

Microarray Data Preprocessing

Affymetrix GeneChip

國立臺灣大學 資訊所

Course: 生物資訊之統計與計算方法

2007/03/29

吳漢銘

hmwu@stat.sinica.edu.tw

<http://idv.sinica.edu.tw/hmwu/>

Institute of Statistical Science, Academia Sinica

中央研究院 統計科學研究所

Outlines

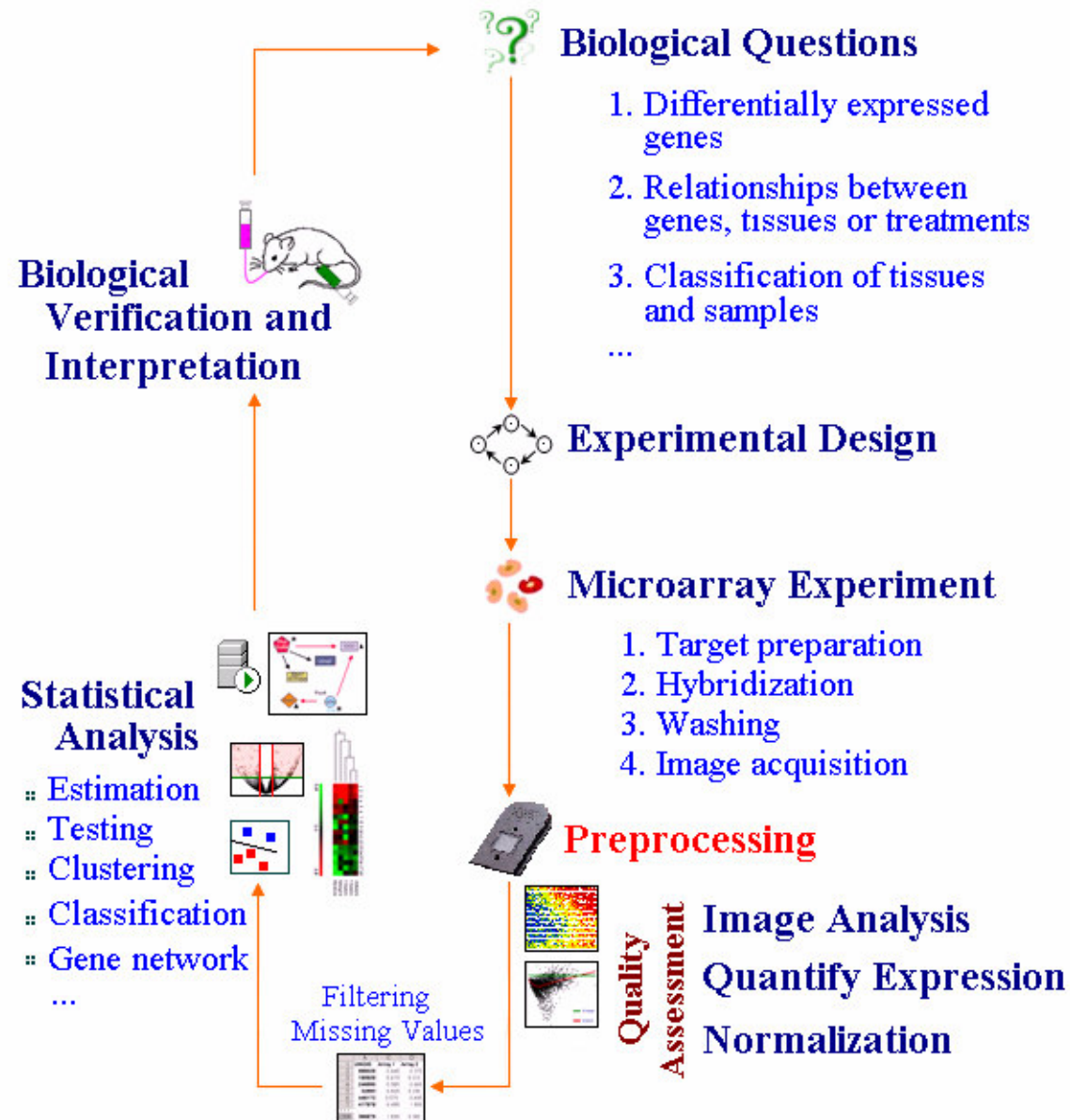
- **Affymetrix GeneChip Technology**
- **Assay and Analysis Flow Chart**
- **Quality Assessment**
- **Low Level Analysis**
(from probe level data to expression value)
- **Reproducibility and False Positive Rates**
- **Software**
- **Useful Links and Reference**



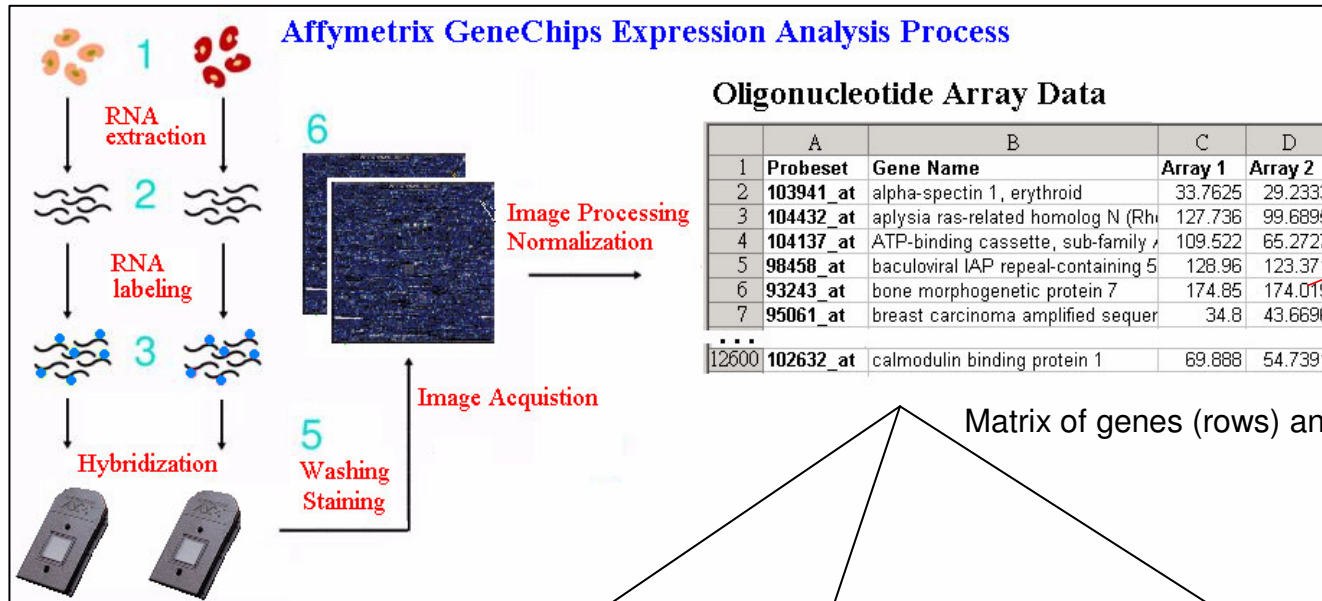
Affymetrix Dominates DNA Microarrays Market (75%~85%)

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

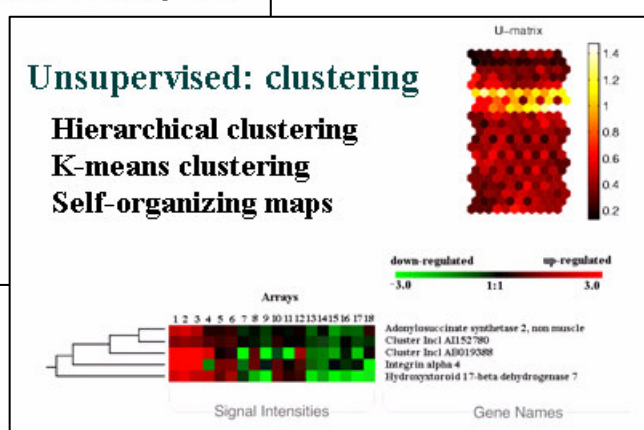
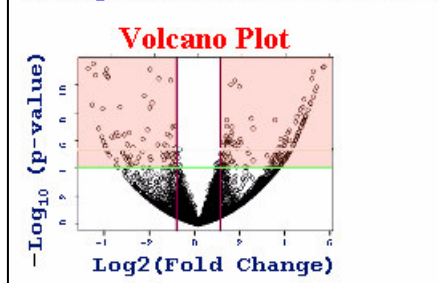
Microarray Life Cycle



Overview of Microarray Analysis



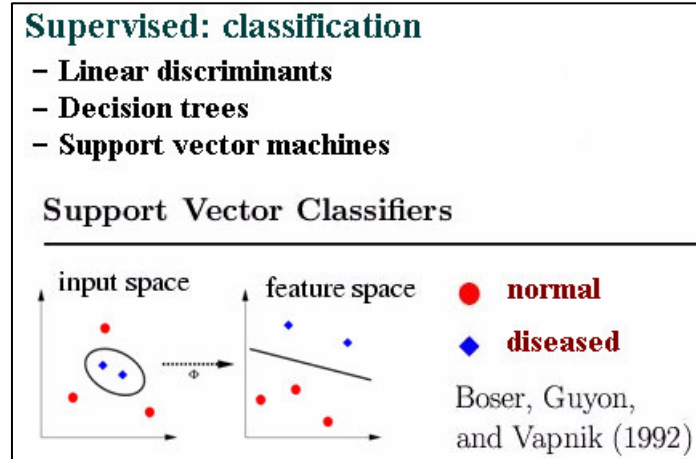
Discovery of differentially expressed genes
Parametric : t-test
Non-parametric : Wilcoxon, Mann-Whitney test



Supervised: classification

- Linear discriminants
- Decision trees
- Support vector machines

Support Vector Classifiers

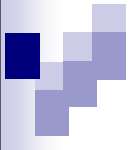


input space

feature space

● normal
 ◆ diseased

Boser, Guyon, and Vapnik (1992)



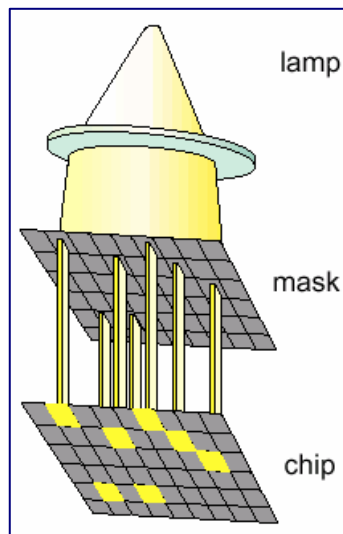
Affymetrix GeneChip Technology

GeneChip Photolithography

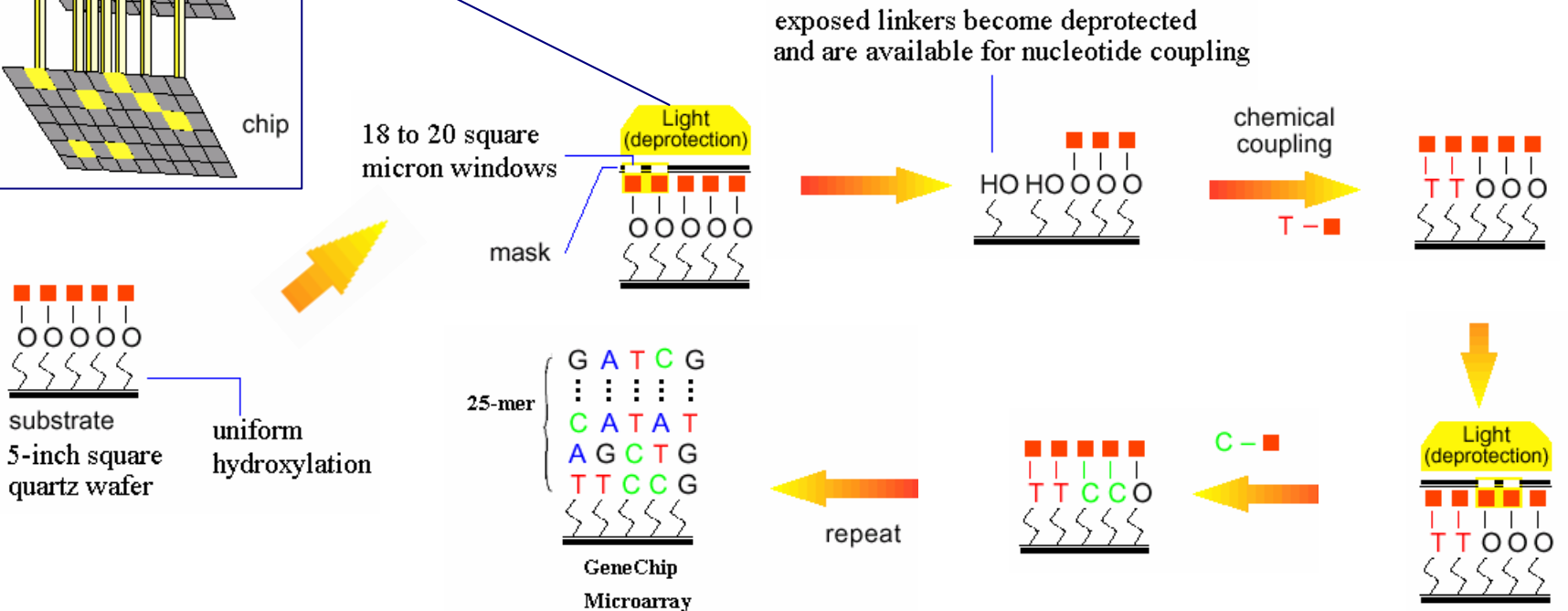


1. 在要作為晶片的玻璃版上放一層對光敏感的標記分子 (X)
 - 將光當作合成反應中的活化物。
 - 這些標記分子經過光照後可以形成羥基 (-OH)。
 - 這些羥基可與核甘中的鹼基序列結合起來，合成一段DNA序列。

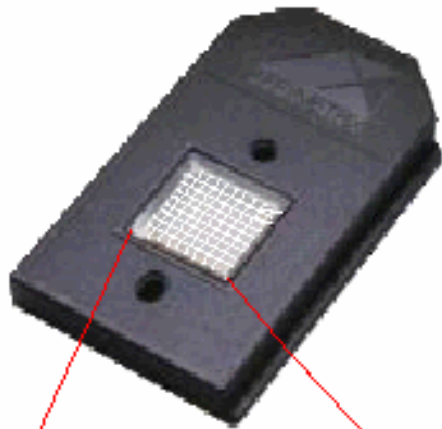
2. 藉由石版照明面罩(Photolithography Mask) 調控照光與不照光的區域
 - 把不接上某個鹼基的部份蓋住。
 - 沒有蓋住的部份經光照後即可形成羥基。



3. 事先將核甘序列中的鹼基 (A、T、C、G) 經過修飾成 3-O-phosphoramidite-activated deoxynucleoside，並在其5'端的羥基處以光標記物質加以保護
 - 在此四個鹼基中選一個，令其流經玻璃表面，此鹼基會與羥基的部份結合起來。
 - 之後再蓋住其他部份，使其他未形成羥基的部份經過光照後產生羥基。
 - 再以另一個鹼基流過玻璃片，重複這些步驟，以接出含各種鹼基序列的核甘序列。



GeneChip Expression Array Design



1.28cm

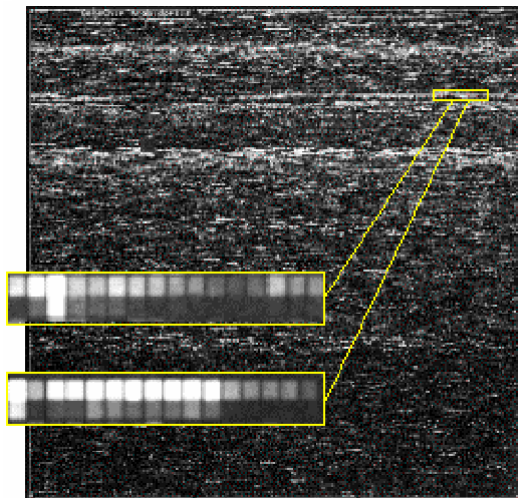
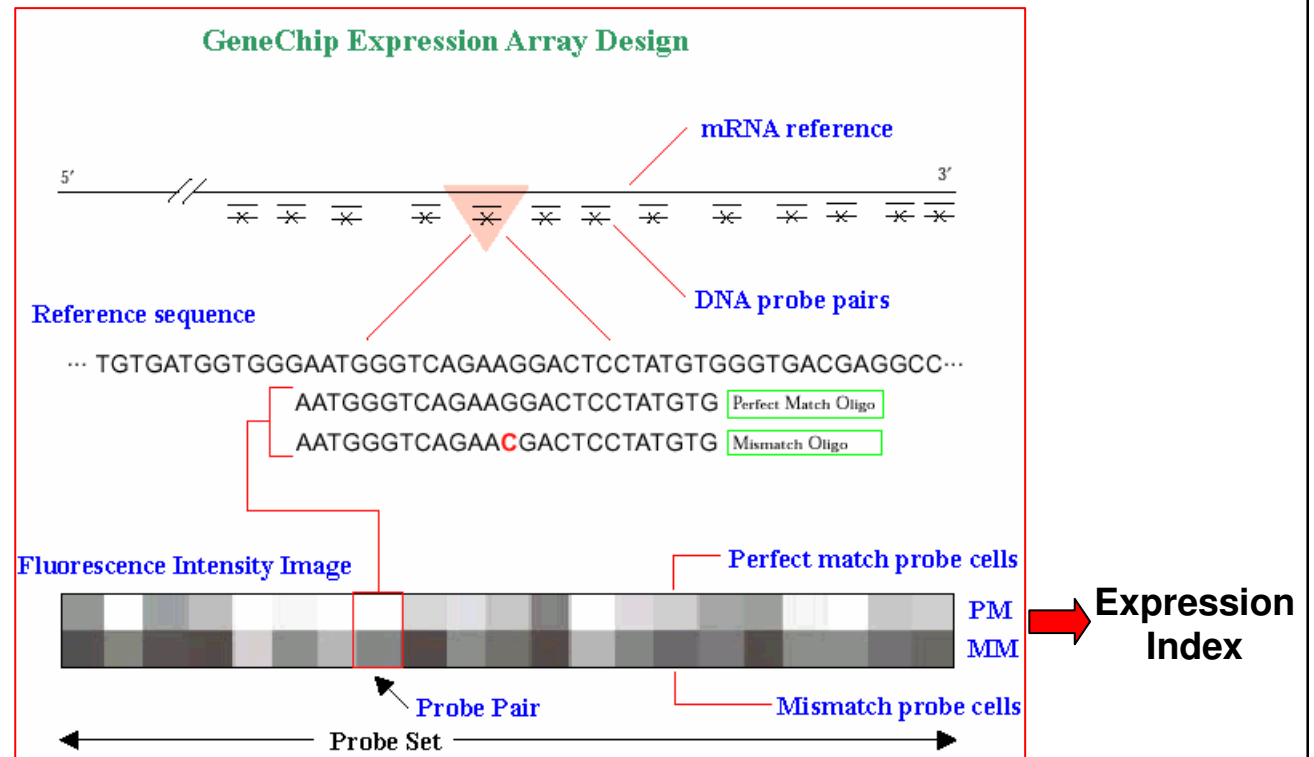
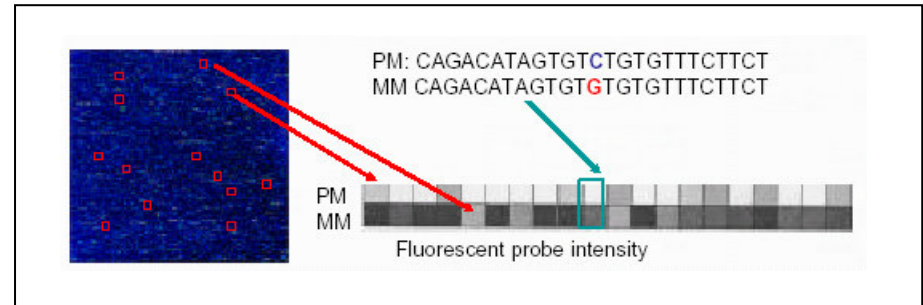


Image of hybridized probe array



GeneChip Expression Analysis Process

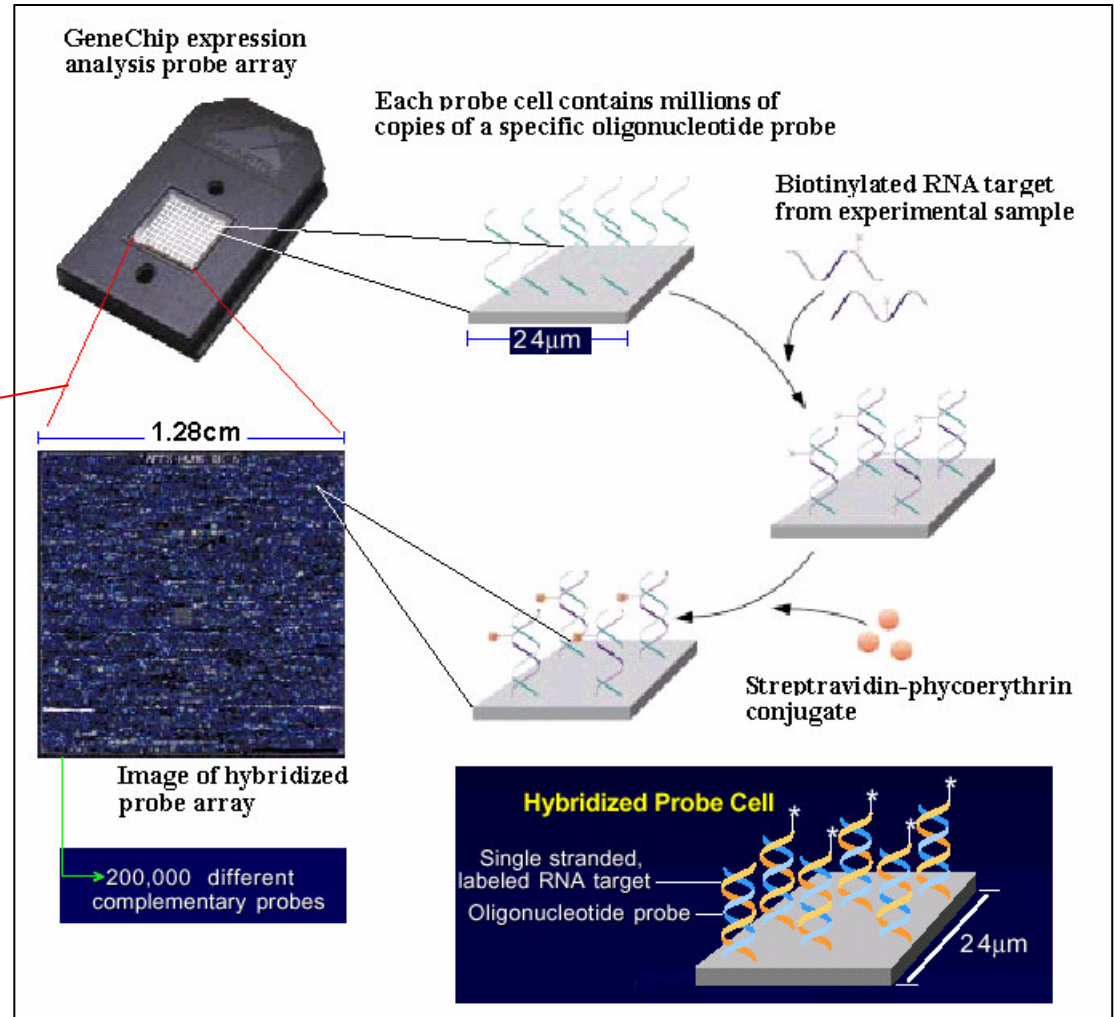
The GeneChip® Instrument System



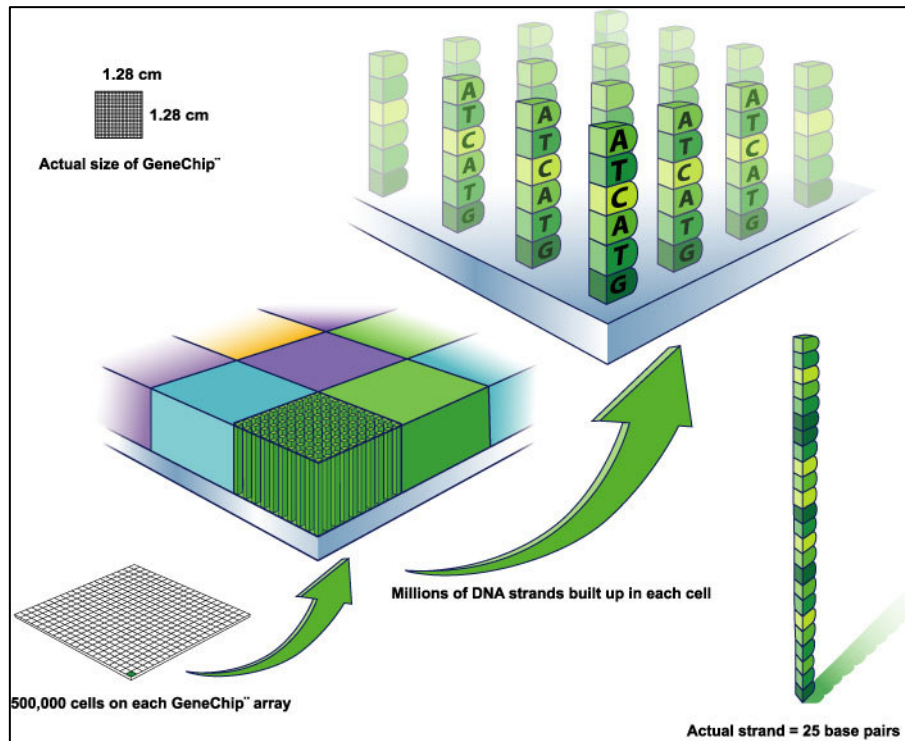
Scan and Quantitate



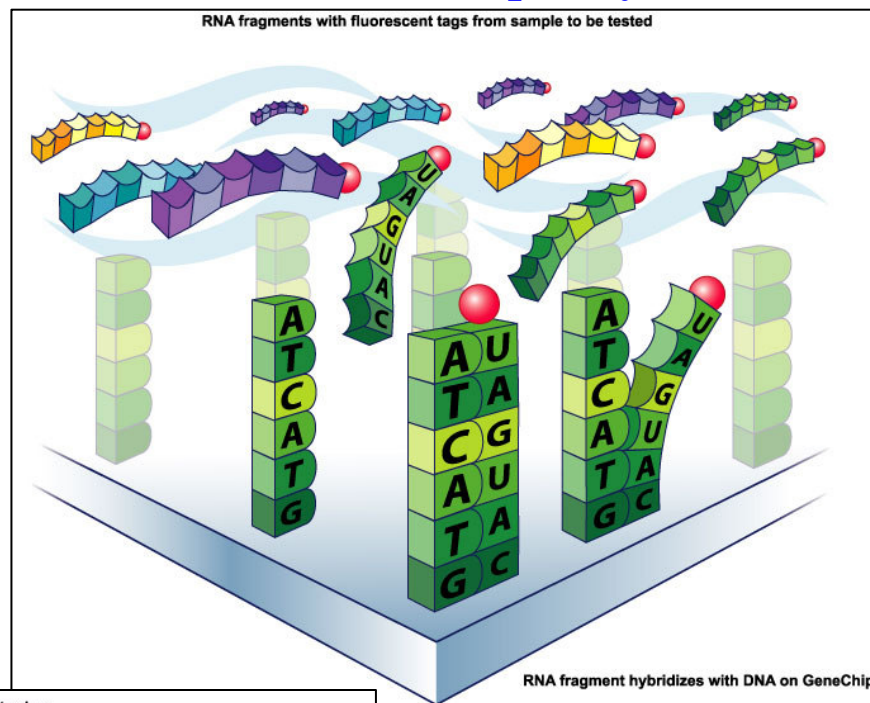
Affymetrix GeneChip®
Scanner 3000 with workstation.



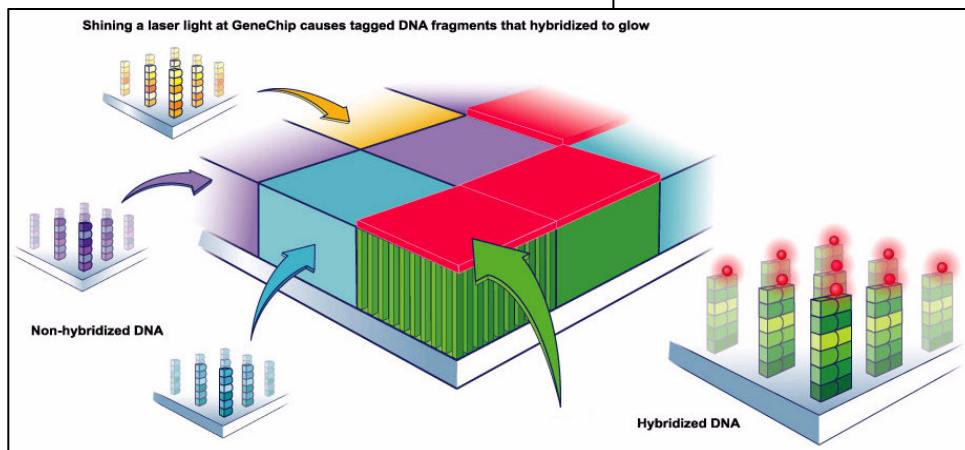
More Figures on Affymetrix Web Site



GeneChip® Hybridization



GeneChip® Single Feature



Hybridized
GeneChip®
Microarray

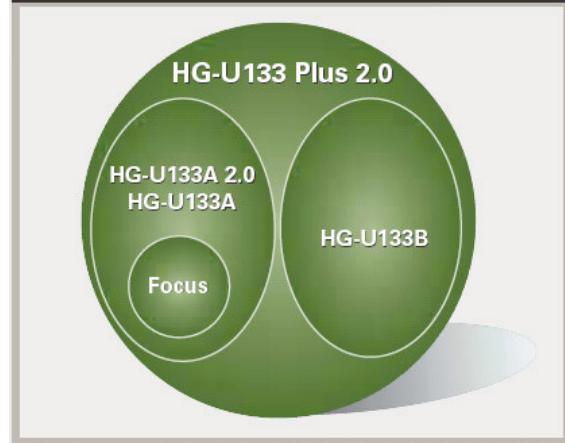
GeneChip® Human Genome Array

Human Affy arrays

Year	Array Name	Genes/Transcripts
1995	HuFL	6800 genes
1998	U95 set	A,B,C,D,E arrays 63,000 genes
2001	U133set	A,B arrays 44,000 transcripts
2003	U133 2.0	

Source: Genomics Core lab, MSKCC

Figure 1. Relationship Among GeneChip® Human Genome Arrays



Critical Specifications for GeneChip® Human Genome Arrays

	Human Genome U133 Plus 2.0 Array	Human Genome U133A 2.0 Array	Human Genome U133 Set	Human Genome Focus Array
Number of arrays in set	1	1	2	1
Number of transcripts	~47,400	18,400	~39,000	~8,500
Number of genes	38,500	14,500	~33,000	~8,400
Number of probe sets	>54,000	>22,000	>45,000	>8,700
Feature size	11 µm	11 µm	18 µm	18 µm
Oligonucleotide probe length	25-mer	25-mer	25-mer	25-mer
Probe pairs/sequence	11	11	11	11
Control sequences included:				
Hybridization controls	<i>bioB, bioC, bioD, cre</i>	<i>bioB, bioC, bioD, cre</i>	<i>bioB, bioC, bioD, cre</i>	<i>bioB, bioC, bioD, cre</i>
Poly-A controls	<i>dap, lys, phe, thr</i>	<i>dap, lys, phe, thr</i>	<i>dap, lys, phe, thr</i>	<i>dap, lys, phe, thr</i>
Normalization control set	100 probe sets	100 probe sets	100 probe sets	100 probe sets
Housekeeping/Control genes	GAPDH, beta-Actin, ISGF-3 (STAT1)	GAPDH, beta-Actin, ISGF-3 (STAT1)	GAPDH, beta-Actin, ISGF-3 (STAT1)	GAPDH, beta-Actin, ISGF-3 (STAT1)
Detection sensitivity	1:100,000*	1:100,000*	1:100,000*	1:100,000*

*As measured by detection of pre-labeled transcripts derived from human cDNA clones in a complex human background.

Terms & Descriptions



11/69

- **Target:** the labeled sample applied to the array (consists of cRNA in vitro transcribed from cDNA which was in turn reverse transcribed from total mRNA extracted from the sample).
- **Background (BG):** a measure of the magnitude of background. For each of 16 sectors, the average intensity of features with intensities falling in the lowest 2% of features within the sector.
- **Noise:** a measure of the variance in background.
- **Feature (Probe):** a 24-50 nm portion of the array on which are synthesized ~10⁷ molecules of a single oligonucleotide. A scan generates one pixel for every 3mm².
- **Perfect match (PM):** an oligonucleotide (~25bp) specific for a region of the cRNA of a gene.
- **Mismatch (MM):** an oligonucleotide (~25bp) specific for a region of the cRNA of a gene with a single mismatched nucleotide in the centre location - always paired with a PM.
 - Capture non-specific hybridization
 - Problem: Approximately 30% of the mismatches are greater than their corresponding PM.
- **Probe pair:** a pair of probes, one PM and its corresponding MM.
- **Probe set:** a set of 20 probe pairs designed to probe for the transcript of a single gene.
Correspond to genes, gene fragments, or ESTs.
- **Fold change (FC):** the magnitude of change observed in a gene's expression from one scan to another.
- **Metrics** - The calculated answer of mathematical equations used by the GeneChip® probe array algorithm software.

Animations



12/69

The Structure of a GeneChip® Microarray

How to Use GeneChip® Microarrays to Study Gene Expression

http://www.affymetrix.com/corporate/outreach/lesson_plan/educator_resources.affx

<http://www.affymetrix.com/corporate/outreach/educator.affx>

Genisphere

http://www.genisphere.com/ed_data_ref.html

HHMI (Howard Hughes Medical Institute)

<http://www.hhmi.org/biointeractive/genomics/video.html>

<http://www.hhmi.org/biointeractive/genomics/animations.html>

<http://www.hhmi.org/biointeractive/genomics/click.html>

DNA Interactive Site from Cold Spring Harbor Labs

<http://www.dnai.org/index.htm>

"Applications", => "Genes and Medicine" => "Genetic Profiling"

Digizyme - Web & Multimedia Design for the Sciences

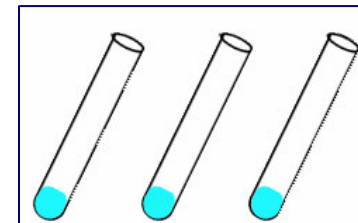
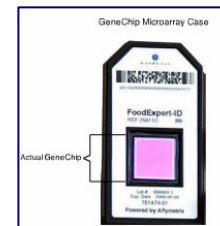
<http://www.digizyme.com/>

<http://www.digizyme.com/portfolio/microarraysfab/index.html>

<http://www.digizyme.com/competition/examples/genechip.swf>

DNA Microarray Virtual Lab

<http://learn.genetics.utah.edu/units/biotech/microarray>





Assay and Analysis Flow Chart

- Image Analysis
- Affymetrix Data Files
- From DAT to CEL

Assay and Analysis Flow Chart



Hybridization + Scanning

EXP File

Experiment Information File



DAT File

Data File:
the image of the scanned array

Image analysis



Cell Intensity File

CEL File

+

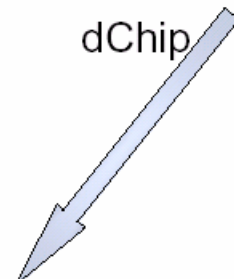
Chip Description Files

CDF File

Preprocessing

1. Background Correction
2. Normalization
3. PM Correction
4. Expression Index

dChip



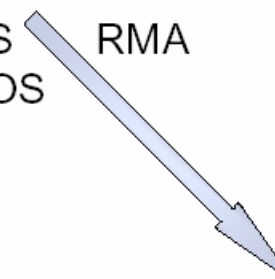
Excel File

MAS
GCOS



CHP File
Intensity value
Absent / Present call

RMA



Text File
Probe ID +
 $\text{Log}_2(\text{Intensity})$

RPT File

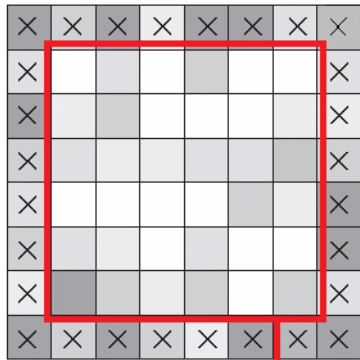


Report File, quality

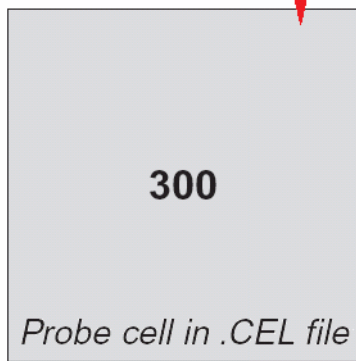
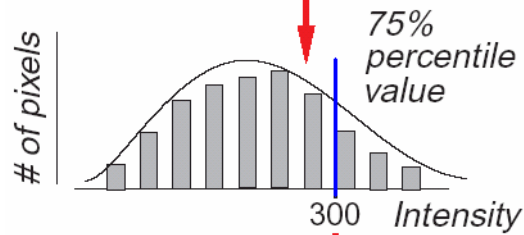
source:

UCSF Shared Functional
Genomics Core Facility

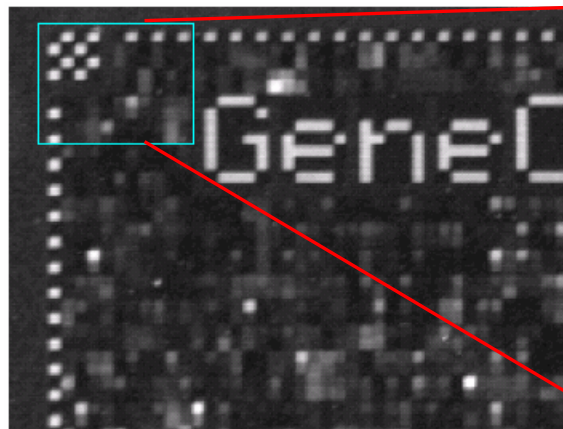
From DAT to CEL



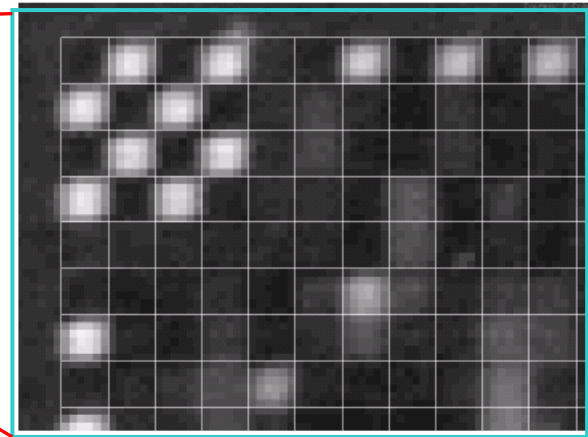
Probe cell in .DAT file



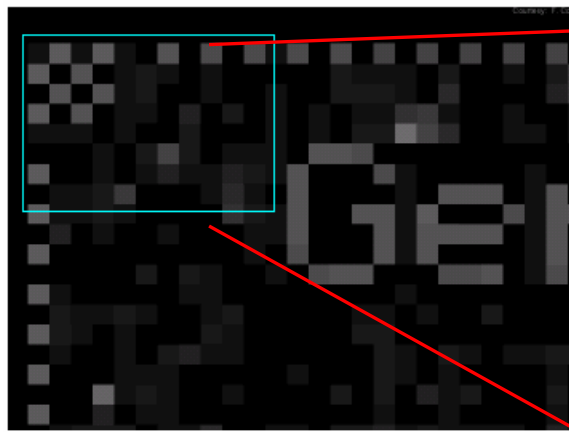
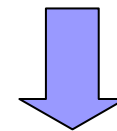
Probe cell in .CEL file



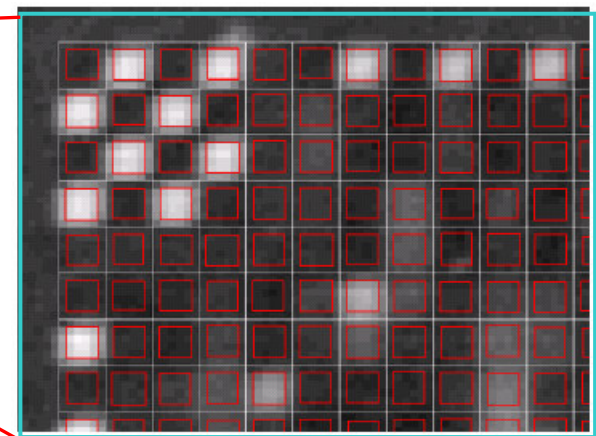
DAT



DAT + Grid



CEL



DAT + Grid - Outer Pixel

MAS5.0 Analysis Output File (*.CHP)



	Analysis Name	Probe Set Name	Stat Pairs	Stat Pairs Used	Signal	Detection	Detection p-value	Stat Comr
1	030606 En test3	Pae_16SrRNA_s_at	16	16	11.3	A	0.872355	
2	030606 En test3	Pae_23SrRNA_s_at	16	16	26.6	A	0.378184	
3	030606 En test3	PA1178_oprH_at	12	12	5.4	A	0.975070	
4	030606 En test3	PA1816_dnaQ_at	12	12	5.9	A	0.805907	
5	030606 En test3	PA3183_zwf_at	12	12	7.9	A	0.708540	
6	030606 En test3	PA3640_dnaE_at	12	12	10.8	A	0.964405	
7	030606 En test3	PA4407_ftsZ_at	12	12	9.5	A	0.921030	
8	030606 En test3	Pae_16SrRNA_s_st	16	16	8.9	A	0.660442	
9	030606 En test3	Pae_23SrRNA_s_st	16	16	22.0	A	0.561639	
10	030606 En test3	PA1178_oprH_st	12	12	35.1	P	0.024930	
11	030606 En test3	PA1816_dnaQ_st	12	12	34.7	A	0.240088	
12	030606 En test3	PA3183_zwf_st	12	12	6.5	A	0.985972	
13	030606 En test3	PA3640_dnaE_st	12	12	87.5	A	0.173261	
14	030606 En test3	PA4407_ftsZ_st	12	12	47.5	A	0.623158	
15	030606 En test3	AFFX-Athal-Actin_5_r_at	16	16	89.8	P	0.013092	

Metrics

	030606 En test3		Descriptions
	Signal	Detection	
Pae_16SrRNA_s_at	11.3	A	
Pae_23SrRNA_s_at	26.6	A	
PA1178_oprH_at	5.4	A	
PA1816_dnaQ_at	5.9	A	
PA3183_zwf_at	7.9	A	
PA3640_dnaE_at	10.8	A	
PA4407_ftsZ_at	9.5	A	
Pae_16SrRNA_s_st	8.9	A	
Pae_23SrRNA_s_st	22.0	A	
PA1178_oprH_st	35.1	P	
PA1816_dnaQ_st	34.7	A	
PA3183_zwf_st	6.5	A	
PA3640_dnaE_st	87.5	A	
PA4407_ftsZ_st	47.5	A	

Pivot



Quality Assessment

- RNA Sample Quality Control
- Array Hybridization Quality Control
- Statistical Quality Control
(Diagnostic Plots)

■ RNA Sample Quality Control

- *Validation of total RNA*
- *Validation of cRNA*
- *Validation of fragmented cRNA*

Two aspects of quality control: detecting poor hybridization and outliers

■ Array Hybridization Quality Control

- Probe Array Image Inspection (DAT, CEL)
- B2 Oligo Performance
- MAS5.0 Expression Report Files (RPT)
 - Scaling and Normalization factors
 - Average Background and Noise Values
 - Percent Genes Present
 - Housekeeping Controls: Internal Control Genes
 - Spike Controls: Hybridization Controls: bioB, bioC, bioD, cre
 - Spike Controls: Poly-A Control: dap, lys, phe, thr, trp

■ Statistical Quality Control (Diagnostic Plots)

◆ Reasons for poor hybridizations

- mRNA degenerated
- one or more experimental steps failed
- poor chip quality, ...

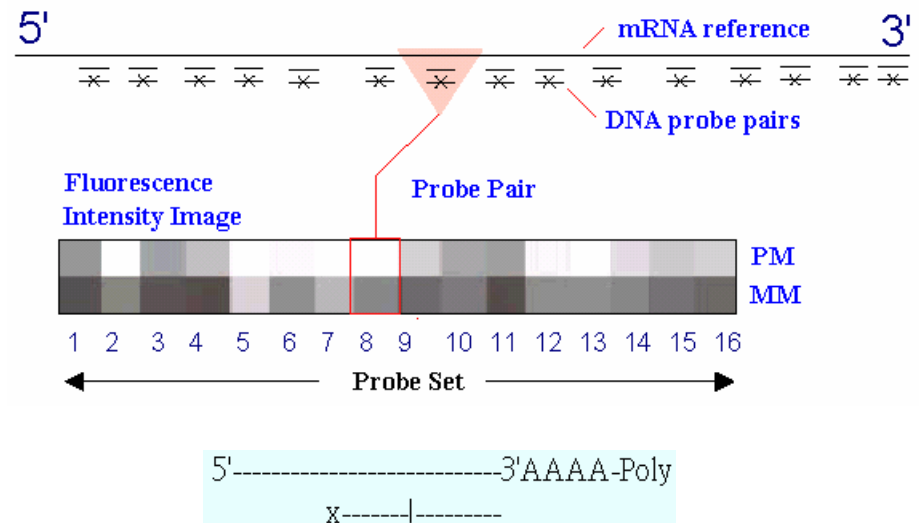
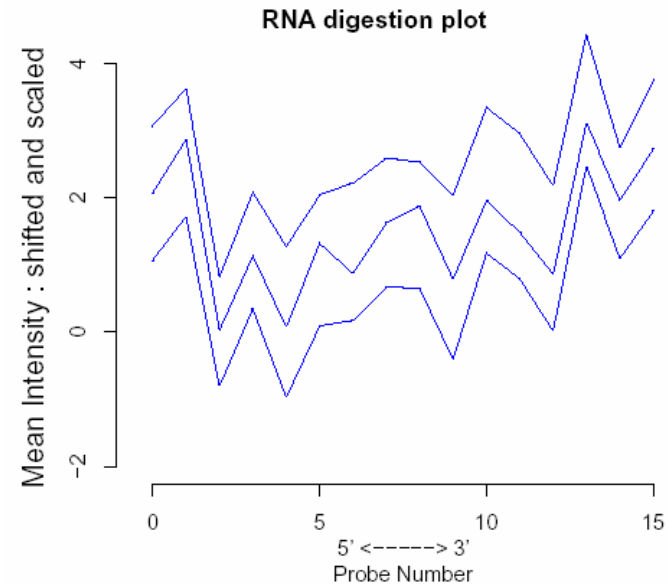
◆ reasons for (biological) outliers

- infiltration with non-tumour tissue
- wrong label
- contamination, ...

RNA Degradation Plots

Assessment of RNA Quality:

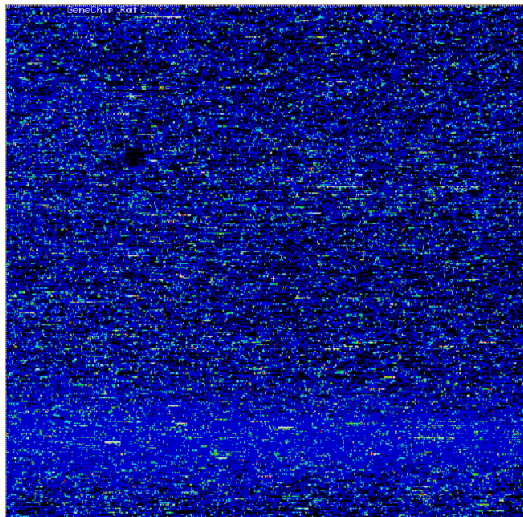
- Individual probes in a probe set are ordered by location relative to the 5' end of the targeted RNA molecule.
- Since RNA degradation typically starts from the 5' end of the molecule, **we would expect probe intensities to be systematically lowered at that end of a probeset when compared to the 3' end.**
- On each chip, probe intensities are averaged by location in probeset, with the average taken over probesets.
- The RNA degradation plot produces a side-by-side plots of these means, making it easy to notice any 5' to 3' trend.



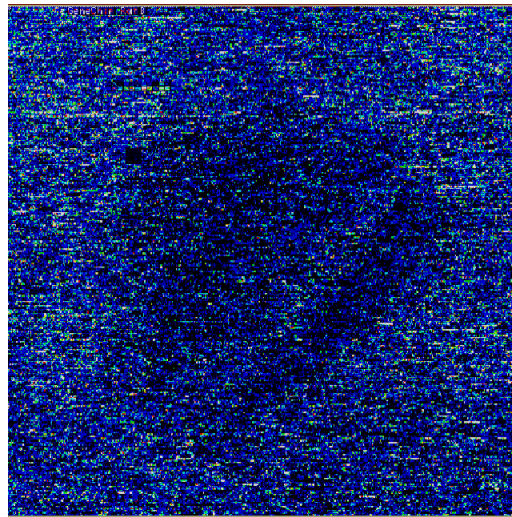
Probe Array Image Inspection

- 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.

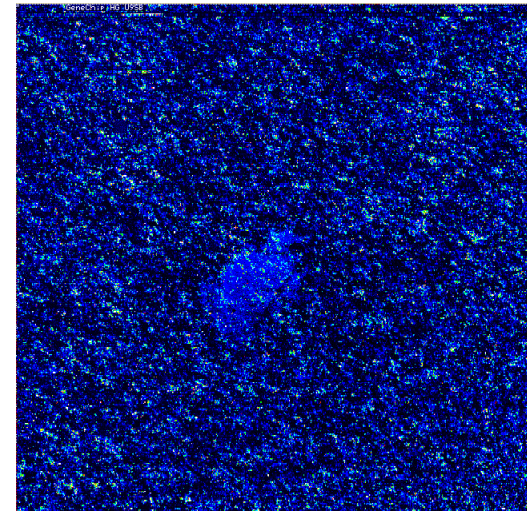
Haze Band



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.

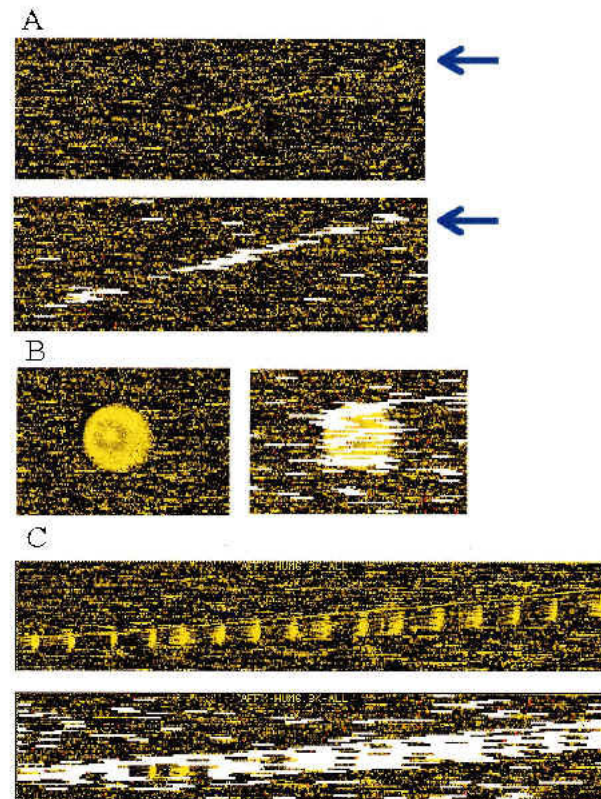


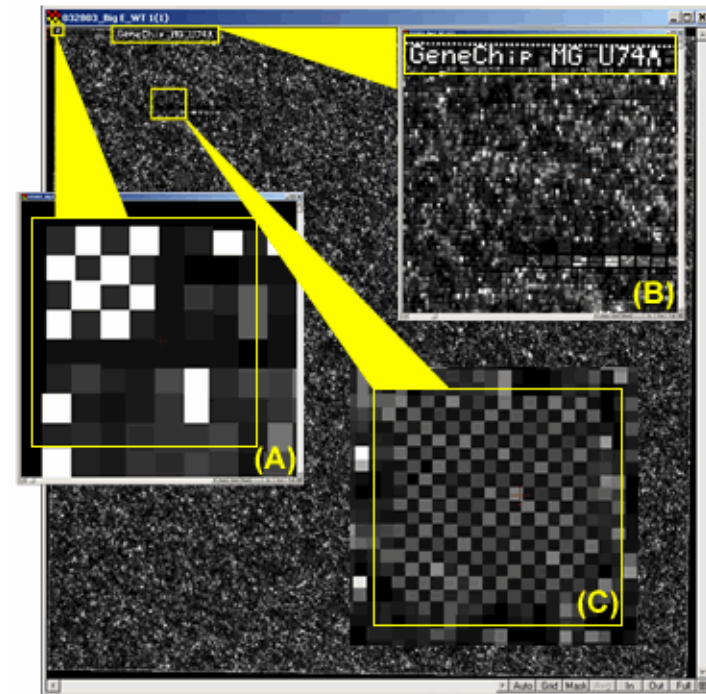
Fig. 5. (A) A long scratch contamination (indicated by arrow) is alleviated by automatic outlier exclusion along this scratch. (B and C) Regional clustering of array outliers (white bars) indicates contaminated regions in the original images. These outliers are automatically detected and accommodated 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



23/69

- 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

MAS5.0 Expression Report File (*.RPT)

Report Type: Expression Report
Date: 04:42PM 02/24/2004

Filename: test.CHIP
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.

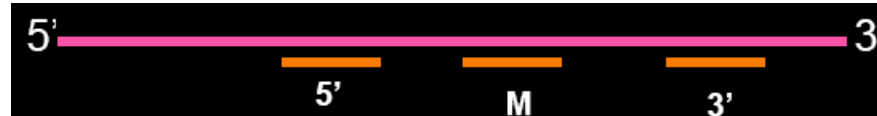
- Average Background: 20-100
- Noise < 4

- 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%.
- 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 and B-ACTIN. In general, they should be less than three.



- 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.

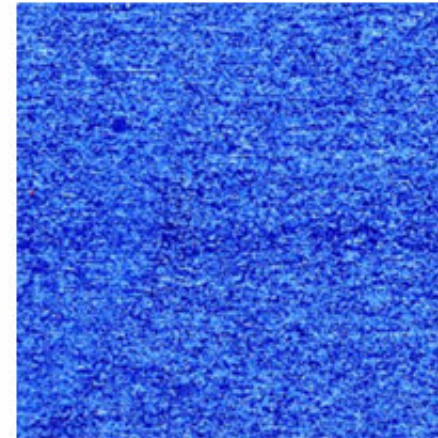
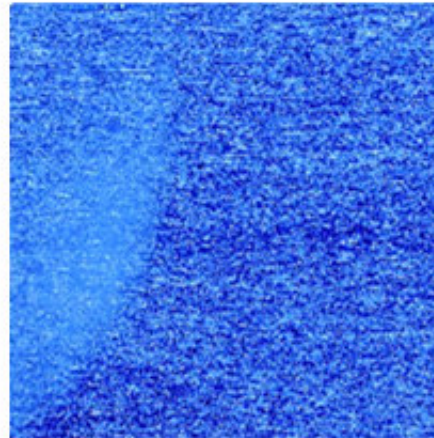
Housekeeping Controls:								
Probe Set	Sig(5')	Det(5')	Sig(M')	Det(M')	Sig(3')	Det(3')	Sig(all)	Sig(3'/5')
AFFX-HUMISGF3A/M97935	272.8	P	856.8	P	1274.5	P	801.36	4.67
AFFX-HUMRGE/M10098	340.6	M	181.3	A	632.6	P	384.80	1.86
AFFX-HUMGAPDH/M33197	13890.6	P	15366.6	P	14060.7	P	14439.32	1.01
AFFX-HSAC07/X00351	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

Statistical Plots

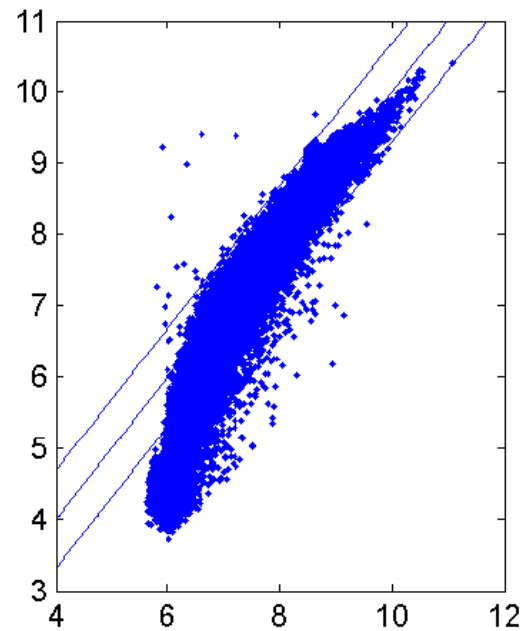
Gradient Correction

GeneChipImage

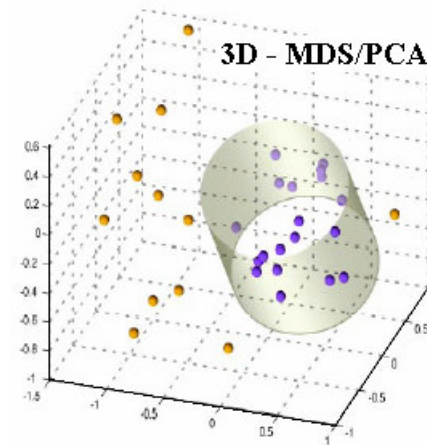


Before

After



Scatterplot



Dimension Reduction
(PCA, MDS)

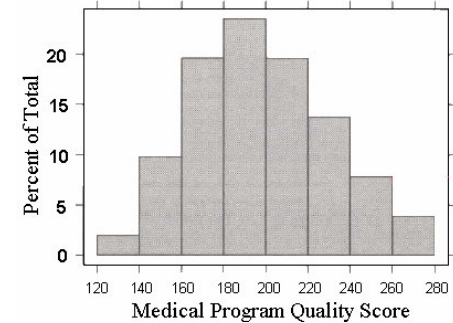
Statistical Plots: Histogram

- $1/2h$ 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.

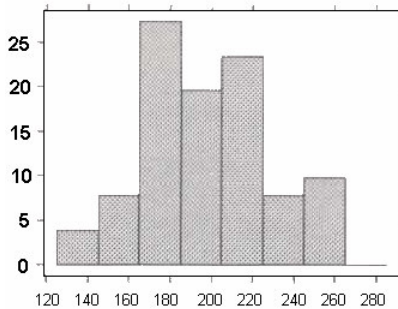
Disadvantage for displaying a variable's distribution:

- 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.

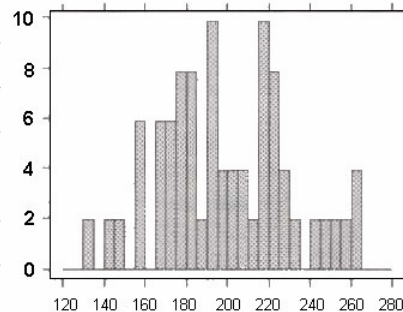
O. Bin origin at 120, bin widths of 20.



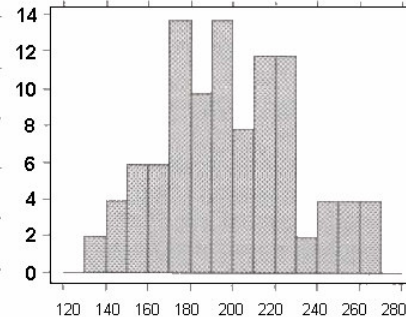
A. Bin origin at 125, bin widths of 20.



B. Bin origin at 120, bin widths of 5.

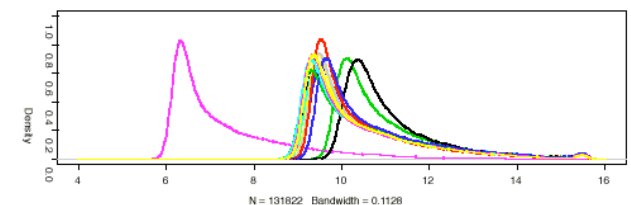


C. Bin origin at 120, bin widths of 10.



Density Plots

density(x = x[, 1], from = 4, to = 16)



density(x = y[, 1], from = 4, to = 16)

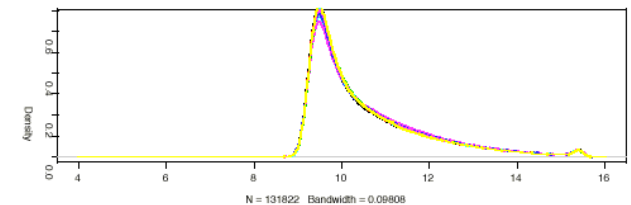
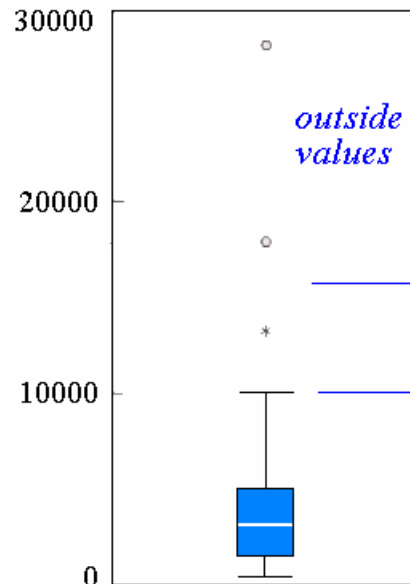
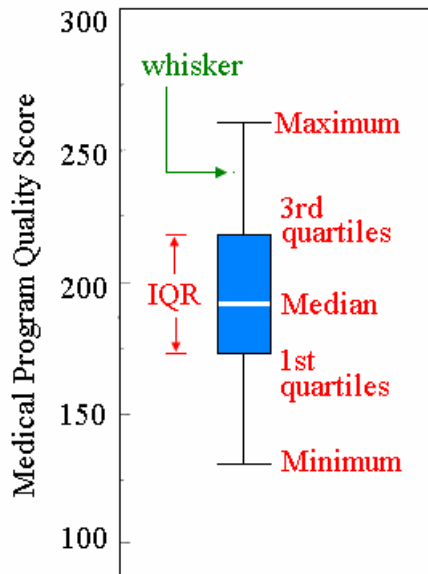
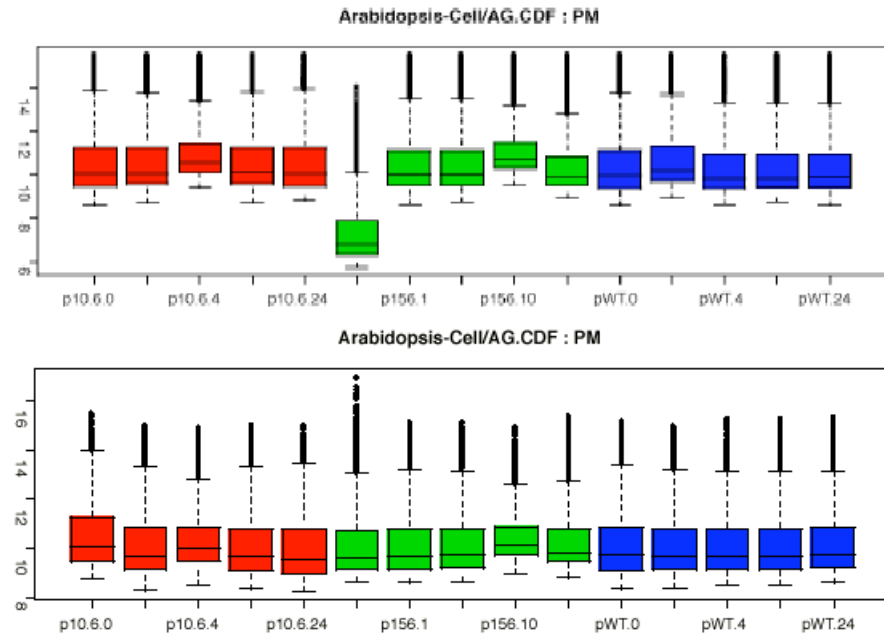


Figure Sources: Jacoby (1997).

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.



Upper Outer Fence:
 $x_{0.75} + 3 \text{ IQR}$

Upper Inner Fence:
 $x_{0.75} + 1.5 \text{ IQR}$

Lower Inner Fence:
 $x_{0.25} - 1.5 \text{ IQR}$

Lower Outer Fence:
 $x_{0.25} - 3 \text{ IQR}$

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:

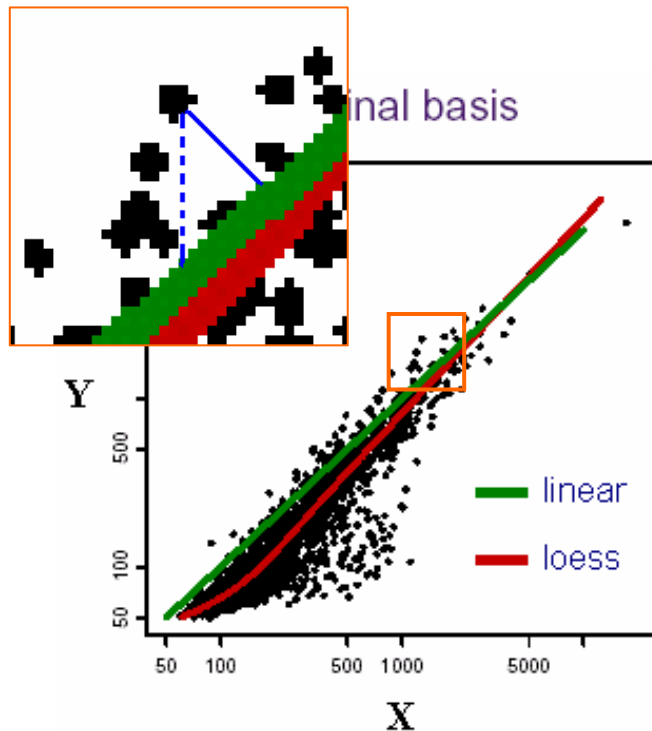
<http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm>

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)

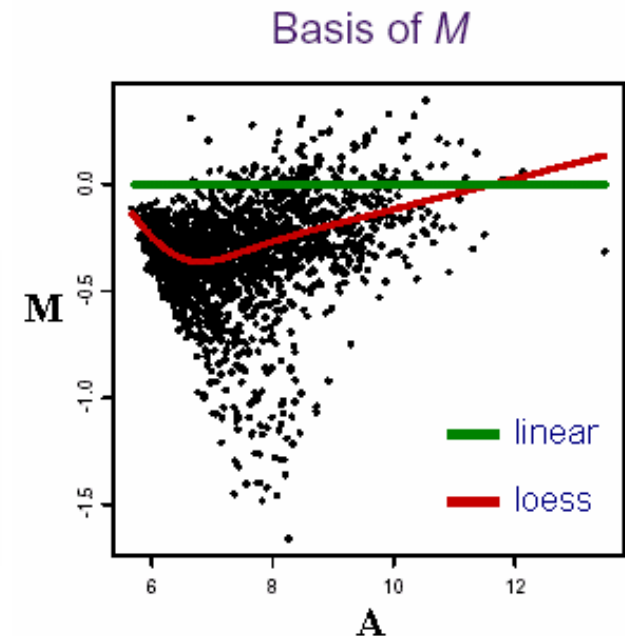
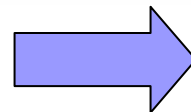
- **Goals:** to identify genes that are differentially regulated between two experimental conditions.



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

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

Oligo	cDNA
X = PM ₁ ,	X = Cy3
Y = PM ₂	Y = Cy5
X = PM ₁ · MM ₁ ,	
Y = PM ₂ · MM ₂	

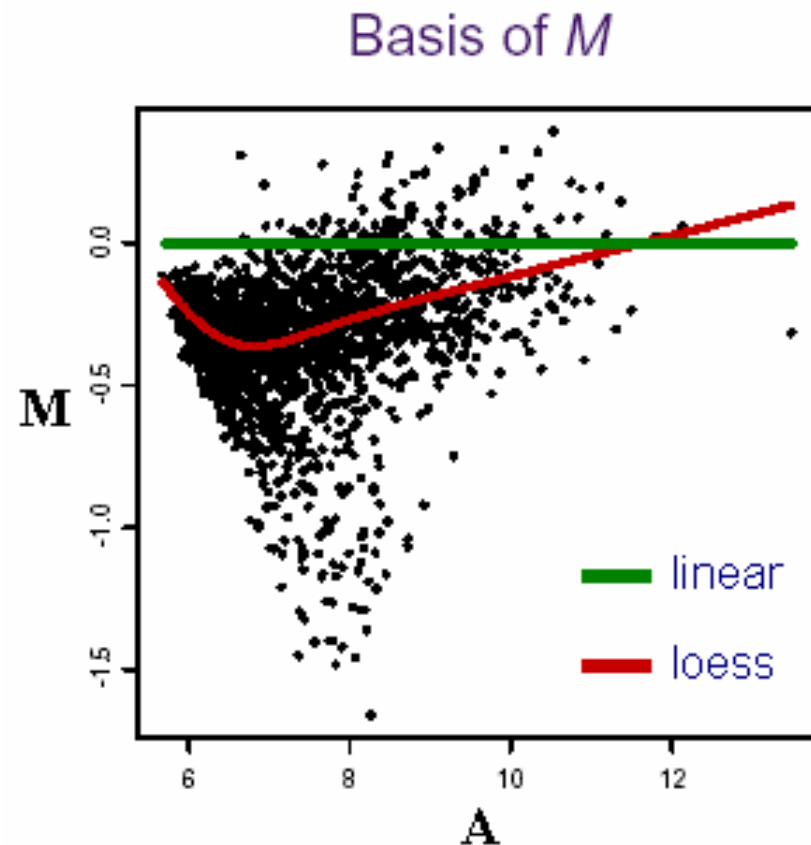


Scatterplot and MA plot (conti.)

- **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**

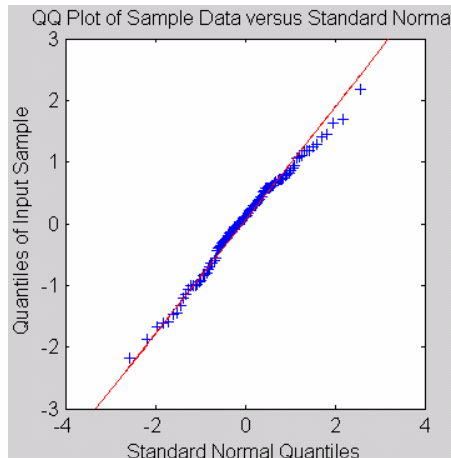
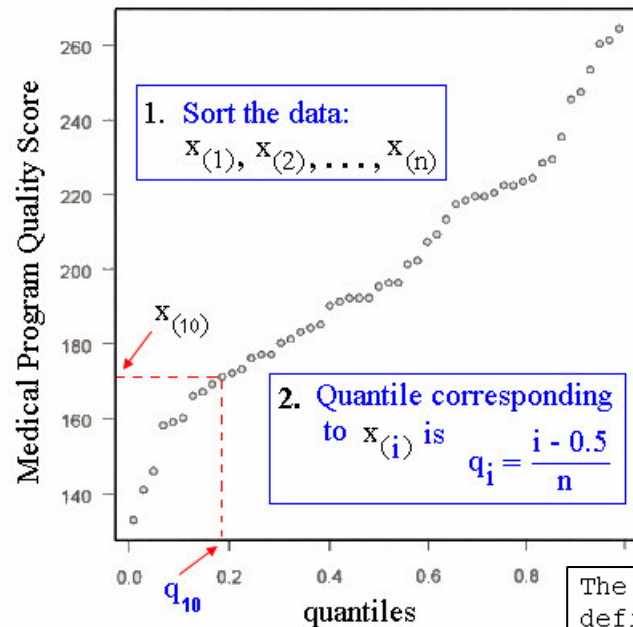
- spreads the data from the lower left corner to a more centered distribution in which the prosperities of the data are easy to analyze.
- easier to describe the fold regulation of genes using a log scale. In log₂ space, the data points are symmetric about 0.



Quantile Plots

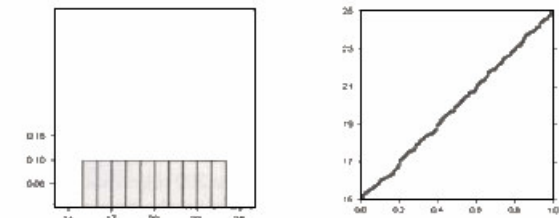


The empirical quantiles

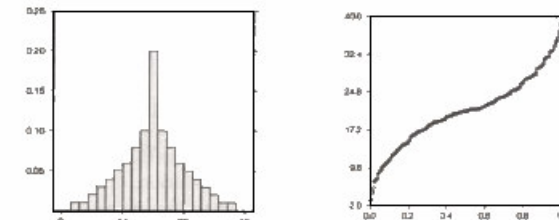


Comparison of histogram and Quantile plots for differently shaped data distribution

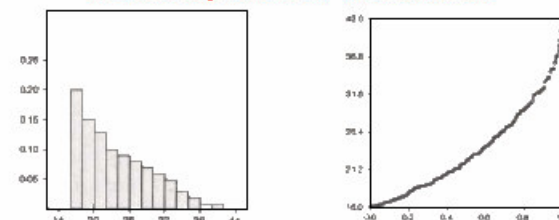
Uniform distribution



Symmetric, bell-shaped distribution



Positively skewed distribution



The q th 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.

- 0.5 is subtracted from each i value to avoid extreme quantiles of exactly 0 or 1.
- The latter would cause problems if empirical quantiles were to be compared against quantiles derived from a theoretical, asymptotic distribution such as the normal.
- This adjustment has no effect on the shape of any graphical display.

Figures modified from Jacoby (1997)

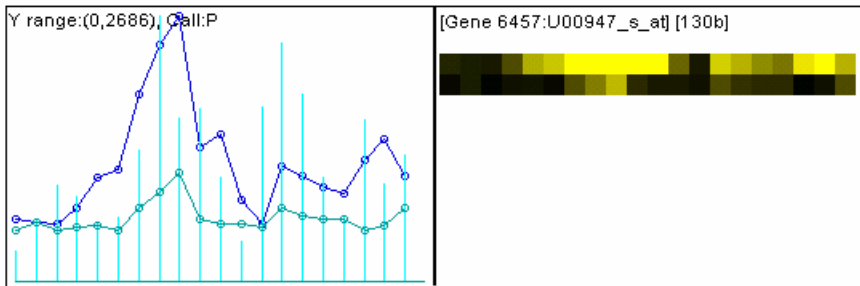
Statistical Plots



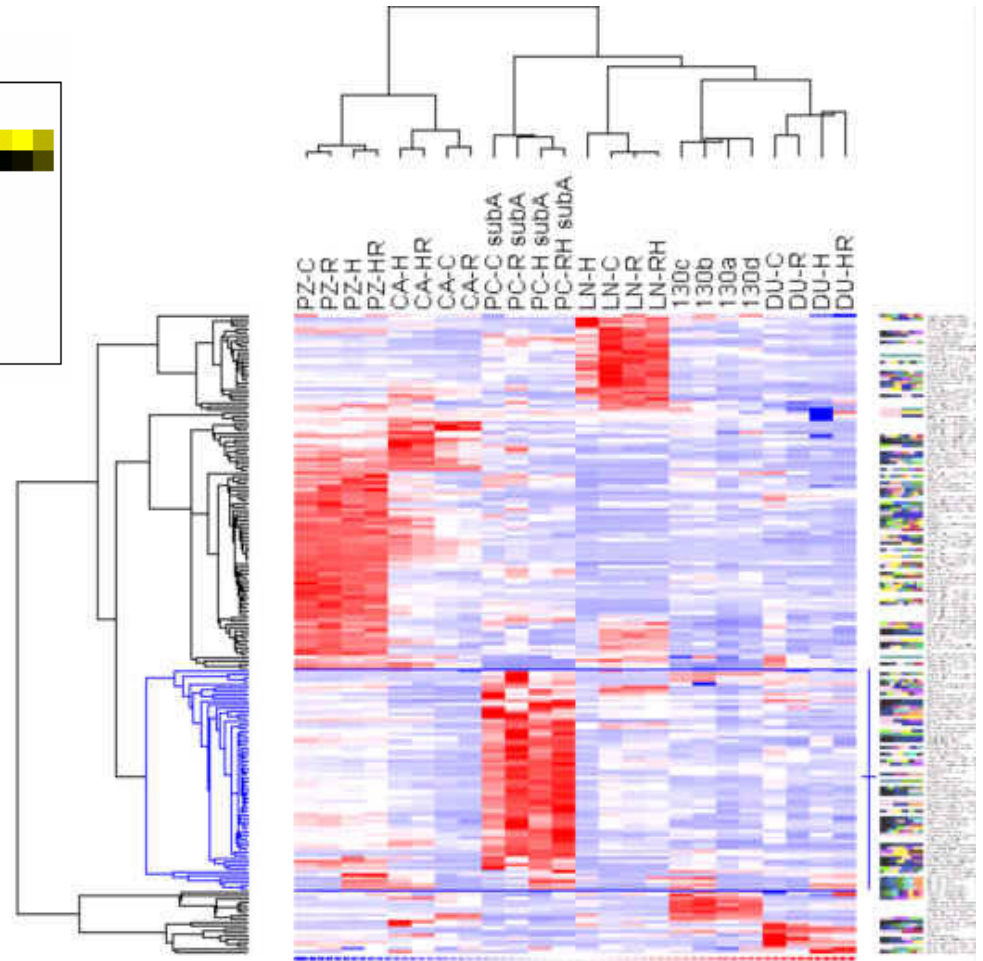
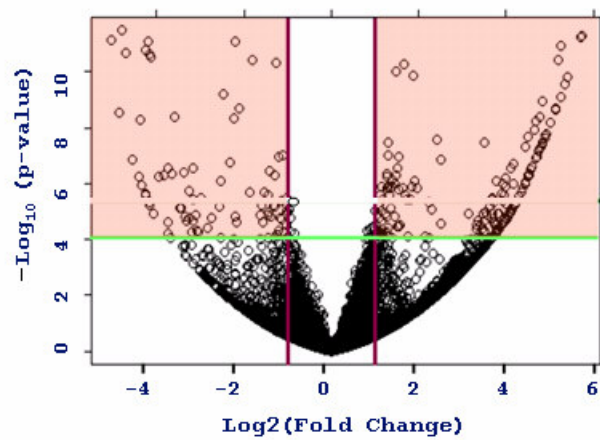
Heatmap with Dendrogram

Line Plots

Profiles Plots



Volcano Plot

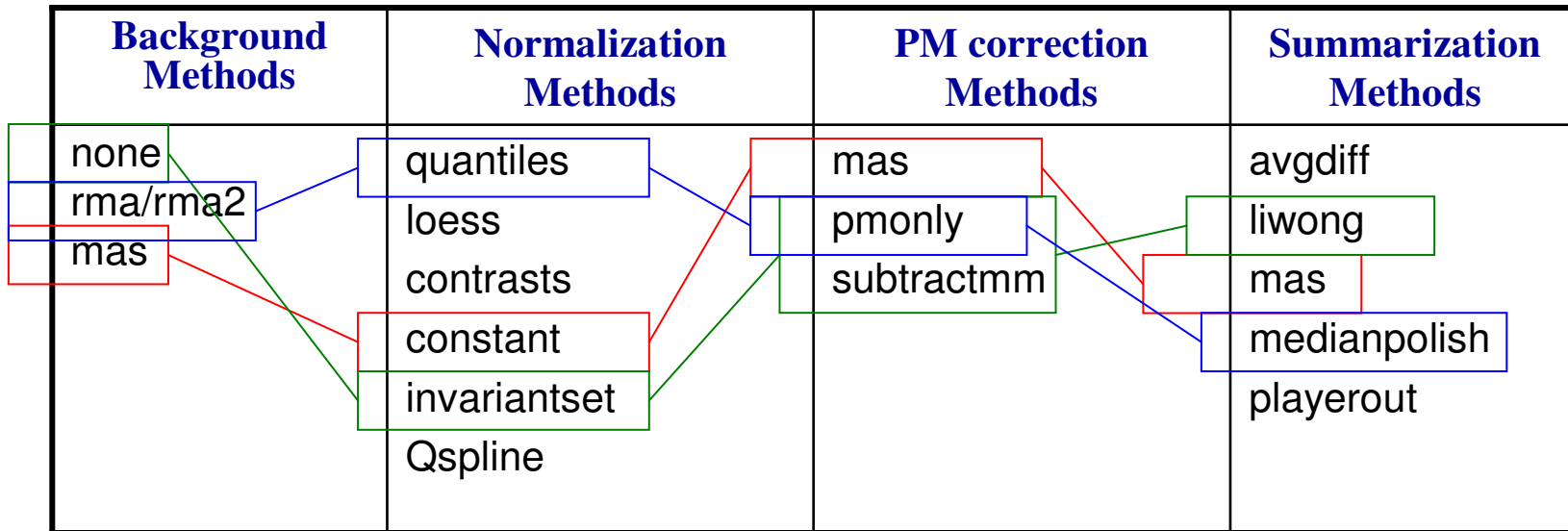




Low level Analysis

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

Low level analysis



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")`

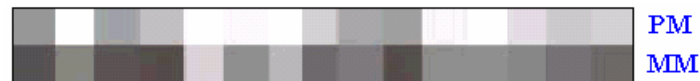
Background Correction

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.



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 explicitly 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?



37/69

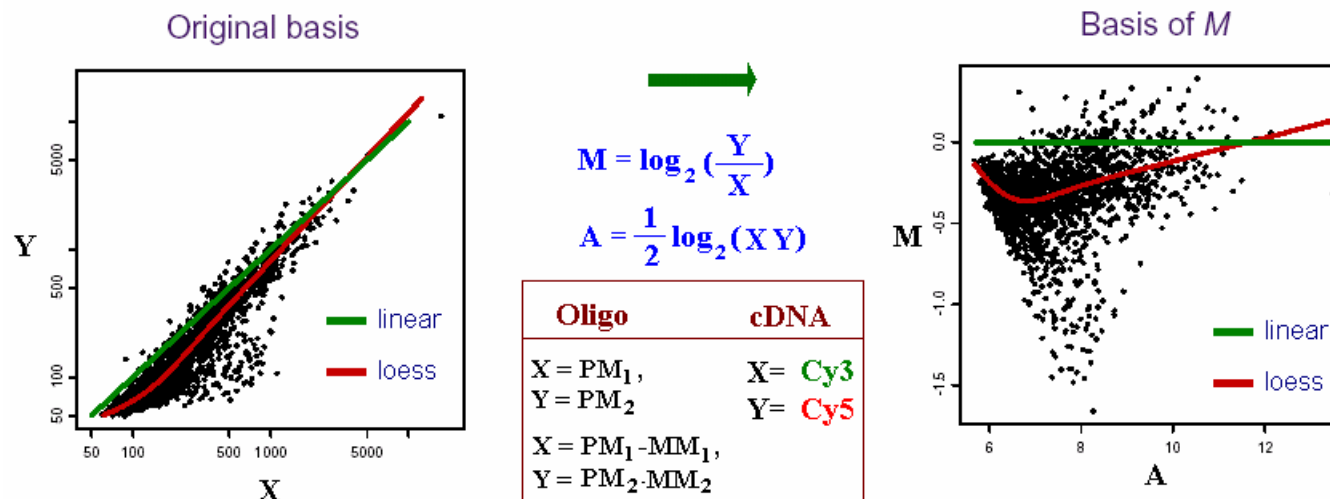
Normalization corrects for overall chip brightness and other factors that may influence the numerical value of expression intensity, enabling the user to more confidently compare gene expression estimates between samples.

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.



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).

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

$$A \times SF = TGT$$

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

Global Normalization

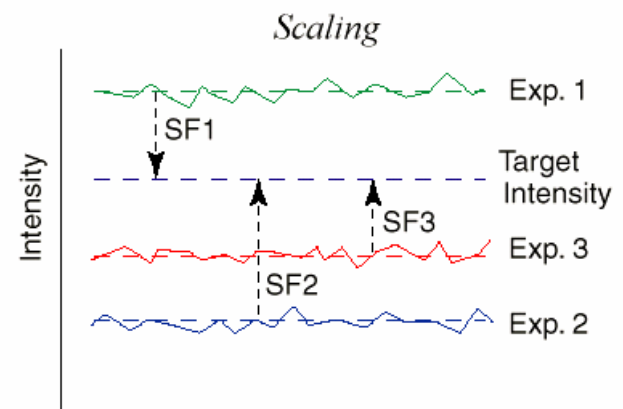
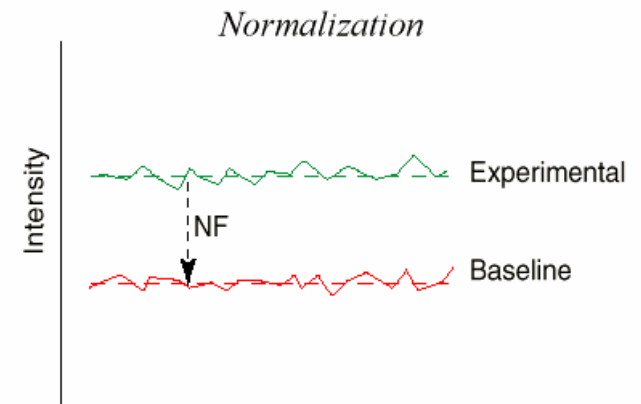
the normalization of the array is multiplied by a Normalization Factor (NF) to make its Average Intensity equivalent to the Average Intensity of the baseline array.

$$A_{exp} \times NF = A_{base}$$

$$\Rightarrow NF = \frac{A_{base}}{A_{exp}}$$

$$nf = \frac{TrimMean(SPVB_i, 0.02, 0.98)}{TrimMean(SPVE_i, 0.02, 0.98)}$$

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.

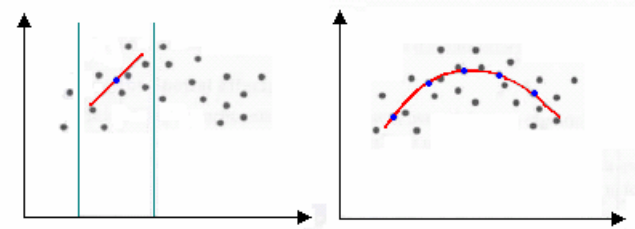


LOESS Normalization



- Loess normalization (Bolstad *et al.*, 2003) is based on **MA plots**. Two arrays are normalized by using a loess 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 $k = 1, \dots, p$ 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.

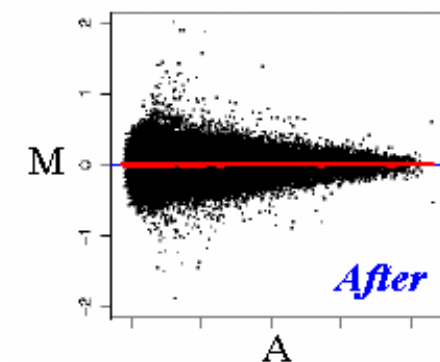
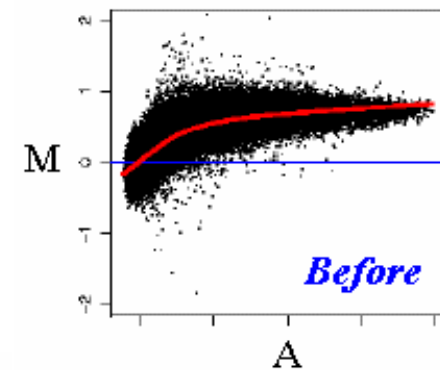
Loess is a method of local regression (see Cleveland and Devlin (1988) for details).

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 intensities are given by $x'_{ki} = 2^{A_k + \frac{M'_k}{2}}$ and $x'_{kj} = 2^{A_k - \frac{M'_k}{2}}$.

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

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

Oligo	cDNA
X = PM ₁ ,	X = Cy3
Y = PM ₂	Y = Cy5
X = PM ₁ · MM ₁ ,	
Y = PM ₂ · MM ₂	



Qspline Normalization



41/69

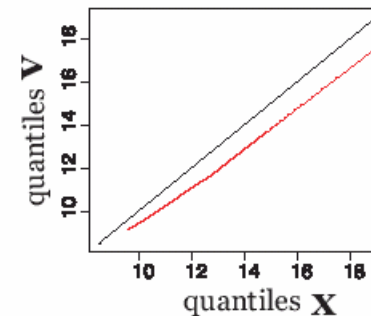
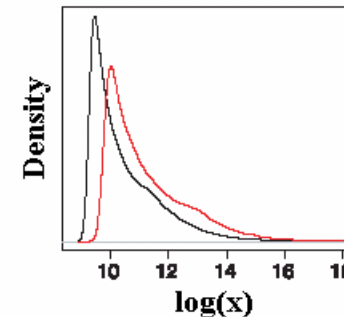
- Qspline normalization (Workman *et al.*,2002) uses a target array (either one of the arrays or a synthetic target), arrays are normalized by fitting **splines** to the **quantiles**, then using the splines to perform the normalization.

Qspline normalization

uses quantiles from array signals \mathbf{x} and target signals \mathbf{v} , to fit smoothing B-splines.

The splines are then used as signal-dependent normalization functions on the signals of \mathbf{x} .

The target signals can be from another array or could be means calculated from multiple arrays



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 conjunction 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 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 Approximately constant CV	All probe sets biased Compression of differential change at the low end
MM treated as additional PM	Background variation is insignificant Abundances moderate to large	Added statistical power Low noise Constant CV	All probe sets biased 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.

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



Low level analysis

- MAS4.0
- MAS5.0
- Li-Wong Model
- RMA

Average Difference

- The average difference for a particular probe set is then defined as the mean of all the (PM-MM) differences.
- The resulting value, or absolute expression value, is then taken as proportional to the actual amount of RNA of the corresponding gene in the sample.
- No longer recommended for use due to many flaws.

$$\text{Difference}_{\text{probepair}} = PM - MM$$

$$\text{Average Difference}_{\text{probe set}} = \sum_{i=1}^n \frac{(PM_i - MM_i)}{n}$$

(Where: n = number of probe pairs for gene X)

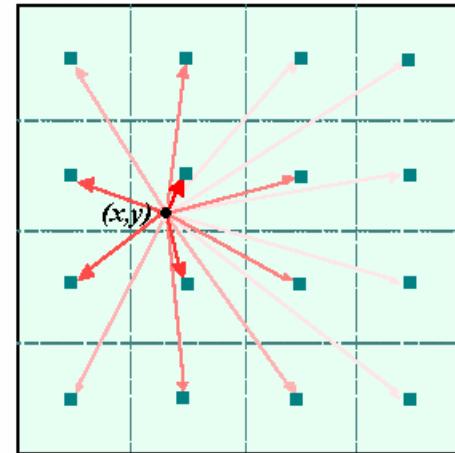
MAS5: Background Methods

MAS5

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

Zone Values

- For purposes of calculating background values, the array is split up into K rectangular zones Z_k ($k = 1, \dots, K$, default $K = 16$).
- Control cells and masked cells are not used in the calculation.
- The cells are ranked and the lowest 2% is chosen as the background b for that zone (bZ_k).
- The standard deviation of the lowest 2% cell intensities is calculated as an estimate of the background variability n for each zone (nZ_k).



Smoothing Adjustment

weights

$$w_k(x, y) = \frac{1}{d_k^2(x, y) + \text{smooth}}$$

background

$$b(x, y) = \frac{1}{\sum_{k=1}^K w_k(x, y)} \sum_{k=1}^K w_k(x, y) bZ_k$$

Noise Correction

noise

$$n(x, y) = \frac{1}{\sum_{k=1}^K w_k(x, y)} \sum_{k=1}^K w_k(x, y) nZ_k$$

adjusted intensity

$$A(x, y) = \max(I'(x, y) - b(x, y), \text{NoiseFrac} * n(x, y))$$

where $I'(x, y) = \max(I(x, y), 0.5)$

- MAS method corrects both PM and MM probes.

Affymetrix: Statistical Algorithm Description Document

MAS5: PM Correction Method

- An **ideal mismatch** is subtracted from PM. The ideal mismatch is documented by Affymetrix (2002).
- The Ideal Mismatch will always be less than the corresponding PM and thus we can safely subtract it without risk of negative values.

To calculate a specific background ratio representative for the probe set, we use the **one-step biweight algorithm** (T_{bi}).

The biweight specific background (SB) for probe pair j in probe set i is:

$$SB_i = T_{bi} \left(\log_2(PM_{i,j}) - \log_2(MM_{i,j}) : j = 1, \dots, n_i \right)$$

$$IM_{i,j} = \begin{cases} MM_{i,j}, & MM_{i,j} < PM_{i,j} \\ \frac{PM_{i,j}}{2^{(SB_i)}}, & MM_{i,j} \geq PM_{i,j} \text{ and } SB_i > \text{contrast}\tau \\ \frac{PM_{i,j}}{2^{\left(\frac{\text{contrast}\tau}{1 + \left(\frac{\text{contrast}\tau - SB_i}{\text{scale}\tau} \right)^2} \right)}}, & MM_{i,j} \geq PM_{i,j} \text{ and } SB_i \leq \text{contrast}\tau \end{cases}$$

default $\text{contrast}\tau = 0.03$, default $\text{scale}\tau = 10$

Probe Value

the probe value PV for every probe pair j in probeset i .

n is the number of probe pairs in the probeset.

$$V_{i,j} = \max(PM_{i,j} - IM_{i,j}, d)$$

default $\delta = 2^{(-20)}$

$$PV_{i,j} = \log_2(V_{i,j}), j = 1, \dots, n_i$$

[Affymetrix: Statistical Algorithm Description Document](#)

One-Step Tukey's Biweight Algorithm **Purpose**

There are several stages in the algorithms in which we want to calculate an average. The biweight algorithm is a method to determine a robust average unaffected by outliers.

MAS5: Summarization Method

Signal is calculated as follows:

1. Cell intensities are preprocessed for global background.
2. An ideal mismatch value is calculated and subtracted to adjust the PM intensity.
3. The adjusted PM intensities are log-transformed to stabilize the variance.
4. The biweight estimator is used to provide a robust mean of the resulting values. Signal is output as the antilog of the resulting value.
5. Finally, Signal is scaled using a trimmed mean.

Probe Value

the probe value PV for every probe pair j in probeset i .
 n is the number of probe pairs in the probeset.

$$V_{i,j} = \max(PM_{i,j} - IM_{i,j}, d) \quad \text{default } \delta = 2^{(-20)}$$

$$PV_{i,j} = \log_2(V_{i,j}), j=1, \dots, n_i$$

Signal Log Value

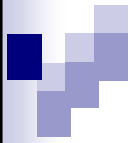
$$SignalLogValue_i = T_{bi}(PV_{i,1}, \dots, PV_{i,n_i})$$

$$sf = \frac{\text{target signal}}{\text{TrimMean}(2^{SignalLogValue_i}, 0.02, 0.98)}$$

$$nf = \frac{\text{TrimMean}(SPVb_i, 0.02, 0.98)}{\text{TrimMean}(SPVe_i, 0.02, 0.98)}$$

The reported value of probe set i is: **Signal**

$$ReportedValue(i) = nf * sf * 2^{(SignalLogValue_i)}$$



PLIER (Affymetrix, 2005)

■ Guide to Probe Logarithmic Intensity Error (PLIER) Estimation

	Previous Generation	2.0 Platform
Array Technology	<ul style="list-style-type: none">• 18-μm features• Edge minimization mask strategy	<ul style="list-style-type: none">• 11-μm features• Chrome setback mask design strategy• 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	<ul style="list-style-type: none">• 3rd-party cDNA reagents• Enzo labeling kits	<ul style="list-style-type: none">• GeneChip[®] One- and Two-Cycle cDNA Kits• GeneChip[®] IVT Labeling Kit

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

Liwong: Normalization

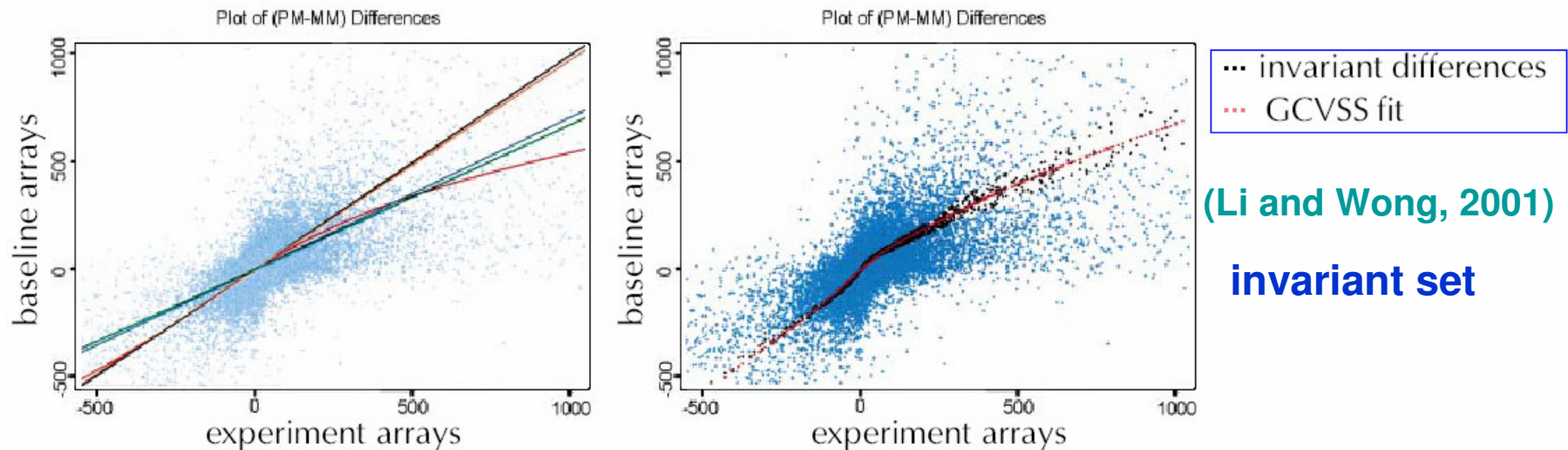
Liwong (PM-only Model)

```
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).



Invariant Set Algorithm

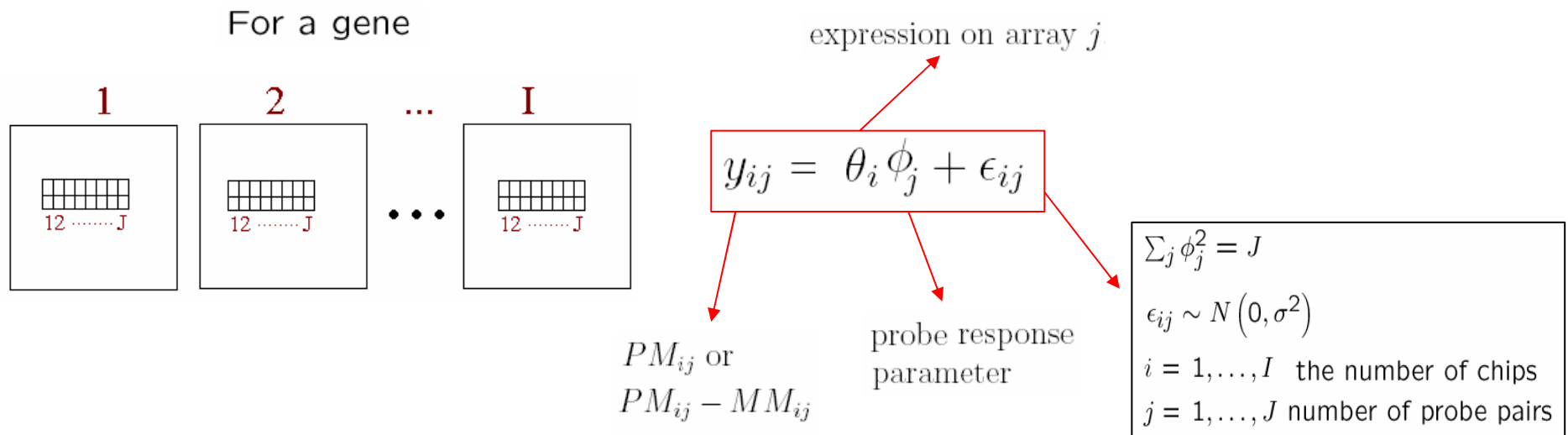


(Li and Wong, 2001)

- Invariant difference selection algorithm (IDS) chooses a subset of PM/MM intensity differences to serve as the basis for fitting a normalization relation.
- A set of probes are said to be invariant if the ordering of these probes according to the PM/MM differences in the experiment array, is the same as that in the baseline array.
- Intuitively, if a gene is truly differentially expressed, then the PM/MM differences for this gene are more likely to have different ranks relative to the other probes, and hence they are not likely to be included in a large invariant set.
- IDS algorithm uses the following expressions to determine the approximately invariant set:
$$R_i = \frac{[L(B_i + E_i) + H(2N - B_i - E_i)]}{2N}$$
$$D_i = \frac{2|B_i - E_i|}{(B_i + E_i)}$$
- L and H are the rank difference thresholds for the low and high ends of the difference intensity range.
- B_i and E_i are the ranks for the i th difference of the baseline and experiment arrays
- N is the total number of differences that were ordered in the current iteration of the algorithm.
- R_i defines the threshold for difference intensity i by linearly interpolating the threshold between a low difference intensity threshold, given by L , and a high difference intensity threshold, given by H .
- D_i is the rank difference test statistic used to determine if the i th difference should be included in the invariant set
- The i th difference is considered approximately invariant if $D_i < R_i$
- Once the approximately invariant set of differences has been selected, the normalization curve is constructed by applying the GCVSS technique to the invariant set

Liwong: Summarization Method

(Model-Based Expression Index , MBEI)



- θ_i : this model computes an expression level on the i th array.
- $SE(\theta)$'s and $SE(\phi)$'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 $SE(\phi_j)$, 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

$$O = S + N$$

Exponential (alpha)

Normal (mu, sigma)

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

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

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

$$b = \sigma$$

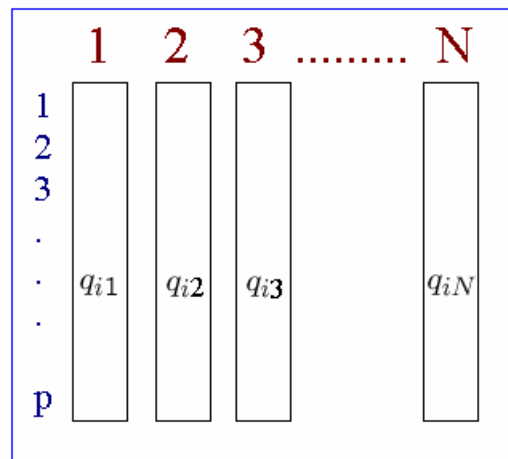
ϕ : standard normal density function

Φ : standard normal distribution function

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

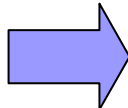
RMA: Normalization

- **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.



X_{sort}

average
quantile


$$\frac{1}{N} \sum_{j=1}^N q_{ij}$$

The q th 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.

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

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

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

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

$i = 1, \dots, I_k$ probes

$j = 1, \dots, J$ arrays

probe effect

log₂ expression value

Comparison of Affymetrix GeneChip Expression Measures



56/69

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

網站(D) 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-IN)
 - original assessment (of DILUTION)
 - entry comparison tool (beta)
 - study archives
- Comparison of Affymetrix GeneChip Expression Measures
- A Benchmark for Affymetrix GeneChip Expression Measures
- R package
- FAQ
- Contact us

Sponsored by: The Hopgene Project

Results as of August 7, 2003 present

IN	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	1.00
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	0.00
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	0.66
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_log / 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 / imacdon	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	gcrma1131 / 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_NBGC / helstad	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

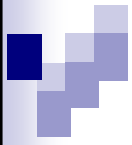
- Download the spike-in and dilution data sets.

Spike-in hgu95a Data

Method	SD	99.9%	low	slope med	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_NBGC	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 Mixed error model PM-MM, PM only etc. Multiple background options Smoothly handles intensities below background	Higher reproducibility of signal (lower coefficient of variation) without loss of accuracy relative to single array analysis Higher differential sensitivity for low expressors Lack of bias	Computationally intensive In cases where feature intensities disagree, may have more than one solution Performance relative to amount of model data provided Variance not stable on log scale
dCHIP	Multiple array analysis Arithmetic error model PM only (stanardly) Multiple background options (no background typical)	Higher reproducibility of signal over single array analysis Good differential change detection Variance stable on log scale with no background	In cases where feature intensities disagree, may have more than one solution Performance relative to amount of model data provided Positive bias at low end (compression of Fold Change)
RMA	Multiple array analysis Multiplicative error PM only Attenuated global background (single global background used to adjust for each intensity)	Higher reproducibility of signal over single array analysis Good differential change detection Variance stable on log scale	In cases where feature intensities disagree, may have more than one solution (mitigated by median polish) Performance relative to amount of model data provided Positive bias at low end (compression of Fold Change)
MAS 5	Single array analysis Multiplicative error PM-MM Background imputed to handle negative differences	Conservative Smooth down-weighting of outliers Positive output values Minimal bias	Limited by single array analysis Variance not stable on log scale Some positive bias

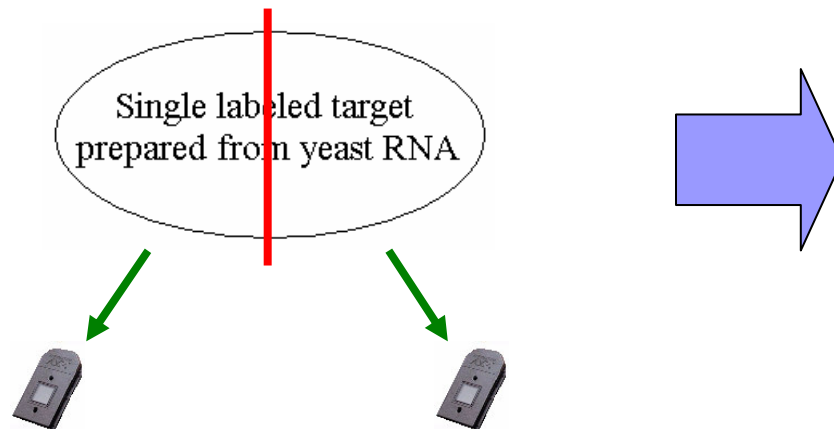
Affymetrix: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation. Edited by: Affymetrix I. Santa Clara, CA, ; 2005.



Reproducibility and False Positive Rates

Reproducibility

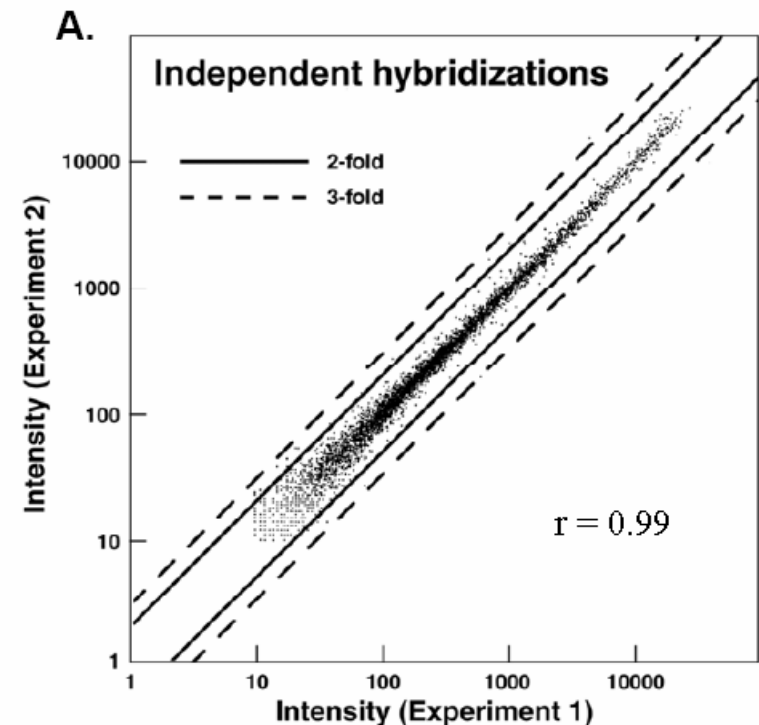
- Reproducibility (再現性) = Repeatability = Precision
- The degree to which repeat measurements of the same quantity will show the same or similar results.
- The precision is usually measured by comparing some measure of dispersion (e.g., standard deviation) with zero.



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

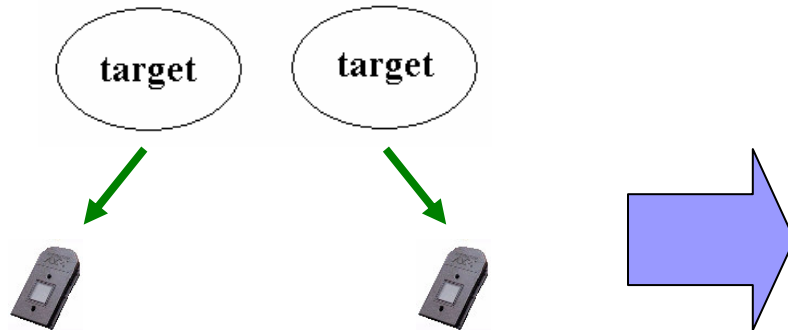
Pearson correlation coefficients = 0.99



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

Reproducibility (conti.)

different individuals from
same pellet of yeast cells

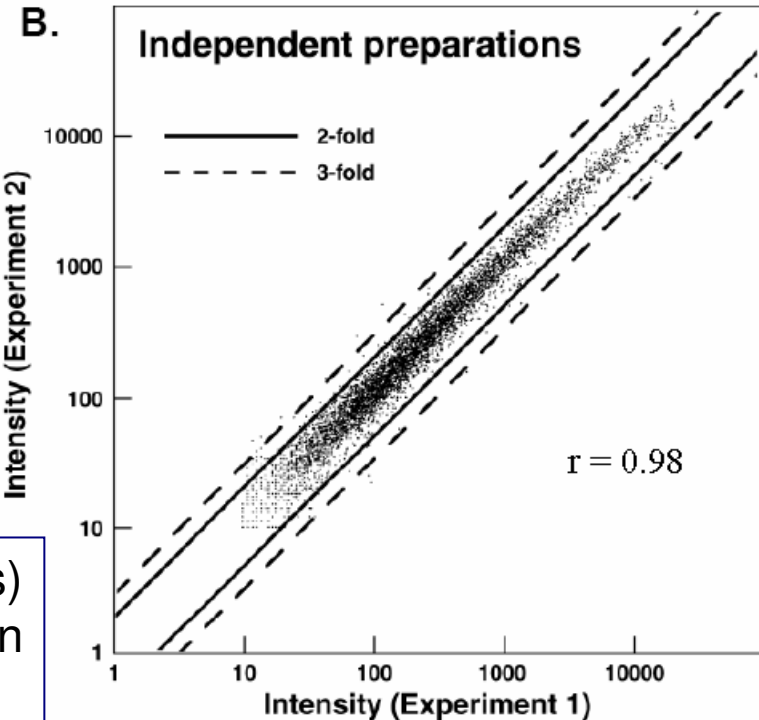


74 probe sets of the more than 6,200 probes sets with an intensity difference of more than 2-fold.

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

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

Pearson correlation coefficients = 0.98

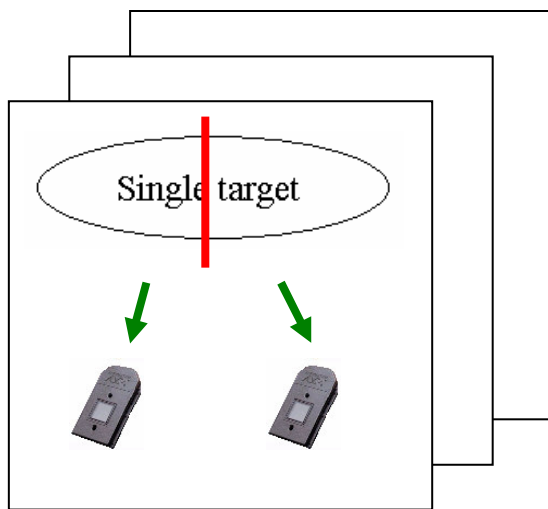


Figures Source: *Modified from*
Affymetrix yeast microarray (Wodicka 1997).

Concordance correlation coefficients

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

```

Average Signal (All): 778.1
Average Signal (All): 763.6
Average Signal (All): 763.6
Number Increase: 233 1.0%
Number Decrease: 610 2.7%
Number MIncrease: 51 0.2%
Number MDecrease: 109 0.5%
Number No Change: 21280 95.5%
Number (A/M->P, MI/I) 53
Number (P->A/M, MD/D) 145 0.7%
    
```

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

Mean Percent Change
Standard Deviation

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.)

Image Analysis/Normalization

Shareware/Freeware

- **Bioconductor** (R, Gentleman)
- DNA-Chip Analyzer (**dChip** v1.3) (Li and Wong)
- **RMExpress**: a simple standalone GUI program for windows for computing the RMA expression measure.

Commercial

- Affymetrix GeneChip Operating Software (**GCOS** v1.0)
- GeneSpring GX v7.3

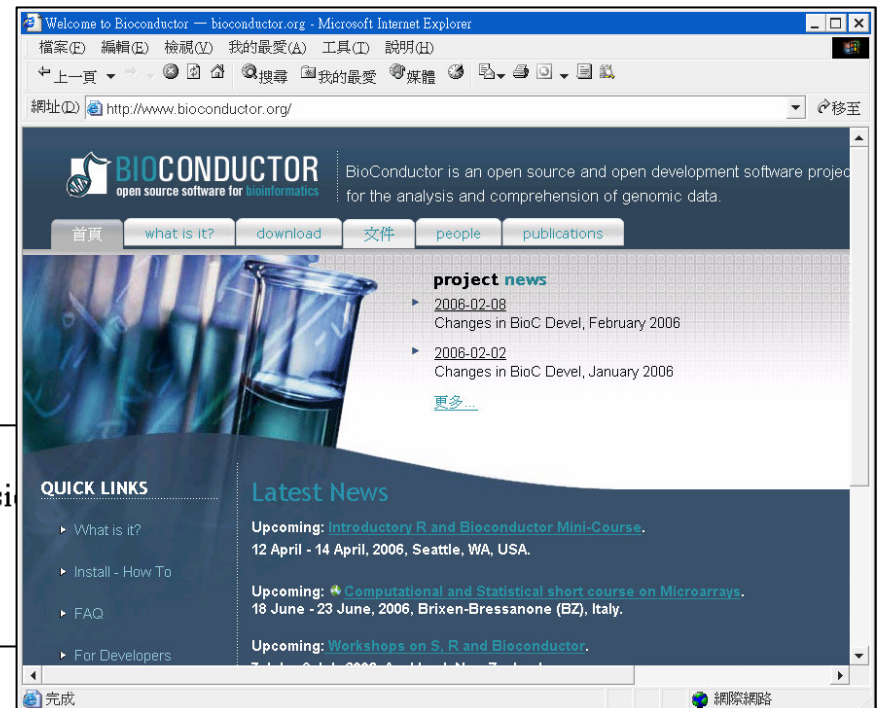
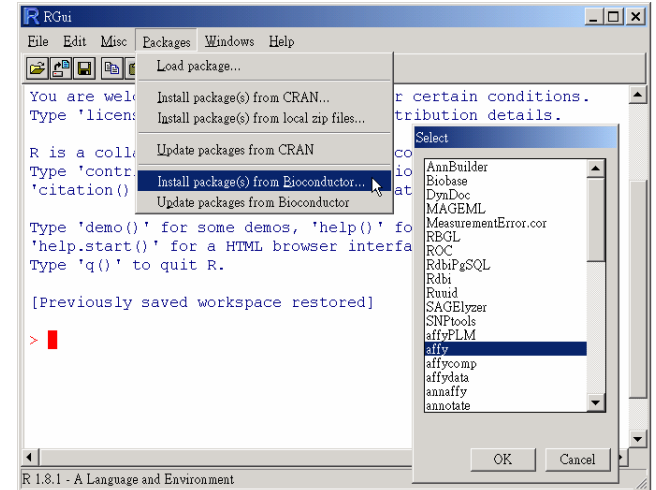
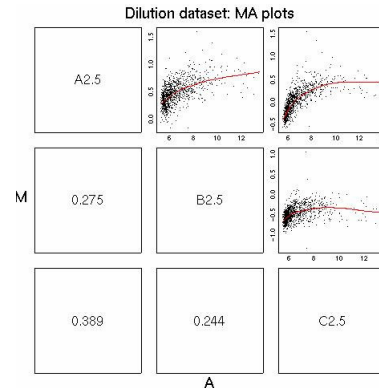
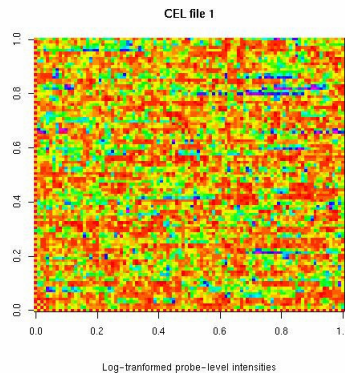
The Bioconductor: affy

The Bioconductor Project
Release 1.7

<http://www.bioconductor.org/>



affypdnn
affyPLM
gcrma
makecdfenv



- [affy](#) Methods for Affymetrix Oligonucleotide Arrays
- [affycomp](#) Graphics Toolbox for Assessment of Affymetrix Expression
- [affydata](#) Affymetrix Data for Demonstration Purpose
- [annaffy](#) Annotation tools for Affymetrix biological metadata
- [AffyExtensions](#) For fitting more general probe level models

The Bioconductor: affy



64/69

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,
  bgcorrect.param = list(),
  | none,
  | mas,
  | rma

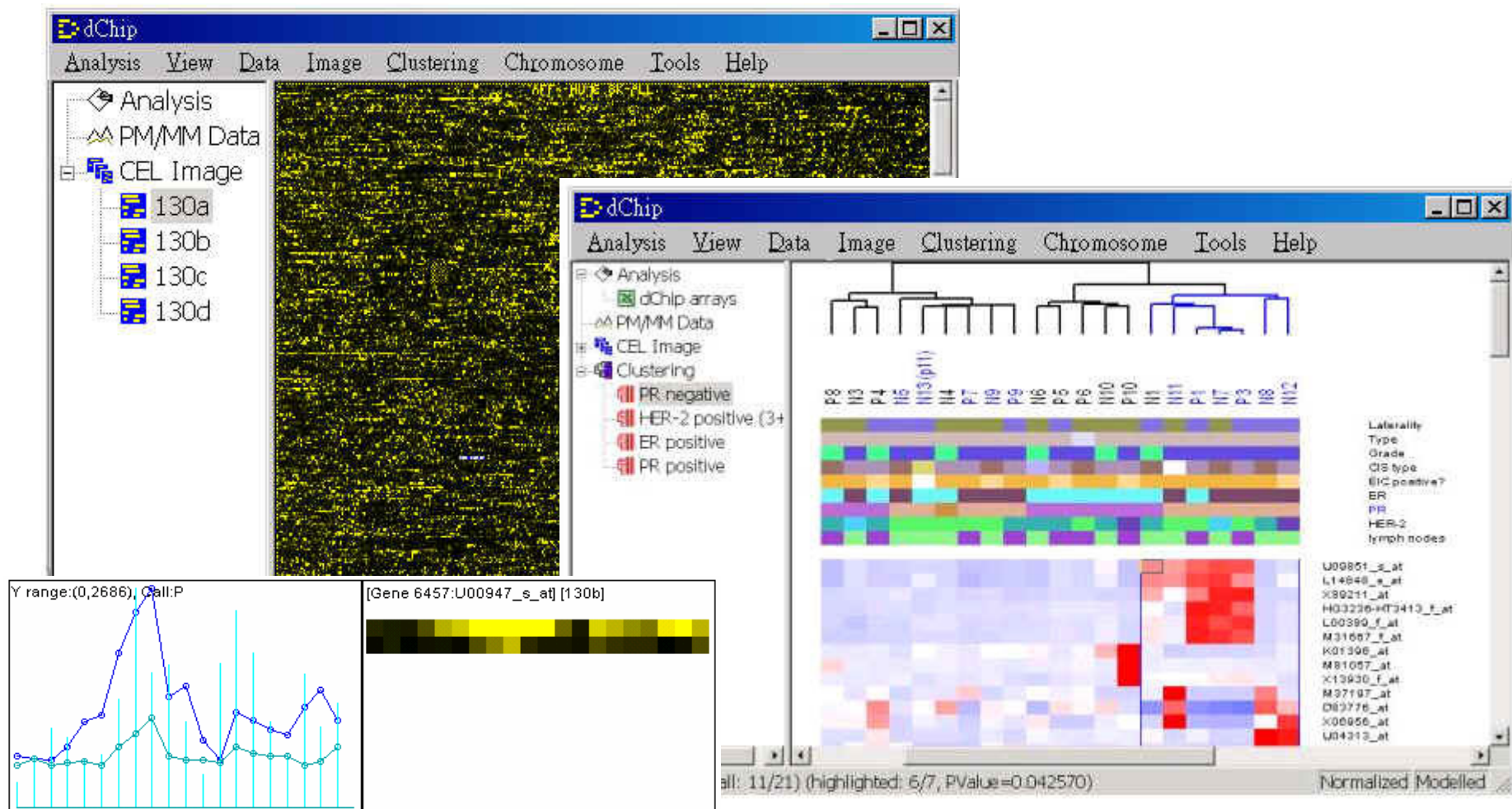
  # normalize
  normalize = TRUE,
  normalize.method = NULL,
  normalize.param = list(),
  | constant,
  | contrasts,
  | invariantset,
  | loess, qspline,
  | quantiles,
  | quantiles.robust

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

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

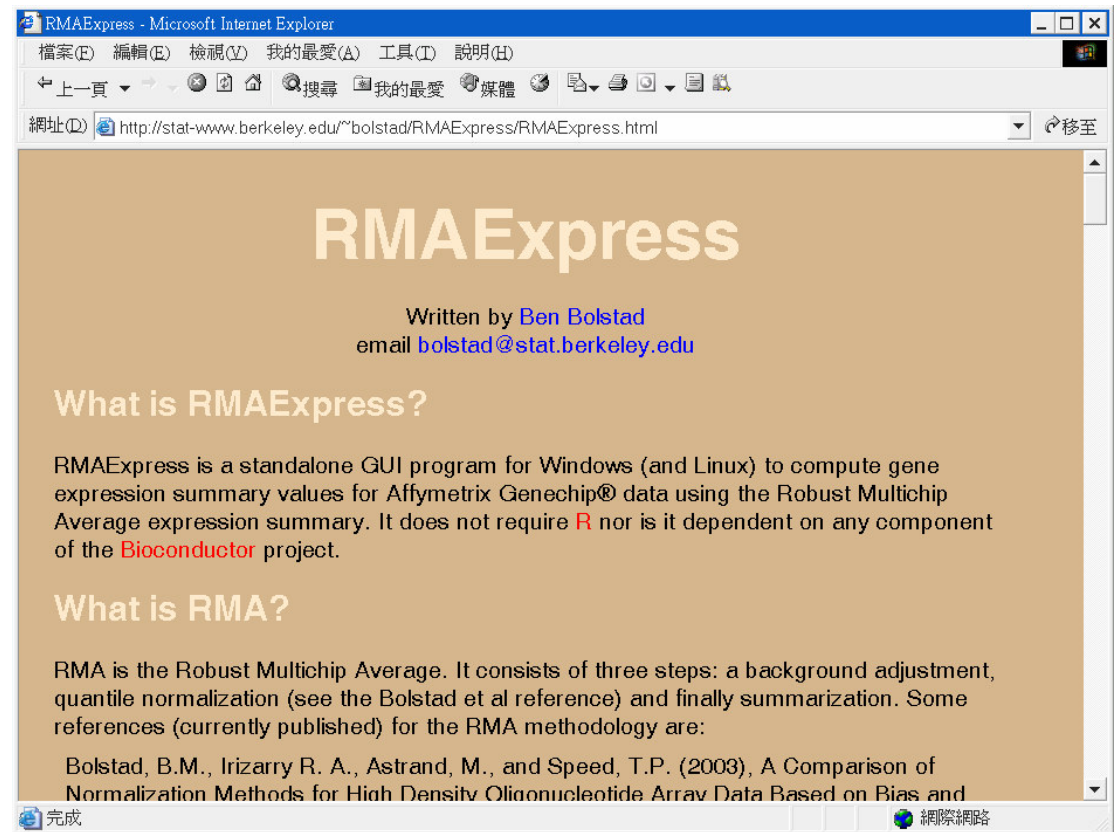
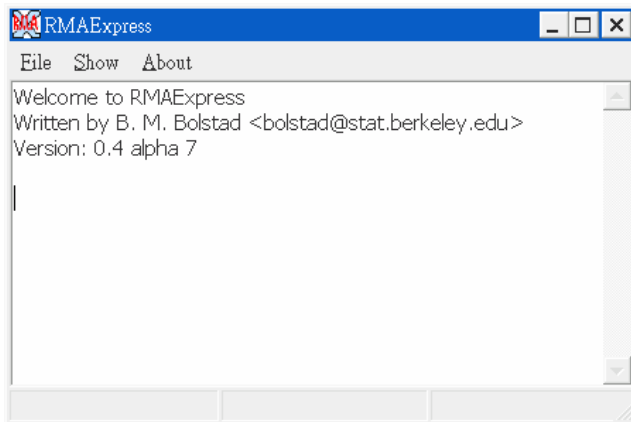
  # misc.
  verbose = TRUE,
  warnings = TRUE,
  widget = FALSE)
```

DNA-Chip Analyzer (dChip2006)

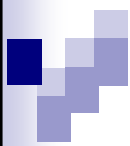


<http://www.biostat.harvard.edu/complab/dchip/>

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



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



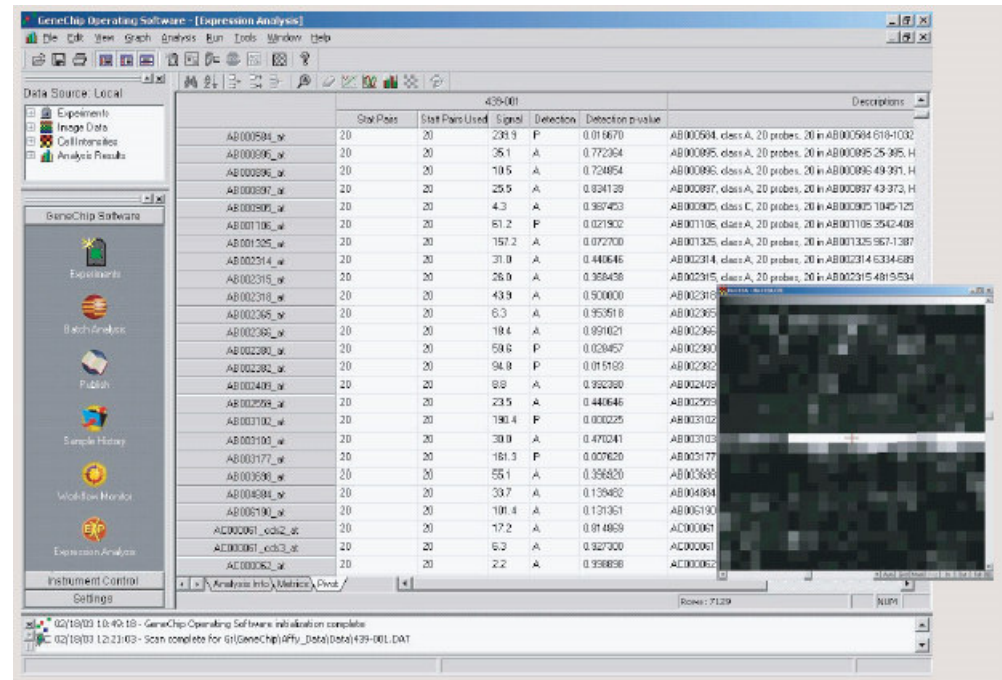
Affymetrix GeneChip Operating Software



<http://www.affymetrix.com>

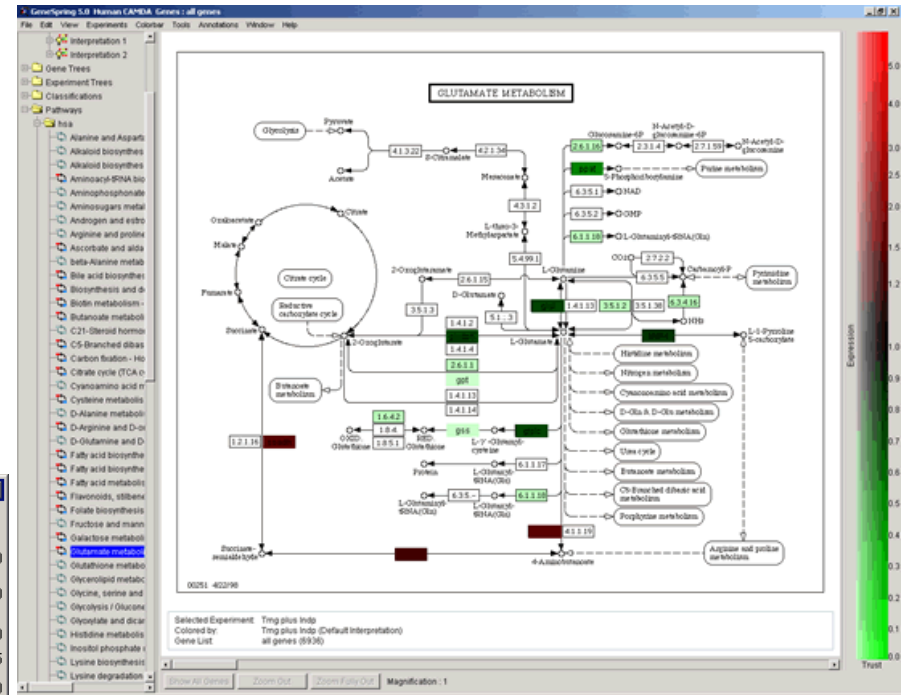
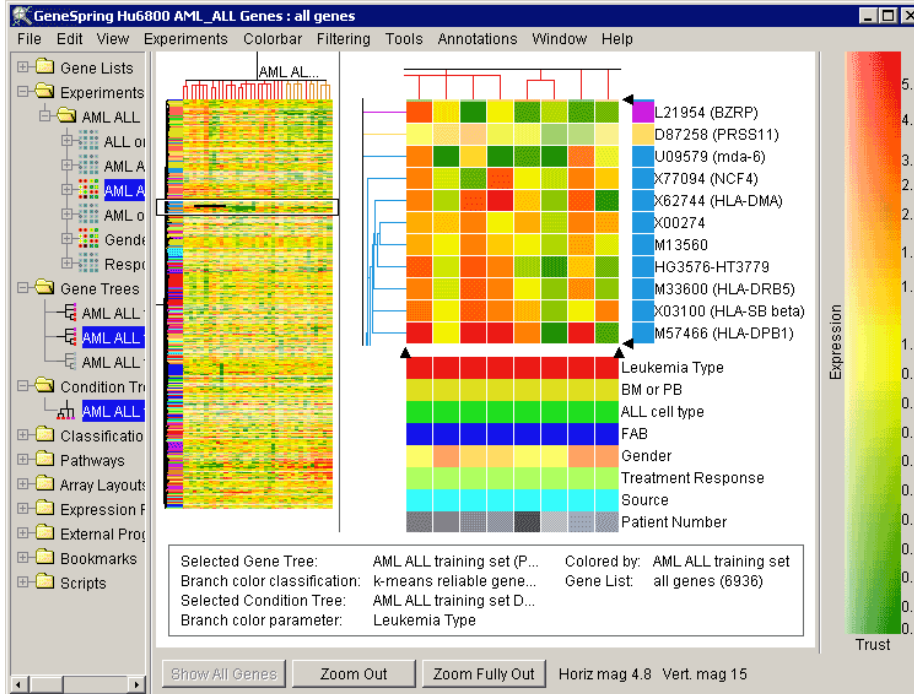
Specifications

Instrument Support	<ul style="list-style-type: none"> Affymetrix GeneChip® Fluidics Station 400 & 450 GeneChip Scanner 3000 GeneArray 2500 Scanner
Affymetrix Software Compatibility	<ul style="list-style-type: none"> Support GeneChip DNA Analysis Software (GDAS) for mapping and resequencing data analysis Support Affymetrix® Data Mining Tool software for statistical and clus analysis
Database Engine	<ul style="list-style-type: none"> Microsoft Data Engine
GCOS Database	<ul style="list-style-type: none"> Process Database Publish Database Gene Information Database
Database Management	<ul style="list-style-type: none"> GCOS Manager GCOS Administrator
Algorithm	<ul style="list-style-type: none"> 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>



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

More than 700 papers

Useful Links and Reference



<http://ihome.cuhk.edu.hk/~b400559/>

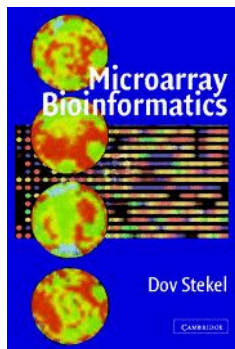


<http://www.affymetrix.com>

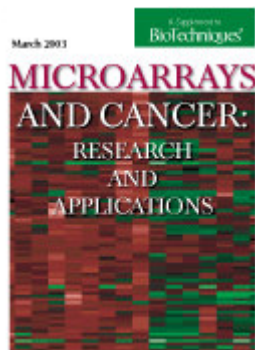
Bibliography on
Microarray Data Analysis
<http://www.nslj-genetics.org/microarray/>



<http://bioinformatics.oupjournals.org>



Stekel, D. (2003).
Microarray
bioinformatics,
New York :
Cambridge
University Press.



Microarrays and Cancer: Research and Applications
<http://www.biotechniques.com/microarrays/>

■ Speed Group Microarray Page: Affymetrix data analysis
http://www.stat.berkeley.edu/users/terry/zarray/Affy/affy_index.html

■ Statistics and Genomics Short Course, Department of Biostatistics Harvard School of Public Health.
<http://www.biostat.harvard.edu/~rgentlem/Wshop/harvard02.html>

■ Statistics for Gene Expression
<http://www.biostat.jhsph.edu/~ririzarr/Teaching/688/>

■ Bioconductor Short Courses
<http://www.bioconductor.org/workshop.htm>

DNA Microarray Data Analysis
http://www.csc.fi/csc/julkaisut/oppaat/arraybook_overview

