

Some issues in low level analysis of Affymetrix Genechip data

Ben Bolstad

Nov 1, 2001

Outline

1. Genechip technology

(a) Chip design

(b) Probes

(c) Probesets

(d) Image analysis

2. Analysis

(a) Data exploration

(b) Background

(c) Normalization

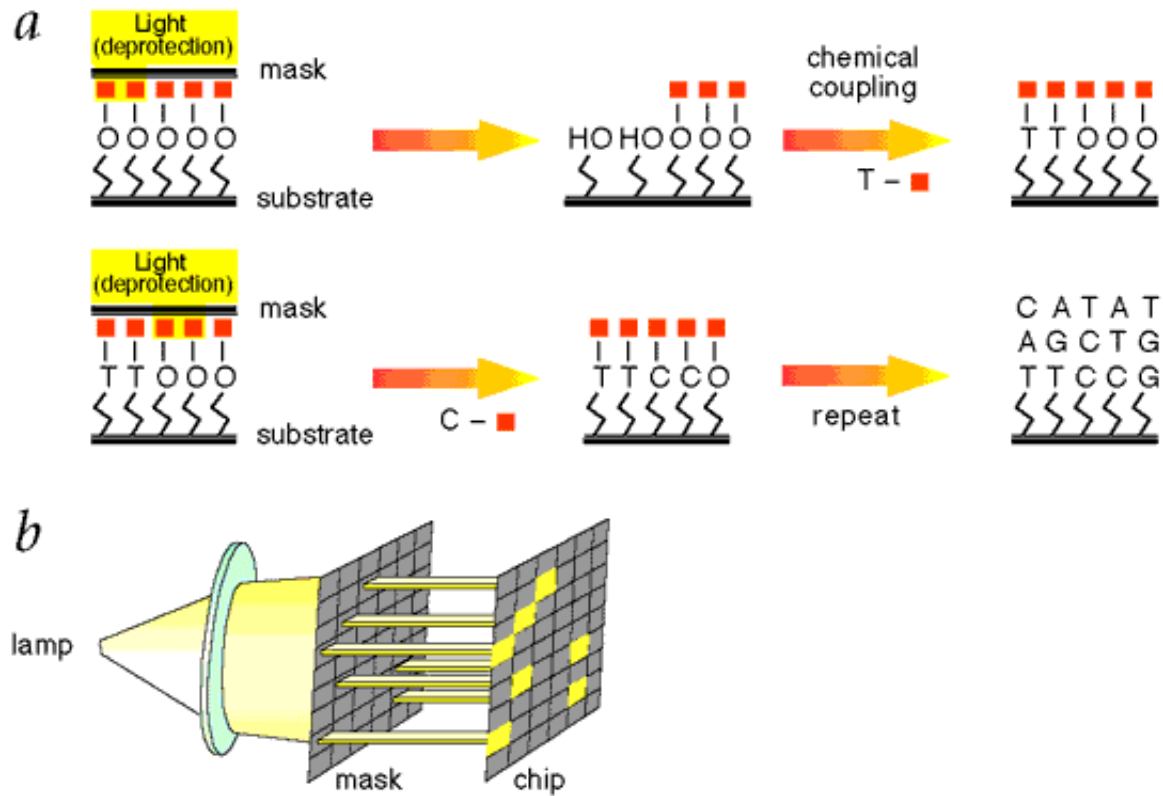
(d) Measures of expression

(e) Quality issues

Affymetrix Genechip technology

1. High density oligonucleotide microarrays
2. Anywhere from 10,000 to 400,000+ probes on a chip
3. Probes are typically 25-mers length
4. Multiple chips in a set to probe more regions of genome
5. Chips are created using photolithographic process.

Photolithographic process



Source: Lipshutz et al (1999) Nature Genetics Supplement: The Chipping Forecast

Probe selection

“Probe design is based upon complementarity to the selected gene or EST reference sequence, uniqueness relative to family members and other genes...”

In addition empirical rules are applied

“based on array hybridization data (which) help improve the odds of choosing oligonucleotides that will hybridize with high affinity and specificity.”

Source: Lipshutz et al (1999) Nature Genetics Supplement: The Chipping Forecast

Purpose of the PM and MM

1. PM: the perfect match - a probe having sequence complementary to selected sequence of gene or EST to be interrogated.
2. MM: the mismatch - identical to the PM except that the central base is different

The idea is that the MM can be used to control background and allow discrimination between “real” signal and those for non specific hybridization. A PM and MM make up a probe pair. PM and MM probe pairs are adjacent on the chip.

Example of a PM and MM

PM: CAGACATAGTGTCTGTGTTTCTTCT

MM: CAGACATAGTGTGTGTGTTTCTTCT

Probesets

To ensure redundancy, multiple probes from different parts of the sequence of a gene are used. These are known as probesets. There are typically 16 to 20 probe pairs in a probeset. Early chips had all the probe pairs in a probeset located adjacent on chip, current chips have probe pairs scattered on different regions of the chip.

Samples to image

Fluorescent tagged nucleic acid sample is injected into a hybridization chamber and hybridizes to the complementary oligonucleotides on the chip. Laser excitation enters through the glass support, focused at interface of array surface and target solution. Fluorescence emission is collected by lens and passes through a set of filters to a detector. By scanning the laser across the chip a two dimensional image is created.

Image Analysis

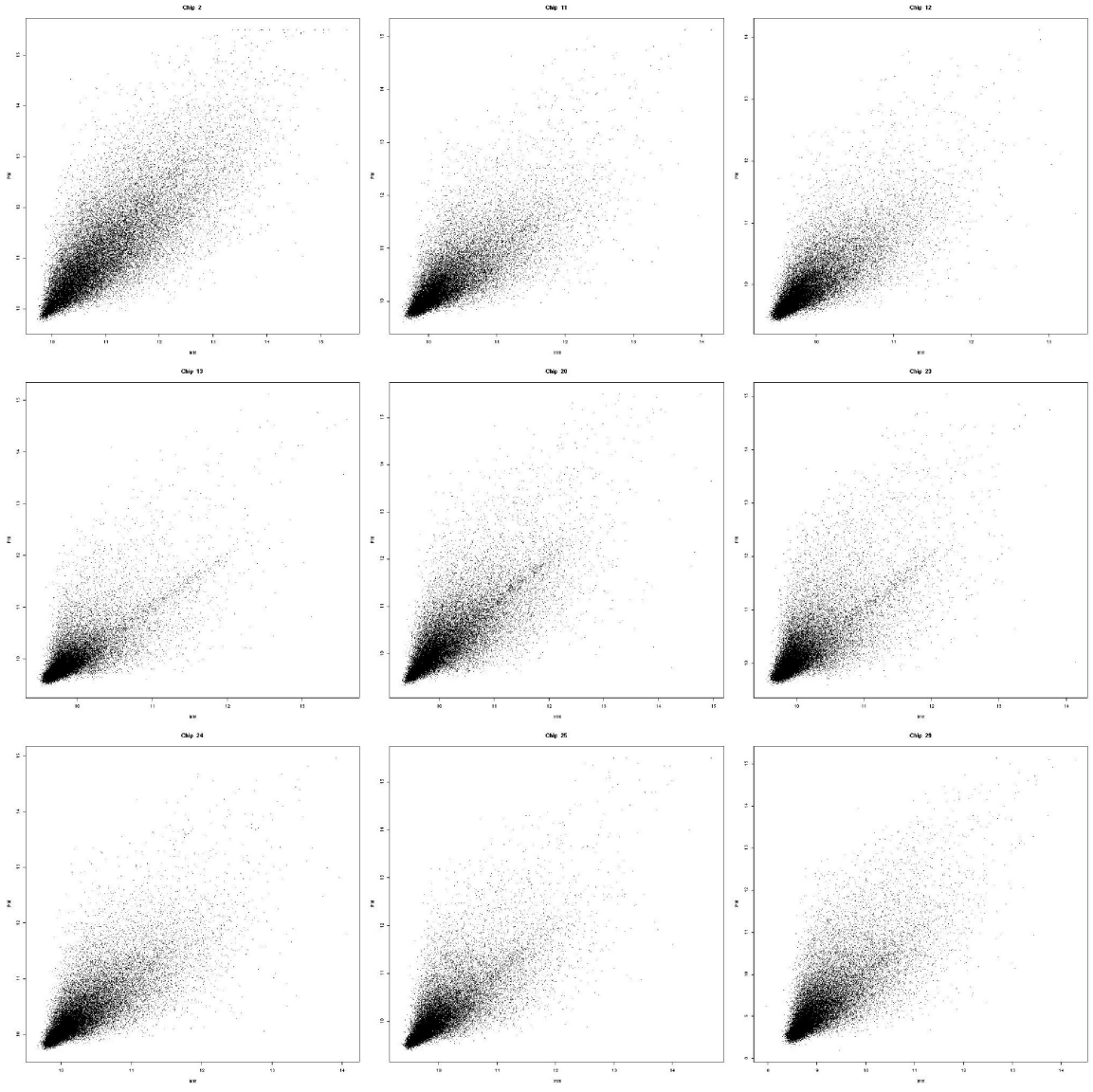
Probe arrays are scanned at high pixel resolution, on average 64 pixels per probe cell. A single intensity for each probe is then derived as follows. First exclude boundary pixels, then calculate the 75th percentile of the remaining pixels and use this as the average intensity for the probe cell.

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

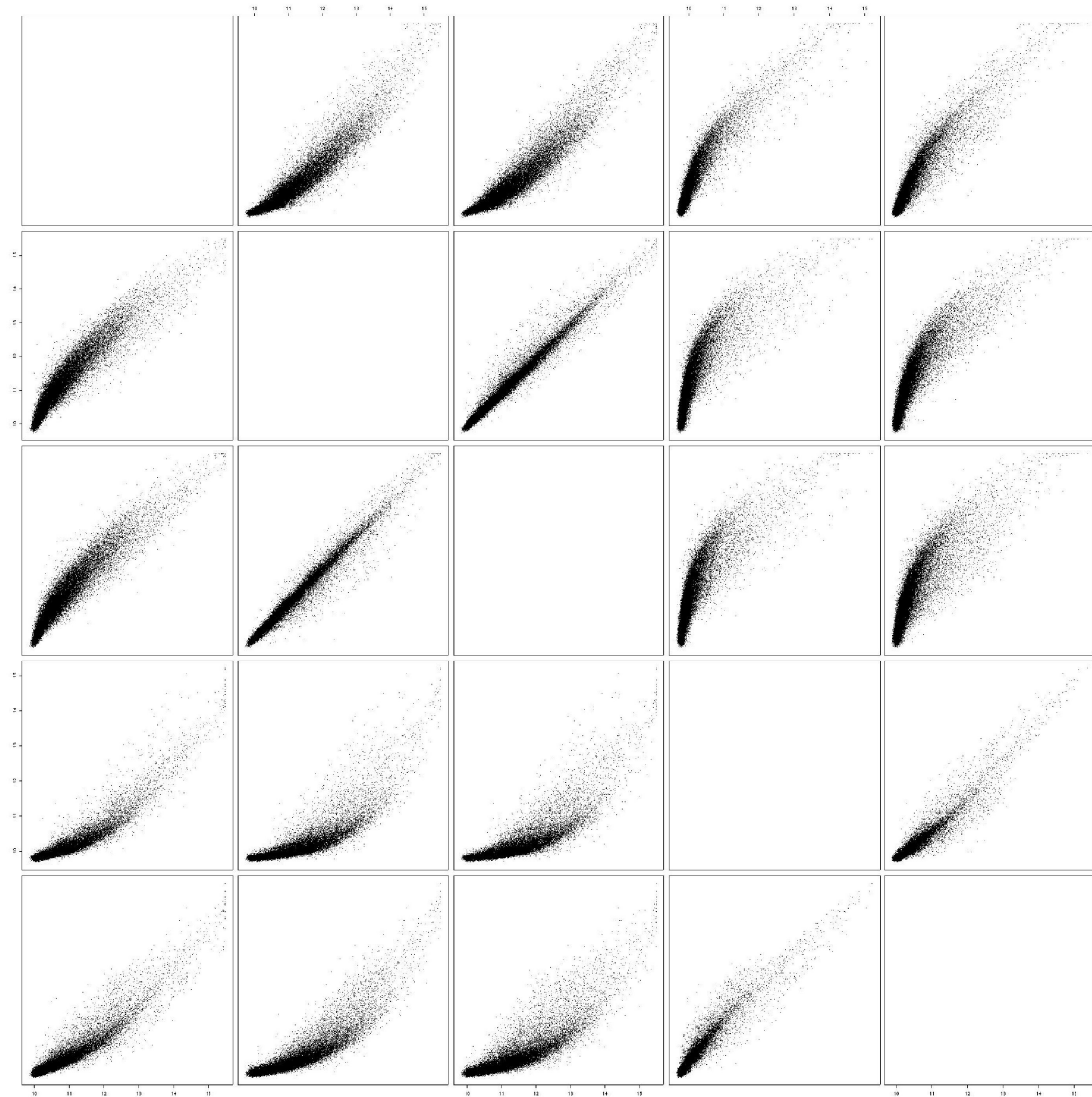
Some data exploration

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

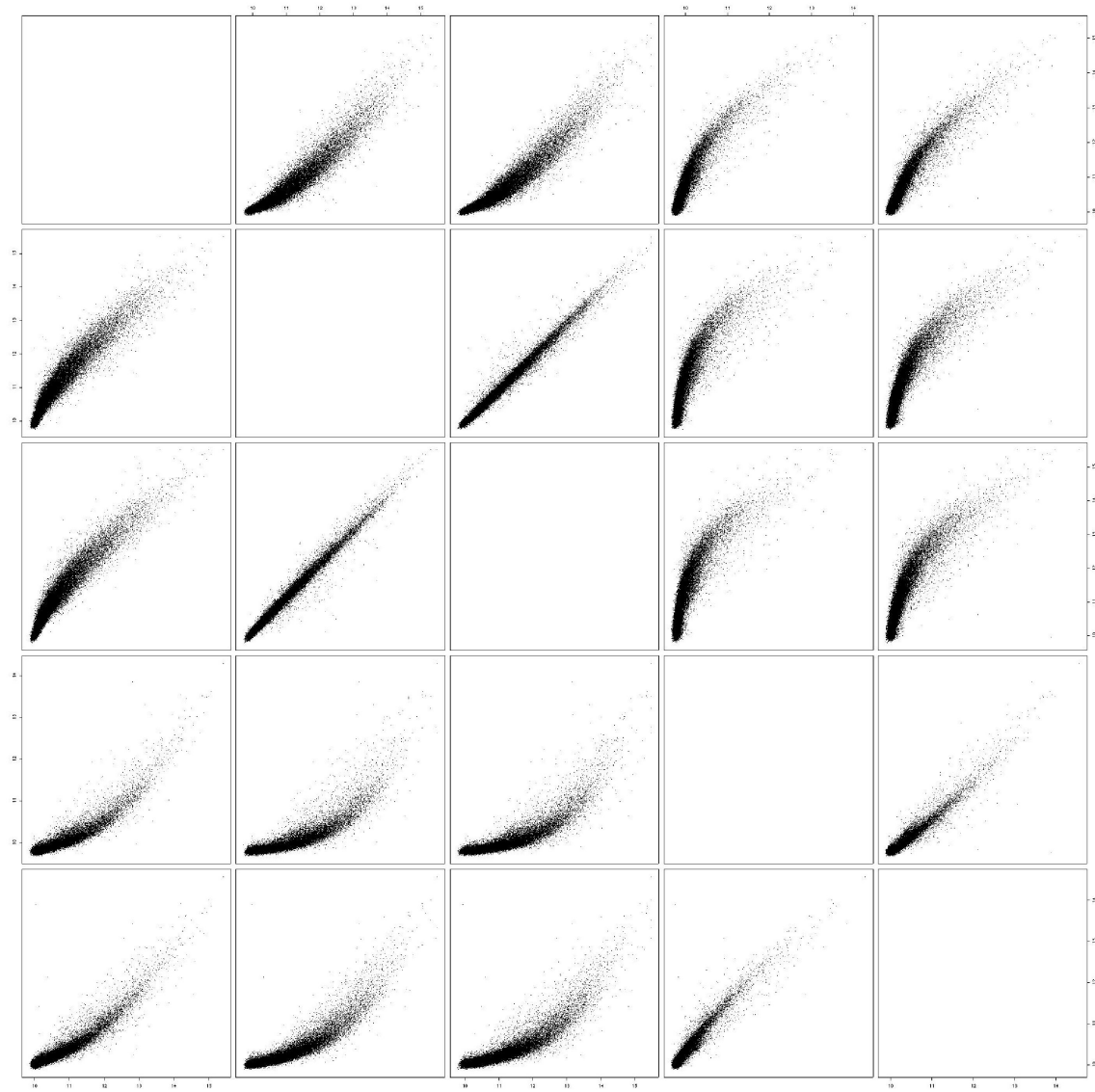
PM vs MM



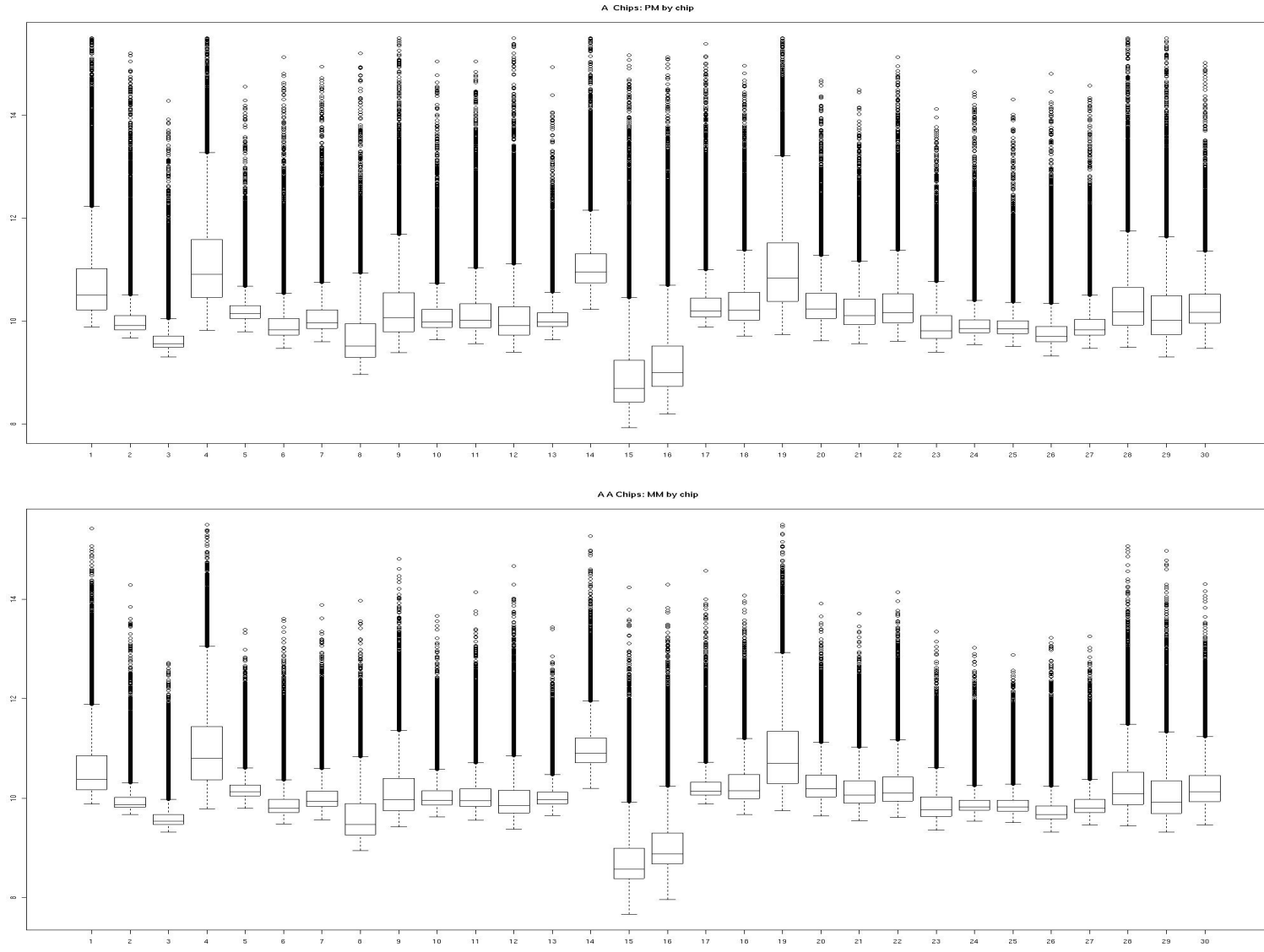
PM vs PM



MM vs MM

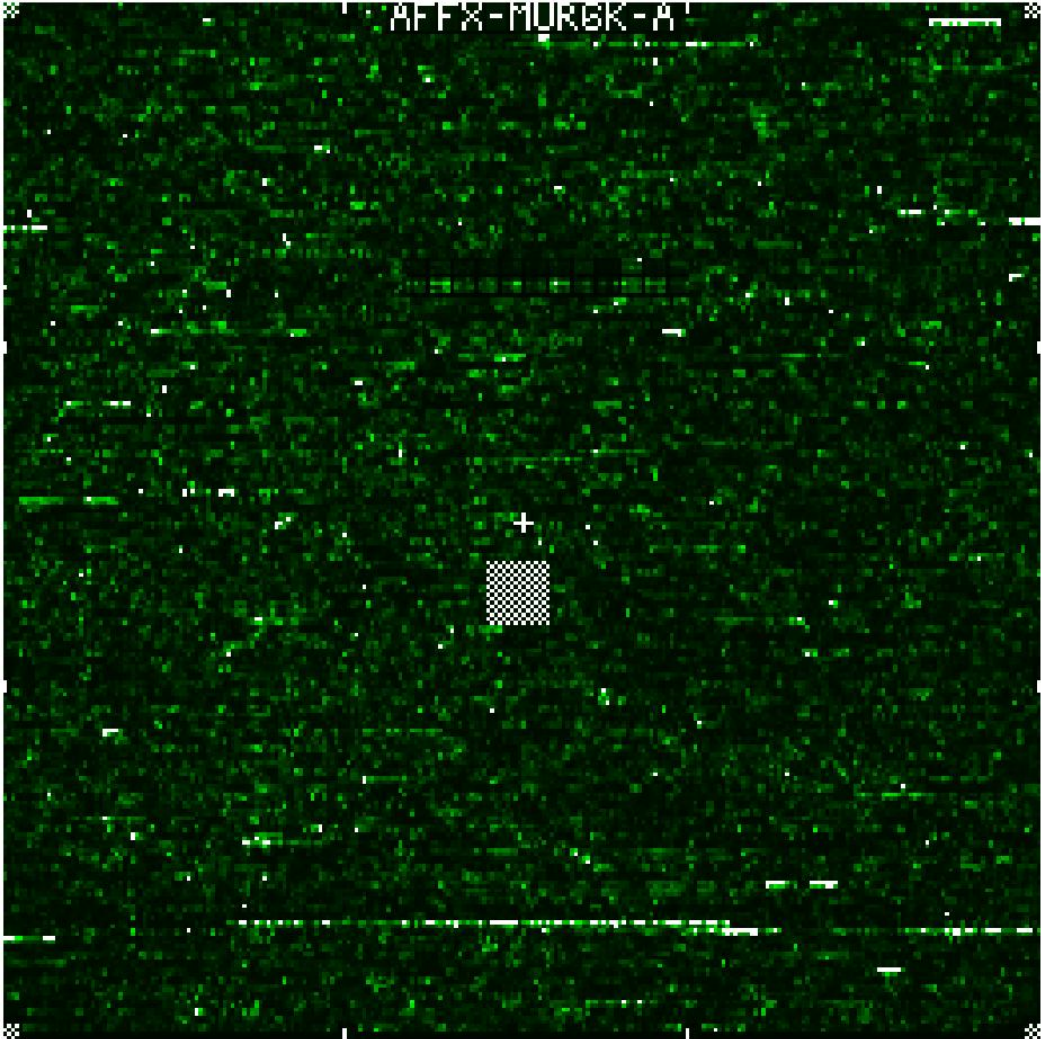


Distribution of intensities across chips



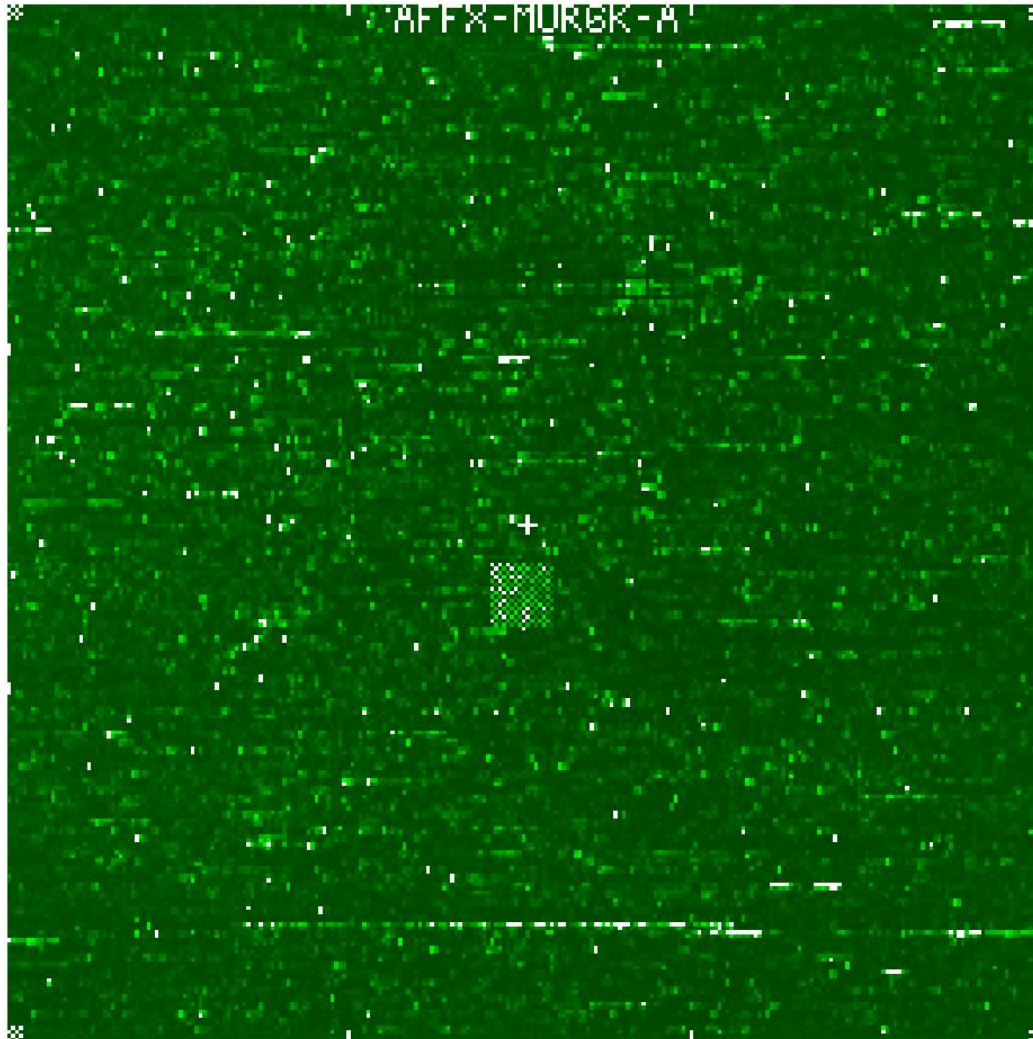
CEL image

Plot of X04A.cel

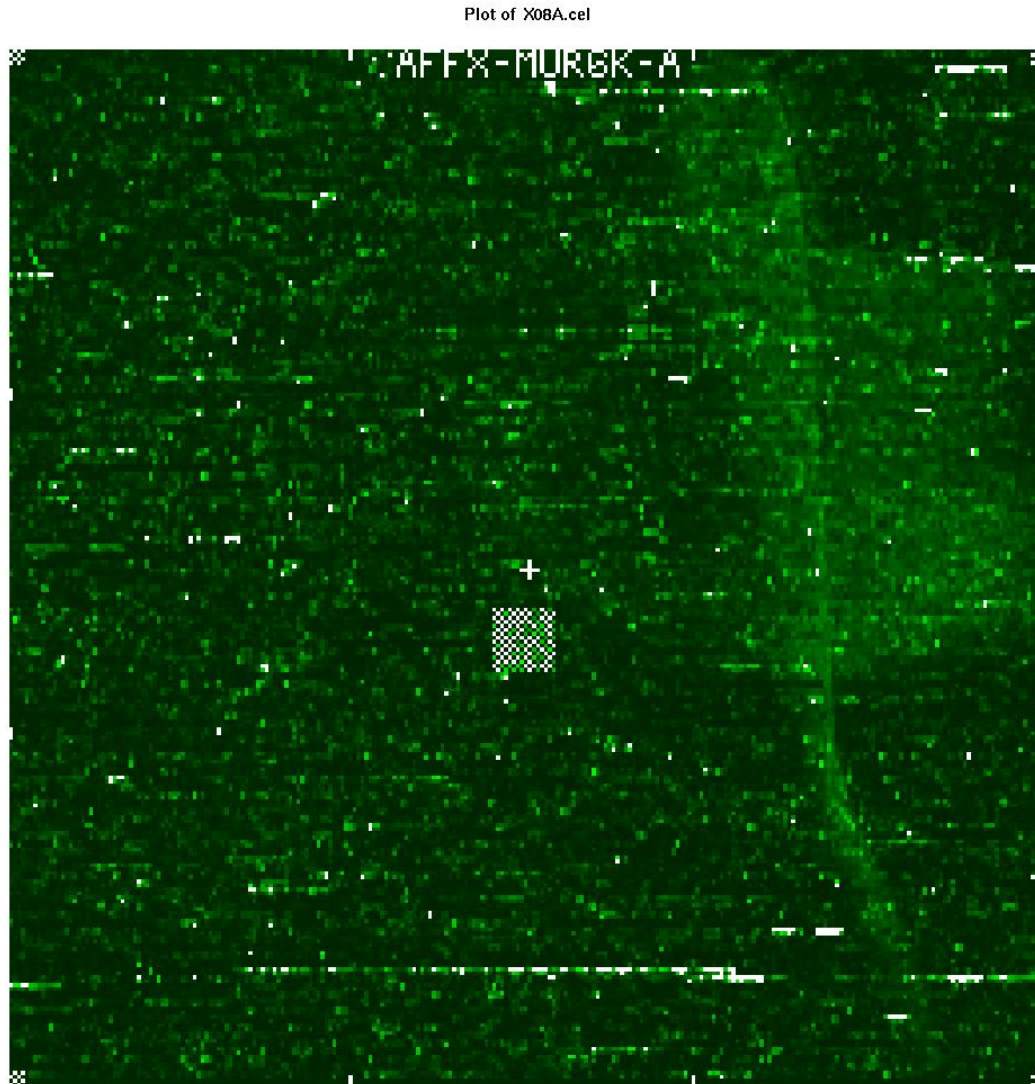


Noiser CEL image

Plot of X05A.cel



CEL image with some artifacts



Background

- A measurement of signal intensity caused by autofluorescence of the array surface and non specific binding.
- In theory the MM should serve as a background correction for the PM.
- 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.

Background - Affy

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.

Background - Naef et al

Naef et al 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 these probes have weak normal distribution. By fitting gaussians, they estimate a mean background. 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 chip₁ against chip₂ 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

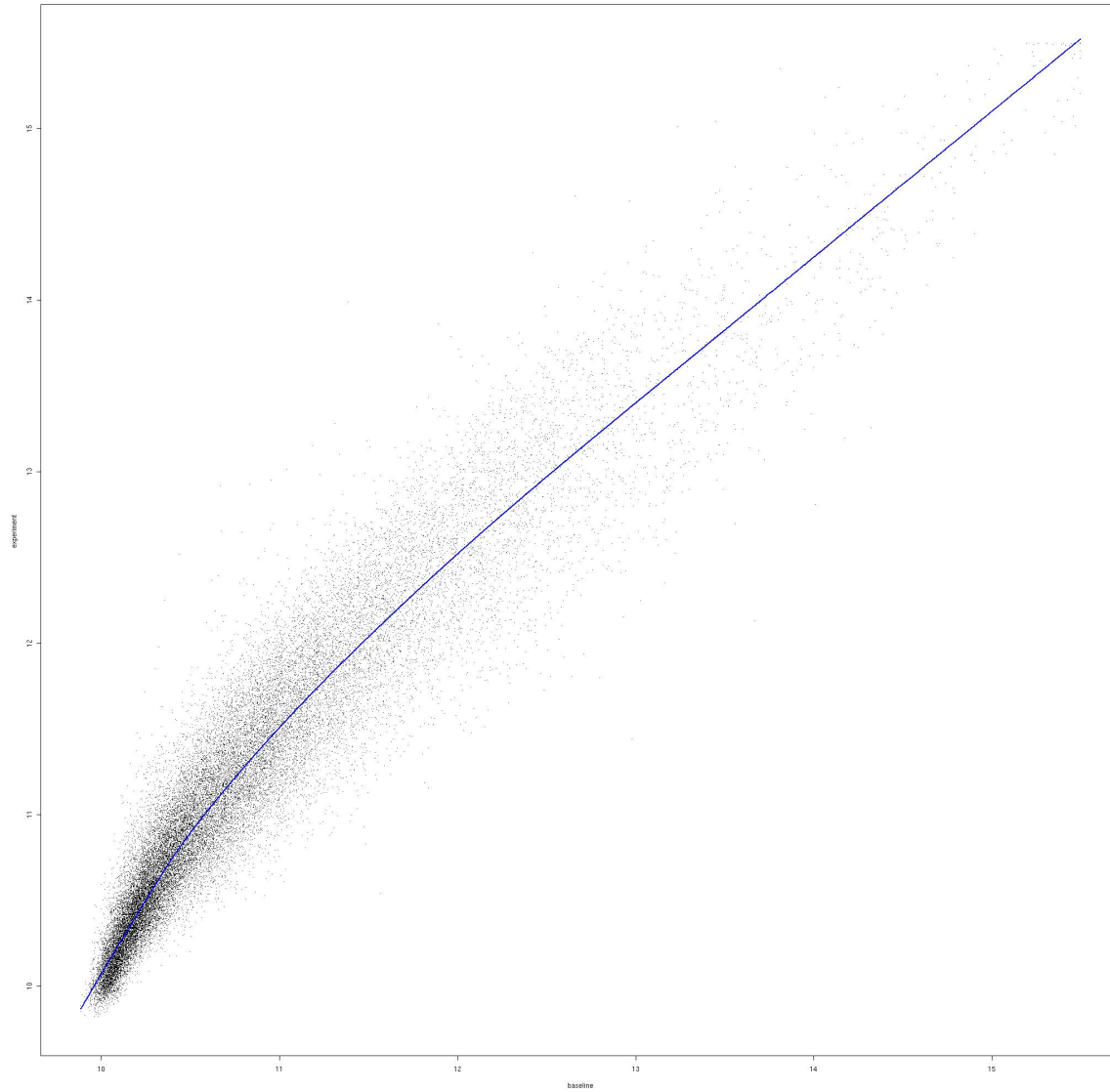
$$(PM - MM)_{new} = \hat{\beta} (PM - MM)_{old}$$

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

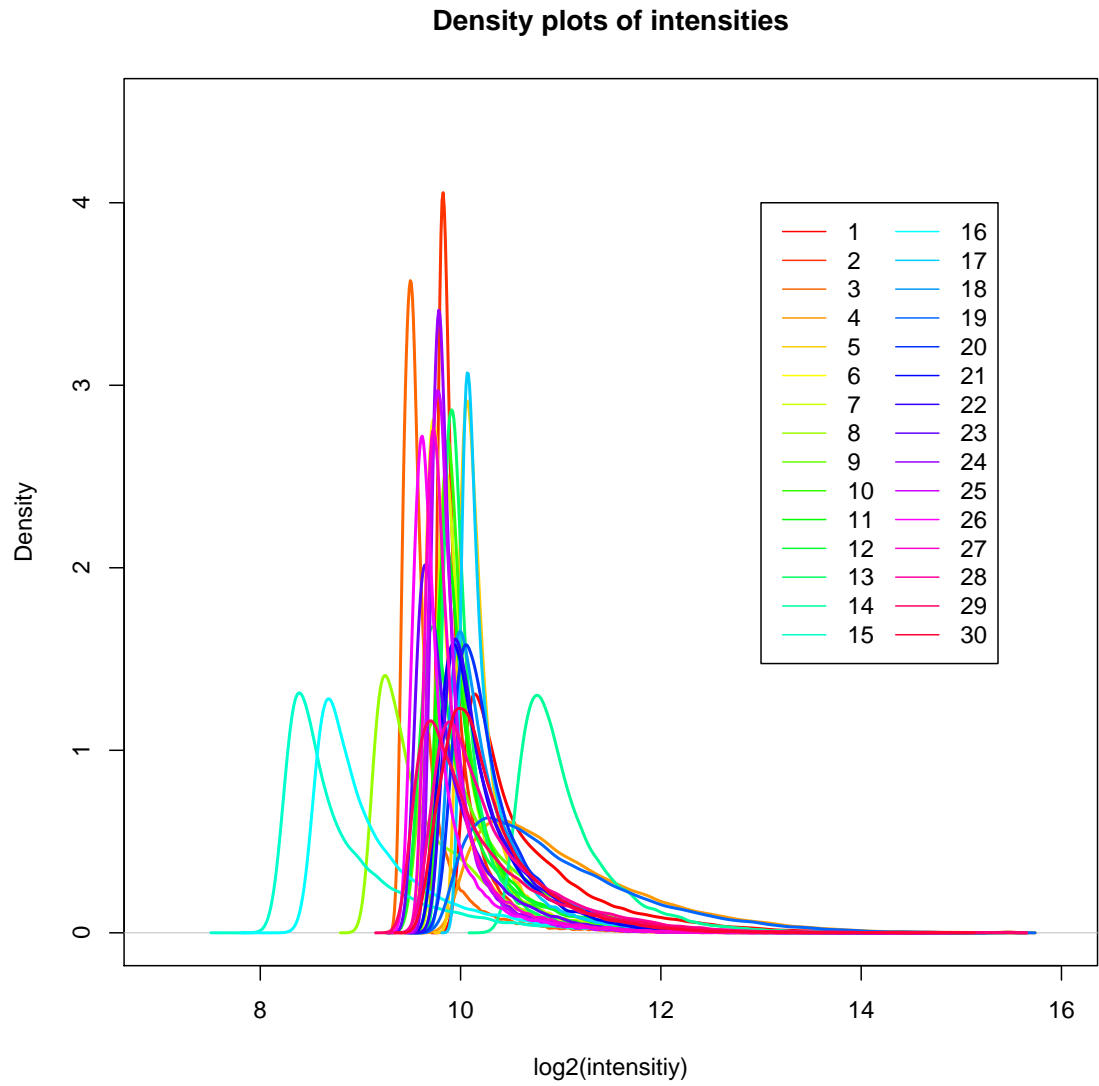
Schadt et al, cont



Normalization - Quantile normalization

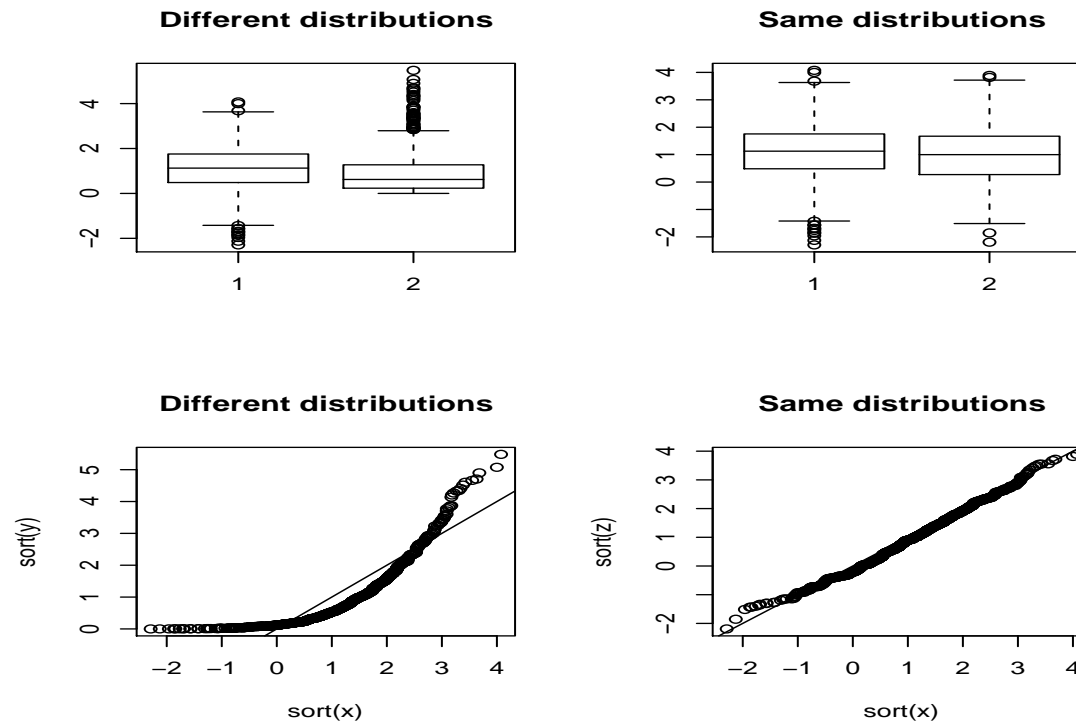
Based upon the assumption that the distribution of intensities for each chip should be the same. That is each chip is really the transformation of an underlying common distribution.

Quantile normalization - Validity of assumption



Quantile normalization - Method

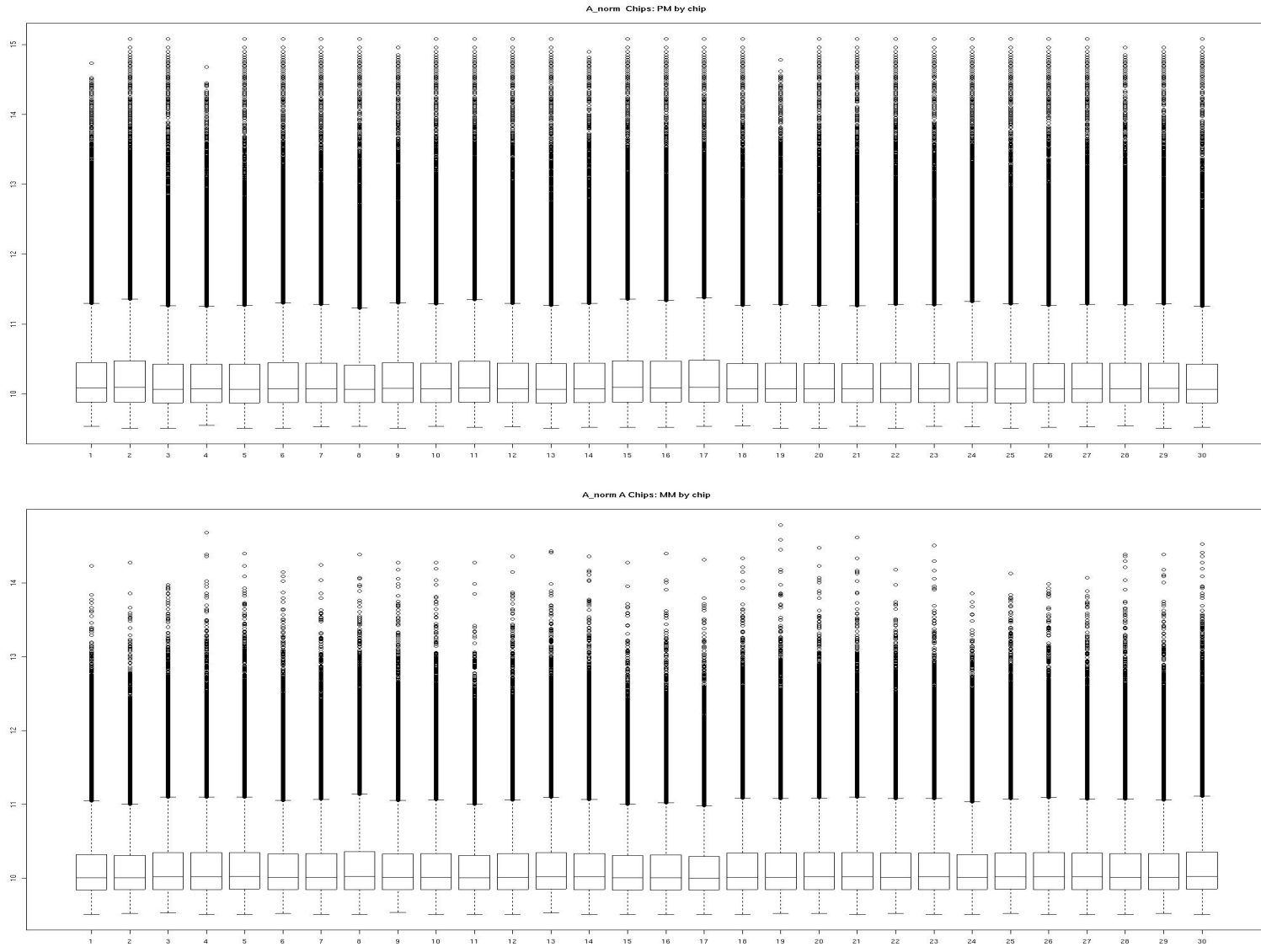
Consider traditional quantile-quantile plot. Two data vectors from the same distribution will have a diagonal line quantile-quantile plot. Use this idea to motivate algorithm.



Quantile normalization - Algorithm

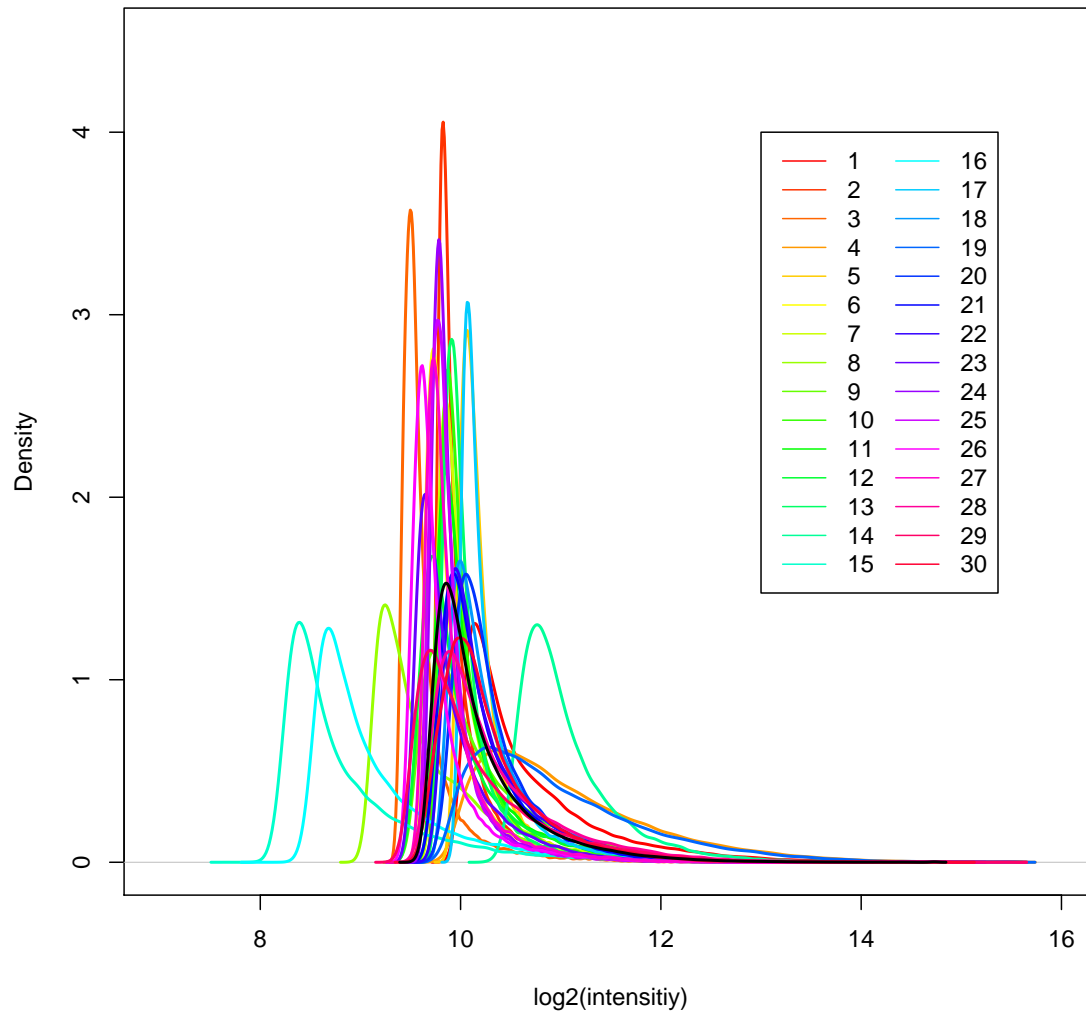
1. Given N datasets of length p form X $p \times N$ where each dataset is a column
2. Set $d = \left(\frac{1}{\sqrt{N}}, \dots, \frac{1}{\sqrt{N}} \right)$
3. Sort each column of X to give X_{sort}
4. Project each row of X_{sort} onto d to get X'_{sort}
5. Get X_{norm} by rearranging each column of X'_{sort}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

Quantile normalization - Results

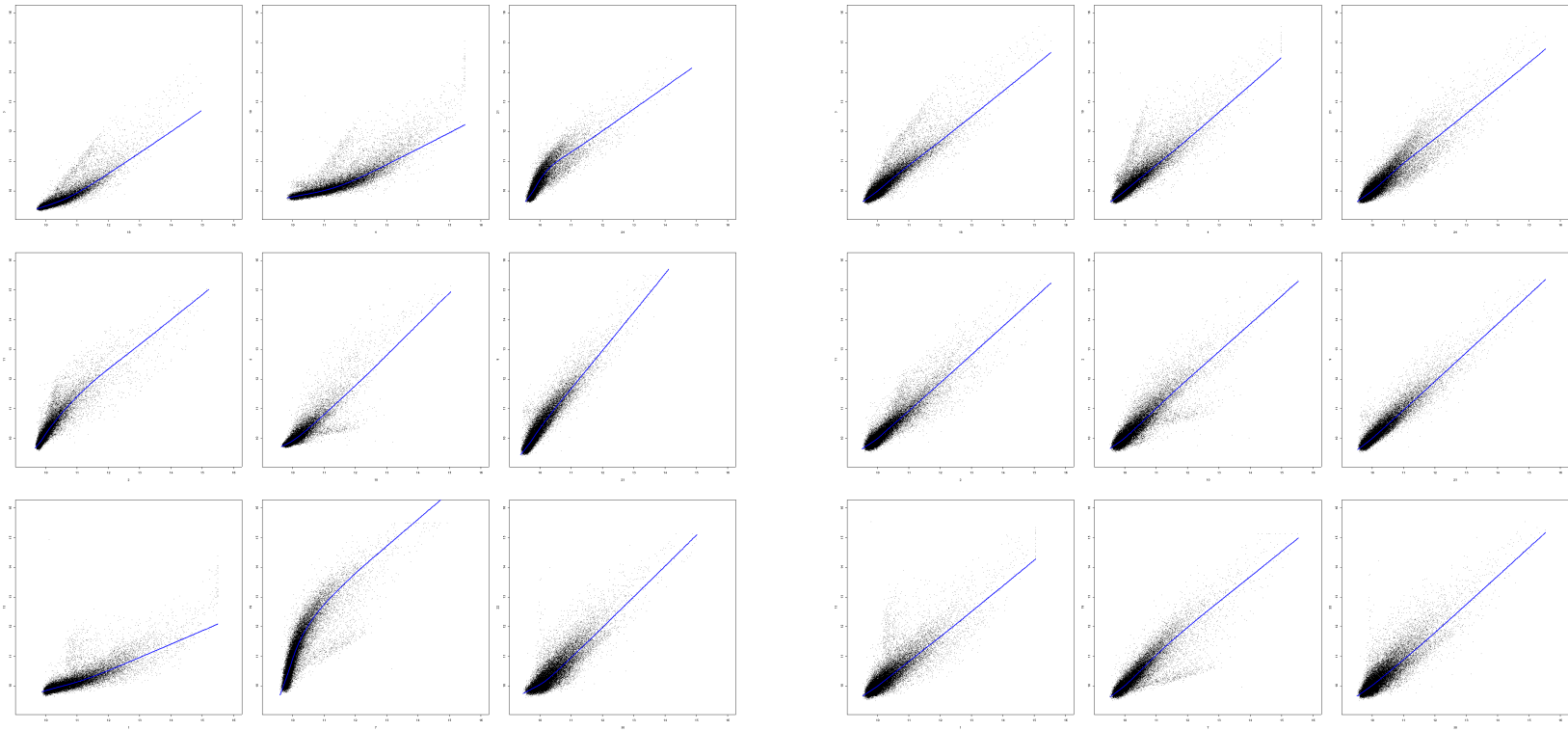


Quantile normalization - Results cont

Density plots of intensities with normalized distribution

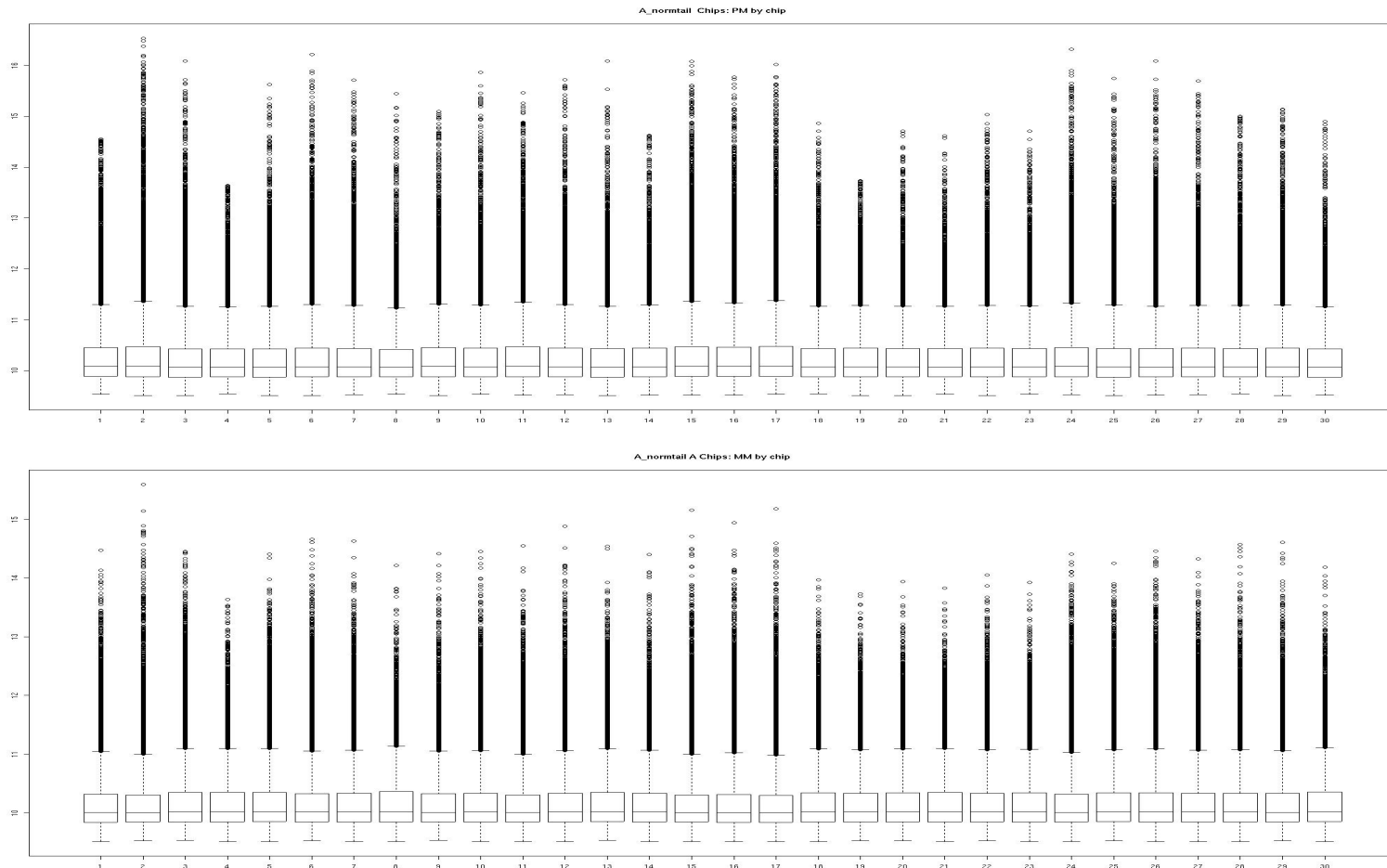


Quantile normalization - Results cont



Quantile normalization - Future direction

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



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 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

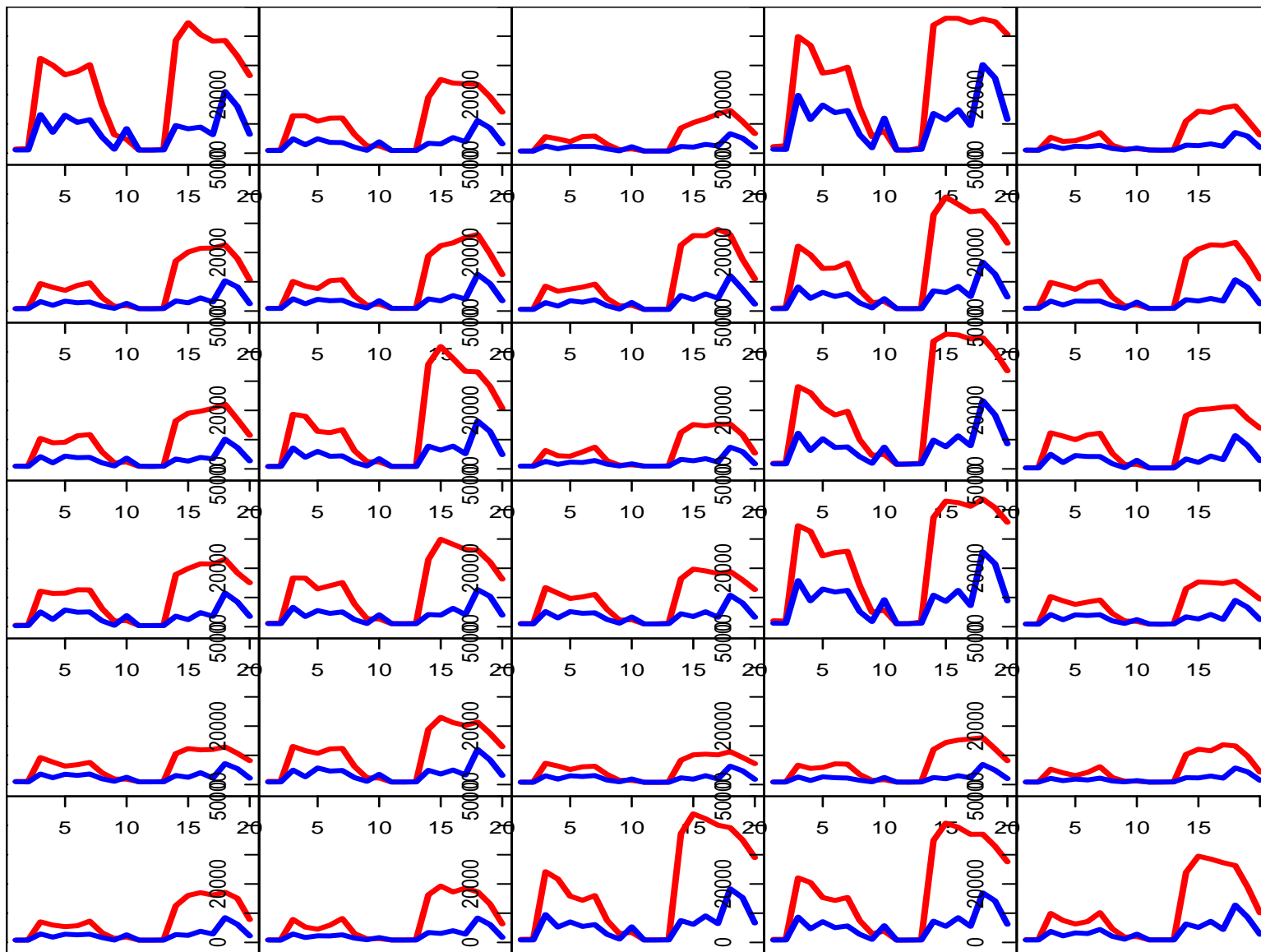
$$y_{ij}^{(n)} = \theta_i^{(n)} \phi_j^{(n)} + \epsilon_{ij}^{(n)}$$

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

Motivation for Li-Wong method



Li-Wong Algorithm

1. Initialize $\phi_j = 1$
2. Initialize $\theta_j = y_i$.
3. Iterate until convergence

(a) Solve $\hat{\theta}_i = \frac{\sum_{j=1}^J y_{ij} \phi_j}{\sum_{j=1}^J \phi_j^2}$

(b) Solve $\hat{\phi}_j = \frac{\sum_{i=1}^I y_{ij} \theta_i}{\sum_{i=1}^I \theta_i^2}$

(c) Rescale $\hat{\phi}_j = \phi_j \sqrt{\frac{\sum_{j=1}^J \phi_j^2}{J}}$

(d) Rescale $\hat{\theta}_i = \theta_i \sqrt{\frac{J}{\sum_{j=1}^J \phi_j^2}}$

Li-Wong Algorithm: Outliers

1. Determine outlier chips

(a) Calculate $se(\hat{\theta}_i) = \sqrt{\sum_{j=1}^J (y_{ij} - \theta_i \phi_j)^2 / J (J - 1)}$

(b) Outlier chips have $\hat{\theta}_i$ where $se(\hat{\theta}_i) > 3\text{median}(se(\hat{\theta}_i))$
or $\hat{\theta}_i^2 / \sum_{i=1}^I \hat{\theta}_i^2 > 0.8$

(c) With outlier chips excluded refit the model

2. Determine outlier probes

(a) Calculate $se(\hat{\phi}_j) = \sqrt{\sum_{i=1}^I (y_{ij} - \hat{\theta}_i \hat{\phi}_j)^2 / I (I - 1)}$

(b) Outlier probes have $\hat{\phi}_j$ where $se(\hat{\phi}_j) > 3\text{median}(se(\hat{\phi}_j))$

or $\hat{\phi}_j^2 / \sum_{j=1}^J \hat{\phi}_j^2 > 0.8$

or $\hat{\phi}_j < 0$

(c) with outlier probes excluded refit the model

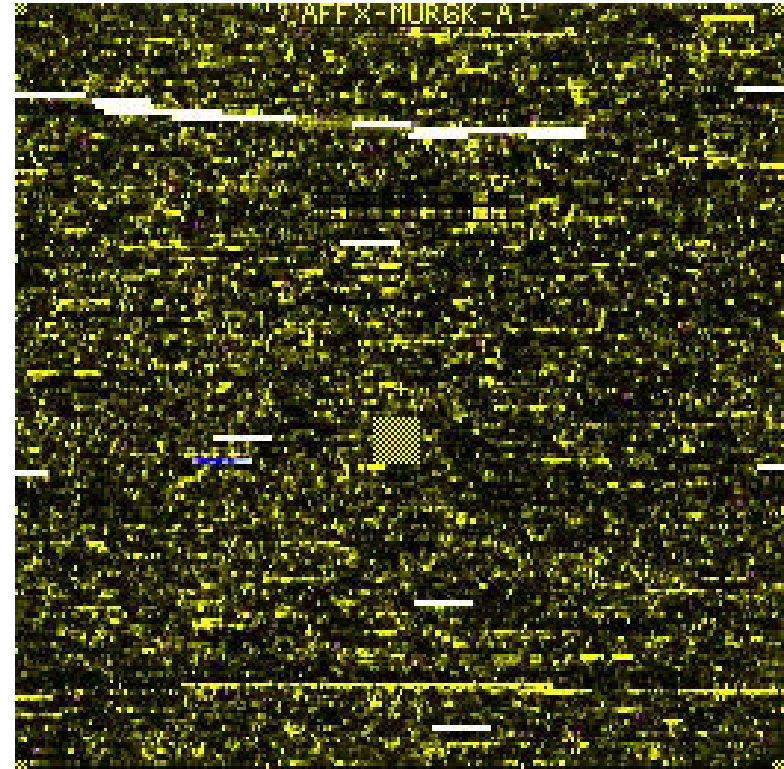
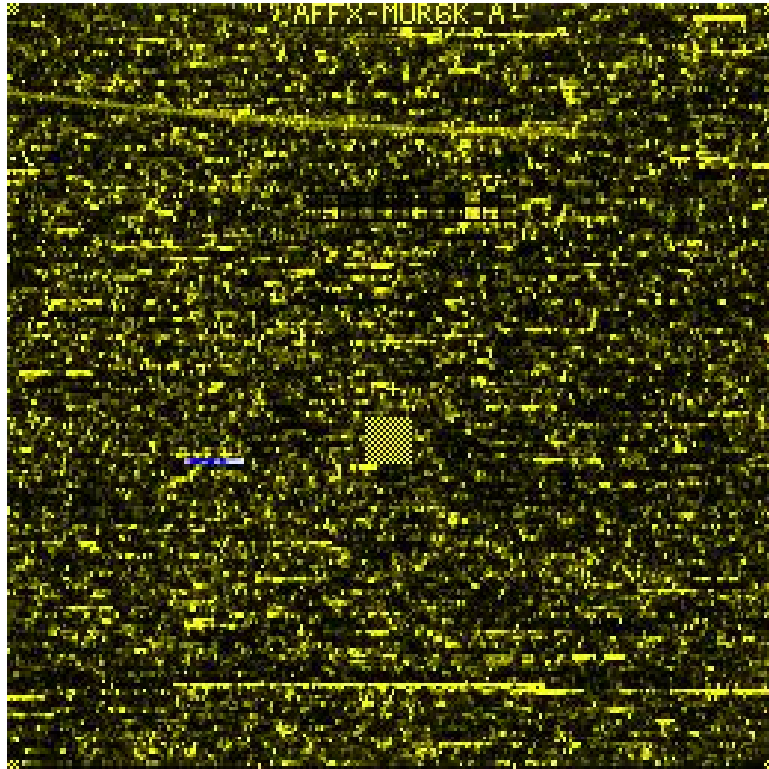
3. Determine individual outliers and refit model excluding them
4. Estimate $\hat{\theta}_i$ for outlier chips using latest ϕ_j .
5. Repeat outlier process and stop when no change in exclusions

Measures of expression - Other ideas

1. Something based on $\log(PM)$ or perhaps background corrected $\log(PM - bg)$, say average
2. Something based on $\log(PM/MM)$, say average $\log(PM/MM)$.
3. New Affymetrix algorithm
http://www.affymetrix.com/products/algorithms_tech.html
4. Li-Wong on PM only

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.



Conclusion

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