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

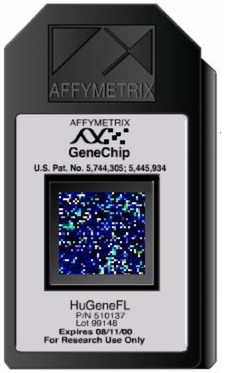
2006/03/28

吳漢銘
hmwu@stat.sinica.edu.tw
http://www.sinica.edu.tw/~hmwu/

Institute of Statistical Science, Academia Sinica
中央研究院 統計科學研究所

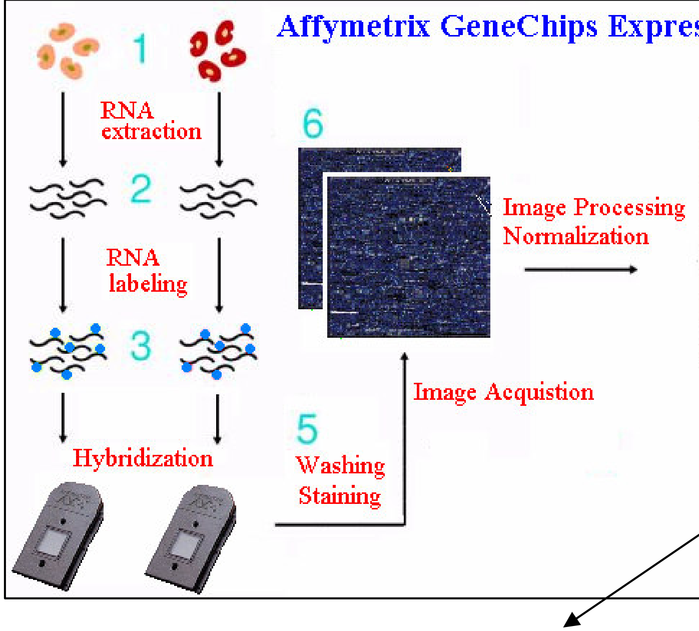
Outlines

- **Affymetrix GeneChip Technology**
  - GeneChip Photolithography, Array Design, Analysis Process
- **Comparison with cDNA Microrrays**
- **Assay and Analysis Flow Chart**
  - Image Analysis, Affymetrix Data Files, from DAT to CEL.
- **Quality Assessment**
  - RNA Sample Quality Control
  - Array Hybridization Quality Control
  - Statistical Quality Control (Diagnostic Plots)
- **Low level Analysis**
  (from probe level data to expression value)
  - Background Correction, Normalization, PM Correction, Expression Index
- **Software**
  - Freeware: BioConductor, dChip, RMAExpress
  - Commercial: GCOS, GeneSpring
- **Useful Links and Reference**

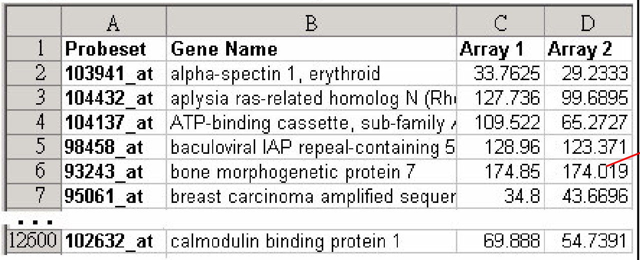


Overview

Affymetrix GeneChips Expression Analysis Process



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 /	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

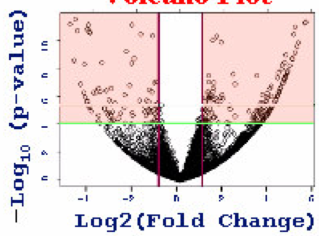
Matrix of genes (rows) and samples (columns)

Discovery of differentially expressed genes

Parametric : t-test

Non-parametric : Wilcoxon, Mann-Whitney test

Volcano Plot

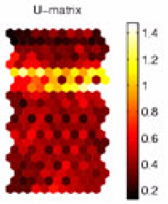


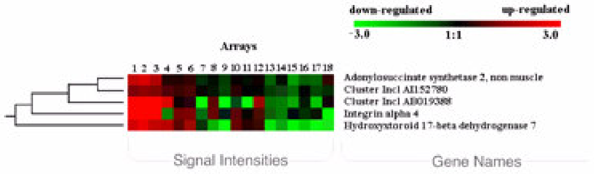
Unsupervised: clustering

Hierarchical clustering

K-means clustering

Self-organizing maps

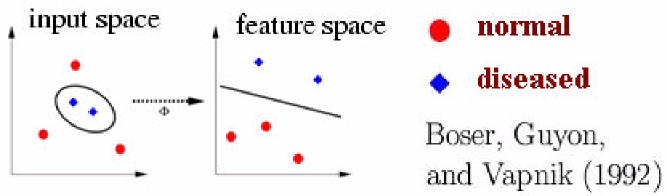




Supervised: classification

- **Linear discriminants**
- **Decision trees**
- **Support vector machines**

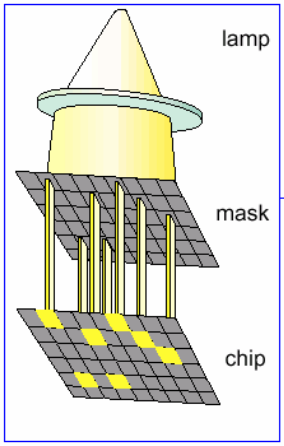
Support Vector Classifiers



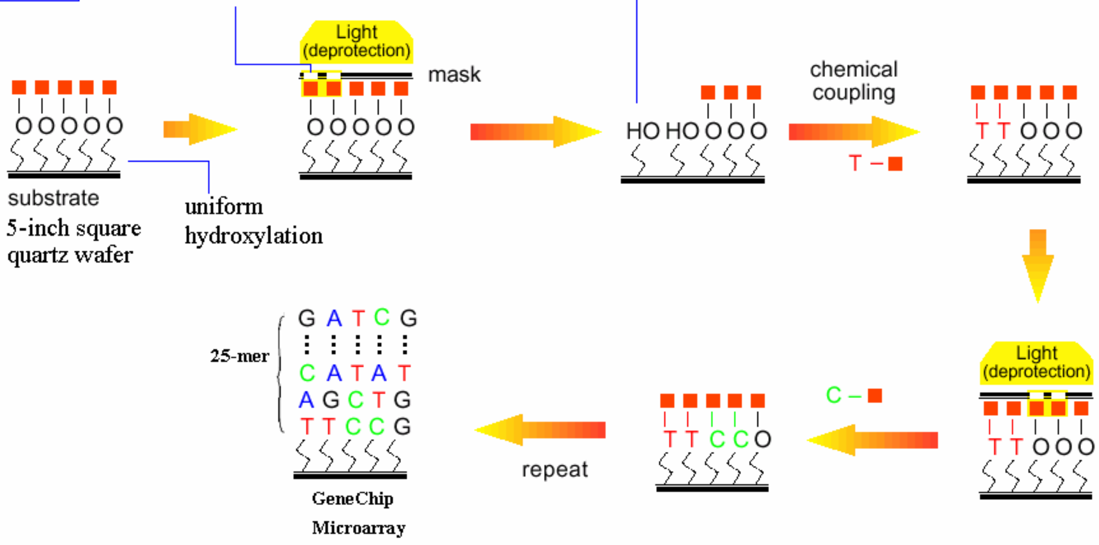
Boser, Guyon, and Vapnik (1992)

GeneChip Photolithography

1. 在要作爲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烴基（-OH）。
   - 這些烴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烴基。



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'端的烴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，此鹼基會與烴基的部份結合起來。
   - 之後再蓋住其他部份，使其他未形成烴基的部份經過光照後產生烴基。
   - 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

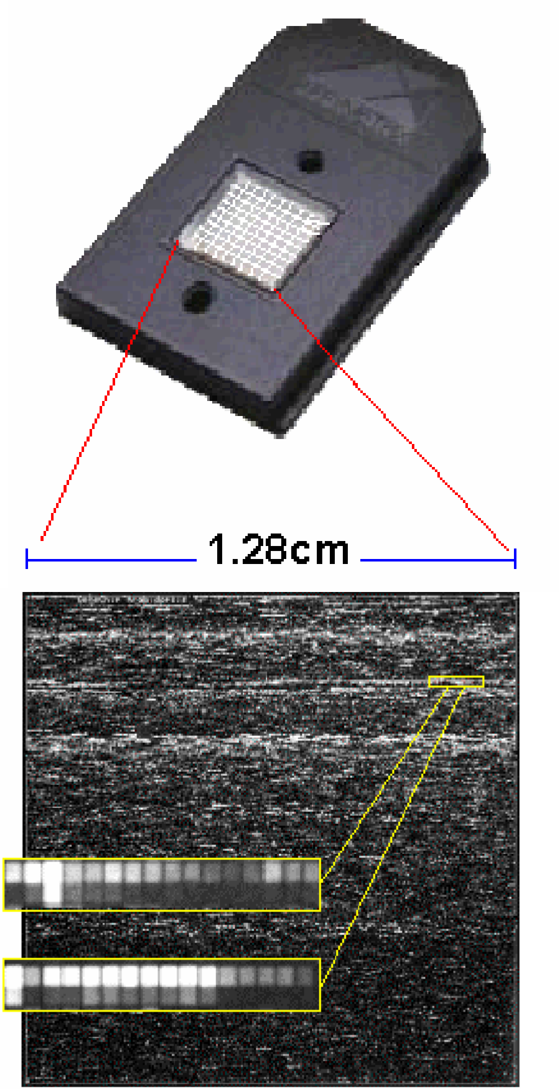
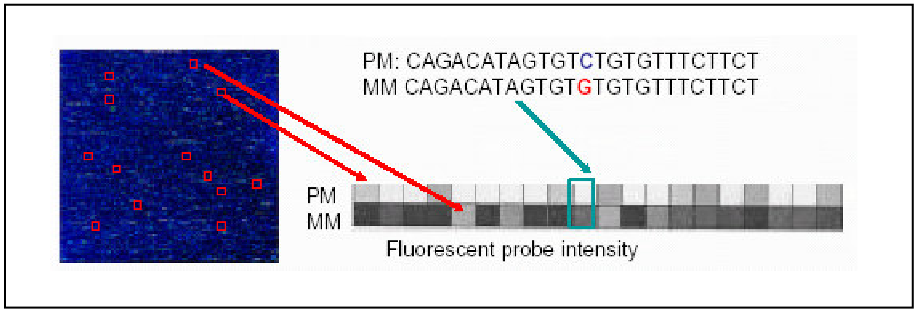
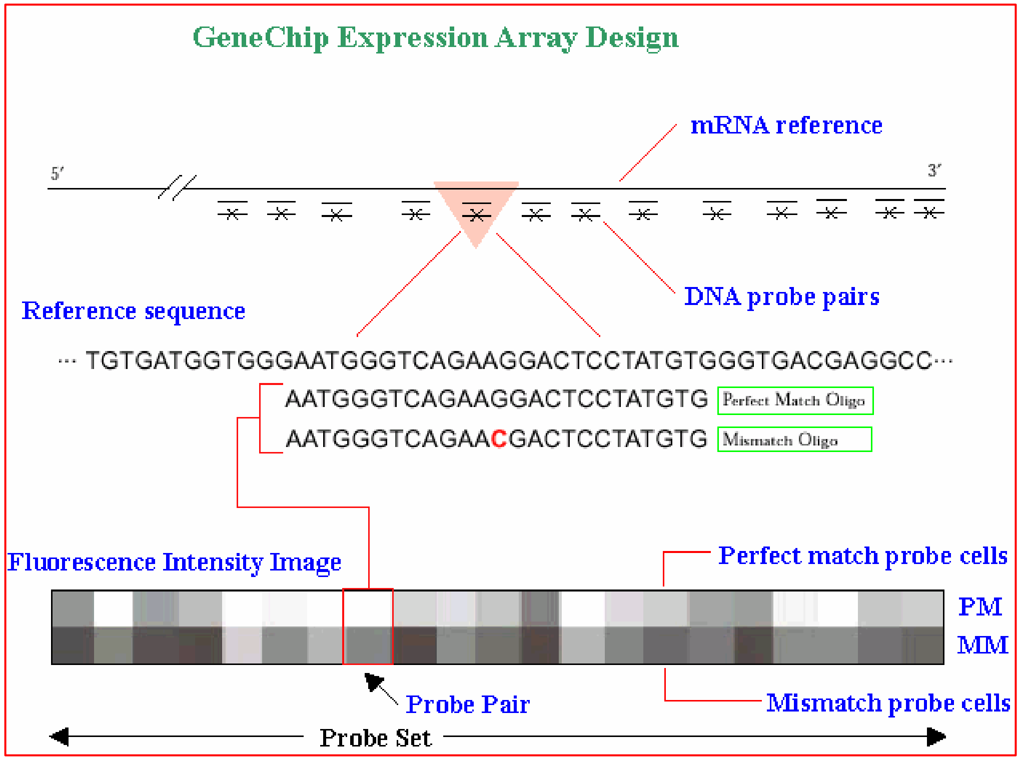


Image of hybridized probe array


PM: CAGACATAGTGTCTGTGTTTCTTCT
MM CAGACATAGTGTGTGTGTTTCTTCT
PM
MM
Fluorescent probe intensity

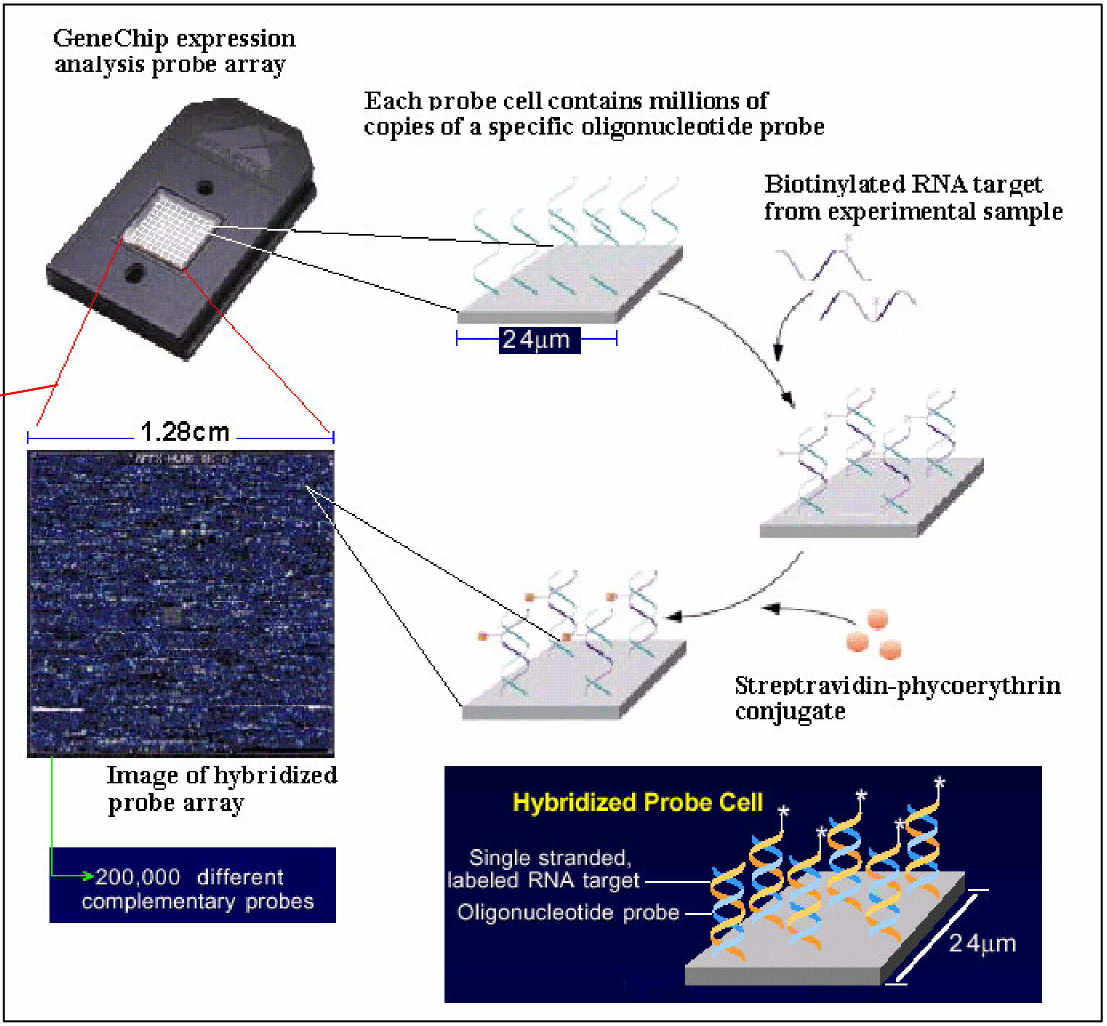

GeneChip Expression Array Design
mRNA reference
5′
3′
DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC···
AATGGGTCAGAAGGACTCCTATGTG
Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTG
Mismatch Oligo
Fluorescence Intensity Image
Perfect match probe cells
PM
MM
Probe Pair
Mismatch probe cells
Probe Set

GeneChip Expression Analysis Process

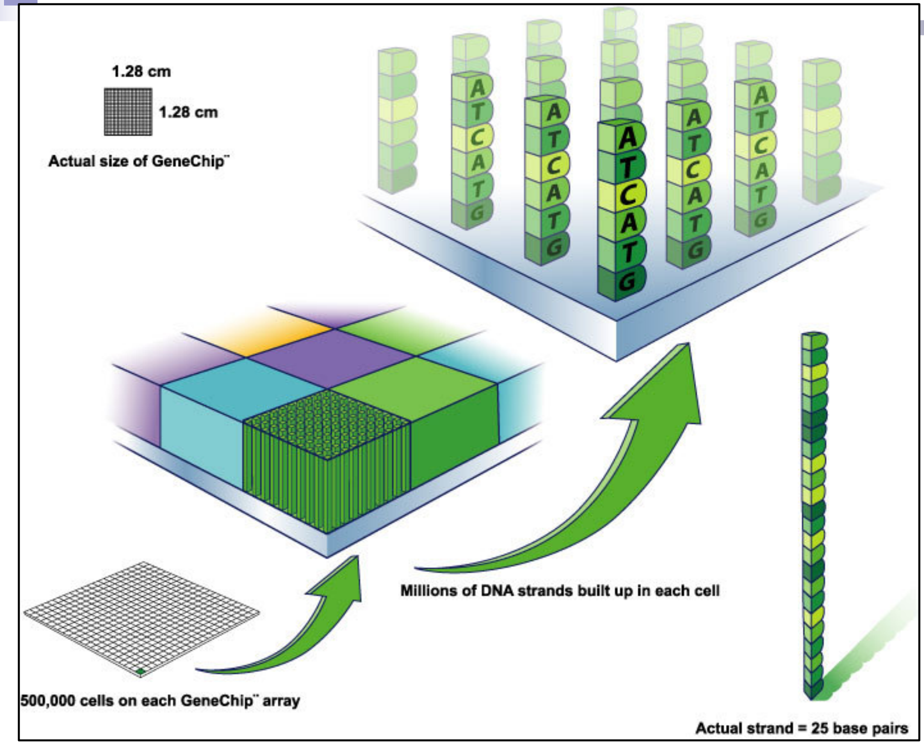
The GeneChip® Instrument System

Scan and Quantitate

Affymetrix GeneChip®
Scanner 3000 with workstation.

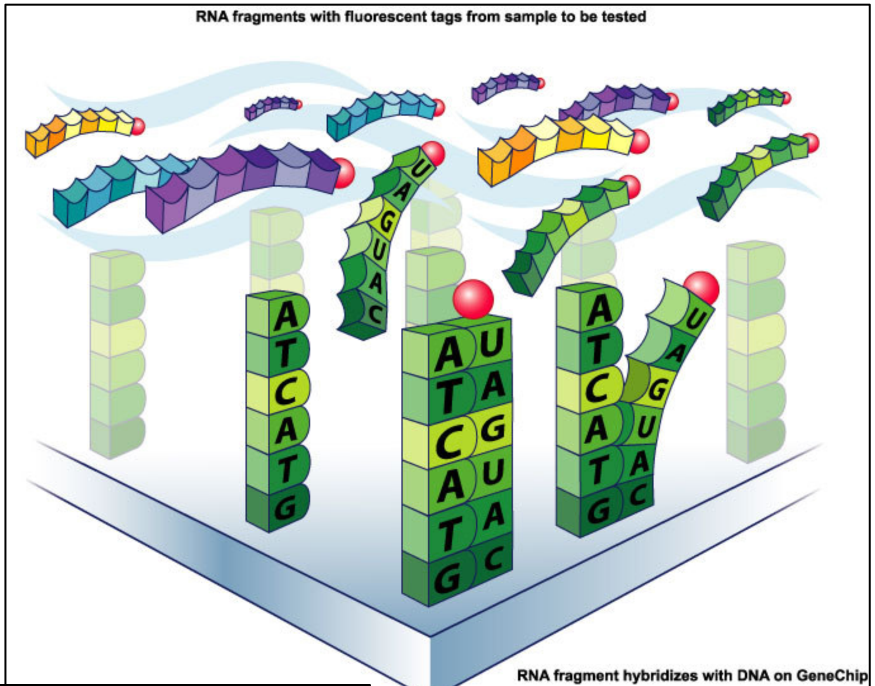


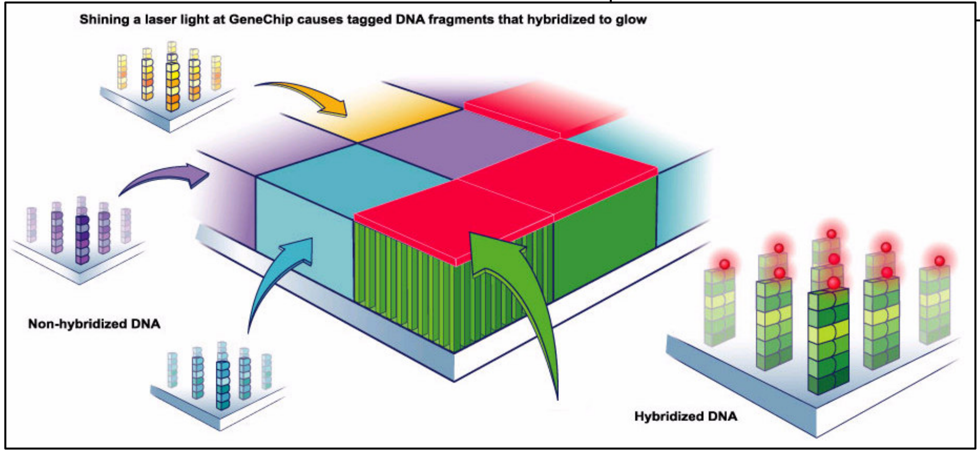
More Figures on Affymetrix Web Site


1.28 cm
1.28 cm
Actual size of GeneChip™
A T C A T G
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip™ array
Actual strand = 25 base pairs

GeneChip® Single Feature

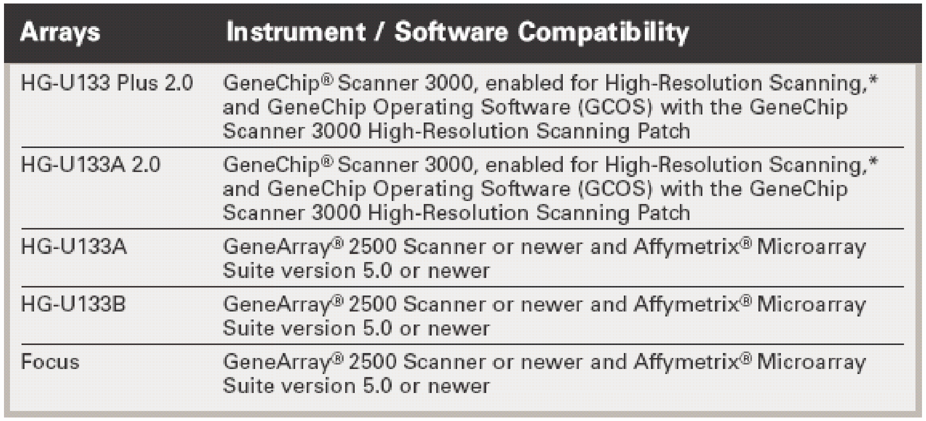
GeneChip® Hybridization


RNA fragments with fluorescent tags from sample to be tested
U A G U A C
A T C A T G
A U
T A
C G
A U
T A
G C
RNA fragment hybridizes with DNA on GeneChip


Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA

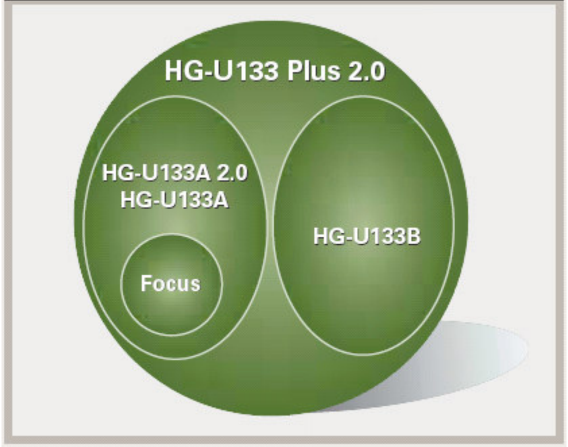
Hybridized GeneChip® Microarray

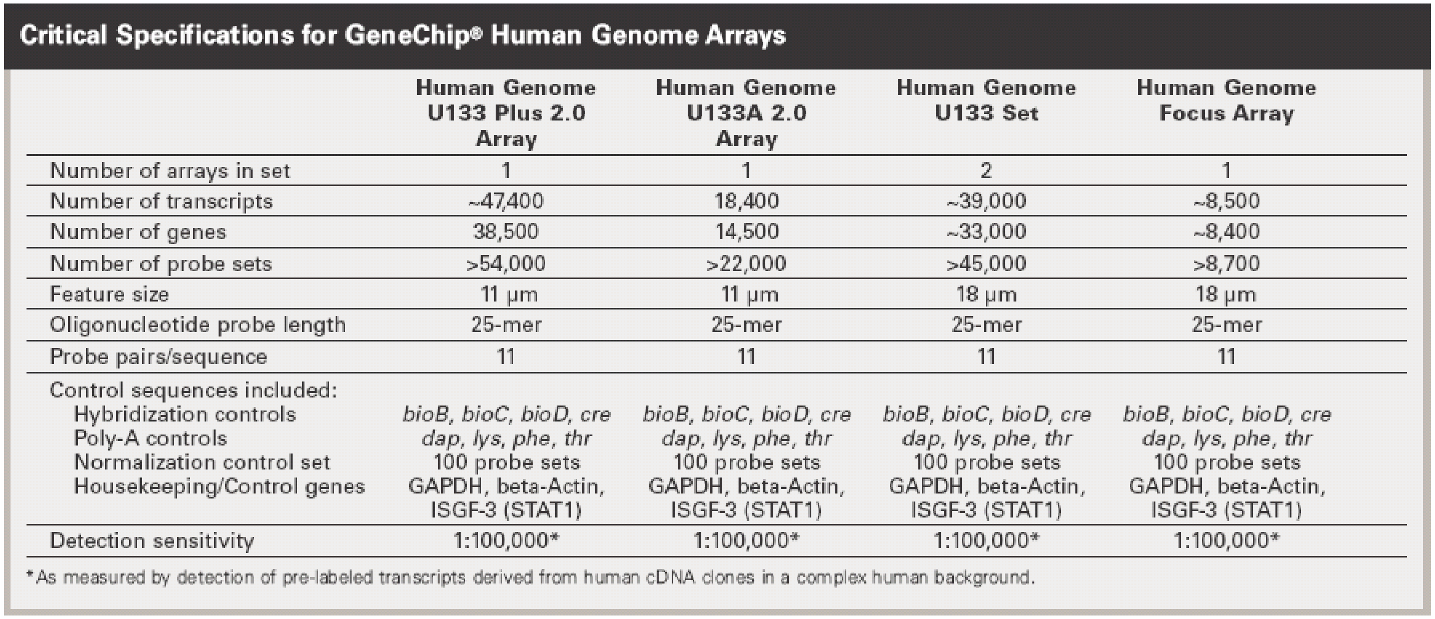
GeneChip® Human Genome Array



Arrays	Instrument / Software Compatibility
HG-U133 Plus 2.0	GeneChip® Scanner 3000, enabled for High-Resolution Scanning,* and GeneChip Operating Software (GCOS) with the GeneChip Scanner 3000 High-Resolution Scanning Patch
HG-U133A 2.0	GeneChip® Scanner 3000, enabled for High-Resolution Scanning,* and GeneChip Operating Software (GCOS) with the GeneChip Scanner 3000 High-Resolution Scanning Patch
HG-U133A	GeneArray® 2500 Scanner or newer and Affymetrix® Microarray Suite version 5.0 or newer
HG-U133B	GeneArray® 2500 Scanner or newer and Affymetrix® Microarray Suite version 5.0 or newer
Focus	GeneArray® 2500 Scanner or newer and Affymetrix® Microarray Suite version 5.0 or newer

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	*bioB, bioC, bioD, cre*	*bioB, bioC, bioD, cre*	*bioB, bioC, bioD, cre*	*bioB, bioC, bioD, cre*
Poly-A controls	*dap, lys, phe, thr*	*dap, lys, phe, thr*	*dap, lys, phe, thr*	*dap, lys, phe, thr*
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.

Comparison with Spotted cDNA Microarray

Spotted cDNA Microarray	**Affymetrix GeneChips**
Probes are cDNA fragments, usually amplified by PCR.	Probes are oligos synthesized in situ using a photolithographic approach
■ At least two samples are hybridized to chip.	■ One target sample per array.
■ One probe one gene.	■16-20 probe-pairs per gene.
■ Probes of varying length	■ Probes are 25-mers.
■ Fluorescence at different wavelengths measured by a scanner.	■The apparatus requires a fluidics station for hybridization and a special scanner.
■Samples (normally poly(A)+ RNA) are labelled using fluorescent dyes.	■ Only a single fluorochrome is used per hybridization.
■Probes are deposited on a solid support, either positively charged nylon or glass slide.	■ Oligonucleotides synthesized in situ on silica wafers.

Comparison with Spotted cDNA Microarray

Advantages of Spotted Arrays

- Can choose the DNA on the array.
- Cheaper.
- Do not need to know the DNA sequence.
- Can hybridize closely related species.

Disadvantages of Spotted Arrays

- More technically variable.
- Less specificity (will cross hybridize to genes ~80% homology).
- Cannot distinguish closely related gene families.
- May need to confirm DNA sequence.
- Repeated amplification and quality control.

Advantages of Affymetrix Chips

- High specificity (small probe length means gene family members can be differentiated.)
- Very well researched technology
- Very robust protocols and results are very reproducible.
- Can use small amount of RNA
- Widely used, so annotation of probe sets is of relatively high quality.

Disadvantages of Affymetrix Chips

- Very expensive to design (~US$300,000).
- Expensive to perform experiments (~US$400 + $300 labeling/hybridization).
- Limited to the species for which there are chips available sequence required.
- No probe manipulation.
- Single target hybridization, so comparison always involves two experiments, and dye swaps are impossible.
- Match/mismatch technology has major limitations: mismatch signal often higher than match, and dose response curve is different for each pair.

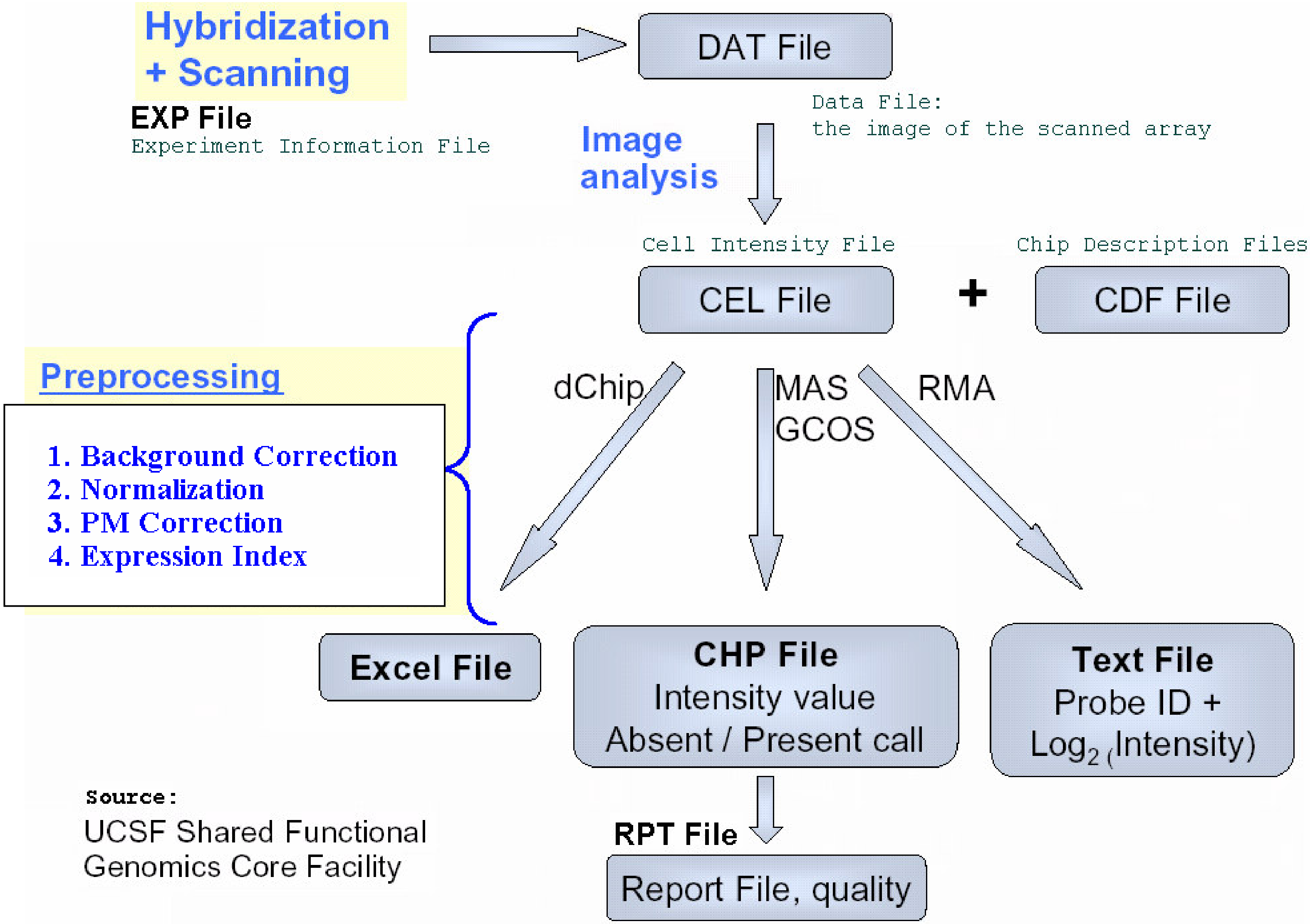
Terms & Descriptions

- **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 mm portion of the array on which are synthesized ~107 molecules of a single oligonucleotide(a.k.a. tile). 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.
- **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.
- **Probe Array Tiling** - The spatial organization of probe array features into probe pairs and sets.

- **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.

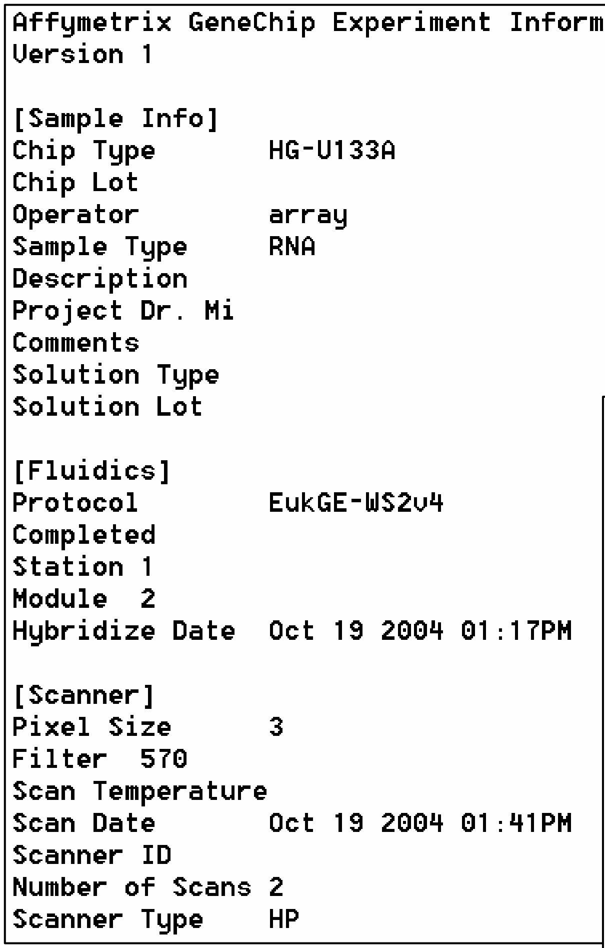
Assay and Analysis Flow Chart



Source:
UCSF Shared Functional
Genomics Core Facility

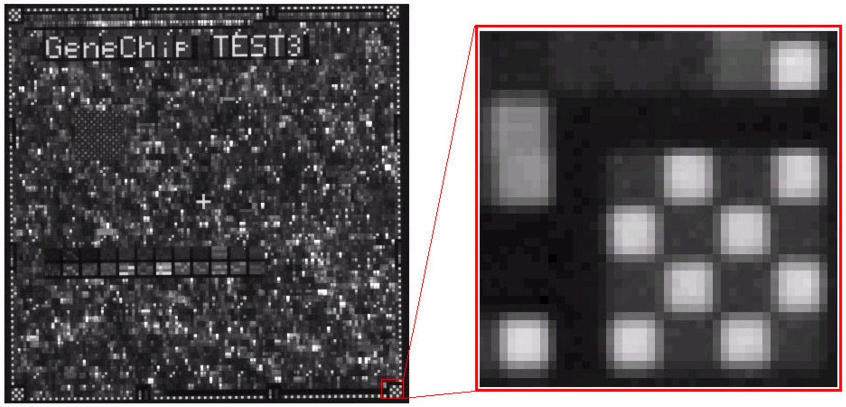
Affymetrix Data Files

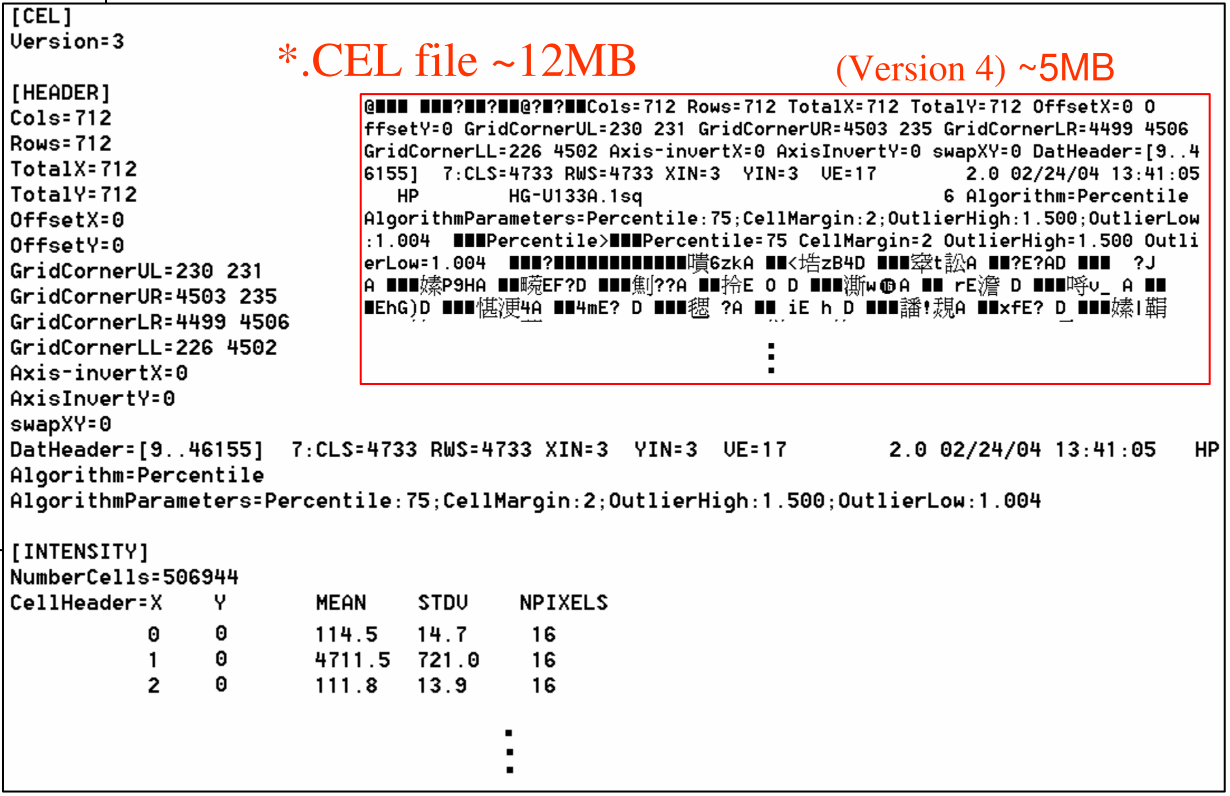
*.EXP file



```
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

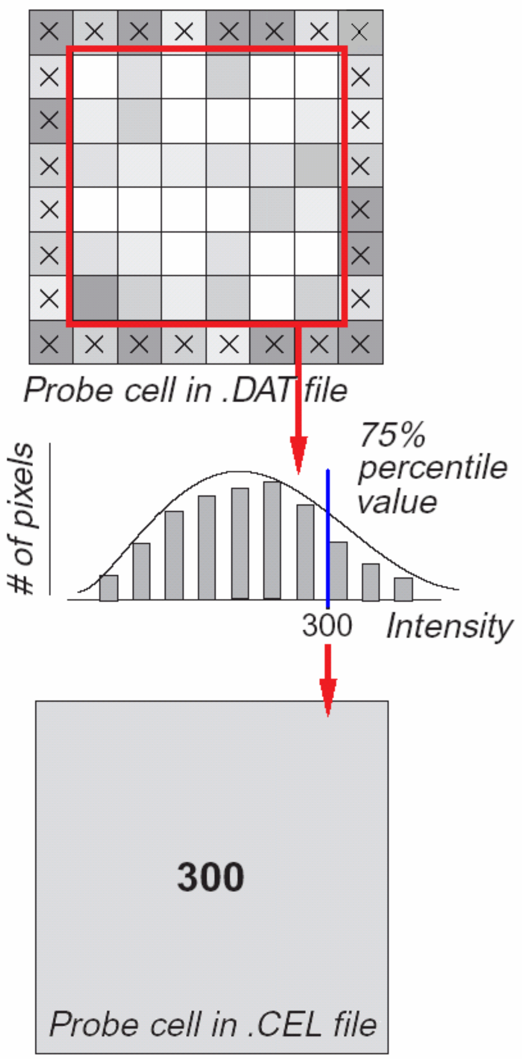
(Version 4) ~5MB

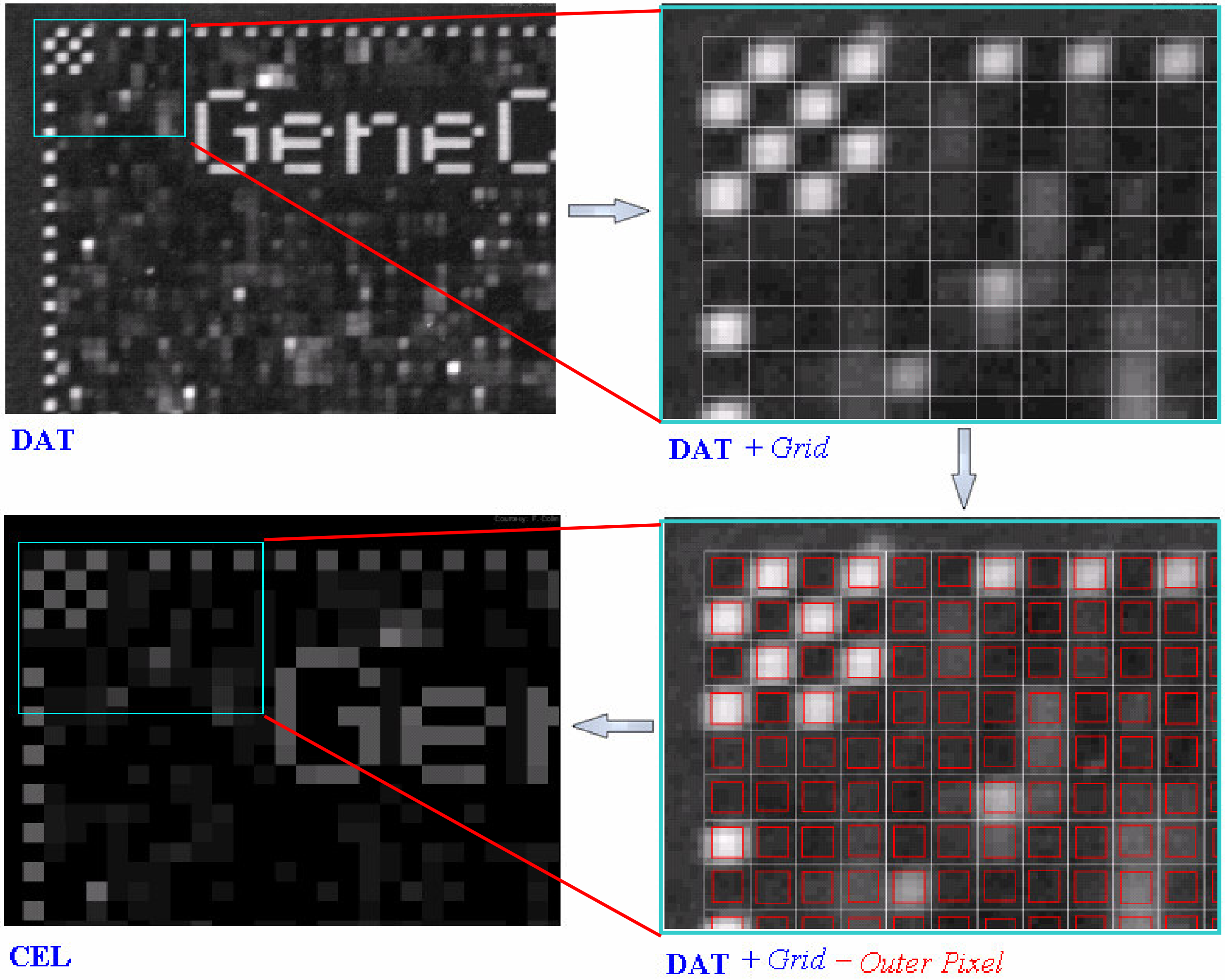
```
[CEL]
Version=3

[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
                              ⋮
```

```
@■■■ ■■■?■■?■■@?■?■■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 ■■<垍zB4D ■■■窣t訟A ■■?E?AD ■■■  ?J
A ■■■嫊P9HA ■■曉EF?D ■■■劁??A ■■拎E O D ■■■澌w⑮A ■■ rE澹 D ■■■哷v_ A ■■
■EhG)D ■■■愖渜4A ■■4mE? D ■■■嵈 ?A ■■ iE h D ■■■譎!覝A ■■xfE? D_■■■嫊I輯
                                     ⋮
```

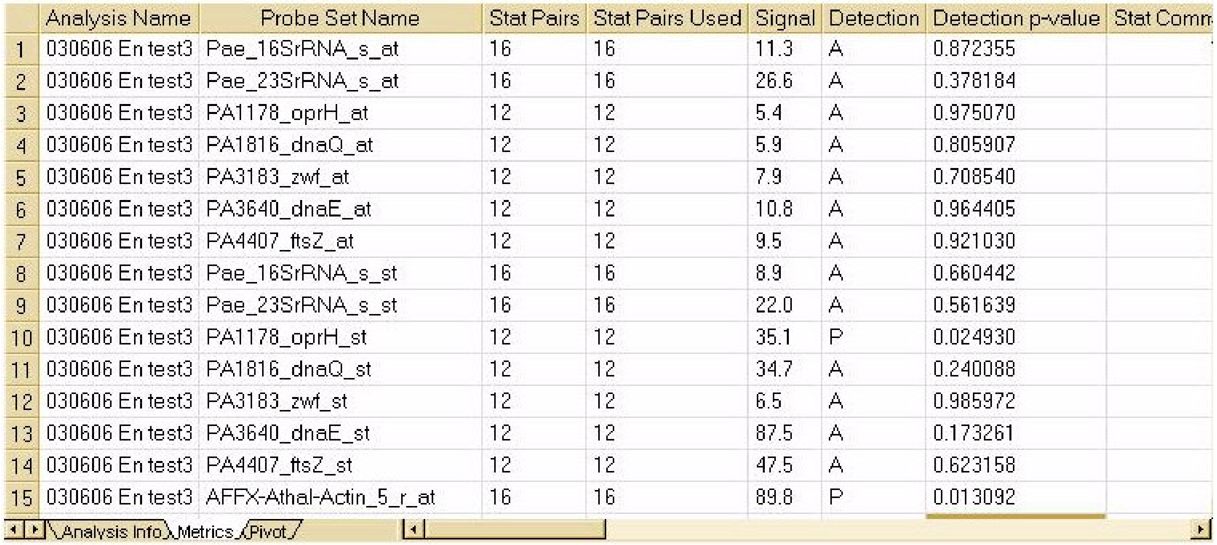
CEL File Conversion Tool

From DAT to CEL



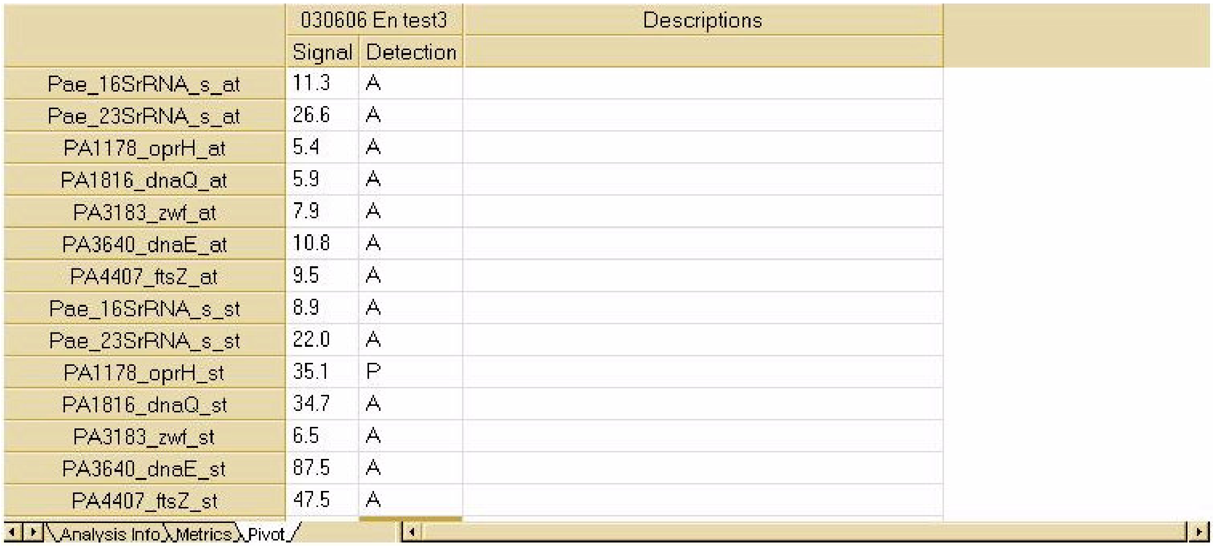


MAS5.0 Analysis Output File (*.CHP)



	Analysis Name	Probe Set Name	Stat Pairs	Stat Pairs Used	Signal	Detection	Detection p-value	Stat Comm
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**
  - □ *Validation of total RNA*
  - □ *Validation of cRNA*
  - □ *Validation of fragmented cRNA*
- **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)**

Two aspects of quality control: detecting poor hybridization and outliers

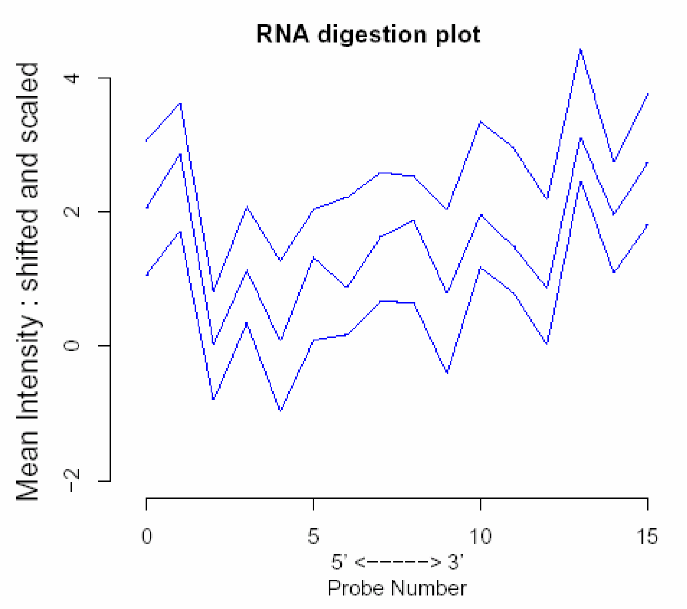
◆ Reasons for poor hybridizations
- mRNA degenerated
- one or more experimental steps failed
- poor chip quality, …

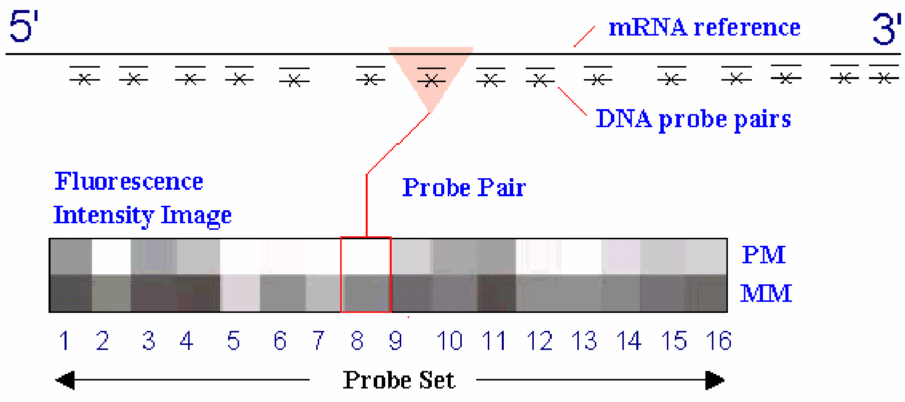
◆reasons for (biological) outliers
- infiltration with non-tumour tissue
- wrong label
- contamination, …

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.



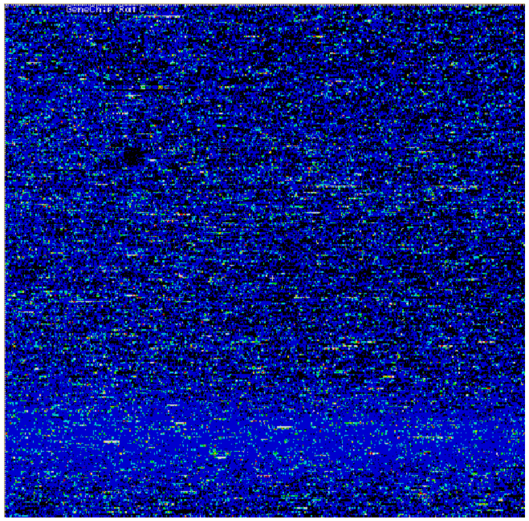


5'---------------------------3'AAAA-Poly
x-------|---------

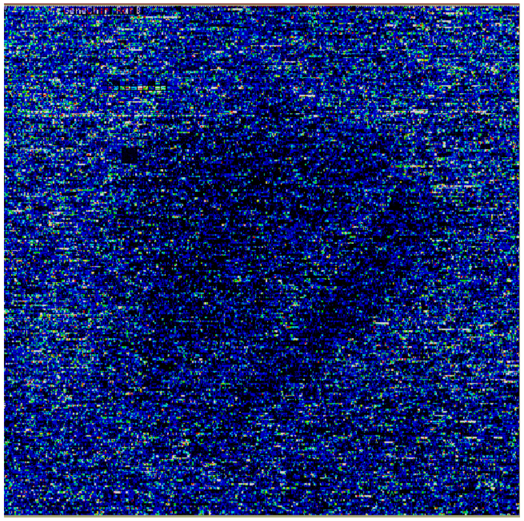
Probe Array Image Inspection

- 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.

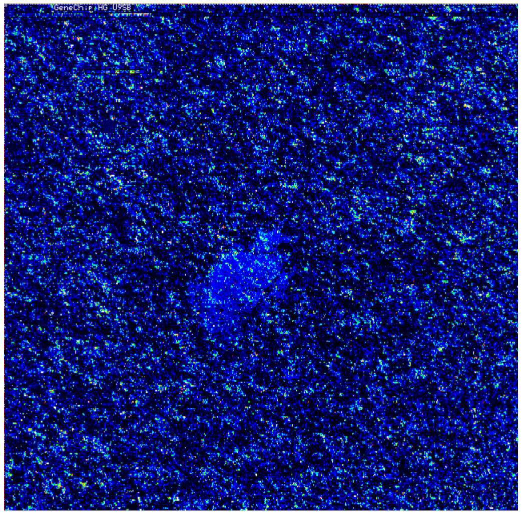
Haze Band



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.

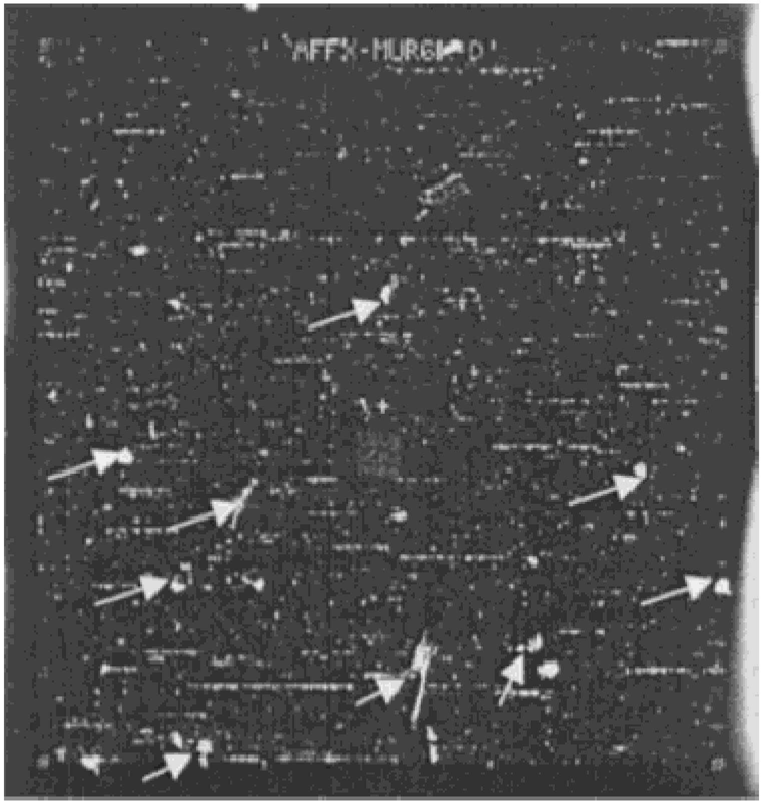


Fig. 1. A contaminated D array from the Murine 6500 Affymetrix GeneChip® set. Several particles are highlighted by arrows and are thought to be torn pieces of the chip cartridge septum, potentially resulting from repeatedly pipetting the target into the array.

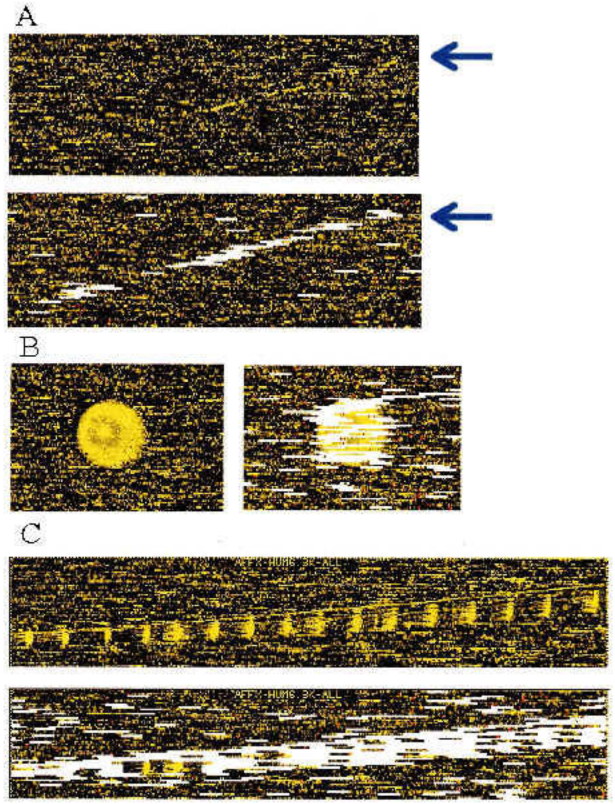
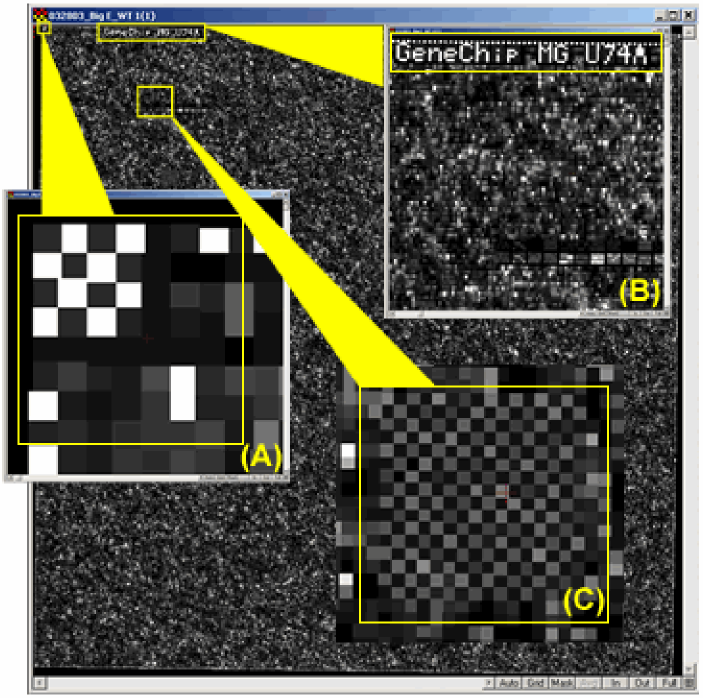


Fig. 5. (*A*) A long scratch contamination (indicated by arrow) is alleviated by automatic outlier exclusion along this scratch. (*B* and *C*) Regional clustering of array outliers (white bars) indicates contaminated regions in the original images. These outliers are automatically detected and accommodated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

PNAS | January 2, 2001 | vol. 98 | no. 1 | 33

B2 Oligo Performance

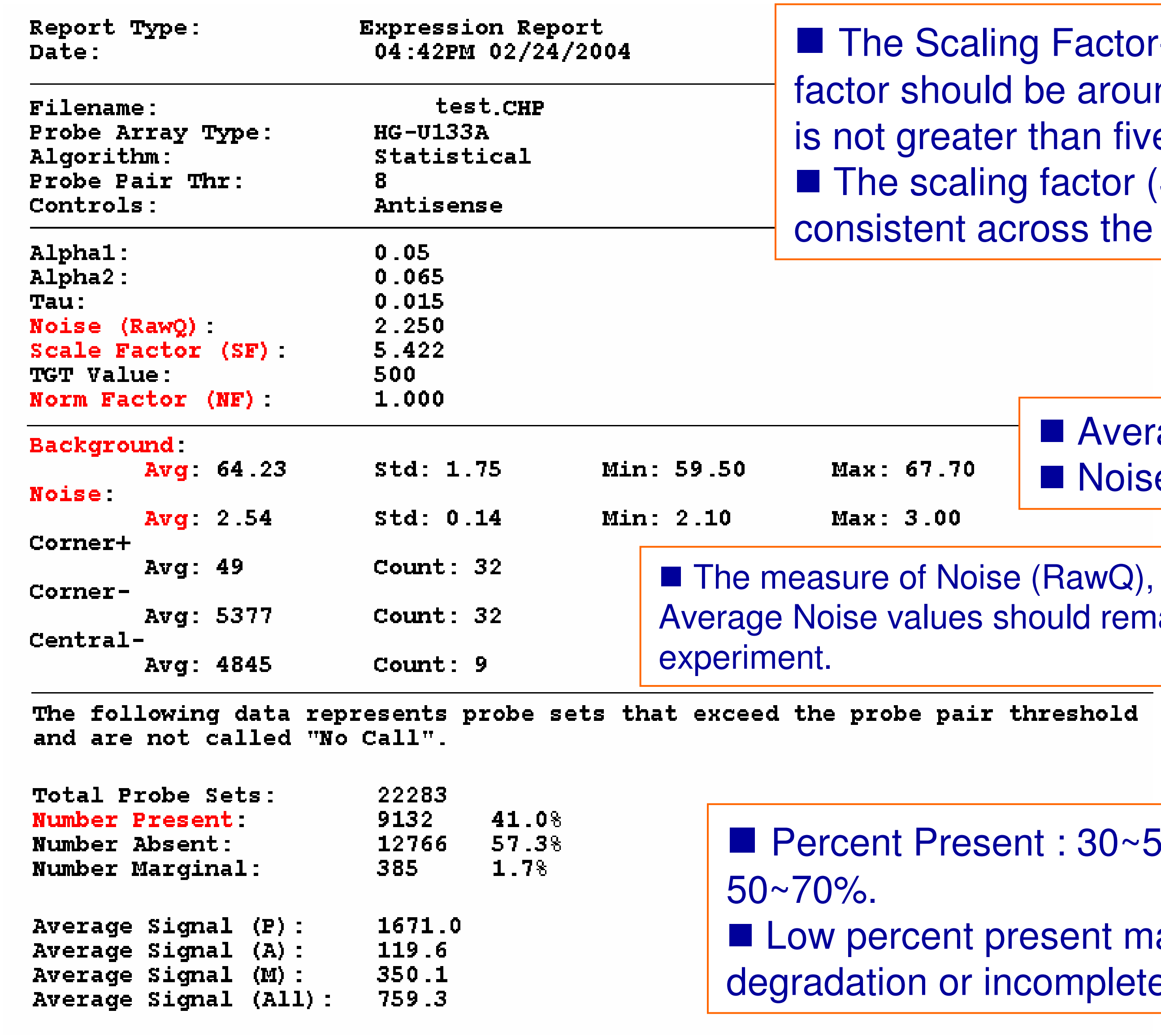
- Make sure the alignment of the grid was done appropriately.
- Look at the spiked in Oligo B2 control in order to check the hybridization uniformity.
- The border around the array, the corner region, the control regions in the center, are all checked to make sure the hybridization was successful.



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

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

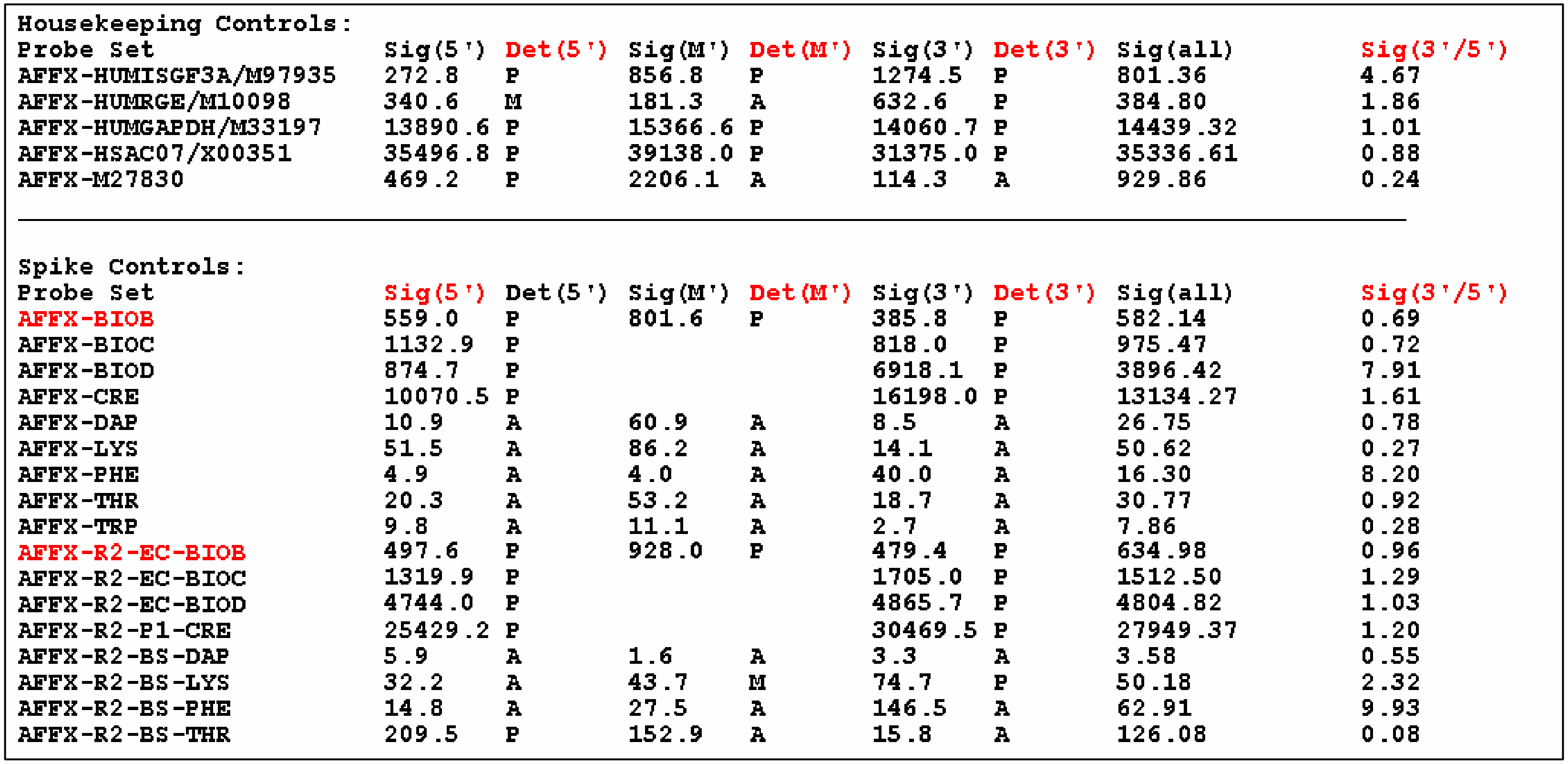
- The measure of Noise (RawQ), Average Background and Average Noise values should remain consistent across the experiment.

- Percent Present : 30~50%, 40~50%, 50~70%.
- 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.

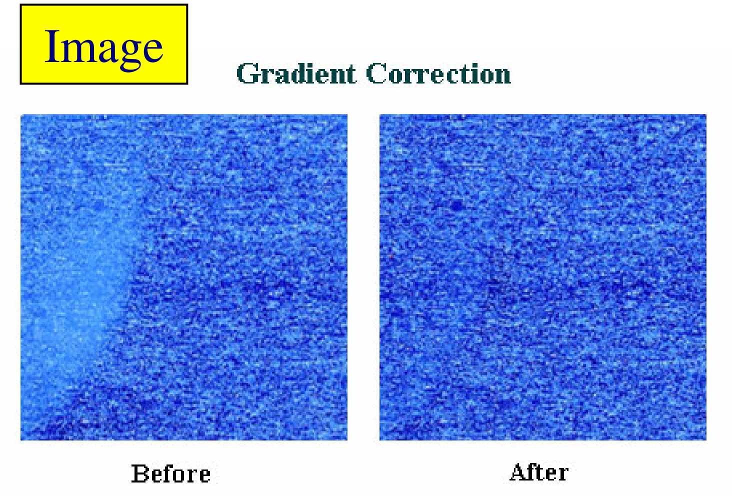
- Spike-In Controls (BioB, BioC, BioD, Cre)- These spike in controls also tell how well your labelling reaction went. BioB is only Present half of the time, but BioC, BioD, & Cre should always have a present (P) call.



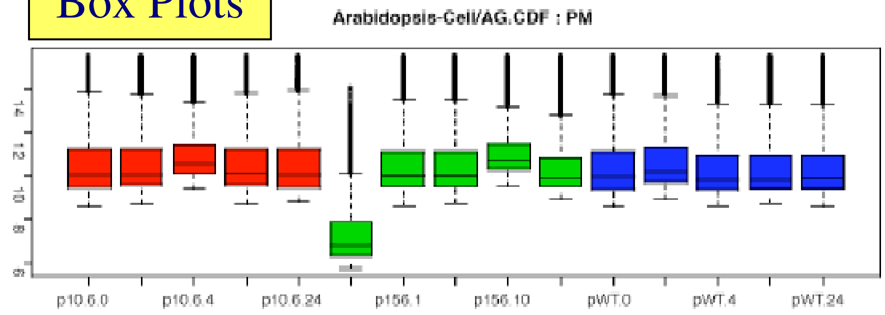
```
Housekeeping Controls:
Probe Set               Sig(5') Det(5') Sig(M') Det(M') Sig(3') Det(3') Sig(all)        Sig(3'/5')
AFFX-HUMISGF3A/M97935   272.8   P       856.8   P       1274.5  P       801.36          4.67
AFFX-HUMRGE/M10098      340.6   M       181.3   A       632.6   P       384.80          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

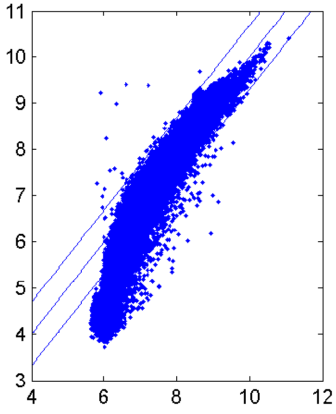
Image

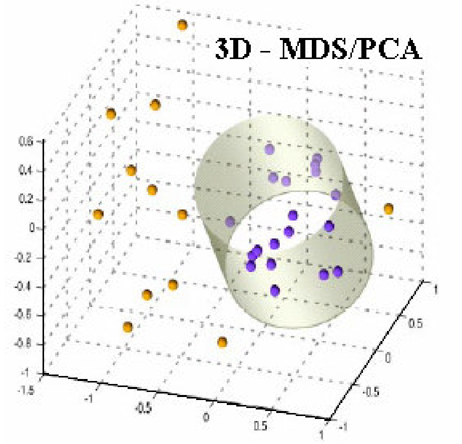


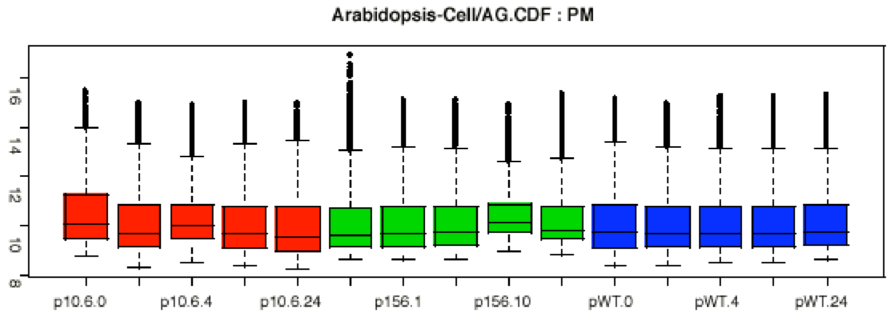
Box Plots



Scatterplot

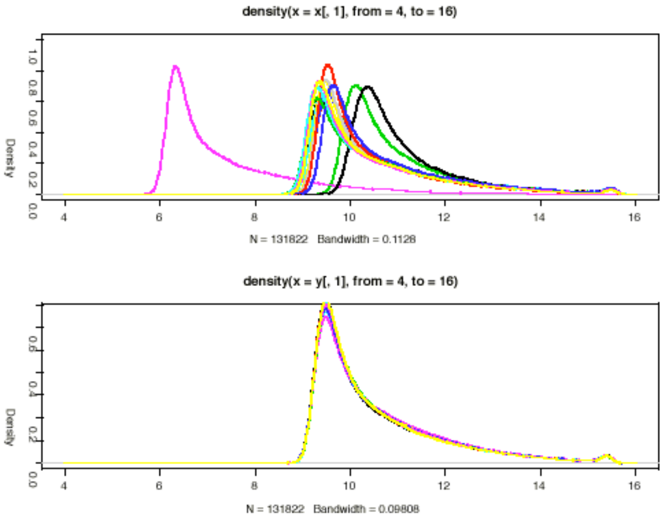




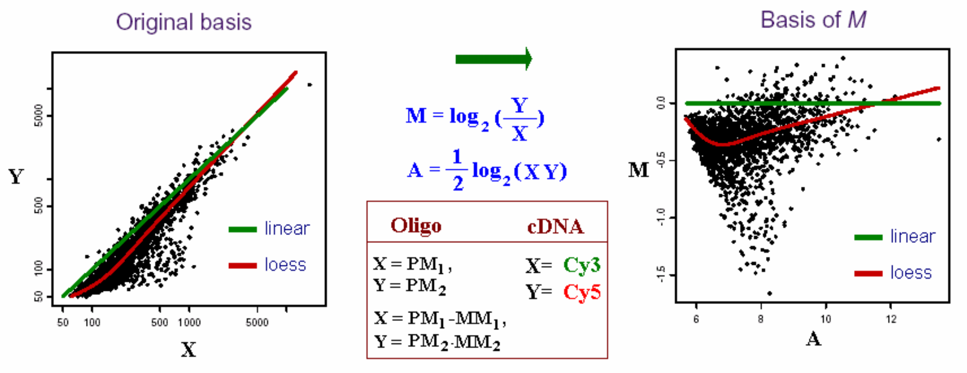


Statistical Plots (conti.)

Density Plots



MA Plot



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

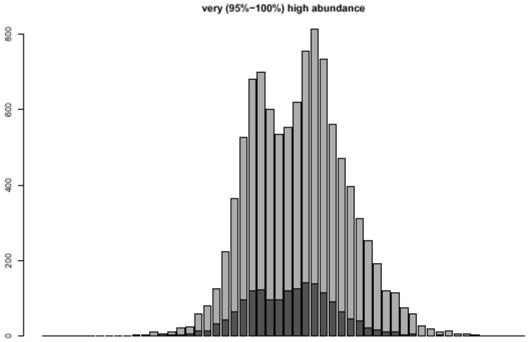
$$A = \frac{1}{2}\log_2(XY)$$

Oligo	cDNA

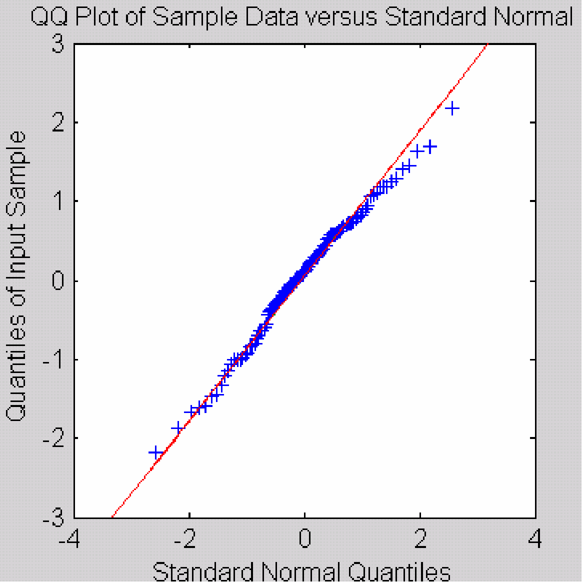
$X = PM_1$,
$Y = PM_2$
	X= Cy3
	Y= Cy5

$X = PM_1 - MM_1$, $Y = PM_2 - MM_2$	

Histogram

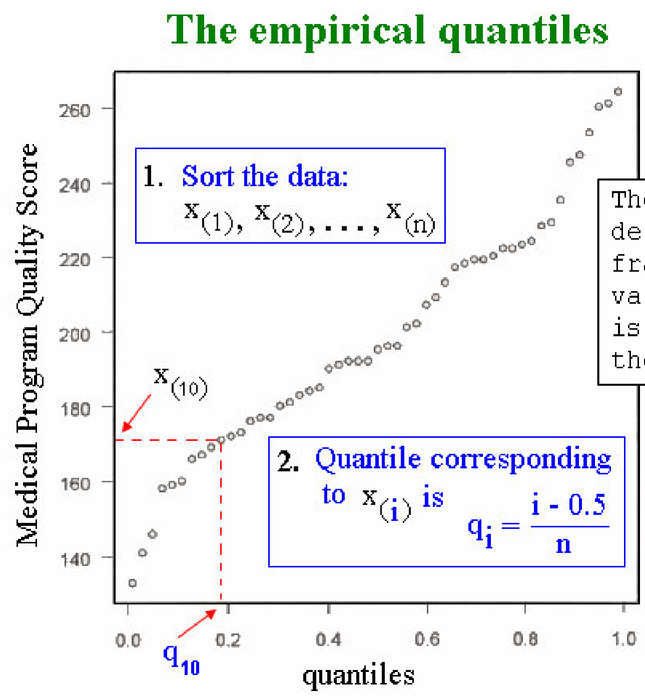


QQ-Plot



Quantile Plots

The empirical quantiles

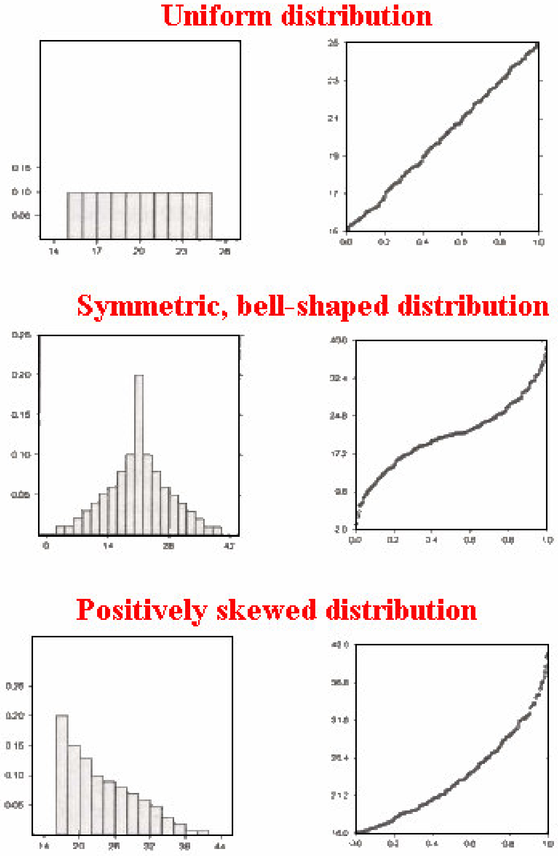


1. Sort the data: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

2. Quantile corresponding to $x_{(i)}$ is $q_i = \dfrac{i - 0.5}{n}$

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.

Comparison of histogram and Quantile plolts for differently shaped data distribution


Uniform distribution

Symmetric, bell-shaped distribution

Positively skewed distribution

Figures modified from Jacoby (1997)

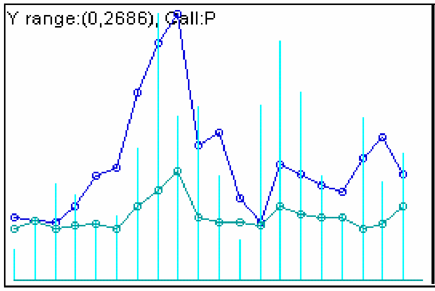
- 0.5 is subtracted from each i value to avoid extreme quantiles of exactly 0 or 1.
- The latter would cause problems if empirical quantiles were to be compared against quantiles derived from a theoretical. asymptotic distribution such as the normal.
- This adjustment has no effect on the shape of any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

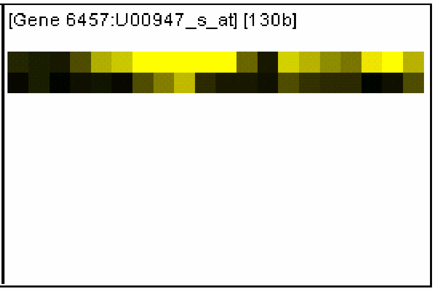
Statistical Plots (conti.)

Heatmap with Dendrogram

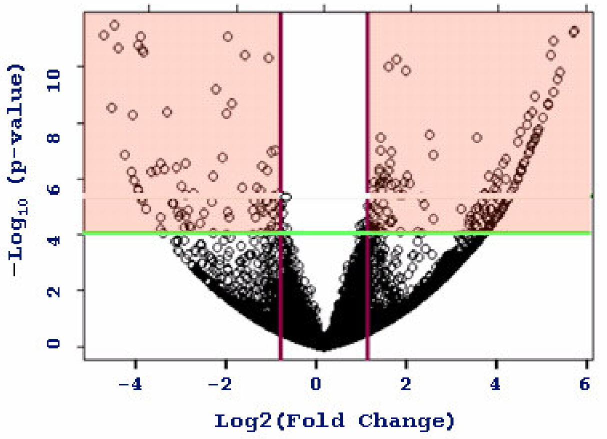
Line Plots

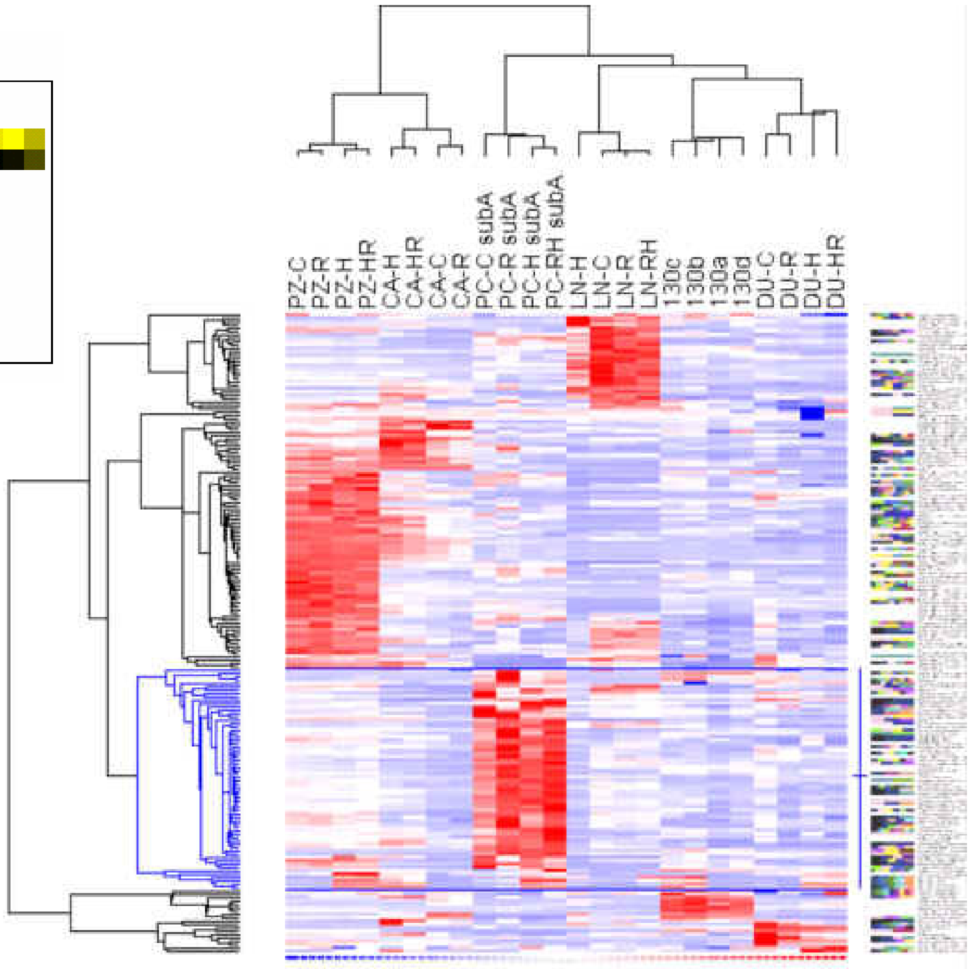
Profiles Plots





Volcano Plot

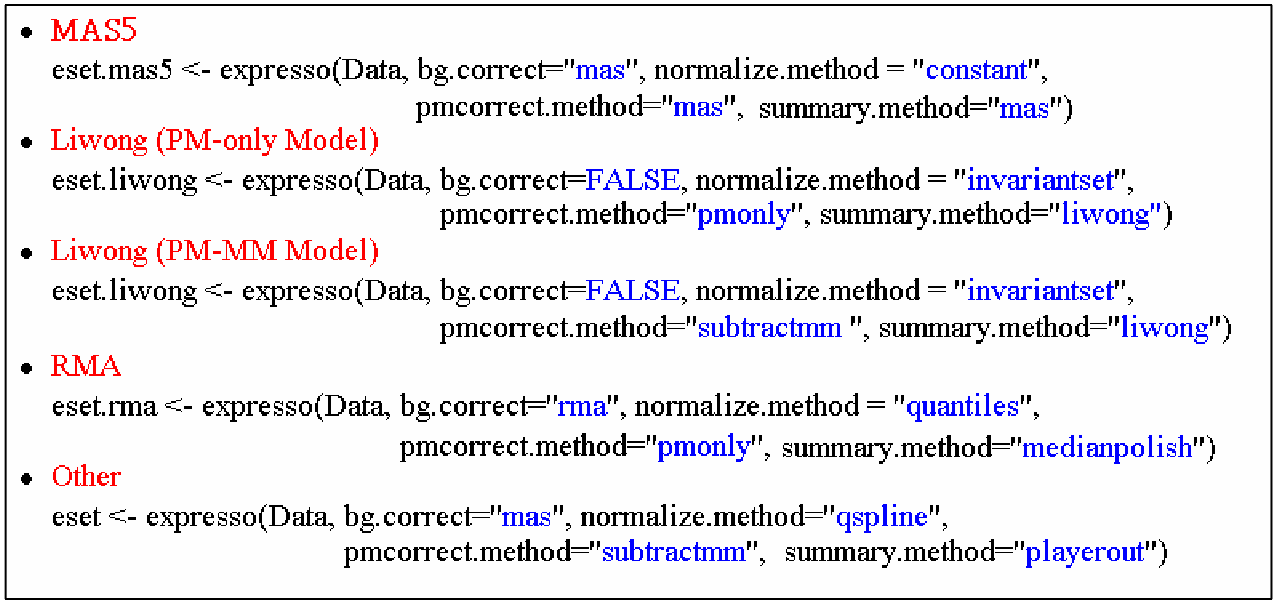




Low level analysis

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

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/Adjustment

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.

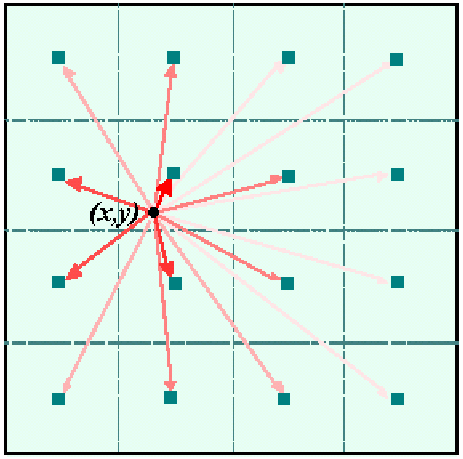
What is background correction?

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

Background Methods: MAS5

Zone Values

- For purposes of calculating background values, the array is split up into K rectangular zones Z_k (k = 1, …, 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

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

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

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

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

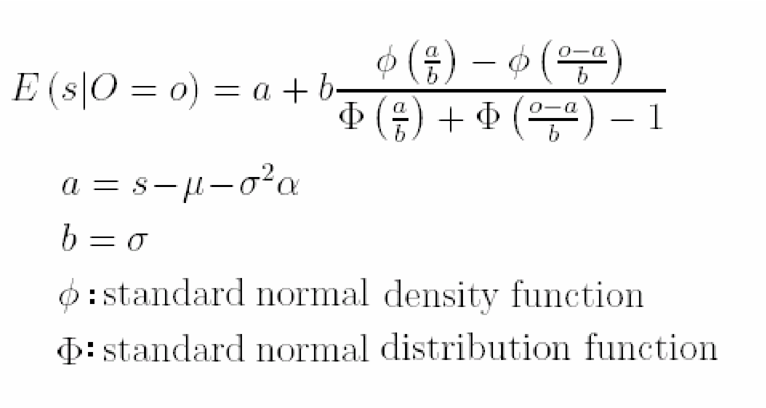
- MAS method corrects both PM and MM probes.

Affymetrix: Statistical Algorithm Description Document

Background Methods: RMA

RMA: Robust Multichip Average (Irizarry and Speed, 2003)

- Assumes PM probes are a convolution of normal and exponentional.
- ObservedPM = Signal + Noise, (O = S + N).
- **Assume**
  - Signal is exponentional (alpha)
  - Noise (background) is Normal (mu, sigma).
- Use E[S|O=o, S>0] as the backround corrected PM.
- MM probe intensities are not corrected by RMA/RMA2.



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

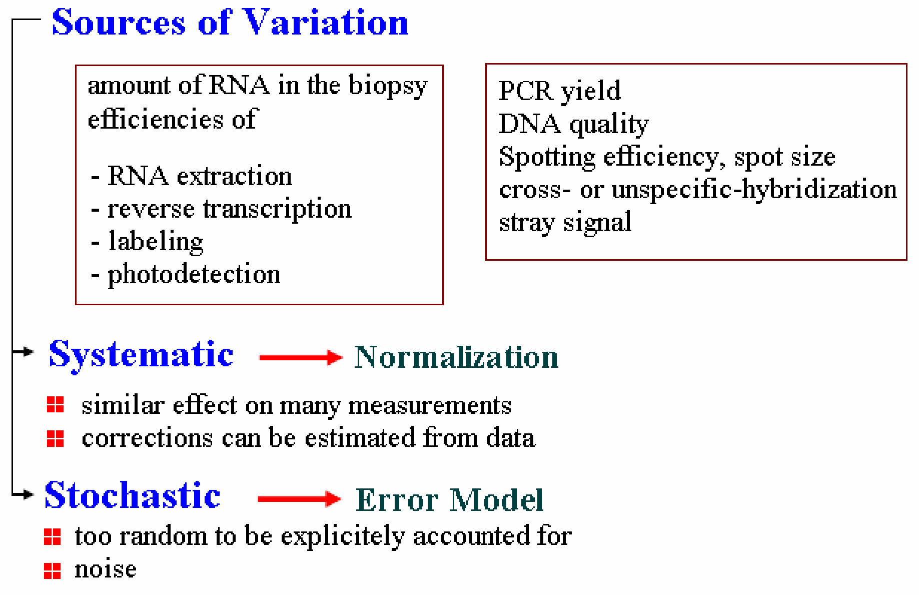
$a=s-\mu-\sigma^2\alpha$

$b=\sigma$

ϕ: standard normal density function

Φ: standard normal distribution function

Normalization



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

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.

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.

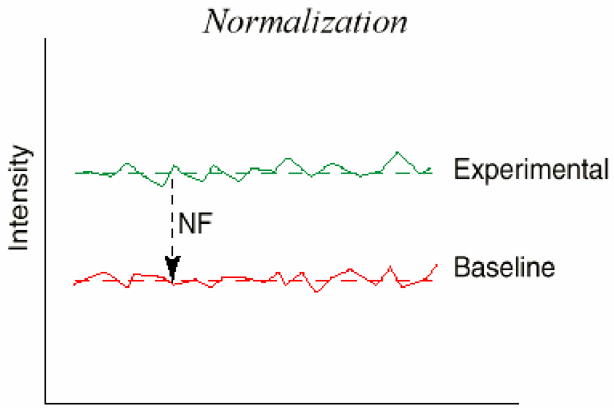
The Options on Normalization

- 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

Normalization Methods: constant

Normalization and Scaling

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

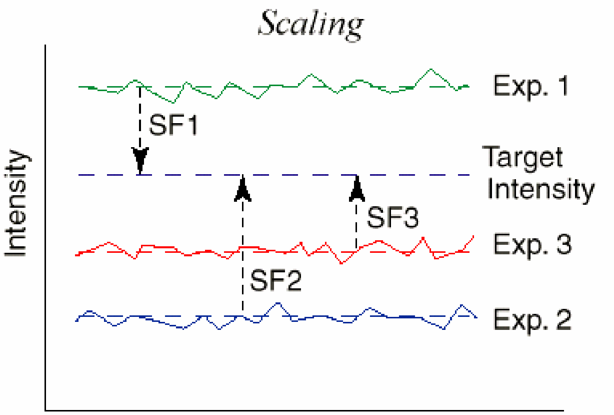

Normalization

- **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).

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

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

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


Scaling

- **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.

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

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

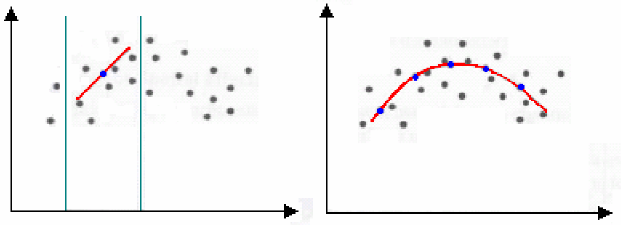
$$\Rightarrow \mathbf{NF} = \frac{\mathbf{A}_{base}}{\mathbf{A}_{exp}}$$

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.

Normalization Methods: loess

- 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 $x'_{kj} = 2^{A_K - \frac{M'_k}{2}}$.

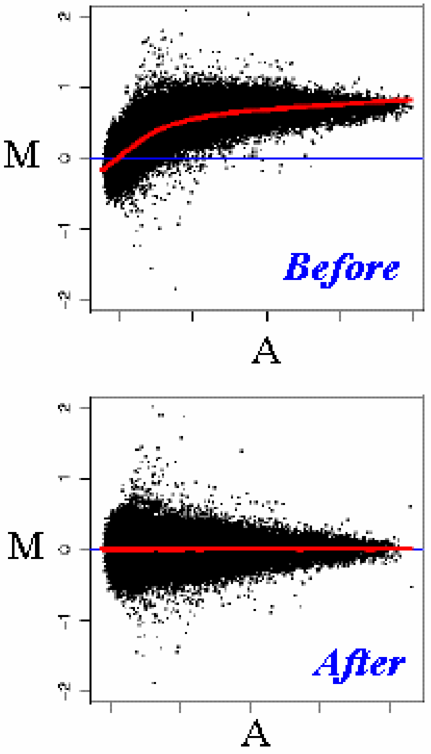
$$M = \log_2 (\frac{Y}{X})$$

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

Oligo	cDNA

X = PM$_1$,
Y = PM$_2$
	X= Cy3
	Y= Cy5

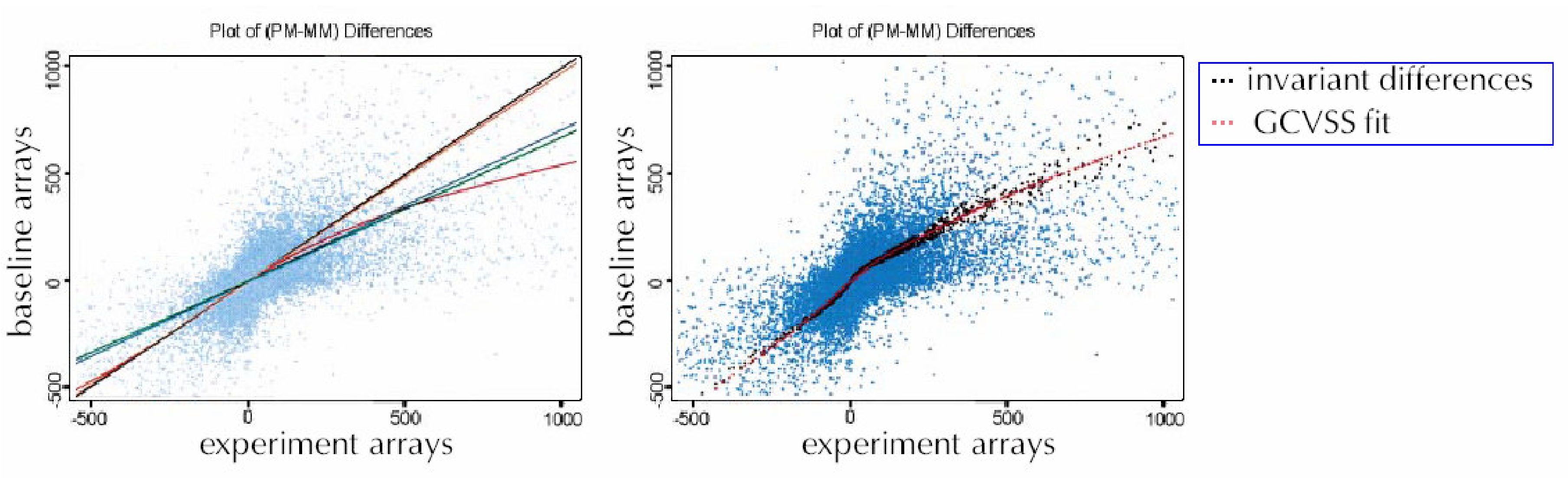
X = PM$_1$ -MM$_1$,
Y = PM$_2$-MM$_2$	



Normalization Methods: invariant set

(Li and Wong, 2001)

- 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.
- The non-linear relationship is used to carry out the normalization.
- 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:

$$R_i = \frac{[L(B_i + E_i) + H(2N - B_i - E_i)]}{2N}$$

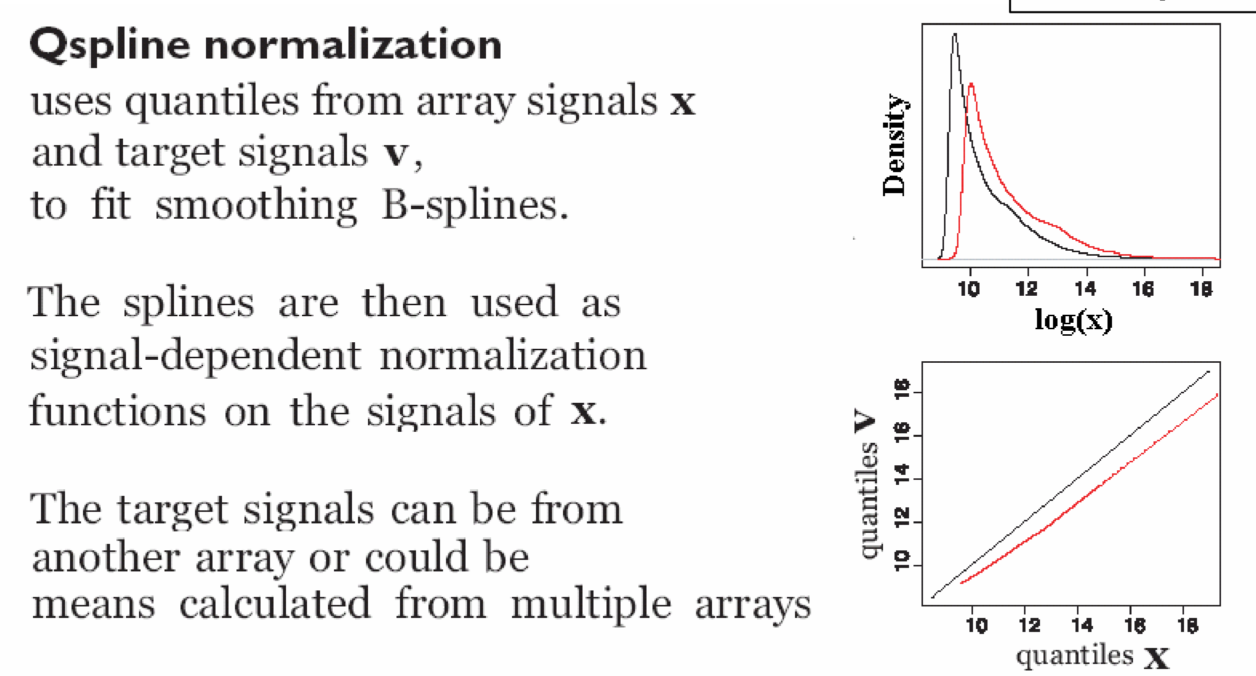
$$D_i = \frac{2|B_i - E_i|}{(B_i + E_i)}$$

- 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 ith 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 ith difference should be included in the invariant set
- The ith 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

Normalization Methods: qspline

- 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.

```
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.
```



Qspline normalization uses quantiles from array signals **x** and target signals **v**, to fit smoothing B-splines.

The splines are then used as signal-dependent normalization functions on the signals of **x**.

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

PM Correction Methods

- **PM only**

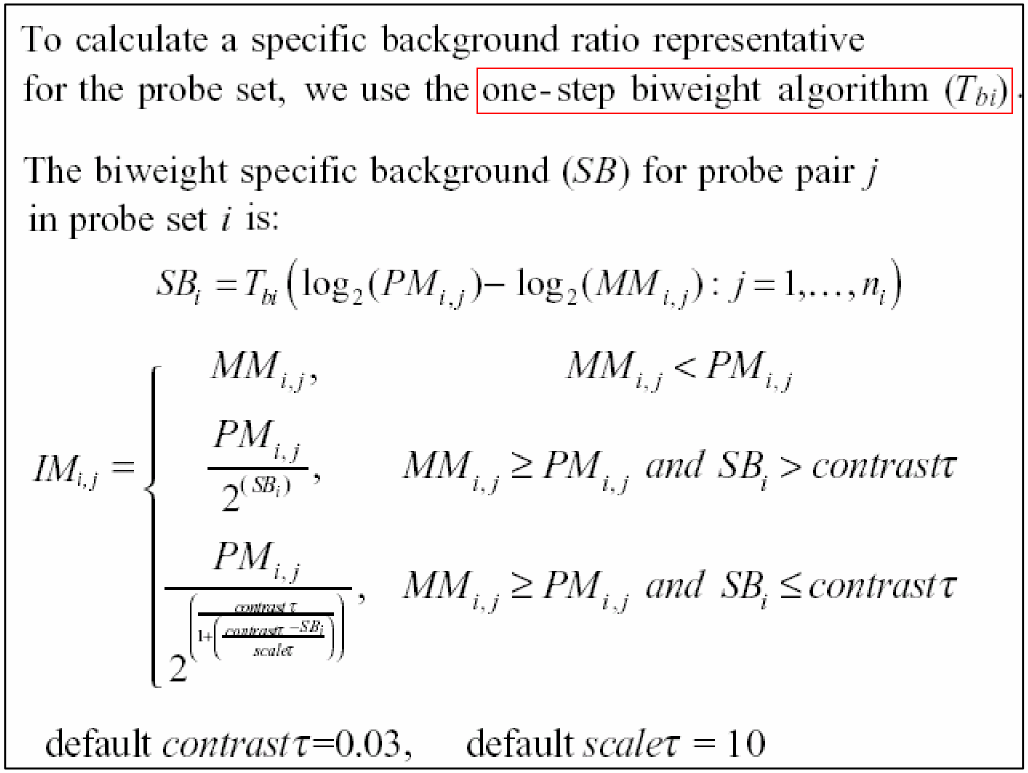
  make no adjustment to the PM values.

- **Subtract MM from PM**

  This would be the approach taken in MAS 4.0 Affymetrix (1999). It could also be used in conjuntion with the liwong model.

PM Correction Methods: MAS 5.0

- 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.



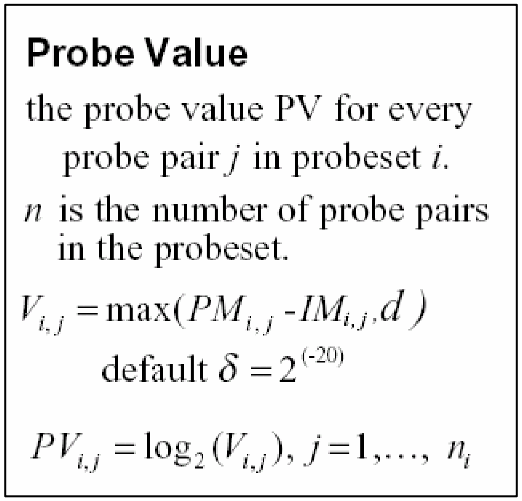
To calculate a specific background ratio representative for the probe set, we use the one-step biweight algorithm (T_{bi}).

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

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

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

default $contrast\tau$=0.03, default $scale\tau$ = 10



Probe Value

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

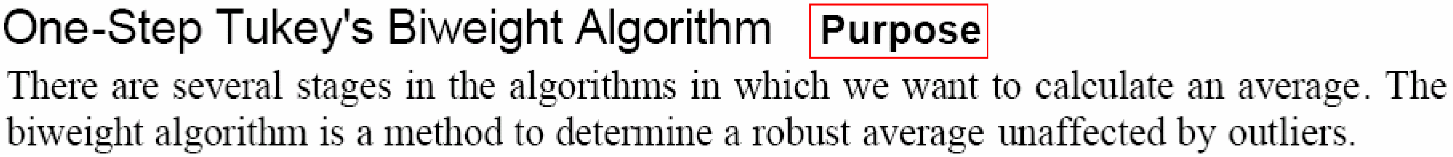
n is the number of probe pairs in the probeset.

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

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.

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

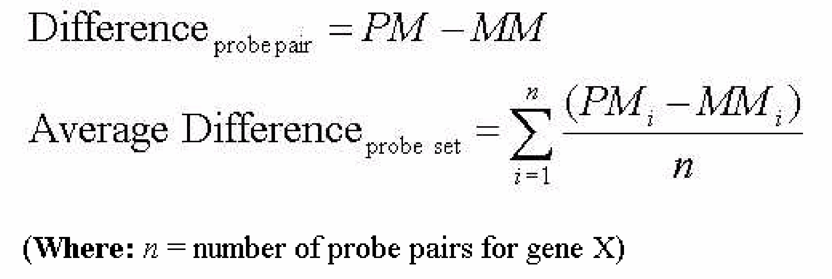
Summarization Methods: avgdiff (MAS4.0)

- **Average Difference**
  The mean intensity of a particular probe set after control correction (perfect match minus mismatch for each probe pair).

- **Absolute Expression Value**
  A value derived by certain statistical methods (depending on the array type and other factors) that is representative of the amount of RNA hybridised to the array for a particular gene. The statistical algorithms are usually provided with the scanning software.

- 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.

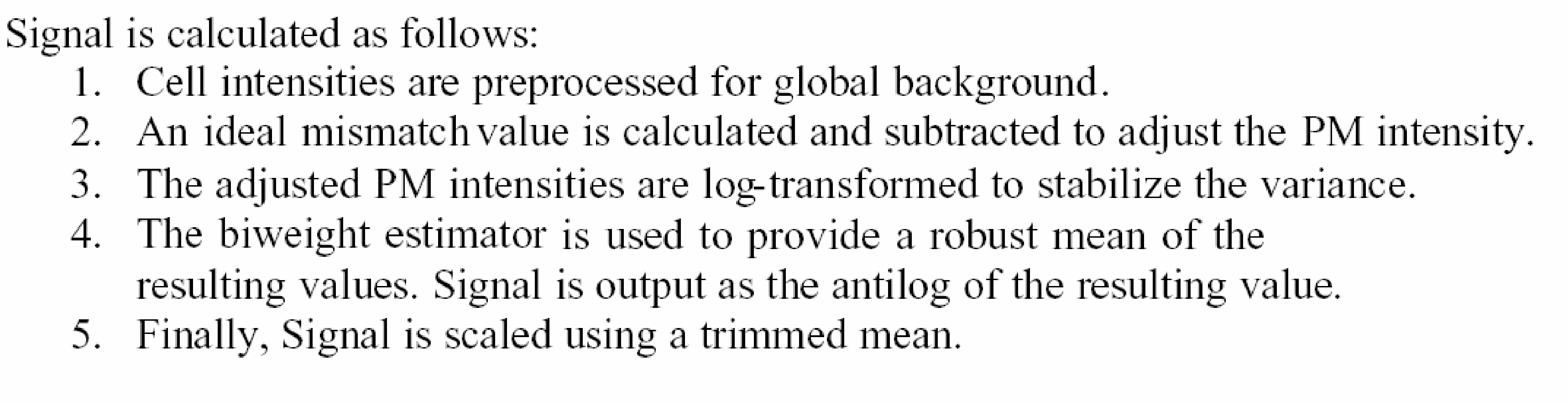


$$\text{Difference}_{\text{probe pair}} = PM - MM$$

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

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

Summarization Methods: MAS5



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.

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

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

Signal Log Value

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

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

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

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

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

Affymetrix: Statistical Algorithm Description Document

Summarization Methods: liwong

(Model-Based Expression Index , MBEI)

- If there are multiple arrays from the same experiment available, this model provides an intuitive estimate of the mean and standard error of the θ s and φ s.
  - □ The standard error estimates of the θ s and φ s can be used to identify outlier arrays and probes that will consequently be excluded from the final estimation of the probe response pattern. For each array, this model computes an expression level on the ith array θ i.
  - □ If a specific array has a large standard error relative to other arrays, possibly due to external factors like the imaging process, then this is called an **outlier array**.
  - □ Similarly, if the estimate of φ j for the jth probe has a large standard error, possibly due to non-specific cross-hybridization, it is called an **outlier probe**.
  - □ Individual PM-MM differences might also be identified by large residuals compared with the fit; these **single outliers** are regarded as missing values in the model-fitting algorithm.
- Cross-hybridization is more likely to occur at the MM probes, rather than the PM probes, and so a PM-only model exists that calculates expression values that are always positive (Li and Wong 2001). Studies suggest that the PM-only model is more robust to cross-hybridization than the PM-MM

For a gene

$$y_{ij} = \phi_i\theta_j + \epsilon_{ij}$$

y_{ij} is PM_{ij} or the difference between $PM_{ij} - MM_{ij}$.

ϕ_i is a probe response parameter

θ_j is the expression on array j.

$\sum_j \phi_j^2 = J$

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

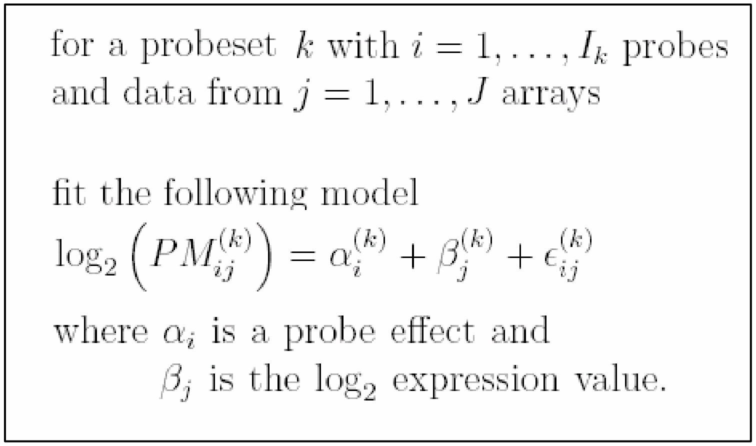
$i = 1, \ldots, I$ the number of chips

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

Other Summarization Methods

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 with $i = 1, \ldots, I_k$ probes and data from $j = 1, \ldots, J$ arrays

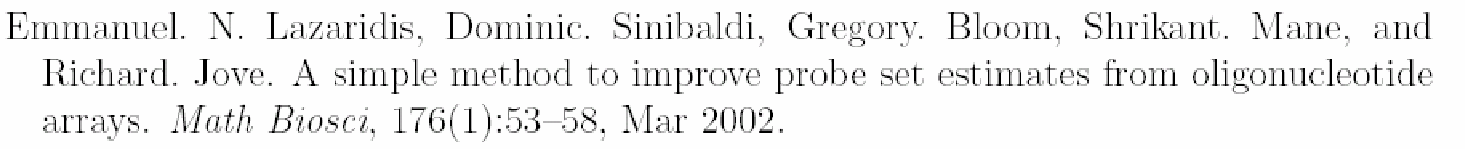
fit the following model

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

where α_i is a probe effect and β_j is the $\log_2$ expression value.

Playerout

- This method is described in Lazaridis et al. (2002).
- A non parameteric method is used to determine weights.
- The expression value is then the weighted average.



Emmanuel. N. Lazaridis, Dominic. Sinibaldi, Gregory. Bloom, Shrikant. Mane, and Richard. Jove. A simple method to improve probe set estimates from oligonucleotide arrays. *Math Biosci*, 176(1):53–58, Mar 2002.

Software

Image Analysis/Normalization

Shareware/Freeware

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

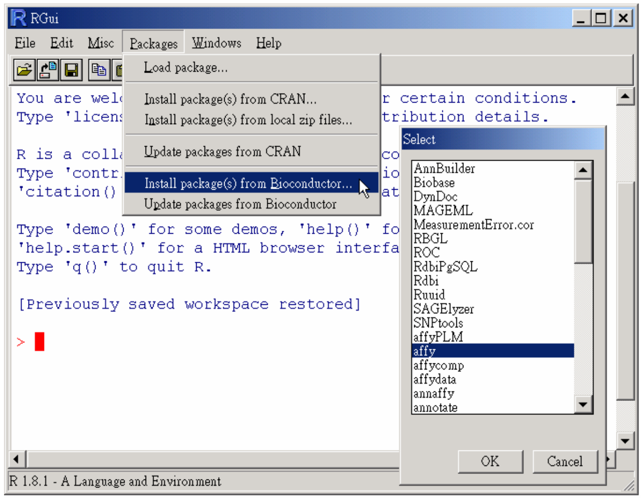
Commercial

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

The Bioconductor: affy

The Bioconductor Project
Release 1.7
http://www.bioconductor.org/

The R Project for Statistical Computing


RGui
File
Edit
Misc
Packages
Windows
Help
Load package...
Install package(s) from CRAN...
Install package(s) from local zip files...
Update packages from CRAN
Install package(s) from Bioconductor...
Update packages from Bioconductor
Select
OK
Cancel

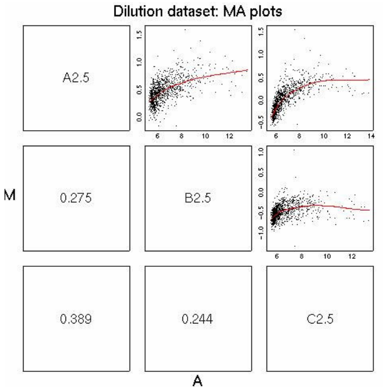
```
You are wel                                  r certain conditions.
Type 'licens                                 tribution details.

R is a colla                                 co
Type 'contr.                                 io
'citation()                                  at

Type 'demo()' for some demos, 'help()' fo
'help.start()' for a HTML browser interfa
Type 'q()' to quit R.

[Previously saved workspace restored]

>
```

AnnBuilder
Biobase
DynDoc
MAGEML
MeasurementError.cor
RBGL
ROC
RdbiPgSQL
Rdbi
Ruuid
SAGElyzer
SNPtools
affyPLM
affy
affycomp
affydata
annaffy
annotate

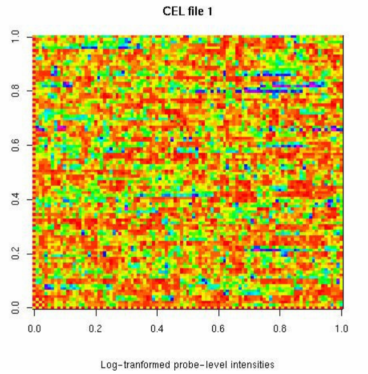
R 1.8.1 - A Language and Environment

affypdnn
affyPLM
gcrma
makecdfenv

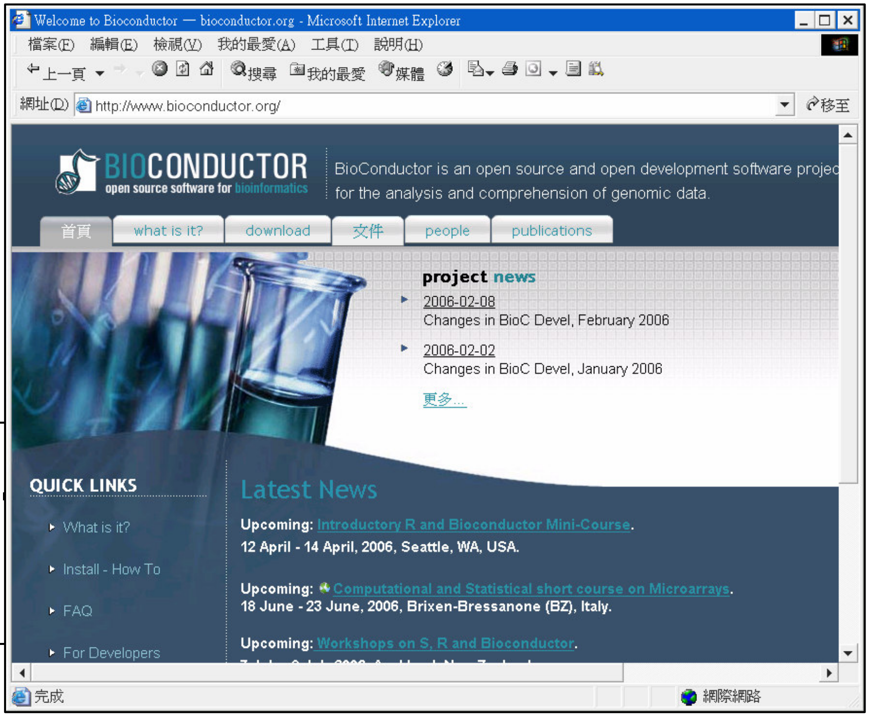
Dilution dataset: MA plots



CEL file 1



Log–tranformed probe–level intensities



http://www.bioconductor.org/

BioConductor is an open source and open development software projec
for the analysis and comprehension of genomic data.

project news

- 2006-02-08
  Changes in BioC Devel, February 2006
- 2006-02-02
  Changes in BioC Devel, January 2006

更多...

QUICK LINKS

- What is it?
- Install - How To
- FAQ
- For Developers

Latest News

Upcoming: Introductory R and Bioconductor Mini-Course.
12 April - 14 April, 2006, Seattle, WA, USA.

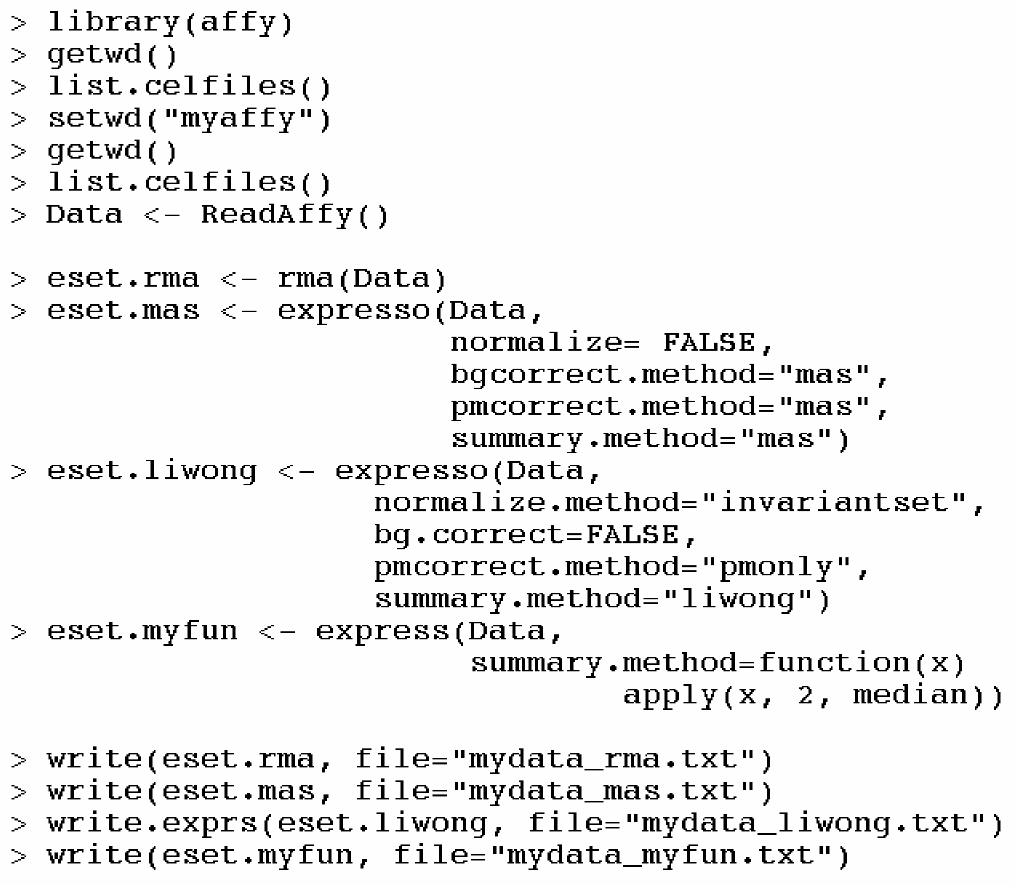
Upcoming: Computational and Statistical short course on Microarrays.
18 June - 23 June, 2006, Brixen-Bressanone (BZ), Italy.

Upcoming: Workshops on S, R and Bioconductor.

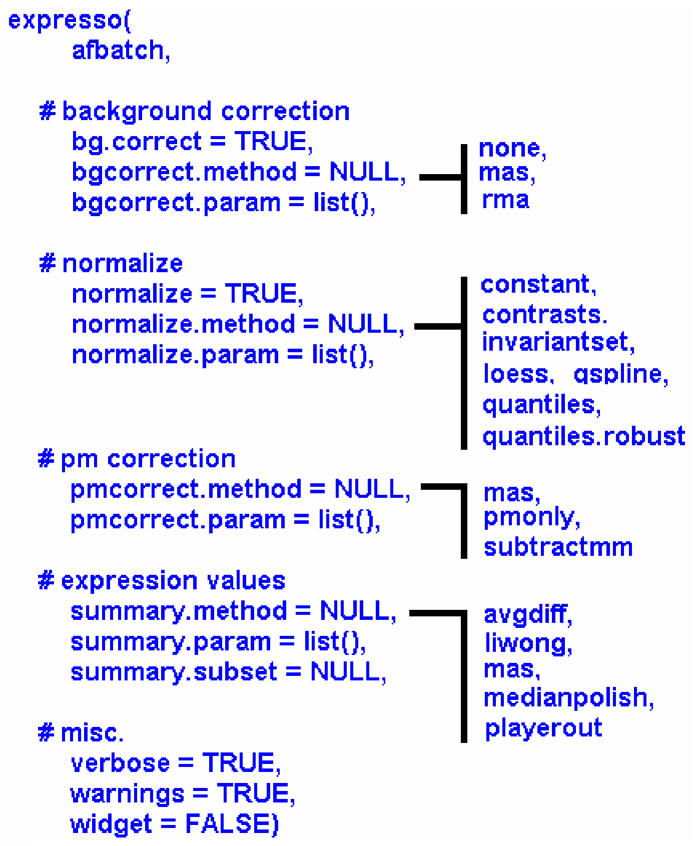
affy	Methods for Affymetrix Oligonucleotide Arrays
affycomp	Graphics Toolbox for Assessment of Affymetrix Expressi
affydata	Affymetrix Data for Demonstration Purpose
annaffy	Annotation tools for Affymetrix biological metadata
AffyExtensions	For fitting more general probe level models

The Bioconductor: affy

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

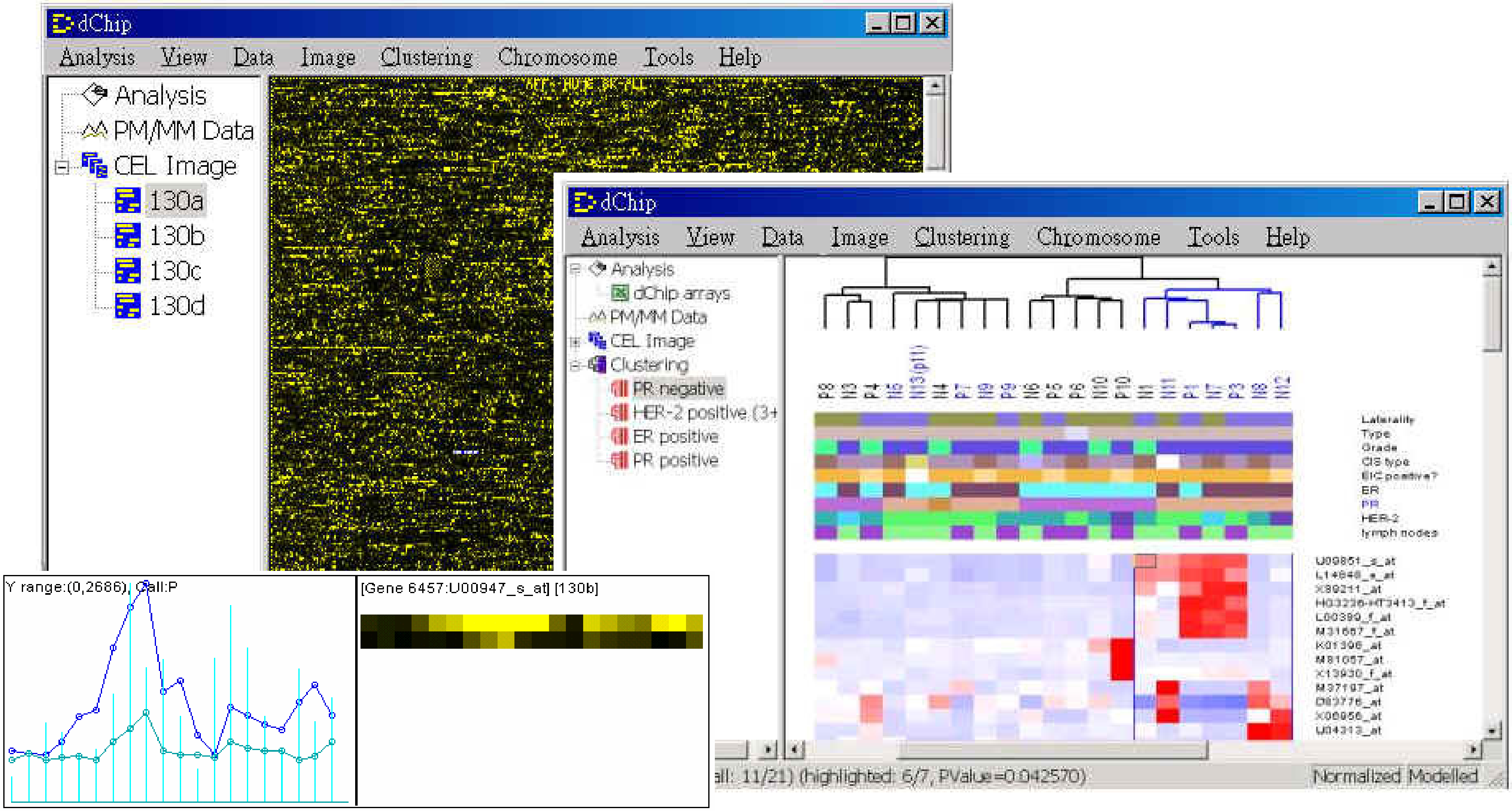


```
> library(affy)
> getwd()
> list.celfiles()
> setwd("myaffy")
> getwd()
> list.celfiles()
> Data <- ReadAffy()

> 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,                    none,
     bgcorrect.method = NULL,              mas,
     bgcorrect.param = list(),             rma

  # normalize
     normalize = TRUE,                     constant,
     normalize.method = NULL,              contrasts.
     normalize.param = list(),             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,
  # misc.                                  playerout
     verbose = TRUE,
     warnings = TRUE,
     widget = FALSE)
```

DNA-Chip Analyzer (dChip v1.3)

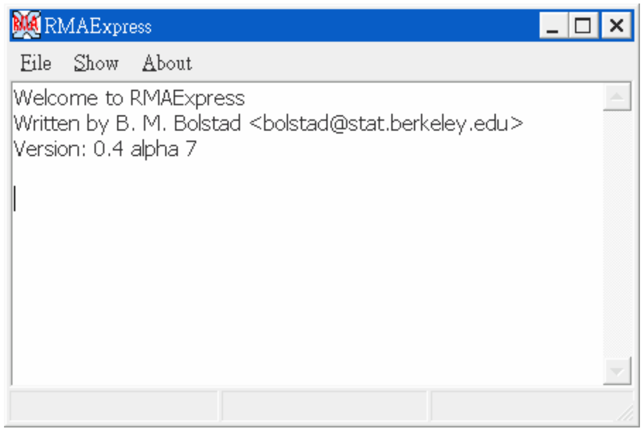


all: 11/21) (highlighted: 6/7, PValue=0.042570)

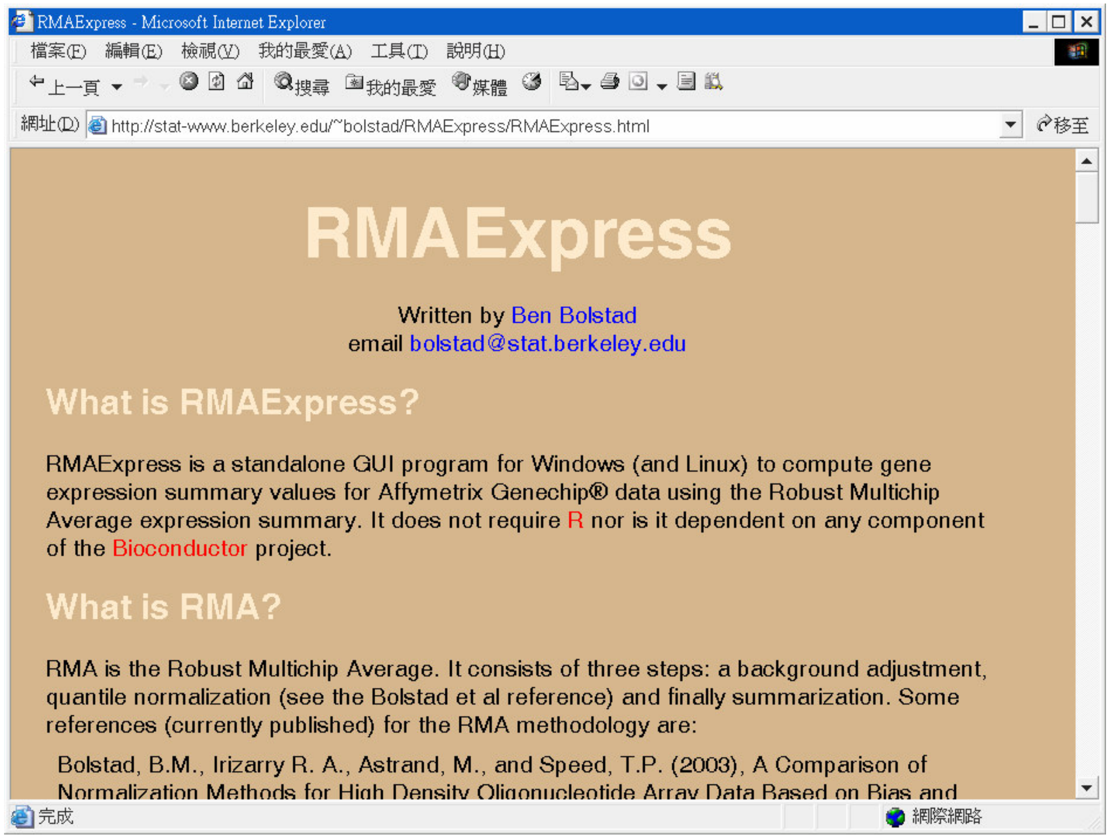
http://www.biostat.harvard.edu/complab/dchip/

RMAExpress

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



Welcome to RMAExpress
Written by B. M. Bolstad <bolstad@stat.berkeley.edu>
Version: 0.4 alpha 7


RMAExpress - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
媒體
網址(D)
http://stat-www.berkeley.edu/~bolstad/RMAExpress/RMAExpress.html
移至
完成
網際網路

RMAExpress

Written by Ben Bolstad
email bolstad@stat.berkeley.edu

What is RMAExpress?

RMAExpress is a standalone GUI program for Windows (and Linux) to compute gene expression summary values for Affymetrix Genechip® data using the Robust Multichip Average expression summary. It does not require R nor is it dependent on any component of the Bioconductor project.

What is RMA?

RMA is the Robust Multichip Average. It consists of three steps: a background adjustment, quantile normalization (see the Bolstad et al reference) and finally summarization. Some references (currently published) for the RMA methodology are:

Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and

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

GCOS V1.0

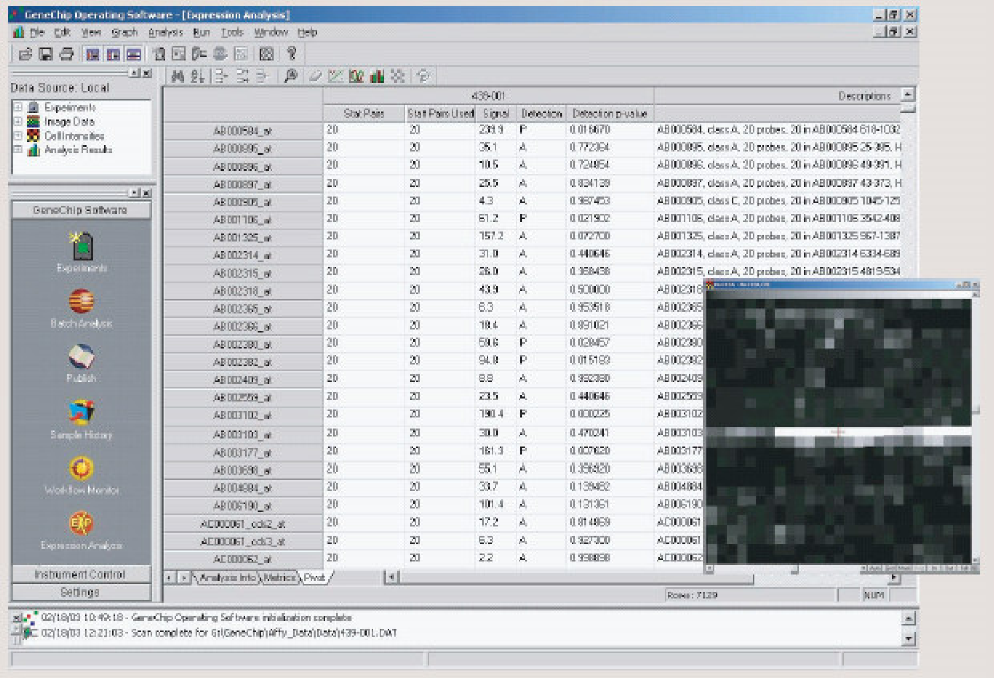
Affymetrix GeneChip Operating Software

AFFYMETRIX

http://www.affymetrix.com

Specifications

Instrument Support	• Affymetrix GeneChip® Fluidics Station 400 & 450 • GeneChip Scanner 3000 • GeneArray 2500 Scanner
Affymetrix Software Compatibility	• 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	• Microsoft Data Engine
GCOS Database	• Process Database • Publish Database • Gene Information Database
Database Management	• GCOS Manager • GCOS Administrator
Algorithm	• Affymetrix Statistical Expression Algorithm



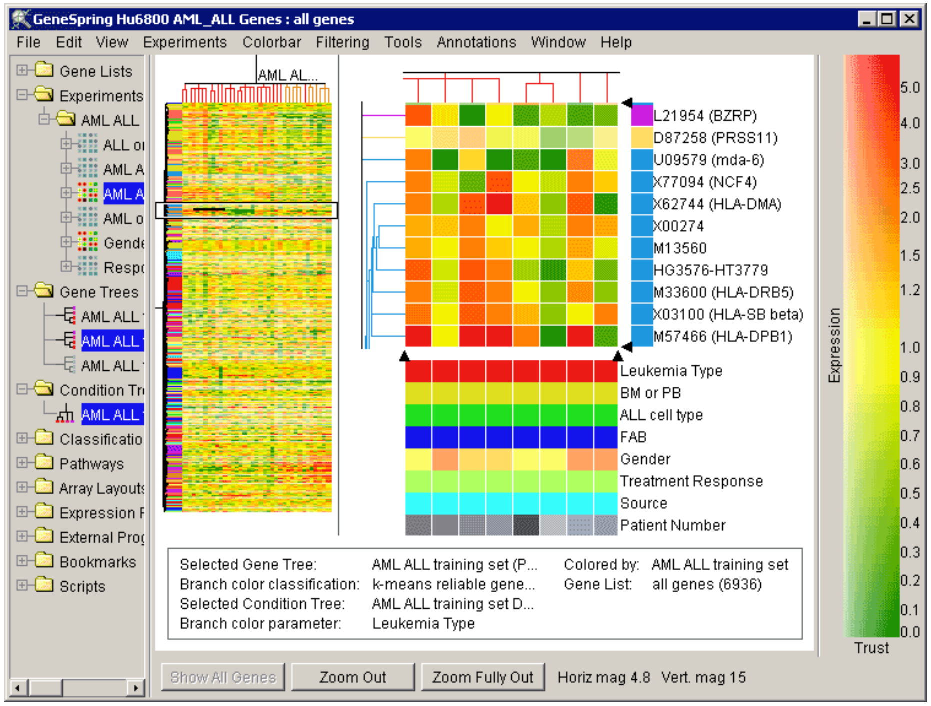
	439-001					Descriptions
	Stat Pairs	Stat Pairs Used	Signal	Detection	Detection p-value	
AB000584_at	20	20	238.9	P	0.016670	AB000584, class A, 20 probes, 20 in AB000584 618-1032
AB000895_at	20	20	35.1	A	0.772364	AB000895, class A, 20 probes, 20 in AB000895 25-395, H
AB000896_at	20	20	10.5	A	0.724854	AB000896, class A, 20 probes, 20 in AB000896 49-391, H
AB000897_at	20	20	25.5	A	0.834139	AB000897, class A, 20 probes, 20 in AB000897 43-373, H
AB000905_at	20	20	4.3	A	0.987453	AB000905, class C, 20 probes, 20 in AB000905 1045-125
AB001106_at	20	20	61.2	P	0.021902	AB001106, class A, 20 probes, 20 in AB001106 3542-408
AB001325_at	20	20	157.2	A	0.072700	AB001325, class A, 20 probes, 20 in AB001325 967-1387
AB002314_at	20	20	31.0	A	0.440646	AB002314, class A, 20 probes, 20 in AB002314 6334-689
AB002315_at	20	20	26.0	A	0.368438	AB002315, class A, 20 probes, 20 in AB002315 4819-534
AB002318_at	20	20	43.9	A	0.500000	AB00231
AB002365_at	20	20	6.3	A	0.953518	AB00236
AB002366_at	20	20	18.4	A	0.891021	AB00236
AB002380_at	20	20	59.6	P	0.028457	AB00238
AB002382_at	20	20	94.8	P	0.015193	AB00238
AB002409_at	20	20	8.8	A	0.992380	AB00240
AB002559_at	20	20	23.5	A	0.440646	AB00255
AB003102_at	20	20	190.4	P	0.000225	AB00310
AB003103_at	20	20	30.0	A	0.470241	AB00310
AB003177_at	20	20	161.3	P	0.007620	AB00317
AB003698_at	20	20	55.1	A	0.356820	AB00369
AB004884_at	20	20	33.7	A	0.139482	AB00488
AB006190_at	20	20	101.4	A	0.131361	AB00619
AC000061_cds2_at	20	20	17.2	A	0.814859	AC00006
AC000061_cds3_at	20	20	6.3	A	0.927300	AC00006
AC000062_at	20	20	2.2	A	0.938898	AC00006

Rows: 7129

```
02/18/03 10:49:18 - GeneChip Operating Software initialization complete
02/18/03 12:21:03 - Scan complete for G:\GeneChip\Affy_Data\Data\439-001.DAT
```

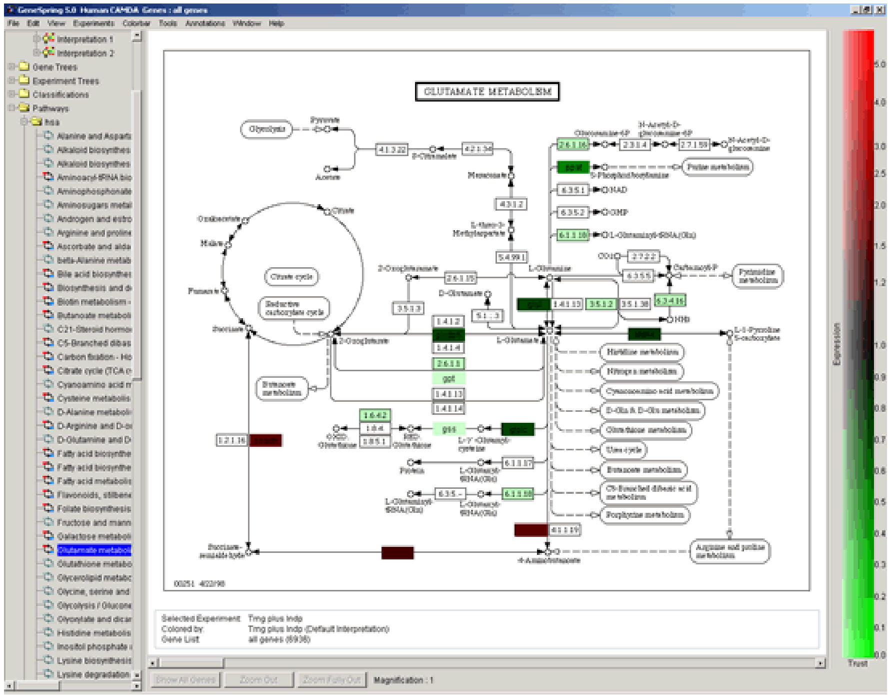
GeneSpring GX v7.3

- 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



Selected Gene Tree: AML ALL training set (P...
Branch color classification: k-means reliable gene...
Selected Condition Tree: AML ALL training set D...
Branch color parameter: Leukemia Type
Colored by: AML ALL training set
Gene List: all genes (6936)

Horiz mag 4.8 Vert. mag 15



Selected Experiment: Tmg plus Indp
Colored by: Tmg plus Indp (Default Interpretation)
Gene List: all genes (6936)

Magnification : 1

Images from
http://www.silicongenetics.com

Agilent Technologies

2004 Articles Citing GeneSpring®

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

More than 700 papers

Useful Links and Reference

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

AFFYMETRIX®

http://www.affymetrix.com

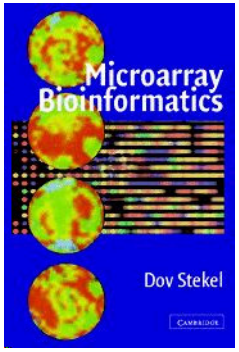
BIOINFORMATICS

http://bioinformatics.oupjournals.org

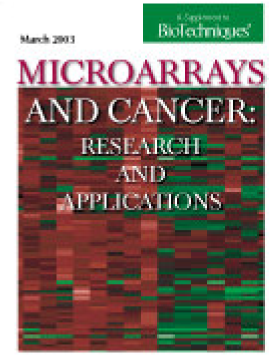
Bibliography on
Microarray Data Analysis

http://www.nslij-genetics.org/microarray/

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



- 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

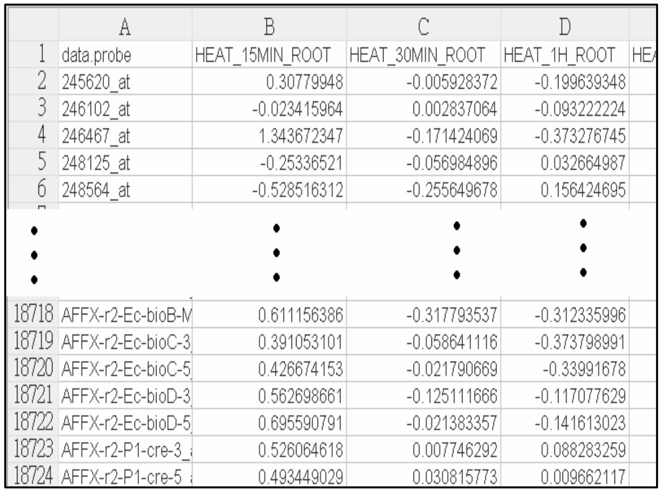


Microarrays and Cancer: Research and Applications
http://www.biotechniques.com/microarrays/

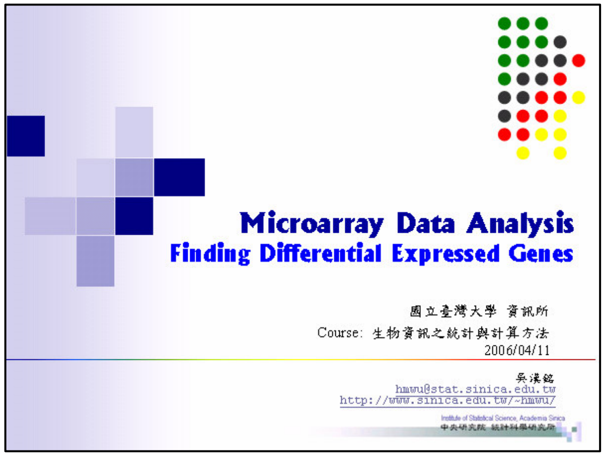
Google™

Next Class on 2006/4/11

The Normalized Data



	A	B	C	D	HE
1	data.probe	HEAT_15MIN_ROOT	HEAT_30MIN_ROOT	HEAT_1H_ROOT	HE
2	245620_at	0.30779948	-0.005928372	-0.199639348	
3	246102_at	-0.023415964	0.002837064	-0.093222224	
4	246467_at	1.343672347	-0.171424069	-0.373276745	
5	248125_at	-0.25336521	-0.056984896	0.032664987	
6	248564_at	-0.528516312	-0.255649678	0.156424695	
⋮		⋮	⋮	⋮	
18718	AFFX-r2-Ec-bioB-M	0.611156386	-0.317793537	-0.312335996	
18719	AFFX-r2-Ec-bioC-3	0.391053101	-0.058641116	-0.373798991	
18720	AFFX-r2-Ec-bioC-5	0.426674153	-0.021790669	-0.33991678	
18721	AFFX-r2-Ec-bioD-3	0.562698661	-0.125111666	-0.117077629	
18722	AFFX-r2-Ec-bioD-5	0.695590791	-0.021383357	-0.141613023	
18723	AFFX-r2-P1-cre-3_	0.526064618	0.007746292	0.088283259	
18724	AFFX-r2-P1-cre-5_	0.493449029	0.030815773	0.009662117	



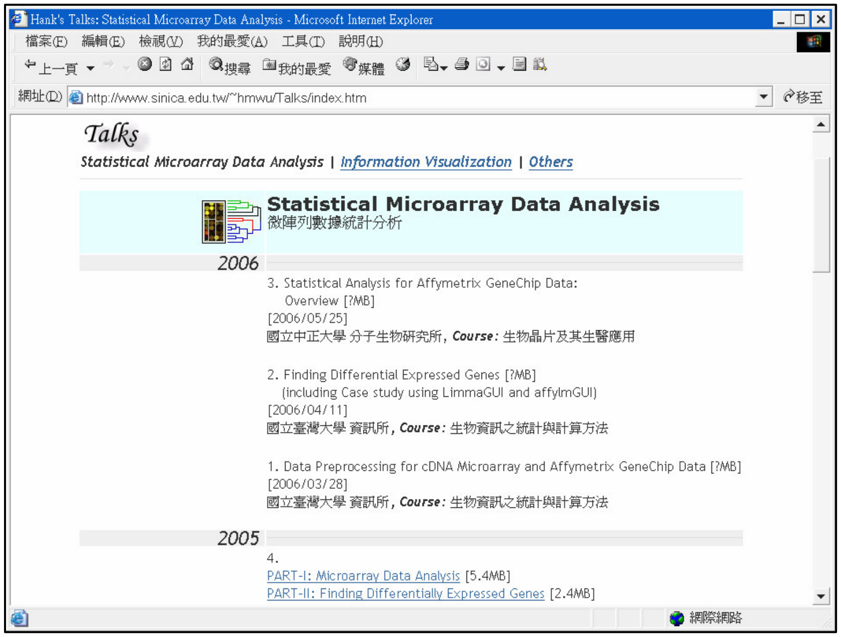
Microarray Data Analysis
Finding Differential Expressed Genes

國立臺灣大學 資訊所
Course: 生物資訊之統計與計算方法
2006/04/11

吳漢銘
hmwu@stat.sinica.edu.tw
http://www.sinica.edu.tw/~hmwu/

Institute of Statistical Science, Academia Sinica
中央研究院 統計科學研究所

- **Gene Filtering**
- **Finding DE Genes**
- **Case Study**



Hank's Talks: Statistical Microarray Data Analysis - Microsoft Internet Explorer

http://www.sinica.edu.tw/~hmwu/Talks/index.htm

Talks

Statistical Microarray Data Analysis | Information Visualization | Others

Statistical Microarray Data Analysis
微陣列數據統計分析

2006

3. Statistical Analysis for Affymetrix GeneChip Data:
   Overview [?MB]
[2006/05/25]
國立中正大學 分子生物研究所, ***Course***: 生物晶片及其生醫應用

2. Finding Differential Expressed Genes [?MB]
   (including Case study using LimmaGUI and affylmGUI)
[2006/04/11]
國立臺灣大學 資訊所, ***Course***: 生物資訊之統計與計算方法

1. Data Preprocessing for cDNA Microarray and Affymetrix GeneChip Data [?MB]
[2006/03/28]
國立臺灣大學 資訊所, ***Course***: 生物資訊之統計與計算方法

2005

4.
PART-I: Microarray Data Analysis [5.4MB]
PART-II: Finding Differentially Expressed Genes [2.4MB]

吳漢銘
hmwu@stat.sinica.edu.tw
http://www.sinica.edu.tw/~hmwu
中央研究院 統計科學研究所
Institute of Statistical Science, Academia Sinica