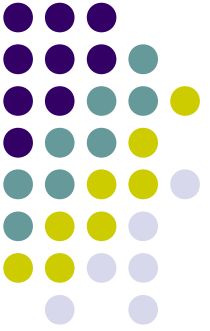
Lecture #5

Normalization and Bioconductor



Outline

- Importance of normalization
- cDNA arrays
  - M v A plots
  - Global
  - Intensity dependent
  - Within-print-tip-group
  - Dye-swap experiment
- Which genes to use?
- Affymetrix arrays
  - GeneChip® MAS 4.0
  - GeneChip® MAS 5.0
  - Li & Wong
  - Robust multi-chip normalization

Why is Normalization Necessary?

- Multiple factors contribute to the variation in sample processing
  - RNA extraction
  - Fluidics modules
  - Diverse protocols
  - Different labeling efficiencies
    - Cy3 and Cy5
  - Scanner differences
  - Chip manufacturing differences
  - Image analysis saturation
  - Other systematic variability
- These factors can depict differences between replicate samples
- Good normalization provides a method of reducing these systematic effects, while maintaining the true biological variability

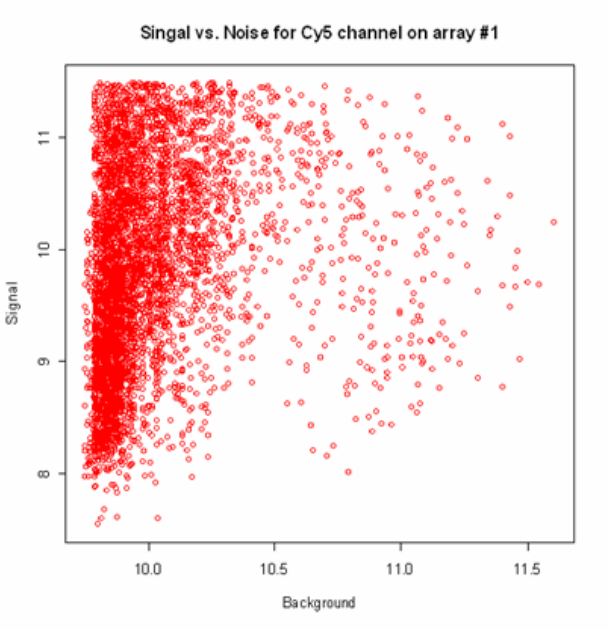
cDNA array image files

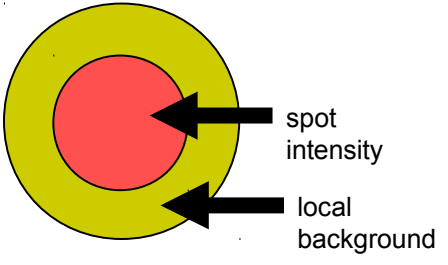
- Two channel arrays
  - Cy5 and Cy3
  - Values are reported as ratio of the two channels

- Image file
  - TIFF (16-bit file)
  - ~20MB per channel
  - ~2,000 x 5,500 pixels per image file
  - Array has mean spot area of 43 pixels
  - Array has median spot area of 32 pixels
  - Standard deviation of spot area is 26 pixels

cDNA Signal vs. Background

- Can plot the signal vs. the background for Cy5 for a single cDNA array





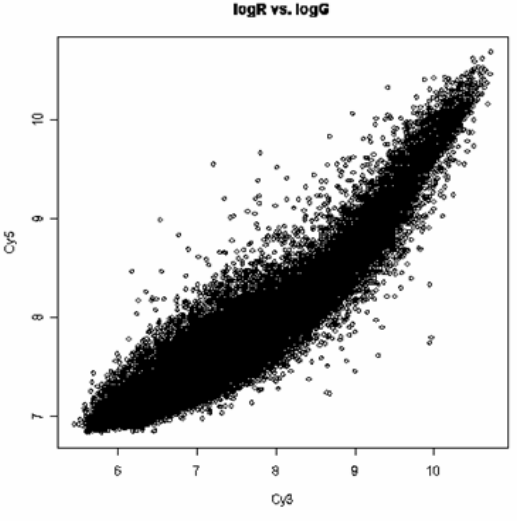
cDNA within-slide normalization

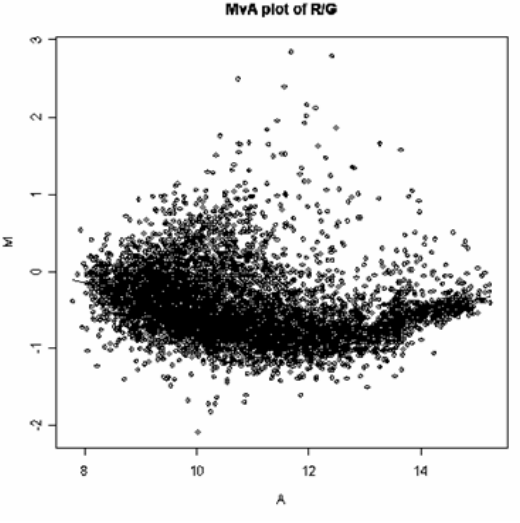
- The expression of a single array is usually plotted using the log ratio of the red dye (Cy5) vs. the green dye (Cy3)
  - This provides the degree of concordance between the two dyes
  - Deviations from a linear relationship depict systematic differences in the intensities

- However, this plot tends to give an unrealistic sense of agreement between the two dyes, so this plot has been adapted to give a better estimate of the agreement

- The MvA plot has taken this place to better represent the agreement between the two dyes
  - $M = \log_2 R/G$ where R is red dye and G is green dye
  - $A = \log_2 \text{sqrt}(RG)$
  - This is essentially a 45 degree rotation of the xy plot

Cy5 vs. Cy3 plot and M v A plot





cDNA Global normalization[1]

- Assumption
  - Provided a large enough sample size, the mean signal on an array does not vary greatly from array to array
  - Red and green dyes are related by a constant factor
    - $R = k * G$

- Methodology

  $\log_2 R/G \rightarrow \log_2 R/G - c = \log_2 R/(kG)$

  commonly, the location parameter, $c = \log_2 k$ is the mean
  - The target mean of all ratios of all the genes on the array is set to a value for scaling

- Drawbacks
  - If the assumption is violated, very large or very small intensities can increase or decrease the global mean
  - Does not account for spatial or intensity-dependent dye biases

cDNA Global normalization (cont.)

- Alternative estimators to the mean
  - The median can be used in cases of aberrant gene intensities
    - More robust to outliers
  - A trimmed mean can be used in cases of high and low extreme intensities
    - The top *n*% and bottom *n*% are excluded from the calculation of the array mean

cDNA intensity dependent normalization[1]

- Assumption
  - Dye bias is dependent upon spot intensity

- Methodology

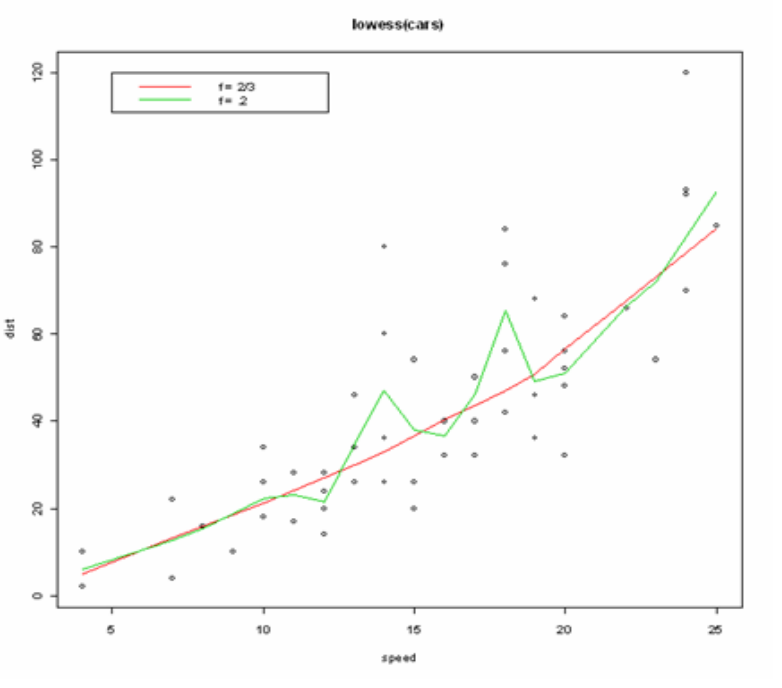
  $\log_2 R/G \rightarrow \log_2 R/G - c(A) = \log_2 R/(k(A)G)$

  where $c(A)$ is the *lowess fit to the M v A plot
  - Lowess smoothing is a robust local linear fit, which uses a specified window size to fit a curve of the data
  - Use the residual values to this smoothing for normalized log-ratio values

- Drawbacks
  - Span smoothing parameter *(f)* may deviate for each array
  - Extreme values can alter the smoothing, making a poor fit

*example illustrating concept of lowess smoothing on next slide

Lowess smoothing example

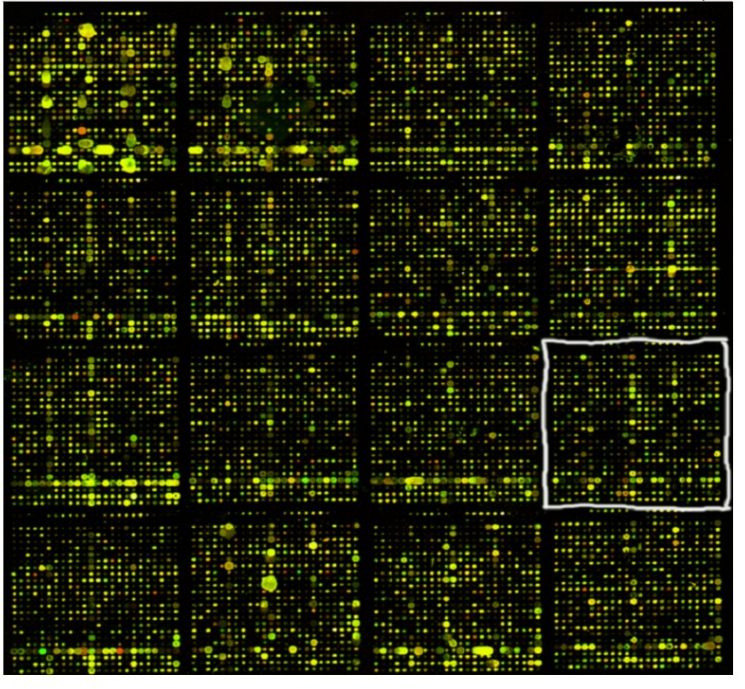


cDNA print-tip groups

4x4 design

24x25 spots per each print-tip group

Total array has 9,600 spots



cDNA within-print-tip-group normalization[1]

- Assumption
  - Differences between arrays can be explained by differences in printing setups
    - Arrayer print-tip format (2x2 or 4x4)
    - Openings or lengths of print tips

- Methodology

  $\log_2 R/G \rightarrow \log_2 R/G - c_i(A) = \log_2 R/(k(A)G)$

  where $c(A)$ is the lowess fit to the M v A plot for the ith grid only (for i=1..I for the number of print tips)
  - Use the residual values to this smoothing for normalized log-ratio values

- Drawbacks
  - Over normalization for a particular array

cDNA within-print-tip-group normalization (scale parameter) [1]

- The location normalization may correct the location of the distribution, but the scale may differ
  - Need to apply scale normalization for within-print-tip group

- Assumption
  - All log-ratios from the *i*th print-tip group are normally distributed with mean = 0 and variance = $a_i^2\sigma^2$
    - Where σ^2 is the variance and a_i^2 is the scale factor for the *i*th print-tip group
  - A relatively small number of genes will vary between the 2 mRNA samples
  - The spread of the distribution for the log-ratios should be similar for all print-tip groups

- Methodology

  a_i follows the constraint $\sum \log a_i^2 = 0$

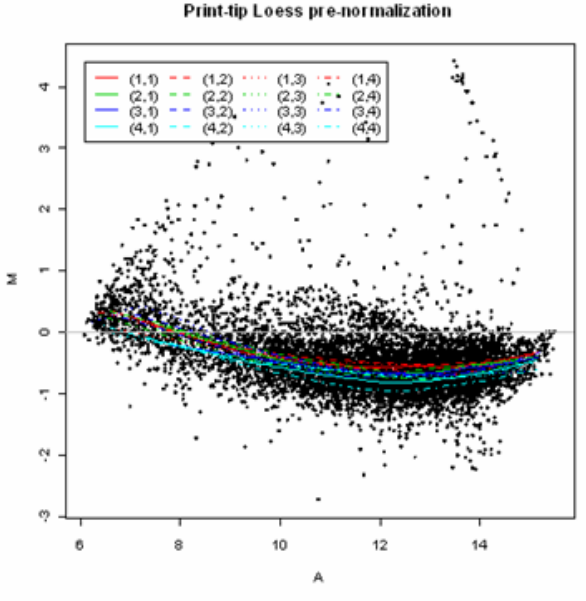
  Then, a_i is estimated by the MLE:

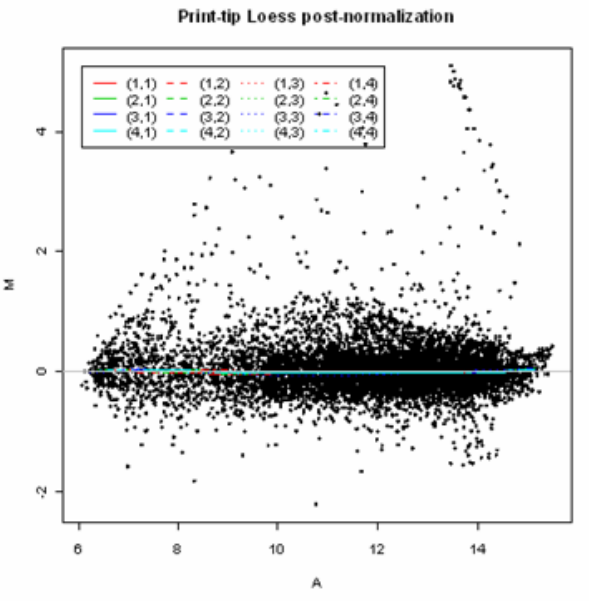
  $a_i = MAD_i / \text{sqrt}(\prod MAD_i)$

  $MADi = \text{median}_j \{|M_{ij} - \text{median}_j (M_{ij})|\}$

  where M_{ij} is the *j*th log-ratio in the *i*th print-tip group

Print-tip normalization (pre and post) [1]





Dye-Swap normalization[1]

- Two hybridizations for two mRNA samples, where the dye assignment is flipped in the second hyb.

- Assumption
  - The normalization functions are the same for the 2 slides
  - Since the assignments are reversed, the normalized log-ratios should be the same and opposite direction on the 2 slides
  - Assumes that the scale parameter is the same for the 2 slides

- Methodology
  - Slide #1: $M = \log_2 (R/G) - L$
  - Slide #2: $M' = \log_2 (R'/G') - L'$
  - $M - M' = [(\log_2 (R/G) - L) - (\log_2 (R'/G') - L)]/2$
  - $= [\log_2 (RG'/GR')] / 2$
  - $c \sim 0.5 * [\log_2 (R/G) + \log_2 (R'/G')]$

    where c=c(*A*) is estimated by the lowess fit to the plot of $0.5*(RG'/GR')$ *vs.* $0.5*(A+A')$

    (*A* is average of M and M')

Which genes to use?

- All genes
- Housekeeping genes
- Control genes

All gene approach

- All genes on the array
  - This assumes that only a fraction of the genes on the array are differentially expressed
  - The remaining genes are thought to have constant expression
  - These remaining genes constitute the majority of the expression values and shouldn't vary much from array to array, so they can be used for normalization

- Assumes
  - The fraction of differentially expressed genes is small from array to array
  - There is a symmetry between up-regulated and down-regulated genes

Housekeeping gene approach

- Constantly expressed genes
  - Use of a small subset of characterized genes that are thought to be expressed in all tissues and samples
  - Beta-actin and GADPH are among some of these genes

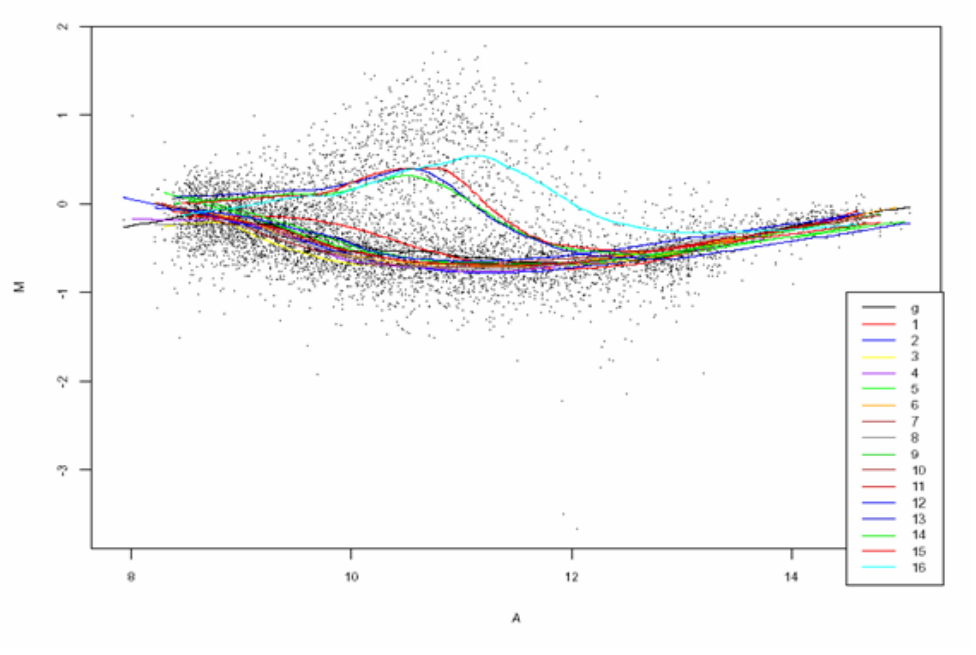
- Assumes
  - This assumes that the genes chosen as housekeeping genes are both highly expressed and somewhat invariant across multiple samples
  - These genes can be over-expressed and sometimes saturated in intensity

Control gene approach

- Control genes
  - Either spiked controls or titration of specific genes to another organism assayed at various concentrations
  - Can calculate a standard curve from the concentration series and use to normalize all other values on the array

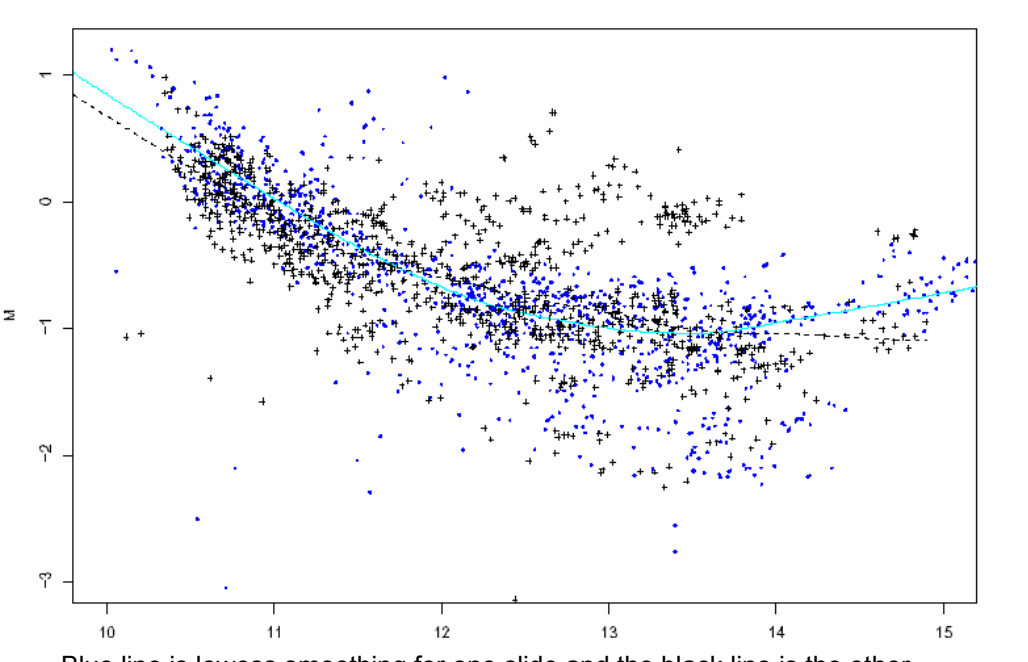
- Assumes
  - Genomic DNA is used because it is supposed to exhibit constant expression across various conditions
  - Weak signal in higher organisms with high intron/exon ratio (e.g. mouse, human) making it technically challenging

cDNA Global Normalization Data[1]



Different lowess smoothing lines for the 16 within-print-tip-groups illustrate the dependence on spot intensity

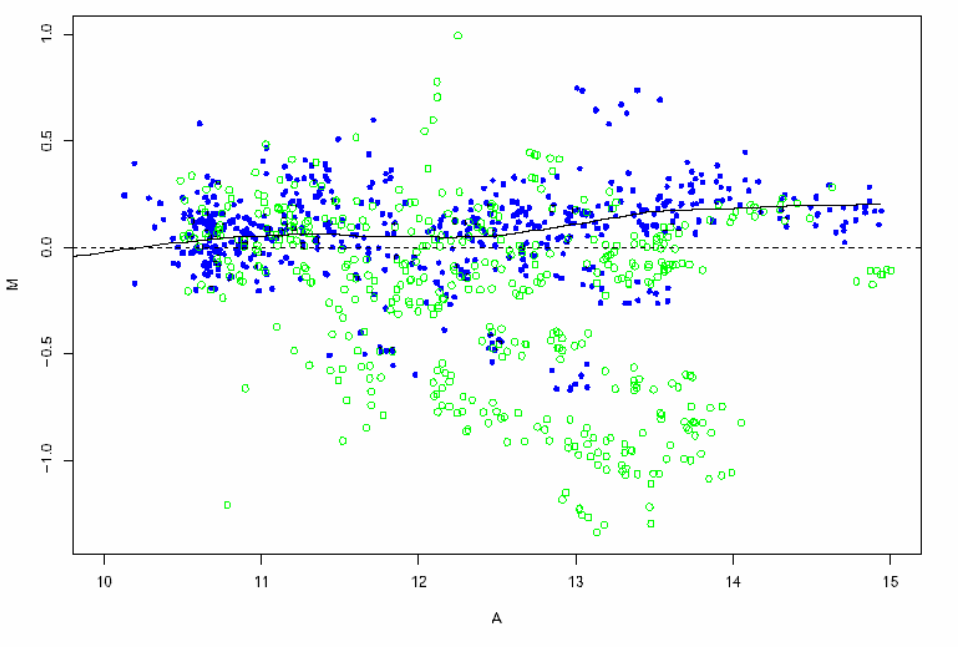
cDNA Dye-Swap Data (pre-normalized) [1]



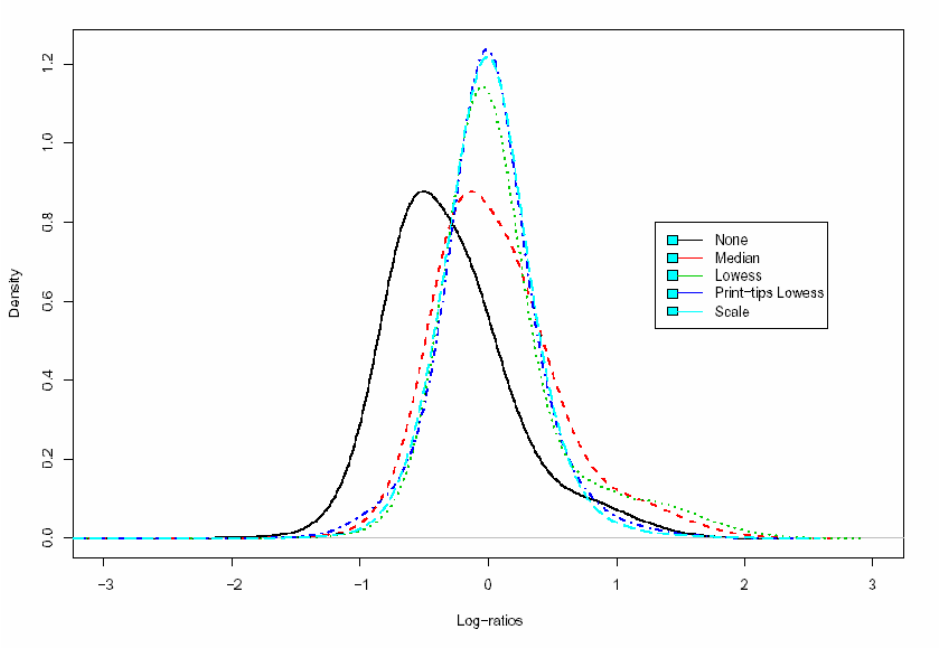
Blue line is lowess smoothing for one slide and the black line is the other

Both lines are similar, suggesting similar dye bias

cDNA Dye-Swap post-normalized[1]



Within-slide normalization density comparisons[1]



Affymetrix array image files

- Three files for each array
  - DAT file: image file with ~10e7 pixels (~50MB file)
  - CEL file: cell intensity file
  - CHP file: normalized expression data file
  - Process: DAT -> CEL -> CHP
- Data
  - Difference is computed between the perfect match (PM) and mismatch (MM) for each probe
  - Usually about 16 to 20 probe pairs for each gene

GeneChip® MAS 4.0 normalization[3]

- Average difference calculation

$$AvDiff = \frac{1}{|\mathrm{A}|} \sum_{j \in \mathrm{A}} (PM_j - MM_j)$$

where A is a set of pairs that fall within 3 SDs of the average difference between PM and MM

and *j* is the *j*th probe for gene *I*

- If MM is larger than PM, negative values will result
  - Background is larger than signal

GeneChip® MAS 5.0 normalization[3]

- Average difference with biweight calculation

$$signal = \text{Tukey Biweight}\{\log(PM_j - MM_j^*)\}$$

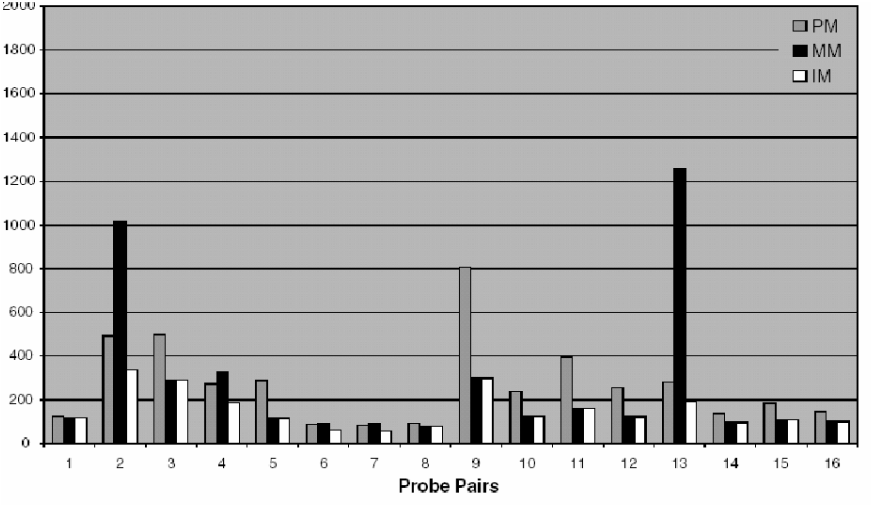
If PM > MM, then MM* = MM

If PM < MM, then MM* = PM – correction value

Correction value: robust mean of probe set using Tukey Biweight calculation

Tukey Biweight: The mean/median is first calculated, then the distance between each point and the mean/median is calculated. These distances determine how each value is weighted in the contribution to the average

GeneChip® MAS 5.0 normalization[3]



he grey bars illustrate the Perfect Match (PM) intensities and black bars the Mismatch (MM) intensities cross a 16-probe pair probe set. The white bars, Idealized Mismatch (IM), are the intensities of the lismatch based on the Signal rules. In this example, most of the Perfect Match intensities are higher ıan the Mismatch intensities and therefore Mismatch values can be used directly (e.g., probe pair 9). /hen the Mismatch is larger than the Perfect Match (e.g., probe pairs 2, 4, and 13) the IM value is sed instead of the Mismatch.

information taken from the Affymetrix manual

Li & Wong normalization[3]

- A model is fit for each probe set

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij},\ \varepsilon_{ij} \propto N(0, \sigma^2)$$

Θ: expression index in chip *i*

Φ: scaling factor characterizing probe pair *j*

ε: random normal error term

Estimates for the parameters are calculated by least squares iteratively fitting Θ and Φ, while treating the other set as known

Robust Multi-chip Analysis (RMA) normalization[3]

- Use a chip background estimate and subtract from the PM probes
  - subtracting the MM from the PM adds more noise to the signal
  - Intensity-dependent normalization

$$\log_2(PM_{ij} - BG) = a_i + b_j + \varepsilon_{ij}$$

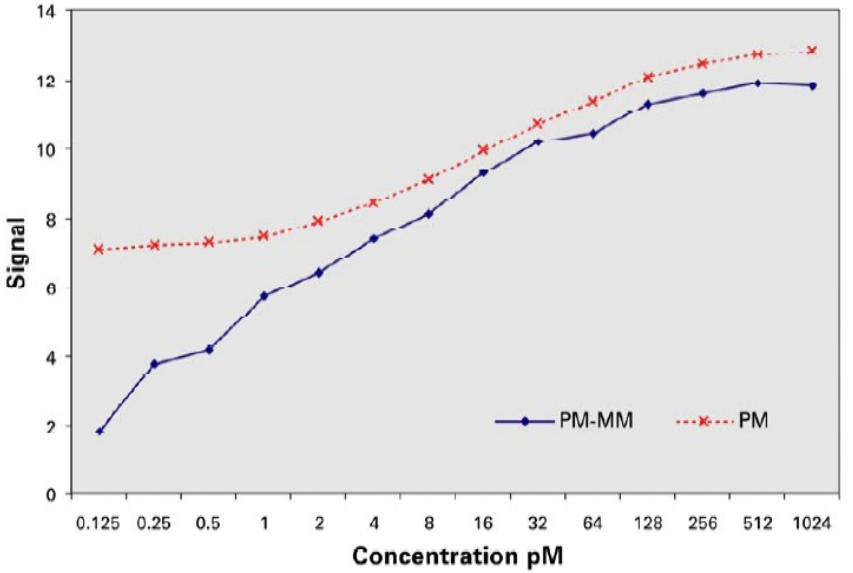
background intensity can be the mode value of the $\log_2$(MM) distribution for a given chip (kernel density estimate)

if PM <= background intensity, use ½ the minimum of $\log_2$(PM, background intensity) for PM > background intensity over all chips and probes

normalized values are log transformed because probe effects are additive on a log scale

  - Estimate RMA = a_i for chip *i* using Tukey's median polish procedure
    - *Iterative fitting, removing row and column medians, accumulating terms, until the process converges*

Sensitivity of PM only versus (PM-MM)[4]



Known concentrations of human transcripts were spiked at various concentrations into tissue samples where the transcripts were originally absent. Labeled samples were then hybridized to Affymetrix U95 microarrays. Hybridization intensities for each of the transcripts were calculated using both the PM and MM probes (solid lines) or the PM probes alone (dashed line) and then plotted against the RNA concentration. (Source = Affymetrix 2001d)

Figure 3. Comparison of the Assay Sensitivity using PM Probes Only or PM-MM Probe Pairs.

Normalization comparison criteria

The score components, along with the corresponding assessment *score number*, are as follows:

1. [1] *Median SD* - median SD across replicates in the dilution data
2. [2] *R2* - average R-squared over all pairs of replicates
3. [3] *1.25v20 corr* - correlation between expression of arrays hybridized to 1.25 micrograms and 20 micrograms of RNA
4. [6] *Median slope* - median slope obtained from regressing expression values on RNA concentrations in the dilution study
5. [7] *Signal detect slope* - slope obtained from regressing expression values on nominal concentrations in the spike-in data
6. [8] *Signal detect R2* - R-squared obtained from regressing expression values on nominal concentrations in the spike-in data
7. [12] *AUC (FP<100)* - area under the ROC curve up to 100 false positives
8. [13] *AFP, call if fc>2* - average false positives if we use fold-change > 2 as a cut-off
9. [14] *ATP, call if fc>2* - average true positives if we use fold-change > 2 as a cut-off
10. [15] *IQR* - interquartile range of log ratios among genes not differentially expressed
11. [16] *Obs-intended-fc slope* - slope obtained from regressing observed log-fold-changes against nominal log-fold-changes
12. [17] *Obs-(low)int-fc slope* - slope obtained from regressing observed log-fold-changes against nominal log-fold-changes for genes with nominal concentrations less than or equal to 2.
13. [21] *FC=2, AUC (FP<100)* - area under the ROC curve up to 100 false positives when comparing arrays with nominal fold changes of 2.
14. [22] *FC=2, AFP, call if fc>2* - average false positives if we use fold-change > 2 as a cut-off when comparing arrays where nominal fold-changes are 2
15. [23] *FC=2, ATP, call if fc>2* - average true positives if we use fold-change > 2 as a cut-off when comparing arrays where nominal fold-changes are 2

http://affycomp.biostat.jhsph.edu/AFFY/TABLES/0.html

Normalization comparison criteria

N	Method / Submitter	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	*(perfection)*	*0.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*0.00*	*16.00*	*0.00*	*1.00*	*1.00*	*1.00*	*0.00*	*16.00*
1	*MAS 5.0* / rafa	0.29	0.89	0.73	0.85	0.71	0.86	0.36	3108.99	12.82	2.66	0.69	0.65	0.07	3072.18	3.71
2	*RMA* / rafa	0.09	0.99	0.94	0.87	0.63	0.80	0.82	15.84	11.98	0.31	0.61	0.36	0.54	1.00	1.71
3	*dChip* / rafa	0.09	0.99	0.91	0.77	0.53	0.85	0.67	36.91	11.43	0.45	0.52	0.32	0.17	28.64	1.25
4	*ZAM2NBG* / magnus.astrand	0.07	0.99	0.94	0.72	0.57	0.77	0.84	2.44	11.70	0.24	0.57	0.32	0.61	0.57	1.14
5	*qm.p5* / cope	0.11	0.98	0.56	0.06	0.42	0.50	0.62	20.30	9.58	0.38	0.43	0.14	0.24	15.75	1.39
6	*vsn_scal* / w.huber	0.08	0.99	0.96	***1.00***	0.77	0.81	0.85	6.69	12.23	0.23	0.75	0.28	0.66	0.43	3.89
7	*vsn* / w.huber	0.06	0.99	0.96	0.67	0.51	0.81	0.85	0.40	10.83	0.15	0.50	0.19	0.66	0.21	1.11
8	*RMAVSN* / thomas.cappola	0.09	0.99	0.94	0.89	0.61	0.81	0.83	17.87	11.79	0.25	0.60	0.32	0.59	0.50	1.61
9	*RMA_NBG* / bolstad	0.04	1.00	0.91	0.56	0.48	0.81	0.85	0.13	10.45	0.12	0.47	0.15	0.68	***0.11***	1.04
10	*GSVDmin* / hzuzan	0.05	0.98	0.97	0.59	0.50	0.83	0.81	4.87	11.09	0.21	0.49	0.24	0.56	2.43	1.00
11	*PLIER* / Earl_Hubbell	0.13	0.09	0.01	0.84	0.71	0.91	0.02	596.77	12.85	4.03	0.72	0.65	0.02	589.96	3.57
12	*GSVDmod* / hzuzan	0.05	1.00	0.97	0.55	0.51	0.85	0.84	0.79	11.19	0.19	0.50	0.24	0.60	0.54	1.11
13	*PLIER+16* / Earl_Hubbell	0.08	0.99	0.88	0.64	0.65	***0.91***	0.81	8.42	12.34	0.34	0.65	0.46	0.46	5.07	2.04
14	*GCRMA* / zwu	0.09	0.99	0.89	0.72	***0.97***	0.84	0.82	7.62	12.97	0.35	0.92	0.66	0.54	7.07	5.29
15	*ChipMan* / plauren	0.31	0.99	0.94	1.26	0.88	0.82	0.67	183.99	***13.03***	0.67	0.87	0.44	0.20	159.86	5.11
16	*ProbePro* / shilmer	0.16	0.70	0.58	0.84	1.45	0.47	0.17	2087.07	12.53	15.70	1.33	1.93	0.07	2046.46	4.93
17	*MMEI* / shibing.deng	***0.02***	***1.00***	0.92	0.52	0.45	0.80	***0.86***	***0.12***	10.41	***0.12***	0.45	0.16	***0.69***	***0.11***	1.00
18	*PM* / zhangli	0.05	0.99	0.97	0.53	0.46	0.87	0.84	1.39	10.67	0.15	0.45	0.18	0.64	0.68	1.00
19	*RMA* / szeto	0.09	0.99	***0.98***	0.68	0.62	0.80	0.82	15.86	11.99	0.31	0.61	0.36	0.54	1.00	1.71
20	*GL* / mai98ftu	0.05	0.99	0.92	0.56	0.48	0.81	0.83	0.15	10.42	0.14	0.47	0.16	0.66	***0.11***	1.18
21	*MAS5+32* / Earl_Hubbell	0.07	0.98	0.93	0.71	0.60	0.88	0.72	20.56	11.76	0.51	0.59	0.33	0.18	19.18	1.68
22	*gMOS v.1* / m.milo	0.32	0.97	0.81	0.64	0.95	0.75	0.54	1358.01	12.75	2.15	***0.94***	***1.04***	0.10	1319.07	***5.36***
0	*(perfection)*	*0.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*0.00*	*16.00*	*0.00*	*1.00*	*1.00*	*1.00*	*0.00*	*16.00*
N	Method / Submitter	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

http://affycomp.biostat.jhsph.edu/AFFY/TABLES/0.html

References

- [1]Yang Y, Dudoit S, Luu P, and Speed T. Normalization for cDNA Microarray Data. (2000*) UC Berkeley Tech Report.*

- [2]Irizarry R, Bolstad B, Collin F, Cope L, Hobbs B, and Speed T. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acid Research*. **31.**

- [3]Dudoit, S., Gentleman, R., Irizarry, R., and Yang, Y. (2002) Pre-processing in DNA microarray experiments. *Bioconductor short course.*

- [4]*http://www.expressionanalysis.com/pdf/Affy-Platform-Comparison-Tech-Note.pdf*

R Code

```
# cDNA array plots
library(marrayInput)
library(marrayNorm)
library(marrayPlots)
library(sma)

# signal vs. noise plot for a single cDNA array
data(MouseArray)        # get mouse array data
plot.svb(mouse.data, "red",image.id=1,col='red',main='Singal vs. Noise for Cy5 channel on array #1')

# Examples use swirl dataset
data(swirl)

# look at image file from swirl data
maImage(swirl)

# look at boxplot from swirl data by print-tip
maBoxplot(swirl[,3])

# one form of an MvA plot
library(sma)

# mouse array
data(MouseArray)
plot.mva(mouse.data, mouse.setup, norm="l", 2, extra.type="pci",plot.type="n")

# Pre-normalization MvA-plot for the Swirl 93 array, with the lowess fits for
# individual print-tip-groups.
# - Default arguments
maPlot(swirl[,1],main='Print-tip Loess pre-normalization')

# Post-normalization using print-tip loess
mnorm<-maNorm(swirl[,1], norm="p", span=0.45)
maPlot(mnorm,main='Print-tip Loess post-normalization')
```

R Code

```
# import eisen data
dat <- read.table("eisen.txt",header=T)
dimnames(dat)[[1]] <- as.character(dat[,1])
dat <- dat[,-1]
dat <- as.data.frame(dat)

# scatter plot
cars.lm <- lm(dist~speed,data=cars)
plot(cars$speed,cars$dist,xlab="speed",ylab="dist",main="regression(cars)")
abline(as.numeric(cars.lm$coefficients[1]),as.numeric(cars.lm$coefficients[2]),col='red',lwd=2)

# lowess smoothing plot
data(cars)
plot(cars, main = "lowess(cars)")
lines(lowess(cars), col = 2,lwd=2)
lines(lowess(cars, f=.2), col = 3,lwd=2)
legend(5, 120, c(paste("f = ", c("2/3", ".2"))), lty = 1, col = 2:3)

# load affy library
library(affy)

 # get data
data(affybatch.example)

# plot data both before and after loess normalization using PM data
x <- pm(affybatch.example)
mva.pairs(x)
x <- normalize.loess(x,subset=1:nrow(x))
mva.pairs(x)
```

R Code

```
# affy normalization parameters for expresso function
> bgcorrect.methods
[1] "mas"  "none" "rma"  "rma2"

> normalize.AffyBatch.methods
[1] "constant" "contrasts" "invariantset" "loess"
[5] "qspline" "quantiles" "quantiles.robust"

> pmcorrect.methods
[1] "mas" "pmonly" "subtractmm"

> express.summary.stat.methods
[1] "avgdiff" "liwong" "mas" "medianpolish" "playerout"


eset <- expresso(affybatch.example,bgcorrect.method="rma",
            normalize.method="quantiles",
            pmcorrect.method="pmonly",
            summary.method="medianpolish")

# look at data frame of RMA values
attributes(eset)$exprs

# first scatter plot of R vs. G and un-normalized MvA plot with Mouse cDNA data
> plot(log(mouse.data$G),log(mouse.data$R),xlab='Cy3',ylab='Cy5',main='logR vs. logG')
> plot.mva(mouse.data, mouse.setup, norm="n", 2, extra.type="p",plot.type="r",main="MvA plot of R/G")
```