*Microarray Data Analysis*

# Normalization Methods for Analysis of Affymetrix GeneChip Microarray

中央研究院 生命科學圖書館

2008 年教育訓練課程

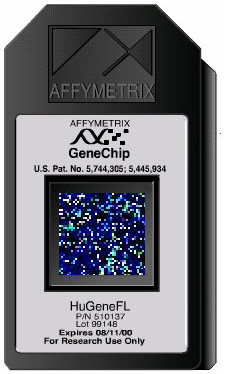
**2008/01/29**

吳漢銘
淡江大學 數學系
hmwu@math.tku.edu.tw
http://www.hmwu.idv.tw

# Outlines



- **Analysis Flow Chart**
- **Quality Assessment**
- **Low Level Analysis**
  **(from probe level data to expression value)**
- **Software**




**Affymetrix Dominates DNA Microarrays Market (75%~85%)**

http://www.gene2drug.com/about/archives.asp?newsId=180

假設:

1. 您對**Affymetrix GeneChip**已有一些了解。
2. 您對統計分析方法並不討厭。

# Overview of Microarray Analysis



## Affymetrix GeneChips Expression Analysis Process

1 RNA extraction

2 RNA labeling

3 Hybridization

5 Washing Staining

6 Image Acquistion

Image Processing Normalization

### Oligonucleotide Array Data

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Probeset | Gene Name | Array 1 | Array 2 |
| 2 | 103941_at | alpha-spectin 1, erythroid | 33.7625 | 29.2333 |
| 3 | 104432_at | aplysia ras-related homolog N (Rh | 127.736 | 99.6895 |
| 4 | 104137_at | ATP-binding cassette, sub-family A | 109.522 | 65.2727 |
| 5 | 98458_at | baculoviral IAP repeal-containing 5 | 128.96 | 123.371 |
| 6 | 93243_at | bone morphogenetic protein 7 | 174.85 | 174.019 |
| 7 | 95061_at | breast carcinoma amplified sequer | 34.8 | 43.6696 |
| ... | | | | |
| 12600 | 102632_at | calmodulin binding protein 1 | 69.888 | 54.7391 |

Expression Index

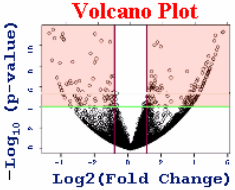
Matrix of genes (rows) and samples (columns)

### Discovery of differentially expressed genes

**Parametric** : t-test

**Non-parametric** : Wilcoxon, Mann-Whitney test

Volcano Plot

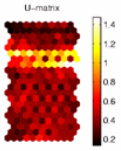


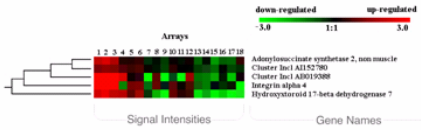

### Unsupervised: clustering

Hierarchical clustering

K-means clustering

Self-organizing maps

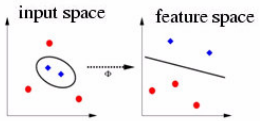



### Supervised: classification

- Linear discriminants
- Decision trees
- Support vector machines

Support Vector Classifiers

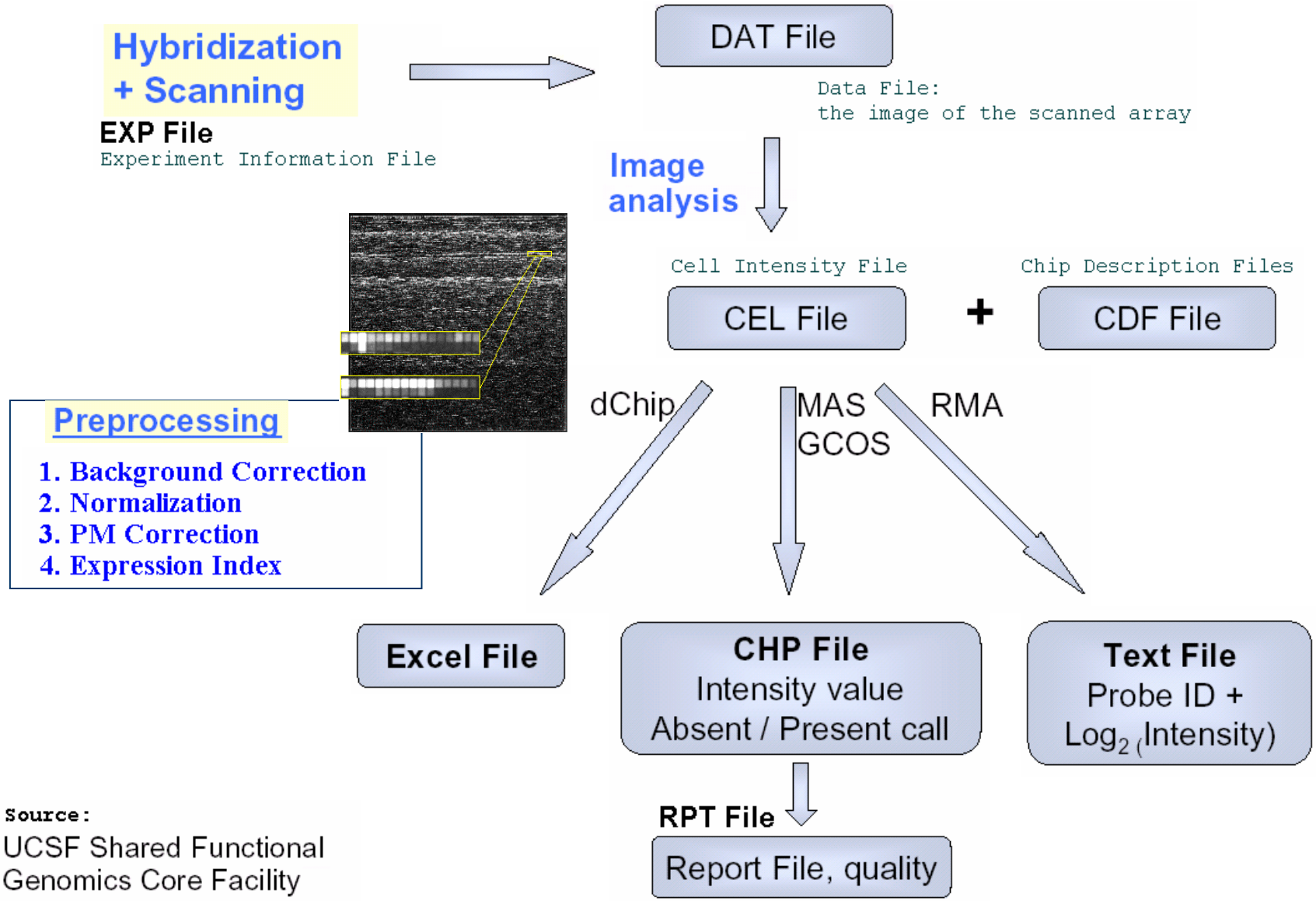



Boser, Guyon, and Vapnik (1992)

Biological Relevance

# Analysis Flow Chart






Source:
UCSF Shared Functional
Genomics Core Facility

# Affymetrix Data Files



*.EXP file

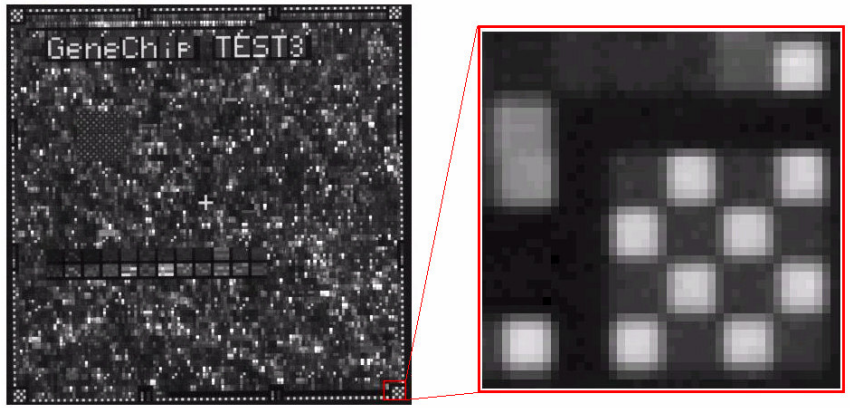
```
Affymetrix GeneChip Experiment Information
Version 1

[Sample Info]
Chip Type        HG-U133A
Chip Lot
Operator         array
Sample Type      RNA
Description
Project Dr. Mi
Comments
Solution Type
Solution Lot

[Fluidics]
Protocol         EukGE-WS2v4
Completed
Station 1
Module  2
Hybridize Date   Oct 19 2004 01:17PM

[Scanner]
Pixel Size       3
Filter  570
Scan Temperature
Scan Date        Oct 19 2004 01:41PM
Scanner ID
Number of Scans 2
Scanner Type     HP
```

*.DAT file ~50MB




*.CEL file ~12MB

(Version 4) ~5MB

```
@■■■ ■■■?■■?■■@?■?■■Cols=712 Rows=712 TotalX=712 TotalY=712 OffsetX=0 O
ffsetY=0 GridCornerUL=230 231 GridCornerUR=4503 235 GridCornerLR=4499 4506
GridCornerLL=226 4502 Axis-invertX=0 AxisInvertY=0 swapXY=0 DatHeader=[9..4
6155]  7:CLS=4733 RWS=4733 XIN=3  YIN=3  VE=17        2.0 02/24/04 13:41:05
   HP        HG-U133A.1sq                          6 Algorithm=Percentile
AlgorithmParameters=Percentile:75;CellMargin:2;OutlierHigh:1.500;OutlierLow
:1.004  ■■■Percentile>■■■Percentile=75 CellMargin=2 OutlierHigh=1.500 Outli
erLow=1.004  ■■■?■■■■■■■■■■■■噴6zkA ■■<培zB4D ■■■窒t訟A ■■?E?AD ■■■  ?J
A ■■■嫊P9HA ■■畹EF?D ■■■劁??A ■■拎E O D ■■■澌w⑯A ■■ rE澹 D ■■■哷v_ A ■■
■EhG)D ■■■愖浭4A ■■4mE? D ■■■趨 ?A ■■ iE h D ■■■譒!覝A ■■xfE? D ■■■嫊I鞙
```

⋮

```
[CEL]
Version=3

[HEADER]
Cols=712
Rows=712
TotalX=712
TotalY=712
OffsetX=0
OffsetY=0
GridCornerUL=230 231
GridCornerUR=4503 235
GridCornerLR=4499 4506
GridCornerLL=226 4502
Axis-invertX=0
AxisInvertY=0
swapXY=0
DatHeader=[9..46155]  7:CLS=4733 RWS=4733 XIN=3  YIN=3  VE=17        2.0 02/24/04 13:41:05   HP
Algorithm=Percentile
AlgorithmParameters=Percentile:75;CellMargin:2;OutlierHigh:1.500;OutlierLow:1.004

[INTENSITY]
NumberCells=506944
CellHeader=X    Y       MEAN    STDV    NPIXELS
           0    0       114.5   14.7    16
           1    0       4711.5  721.0   16
           2    0       111.8   13.9    16
```

⋮

CEL File Conversion Tool

# CDF file



## Chip Description File (E.g., HG-U133_Plus_2.cdf)

```
[CDF]
Version=GC3.0

[Chip]
Name=HG-U133_Plus_2
Rows=1164
Cols=1164
NumberOfUnits=54675
MaxUnit=59076
NumQCUnits=9
ChipReference=

[QC1]
Type=15
NumberCells=2280
CellHeader=X     Y
Cell1=6 0        N
Cell2=8 0        N
Cell3=10         0
Cell4=12         0
Cell5=14         0
Cell6=16         0
Cell7=18         0
Cell8=20         0
Cell9=22         0
Cell10=24        0
Cell11=26        0
Cell12=28        0
Cell13=30        0
Cell14=32        0
Cell15=34        0
Cell16=36        0
Cell17=38        0
Cell18=40        0
Cell19=42        0
Cell20=44        0
Cell21=46        0
Cell22=48        0
Cell23=50        0
Cell24=52        0
Cell25=54        0
Cell26=56        0
Cell27=58        0
Cell28=60        0
Cell29=62        0
Cell30=64        0
```

```
N          25          0          48
N          25          0          50
N          25          0          52
N          25          0          54
N          25          0          56
N          25          0          58
N          25          0          60
N          25          0          62
N          25          0          64
```

```
[Unit1003_Block1]
Name=121_at
BlockNumber=1
NumAtoms=16
NumCells=32
StartPosition=0
StopPosition=15
CellHeader=X    Y       PROBE   FEAT    QUAL    EXPOS   POS     CBASE   PBASE   TBASE   ATOM    INDEX   CODONIND        CODON   REGIONTYPE      REGION
Cell1=656       1012    N       control 121_at  0       13      A       A       A       0       1178624 -1      -1      99
Cell2=656       1011    N       control 121_at  0       13      A       T       A       0       1177460 -1      -1      99
Cell3=1079      93      N       control 121_at  1       13      G       C       G       1       109331  -1      -1      99
Cell4=1079      94      N       control 121_at  1       13      G       G       G       1       110495  -1      -1      99
Cell5=760       940     N       control 121_at  2       13      A       A       A       2       1094920 -1      -1      99
Cell6=760       939     N       control 121_at  2       13      A       T       A       2       1093756 -1      -1      99
Cell7=575       983     N       control 121_at  3       13      G       C       G       3       1144787 -1      -1      99
Cell8=575       984     N       control 121_at  3       13      G       G       G       3       1145951 -1      -1      99
Cell9=122       325     N       control 121_at  4       13      T       A       T       4       378422  -1      -1      99
Cell10=122      326     N       control 121_at  4       13      T       T       T       4       379586  -1      -1      99
Cell11=806      148     N       control 121_at  5       13      A       A       A       5       173078  -1      -1      99
Cell12=806      147     N       control 121_at  5       13      A       T       A       5       171914  -1      -1      99
Cell13=476      65      N       control 121_at  6       13      G       C       G       6       76136   -1      -1      99
Cell14=476      66      N       control 121_at  6       13      G       G       G       6       77300   -1      -1      99
Cell15=922      26      N       control 121_at  7       13      A       A       A       7       31186   -1      -1      99
Cell16=922      25      N       control 121_at  7       13      A       T       A       7       30022   -1      -1      99
Cell17=232      435     N       control 121_at  8       13      T       A       T       8       506572  -1      -1      99
Cell18=232      436     N       control 121_at  8       13      T       T       T       8       507736  -1      -1      99
Cell19=791      176     N       control 121_at  9       13      C       C       C       9       205655  -1      -1      99
Cell20=791      175     N       control 121_at  9       13      C       G       C       9       204491  -1      -1      99
Cell21=928      652     N       control 121_at  10      13      A       A       A       10      759856  -1      -1      99
Cell22=928      651     N       control 121_at  10      13      A       T       A       10      758692  -1      -1      99
Cell23=268      274     N       control 121_at  11      13      A       A       A       11      319204  -1      -1      99
```

# Quality Assessment

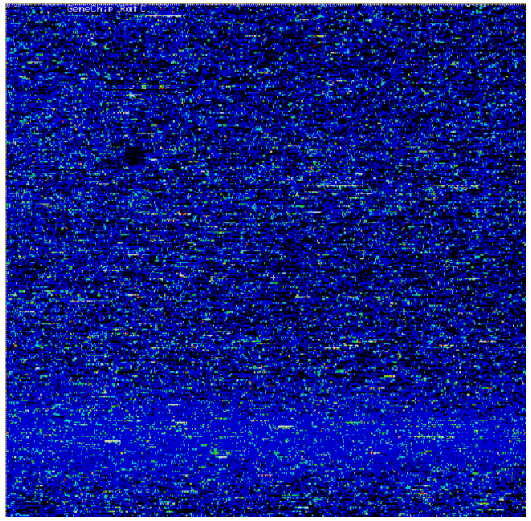
- Array Image Inspection
- RNA Degradation Plots
- MAS5.0 Expression Report File (*.RPT)
- Statistical Quality Control (Diagnostic Plots)
- QC Reference

# Probe Array Image Inspection

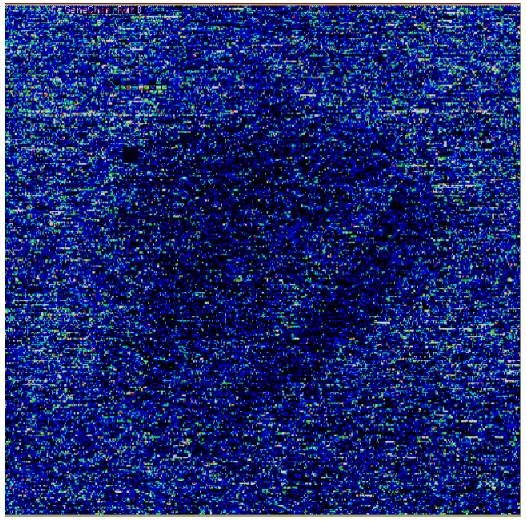


- Saturation: PM or MM cells > 46000
- Defect Classes:
  dimness/brightness, high Background, high/low intensity spots, scratches, high regional, overall background, unevenness, spots, Haze band, scratches, crop circle, cracked, cnow, grid misalignment.
- As long as these areas do not represent more than 10% of the total probes for the chip, then the area can be masked and the data points thrown out as outliers.

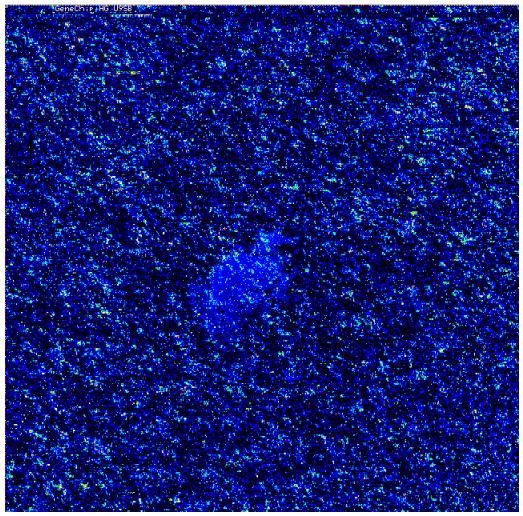
Haze Band



Crop Circles



Spots, Scratches, etc.



Source: Michael Elashoff (GLGC)

# Probe Array Image Inspection (conti.)



Li, C. and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, Proc. Natl. Acad. Sci. Vol. 98, 31-36.

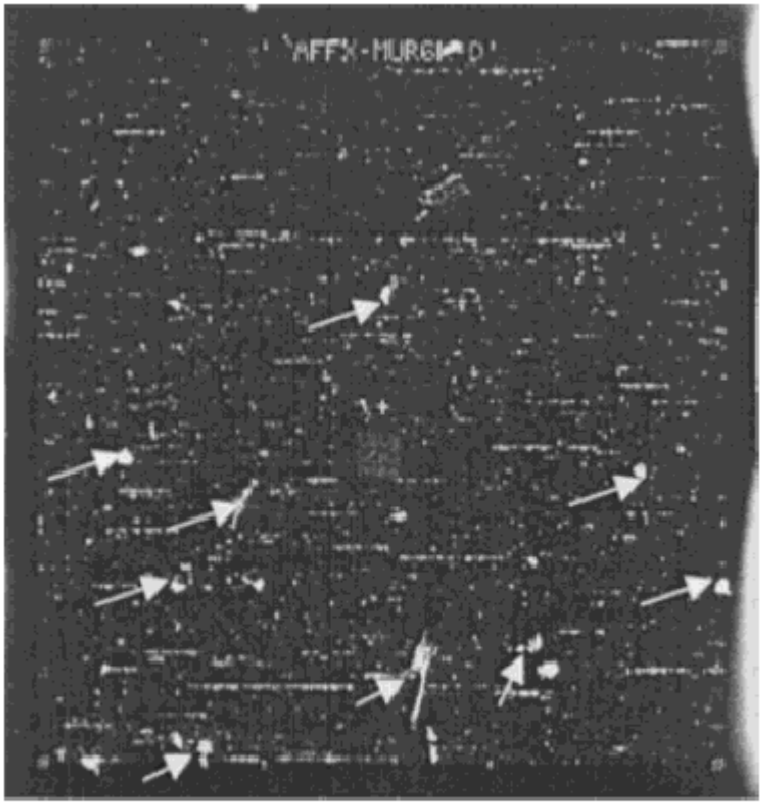


**Fig. 1.** A contaminated D array from the Murine 6500 Affymetrix GeneChip® set. Several particles are highlighted by arrows and are thought to be torn pieces of the chip cartridge septum, potentially resulting from repeatedly pipetting the target into the array.

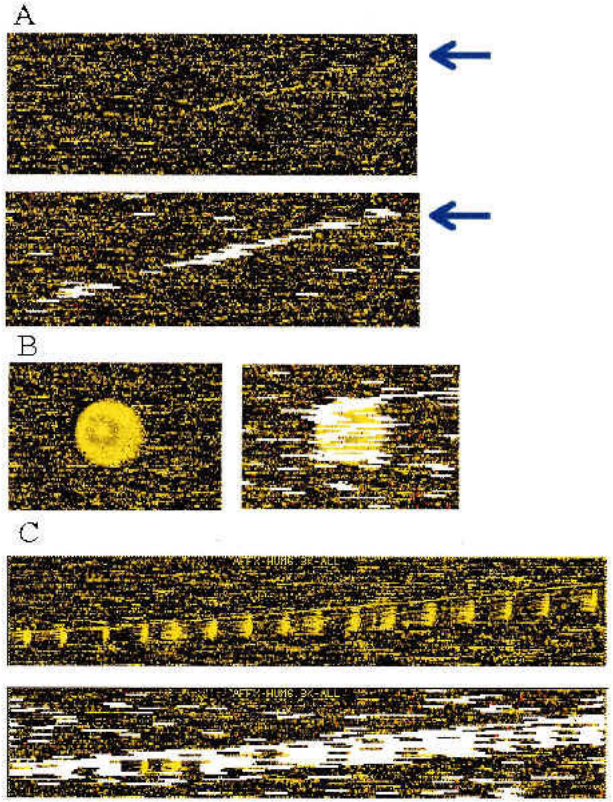
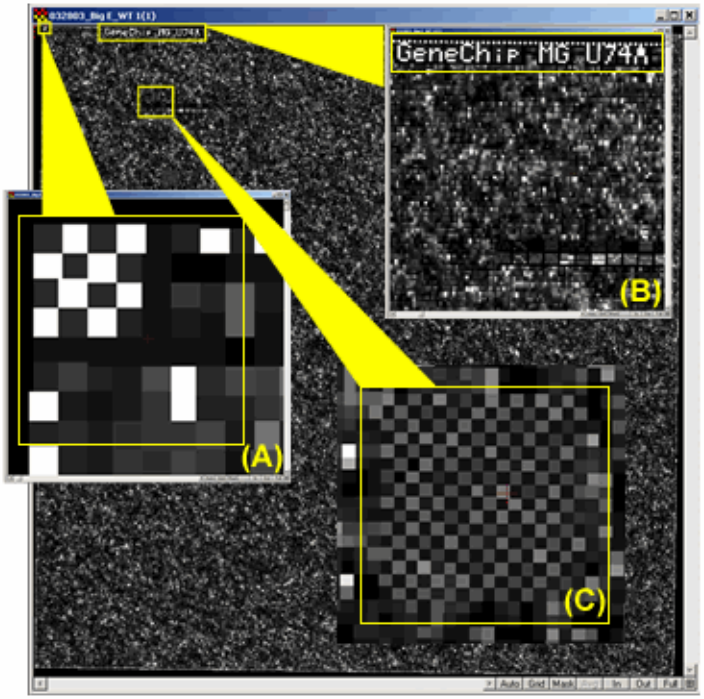



**Fig. 5.** (*A*) A long scratch contamination (indicated by arrow) is alleviated by automatic outlier exclusion along this scratch. (*B* and *C*) Regional clustering of array outliers (white bars) indicates contaminated regions in the original images. These outliers are automatically detected and accommodated in the analysis. Note that some probe sets in the contaminated region are not marked as array outliers, because contamination contributed additively to PM and MM in a similar magnitude and thus cancel in the PM–MM differences, preserving the correct signals and probe patterns.

# B2 Oligo Performance



- Make sure the **alignment** of the grid was done appropriately.
- Look at the spiked in Oligo B2 control in order to check the **hybridization uniformity**.
- The border around the array, the corner region, the control regions in the center, are all checked to make sure the **hybridization** was successful.



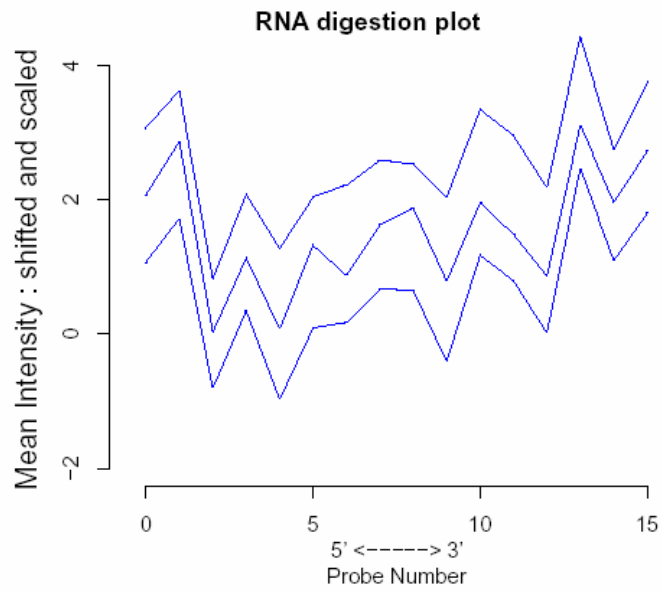

**Affymetrix CEL File Image- Yellow squares highlighting various Oligo B2 control regions: (A) one of the corner regions, (B) the name of the array, and (C) the "checkerboard" region.**

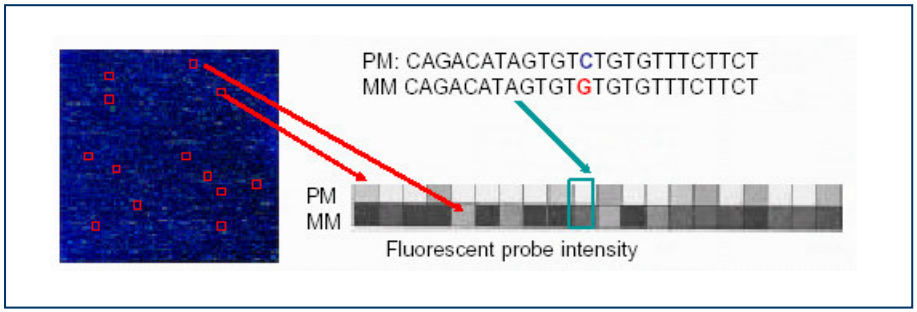
Source: Baylor College of Medicine, Microarray Core Facility

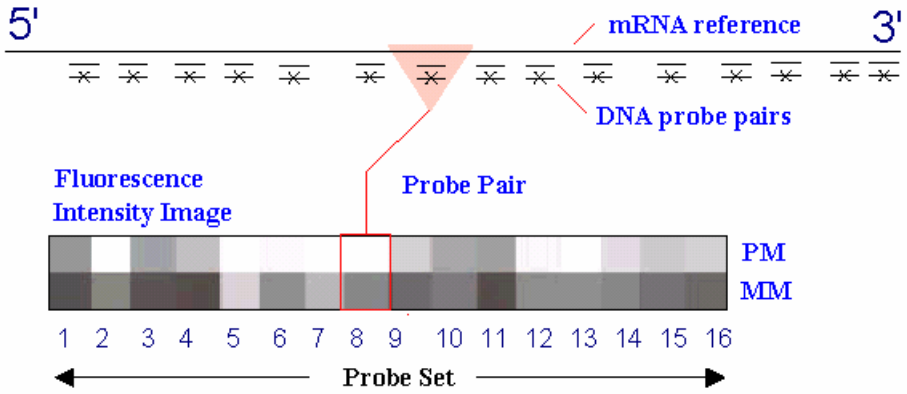
# RNA Degradation Plots

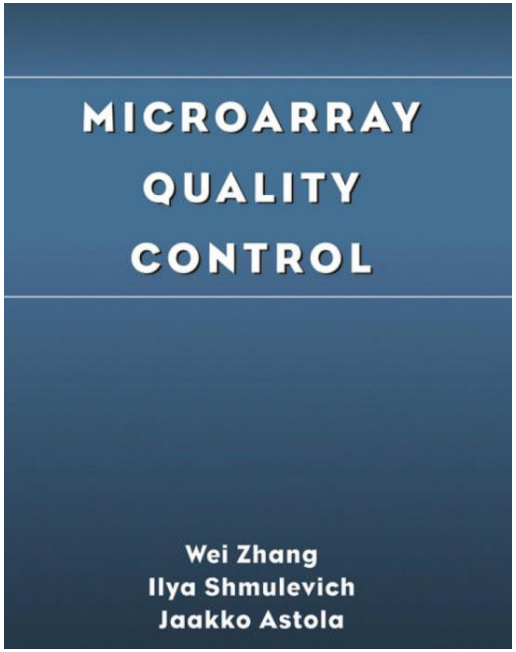


## Assessment of RNA Quality:




5'----------3'AAAA-Poly

x-------|---------

# MAS5.0 Expression Report File (*.RPT)



```
Report Type:            Expression Report
Date:                    04:42PM 02/24/2004
-------------------------------------------------------------------------
Filename:                     test.CHP
Probe Array Type:        HG-U133A
Algorithm:               Statistical
Probe Pair Thr:          8
Controls:                Antisense
-------------------------------------------------------------------------
Alpha1:                  0.05
Alpha2:                  0.065
Tau:                     0.015
Noise (RawQ):            2.250
Scale Factor (SF):       5.422
TGT Value:               500
Norm Factor (NF):        1.000
-------------------------------------------------------------------------
Background:
        Avg: 64.23       Std: 1.75       Min: 59.50      Max: 67.70
Noise:
        Avg: 2.54        Std: 0.14       Min: 2.10       Max: 3.00
Corner+
        Avg: 49          Count: 32
Corner-
        Avg: 5377        Count: 32
Central-
        Avg: 4845        Count: 9
-------------------------------------------------------------------------
The following data represents probe sets that exceed the probe pair threshold
and are not called "No Call".

Total Probe Sets:        22283
Number Present:          9132     41.0%
Number Absent:           12766    57.3%
Number Marginal:         385      1.7%

Average Signal (P):      1671.0
Average Signal (A):      119.6
Average Signal (M):      350.1
Average Signal (All):    759.3
```

- The Scaling Factor- In general, the scaling factor should be around three, but as long as it is not greater than five, the chip should be okay.
- The scaling factor (SF) should remain consistent across the experiment.

- Average Background: 20-100
- Noise < 4

- The measure of Noise (RawQ), Average Background and Average Noise values should remain consistent across the experiment.

- Percent Present : 30~50%, 40~50%, 50~70%. (should be consistent)
- Low percent present may also indicate degradation or incomplete synthesis.

# MAS5.0 Expression Report File (*.RPT)



- Sig (3'/5')- This is a ratio which tells us how well the labeling reaction went. The two to really look at are your 3'/5' ratio for GAPDH (around 1) and B-ACTIN (around 3).


- Spike-In Controls (BioB, BioC, BioD, Cre)- These spike in controls also tell how well your labelling reaction went. BioB is only Present half of the time, but BioC, BioD, & Cre should always have a present (P) call.

Housekeeping Controls:

| Probe Set | Sig(5') | Det(5') | Sig(M') | Det(M') | Sig(3') | Det(3') | Sig(all) | Sig(3'/5') |
|---|---|---|---|---|---|---|---|---|
| AFFX-HUMISGF3A/M97935 | 272.8 | P | 856.8 | P | 1274.5 | P | 801.36 | 4.67 |
| AFFX-HUMRGE/M10098 | 340.6 | M | 181.3 | A | 632.6 | P | 384.80 | 1.86 |
| AFFX-HUMGAPDH/M33197 | 13890.6 | P | 15366.6 | P | 14060.7 | P | 14439.32 | 1.01 |
| AFFX-HSAC07/X00351 | 35496.8 | P | 39138.0 | P | 31375.0 | P | 35336.61 | 0.88 |
| AFFX-M27830 | 469.2 | P | 2206.1 | A | 114.3 | A | 929.86 | 0.24 |

Spike Controls:

| Probe Set | Sig(5') | Det(5') | Sig(M') | Det(M') | Sig(3') | Det(3') | Sig(all) | Sig(3'/5') |
|---|---|---|---|---|---|---|---|---|
| AFFX-BIOB | 559.0 | P | 801.6 | P | 385.8 | P | 582.14 | 0.69 |
| AFFX-BIOC | 1132.9 | P | | | 818.0 | P | 975.47 | 0.72 |
| AFFX-BIOD | 874.7 | P | | | 6918.1 | P | 3896.42 | 7.91 |
| AFFX-CRE | 10070.5 | P | | | 16198.0 | P | 13134.27 | 1.61 |
| AFFX-DAP | 10.9 | A | 60.9 | A | 8.5 | A | 26.75 | 0.78 |
| AFFX-LYS | 51.5 | A | 86.2 | A | 14.1 | A | 50.62 | 0.27 |
| AFFX-PHE | 4.9 | A | 4.0 | A | 40.0 | A | 16.30 | 8.20 |
| AFFX-THR | 20.3 | A | 53.2 | A | 18.7 | A | 30.77 | 0.92 |
| AFFX-TRP | 9.8 | A | 11.1 | A | 2.7 | A | 7.86 | 0.28 |
| AFFX-R2-EC-BIOB | 497.6 | P | 928.0 | P | 479.4 | P | 634.98 | 0.96 |
| AFFX-R2-EC-BIOC | 1319.9 | P | | | 1705.0 | P | 1512.50 | 1.29 |
| AFFX-R2-EC-BIOD | 4744.0 | P | | | 4865.7 | P | 4804.82 | 1.03 |
| AFFX-R2-P1-CRE | 25429.2 | P | | | 30469.5 | P | 27949.37 | 1.20 |
| AFFX-R2-BS-DAP | 5.9 | A | 1.6 | A | 3.3 | A | 3.58 | 0.55 |
| AFFX-R2-BS-LYS | 32.2 | A | 43.7 | M | 74.7 | P | 50.18 | 2.32 |
| AFFX-R2-BS-PHE | 14.8 | A | 27.5 | A | 146.5 | A | 62.91 | 9.93 |
| AFFX-R2-BS-THR | 209.5 | P | 152.9 | A | 15.8 | A | 126.08 | 0.08 |

# Suggestions



- Affymetrix arrays with high **background** are more likely to be of poor quality.
  - Cutoff would be to exclude arrays with a value more than 100.

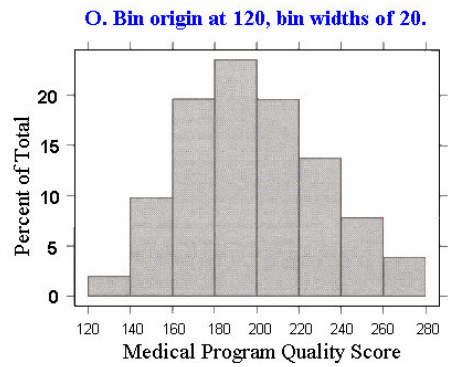
- **Raw noise score (Q)**: a measure of the variability of the pixel values within a probe cell averaged over all of the probe cells on an array.
  - Exclude those arrays that have an unusually high Q-value relative to other arrays that were processed with the same scanner.

- **BioB**: is included at a concentration that is close to the level of detection of the array, and so should be indicated as present about 50% of the time.
- Other spike controls are included at increasingly greater levels of concentration. Therefore, they should all be indicated as present, and also should have increasingly large signal values:
  - Signal(bioB) < Signal(bioC) < Signal(bioD) < Signal(cre)

# Statistical Plots: Histogram



- 1/2h adjusts the height of each bar so that the total area enclosed by the entire histogram is 1.
- The area covered by each bar can be interpreted as the probability of an observation falling within that bar.




## Disadvantage for displaying a variable's distribution:

- selection of origin of the bins.
- selection of bin widths.
- the very use of the bins is a distortion of information because any data variability within the bins cannot be displayed in the histogram.

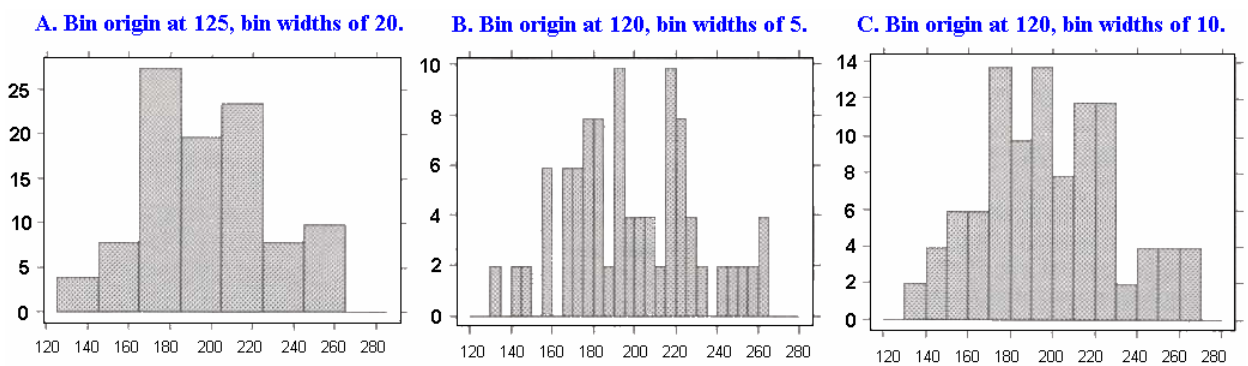
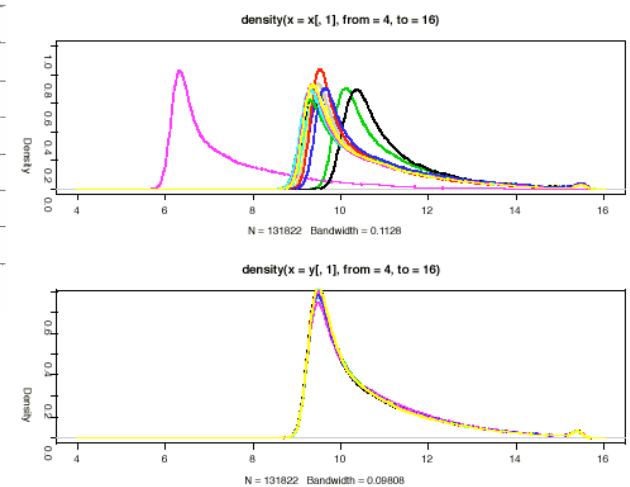



Figure Sources: Jacoby (1997).

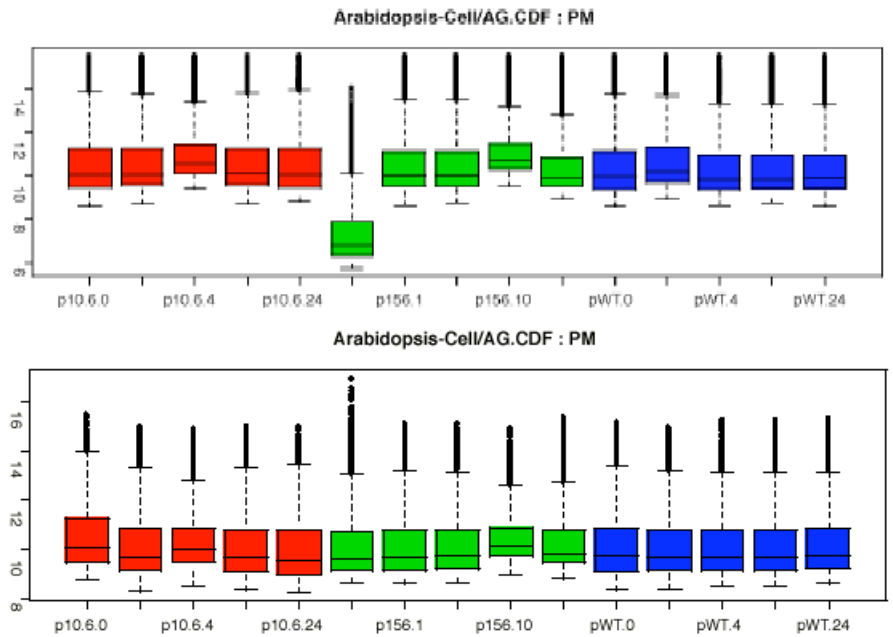
Density Plots

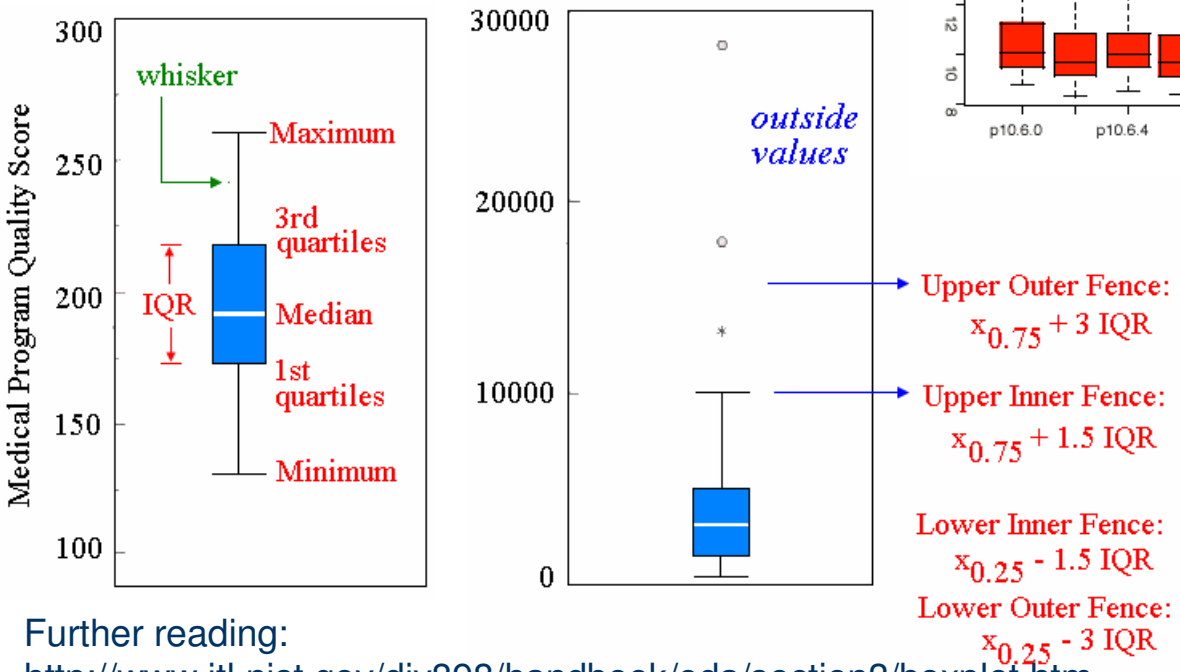


# Statistical Plots: Box Plots



- Box plots (Tukey 1977, Chambers 1983) are an excellent tool for conveying **location and variation** information in data sets.
- For detecting and illustrating location and variation changes between different groups of data.




**The box plot can provide answers to the following questions:**

- Is a factor significant?
- Does the location differ between subgroups?
- Does the variation differ between subgroups?
- Are there any outliers?

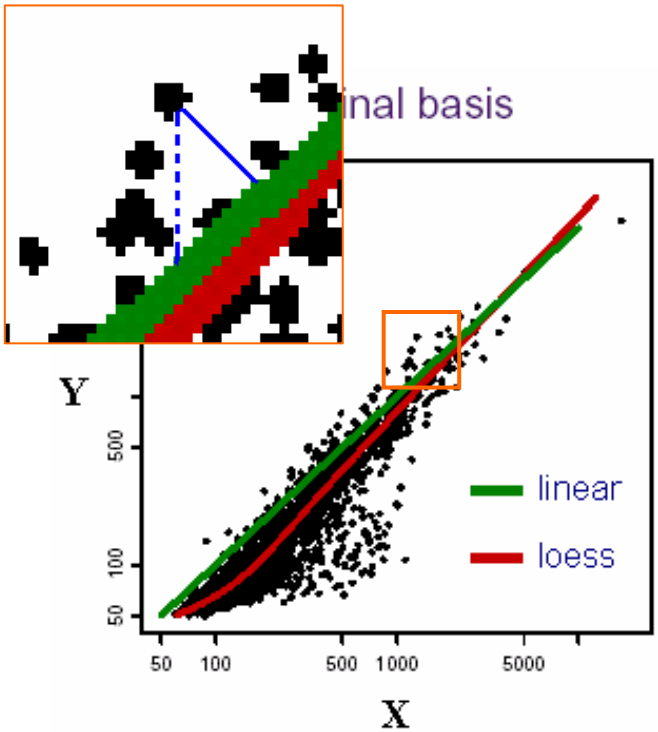
Further reading:
http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm

# Scatterplot and MA plot



- **Features of scatterplot.**
  - the substantial **correlation** between the expression values in the two conditions being compared.
  - the preponderance of low-intensity values.
    (the majority of genes are expressed at only a low level, and relatively few genes are expressed at a high level)

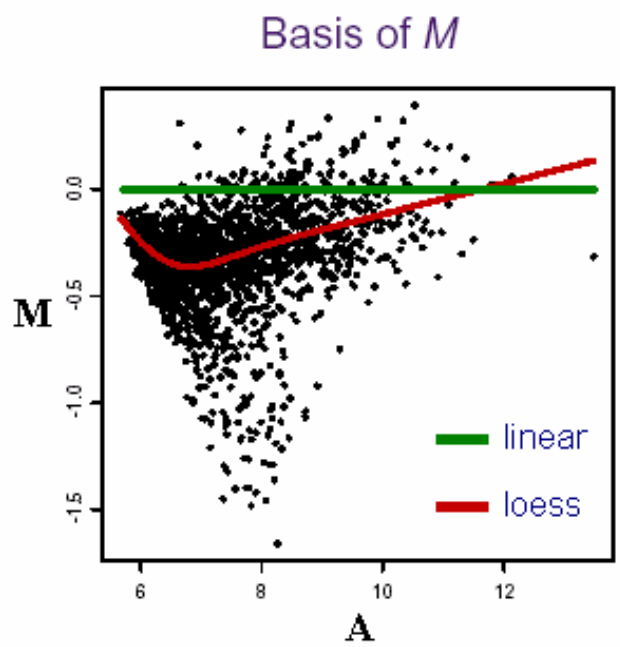
- **Goals:** to identify genes that are differentially regulated between two experimental conditions.




$$M = \log_2\left(\frac{Y}{X}\right)$$

$$A = \frac{1}{2}\log_2(X\,Y)$$

| Oligo | cDNA |
|---|---|
| $X = PM_1$, $Y = PM_2$ | X= Cy3, Y= Cy5 |
| $X = PM_1 - MM_1$, $Y = PM_2 - MM_2$ | |

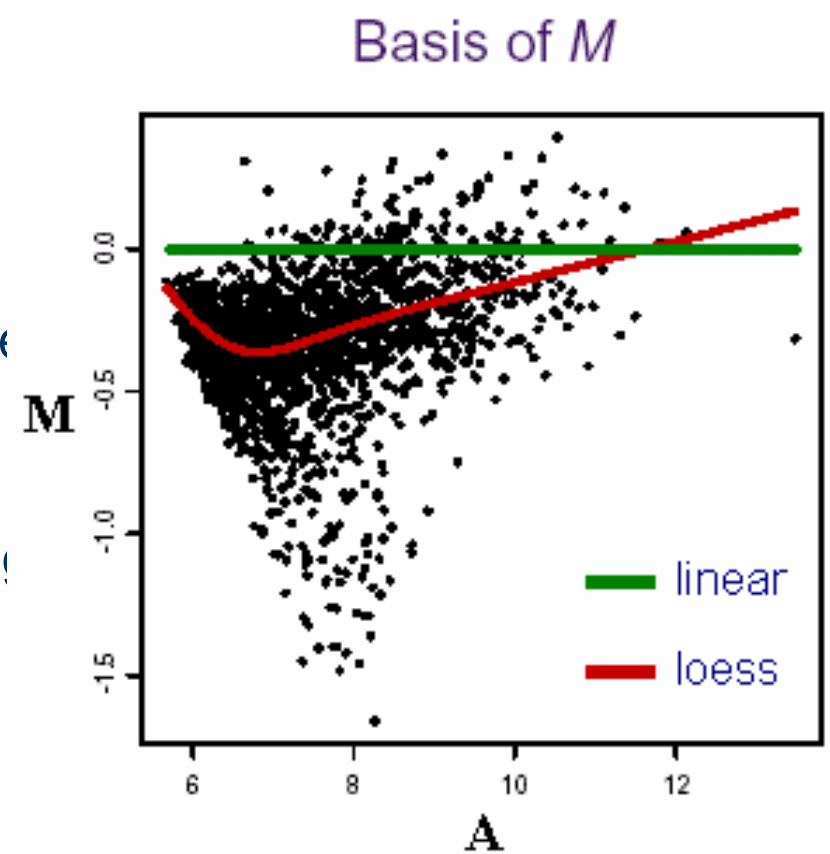


# Scatterplot and MA plot (conti.)



- **MA plots** can show the intensity-dependant ratio of raw microarray data.
  - x-axis (mean log2 intensity): average intensity of a particular element across the control and experimental conditions.
  - y-axis (ratio): ratio of the two intensities. (fold change)

- **Outliers in logarithm scale**
  - spreads the data from the lower left corner to a more centered distribution in which the prosperities of the data are easy to analyze.
  - easier to describe the fold regulation of genes using a log scale. In log2 space, the data points are symmetric about 0.



# MAQC project



NCTR Center for Toxicoinformatics - MAQC Project - Windows Internet Explorer

http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/

U.S. Food and Drug Administration

NATIONAL CENTER FOR TOXICOLOGICAL RESEARCH

FDA Home Page | NCTR Home | NCTR Initiatives | NCTR Science | NCTR Site Map

*MicroArray Quality Control (MAQC) Project*

| Toxicoinformatics Home | MAQC Home | Working Groups | Presentations | Participating Organizations | Study Guidance |

**Executive Summary**

[Word] [PDF]

**Purpose**

The purpose of the MAQC project is to provide quality control tools to the microarray community in order to avoid procedural failures and to develop guidelines for microarray data analysis by providing the public with large reference datasets along with readily accessible reference RNA samples.

**Project Description**

FDA's Critical Path Initiative identifies pharmacogenomics and toxicogenomics as key opportunities in advancing medical product development and personalized medicine, and the "Guidance for Industry: Pharmacogenomic Data Submissions" [PDF] [WORD] has been released. Microarrays represent a core technology in pharmacogenomics and toxicogenomics; however, before this technology can successfully and reliably be used in clinical practice and regulatory decision-making, standards and quality measures need to be developed.

The MicroArray Quality Control (MAQC) project involves six FDA Centers, major providers of microarray platforms and RNA samples, EPA, NIST, academic laboratories, and other stakeholders. The MAQC project aims to establish QC metrics and thresholds for objectively assessing the performance achievable by various microarray platforms and evaluating the advantages and disadvantages of various

MAQC Consortium, 2006, The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnology 24(9):1151-61.

# QC Reference




- G. V. Cohen Freue, Z. Hollander, E. Shen, R. H. Zamar, R. Balshaw, A. Scherer, B. McManus, P. Keown, W. R. McMaster, and R. T. Ng, 2007, **MDQC**: a new quality assessment method for microarrays based on quality control reports, Bioinformatics 23(23): 3162 - 3169.

- Steffen heber and Beate Sick, 2006, Quality Assessment of Affymetrix GeneChip Data, OMICS A Journal of Integrative Biology, Volume 10, Number 3, 358-368.
- Kyoungmi Kim , Grier P Page , T Mark Beasley , Stephen Barnes , Katherine E Scheirer and David B Allison, 2006, A proposed metric for assessing the measurement quality of individual microarrays, BMC Bioinformatics 7:35.

- Claire L. Wilson and Crispin J. Miller, 2005, **Simpleaffy**: a BioConductor package for Affymetrix Quality Control and data analysis, Bioinformatics 21: 3683 - 3685.

- **affyQCReport**: A Package to Generate QC Reports for Affymetrix Array Data
- **affyPLM**: Model Based QC Assessment of Affymetrix GeneChips


**Red color: R package at Bioconductor.**

# Low level Analysis

- Background correction (local vs. global)
- Normalization (baseline array vs. complete data)
- PM Correction
- Summarization [Expression Index] (single vs. multiple chips)

# Low level analysis



| Background Methods | Normalization Methods | PM correction Methods | Summarization Methods |
|---|---|---|---|
| none | quantiles | mas | avgdiff |
| rma/rma2 | loess | pmonly | liwong |
| mas | contrasts | subtractmm | mas |
| | constant | | medianpolish |
| | invariantset | | playerout |
| | Qspline | | |

The Bioconductor: affy package

- MAS5

```
eset.mas5 <- expresso(Data, bg.correct="mas", normalize.method = "constant",
                      pmcorrect.method="mas", summary.method="mas")
```

- Liwong (PM-only Model)

```
eset.liwong <- expresso(Data, bg.correct=FALSE, normalize.method = "invariantset",
                        pmcorrect.method="pmonly", summary.method="liwong")
```

- Liwong (PM-MM Model)

```
eset.liwong <- expresso(Data, bg.correct=FALSE, normalize.method = "invariantset",
                        pmcorrect.method="subtractmm ", summary.method="liwong")
```

- RMA

```
eset.rma <- expresso(Data, bg.correct="rma", normalize.method = "quantiles",
                     pmcorrect.method="pmonly", summary.method="medianpolish")
```

- Other

```
eset <- expresso(Data, bg.correct="mas", normalize.method="qspline",
                 pmcorrect.method="subtractmm", summary.method="playerout")
```

# 1. Background Correction

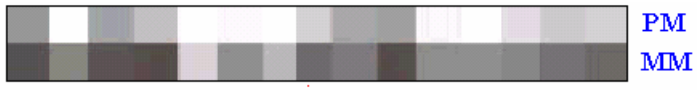


## What is background?

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

## What is background correction?

- A method for removing background noise from signal intensities using information from only one chip.



# 2. Normalization



## What is Normalization?

- ***Non-biological factor*** 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.

**Sources of Variation**

amount of RNA in the biopsy efficiencies of

- RNA extraction
- reverse transcription
- labeling
- photodetection

PCR yield
DNA quality
Spotting efficiency, spot size
cross- or unspecific-hybridization
stray signal

**Systematic** → **Normalization**

- similar effect on many measurements
- corrections can be estimated from data

**Stochastic** → **Error Model**

- too random to be explicitely accounted for
- noise

Systematic

- Amount of RNA in biopsy extraction, Efficiencies of RNA extraction, reverse transcription, labeling, photodetection, GC content of probes
- Similar effect on many measurements
- Corrections can be estimated from data
- Calibration corrections

Stochastic

- PCR yield, DNA quality, Spotting efficiency, spot size,
- Non-specific hybridization, Stray signal
- Too random to be explicitly accounted for in a model
- Noise components & “Schmutz” (dirt)

# Why Normalization?



Normalization corrects for overall chip brightness and other factors that may influence the numerical value of expression intensity, enabling the user to more confidently compare gene expression estimates between samples.

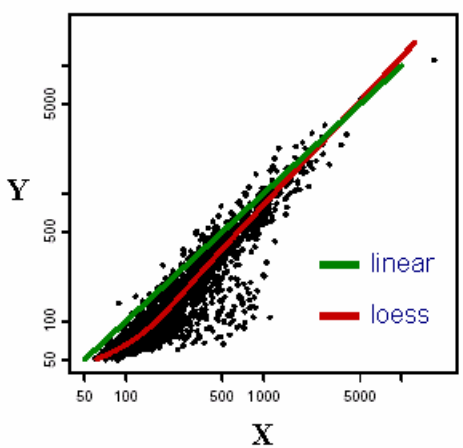
***Main idea***

Remove the systematic bias in the data as completely possible while preserving the variation in the gene expression that occurs because of biologically relevant changes in transcription.

***Assumption***

- The average gene does not change in its expression level in the biological sample being tested.
- Most genes are not differentially expressed or up- and down-regulated genes roughly cancel out the expression effect.

Original basis

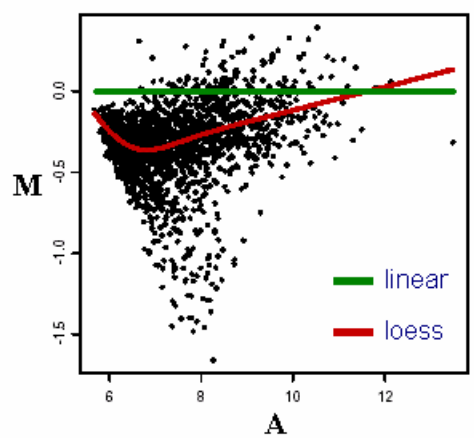



$$M = \log_2\left(\frac{Y}{X}\right)$$

$$A = \frac{1}{2}\log_2(X\,Y)$$

| Oligo | cDNA |
|---|---|
| $X = PM_1$, $Y = PM_2$ | X= Cy3, Y= Cy5 |
| $X = PM_1 - MM_1$, $Y = PM_2 - MM_2$ | |

Basis of *M*



# The Options on Normalization



- **Levels**
  - PM&MM, PM-MM, Expression indexes
- **Features**
  - All, Rank invariant set, Spike-ins, housekeeping genes.
- **Methods**
  - Complete data: no reference chip, information from all arrays used: Quantiles Normalization, MVA Plot + Loess
  - Baseline: normalized using reference chip: MAS 4.0, MAS 5.0, Li-Wong's Model-Based, Qspline

# Constant Normalization



## Normalization and Scaling

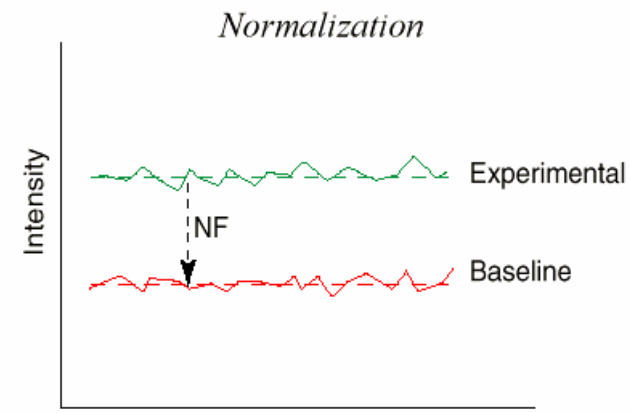
- The data can be normalized from:
  - a limited group of probe sets.
  - all probe sets.

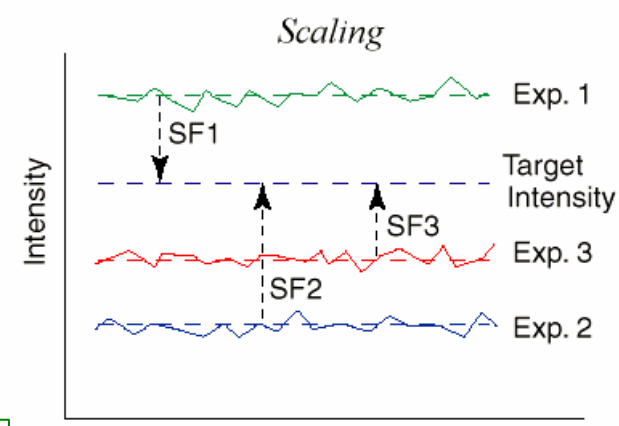
- **Global Scaling**
  the average intensities of all the arrays that are going to be compared are multiplied by scaling factors so that all average intensities are made to be numerically equivalent to a preset amount
  (termed target intensity).

$$\mathbf{SF} = \frac{\mathrm{TGT}}{TrimMean\left(2^{SignalLogValue_i}, 0.02, 0.98\right)}$$

$$\mathbf{A} \times \mathbf{SF} = \mathbf{TGT}$$

$$\Rightarrow \mathbf{SF} = \frac{\mathbf{TGT}}{\mathbf{A}}$$




- **Global Normalization**
  the normalization of the array is multiplied by a Normalization Factor (NF) to make its Average Intensity equivalent to the Average Intensity of the baseline array.

$$nf = \frac{TrimMean\left(SPVb_{i,}\ 0.02, 0.98\right)}{TrimMean\left(SPVe_{i,}\ 0.02, 0.98\right)}$$

$$\mathbf{A}_{exp} \times \mathbf{NF} = \mathbf{A}_{base}$$

$$\Rightarrow \mathbf{NF} = \frac{\mathbf{A}_{base}}{\mathbf{A}_{exp}}$$

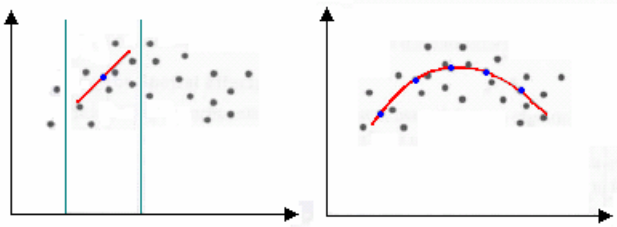
**Average intensity** of an array is calculated by averaging all the Average Difference values of every probe set on the array, excluding the highest 2% and lowest 2% of the values.

# LOESS Normalization



- Loess normalization (Bolstad *et al.*, 2003) is based on **MA plots**. Two arrays are normalized by using a lowess smoother.
- **Skewing** reflects experimental artifacts such as the
  - contamination of one RNA source with genomic DNA or rRNA,
  - the use of unequal amounts of radioactive or fluorescent probes on the microarray.
- Skewing can be corrected with local normalization: fitting a local regression curve to the data.

Loess regression
(locally weighted polynomial regression)



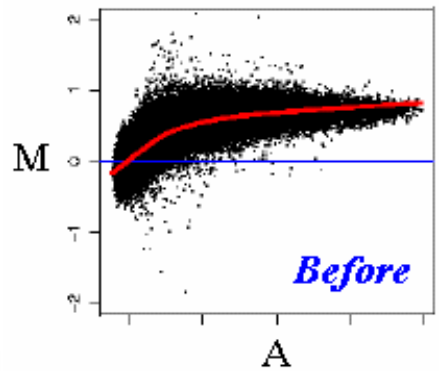
1. For any two arrays $i,j$ with probe intensities $x_{ki}$ and $x_{kj}$ where $k = 1, \ldots, p$ represents the probe
2. we calculate $M_k = \log_2(x_{ki}/x_{kj})$ and $A_k = \frac{1}{2}\log_2(x_{ki}x_{kj})$.
3. A normalization curve is fitted to this $M$ versus $A$ plot using loess.

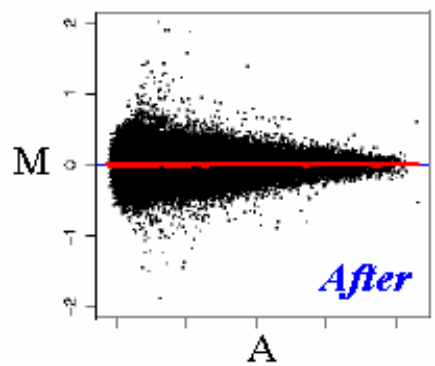
   Loess is a method of local regression (see Cleveland and Devlin (1988) for details).
4. The fits based on the normalization curve are $\hat{M}_k$
5. the normalization adjustment is $M'_k = M_k - \hat{M}_k$.
6. Adjusted probe intensites are given by $x'_{ki} = 2^{A_k + \frac{M'_K}{2}}$ and $x'_{ki} = 2^{A_K - \frac{M'_k}{2}}$.

$$M = \log_2\left(\frac{Y}{X}\right)$$

$$A = \frac{1}{2}\log_2(XY)$$

| Oligo | cDNA |
|---|---|
| $X = PM_1$, $Y = PM_2$ | X= Cy3, Y= Cy5 |
| $X = PM_1 - MM_1$, $Y = PM_2 - MM_2$ | |



# 3. PM Correction Methods



- **PM only**
  make no adjustment to the PM values.
- **Subtract MM from PM**
  This would be the approach taken in MAS 4.0 Affymetrix (1999). It could also be used in conjuntion with the liwong model.

**Table 1: Summary Table**

| Method | Assumptions | Benefits | Drawbacks |
|---|---|---|---|
| **PM-MM** | Background effects are large and potentially variable between features across experiments relative to effects of interest | Background effects minimized due to low bias<br>Sensitivity to low expressors | Slightly noisier when signal is higher than background |
| **PM-B** | Features have approximately the same background | Low noise | May not represent all probe sets accurately, typically leading to underestimated differential change |
| **PM Only** | Background variation is insignificant | Low noise<br>Approximately constant CV | All probe sets biased<br>Compression of differential change at the low end |
| **MM treated as additional PM** | Background variation is insignificant<br>Abundances moderate to large | Added statistical power<br>Low noise<br>Constant CV | All probe sets biased<br>Compression of differential change at the low end |

Affymetrix: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation. Edited by: Affymetrix I. Santa Clara, CA, ; 2005.

# 4. Expression Index Estimates



## Summarization

- Reduce the 11-20 probe intensities on each array to a single number for gene 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.

## Single Chip

- avgDiff : no longer recommended for use due to many flaws.
- **Signal** (MAS5.0): use One-Step Tukey biweight to combine the probe intensities in log scale
- average log 2 (PM - BG)

## Multiple Chip

- **MBEI** (li-wong): a multiplicative model
- **RMA**: a robust multi-chip linear model fit on the log scale

# Three Well-Known Methods

- MAS5 & PLIER
- Li-Wong Model
- RMA/GC-RMA

# MAS5 & PLIER (Affymetrix, 2005)



- Guide to Probe Logarithmic Intensity Error (PLIER) Estimation

| | Previous Generation | 2.0 Platform |
|---|---|---|
| **Array Technology** | • 18-µm features<br>• Edge minimization mask strategy | • 11-µm features<br>• Chrome setback mask design strategy<br>• ARC |
| **Image Analysis** | Global gridding | Feature extraction (in addition to global gridding) |
| **Data Management** | MAS / LIMS | GCOS Client / Server |
| **Analysis** | MAS Statistical Algorithm | GREX including PLIER algorithm (in addition to MAS Statistical Algorithm) |
| **Scanning Technology** | GeneArray® 2500 or GeneChip® Scanner 3000 | GeneChip® Scanner 3000 (high resolution) |
| **Fluidics** | Fluidics Station 400/Fluidics Station 450 | Fluidics Station 450 |
| **AutoLoader** | Not available on GeneArray® 2500 (optional for GeneChip® Scanner 3000) | Optional for GeneChip® Scanner 3000 |
| **Reagents** | • 3rd-party cDNA reagents<br>• Enzo labeling kits | • GeneChip® One- and Two-Cycle cDNA Kits<br>• GeneChip® IVT Labeling Kit |

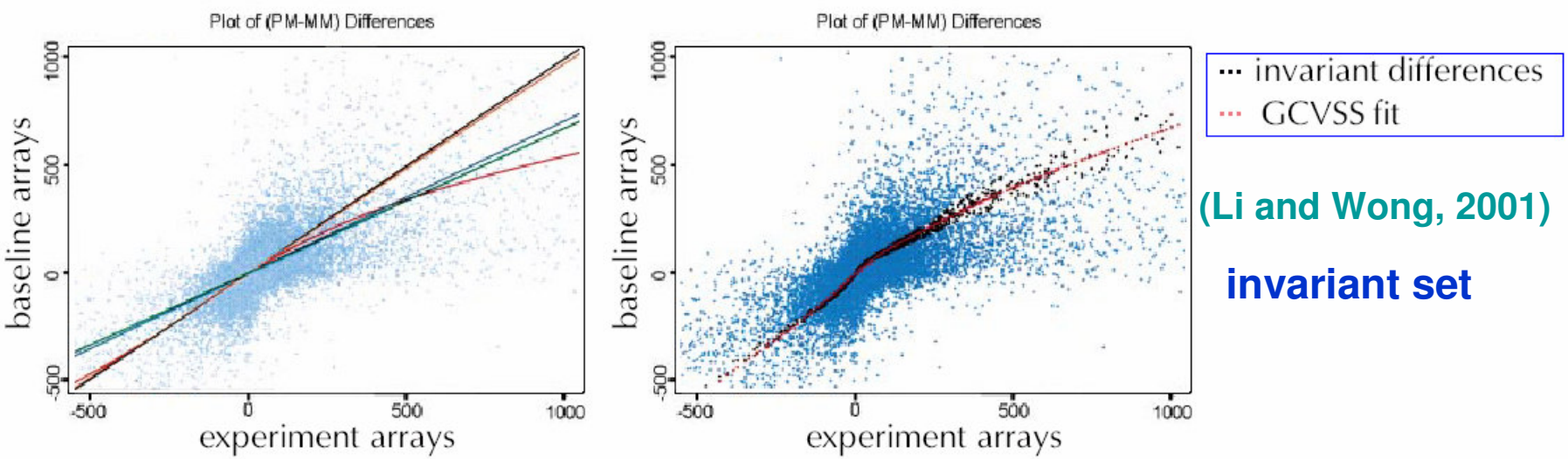
Affymetrix: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation. Edited by: Affymetrix I. Santa Clara, CA, ; 2005.

# Liwong: Normalization



```
Liwong (PM-only Model)
eset.liwong <- expresso(Data, bg.correct=FALSE, normalize.method = "invariantset",
                        pmcorrect.method="pmonly", summary.method="liwong")
Liwong (PM-MM Model)
eset.liwong <- expresso(Data, bg.correct=FALSE, normalize.method = "invariantset",
                        pmcorrect.method="subtractmm ", summary.method="liwong")
```

- Using a baseline array, arrays are normalized by selecting invariant sets of genes (or probes) then using them to fit a *non-linear relationship* between the "treatment" and "baseline" arrays.
- A set of probe is said to be invariant if ordering of probe in one chip is same in other set.
- Fit the non-linear relation using cross validated smoothing splines (GCVSS).



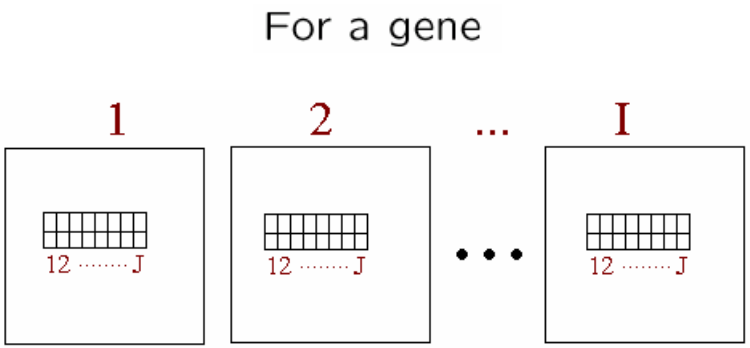

**(Li and Wong, 2001)**

**invariant set**

# Liwong: Summarization Method



## (Model-Based Expression Index , MBEI)




expression on array $j$

$$y_{ij} = \theta_i \phi_j + \epsilon_{ij}$$

$PM_{ij}$ or $PM_{ij} - MM_{ij}$

probe response parameter

$\sum_j \phi_j^2 = J$

$\epsilon_{ij} \sim N\left(0, \sigma^2\right)$

$i = 1, \ldots, I$ the number of chips

$j = 1, \ldots, J$ number of probe pairs

- $\theta$ i: this model computes an expression level on the ith array.
- SE( $\theta$ ) 's and SE( $\varphi$ ) 's: can be used to identify outlier arrays and probes that will consequently be excluded from the final estimation of the probe response pattern.

- **Outlier array:** large SE( $\theta$ i), possibly due to external factors like the imaging process.
- **Outlier probe:** large SE( $\varphi$ j), possibly due to non-specific cross-hybridization.
- **Single outliers:** individual PM-MM differences might also be identified by large residuals compared with the fit. (these are regarded as missing values in the model-fitting algorithm).

# RMA: Background Correction



```
RMA
eset.rma <- expresso(Data, bg.correct="rma", normalize.method = "quantiles",
                     pmcorrect.method="pmonly", summary.method="medianpolish")
```

RMA: Robust Multichip Average (Irizarry and Speed, 2003):
assumes PM probes are a convolution of Normal and Exponential.

Observed PM = Signal + Noise $\boxed{O = S + N}$

Exponential (alpha) Normal (mu, sigma)

Use $E[S|O=o, S>0]$ as the background corrected PM.

$$E(s|O=o) = a + b\frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{o-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{o-a}{b}\right) - 1}$$

$$a = s - \mu - \sigma^2\alpha$$

$$b = \sigma$$

$\phi$ : standard normal density function

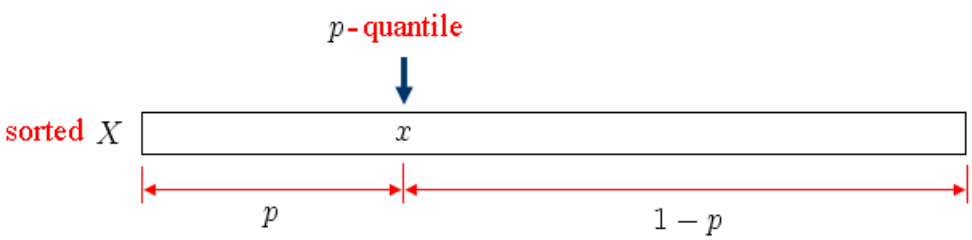
$\Phi$ : standard normal distribution function

Ps. MM probe intensities are not corrected by RMA/RMA2.

# Quantiles

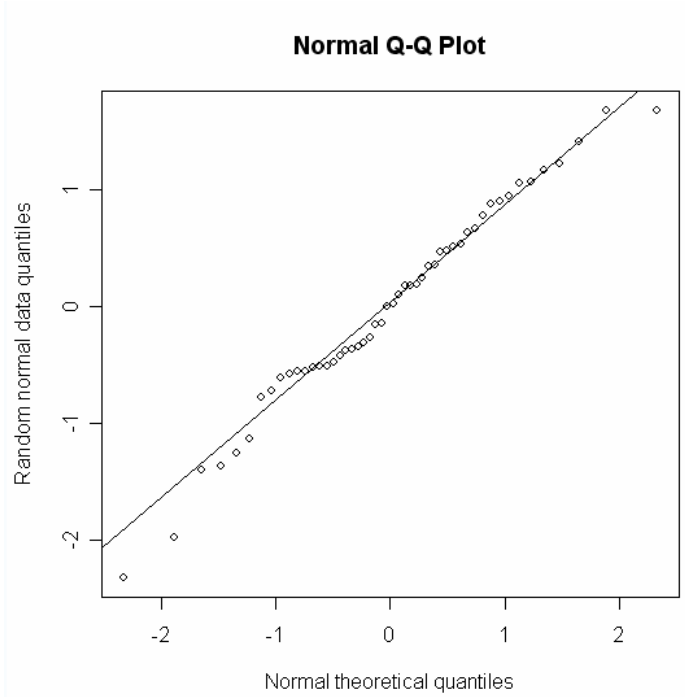


```
The qth quantile of a data set is
defined as that value where a q
fraction of the data is below that
value and (1-q) fraction of the data
is above that value. For example,
the 0.5 quantile is the median.
```




$$P(X < x) \leq p \text{ and } P(X > x) \leq 1 - p.$$

(e.g., the .5 quantile = the 50% point = the median).

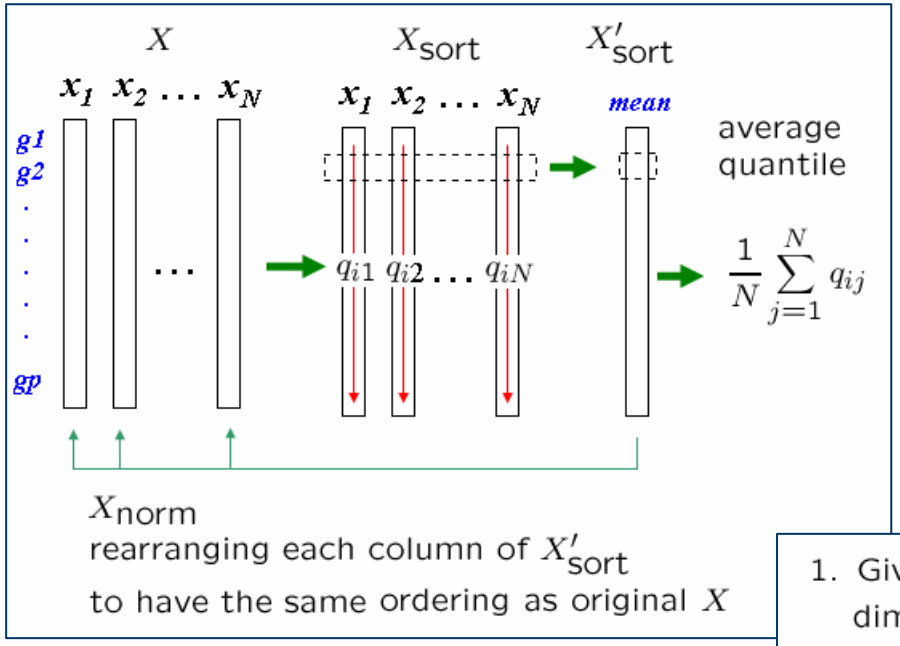



A quantile-quantile plot (or q-q plot) is a graphical data analysis technique for comparing the distributions of 2 data sets.

# RMA: Normalization



- **Quantiles Normalization** (Bolstad *et al*, 2003) is a method to make the distribution of probe intensities the same for every chip.
- Each chip is really the transformation of an underlying common distribution.




1. Given $N$ datasets of length $p$ form $X$ of dimension $p \times N$ where each dataset is a column
2. Set $d = \left(\frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}}\right)$
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 $X$

- The two distribution functions are effectively estimated by the sample quantiles.
- The normalization distribution is chosen by averaging each quantile across chips.

# RMA: Summarization Method



## MedianPolish

- This is the summarization used in the RMA expression summary Irizarry et al. (2003).
- A **multichip linear model** is fit to data from each probeset.
- The medianpolish is an algorithm (see Tukey (1977)) for fitting this model robustly.
- Please note that expression values you get using this summary measure will be in log2 scale.

for a probeset $k$

$$\boxed{\log_2\left(PM_{ij}^{(k)}\right) = \alpha_i^{(k)} + \beta_j^{(k)} + \epsilon_{ij}^{(k)}}$$

$i = 1, \ldots, I_k$ probes

$j = 1, \ldots, J$ arrays

probe effect

$\log_2$ expression value

# GC-RMA



- Robust multi-chip average with GC-content background correction
- Background correction: account for background noise as well as non-specific binding.

optical noise, logNormal

quantity proportional to RNA expression

$$PM = O_{PM} + N_{PM} + S,$$

$$MM = O_{MM} + N_{MM}$$

Observed PM, MM

non-specific binding noise, Bi-variate Normal

Ps. Probe affinity is modeled as a sum of position-dependent base effects and can be calculated for each PM and MM value, based on its corresponding sequence information.

Zhijin Wu; Rafael A. Irizarry; Robert Gentleman; Francisco Martinez-Murillo; Forrest Spencer, 2004, A Model-Based Background Adjustment for Oligonucleotide Expression Arrays, Journal of the American Statistical Association 99(468), 909-917.

# Comparison of Affymetrix GeneChip Expression Measures



Affycomp II: A Benchmark for Affymetrix GeneChip Expression Measures - Microsoft Internet Explorer

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

## Affycomp II

A Benchmark for Affymetrix GeneChip Expression Measures

- Background
- Data and instructions
- Submission form
- Competition results
  - new assessment (of SPIKE-
  - original assessment (of DIL
  - entry comparison tool (beta
  - study archives
- Comparison of Affymetrix GeneC
- A Benchmark for Affymetrix Gen
- R package
- FAQ
- Contact us

Sponsored by: The Hopgene Pr

Results as of August 7, 2003 pre

**dilution study (GeneLogic)**
**spike-in study (Affymetrix)**

| N | Method / Submitter | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | (perfection) | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.0 |
| 1 | MAS_5.0 / rafa | 0.29 | 0.47 | 4.01 | 0.91 | 0.77 | 0.58 | 0.73 | 0.77 | 0.77 | 0.64 | 0.09 | 0.00 | 0.00 | 0.0 |
| 2 | RMA / rafa | 0.07 | 0.13 | 0.40 | 0.90 | 0.68 | 0.20 | 0.71 | 0.80 | 0.68 | 0.31 | 0.57 | 0.91 | 0.96 | 0.6 |
| 8 | RMA VSN / thomas.cappola | 0.02 | 0.04 | 0.15 | 0.89 | 0.12 | 0.06 | 0.13 | 0.10 | 0.12 | 0.08 | 0.46 | 0.59 | 0.43 | 0.4 |
| 23 | rsvd / jack liu | 0.14 | 0.12 | 0.73 | 0.94 | 0.74 | 0.31 | 0.78 | 0.73 | 0.74 | 0.43 | 0.53 | 0.73 | 0.71 | 0.5 |
| 25 | rsvd.pm / jack liu | 0.06 | 0.11 | 0.34 | 0.89 | 0.53 | 0.12 | 0.53 | 0.77 | 0.53 | 0.16 | 0.42 | 0.90 | 0.96 | 0.5 |
| 26 | rma-tog / dgreco | 0.07 | 0.13 | 0.40 | 0.90 | 0.68 | 0.20 | 0.71 | 0.80 | 0.68 | 0.31 | 0.57 | 0.91 | 0.96 | 0.65 |
| 27 | rma-sep / dgreco | 0.18 | 0.28 | 0.96 | 0.90 | 0.71 | 0.27 | 0.72 | 0.84 | 0.71 | 0.39 | 0.38 | 0.53 | 0.63 | 0.42 |
| 28 | LW1 / dgreco | 0.08 | 0.14 | 1.18 | 0.91 | 0.59 | 0.19 | 0.62 | 0.74 | 0.59 | 0.25 | 0.23 | 0.47 | 0.55 | 0.29 |
| 29 | LW2 / dgreco | 0.14 | 0.25 | 13.88 | 0.56 | 1.08 | 1.50 | 0.80 | 0.68 | 1.08 | 1.45 | 0.19 | 0.00 | 0.00 | 0.14 |
| 30 | rsvd.bgc / jack liu | 0.08 | 0.14 | 0.52 | 0.89 | 0.58 | 0.16 | 0.59 | 0.79 | 0.58 | 0.22 | 0.38 | 0.80 | 0.90 | 0.49 |
| 31 | cor523 / cope | 0.02 | 0.03 | 0.12 | 0.88 | 0.12 | 0.06 | 0.13 | 0.10 | 0.12 | 0.08 | 0.54 | 0.77 | 0.61 | 0.60 |
| 33 | UM-Tr-Mn / jmacdon | 0.15 | 0.25 | 1.86 | 0.93 | 0.70 | 0.36 | 0.72 | 0.70 | 0.70 | 0.44 | 0.18 | 0.10 | 0.10 | 0.16 |
| 34 | GS_RMA / thon | 0.07 | 0.13 | 0.40 | 0.90 | 0.68 | 0.20 | 0.71 | 0.80 | 0.68 | 0.30 | 0.56 | 0.91 | 0.96 | 0.65 |
| 35 | GS_GCRMA / thon | 0.07 | 0.09 | 0.65 | 0.93 | 0.93 | 0.37 | 0.96 | 0.96 | 0.93 | 0.55 | 0.59 | 0.87 | 0.90 | 0.66 |
| 36 | gcrma113 / zwu | 0.06 | 0.04 | 0.61 | 0.91 | 1.00 | 0.25 | 1.13 | 0.97 | 1.00 | 0.48 | 0.45 | 0.91 | 0.92 | 0.57 |
| 37 | rsvd2 / jack.liu | 0.17 | 0.28 | 1.74 | 0.91 | 0.75 | 0.46 | 0.74 | 0.81 | 0.75 | 0.52 | 0.29 | 0.16 | 0.21 | 0.26 |
| 38 | W237 / dario.greco | 0.02 | 0.04 | 0.17 | 0.87 | 0.12 | 0.05 | 0.13 | 0.10 | 0.12 | 0.07 | 0.35 | 0.54 | 0.39 | 0.39 |
| 39 | RMA_NBG / bolstad | 0.01 | 0.02 | 0.06 | 0.90 | 0.09 | 0.02 | 0.09 | 0.10 | 0.09 | 0.04 | 0.54 | 0.90 | 0.93 | 0.63 |

Data and instructions

1. Download the spike-in and dilution data sets.
   - **Spike-in hgu95a Data**

| Method | SD | 99.9% | slope low | slope med | slope high | AUC |
|---|---|---|---|---|---|---|
| GCRMA | 0.08 | 0.74 | 0.66 | 1.06 | 0.56 | 0.70 |
| GS_GCRMA | 0.10 | 0.79 | 0.62 | 1.03 | 0.55 | 0.66 |
| MMEI | 0.04 | 0.23 | 0.16 | 0.54 | 0.46 | 0.62 |
| GL | 0.05 | 0.25 | 0.16 | 0.55 | 0.46 | 0.62 |
| RMA_NBG | 0.04 | 0.24 | 0.16 | 0.56 | 0.46 | 0.61 |
| RSVD | 0.00 | 0.58 | 0.42 | 0.85 | 0.40 | 0.61 |
| ZL | 0.22 | 0.52 | 0.35 | 0.71 | 0.45 | 0.61 |
| VSN_scale | 0.09 | 0.43 | 0.28 | 0.91 | 0.70 | 0.59 |
| VSN | 0.06 | 0.28 | 0.18 | 0.6 | 0.46 | 0.59 |
| RMA_VSN | 0.09 | 0.48 | 0.31 | 0.74 | 0.46 | 0.57 |
| GLTRAN | 0.07 | 0.42 | 0.23 | 0.61 | 0.45 | 0.55 |
| ZAM | 0.09 | 0.50 | 0.30 | 0.70 | 0.47 | 0.54 |
| RMA_GNV | 0.11 | 0.58 | 0.35 | 0.76 | 0.47 | 0.52 |
| RMA | 0.11 | 0.57 | 0.35 | 0.76 | 0.47 | 0.52 |
| GSrma | 0.11 | 0.57 | 0.35 | 0.76 | 0.47 | 0.52 |
| GSVDmod | 0.07 | 0.44 | 0.22 | 0.64 | 0.42 | 0.51 |
| PerfectMatch | 0.05 | 0.40 | 0.18 | 0.56 | 0.43 | 0.50 |
| PLIER+16 | 0.13 | 0.83 | 0.49 | 0.80 | 0.46 | 0.48 |
| GSVDmin | 0.08 | 0.60 | 0.22 | 0.62 | 0.41 | 0.41 |
| MAS 5.0+32 | 0.14 | 1.07 | 0.35 | 0.71 | 0.44 | 0.12 |
| ChipMan | 0.27 | 2.26 | 0.44 | 1.11 | 0.68 | 0.12 |
| qn.p5 | 0.12 | 1.09 | 0.13 | 0.50 | 0.52 | 0.11 |
| dChip | 0.13 | 1.44 | 0.31 | 0.67 | 0.39 | 0.09 |
| mmgMOSgs | 0.40 | 3.27 | 1.34 | 1.13 | 0.45 | 0.07 |
| gMOSv.1 | 0.29 | 3.35 | 0.98 | 1.12 | 0.42 | 0.06 |
| ProbeProfi ler | 0.31 | 18.75 | 1.61 | 1.57 | 0.39 | 0.03 |
| dChip PM-MM | 0.23 | 14.83 | 1.40 | 0.86 | 0.35 | 0.02 |
| mgMOS_gs | 0.36 | 2.86 | 0.83 | 0.86 | 0.43 | 0.01 |
| MAS 5.0 | 0.63 | 4.48 | 0.69 | 0.81 | 0.45 | 0.00 |
| PLIER | 0.19 | 123.27 | 0.75 | 0.85 | 0.46 | 0.00 |
| UM-Tr-Mn | 0.32 | 2.92 | 0.58 | 0.83 | 0.42 | 0.00 |

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


- Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP. A benchmark for Affymetrix GeneChip expression measures, Bioinformatics. 2004 Feb 12;20(3):323-31.
- Irizarry RA, Wu Z, Jaffee HA. Comparison of Affymetrix GeneChip expression measures. Bioinformatics. 2006 Apr 1;22(7):789-94.

# Methods Comparison



**Table 2: Other analysis methods**

| Method | Assumptions | Benefits | Drawbacks |
|---|---|---|---|
| **PLIER** | Multiple array analysis<br>Mixed error model<br>PM-MM, PM only etc.<br>Multiple background options<br>Smoothly handles intensities below background | Higher reproducibility of signal (lower coefficient of variation) without loss of accuracy relative to single array analysis<br>Higher differential sensitivity for low expressors<br>Lack of bias | Computationally intensive<br>In cases where feature intensities disagree, may have more than one solution<br>Performance relative to amount of model data provided<br>Variance not stable on log scale |
| **dCHIP** | Multiple array analysis<br>Arithmetic error model<br>PM only (stanardly)<br>Multiple background options (no background typical) | Higher reproducibility of signal over single array analysis<br>Good differential change detection<br>Variance stable on log scale with no background | In cases where feature intensities disagree, may have more than one solution<br>Performance relative to amount of model data provided<br>Positive bias at low end (compression of Fold Change) |
| **RMA** | Multiple array analysis<br>Multiplicative error<br>PM only<br>Attenuated global background (single global background used to adjust for each intensity) | Higher reproducibility of signal over single array analysis<br>Good differential change detection<br>Variance stable on log scale | In cases where feature intensities disagree, may have more than one solution (mitigated by median polish)<br>Performance relative to amount of model data provided<br>Positive bias at low end (compression of Fold Change) |
| **MAS 5** | Single array analysis<br>Multiplicative error<br>PM-MM<br>Background imputed to handle negative differences | Conservative<br>Smooth down-weighting of outliers<br>Positive output values<br>Minimal bias | Limited by single array analysis<br>Variance not stable on log scale<br>Some positive bias |

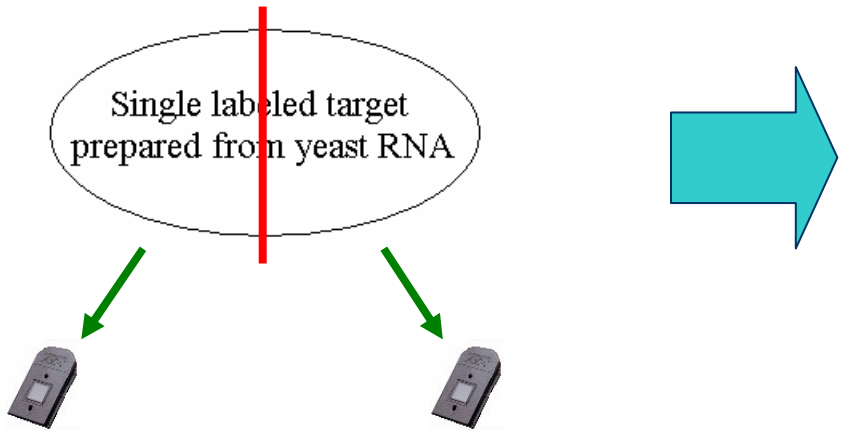
Affymetrix: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation. Edited by: Affymetrix I. Santa Clara, CA, ; 2005.

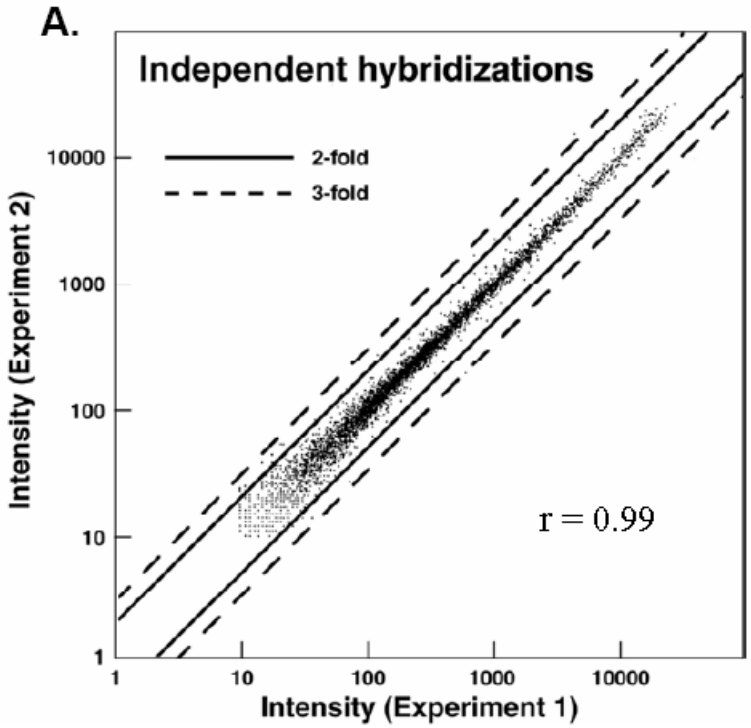
# Reproducibility
# and
# False Positive Rates

# Reproducibility



- Reproducibility (再現性) = Repeatability = Precision
- The degree to which repeat measurements of the same quantity will show the same or similar results.
- The precision is usually measured by comparing some measure of dispersion (e.g., standard deviation) with zero.




Only 14 of the more than 6,200 probes sets showed a difference of more than 2-fold between repeat measurements

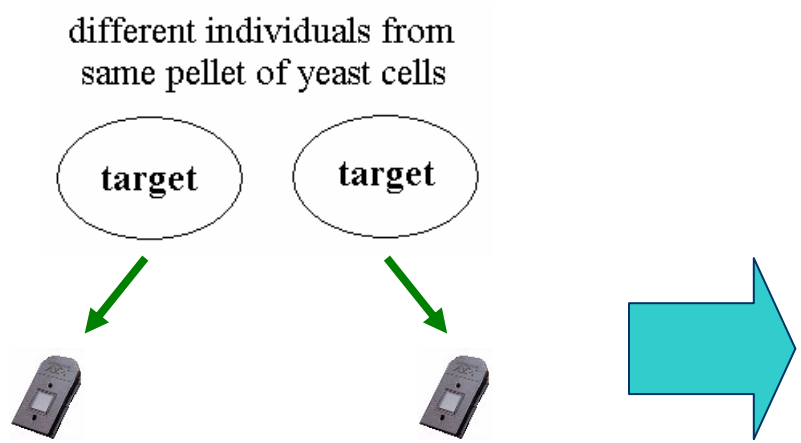
Maximum observed change was 3.4-fold

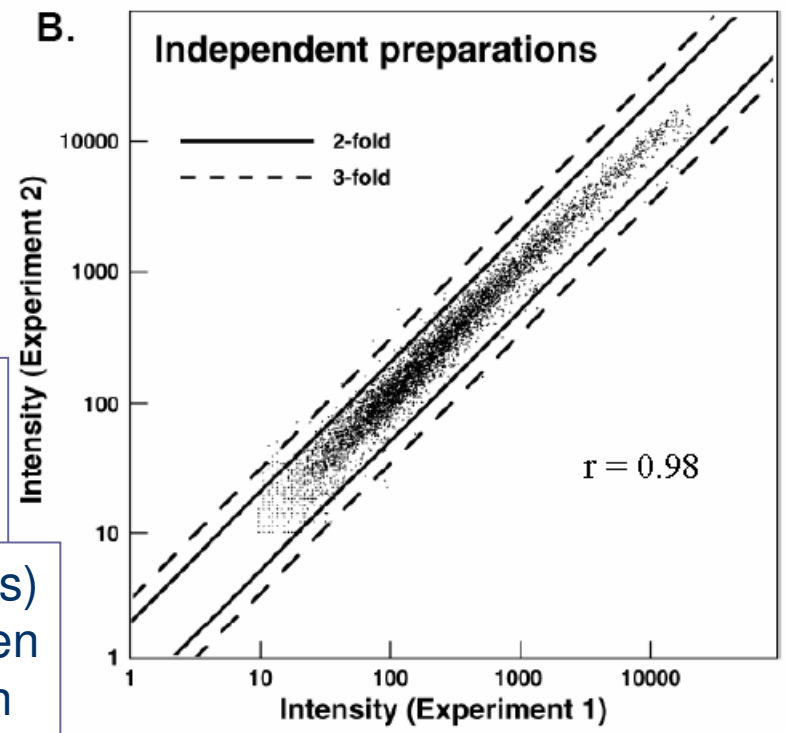
Pearson correlation coefficients = 0.99

Figures Source: *Modified from* Affymetrix yeast microarray (Wodicka 1997)

# Reproducibility (conti.)






74 probe sets of the more than 6,200 probes sets with an intensity difference of more than 2-fold.

the level of variation (i.e. false positives) between identical samples and between independent hybridizations is less than 2%.

Figures Source: *Modified from* Affymetrix yeast microarray (Wodicka 1997).

Only 6 showed a difference of at least 3-fold.

Pearson correlation coefficients = 0.98

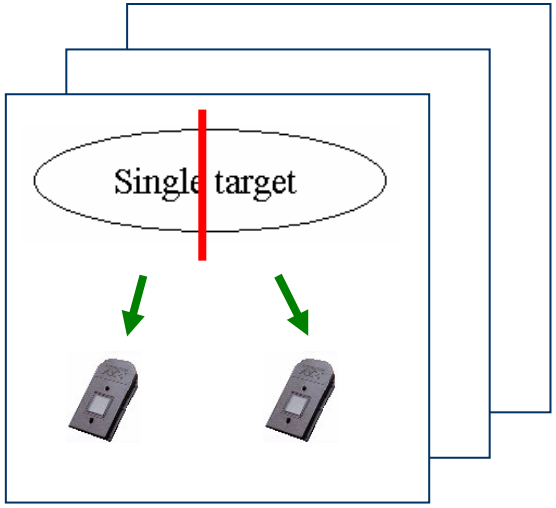
Concordance correlation coefficients

Lin L. A concordance correlation coefficient to evaluate reproducibility. Biometrics1989;45:255-268.

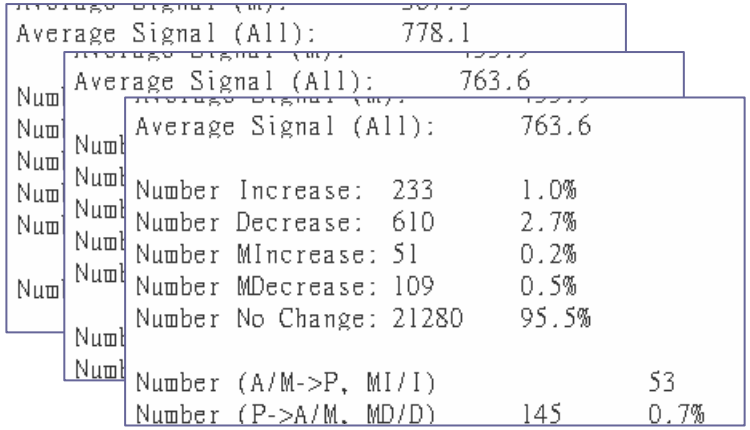
# False Positives Rates



- **Comparison Expression Analysis**: a pairwise comparison of all probe pairs for each transcript to identify Increase or decrease in expression between two samples.




Comparison Expression Analysis




Mean Percent Change
Standard Deviation

**Table 2. Reproducibility Studies on Lots of Affymetrix Arrays.**

| GeneChip | Lot Number | Mean Percent Change | Standard Deviation |
|---|---|---|---|
| HG-U133A | 1008682 | 0.28 | 0.17 |
| | 1008684 | 0.13 | 0.03 |
| | 1008685 | 0.52 | 0.06 |

A false change is defined as the percent of transcripts that demonstrate an Increase or Decrease in expression between the two samples as determined by the ArraySuite Comparison software.
(Source = Dr. Elizabeth Kerr, Marketing Director for Gene Expression, Affymetrix, Inc.)

# Software



## Shareware/Freeware

- Bioconductor (R, Gentleman)
- BRB-ArrayTools (embedded in Excel)
- DNA-Chip Analyzer (dChip) (Li and Wong)
- RMAExpress

## Commercial

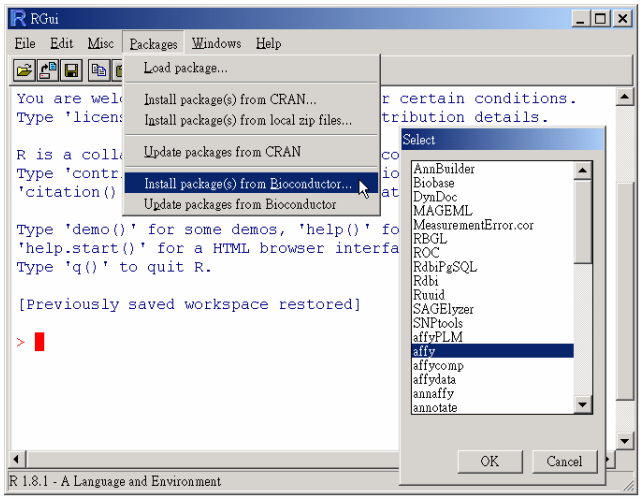
- Affymetrix GeneChip Operating Software (GCOS v1.4)
- GeneSpring GX v7.3
- Spotfire

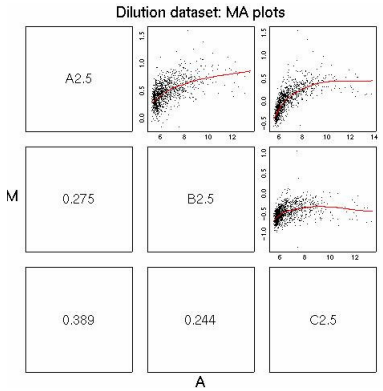
# The Bioconductor: affy

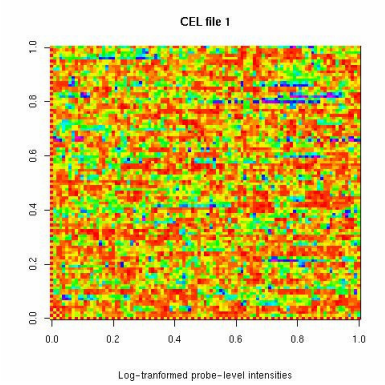


The Bioconductor Project
Release 2.1
http://www.bioconductor.org/

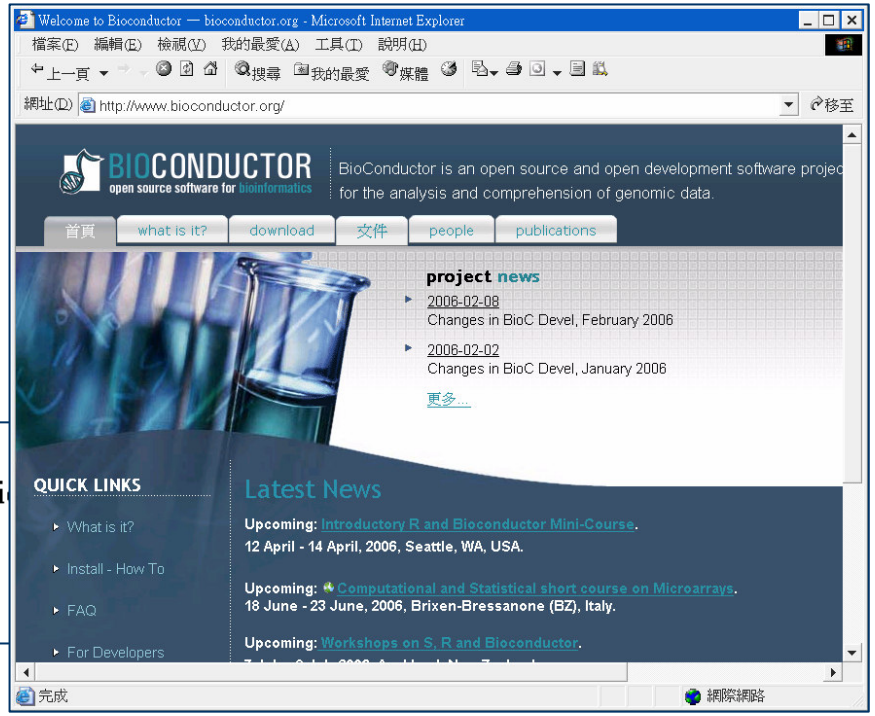
affypdnn
affyPLM
gcrma
makecdfenv


| | |
|---|---|
| affy | Methods for Affymetrix Oligonucleotide Arrays |
| affycomp | Graphics Toolbox for Assessment of Affymetrix Expressi |
| affydata | Affymetrix Data for Demonstration Purpose |
| annaffy | Annotation tools for Affymetrix biological metadata |
| AffyExtensions | For fitting more general probe level models |



# The Bioconductor: affy



*Quick Start:* probe level data (*.cel) to expression measure.

```
> library(affy)
> getwd()
> list.celfiles()
> setwd("myaffy")
> getwd()
> list.celfiles()
> Data <- ReadAffy()

> eset.rma <- rma(Data)
> eset.mas <- expresso(Data,
                       normalize= FALSE,
                       bgcorrect.method="mas",
                       pmcorrect.method="mas",
                       summary.method="mas")
> eset.liwong <- expresso(Data,
                    normalize.method="invariantset",
                    bg.correct=FALSE,
                    pmcorrect.method="pmonly",
                    summary.method="liwong")
> eset.myfun <- express(Data,
                        summary.method=function(x)
                                apply(x, 2, median))

> write(eset.rma, file="mydata_rma.txt")
> write(eset.mas, file="mydata_mas.txt")
> write.exprs(eset.liwong, file="mydata_liwong.txt")
> write(eset.myfun, file="mydata_myfun.txt")
```

```
expresso(
    afbatch,

  # background correction
    bg.correct = TRUE,
    bgcorrect.method = NULL,      none,
    bgcorrect.param = list(),     mas,
                                  rma

  # normalize
    normalize = TRUE,
    normalize.method = NULL,      constant,
    normalize.param = list(),     contrasts.
                                  invariantset,
                                  loess, qspline,
                                  quantiles,
                                  quantiles.robust

  # pm correction
    pmcorrect.method = NULL,      mas,
    pmcorrect.param = list(),     pmonly,
                                  subtractmm

  # expression values
    summary.method = NULL,        avgdiff,
    summary.param = list(),       liwong,
    summary.subset = NULL,        mas,
                                  medianpolish,
                                  playerout

  # misc.
    verbose = TRUE,
    warnings = TRUE,
    widget = FALSE)
```

# Browse the Packages by Task Views



http://www.bioconductor.org/packages/2.1/BiocViews.html

## Bioconductor Task View: BiocViews

### Subviews

- Softwa
- Annota
- Experi

## Bioconductor Task V

### Subview of

- BiocViews

### Subviews

- Microarray
- Annotation
- Visualization
- Statistics
- GraphsAndNetworks
- Technology
- Infrastructure
- GUI

http://www.bioconductor.org/packages/2.1/Microarray.html - Windows Internet Explorer

## Bioconductor Task View: Microarray

### Subview of

- Software
- Technology

### Subviews

- OneChannel
- TwoChannel
- DataImport
- QualityControl
- Preprocessing
- Transcription
- DNACopyNumber
- SNPsAndGeneticVariability
- CpGIsland

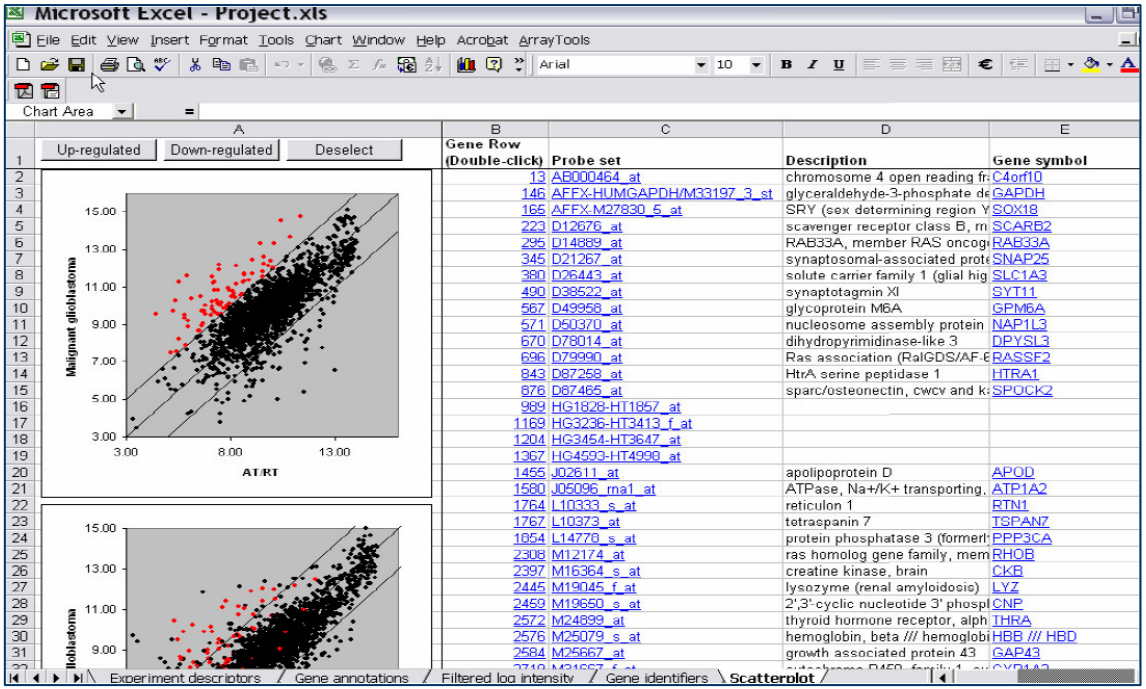
### Packages in view

| Package | Maintainer | Title |
|---|---|---|
| ABarray | Yongming Andrew Sun | Microarray QA and statistical data analysis for Applied Biosystems Genome Survey Micorarray (AB1700) gene expression data. |
| aCGH | Jane Fridlyand | Classes and functions for Array Comparative Genomic Hybridization data. |
| adSplit | Claudio Lottaz | Annotation-Driven Clustering |
| affxparser | Kasper Daniel Hansen | Affymetrix File Parsing SDK |
| affy | Rafael A. Irizarry | Methods for Affymetrix Oligonucleotide Arrays |
| affycomp | Rafael A. Irizarry | Graphics Toolbox for Assessment of Affymetrix Expression Measures |
| affycoretools | James W. MacDonald | Functions useful for those doing repetitive analyses with Affymetrix GeneChips. |
| AffyExpress | Xuejun Arthur Li | Affymetrix Quality Assessment and Analysis Tool |
| affyio | Benjamin Milo Bolstad | Tools for parsing Affymetrix data files |
| affylmGUI | Keith Satterley | GUI for affy analysis using limma package |
| | | Probe Dependent Nearest Neighbours (PDNN) for the affy |

# BRB-ArrayTools



An Integrated Software Tool for DNA Microarray Analysis




http://linus.nci.nih.gov/BRB-ArrayTools.html

**Requirement:**

1. Java Virtual Machine
2. R base (version 2.6.0)
3. RCOM 2.5

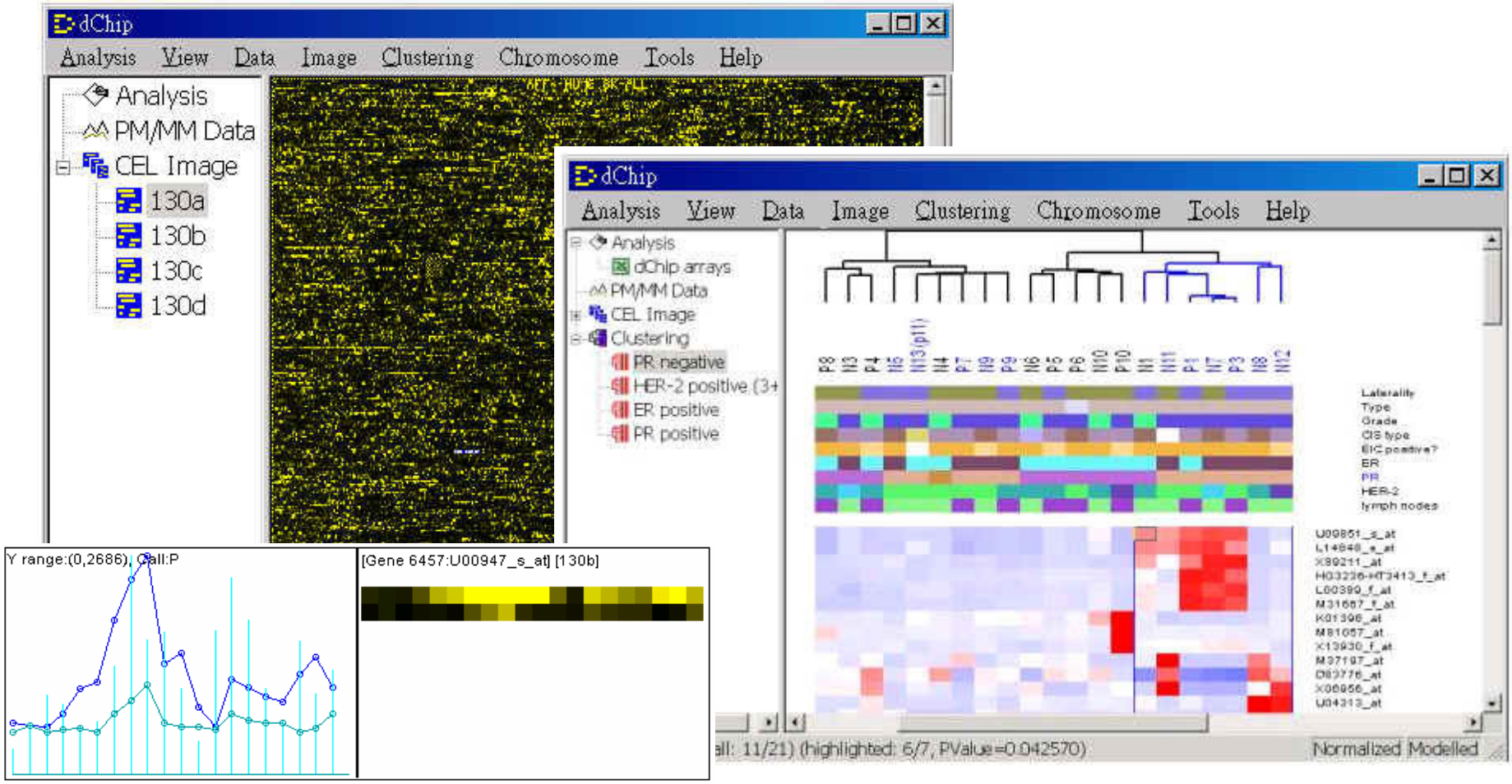
- Software was developed with the purpose of deploying powerful statistical tools for use by biologists.
- Analyses are launched from user-friendly Excel interface.

- Normalization: call RMA, GC-RMA from Bioconductor.
- Affymetrix Quality Control for CEL files: call “simpleaffy” and “affy” from Bioconductor.

# DNA-Chip Analyzer (dChip)



dChip Software: Analysis and visualization of gene expression and SNP microarrays



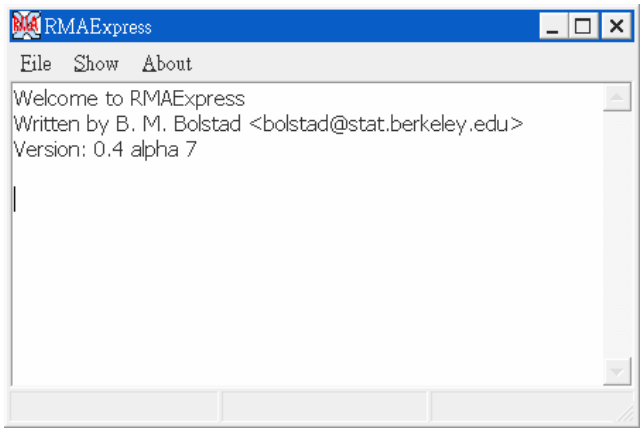

http://www.biostat.harvard.edu/complab/dchip/

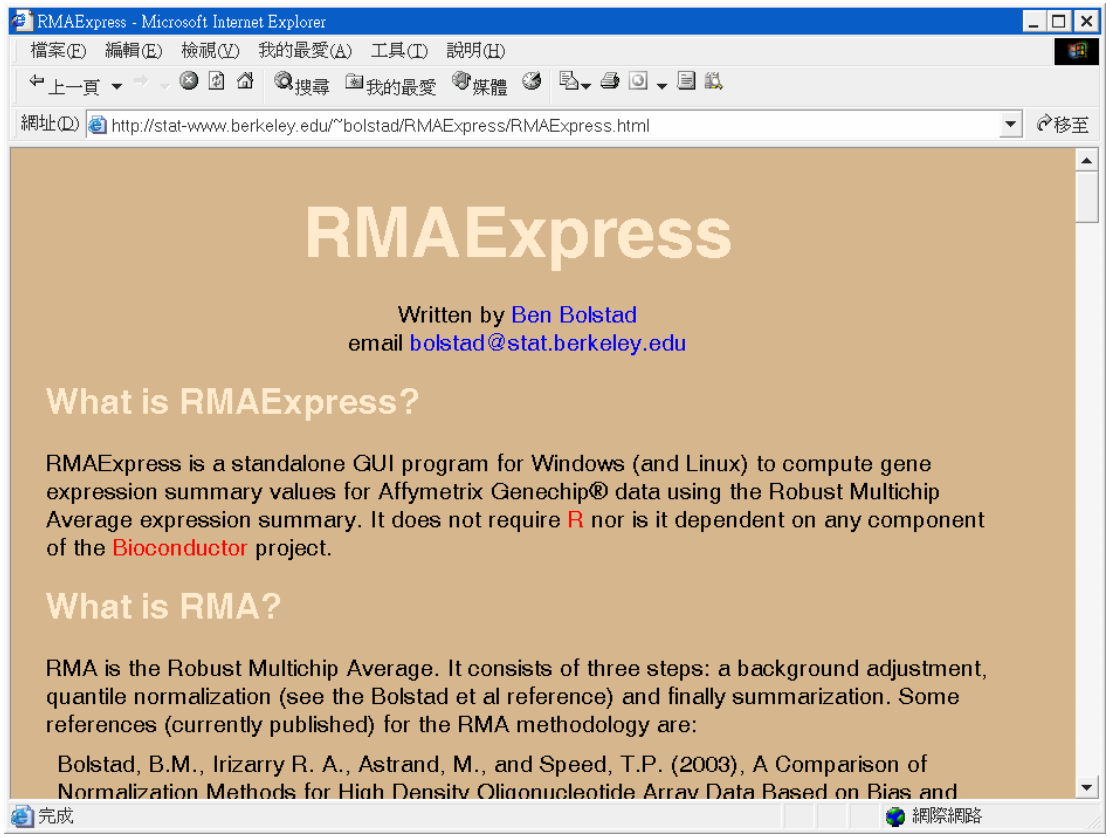
# RMAExpress



Ben Bolstad
Biostatistics,
University Of California, Berkeley
http://stat-www.berkeley.edu/~bolstad/

**Talks Slides**




http://stat-www.berkeley.edu/~bolstad/RMAExpress/RMAExpress.html

# GCOS V1.4

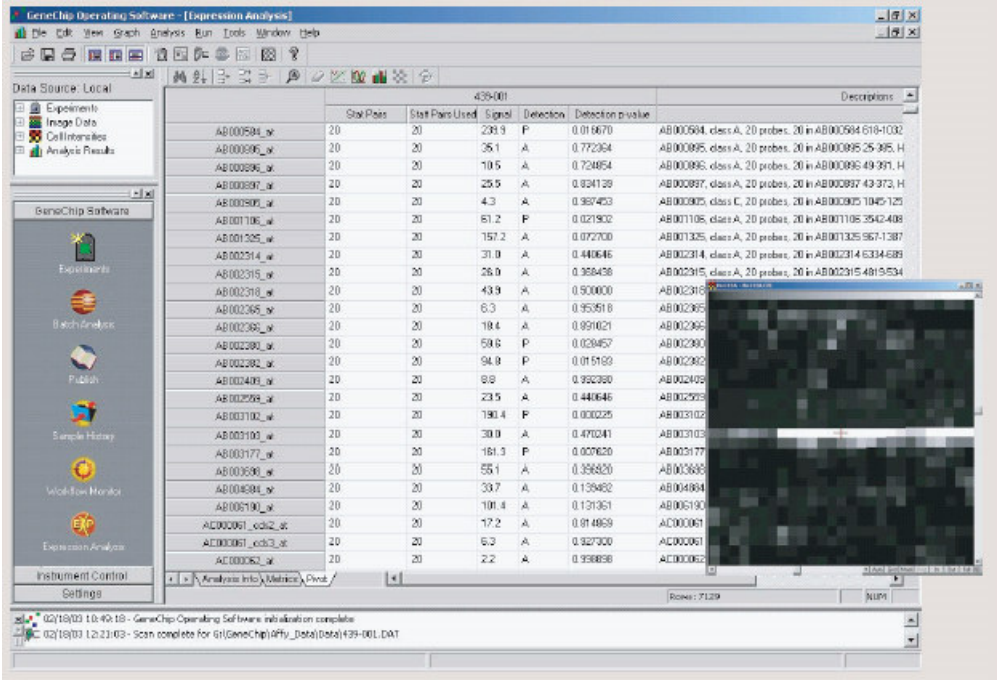


## Affymetrix GeneChip Operating Software

http://www.affymetrix.com/support/technical/software_downloads.affx


http://www.affymetrix.com

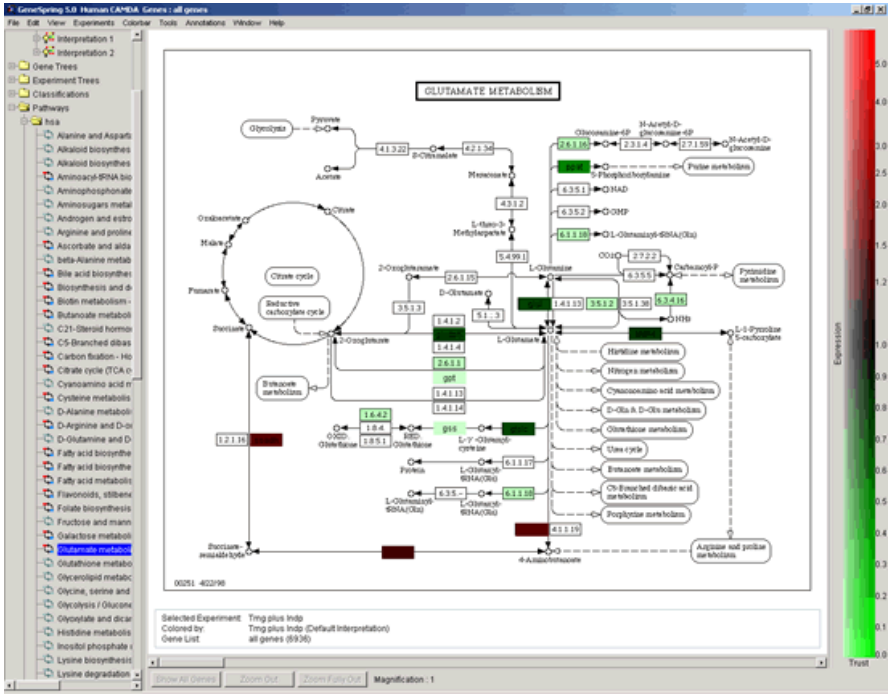
| Specifications | |
|---|---|
| Instrument Support | • Affymetrix GeneChip® Fluidics Station 400 & 450<br>• GeneChip Scanner 3000<br>• GeneArray 2500 Scanner |
| Affymetrix Software Compatibility | • Support GeneChip DNA Analysis Software (GDAS) for mapping and resequencing data analysis<br>• Support Affymetrix® Data Mining Tool software for statistical and clus analysis |
| Database Engine | • Microsoft Data Engine |
| GCOS Database | • Process Database<br>• Publish Database<br>• Gene Information Database |
| Database Management | • GCOS Manager<br>• GCOS Administrator |
| Algorithm | • Affymetrix Statistical Expression Algorithm |

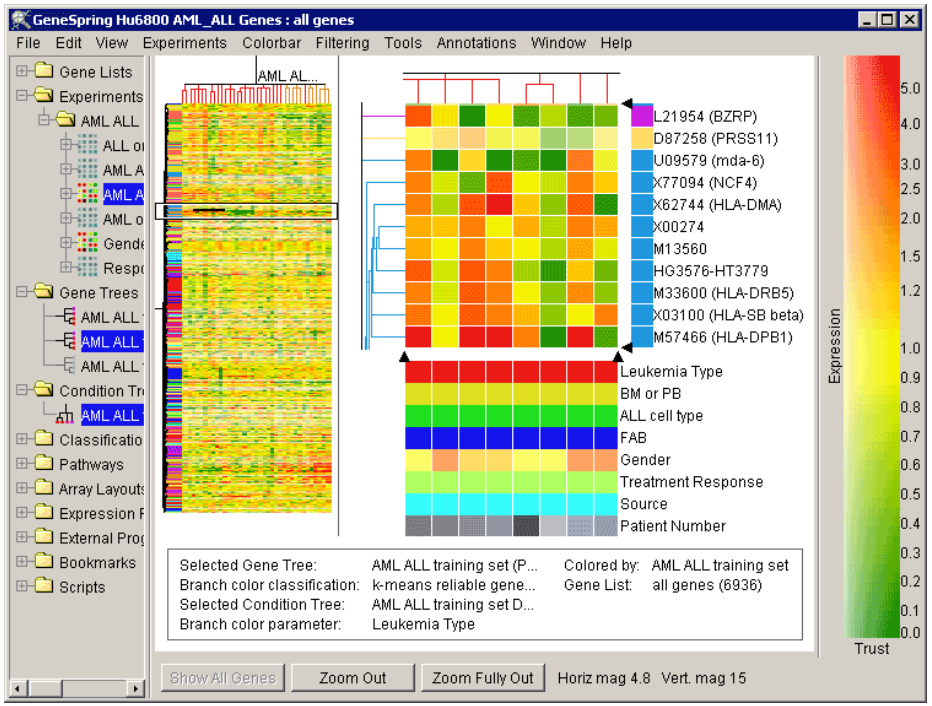


# GeneSpring GX v7.3.1



- RMA or GC-RMA probe level analysis
- Advanced Statistical Tools
- Data Clustering
- Visual Filtering
- 3D Data Visualization
- Data Normalization (Sixteen)
- Pathway Views
- Search for Similar Samples
- Support for MIAME Compliance
- Scripting
- MAGE-ML Export




Images from
http://www.silicongenetics.com

Agilent Technologies

2004 Articles Citing GeneSpring®

**2004** : 2003 : 2002 : 2001 : pre-2001 : Reviews

More than 700 papers

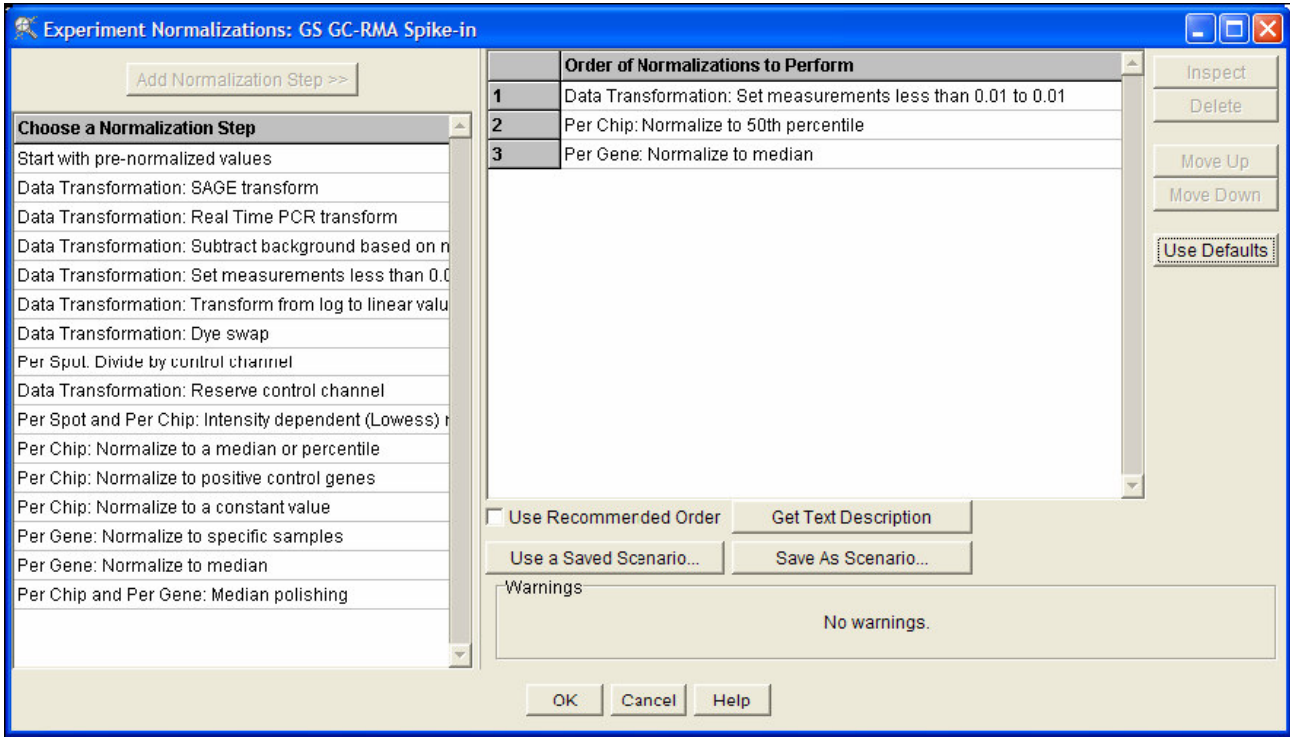
# Normalization in GeneSpring



After RMA or GC-RMA analysis: ensure that these are no negative values and that the data is centered on the value 1.

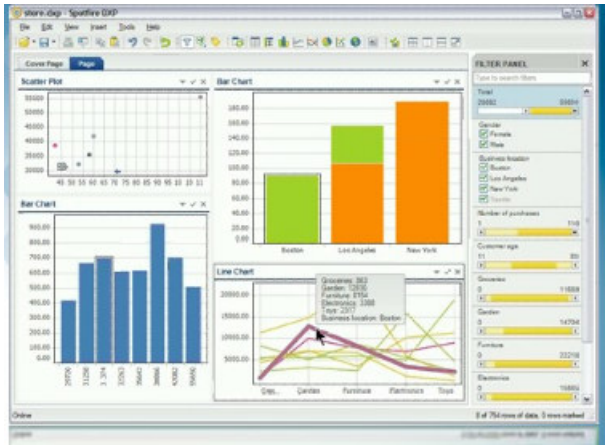
- Data transformation: set measurements less than 0.01 to 0.01.
- Per-gene: normalize to 50th percentile.
- Per-gene: normalize to median.

$$\frac{\text{(signal strength of gene A in sample X)}}{\text{(median of every measurement taken for gene A throughout the experiment}}$$



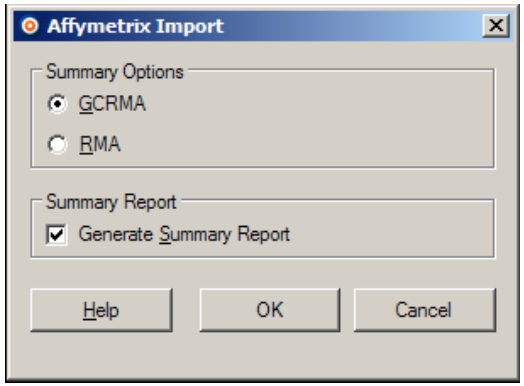
# TIBCO® Spotfire® DecisionSite® 9.1 for Microarray Analysis





http://spotfire.tibco.com/

Affymetrix CEL File Import Summarization Dialog




## Data Transformation

- Normalize by mean
- Normalize by percentile
- Normalize by trimmed mean
- Normalize by Z-score
- …

- Column normalization
- Row summation

The normalized value of $e_i$ for variable E in the i$^{th}$ record is calculated as

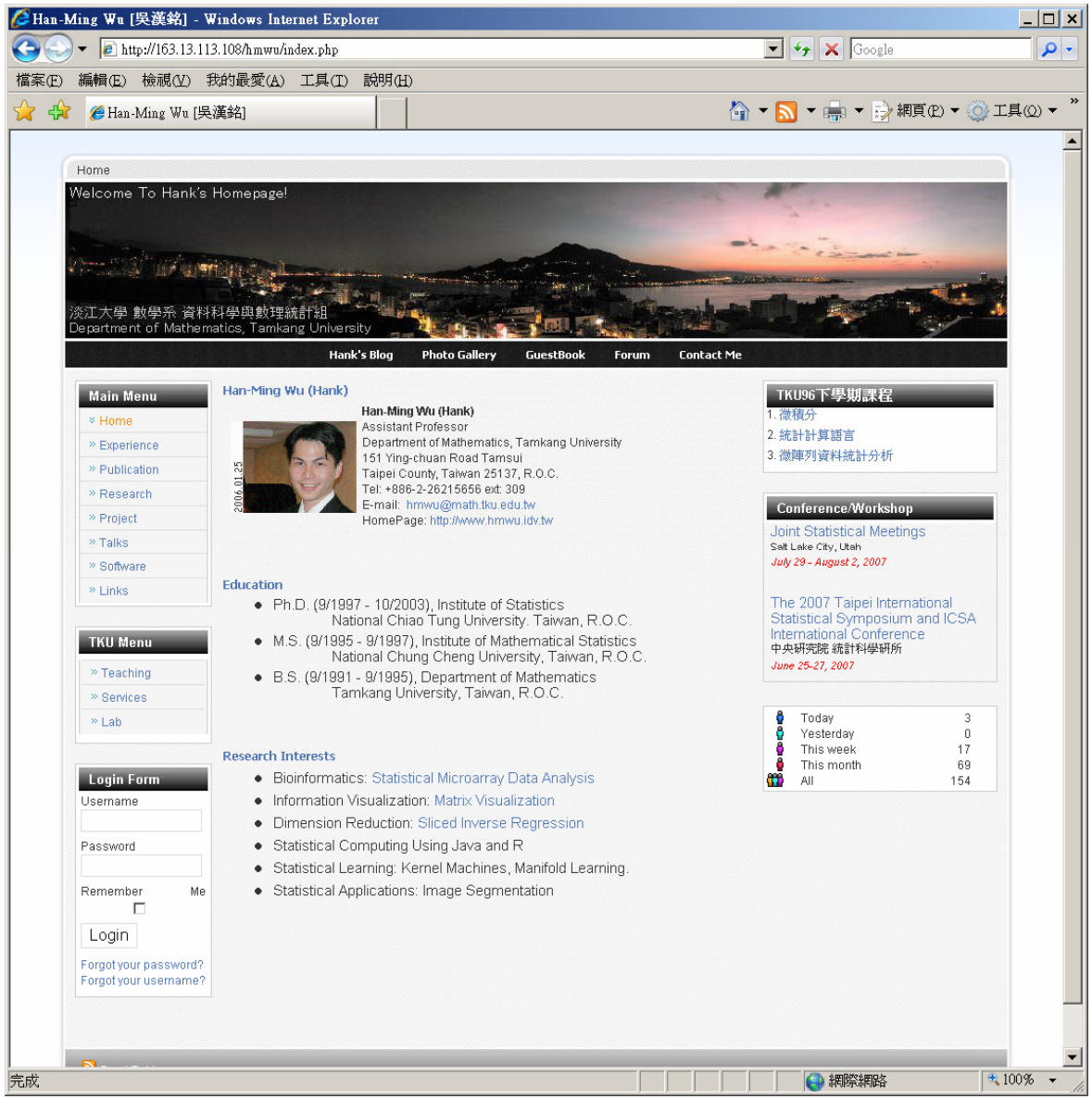
$$\text{Normalized}(e_i) = \frac{e_i}{\frac{1}{p}\sum_{j=1}^{p} e_j}$$

$$\text{Normalized}(e_i) = \frac{e_i}{q_{E,X\%}}$$

$$\text{Normalized}(e_i) = \frac{e_i - \bar{E}}{\text{std}(E)}$$

# Questions?






# Thanks!

吳漢銘
hmwu@math.tku.edu.tw
http://www.hmwu.idv.tw