorithms\\_tech\\_content.html](http://www.affymetrix.com/products/algorithms_tech_content.html)

[http://www.affymetrix.com/products/statistical\\_algorithms\\_reference\\_guide.html](http://www.affymetrix.com/products/statistical_algorithms_reference_guide.html)

## Quality issues

Modifications and extensions of Li-Wong MBEI outlier criteria can be used for assessing probe/chip quality. The original image is on the left, the exclusions given by MBEI are white bars on the right. Implemented in dchip software and by others.



## Conclusions

- Many areas still need attention, the standard analysis does not seem satisfactory.
- Still not a lot of consensus on what is the right approach. Need methodology to compare different methods.
- The value of MM and whether it should be used at all is of interest.