Statistical Microarray Data Analysis

陽明大學 臨床醫學研究所

Course:生物資訊學在醫學研究的應用

2008/04/24

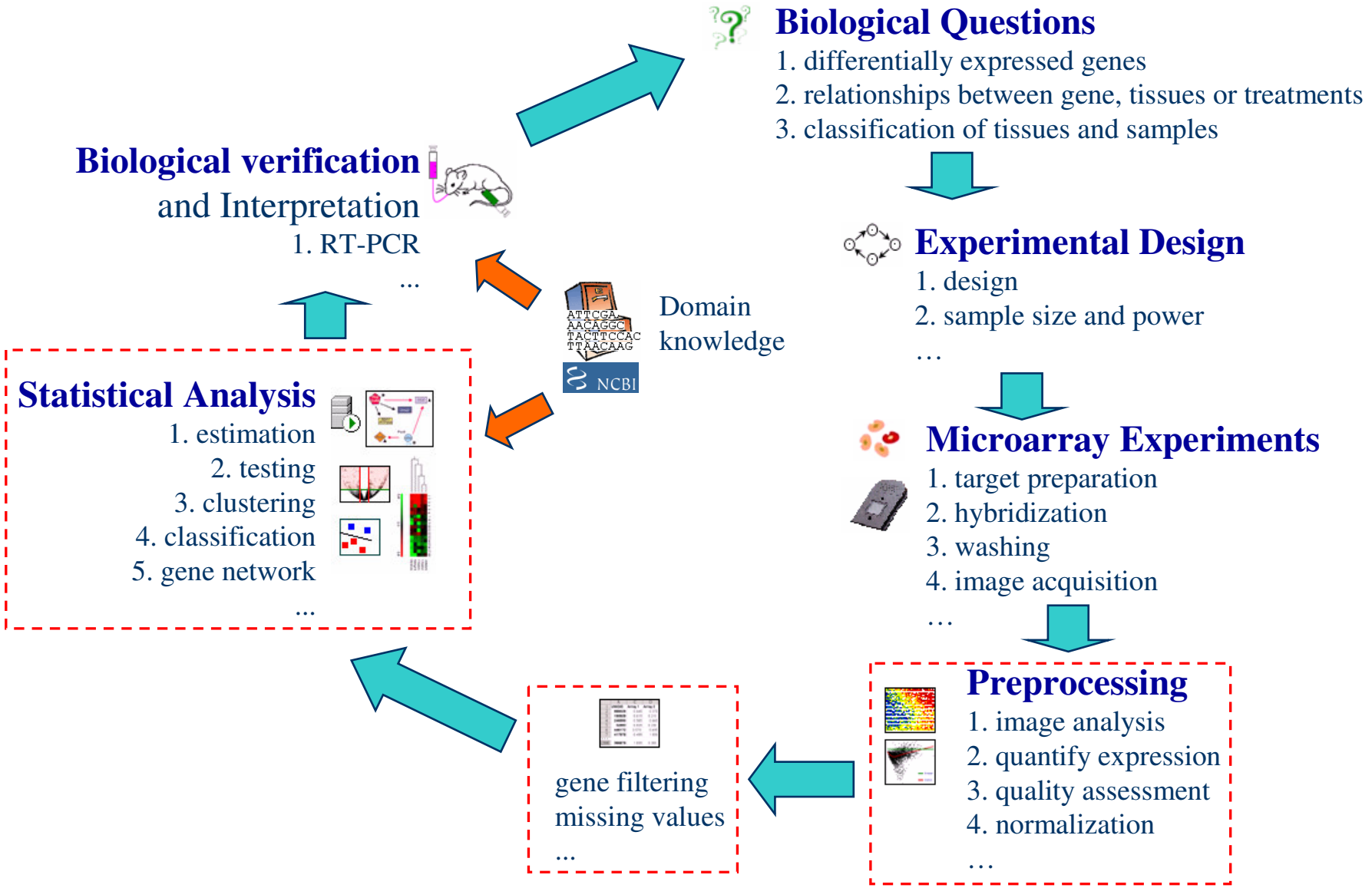
淡江大學 數學系

吳漢銘 助理教授

hmwu@mail.tku.edu.tw

http://www.hmwu.idv.tw

Microarray Life Cycle



Statistical Issues

Basic Issues:

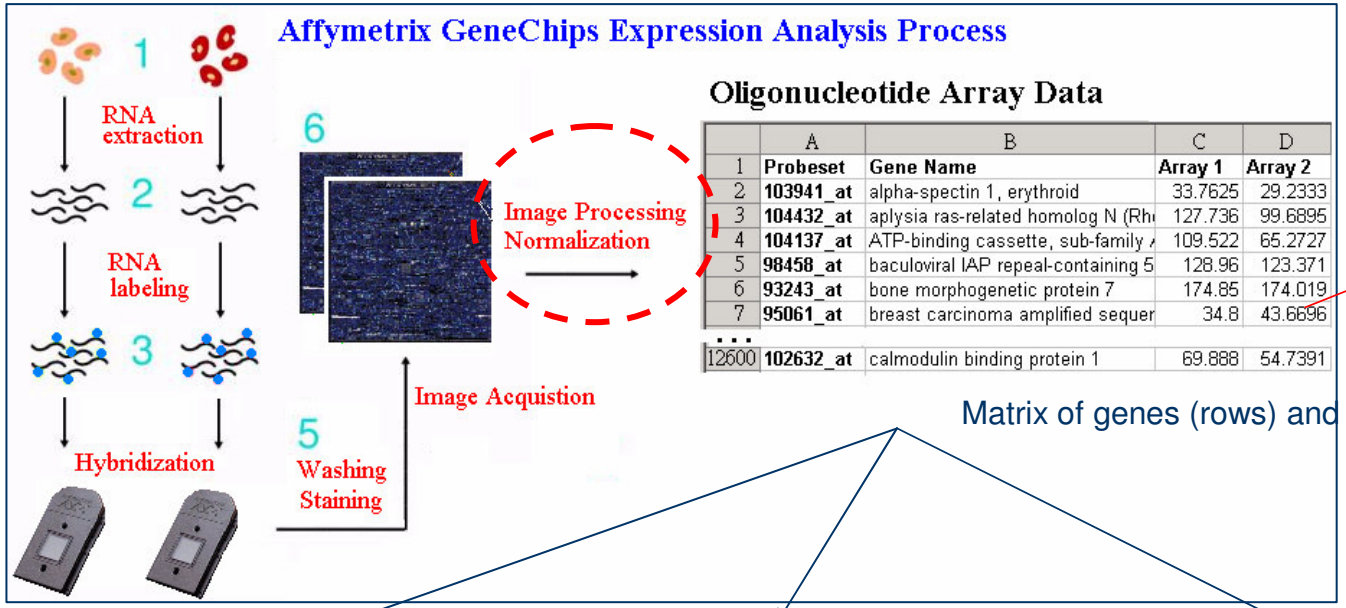
- Data Preprocessing
- Gene Filtering, Missing Values Imputation
- Finding Differential Expressed Genes
- Visualization
- Clustering
- Classification
- ...

Advance Issues:

- Experimental Design
- Time Course Microarray Experiments
- Gene Regulatory Networks/Pathway
- Annotations/Databases
- Comparisons, Sample Size, Dye Swap, Replicates, …
- Web Resource, Software Design
- ...

Data Preprocessing for GeneChip Microarray Data

Overview of Microarray Analysis



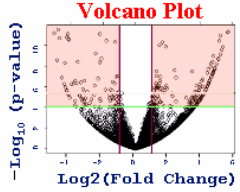
	A	B	C	D
1	Probeset	Gene Name	Array 1	Array 2
2	103941_at	alpha-spectin 1, erythroid	33.7625	29.2333
3	104432_at	aplysia ras-related homolog N (Rh	127.736	99.6895
4	104137_at	ATP-binding cassette, sub-family A	109.522	65.2727
5	98458_at	baculoviral IAP repeal-containing 5	128.96	123.371
6	93243_at	bone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

Discovery of differentially expressed genes

Parametric : t-test

Non-parametric : Wilcoxon, Mann-Whitney test

Volcano Plot

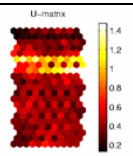


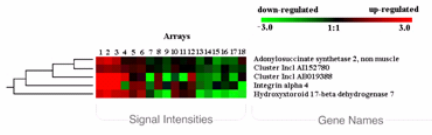
Unsupervised: clustering

Hierarchical clustering

K-means clustering

Self-organizing maps

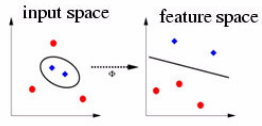




Supervised: classification

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

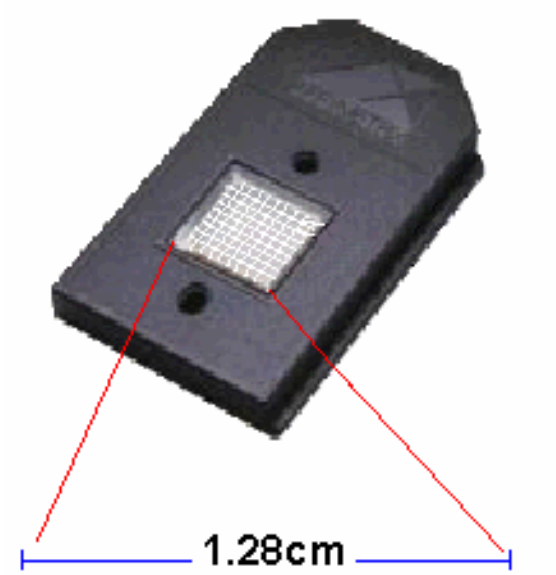
Support Vector Classifiers



Boser, Guyon, and Vapnik (1992)

Biological Relevance

GeneChip Expression Array Design



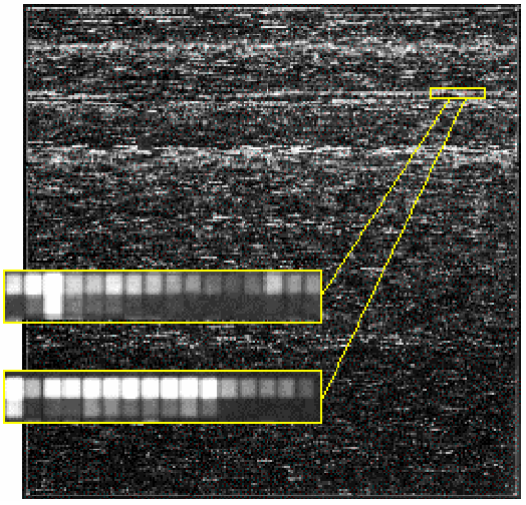
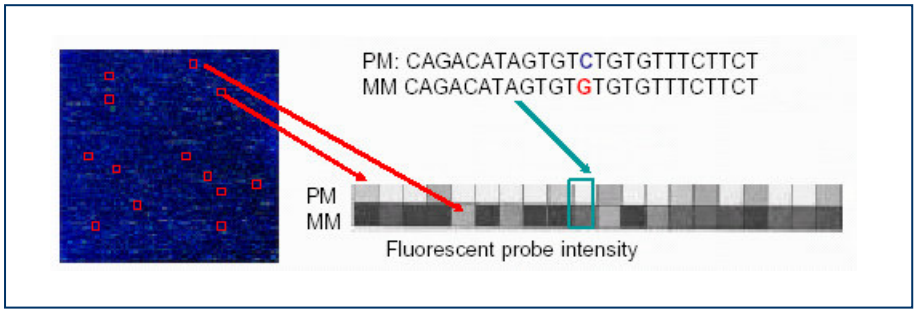
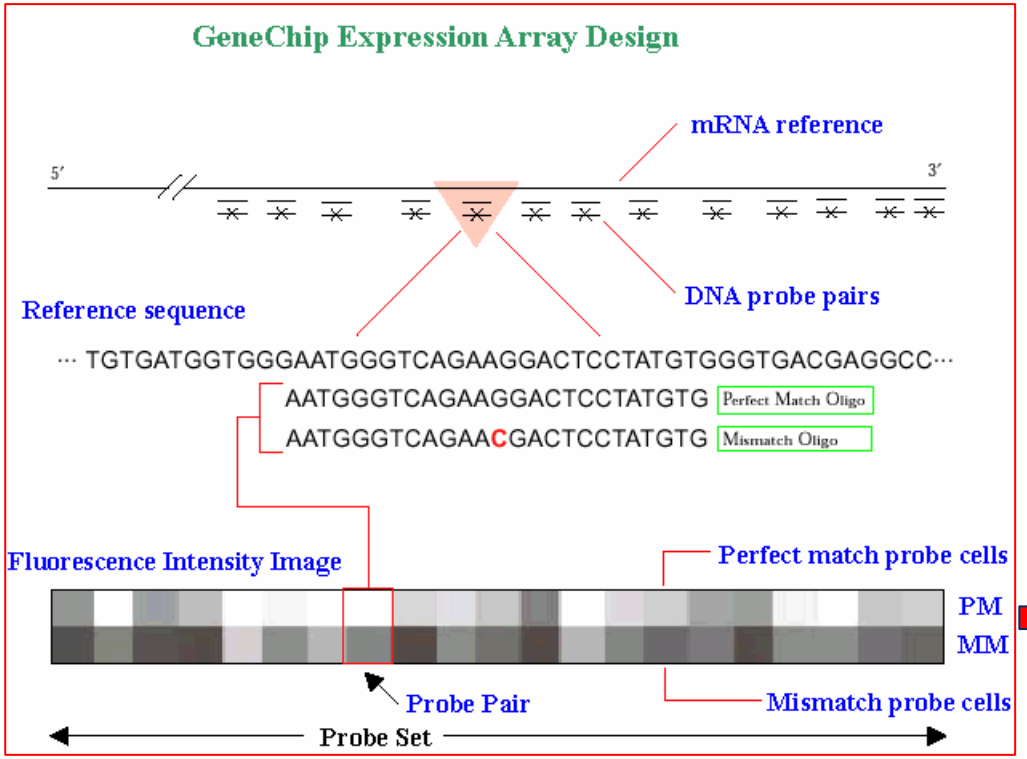
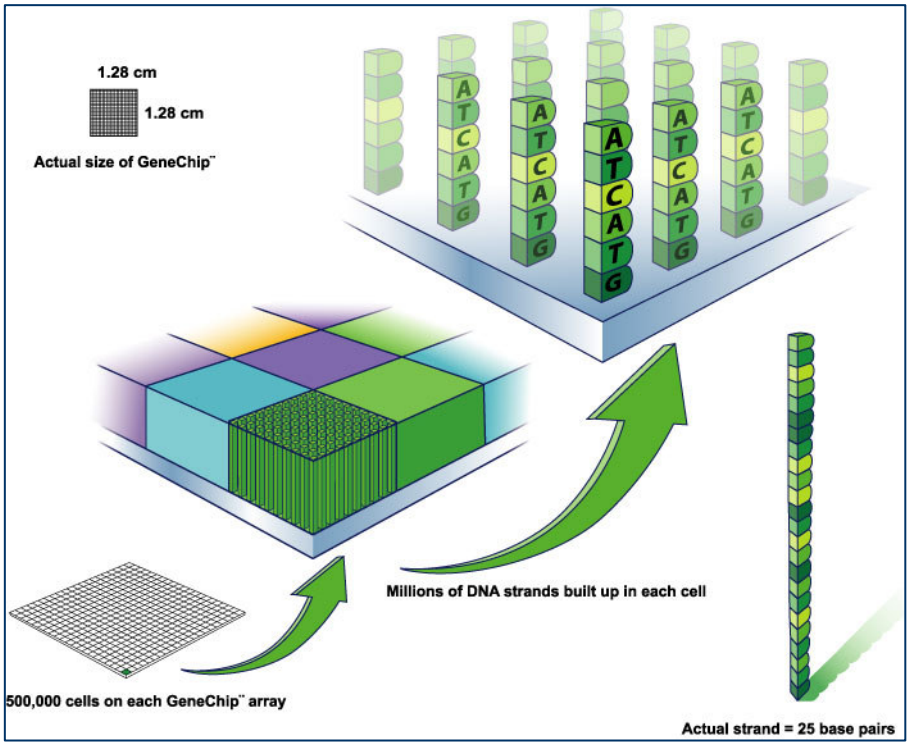


Image of hybridized probe array

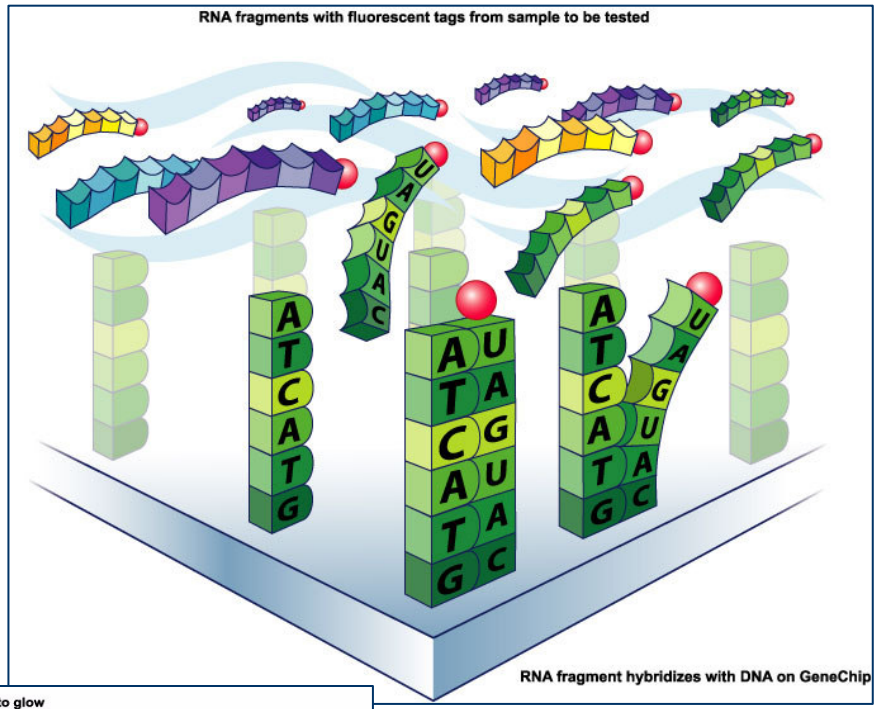




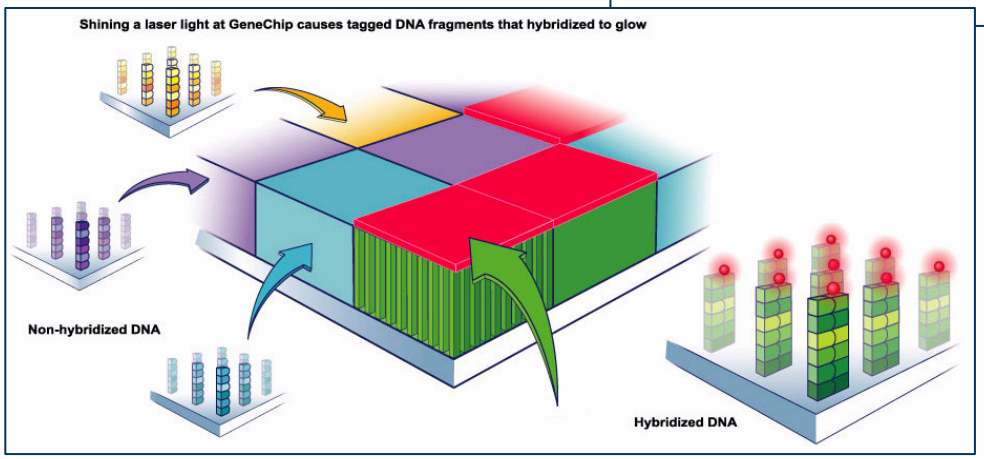
More Figures on Affymetrix Web Site



GeneChip® Hybridization



GeneChip® Single Feature



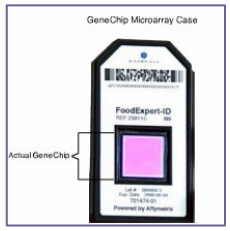
Hybridized GeneChip® Microarray

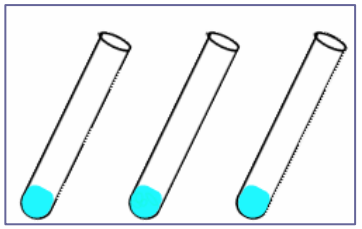
Animations

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"

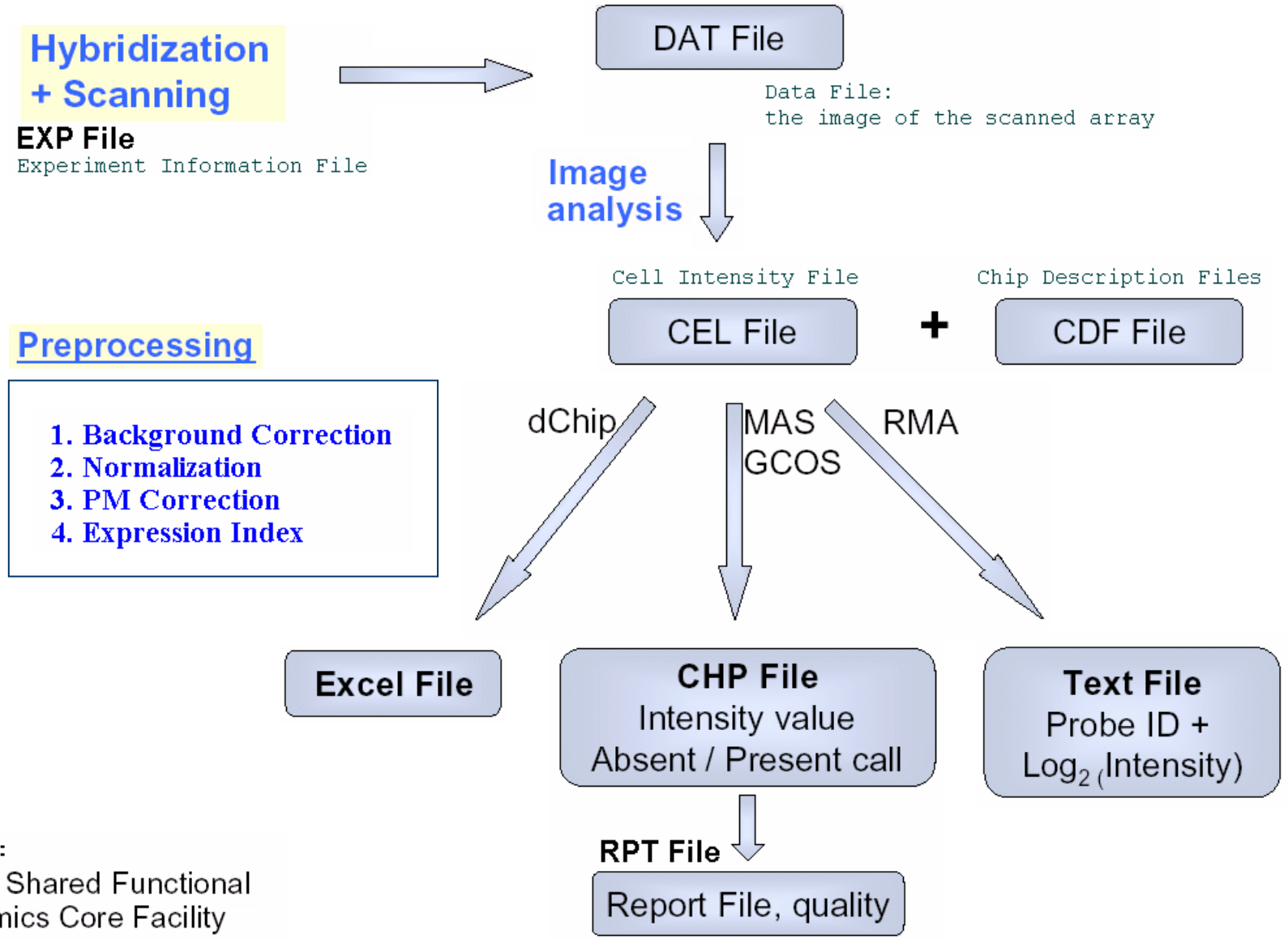
Digizyme - Web & Multimedia Design for the Sciences

http://www.digizyme.com/
http://www.digizyme.com/portfolio/microarraysfab/index.html
http://www.digizyme.com/competition/examples/genechip.swf

DNA Microarray Virtual Lab

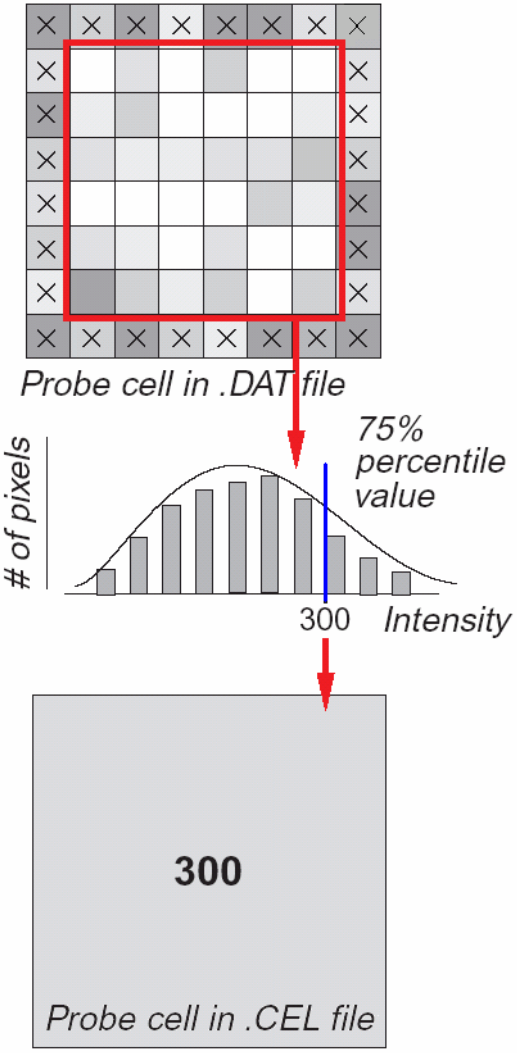
http://learn.genetics.utah.edu/units/biotech/microarray

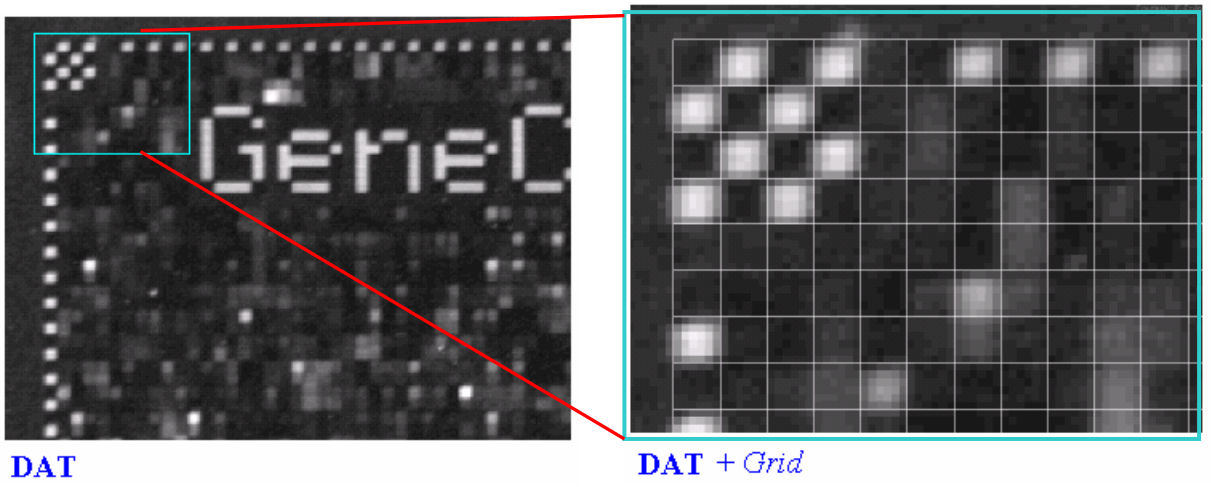
Assay and Analysis Flow Chart

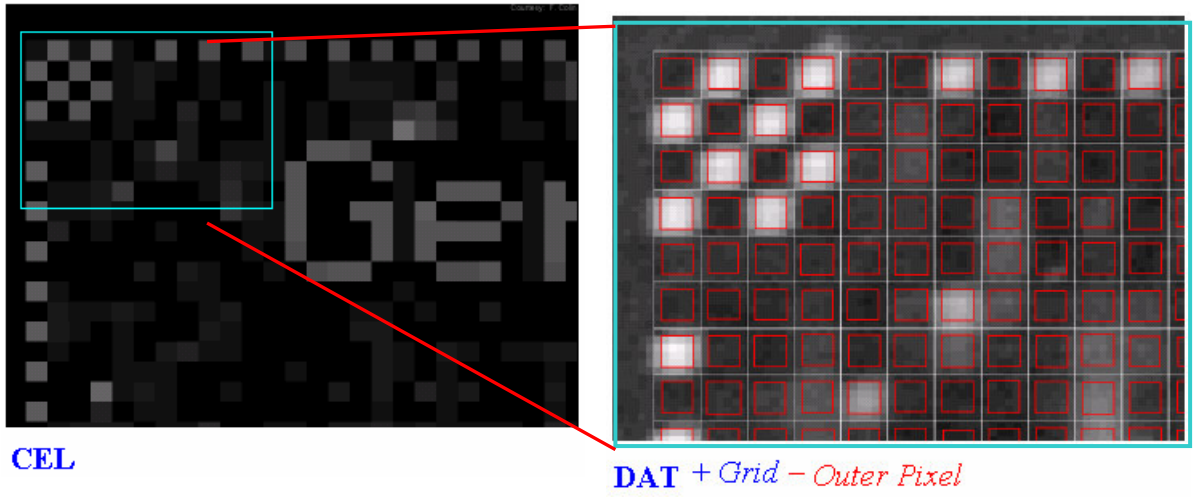


Source:
UCSF Shared Functional
Genomics Core Facility

From DAT to CEL







CDF file

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

```
[CDF]
Version=GC3.0

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

[QC1]
Type=15
NumberCells=2280
CellHeader=X    Y       PROBE
Cell1=6 0       N       25
Cell2=8 0       N       25
Cell3=10        0       N
Cell4=12        0       N
Cell5=14        0       N
Cell6=16        0       N
Cell7=18        0       N
Cell8=20        0       N
Cell9=22        0       N
Cell10=24       0       N
Cell11=26       0       N
Cell12=28       0       N
Cell13=30       0       N
Cell14=32       0       N
Cell15=34       0       N
Cell16=36       0       N
Cell17=38       0       N
Cell18=40       0       N
Cell19=42       0       N
Cell20=44       0       N
Cell21=46       0       N       25      0       46
Cell22=48       0       N       25      0       48
Cell23=50       0       N       25      0       50
Cell24=52       0       N       25      0       52
Cell25=54       0       N       25      0       54
Cell26=56       0       N       25      0       56
Cell27=58       0       N       25      0       58
Cell28=60       0       N       25      0       60
Cell29=62       0       N       25      0       62
Cell30=64       0       N       25      0       64
```

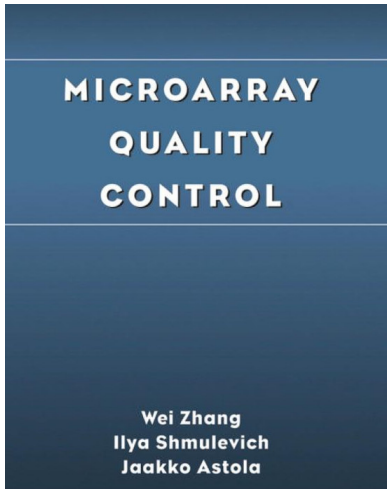
```
[Unit1003_Block1]
Name=121_at
BlockNumber=1
NumAtoms=16
NumCells=32
StartPosition=0
StopPosition=15
CellHeader=X    Y       PROBE   FEAT    QUAL    EXPOS   POS     CBASE   PBASE   TBASE   ATOM    INDEX   CODONIND        CODON   REGIONTYPE      REGION
Cell1=656       1012    N       control 121_at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

Quality Assessment

Two aspects of quality control: detecting poor hybridization and outliers.

- **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)**

- **Reasons for poor hybridizations**
  - mRNA degenerated
  - one or more experimental steps failed
  - poor chip quality, …
- **Reasons for (biological) outliers**
  - infiltration with non-tumor tissue
  - wrong label
  - contamination, …

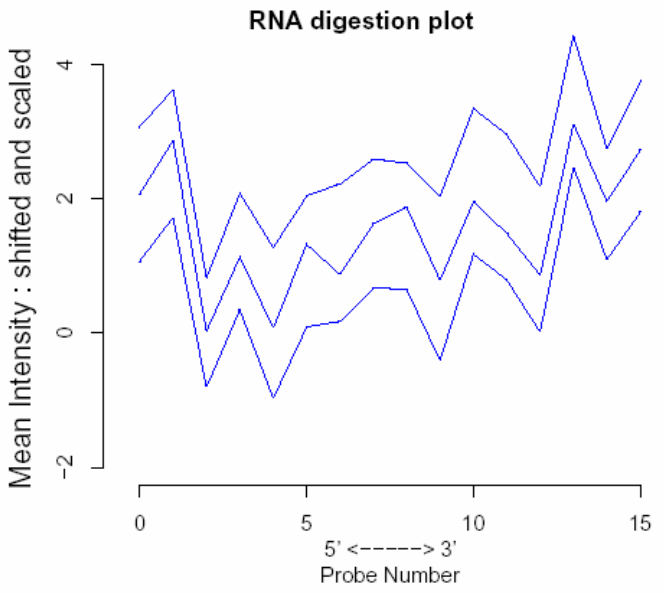


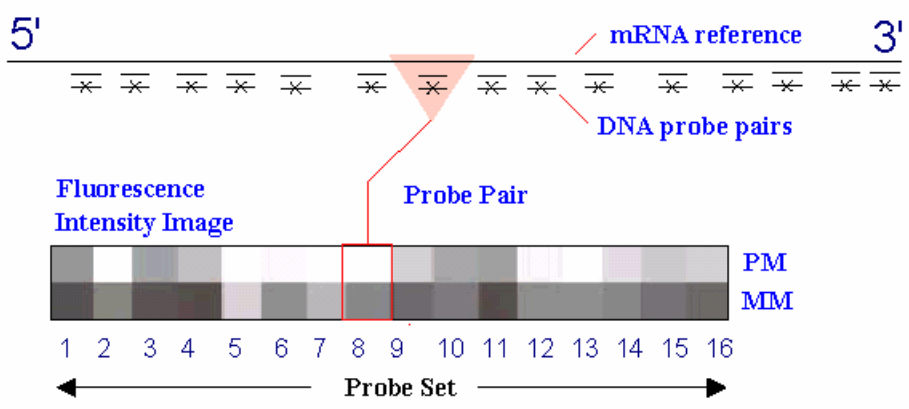
Quality Assessment

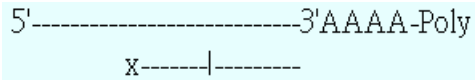
RNA Degradation Plots

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.



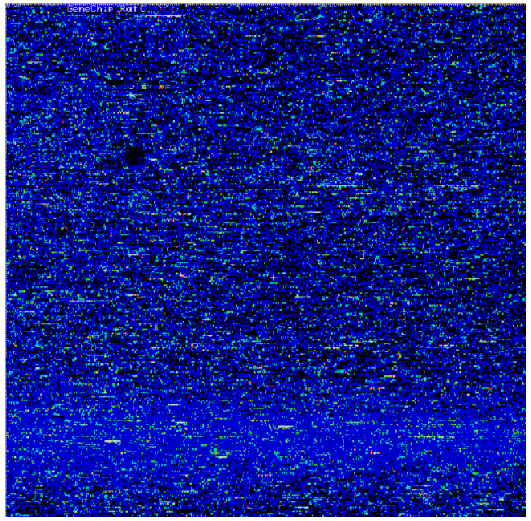




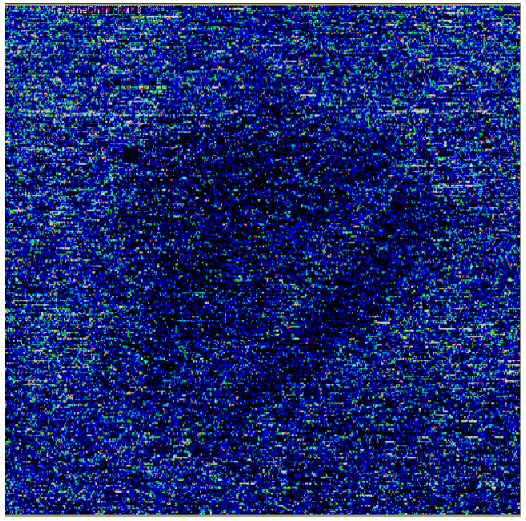
Probe Array Image Inspection

- 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, grid misalignment.
- As long as these areas do not represent more than 10% of the total probes for the chip, then the area can be masked and the data points thrown out as outliers.

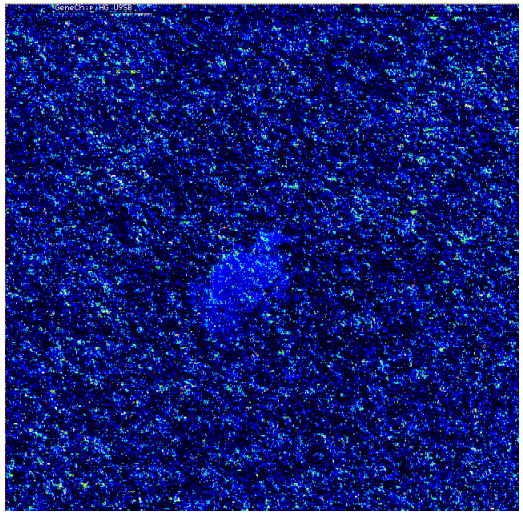
Haze Band



Crop Circles



Spots, Scratches, etc.



Source: Michael Elashoff (GLGC)

Probe Array Image Inspection (conti.)

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

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

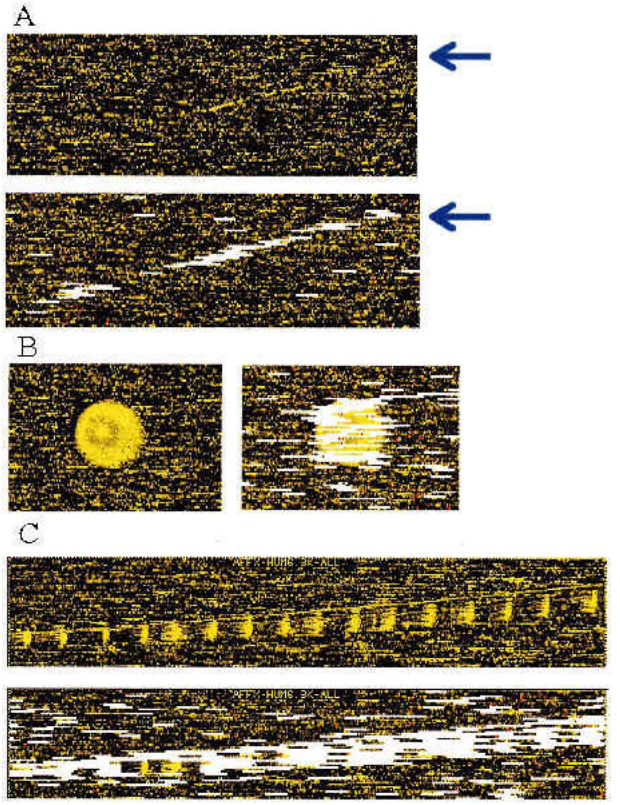
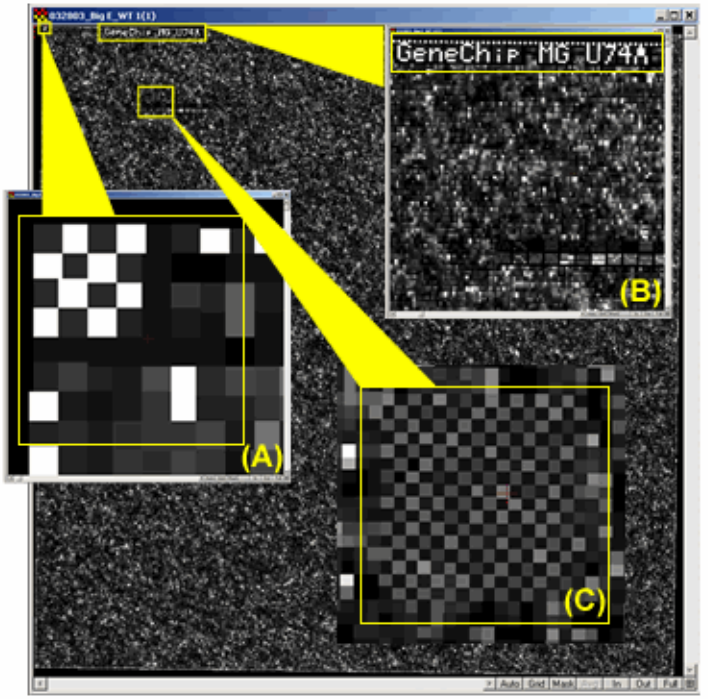


Fig. 5. (*A*) A long scratch contamination (indicated by arrow) is alleviated by automatic outlier exclusion along this scratch. (*B* and *C*) Regional clustering of array outliers (white bars) indicates contaminated regions in the original images. These outliers are automatically detected and accommodated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

B2 Oligo Performance

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



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

Source: Baylor College of Medicine, Microarray Core Facility

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:

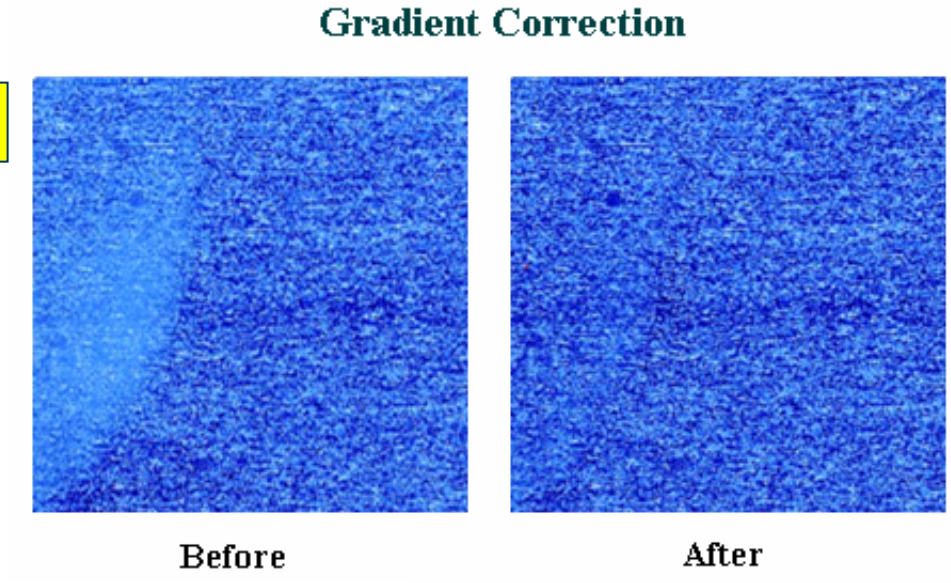
Probe Set	Sig(5')	Det(5')	Sig(M')	Det(M')	Sig(3')	Det(3')	Sig(all)	Sig(3'/5')
AFFX-BIOB	559.0	P	801.6	P	385.8	P	582.14	0.69
AFFX-BIOC	1132.9	P			818.0	P	975.47	0.72
AFFX-BIOD	874.7	P			6918.1	P	3896.42	7.91
AFFX-CRE	10070.5	P			16198.0	P	13134.27	1.61
AFFX-DAP	10.9	A	60.9	A	8.5	A	26.75	0.78
AFFX-LYS	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

Suggestions

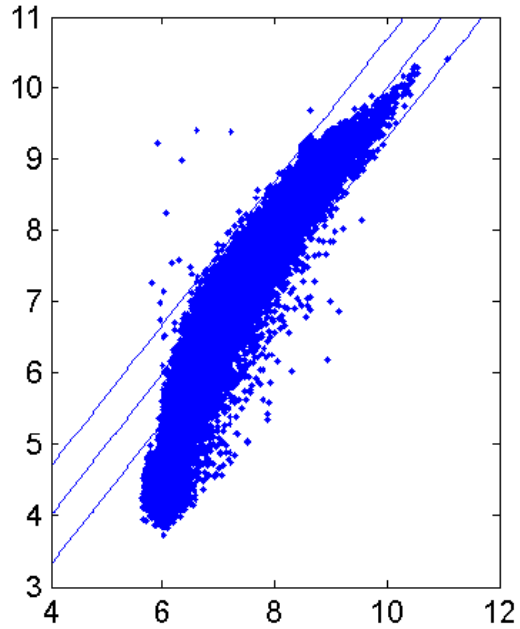
- Affymetrix arrays with high **background** are more likely to be of poor quality.
  - Cutoff would be to exclude arrays with a value more than 100.
- **Raw noise score (Q)**: a measure of the variability of the pixel values within a probe cell averaged over all of the probe cells on an array.
  - Exclude those arrays that have an unusually high Q-value relative to other arrays that were processed with the same scanner.
- **BioB**: is included at a concentration that is close to the level of detection of the array, and so should be indicated as present about 50% of the time.
- Other spike controls are included at increasingly greater levels of concentration. Therefore, they should all be indicated as present, and also should have increasingly large signal values:
  - Signal(bioB) < Signal(bioC) < Signal(bioD) < Signal(cre)

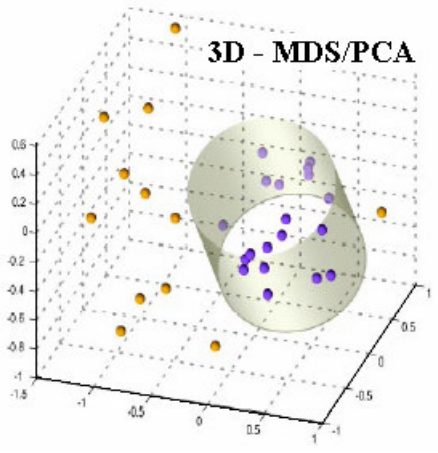
Statistical Plots

GeneChipImage



Scatterplot





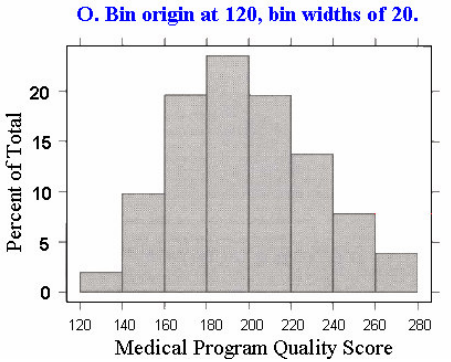
Dimension Reduction
(PCA, MDS)

Statistical Plots: Histogram

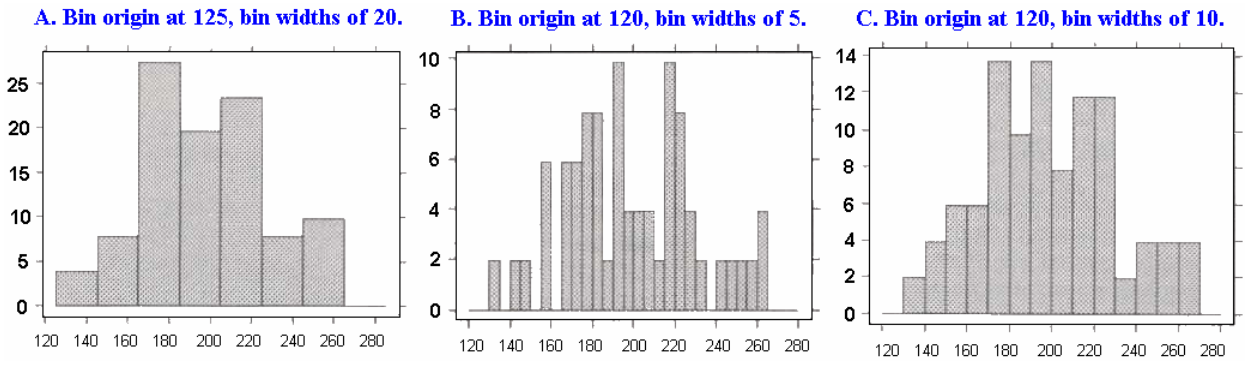
- 1/2h adjusts the height of each bar so that the total area enclosed by the entire histogram is 1.
- The area covered by each bar can be interpreted as the probability of an observation falling within that bar.

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.



Density Plots



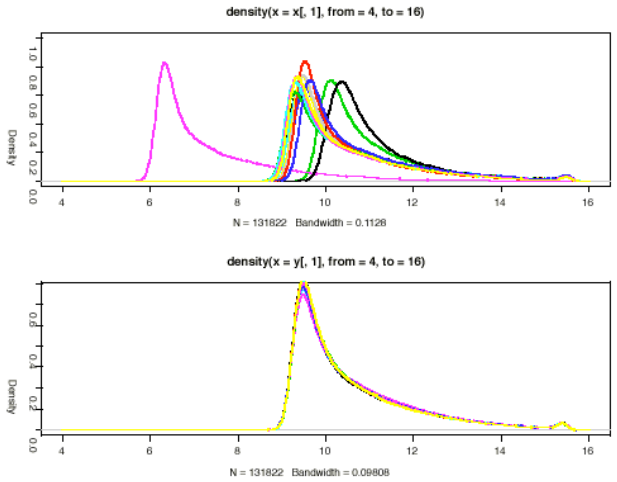
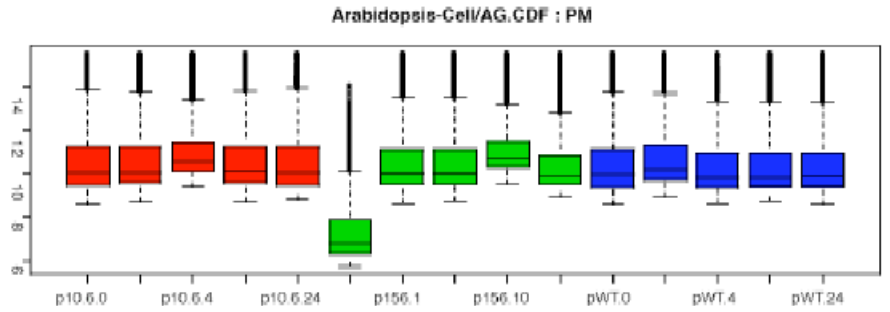
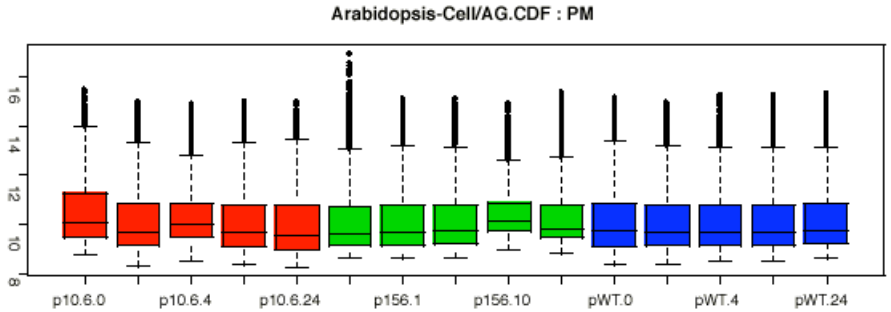


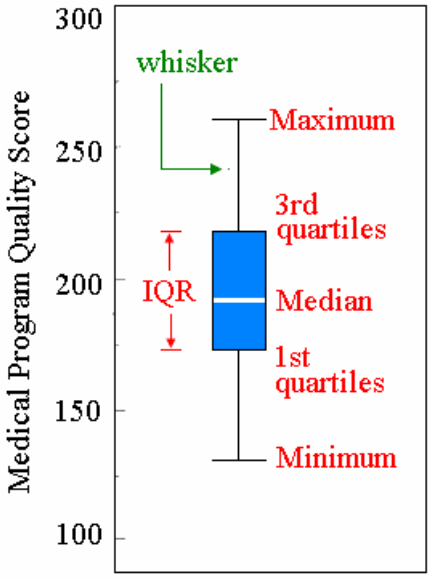
Figure Sources: Jacoby (1997).

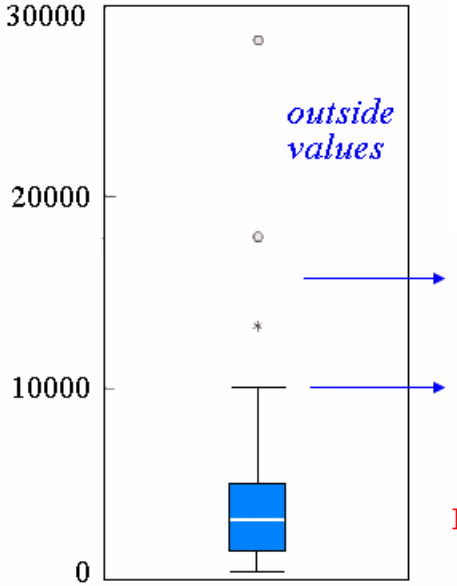
Statistical Plots: Box Plots

- 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: $x_{0.75} + 3\ IQR$

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

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

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

The box plot can provide answers to the following questions:

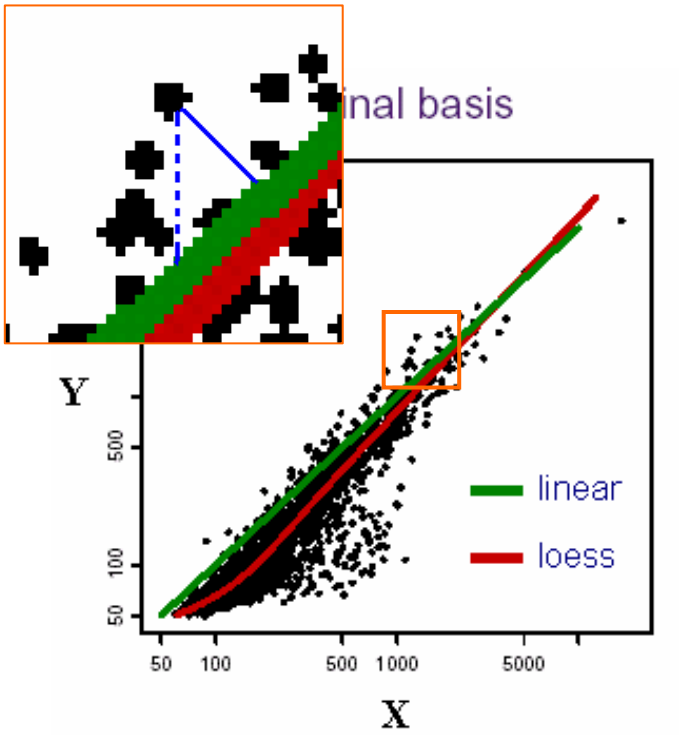
- Is a factor significant?
- Does the location differ between subgroups?
- Does the variation differ between subgroups?
- Are there any outliers?

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

Scatterplot

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

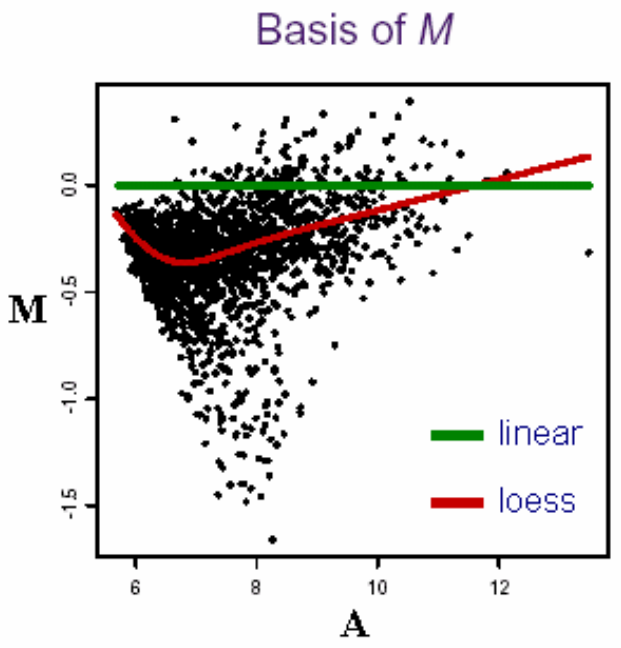
- **Goals:** to identify genes that are differentially regulated between two experimental conditions.



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

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

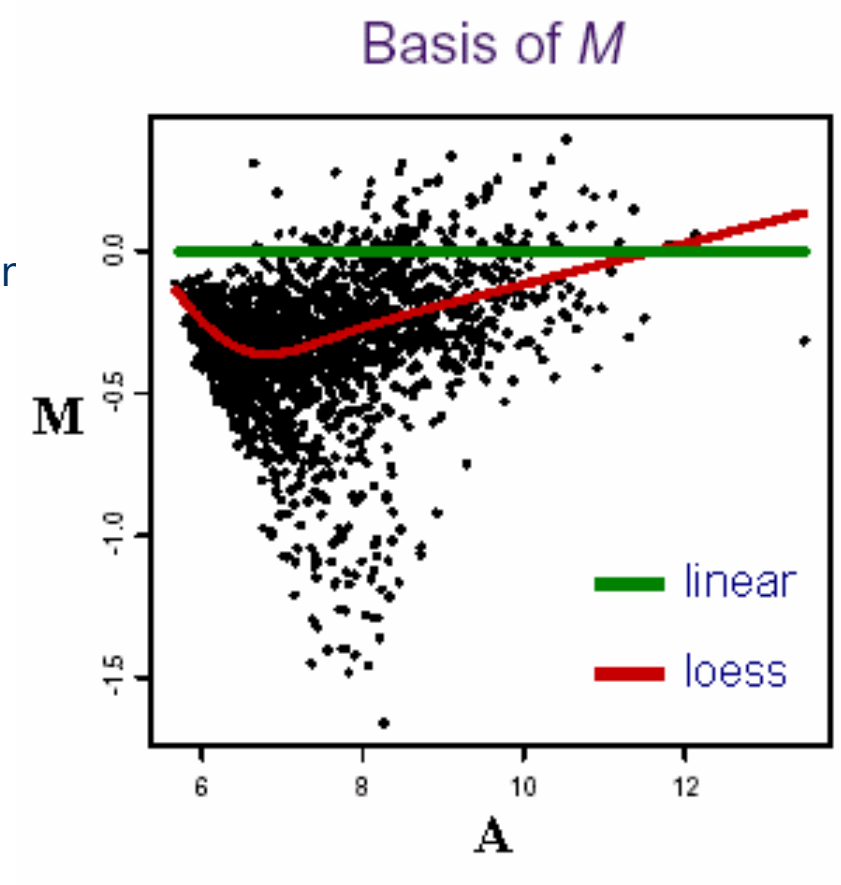
Oligo	cDNA
$X = PM_1$, $Y = PM_2$	X= Cy3, Y= Cy5
$X = PM_1\text{-}MM_1$, $Y = PM_2\text{-}MM_2$	



MA plot

- **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 distributior in which the prosperities of the data are easy to analyze.
  - easier to describe the fold regulation of genes using a log scale.
  - In log2 space, the data points are symmetric about 0.



MAQC project

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

QC Reference

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

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

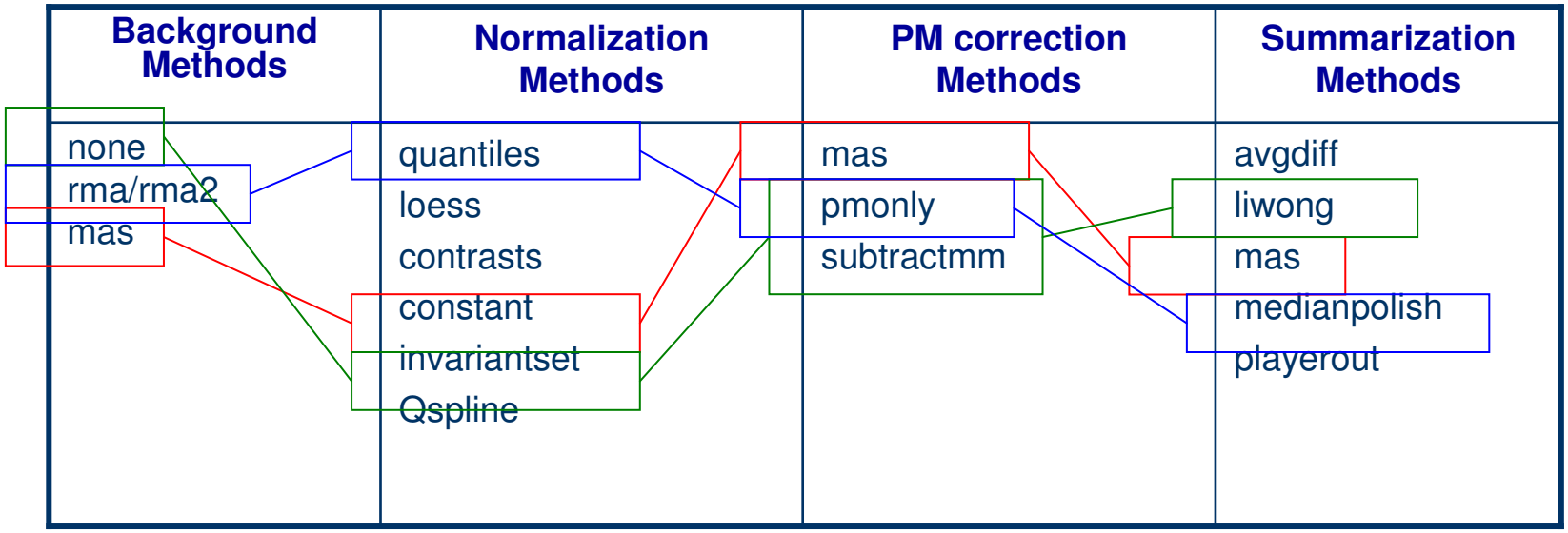
- Claire L. Wilson and Crispin J. Miller, 2005, **Simpleaffy**: a BioConductor package for Affymetrix Quality Control and data analysis, Bioinformatics 21: 3683 - 3685.

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

Red color: R package at Bioconductor.

Low-level Analysis

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")
```

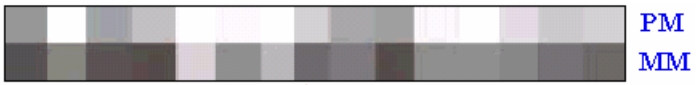
Background Correction

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 explicitely accounted for
- noise

Systematic

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

Stochastic

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

Why Normalization?

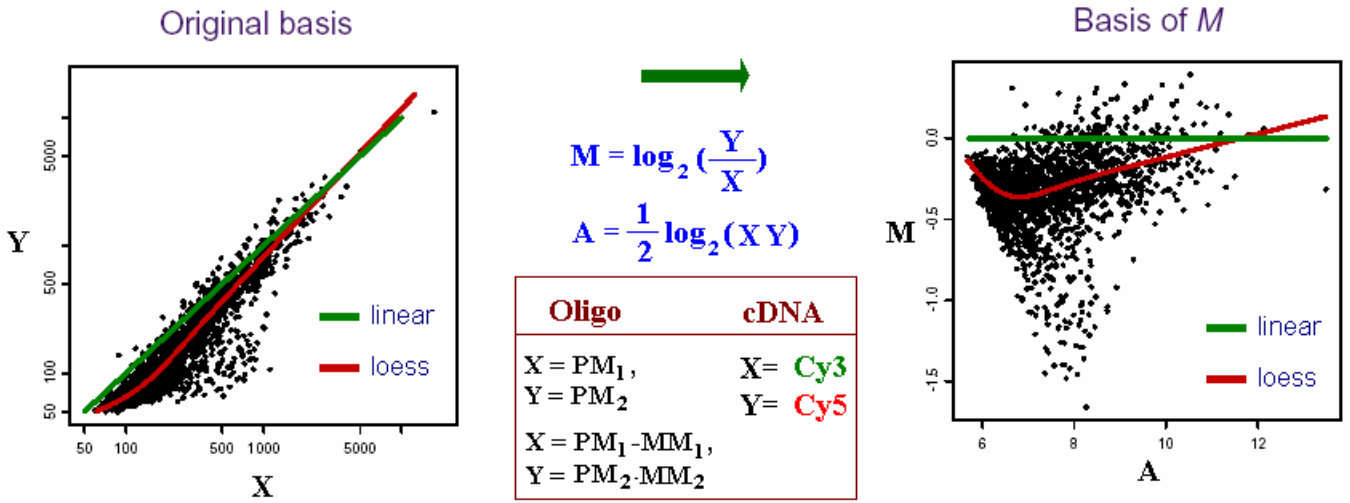
Normalization corrects for overall chip brightness and other factors that may influence the numerical value of expression intensity, enabling the user to more confidently compare gene expression estimates between samples.

Main idea

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

Assumption

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



Constant Normalization

Normalization and Scaling

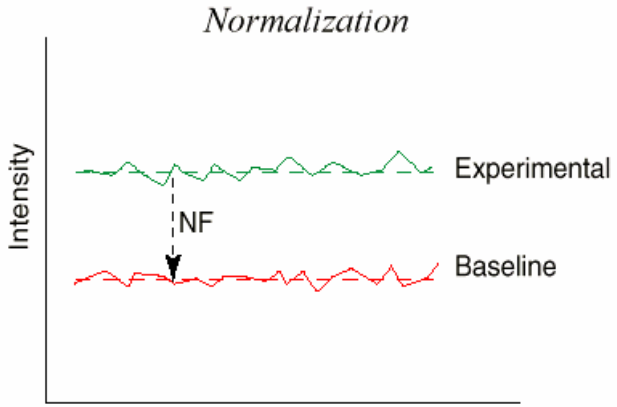
- The data can be normalized from:
  - a limited group of probe sets.
  - all probe sets.

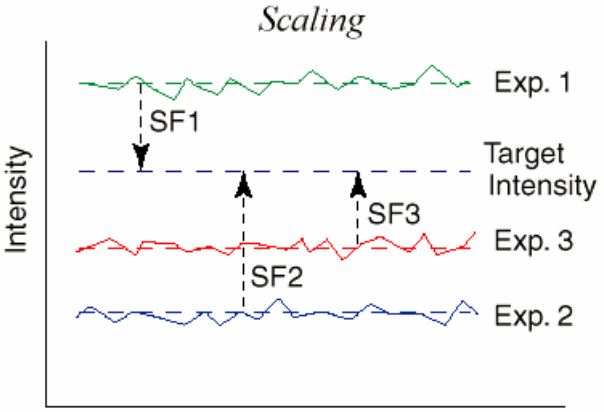
- **Global Scaling**
  the average intensities of all the arrays that are going to be compared are multiplied by scaling factors so that all average intensities are made to be numerically equivalent to a preset amount
  (termed target intensity).

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

$$A \times SF = TGT$$

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





- **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.

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

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

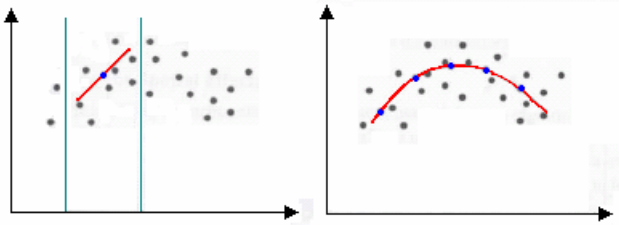
$$nf = \frac{TrimMean\left(SPVb_{i,}\ 0.02, 0.98\right)}{TrimMean\left(SPVe_{i,}\ 0.02, 0.98\right)}$$

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

LOESS Normalization

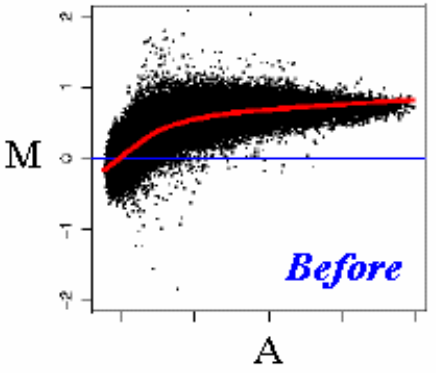
- Loess normalization (Bolstad *et al.*, 2003) is based on **MA plots**. Two arrays are normalized by using a lowess smoother.
- **Skewing** reflects experimental artifacts such as the
  - contamination of one RNA source with genomic DNA or rRNA,
  - the use of unequal amounts of radioactive or fluorescent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

Loess regression
(locally weighted polynomial regression)



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

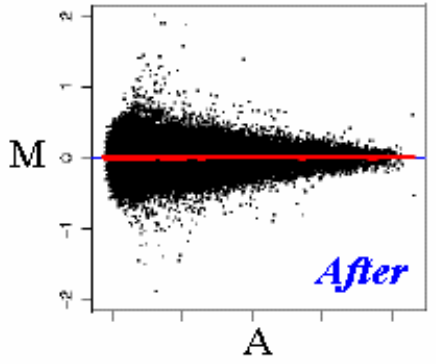
   Loess is a method of local regression (see Cleveland and Devlin (1988) for details).
4. The fits based on the normalization curve are $\hat{M}_k$
5. the normalization adjustment is $M'_k = M_k - \hat{M}_k$.
6. Adjusted probe intensites are given by $x'_{ki} = 2^{A_k + \frac{M'_K}{2}}$ and $x'_{ki} =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



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

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

Oligo	cDNA
$X = PM_1$, $Y = PM_2$	X= Cy3, Y= Cy5
$X = PM_1\text{-}MM_1$, $Y = PM_2\text{-}MM_2$	



PM Correction Methods

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

Table 1: Summary Table

Method	Assumptions	Benefits	Drawbacks
PM-MM	Background effects are large and potentially variable between features across experiments relative to effects of interest	Background effects minimized due to low bias 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, gc-RMA**: a robust multi-chip linear model fit on the log scale.

RMA

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.

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

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

$$b = \sigma$$

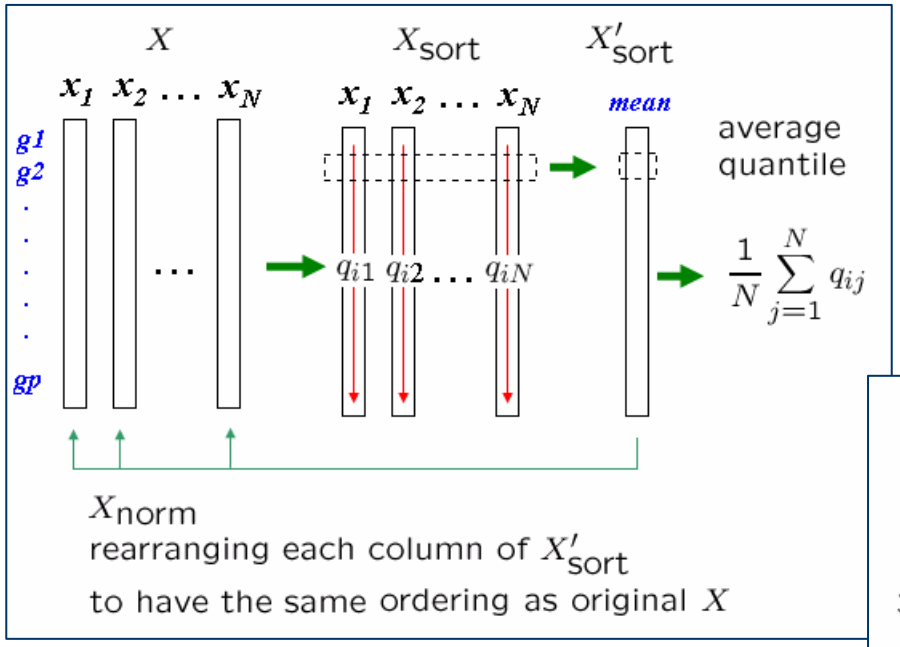
ϕ: standard normal density function

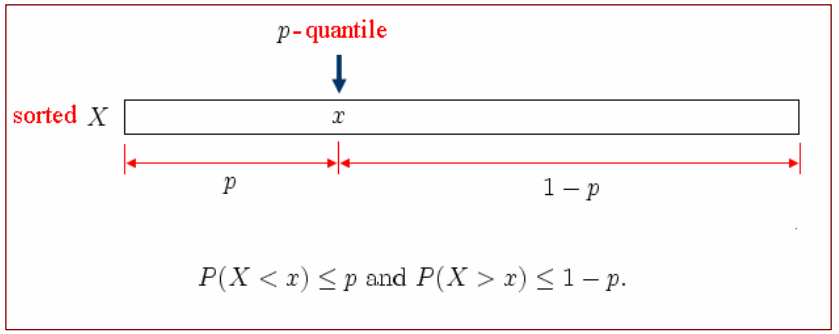
Φ: standard normal distribution function

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.





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

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

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

RMA: Summarization Method

MedianPolish

- This is the summarization used in the RMA expression summary Irizarry et al. (2003).
- A **multichip linear model** is fit to data from each probeset.
- The median polish 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

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

$i = 1, \ldots, I_k$ probes

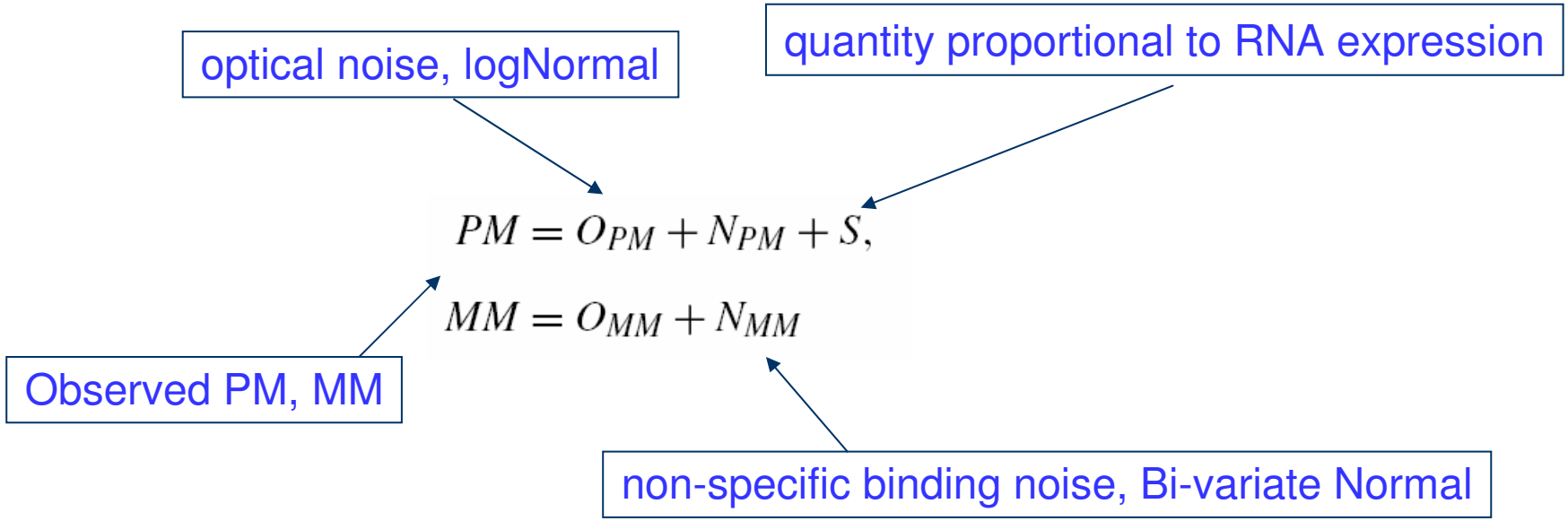
$j = 1, \ldots, J$ arrays

probe effect

$\log_2$ expression value

GC-RMA

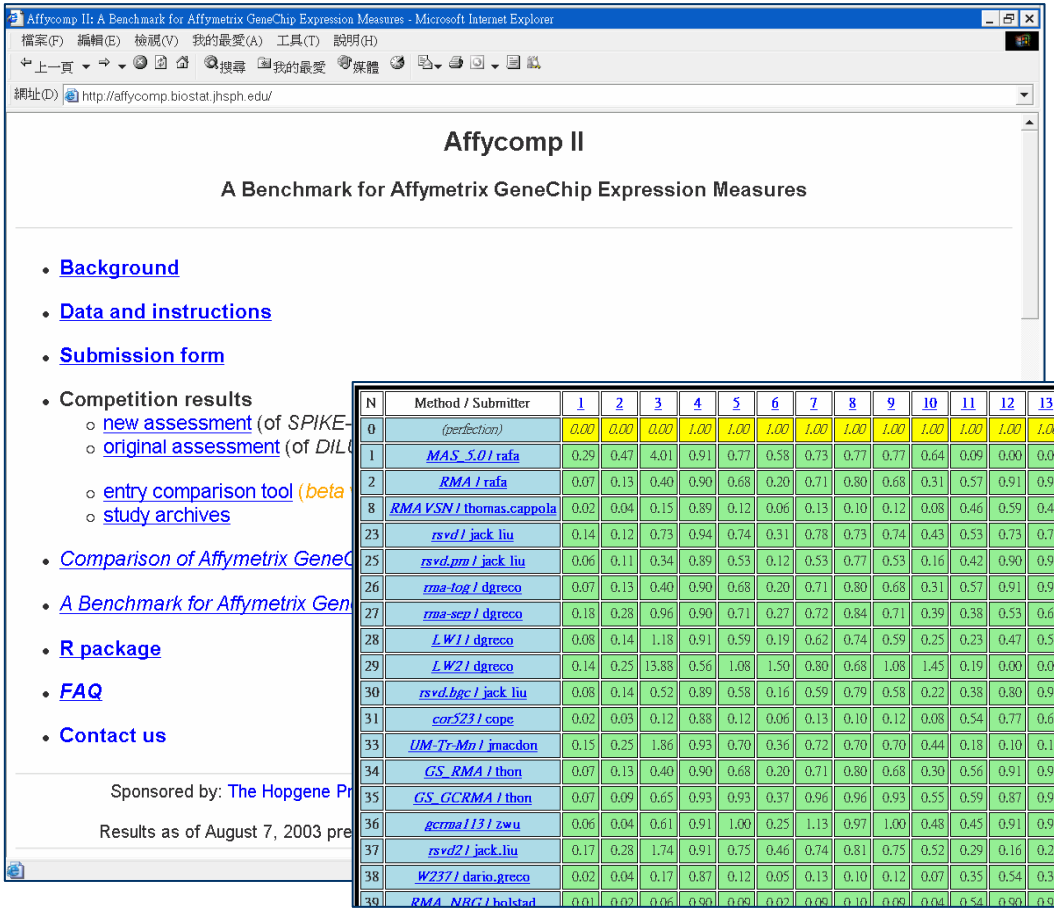
- Robust multi-chip average with GC-content (guanine-cytosine content) background correction
- Background correction: account for background noise as well as non-specific binding.



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

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

Comparison of Affymetrix GeneChip Expression Measures



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

Data and instructions

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

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

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

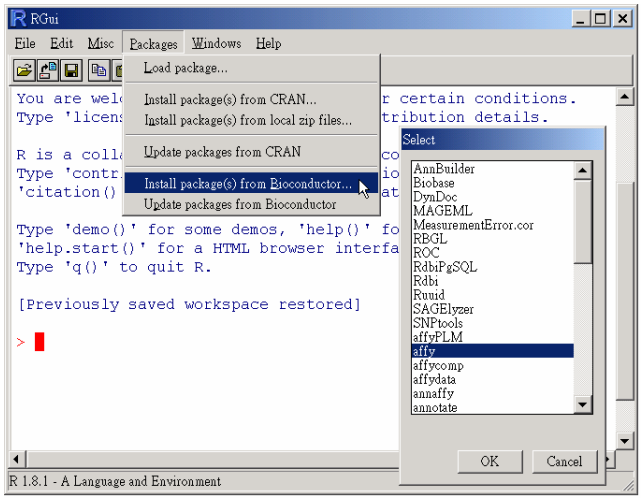
- Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP. A benchmark for Affymetrix GeneChip expression measures, Bioinformatics. 2004 Feb 12;20(3):323-31.
- Irizarry RA, Wu Z, Jaffee HA. Comparison of Affymetrix GeneChip expression measures. Bioinformatics. 2006 Apr 1;22(7):789-94.

Software for Image Analysis and Normalization

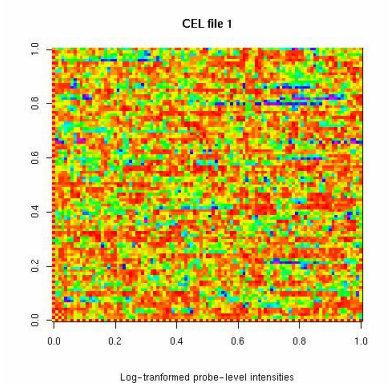
The Bioconductor: affy

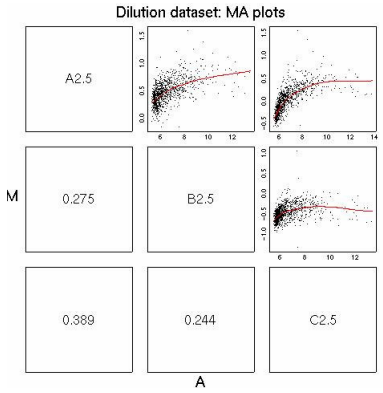
The Bioconductor Project
Release 1.7
http://www.bioconductor.org/

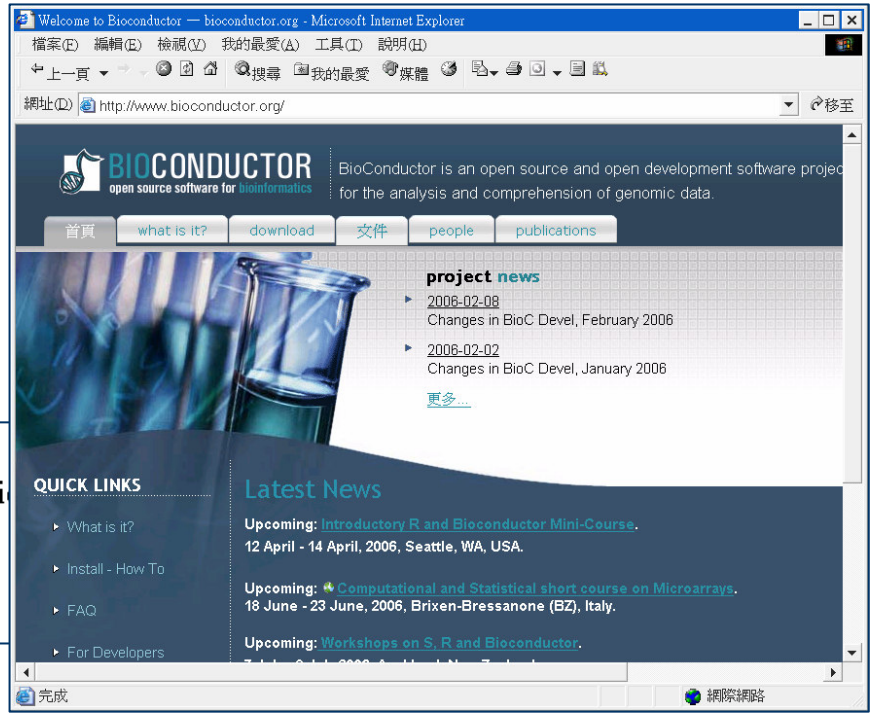
The R Project for Statistical Computing



affypdnn
affyPLM
gcrma
makecdfenv







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

The Bioconductor: affy

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

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

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

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

```
expresso(
    afbatch,

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

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

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

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

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

Browse the Packages by Task Views

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

Bioconductor Task View: BiocViews

Subviews

- Softwa
- Annota
- Experi

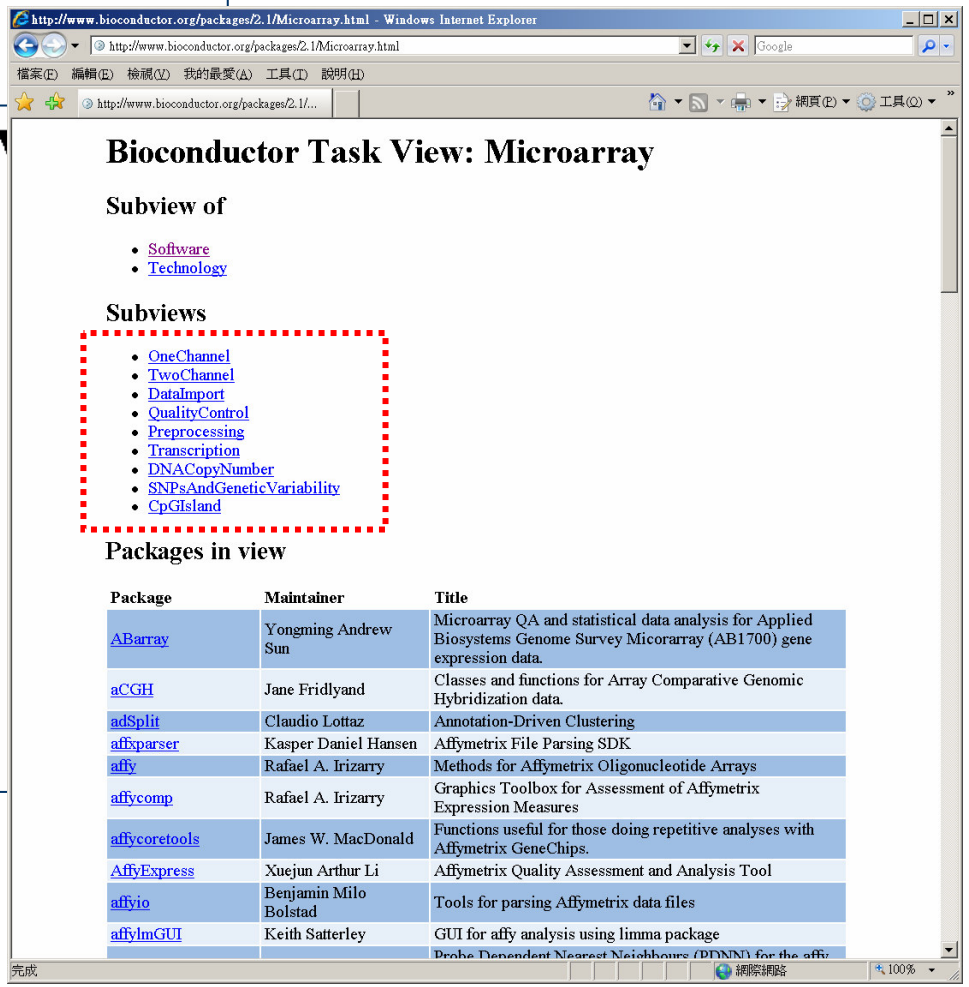
Bioconductor Task V

Subview of

- BiocViews

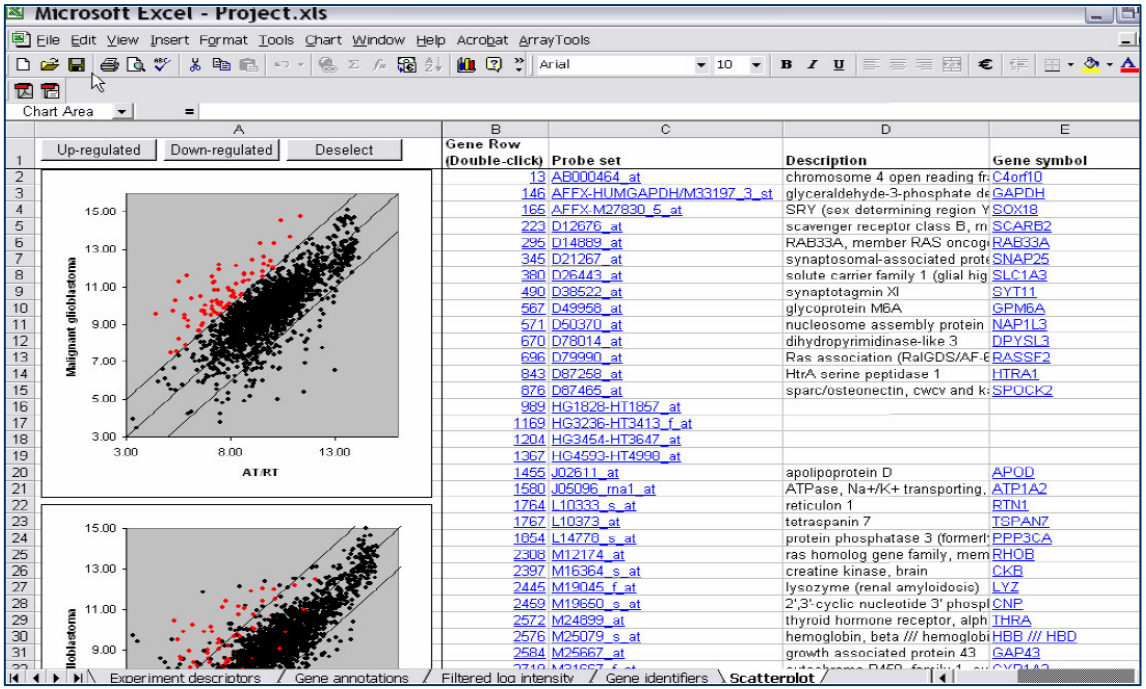
Subviews

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



BRB-ArrayTools

An Integrated Software Tool for DNA Microarray Analysis



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

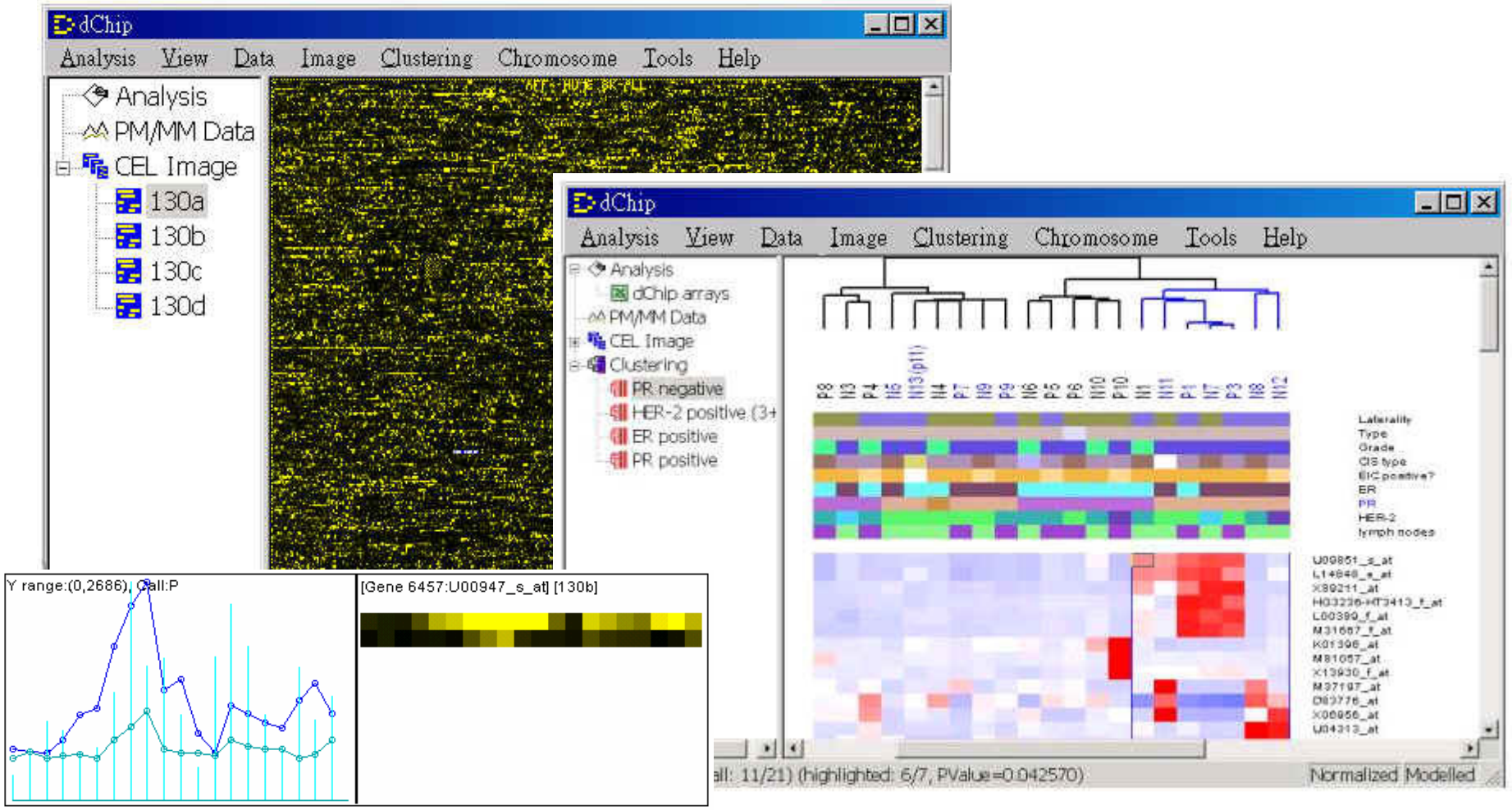
Requirement:

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

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

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

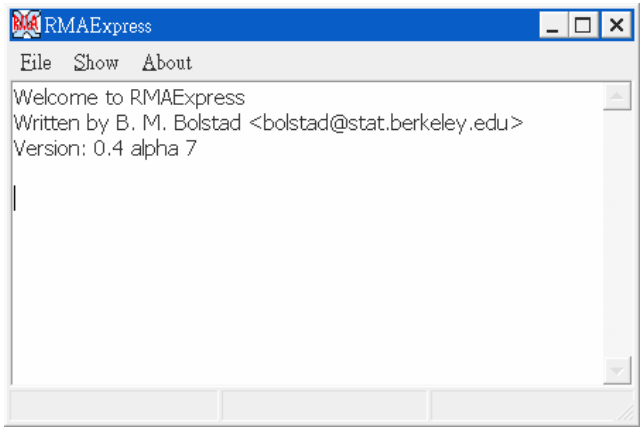
DNA-Chip Analyzer (dChip2006)

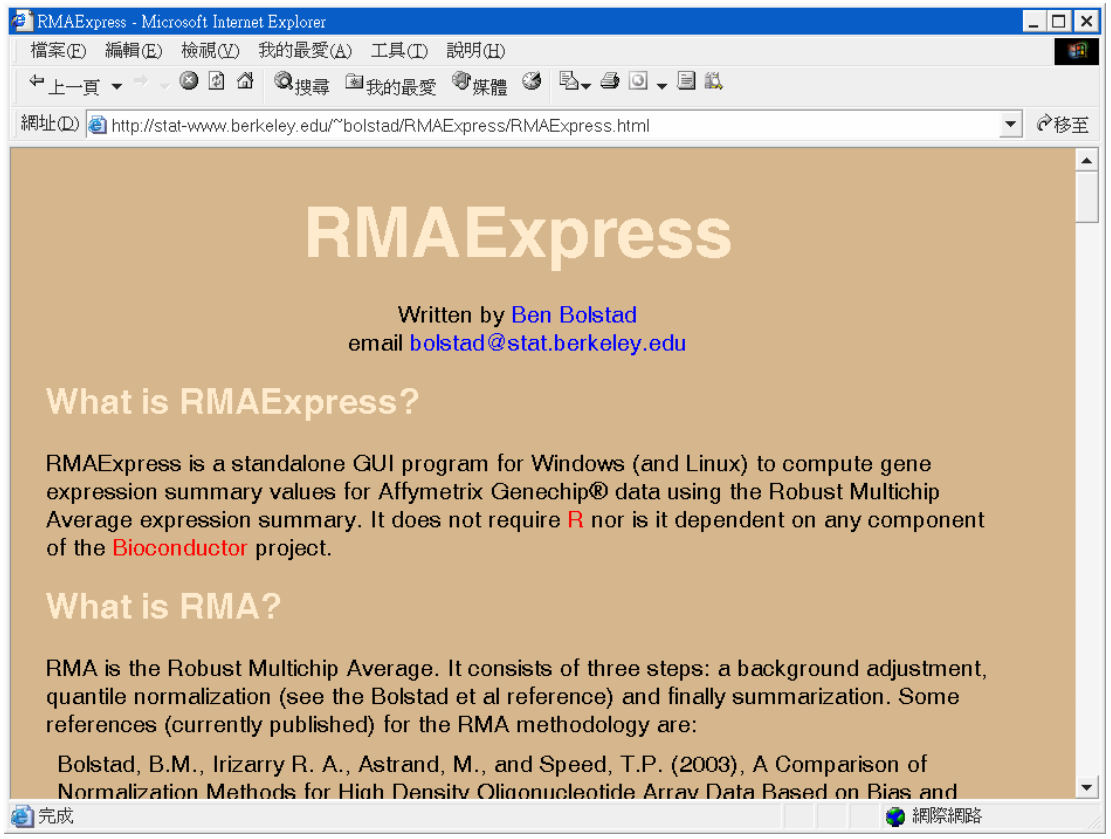


http://www.biostat.harvard.edu/complab/dchip/

RMAExpress

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





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

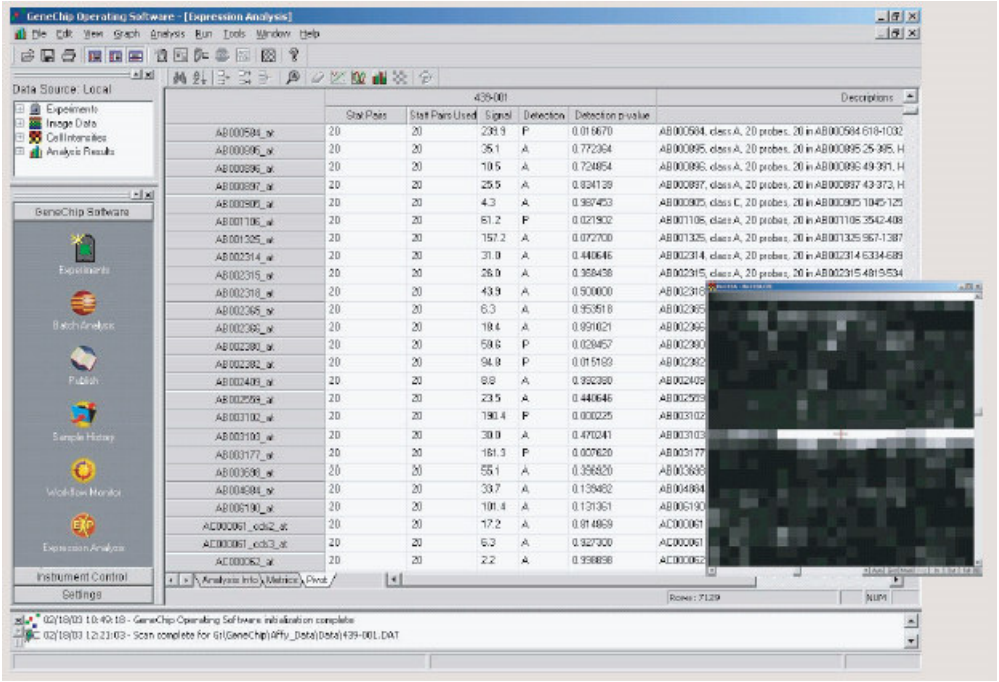
Affymetrix GeneChip Operating Software

AFFYMETRIX

http://www.affymetrix.com

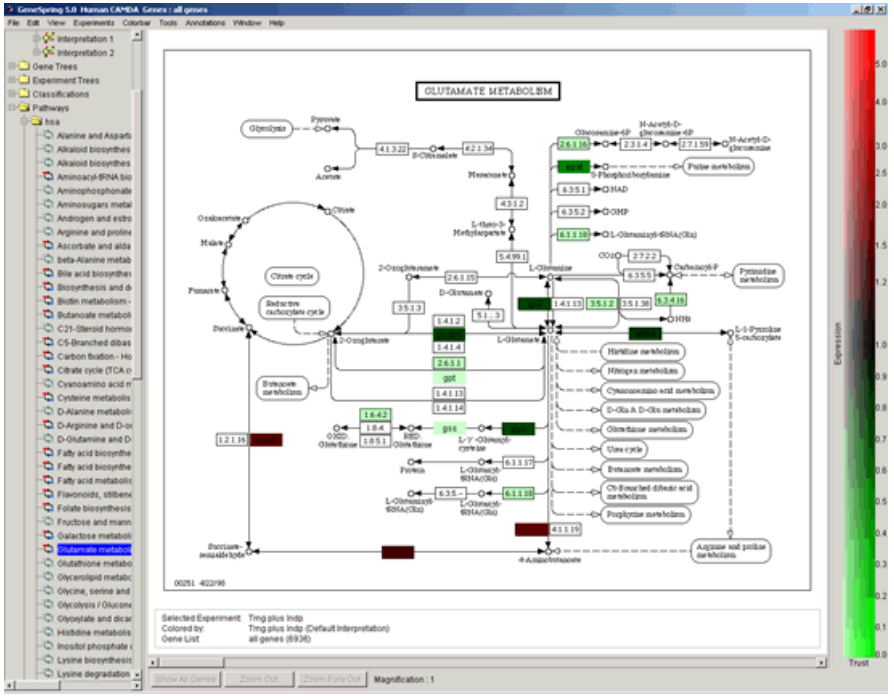
Specifications

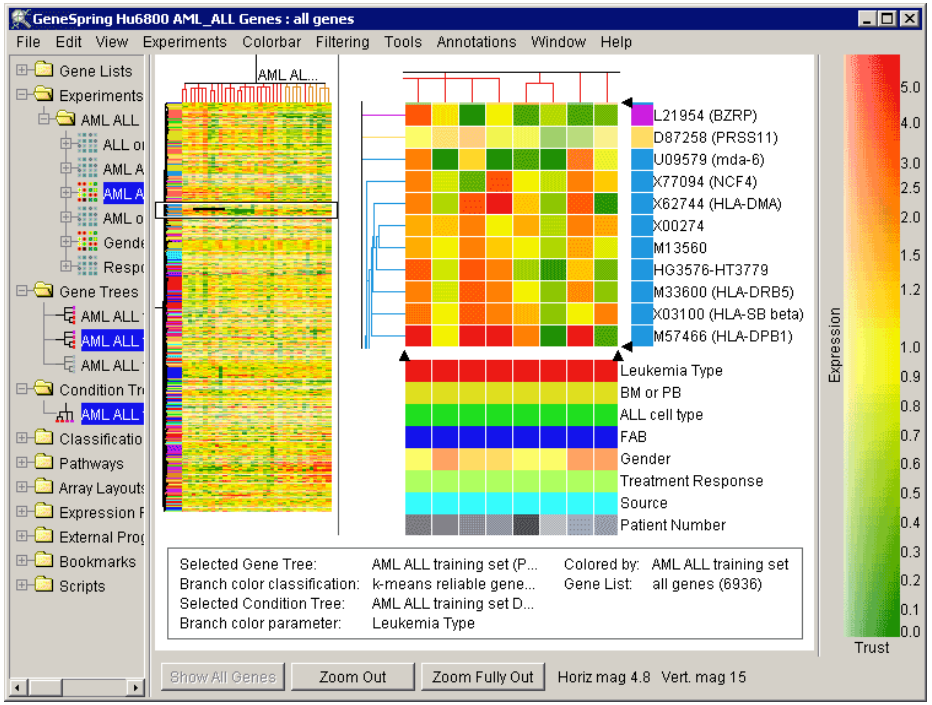
Instrument Support	• Affymetrix GeneChip® Fluidics Station 400 & 450 • GeneChip Scanner 3000 • GeneArray 2500 Scanner
Affymetrix Software Compatibility	• Support GeneChip DNA Analysis Software (GDAS) for mapping and resequencing data analysis • Support Affymetrix® Data Mining Tool software for statistical 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



GeneSpring GX v7.3.1

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





Images from
http://www.silicongenetics.com

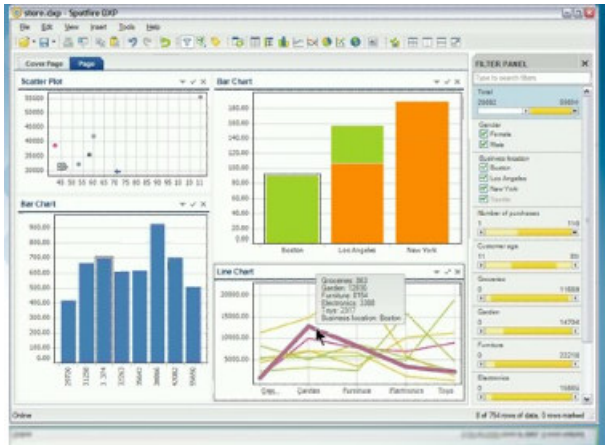
Agilent Technologies

2004 Articles Citing GeneSpring®

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

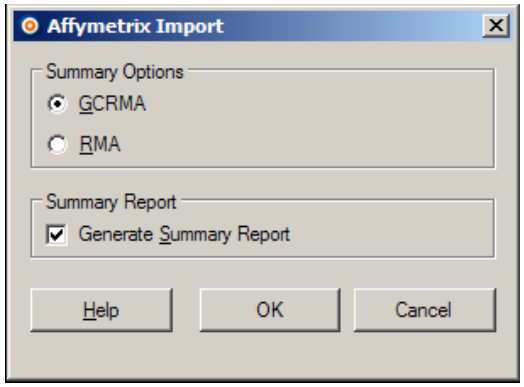
More than 700 papers

TIBCO® Spotfire® DecisionSite® 9.1 for Microarray Analysis



http://spotfire.tibco.com/

Affymetrix CEL File Import Summarization Dialog



Data Transformation

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

- Column normalization
- Row summation

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

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

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

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

Gene Filtering and Missing Values Imputation

MSA5: Detection Calls

- **Answers**: "Is the transcript of a particular gene Present or Absent?"
- **Absent** means that the expression level is below the threshold of detection. That is, the expression level is not provably different from zero.
- **Advantage**: easy to filter and easy to interpret: we may only want to look at genes whose transcripts are detectable in a particular experiment.

	030606 En test3	
	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

Analysis Info \ Metrics \ Pivot

Saturation

If a mismatch cell is saturated $MM \geq 46000$, the corresponding probe pair is not used in further computations. We also discard pairs where PM and MM are within *tau* of each other.

Method

There are four steps to the method:

1. Remove saturated probe pairs and ignore probe pairs wherein PM ~ MM + *tau*
2. Calculate the discrimination scores. (This tells us how different the PM and MM cells are.)
3. Use Wilcoxon's rank test to calculate a significance or *p*-value. (This tells us how confident we can be about a certain result.)
4. Compare the *p*-value with our preset significance levels to make the call.

MSA5: Detection Calls

Discrimination Score

$$R_i = \frac{PM_i - MM_i}{PM_i + MM_i}$$

The null hypothesis is that the target is absent (zero effect on the probes).

Computing *p*-values:

The one-sided Wilcoxon's Signed Rank Test

H_0: median$(R_i - \tau) = 0$

H_1: median$(R_i - \tau) > 0$

Making the call

We set two significance levels α_1 and α_2 such that $0 < \alpha_1 < \alpha_2 < 0.5$

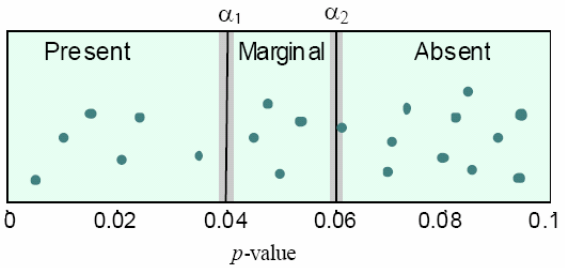
default $\alpha_1 = 0.04$ (16-20 probe pairs)

default $\alpha_2 = 0.06$ (16-20 probe pairs)

Method

There are four steps to the method:

1. Remove saturated probe pairs and ignore probe pairs wherein PM ~ MM + *tau*
2. Calculate the discrimination scores. (This tells us how different the PM and MM cells are.)
3. Use Wilcoxon's rank test to calculate a significance or *p*-value. (This tells us how confident we can be about a certain result.)
4. Compare the *p*-value with our preset significance levels to make the call.



Significance levels α_1 and α_2 define cut-offs of *p*-values for making calls.

dChip: Filter Genes

1. A < SD/mean < B
A < SD (for logged data) < B

A gene is variable enough compared to its mean expression level to contain interesting information (> A), but not so variable that nothing can be learned (< B).

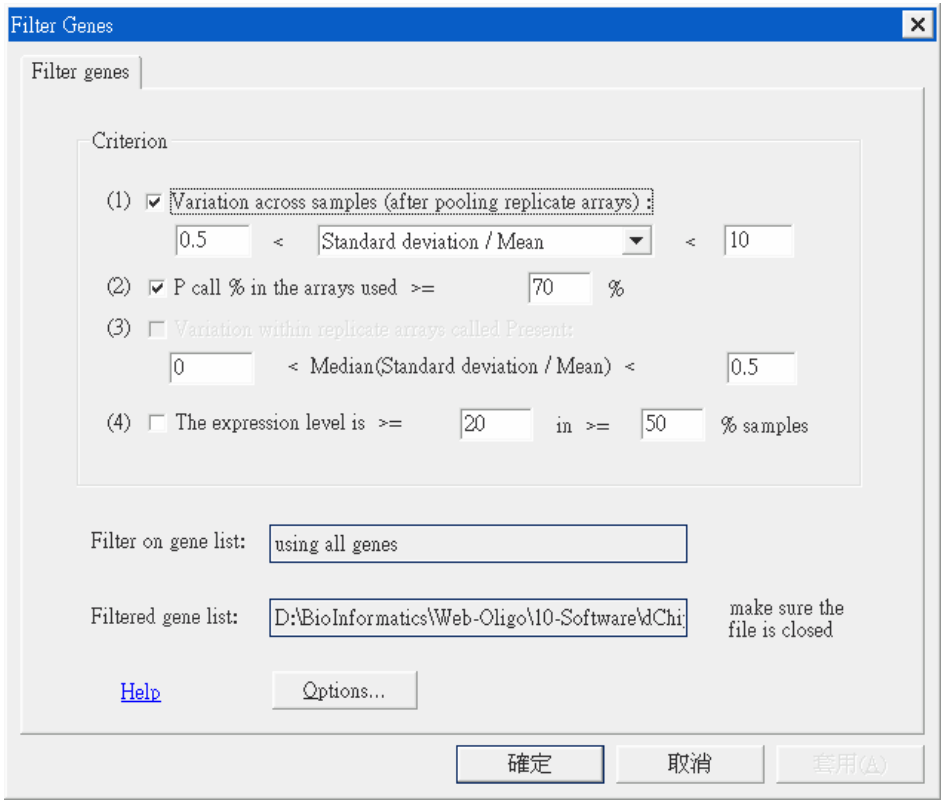
2. Presence call > X%

Narrows genes with a positive presence call in a certain percentage (> X%) of the samples.

3. A < Median(SD/Mean) < B

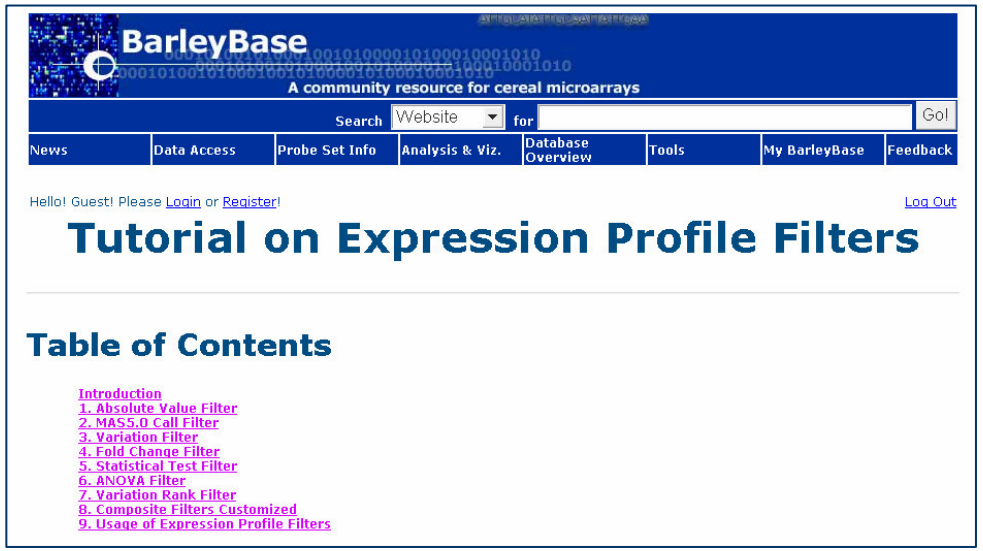
4. Expression level > Y in X%

Since low expression estimates are sometimes unreliable, we may want to limit our analysis to genes that are expressed above some threshold (>Y) in a certain percentage (X%) of the samples.



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

Useful Reference



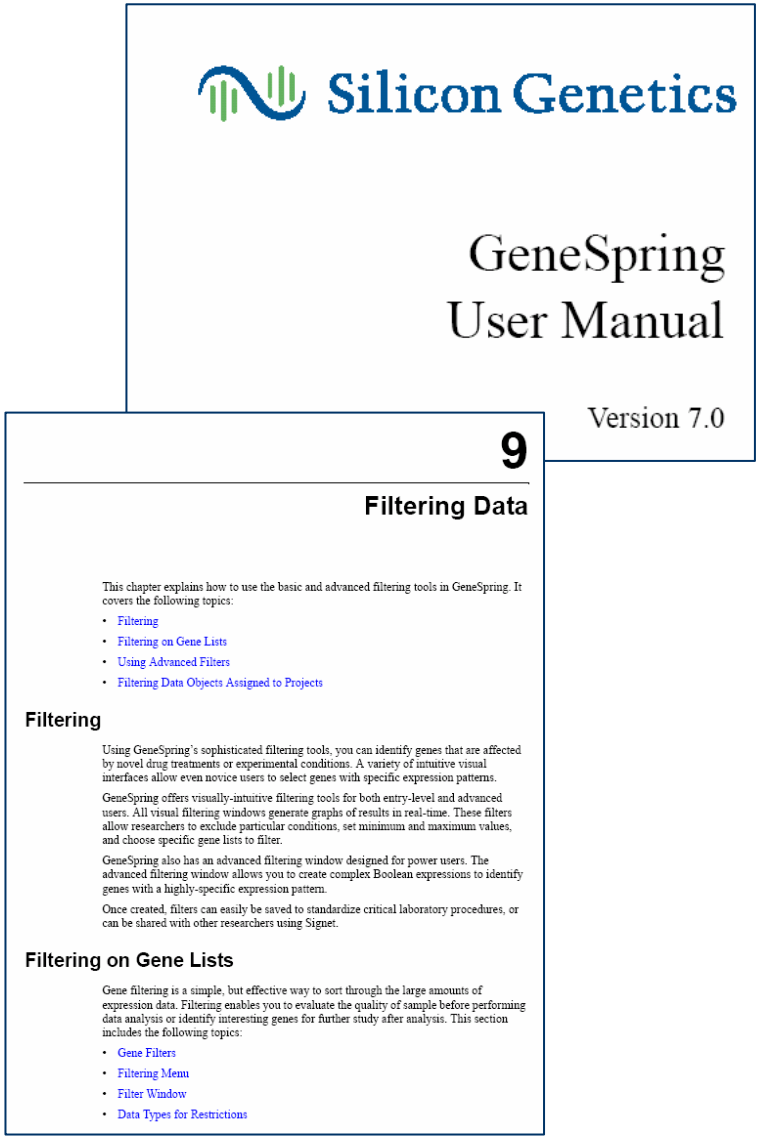
http://www.barleybase.org/filtertut.php

GeneSpring Tutorials

http://www.chem.agilent.com/Scripts/Generic.ASP?lPage=34743&indcol=Y&prodcol=Y

GeneSpring User Manual

http://www.chem.agilent.com/cfusion/faq/faq2.cfm?subsection=78&section=20&faq=1118&lang=en



Missing Values Imputation

Missing Values Imputation for Microarray Data

Missing values imply a loss of information

- Many analysis techniques that require complete data matrices: such as hierarchical clustering, k-means clustering, and self-organizing maps.
- May benefit from using more accurately estimated missing values.

Possible Solution

1. Exclude missing values from subsequent analysis.

May be of scientific interest !

2. Repeat the experiment
3. Missing values in replicated design **Expensive.**

4. Adjust dissimilarity measures. (e.g., pairwise deletion.)
5. Modify clustering methods that can deal with missing values.

6. Imputation of missing values.

Sources of Missing Values

- **Various Reasons**
  - a feature of the robotic apparatus may fail,
  - a scanner may have insufficient resolution,
  - simply dust or scratches on the slide (image corruption),
  - spots with dust particles, irregularities, …
- **Mathematical transformation**
  - undefined mathematical transformed:

    e.g., corrected intensities values that are negative or zero, a subsequent log-transformation will yield missing values.

Sources of Missing Values

Flag

Spots may be flagged as *absent* or *feature not found* when nothing is printed in the location of a spot.

- the imaging software cannot detect any fluorescence at the spot.
- expression readings that are barely above the background correction.
- the expression intensity ratio is undefined: */0, 0/*.

GenePix

Good=100. Bad=-100. Not Found=-50. Absent=-75. unflagged=0.

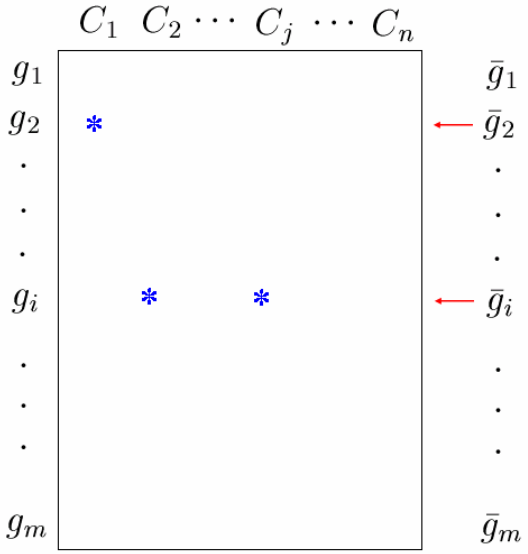
Imputation of Missing Values

- Missing log2 transformed data are replaced by *zeros* or by an *average* expression over the row ("row average“).

- Row average assumes that the expression of a gene in one of the experiments is similar to its expression in a different experiment, which is often not true in microarray experiments.

  - **Main weakness:**
    - it makes no serious attempt to model the connection of the missing values to the observed data.
    - since these methods do not take into consideration the correlation structure of the data.
    - not very effective (Troyanskaya et al, 2001)

  - **Useful:** where an initial imputation is required an iterative imputation method.



$$\widehat{C_j(g_i)} = \bar{g}_i$$

$i = 1, 2, \cdots, m.$

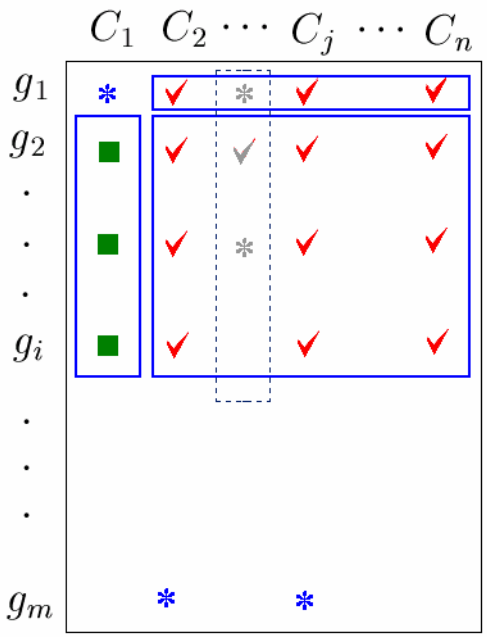
$j = 1, 2, \cdots, n.$

zero's
row average
row median

K-Nearest Neighbors Imputation

KNNImpute: a missing value estimation method to minimize data modeling assumptions and take advantage of the correlation structure of the gene expression data.

- Results are adequate and relatively insensitive to values of k between 10 and 20. (Troyanskaya et al, 2001)
- Euclidean distance appeared to be a sufficiently accurate norm.



KNNimpute

Model:

$$\{g_{(k)}, k = 1, 2, \cdots, K\} = \underset{k}{\text{args}} \max_{i \in C} \text{Corr}(g_1, g_i)$$

$$\{g_{(k)}, k = 1, 2, \cdots, K\} = \underset{k}{\text{args}} \min_{i \in C} \text{Dist}(g_1, g_i)$$

C: Observed C_i's without missing values

Imputation:

$$\text{Average} \quad \widehat{C_1(g_1)} = \frac{1}{K} \sum_{k=1}^{K} C_1(g_k)$$

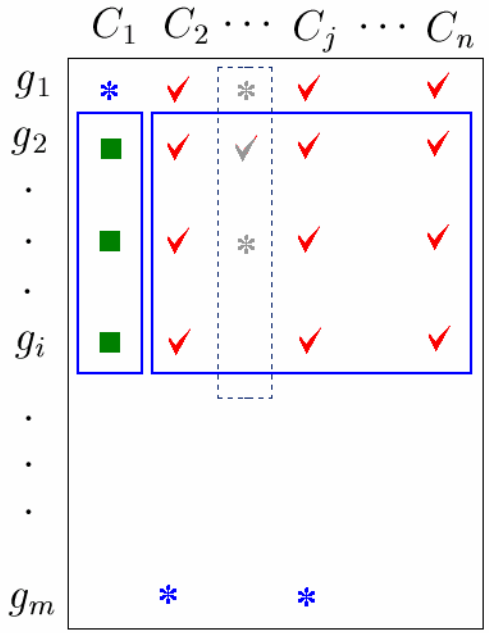
$$\text{Weighted Average} \quad \widehat{C_1(g_1)} = \frac{\sum_{k=1}^{K} w_k C_1(g_k)}{\sum_{k=1}^{K} w_k}$$

$$w_k = \frac{1}{\sum_{j \in C} [C_j(g_k) - C_1(g_1)]^2}$$

- Euclidean distance measure is often sensitive to outliers, which could be present in microarray data.
- Log-transformed data seems to sufficiently reduce the effect of outliers on genes similarity determination.

Regression Methods

- Using fitted regression values to replace missing values.
- The regression model can be applied to the original expression intensities or to transformed values.
- The model must be chosen so that it does not yields invalid fitted values. e.g., negative values.



Regression

Model:

$$C_1 = \beta_0 + \sum_{j \in C} \beta_j C_j$$

C: Observed C_i's without missing values

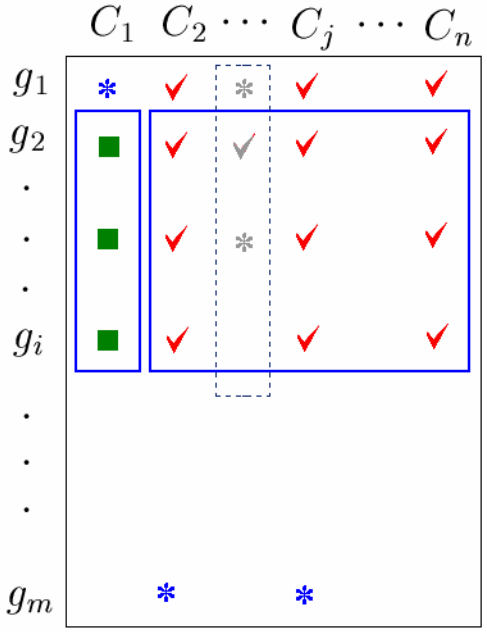
Imputation:

$$\widehat{C_1(g_1)} = \hat{\beta}_0 + \sum_{j \in C} \hat{\beta}_j C_j(g_1)$$

Regression Methods

Using the principal components as regressors.

- Each gene vector is estimated by a suitable regression combination of one or more of the most important principal components.
- The complete set can be obtained by row average method.
- These initial imputations are replaced by imputed values provided by the first application of the principal component method.
- The imputation can proceed through several iterations of the principal component method until the imputations converge to stable values.



Regression

Model:

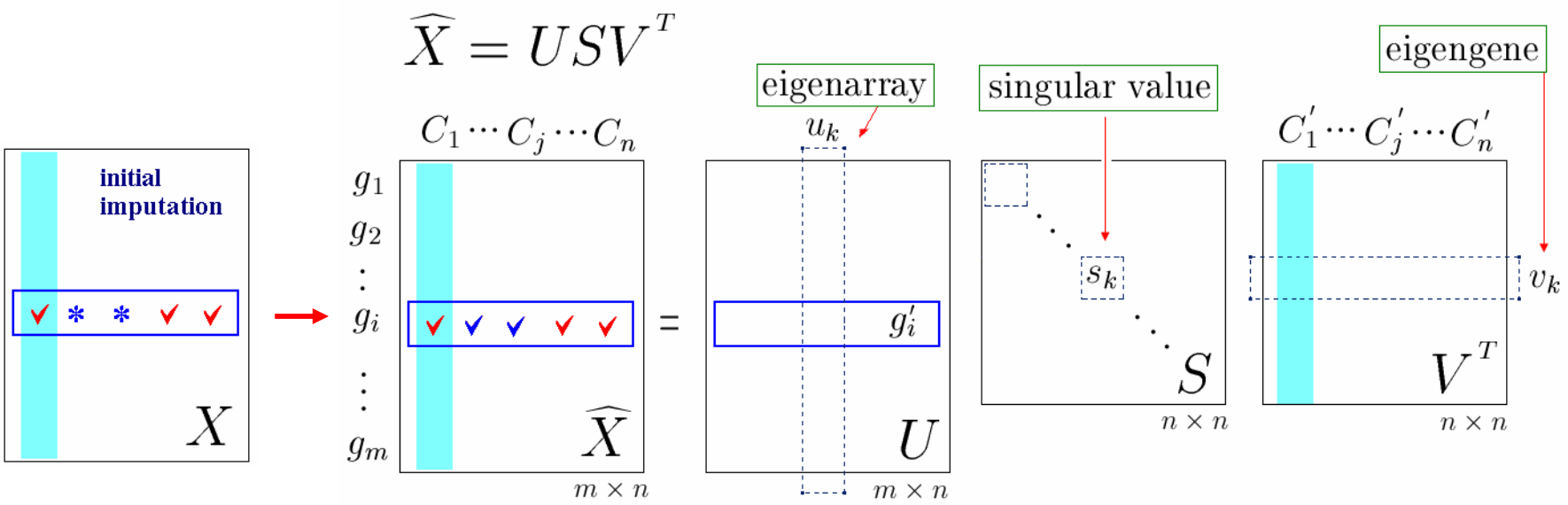
$$C_1 = \beta_0 + \sum_{j \in C} \beta_j C_j$$

C: Observed C_i's without missing values

Imputation:

$$\widehat{C_1(g_1)} = \hat{\beta}_0 + \sum_{j \in C} \hat{\beta}_j C_j(g_1)$$

Singular Value Decomposition Imputation



SVDimpute

Model:

$$g_i(\mathrm{C}) = \beta_0 + \sum_{k=1}^{K} \beta_k v_{(k)}(\mathrm{C})$$

C: Observed C_i's without missing values

Imputation:

$$\widehat{g_i(\sim \mathrm{C})} = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k v_{(k)}(\sim \mathrm{C})$$

Could Extend to Iterative approach

- Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. (2001), Missing value estimation methods for DNA microarrays. Bioinformatics 17(6), 520-525.
- Trevor Hastie , Robert Tibshirani, Gavin Sherlock , Michael Eisen , Patrick Brown , David Botstein. (1999). Imputing Missing Data for Gene Expression Arrays, Technical Report.

Reference for Missing Values Imputation

Singular Value Decomposition Imputation

- Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. (2001), Missing value estimation methods for DNA microarrays. Bioinformatics 17(6), 520-525.
- Trevor Hastie , Robert Tibshirani, Gavin Sherlock , Michael Eisen , Patrick Brown , David Botstein. (1999). Imputing Missing Data for Gene Expression Arrays, Technical Report.

Local Least Square Imputation

- Bo TH, Dysvik B, Jonassen I. LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Res. 2004 Feb 20;32(3):e34.
- Hyunsoo Kimy, Gene H. Golubz, and Haesun Parky. (2004). Missing Value Estimation for DNA Microarray Gene Expression Data: Local Least Squares Imputation, Bioinformatics Advance Access published August 27, 2004.

Bayesian

- Oba S, Sato M-A, Takemasa I, Monden M, Matsubara K-I, Ishii S: A Bayesian missing value estimation method for gene expression profile data,. Bioinformatics 2003, 19:2088-2096.
- Zhou X, Wang X, Dougherty ER: Missing-value estimation using linear and non-linear regression with Bayesian gene selection. Bioinformatics 2003, 19:2302-2307.

GMCimpute

- Ouyang M, Welsh WJ, Georgopoulos P. Gaussian mixture clustering and imputation of microarray data. Bioinformatics. 2004 Apr 12;20(6):917-23. Epub 2004 Jan 29.

Others

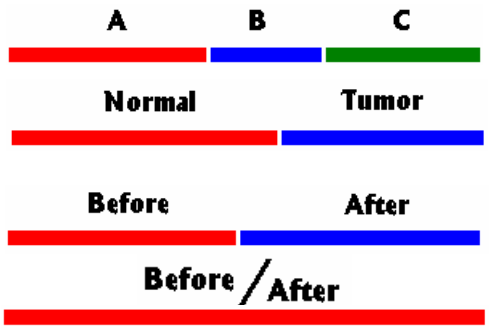
- Kim KY, Kim BJ, Yi GS. Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics. 2004 Oct 26;5(1):160.
- Shmuel Friedland, Amir Niknejad, and Laura Chiharaz. (2004). A Simultaneous Reconstruction of Missing Data in DNA Microarrays, Institute for Mathematics and its Applications,.
- Alexandre G de Brevern, Serge Hazout and Alain Malpertuy. (2004). Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering, BMC Bioinformatics Volume 5.

Which Imputation Method?

- KNN is the most widely-used.
- Characteristics of data that may affect choice of imputation method:
  - dimensionality
  - percentage of values missing
  - experimental design (time series, case/control, etc.)
  - patterns of correlation in data
- Suggestion
  - add artificial missing values to your data set
  - impute them with various methods
  - see which is best (since you know the real value)

Finding Differential Expressed Genes

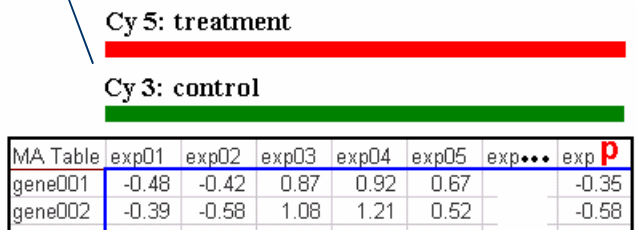
Finding Differentially Expressed Genes



→ More than two samples

→ Two-sample (independent)

→ Paired-sample (dependent)



MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58

MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p	p-values	
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35	0.067	
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58	0.052	
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13	0.013	*
gene004	1.57	1.03	1.22	0.31	0.16		-1.02	0.016	*
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44	0.112	
gene006	0.04	-0.12	0.31	0.16	0.17		0.08	0.017	*
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76	0.059	
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55	0.063	
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03	0.516	
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45	0.009	*
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01	0.068	
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63	0.030	*
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04	0.002	*
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21	0.423	
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64	0.084	
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27	0.048	
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02	0.018	*
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05	0.538	
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38	0.053	
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70	0.074	
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40	0.764	
gene022	-0.11	0.13	0.41	0.60	0.23		0.19	0.423	
gene•••									
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66	0.723	

Microarray Data Matrix

gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
⋮							
gene022	-0.11	0.13	0.41	0.60	0.23		0.19

- Select a statistic which will rank the genes in order of evidence for differential expression, from strongest to weakest evidence.

(Primary Importance): only a limited number of genes can be followed up in a typical biological study.

- Choose a critical-value for the ranking statistic above which any value is considered to be significant.

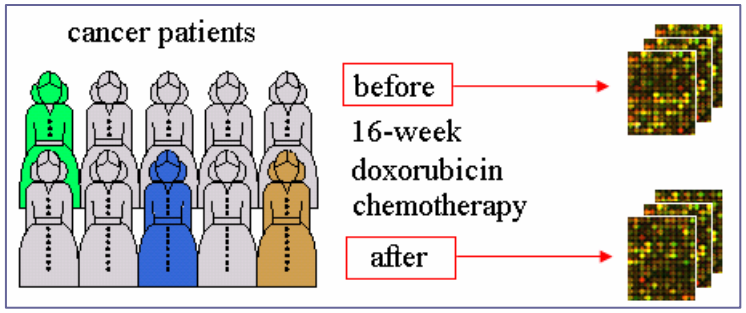
Example 1: Breast Cancer Dataset

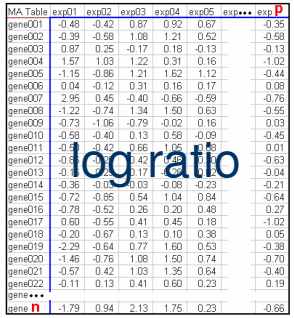
cDNA microarrays

- Samples are taken from 20 breast cancer patients, before and after a 16 week course of doxorubicin chemotherapy, and analyzed using microarray. There are 9216 genes.

- **Paired data**: there are two measurements from each patient, one before treatment and one after treatment.

- These two measurements relate to one another, we are interested in the difference between the two measurements (the log ratio) to determine whether a gene has been up-regulated or down-regulated in breast cancer following that treatment.





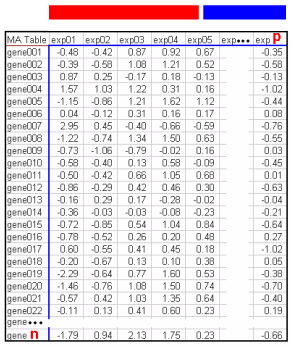
9216 x 20

Perou CM, et al, (2000), Molecular portraits of human breast tumours. Nature 406:747-752.
Stanford Microarray Database: http://genome-www.stanford.edu/breast_cancer/molecularportraits/

Example 2: Leukemia Dataset

Affymetrix

- Bone marrow samples are taken from
  - 27 patients suffering from acute lymphoblastic leukemia (ALL，急性淋巴細胞白血病) and
  - 11 patients suffering from acute myeloid leukemia (AML，急性骨髓性白血病) and analyzed using Affymetrix arrays.
  - There are 7070 genes.
- **Unpaired data**: there are two groups of patients (ALL, AML).



MA Table	exp01	exp02	exp03	exp04	exp05	exp...	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13
gene004	1.57	1.03	1.22	0.31	0.16		-1.02
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44
gene006	0.04	-0.12	0.31	0.16	0.17		0.08
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40
gene022	-0.11	0.13	0.41	0.60	0.23		0.19
gene...							
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66

- We wish to identify the genes that are up- or down-regulated in ALL relative to AML. (i.e., to see if a gene is differentially expressed between the two groups.)

7070 x (27+11)

Golub, T.R et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531--537.
Cancer Genomics Program at Whitehead Institute for Genome Research
http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

Example 3: Small Round Blue Cell Tumors (SRBCT) Dataset

cDNA microarrays

- There are four types of small round blue cell tumors of childhood:
  - Neuroblastoma (NB) (12),
  - Non-Hodgkin lymphoma (NHL) (8),
  - Rhabdomyosarcoma (RMS) (20) and
  - Ewing tumours (EWS) (23).
  - Sixty-three samples from these tumours have been hybridized to microarray.
- We want to identify genes that are differentially expressed in one or more of these four groups.

More on SRBCT:
http://www.thedoctorsdoctor.com/diseases/small_round_blue_cell_tumor.htm

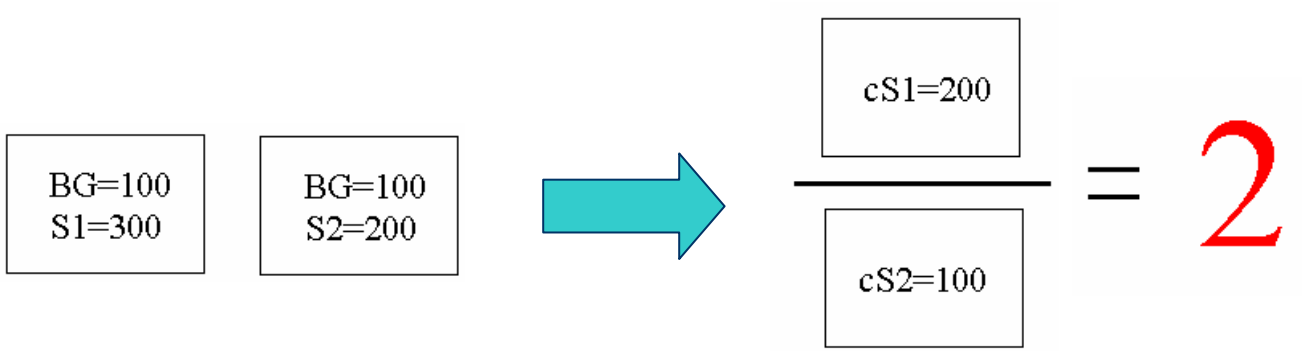
Khan J, Wei J, Ringner M, Saal L, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu C, Peterson C and Meltzer P. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 2001, 7:673-679
Stanford Microarray Database

Fold-Change Method

Calculate the expression ratio in control and experimental cases and to rank order the genes. Chose a threshold, for example at least 2-fold up or down regulation, and selected those genes whose average differential expression is greater than that threshold.

Problems: it is an arbitrary threshold.

- In some experiments, no genes (or few gene) will meet this criterion.
- In other experiments, thousands of genes regulated.



- s2 close to BG, the difference could represent noise.
- It is more credible that a gene is regulated 2-fold with 10000, 5000 units)

- The average fold ratio does not take into account the extent to which the measurements of differential gene expression vary between the individuals being studied.

- The average fold ratio does not take into account the number of patients in the study, which statisticians refer to as the sample size.

Fold-Change Method (conti.)

Define which genes are significantly regulated might be to choose 5% of genes that have the largest expression ratios.

Problems:

- It applies no measure of the extent to which a gene has a different mean expression level in the control and experimental groups.
- Possible that no genes in an experiment have statistically significantly different gene expression.

Hypothesis Testing

Hypothesis Testing

A *hypothesis test* is a procedure for determining if an assertion about a characteristic of a population is reasonable.

Example

someone says that the average price of a gallon of regular unleaded gas in Massachusetts is $2.5.

How would you decide whether this statement is true?

- find out what every gas station in the state was charging and how many gallons they were selling at that price.
- find out the price of gas at a small number of randomly chosen stations around the state and compare the average price to $2.5.

- Of course, the average price you get will probably not be exactly $2.5 due to variability in price from one station to the next.

Suppose your average price was $2.23. Is this three cent difference a result of chance variability, or is the original assertion incorrect?

A **hypothesis test** can provide an answer.

Terminology

- The ***null hypothesis:***
  - H**0**: μ = 2.5. (the average price of a gallon of gas is $2.5)
- The ***alternative hypothesis:***
  - H**1**: μ > 2.5. (gas prices were actually higher)
  - H**1**: μ < 2.5.
  - H**1**: μ != 2.5.
- The ***significance level (alpha)***
  - Alpha is related to the degree of certainty you require in order to reject the null hypothesis in favor of the alternative.
  - Decide in advance to reject the null hypothesis if the probability of observing your sampled result is less than the significance level.
  - Alpha = 0.05: the probability of incorrectly rejecting the null hypothesis when it is actually true is 5%.
  - If you need more protection from this error, then choose a lower value of alpha .

Example

H_0: No differential expressed.

H_0: There is no difference in the mean gene expression in the group tested.

H_0: The gene will have equal means across every group.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 (\ldots = \mu_n)$

The *p*-values

- *p* is the probability of observing your data under the assumption that the null hypothesis is true.
- *p* is the probability that you will be in error if you reject the null hypothesis.
- *p* represents the probability of **false positives** (Reject $\mathbf{H_0}$ | $\mathbf{H_0}$ true).

p=0.03 indicates that you would have only a 3% chance of drawing the sample being tested if the null hypothesis was actually true.

Decision Rule

- Reject H_0 if *P* is less than alpha.
- $P < 0.05$ commonly used. (Reject $\mathbf{H_0}$, the test is significant)
- The lower the p-value, the more significant the difference between the groups.

P is *not* the probability that the null hypothesis is true!

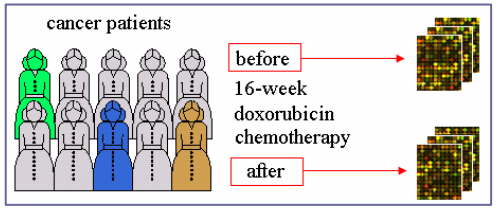
Type I Error (alpha): calling genes as differentially expressed when they are NOT

Type II Error: NOT calling genes as differentially expressed when they ARE

$$\text{Power} = 1 - \beta.$$

Hypothesis Testing		Truth	
		H0	**H1**
Decision	**Reject H0**	Type I Error (alpha) (false positive)	Right Decision (true positive)
	Don't Reject H0	Right Decision	Type II Error (beta)

Hypothesis Testing



Dependent samples

Independent samples

Comparison	Two Groups		More than two Groups
Hypothesis Testing	**Paired data**	**Unpaired data**	**Complex data**
Parametric (variance equal)	One sample t-test	Two-sample t-test	One-Way Analysis of Variance (ANOVA)
Parametric (variance not equal)	Welch t-test		Welch ANOVA
Non-Parametric (無母數檢定)	Wilcoxon Signed-Rank Test	Wilcoxon Rank-Sum Test (Mann-Whitney U Test)	Kruskal-Wallis Test

Steps of Hypothesis Testing

1. Determine the null and alternative hypothesis, using mathematical expressions if applicable.
2. Select a significance level (alpha).
3. Take a random sample from the population of interest.
4. Calculate a test statistic from the sample that provides information about the null hypothesis.
5. Decision

Hypothesis Testing:

two-sided z-test & p-value

H_0: $\mu = 35$ null hypothesis

H_1: $\mu \neq 35$ alternative hypothesis ($\mu > 35$; $\mu < 35$) one-sided

α signifcant level: $= 0.05$

test statistic $z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

Reject H_0 if $|z| > z_{0.05}$

$$H_0 : \mu = m$$
$$H_1 : \mu \neq m$$
$$\alpha = P_{H_0}(|Z| > z_{\alpha/2})$$

Sample Data: $=33.6$

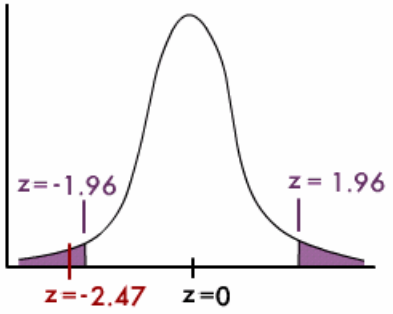
test statistic: $z = -2.47$

$(1-\alpha)100\%$ Confidence Interval:

$$P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha$$

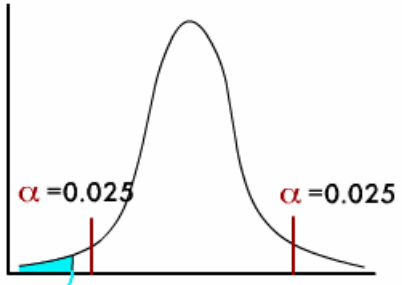
p-value $= P_{H_0}(|Z| > z_0)$, $z_0 = \frac{\bar{X} - m}{\sigma/\sqrt{n}}$

The Classical Approach



Conclusion: since the z value of the test statistic (-2.47) is less than the critical value of z= -1.96, we reject the null hypothesis.

The P-Value Approach



P -value = 0.0068 times 2 (for a 2-sided test) = 0.0136

Conclusion: since the P -value of 0.0136 is less than the significance level of α=0.05, we reject the null hypothesis.

Hypothesis Tests on Microarray Data

- The null hypothesis is that there is no biological effect.
  - For a gene in Breast Cancer Dataset, it would be that this gene is not differentially expressed following doxorubicin chemotherapy.
  - For a gene in Leukemia Dataset, it would be that this gene is not differentially expressed between ALL and AML patients.
- If the null hypothesis were true, then the variability in the data does not represent the biological effect under study, but instead results from difference between individuals or measurement error.
- The smaller the p-value, the less likely it is that the observed data have occurred by chance, and the more significant the result.
- *p=0.01* would mean there is a 1% chance of observing at least this level of differential gene expression by random chance.
- We then select differentially expressed genes not on the basis of their fold ratio, but on the basis of their p-value.

H_0: no differential expressed.

- The test is significant
  = Reject **H_0**
- False Positive
  = (Reject **H_0** | **H_0** true)
  = concluding that a gene is differentially expressed when in fact it is not.

- **A p-value=0.05 indicates that you would have only a 5% chance of drawing the sample being tested if the null hypothesis was actually true.**
- **The p-value is the smallest level of significance at which a null hypothesis may be rejected**

One Sample t-test

The One-Sample t-test compares the mean score of a sample to a known value. Usually, the known value is a population mean.

Assumption: the variable is normally distributed.

One sample t-test

$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$ (two-tailed).
μ: population mean.
α: significant level (e.g., 0.05).
Test Statistic:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}, \quad t_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$\bar{X}$: sample mean.

S: sample standard deviation.

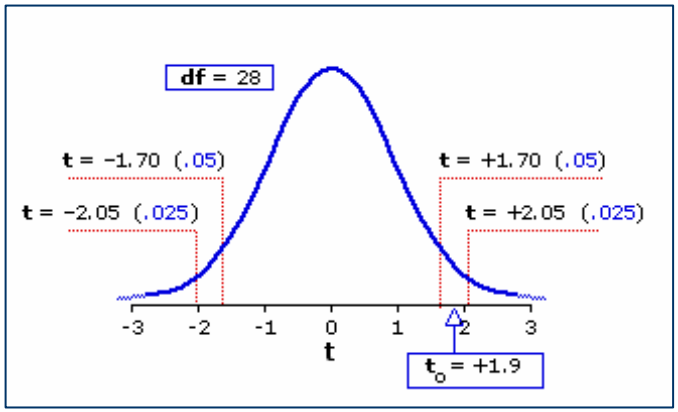
n: number of observations in the sample.

- Reject H_0 if $|t_0| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ: $\bar{X} - t_{\alpha/2}S/\sqrt{n} \leq \mu < \bar{X} + t_{\alpha/2}S/\sqrt{n}$
- $p\text{-}value = P_{H_0}(|\mathbf{T}| > t_0)$, $\mathbf{T} \sim t_{n-1}$.

Question

- whether a gene is differentially expressed for a condition with respect to baseline expression?
- H_0: $\mu = 0$ (log ratio)

MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13



Two Sample t-test

Paired Sample t-test

$H_0 : \mu_d = \mu_0$
$H_1 : \mu_d \neq \mu_0$ (two-tailed).
μ_d: mean of population differences.
α: significant level (e.g., 0.05).
Test Statistic:

$$T_d = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}}, \quad t_d = \frac{\bar{d} - \mu_0}{S_d/\sqrt{n}}$$

$\bar{d}$: average of sample differences.

S_d: standard deviation of sample difference

n: number of pairs.

- Reject H_0 if $|t_d| > t_{\alpha/2,n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ_d: $\bar{d} - t_{\alpha/2}S/\sqrt{n} \leq \mu_d < \bar{d} + t_{\alpha/2}S/\sqrt{n}$
- $p\text{-}value = P_{H_0}(|\mathbf{T}| > t_d)$, $\mathbf{T} \sim t_{n-1}$.

Two Sample t-test (Unpaired)

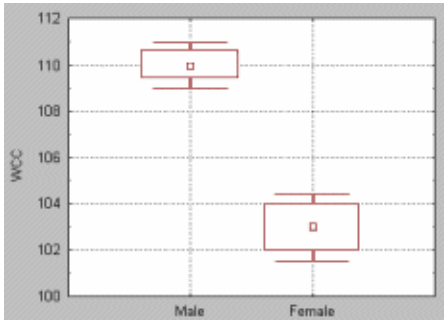
$H_0 : \mu_x - \mu_y = \mu_0$
$H_0 : \mu_x - \mu_y \neq \mu_0$
α: significant level (e.g., 0.05).
Test Statistic:

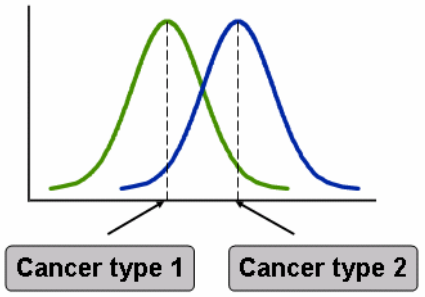
$$t_0 = \frac{(\bar{X} - \bar{Y}) - \mu_0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

for homogeneous variances:
$df = n + m - 2$

for heterogeneous variances:
adjusted df

Reject H_0 if $|t_0| > t_{\alpha/2,\, df}$





Paired t-test
Applied to a gene From Breast Cancer Data

- The gene acetyl-Coenzyme A acetyltransferase 2 (ACAT2) is on the microarray used for the breast cancer data.
- We can use a paired t-test to determine whether or not the gene is differentially expressed following doxoruicin chemotherapy.
- The samples from before and after chemotherapy have been hybridized on separate arrays, with a reference sample in the other channel.
  - Normalize the data.
  - Because this is a reference sample experiment, we calculate the log ratio of the experimental sample relative to the reference sample for before and after treatment in each patient.
  - Calculate a single log ratio for each patient that represents the difference in gene expression due to treatment by subtracting the log ratio for the gene before treatment from the log ratio of the gene after treatment.
  - Perform the t-test. t=3.22 compare to t(19).
  - The p-value for a two-tailed one sample t-test is 0.0045, which is significant at a 1% confidence level.
- Conclude: this gene has been significantly down-regulated following chemotherapy at the 1% level.

Unpaired t-test
Applied to a Gene From Leukemia Dataset

- The gene metallothionein IB is on the Affymetrix array used for the leukemia data.
  - To identify whether or not this gene is differentially expressed between the AML and ALL patients.
  - To identify genes which are up- or down-regulation in AML relative to ALL.
- Steps
  - the data is log transformed.
  - t=-3.4177, p=0.0016
- Conclude that the expression of metallothionein IB is significantly higher in AML than in ALL at the 1% level.

Assumptions of t-test

- The distribution of the data being tested is normal.
  - For paired t-test, it is the distribution of the subtracted data that must be normal.
  - For unpaired t-test, the distribution of both data sets must be normal.

- **Plots**: Histogram, Density Plot, QQplot,…
- **Test for Normality**: Jarque-Bera test, Lilliefors test, Kolmogorov-Smirnov test.

- Homogeneous: the variances of the two population are equal.
- Test for equality of the two variances: Variance ratio F-test.

Note:

◆ If the two populations are symmetric, and if the variances are equal, then the *t* test may be used.

◆ If the two populations are symmetric, and the variances are not equal, then use the two-sample unequal variance t-test or Welch's *t* test.

Other t-Statistics

B-statistic

Lonnstedt and Speed, Statistica Sinica 2002: parametric empirical Bayes approach.

- B-statistic is an estimate of the posterior log-odds that each gene is DE.
- B-statistic is equivalent for the purpose of ranking genes to the penalized t-statistic $t = \frac{\bar{M}}{\sqrt{(a+s^2)/n}}$, where a is estimated from the mean and standard deviation of the sample variances s^2.

$$M_{gj}|\mu_g, \sigma_g \sim N(\mu_g, \sigma_g^2)$$

$$B_g = \log \frac{P(\mu_g \neq 0|M_{gj})}{P(\mu_g = 0|M_{gj})}$$

Lonnstedt, I. and Speed, T.P. Replicated microarray data. *Statistica Sinica* , 12: 31-46, 2002

Penalized t-statistic

Tusher et al (2001, PNAS, SAM)
Efron et al (2001, JASA)

$$t = \frac{\bar{M}}{(a+s)/\sqrt{n}}$$

General Penalized t-statistic

(Lonnstedt et al 2001)

$$t = \frac{b}{s^* \times SE}$$

multiple regression model

Penalized two-sample t-statistic

$$t = \frac{\bar{M}_A - \bar{M}_B}{s^* \times \sqrt{1/n_A + 1/n_B}}, \quad \text{where } s^* = \sqrt{a + s^2}$$

Robust General Penalized t-statistic

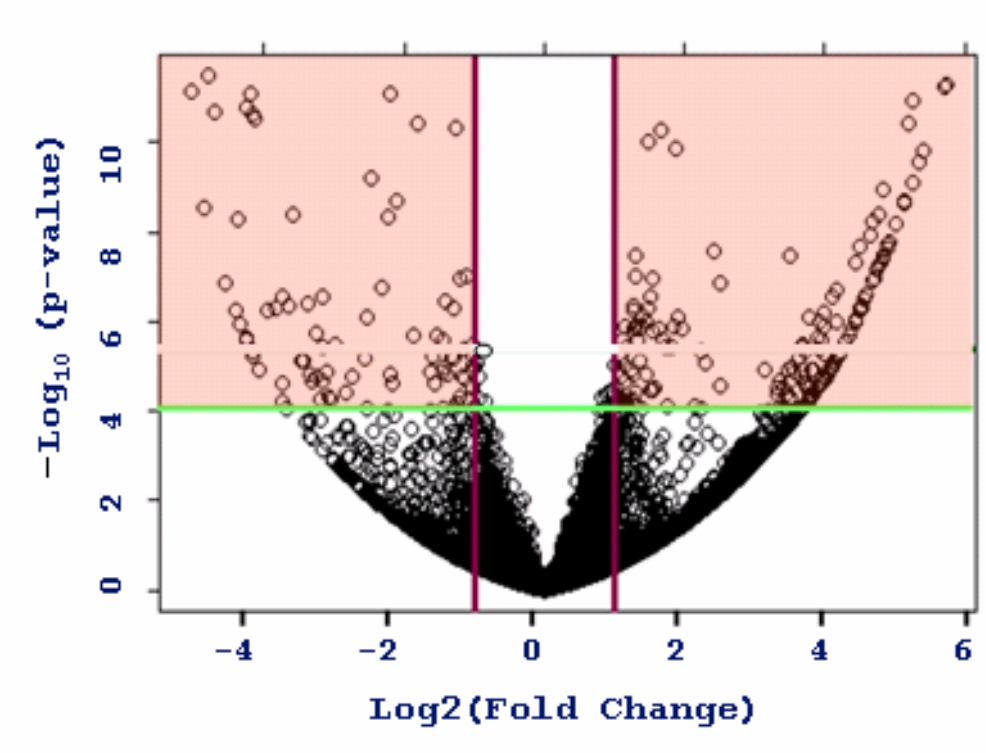
Non-parametric Statistics

- Do not assume that the data is normally distributed.
- There are two good reasons to use non-parametric statistic.
  - *Microarray data is noisy:*
    - there are many sources of variability in a microarray experiment and outliers are frequent.
    - The distribution of intensities of many genes may not be normal.
    - Non-parametric methods are robust to outliers and noisy data.
  - *Microarray data analysis is high throughput:*
    - When analysising the many thousands of genes on a microarray, we would need to check the normality of every gene in order to ensure that t-test is appropriate.
    - Those genes with outliers or which were not normally distributed would then need a different analysis.
    - It makes more sense to apply a test that is distribution free and thus can be applied to all genes in a single pass.

Volcano Plot

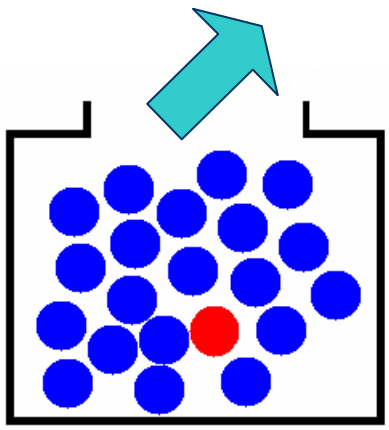
The Y variate is typically a probability (in which case a -log10 transform is used) or less commonly a p-value.

The X variate is usually a measure of differential expression such as a log-ratio.



Multiple Testing

Multiple Testing



Imagine a box with 20 marbles: 19 are blue and 1 is red.

What are the odds of randomly sampling the red marble by chance?
It is 1 out of 20.

Now sample a single marble (and put it back into the box) 20 times.
Have a much higher chance to sample the red marble.
This is exactly what happens when testing several thousand genes at the same time:

Imagine that the red marble is a false positive gene:
the chance that false positives are going to be sampled is higher the more genes you apply a statistical test on.

X: false positive gene

$$P(X \geq 1)$$

$$= 1 - P(X = 0)$$

$$= 1 - 0.95^n$$

Multiplicity of Testing

Number of genes tested (N)	False positives incidence	Probability of calling 1 or more false positives by chance ($100(1-0.95^N)$)
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.4%

Multiplicity of Testing

- There is a serious consequence of performing statistical tests on many genes in parallel, which is known as multiplicity of p-values.

- Take a large supply of reference sample, label it with Cy3 and Cy5: no genes are differentially expressed: all measured differences in expression are experimental error.
  - By the very definition of a p-value, each gene would have a 1% chance of having a p-value of less than 0.01, and thus be significant at the 1% level.
  - Because there are 10000 genes on this imaginary microarray, we would expect to find 100 significant genes at this level.
  - Similarly, we would expect to find 10 genes with a p-value less than 0.001, and 1 gene with p-value less than 0.0001
  - The p-value is the probability that a gene's expression level are different between the two groups due to chance.

Question:

1. How do we know that the genes that appear to be differentially expressed are truly differentially expressed and are not just artifact introduced because we are analyzing a large number of genes?
2. Is this gene truly differentially expressed, or could it be a false positive results?

Types of Error Control

- **Multiple testing** correction adjusts the p-value for each gene to keep the overall error rate (or false positive rate) to less than or equal to the user-specified p-value cutoff or error rate individual.

Multiple Testing

	# Reject H_0	# not Reject H_0	
# true H_{0j}	V	U	m_0
# true H_{1j}	S	T	m_1
	R	m - R	m

V : *false positives = Type I errors*

T : *false negatives = Type II errors*

Type One Errors Rates

$$\text{PCER} = \frac{E[V]}{m}$$

$$\text{PFER} = E[V]$$

$$\text{FWER} = p(V \geq 1)$$

$$\text{FDR} = E[\frac{V}{R}] \quad \text{if } R > 0$$

Power = Reject the false null hypothesis

$$\text{Any-pair Power} = p(S \geq 1)$$

$$\text{Per-pair Power} = \frac{E[S]}{m_1}$$

$$\text{All-pair Power} = p(S = m_1)$$

Multiple Testing Corrections

Test Type	Type of Error control	Genes identified by chance after correction
Bonferroni	Family-wise error rate	If error rate equals 0.05, expects **0.05** genes to be significant by chance
Bonferroni Step-down		
Westfall and Young permutation		
Benjamini and Hochberg	False Discovery Rate	If error rate equals 0.05, **5%** of genes considered statistically significant (that pass the restriction after correction) will be identified by chance (false positives).

- The more stringent a multiple testing correction, the less false positive genes are allowed.
- The trade-off of a stringent multiple testing correction is that the rate of *false negatives* (genes that are called non-significant when they are) is very high.
- FWER is the overall probability of false positive in all tests.
  - Very conservative
  - False positives not tolerated
- False discovery error rate allows a percentage of called genes to be false positives.

(1) Bonferroni

- The p-value of each gene is multiplied by the number of genes in the gene list.
- If the corrected p-value is still below the error rate, the gene will be significant:
  - Corrected p-value= p-value * n <0.05.
  - If testing 1000 genes at a time, the highest accepted individual un-corrected p-value is 0.00005, making the correction very stringent.
- With a Family-wise error rate of 0.05 (i.e., the probability of at least one error in the family), the expected number of false positives will be 0.05.

(4) Benjamini and Hochberg FDR

- This correction is the least stringent of all 4 options, and therefore tolerates more false positives.
- There will be also less false negative genes.
- The correction becomes more stringent as the p-value decreases, similarly as the Bonferroni Step-down correction.
- This method provides a good alternative to Family-wise error rate methods.
- The error rate is a proportion of the number of called genes.
- FDR: Overall proportion of false positives relative to the total number of genes declared significant.

Corrected P-value= p-value * $(n / R_i) < 0.05$

Let n=1000, error rate=0.05

Gene name	p-value (from largest to smallest)	Rank	Correction	Is gene significant after correction?
A	0.1	1000	No correction	0.1 > 0.05 ➔ No
B	0.06	999	1000/999*0.06 = 0.06006	0.06006 > 0.05 ➔ No
C	0.04	998…	1000/998*0.04 = 0.04008	0.04008 < 0.05 ➔ Yes

Recommendations

- The default multiple testing correction in GeneSpring is the Benjamini and Hochberg False Discovery Rate.
- It is the least stringent of all corrections and provides a good balance between discovery of statistically significant genes and limitation of false positive occurrences.
- The Bonferroni correction is the most stringent test of all, but offers the most conservative approach to control for false positives.
- The Westfall and Young Permutation is the only correction accounting for genes coregulation. However, it is very slow and is also very conservative.
- As multiple testing corrections depend on the number of tests performed, or number of genes tested, it is recommended to select a prefiltered gene list.

If There Are No Results with MTC

- increase p-cutoff value
- increase number of replicates
- use less stringent or no MTC
- add cross-validation experiments

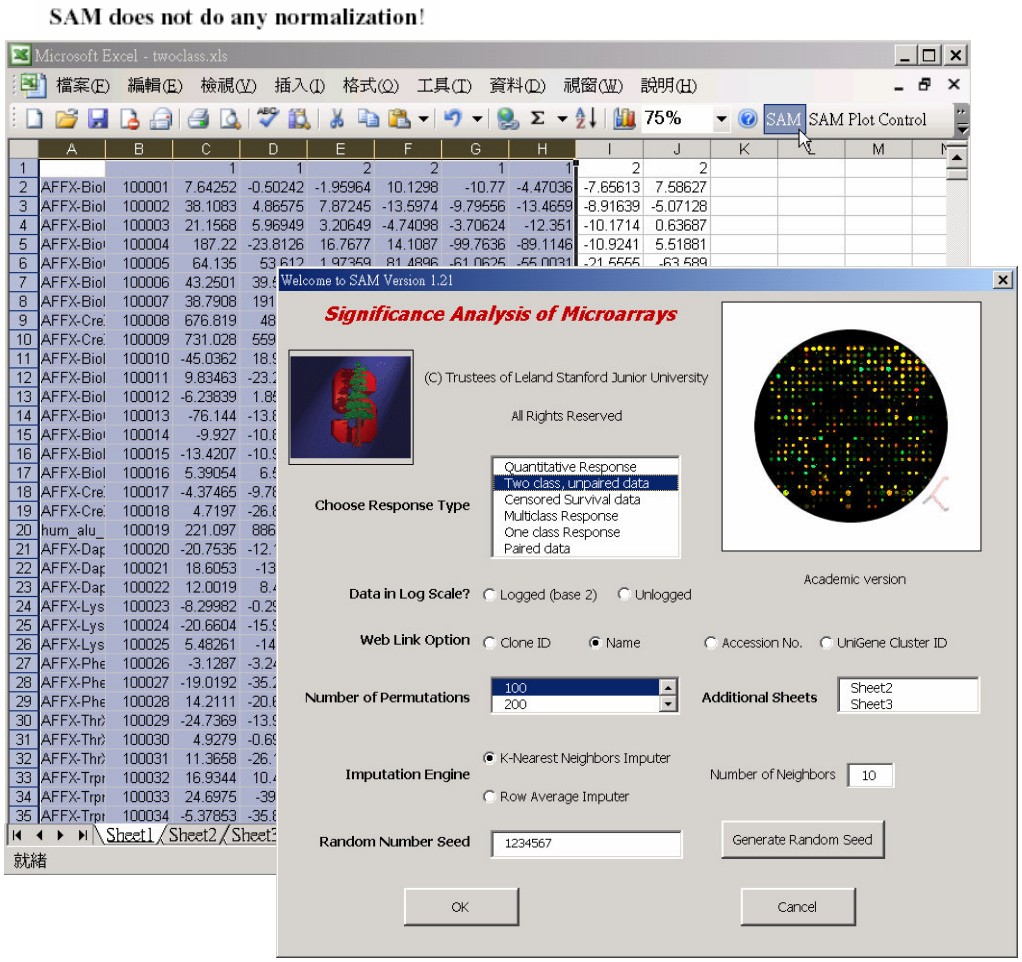
SAM

SAM: Significance Analysis of Microarrays

http://www-stat.stanford.edu/~tibs/SAM/

SAM assigns a score to each gene in a microarray experiment based upon its change in gene expression relative to the standard deviation of repeated measurements.

- ***SAM plot***: the number of observed genes versus the expected number. This visualizes the outlier genes that are most dramatically regulated.
- ***False discovery rate*****:** is the percent of genes that are expected to be identified by chance.
- ***q-value***: the lowest false discovery rate at which a gene is described as significantly regulated.



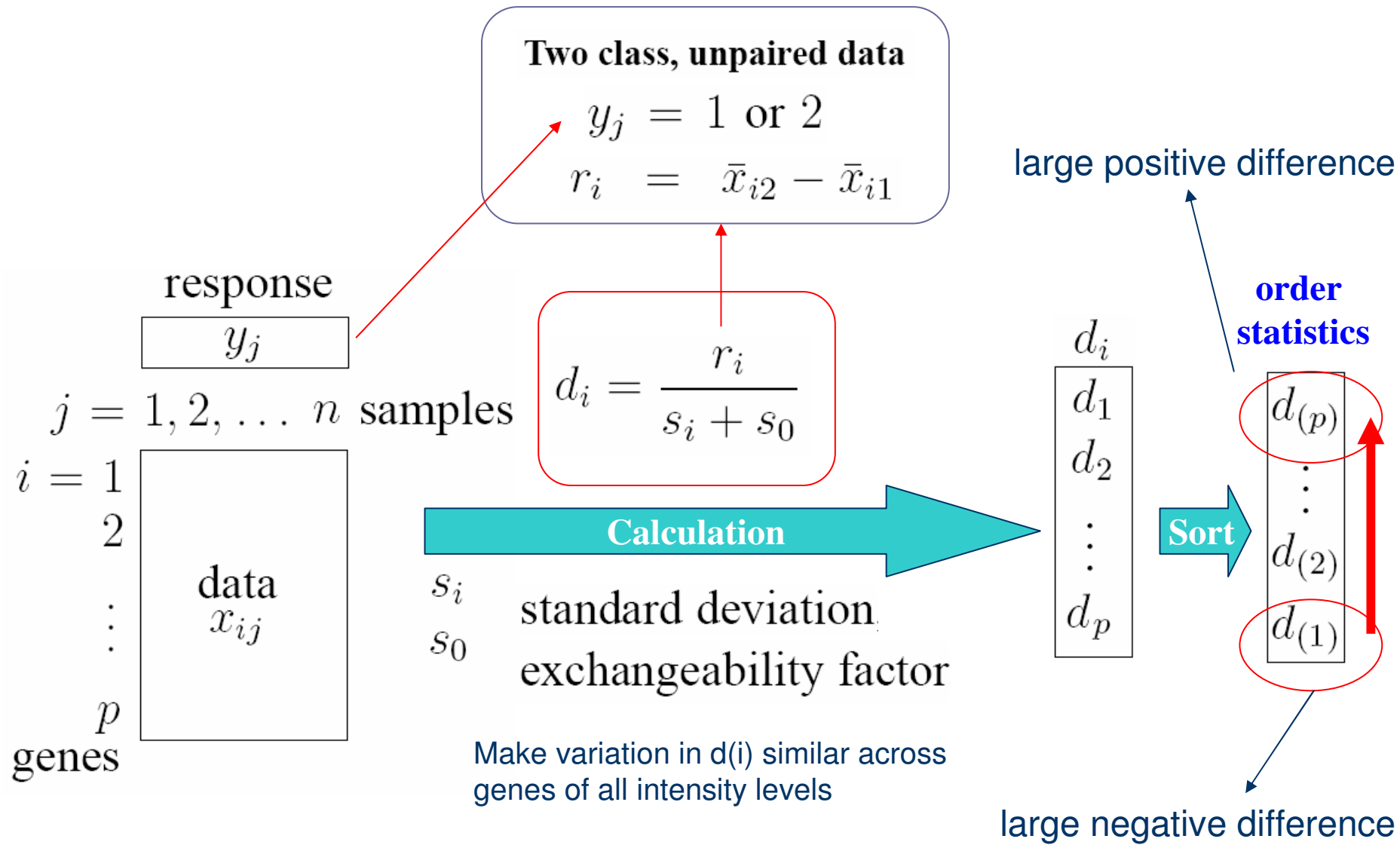
Tusher VG, Tibshirani R, Chu G.(2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci* 98(9):5116-21.

SAM: Response Type

Response type	Coding
Quantitative	Real number eg 27.4 or -45.34
Two class (unpaired)	Integer 1, 2
Multiclass	Integer 1, 2, 3, ...
Paired	Integer -1, 1, -2, 2, etc. eg - means Before treatment, + means after treatment -1 is paired with 1, -2 is paired with 2, etc.
Survival data	(Time, status) pair like (50,1) or (120,0) First number is survival time, second is status (1=died, 0=censored)
One class	Integer, every entry equal to 1
Time course, two class (unpaired)	(1 or 2)Time(t)[Start or End]
Time course, two class (paired)	(-1 or 1 or -2 or 2 etc)Time(t)[Start or End]
Time course, one class	1Time(t)[Start or End]
Pattern discovery	eigengenek, where k is one of 1,2,... number of arrays

SAM Users guide and technical document

SAM: Significance Analysis of Microarrays



SAM: Expected Test Statistics

response

y_j

$1, 1, \ldots, 2, \ldots, 2$

Permutation

$1, 2, 1, 2, 1, \ldots, 1$

$$r_i^* = \bar{x}_{i2}^* - \bar{x}_{i1}^*$$

$$d_i^* = \frac{r_i^*}{s_i^* + s_0^*}$$

$$\begin{bmatrix} d_{(p)}^{*b} \\ \vdots \\ d_{(2)}^{*b} \\ d_{(1)}^{*b} \end{bmatrix} \quad b = 1, 2, \ldots B$$

$$\bar{d}_{(i)} = (1/B) \sum_b d_{(i)}^{*b}$$

expected order statistics

$$\begin{bmatrix} \bar{d}_{(p)} \\ \vdots \\ \bar{d}_{(2)} \\ \bar{d}_{(1)} \end{bmatrix}$$

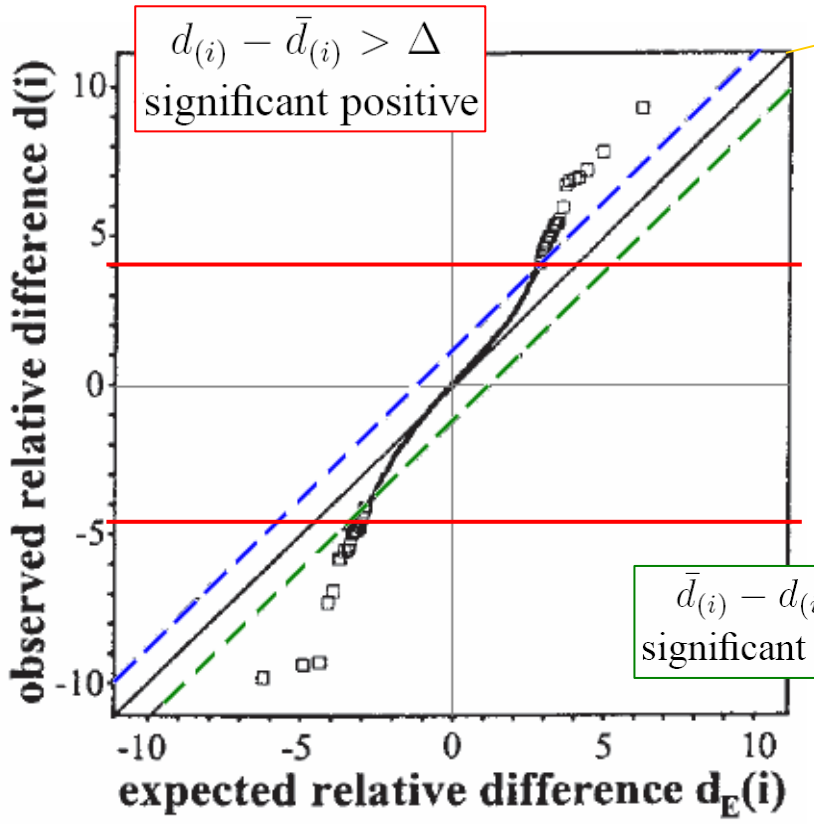
SAM Plot

Points for genes with evidence of induction

$$\begin{bmatrix} d_{(p)} \\ \vdots \\ d_{(2)} \\ d_{(1)} \end{bmatrix}$$

vs

$$\begin{bmatrix} \bar{d}_{(p)} \\ \vdots \\ \bar{d}_{(2)} \\ \bar{d}_{(1)} \end{bmatrix}$$



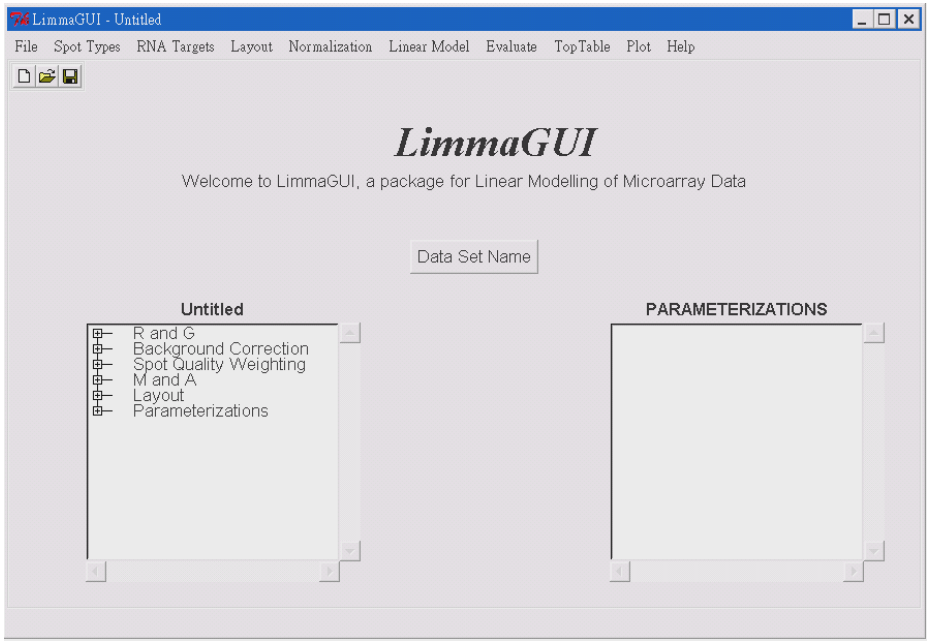
$d_{(i)} = \bar{d}_{(i)}$

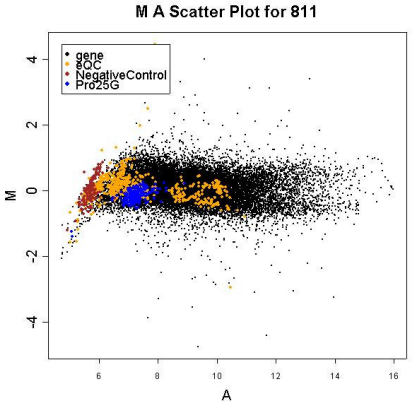
upper cut-point $\text{cut}_{up}(\Delta)$

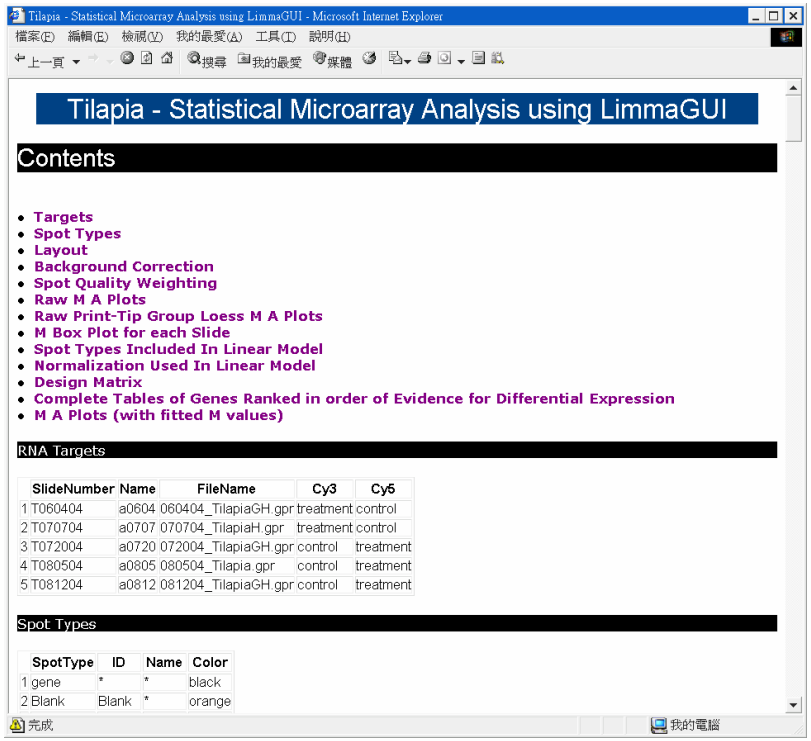
lower cut-point $\text{cut}_{low}(\Delta)$

Points for genes with evidence of repression

Software: Limma, LimmaGUI, affylmGUI







Limma: Linear Models for Microarray Data

http://bioinf.wehi.edu.au/limma/

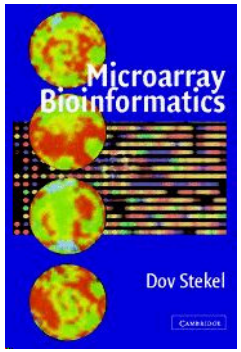
LimmaGUI: a menu driven interface of Limma

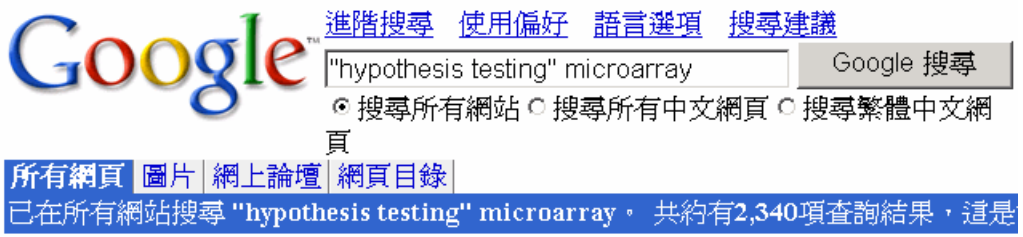
http://bioinf.wehi.edu.au/limmaGUI

- Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, Chapter 23. (To be published in 2005)
- Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3, No. 1, Article 3.

Reference

- Enfron, B. and Tibshirani, R. (1993). An introduction to the bootstrap. Chapman and Hall.
- Jarque, C. M. and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity, and serial independence of regression residuals. Economics Letters 6, 255-9.
- Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data, Journal of Computational Biology, 7: 819-837.
- Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown, The American Statistical Association Journal.
- Martinez, W. L. (2002). Computational statistics handbook with MATLAB, Boca Raton : Chapman & Hall/CRC.
- Runyon, R. P. (1977). Nonparametric statistics : a contemporary approach, Reading, Mass.: Addison-Wesley Pub. Co.
- Statistics Toolbox User's Guide, The MathWorks Inc. http://www.mathworks.com/access/helpdesk/help/toolbox/stats/stats.shtml
- Stekel, D. (2003). Microarray bioinformatics, New York : Cambridge University Press.
- Tsai, C. A., Chen, Y. J. and Chen, J. (2003). Testing for differentially expressed genes with microarray data, Nucleic Acids Research 31, No 9, e52.
- Turner, J. R. and Thayer, J. F. (2001). Introduction to analysis of variance : design, analysis, & interpretation, Thousand Oaks, Calif. : Sage Publications.

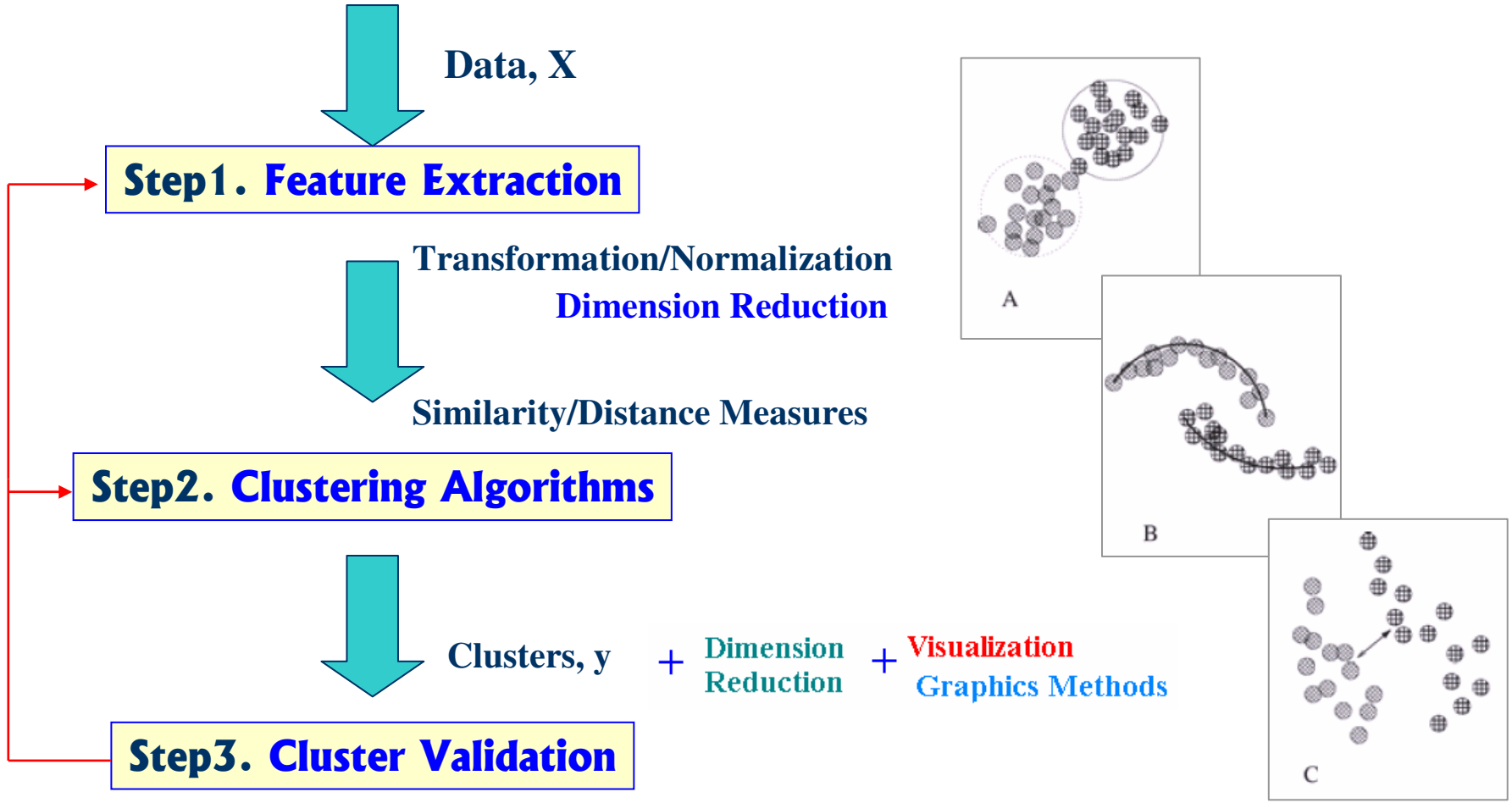




Clustering and Visualization

Cluster Analysis (Unsupervised Learning)

Group a given collection of unlabeled patterns into meaningful clusters.



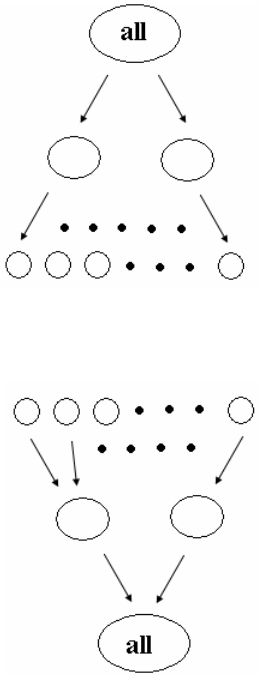
Daxin Jiang, Chun Tang and Aidong Zhang, (2004), **Cluster analysis for gene expression data: a survey**, IEEE Transactions on Knowledge and Data Engineering 16(11), 1370- 1386.

Clustering Analysis

Hierarchical clustering

The result is a tree that depicts the relationships between the objects.

- **Divisive clustering:**
  begin at step 1 with all the data in one cluster.
- **Agglomerative clustering:**
  all the objects start apart., there are n clusters at step 0.



Non-Hierarchical clustering

- k-means, The EM algorithm, K Nearest Neighbor,…

Two important properties of a clustering definition:

1. Most of data has been organized into **non-overlapping clusters**.
2. Each cluster has a within variance and one between variance for each of the other clusters. A good cluster should have a small within variance and large between variance.

Data/Information Visualization

What is Visualization?

- To visualize = to make visible, to transform into pictures.
- Making things/processes visible that are not directly accessible by the human eye.
- Transformation of an abstraction to a picture.
- Computer aided extraction and display of information from data.

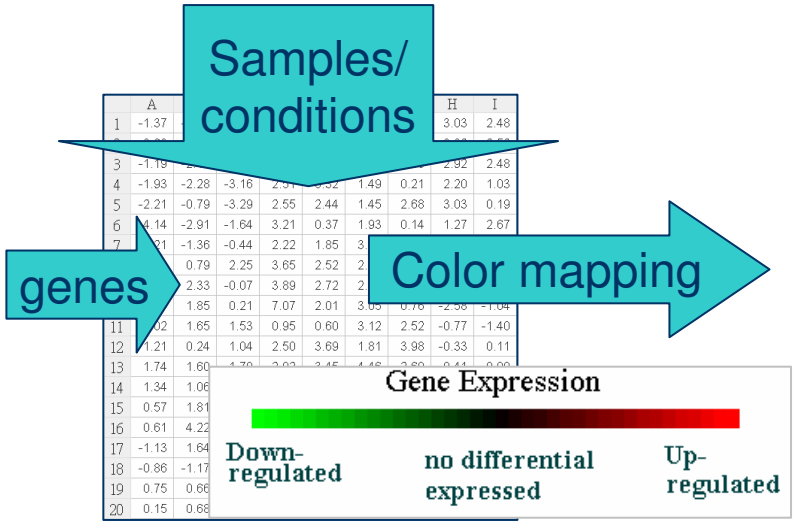
Data/Information Visualization

- Exploiting the human visual system to extract information from data.
- Provides an overview of complex data sets.
- Identifies structure, patterns, trends, anomalies, and relationships in data.
- Assists in identifying the areas of interest.

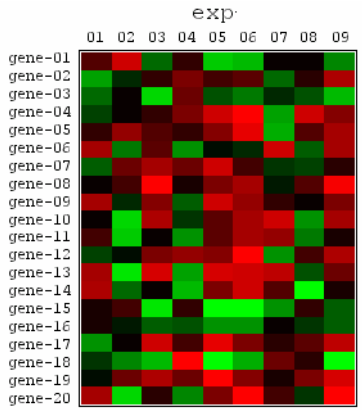
Visualization = Graphing for Data + Fitting + Graphing for Model

Tegarden, D. P. (1999). Business Information Visualization. Communications of AIS 1, 1-38.

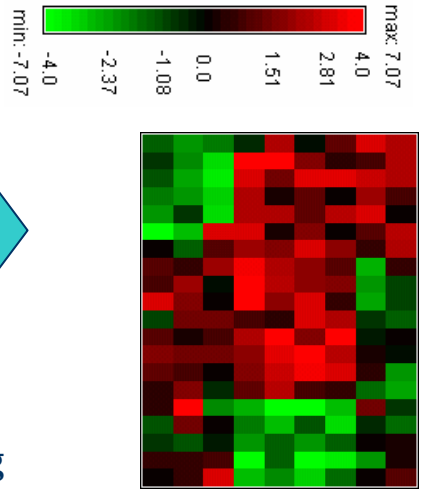
Visualizing Clustering Results: Heat Map



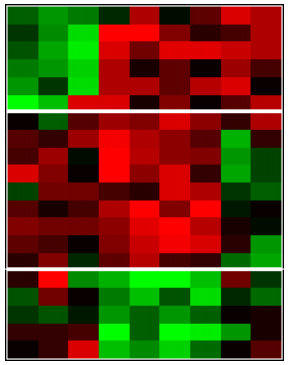
Without ordering



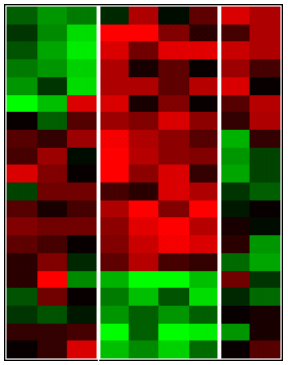
Ordering/ Seriation/ Clustering



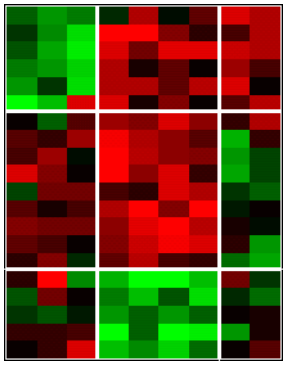
Gene-based clustering



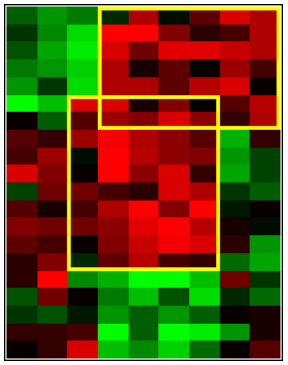
Sample-based clustering



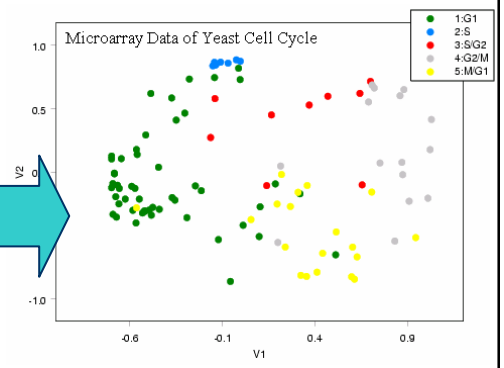
Twoway-based clustering



Subspace clustering



Dimension Reduction



e.g., K-means, SOM, Hierarchical Clustering, Model-based clustering,…

e.g., Bi-clustering

Clustering Analysis in Microarray Experiments

Goals

- Find natural classes in the data
- Identify new classes/gene correlations
- Refine existing taxonomies
- Support biological analysis/discovery

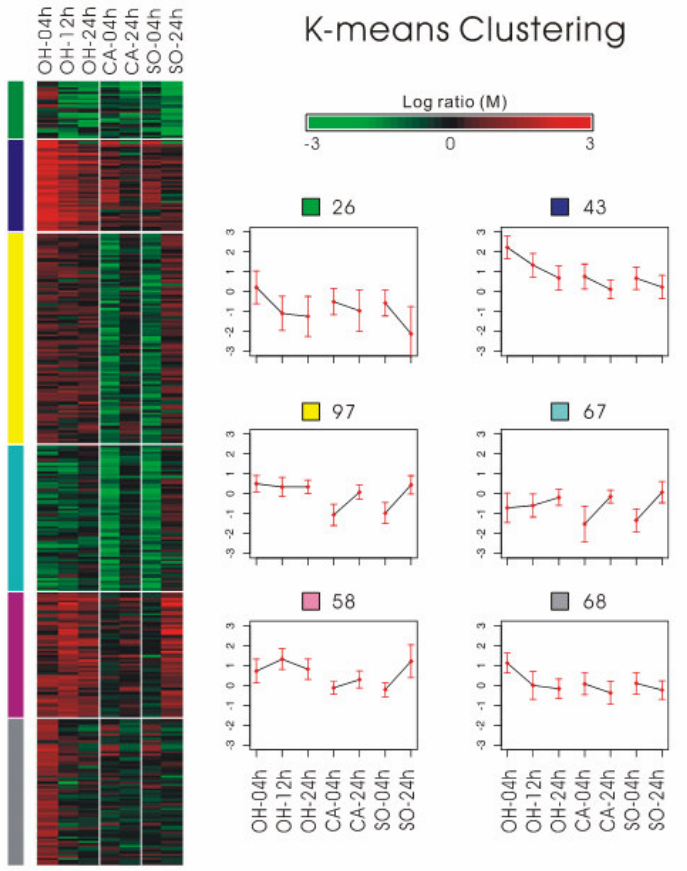
- cluster genes based on samples profiles
- cluster samples based on genes profiles

Hypothesis:

- genes with similar function have similar expression profiles.
- Clustering results in groups of co-expressed genes, groups of samples with a common phenotype, or blocks of genes and samples involved in specific biological processes.

Characteristic of Microarray Data:

- High-throughput, Noise, Outliers



Distance and Similarity Measure

Cov	x1	x2	x3	x4		xp
x1	1.00	0.48	0.10	-0.10		-0.28
x2	0.48	1.00	0.41	0.22		-0.23
x3	0.10	0.41	1.00	0.36		-0.05
x4	-0.10	0.22	0.36	1.00		0.10
xp	-0.28	-0.23	-0.05	0.10		1.00

Proximity Matrix

Data Matrix x y

Data	x1	x2	x3	x4	...	xp
subject01	-0.48	-0.42	0.87	0.92		-0.18
subject02	-0.39	-0.58	1.08	1.21		-0.33
subject03	0.87	0.25	-0.17	0.18		-0.44
subject04	1.57	1.03	1.22	0.31		-0.49
subject05	-1.15	-0.86	1.21	1.62		0.16
subject06	0.04	-0.12	0.31	0.16		-0.06
subject07	2.95	0.45	-0.40	-0.66		-0.38
subject08	-1.22	-0.74	1.34	1.50		0.29
subject09	-0.73	-1.06	-0.79	-0.02		0.44
subject10	-0.58	-0.40	0.13	0.58		0.02
subject11	-0.50	-0.42	0.66	1.05		0.06
subject12	-0.86	-0.29	0.42	0.46		0.10
subject13	-0.16	0.29	0.17	-0.28		-0.55
subject14	-0.36	-0.03	-0.03	-0.08		-0.25
subject15	-0.72	-0.85	0.54	1.04		0.24
subject16	-0.78	-0.52	0.26	0.20		0.48
subject17	0.60	-0.55	0.41	0.45		-0.66
⋮						
subject n	-2.29	-0.64	0.77	1.60		0.55
mean	0.07	-0.04	0.44	0.31	...	-0.21

$$x = (x_1, x_2, \cdots, x_n)$$

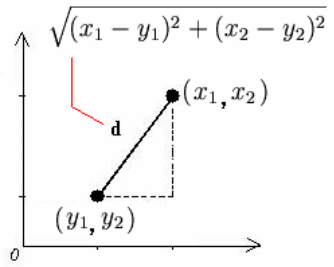
$$y = (y_1, y_2, \cdots, y_n)$$

Pearson Correlation Coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Euclidean Distance

$$d_{xy} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



- The standard transformation from a similarity matrix C to a distance matrix D is given by $d_{rs} = (c_{rr} - 2c_{rs} + c_{ss})^{1/2}$.
- (Eisen *et al.* 1998) $d_{rs} = 1 - c_{rs}$
- Other transformations (Chatfield and Collins 1980, Section 10.2)

K-Means Clustering

- K-means is a partition methods for clustering.
- Data are classified into k groups as specified by the user.
- Two different clusters cannot have any objects in common, and the k groups together constitute the full data set.

Optimization problem:

Minimize the sum of squared within-cluster distances

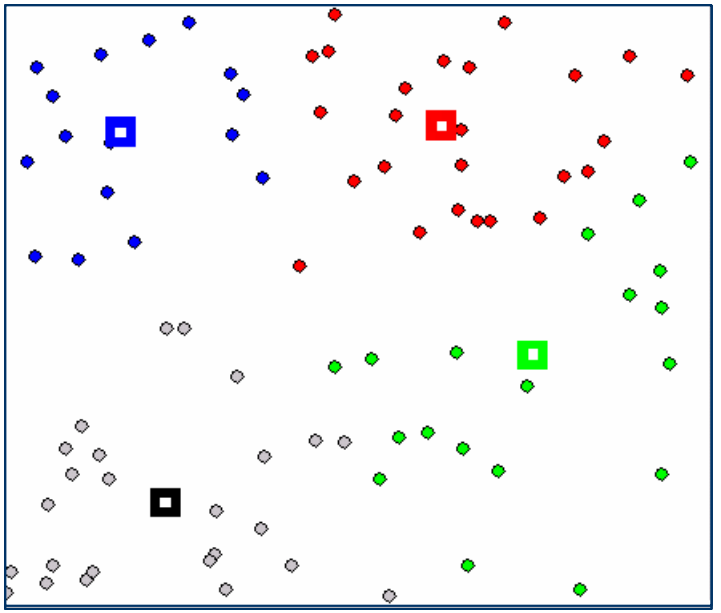
$$W(C) = \frac{1}{2}\sum_{k=1}^{K} \sum_{C(i)=C(j)=k} d_E(x_i, x_j)^2$$

The K-Means Algorithm

1. The data points are randomly assigned to one of the K clusters.
2. The position of the K centroids are determined (initial group centroids).
3. For each data point:
   - Calculate the distance from the data point to each cluster.
   - Assign data point to the cluster that has the closest centroid.
4. Repeat the above step until the centroids no longer move.

The choice of initial partition can greatly affect the final clusters that result.

Converged



Dimension Reduction

Visualizing Clustering Results

Dimension Reduction Techniques

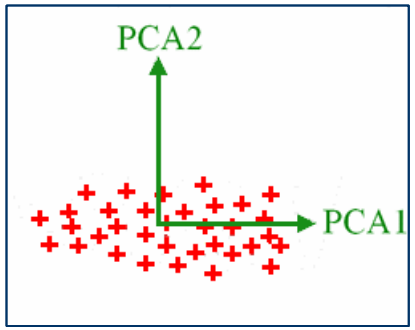
- **Principal Component Analysis (PCA)**
- **Multidimensional Scaling (MDS)**

Dimension reduction visualization is often adopted for presenting grouping structure for methods such as K-means.

Principal Component Analysis (PCA)

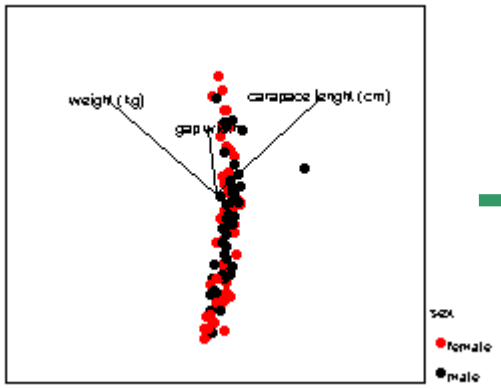
(Pearson 1901; Hotelling 1933; Jolliffe 2002)

PCA is a method that reduces data dimensionality by finding the new variables (major axes, principal components).



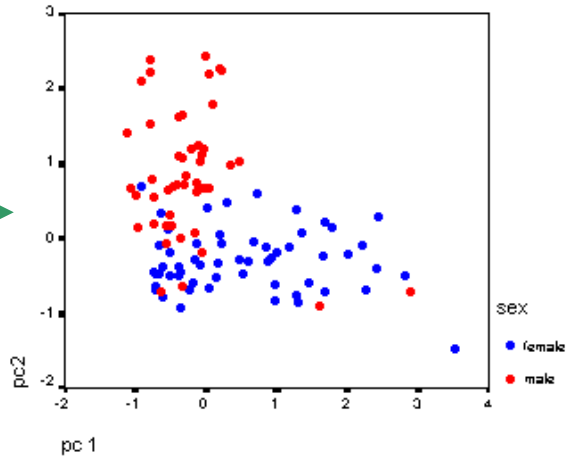
$$\mathrm{PCA}_1 = a_1\mathbf{X} + b_1\mathbf{Y}$$

$$\mathrm{PCA}_2 = a_2\mathbf{X} + b_2\mathbf{Y}$$



$$\mathrm{PCA}_1 = a_1\mathbf{X} + b_1\mathbf{Y} + c_1\mathbf{Z}$$

$$\mathrm{PCA}_2 = a_2\mathbf{X} + b_2\mathbf{Y} + c_2\mathbf{Z}$$



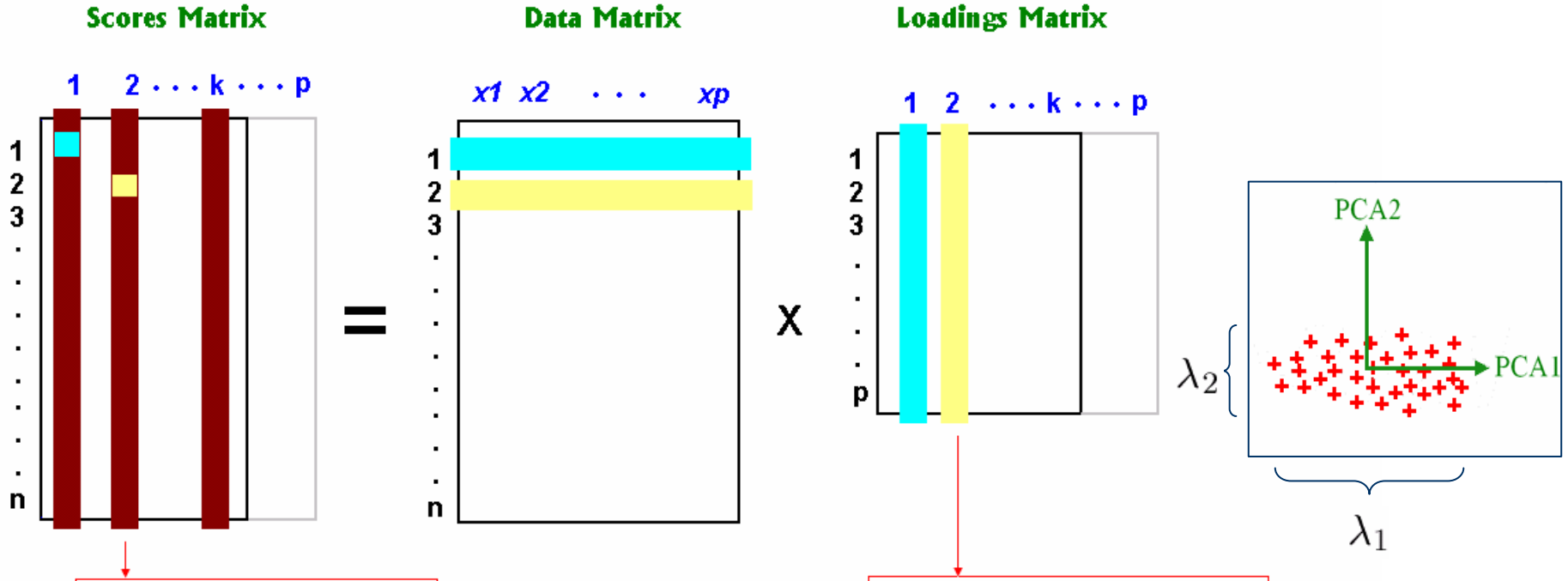
Amongst all possible projections, PCA finds the projections so that the maximum amount of information, measured in terms of variability, is retained in the smallest number of dimensions.

$$\mathrm{PCA}_1 = a_{11}\mathbf{X}_1 + a_{12}\mathbf{X}_2 + \cdots + a_{1p}\mathbf{X}_p$$

$$\mathrm{PCA}_2 = a_{21}\mathbf{X}_1 + a_{22}\mathbf{X}_2 + \cdots + a_{2p}\mathbf{X}_p$$

PCA: Loadings and Scores

$$\mathbf{Z} = \mathbf{X}\,\mathbf{W}$$



The ith principal component of $\mathbf{X}$ is $\mathbf{X}\mathbf{w}_i$, where $\mathbf{w}_i$ is the ith normalized eigenvector of $\Sigma_{\mathbf{x}}$ corresponding to the ith largest eigenvaules.

Eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$

$$\text{proportion} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

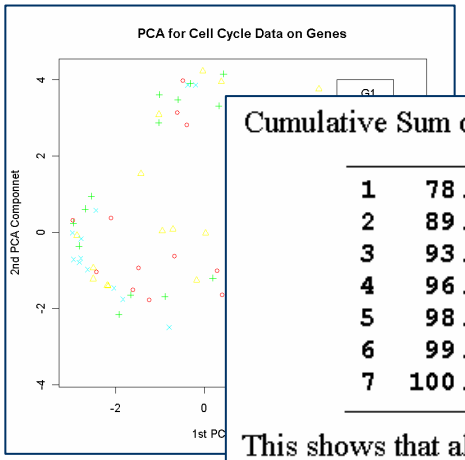
PCA (conti.)

Microarray Data Matrix

MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
gene001	-0.48	-0.42	0.87	0.92	0.67		-0.35
gene002	-0.39	-0.58	1.08	1.21	0.52		-0.58
gene003	0.87	0.25	-0.17	0.18	-0.13		-0.13
gene004	1.57	1.03	1.22	0.31	0.16		-1.02
gene005	-1.15	-0.86	1.21	1.62	1.12		-0.44
gene006	0.04	-0.12	0.31	0.16	0.17		0.08
gene007	2.95	0.45	-0.40	-0.66	-0.59		-0.76
gene008	-1.22	-0.74	1.34	1.50	0.63		-0.55
gene009	-0.73	-1.06	-0.79	-0.02	0.16		0.03
gene010	-0.58	-0.40	0.13	0.58	-0.09		-0.45
gene011	-0.50	-0.42	0.66	1.05	0.68		0.01
gene012	-0.86	-0.29	0.42	0.46	0.30		-0.63
gene013	-0.16	0.29	0.17	-0.28	-0.02		-0.04
gene014	-0.36	-0.03	-0.03	-0.08	-0.23		-0.21
gene015	-0.72	-0.85	0.54	1.04	0.84		-0.64
gene016	-0.78	-0.52	0.26	0.20	0.48		0.27
gene017	0.60	-0.55	0.41	0.45	0.18		-1.02
gene018	-0.20	-0.67	0.13	0.10	0.38		0.05
gene019	-2.29	-0.64	0.77	1.60	0.53		-0.38
gene020	-1.46	-0.76	1.08	1.50	0.74		-0.70
gene021	-0.57	0.42	1.03	1.35	0.64		-0.40
gene022	-0.11	0.13	0.41	0.60	0.23		0.19
gene•••							
gene n	-1.79	0.94	2.13	1.75	0.23		-0.66

PCA on Conditions

MA Table	PCA-1	PCA-2	PCA-3
gene001	-0.18	-0.11	-0.03
gene002	0.51	-0.53	0.54
gene003	-0.35	-0.39	0.26
gene004	-0.18	-1.08	0.41
gene005	-0.62	-0.8	0.13
gene006	-0.09	-0.23	0.77
gene007	-0.38	-0.32	1.08
gene008	-0.88	-0.55	1.03
gene009	-1.26	0.45	0.41
gene010	0.12	-0.36	-0.16
gene011	-0.28	-0.44	2.13
gene012	-0.45	-0.23	0.82
gene013	-0.2	-0.43	0.44
gene014	0.03	-0.26	-0.68
gene015	-0.7	-0.76	0.5
gene016	-0.61	0.07	-0.04
gene017	-0.23	-0.71	0.01
gene018	0.1	0.1	0.11
gene019	-0.94	-0.97	0.24
gene020	-0.55	-0.53	0.86
gene021	-0.47	-0.87	-0.02
gene022	-0.34	-1.1	0.51
gene•••	-0.49	-0.2	0.91
gene n	-0.15	-1.04	-0.01

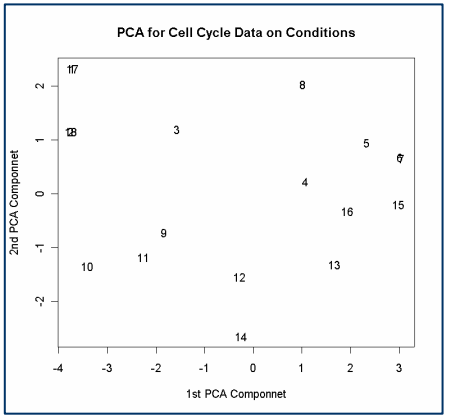


Cumulative Sum of the Variances:

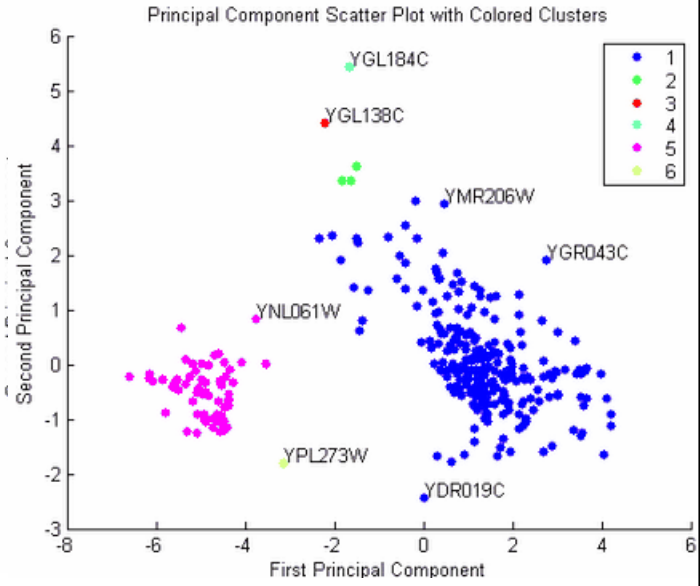
1	78.3719
2	89.2140
3	93.4357
4	96.0831
5	98.3283
6	99.3203
7	100.0000

This shows that almost 90% of the variance is accounted for by the first two principal components.

PCA on Genes



MA Table	exp01	exp02	exp03	exp04	exp05	exp•••	exp p
PCA-1	0.18	0.3	-0.12	-0.44	0.19	-0.39	-0.61
PCA-2	-0.16	-0.58	-0.43	-0.22	0.53	0.69	0.08
PCA-3	0.16	-0.44	-0.93	-1.23	-0.62	0.62	1.3



Yeast Microarray Data is from
DeRisi, JL, Iyer, VR, and Brown, PO.(1997). "Exploring the metabolic and genetic control of gene expression on a genomic scale"; Science, Oct 24;278(5338):680-6.

Multidimensional Scaling (MDS)

(Torgerson 1952; Cox and Cox 2001)

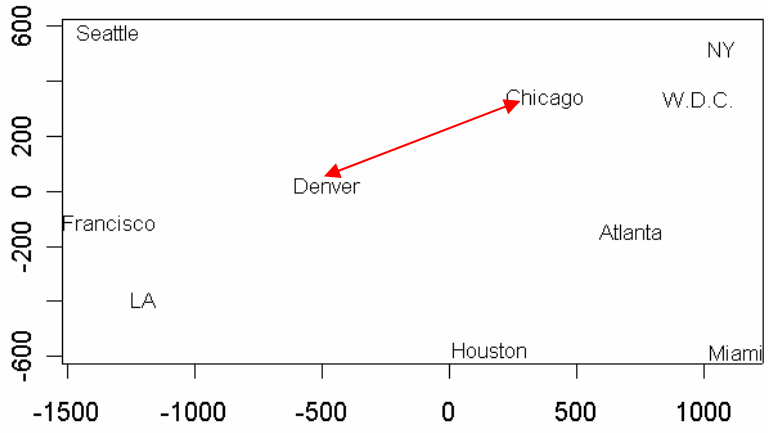
http://www.lib.utexas.edu/maps/united_states.html

- Classical MDS takes a set of dissimilarities and returns a set of points such that the distances between the points are approximately equal to the dissimilarities.
- projection from some unknown dimensional space to 2-d dimension.

Flying Mileages Between Ten U.S. Cities

0										Atlanta
587	0									Chicago
1212	920	0								Denver
701	940	879	0							Houston
1936	1745	831	1374	0						Los Angeles
604	1188	1726	968	2339	0					Miami
748	713	1631	1420	2451	1092	0				New York
2139	1858	949	1645	347	2594	2571	0			San Francisco
2182	1737	1021	1891	959	2734	2408	678	0		Seattle
543	597	1494	1220	2300	923	205	2442	2329	0	Washington D.C.

MDS



MDS: Metric and Non-Metric Scaling

Question

Given a *dissimilarity matrix* D of certain objects, can we construct points in k-dimensional (often 2-dimensional) space such that

Goal of metric scaling

the Euclidean distances between these points approximate the entries in the dissimilarity matrix?

Goal of non-metric scaling

the order in distances coincides with the order in the entries of the dissimilarity matrix approximately?

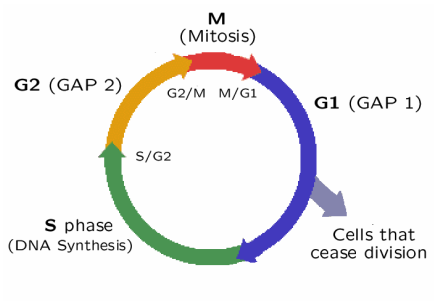
$$S = \sum_{i,j} (\hat{d}_{ij} - d_{ij})^2$$

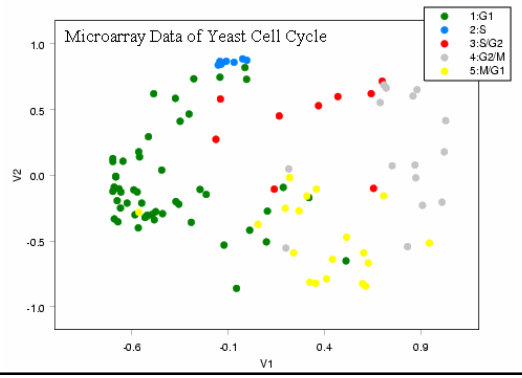
Mathematically: for given k, compute points $x_1,\ldots,x_n$ in k-dimensional space such that the object function is minimized.

$$Stress = \sqrt{\frac{\sum_{i,j} (\hat{d}_{ij} - d_{ij})^2}{\sum_{i,j} d_{ij}^2}}$$

Microarray Data of Yeast Cell Cycle

- **Synchronized by alpha factor arrest method (Spellman et al. 1998; Chu et al. 1998)**
- **103 known genes: every 7 minutes and totally 18 time points.**
- **2D MDS Configuration Plot for 103 known genes.**

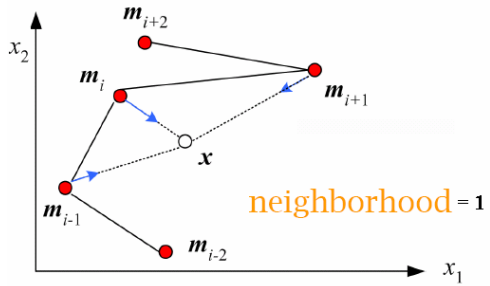




Clustering and Visualization

Self-Organizing Maps (SOM)

- SOMs were developed by **Kohonen** in the early **1980's**, original area was in the area of speech recognition.
- ***Idea:*** Organise data on the basis of **similarity** by putting entities **geometrically** close to each other.



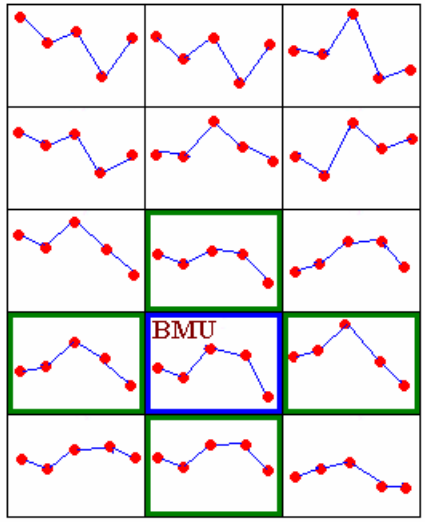
- SOM is unique in the sense that it combines both aspects. It can be used at the same time both to reduce the amount of data by clustering, and to construct a nonlinear projection of the data onto a low-dimensional display.

Step 0:
Initialize weights $\mathbf{w}_i(t)$.
Set $\alpha(t)$ and $h_{ci}(t)$.

Learning process:

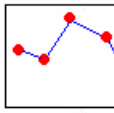
$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + h_{ci}(t)[\,\mathbf{x}(t) - \mathbf{w}_i(t)\,] & i \in N_c(t) \\ \mathbf{w}_i(t), \quad \text{o.w.} \end{cases}$$

5 x 3 output node



Data Matrix

Table	X01	X02	X03	•••	Xp
obs 001	-0.48	-0.42	0.87		-0.35
obs 002	-0.39	-0.58	1.08		-0.58
obs 003	0.87	0.25	-0.17		-0.13
obs 004	1.57	1.03	1.22		-1.02
obs 005	-1.15	-0.86	1.21		-0.44
obs 006	0.04	-0.12	0.31		0.08
obs 007	2.95	0.45	-0.40		-0.76
obs 008	-1.22	-0.74	1.34		-0.55
obs 009	-0.73	-1.06	-0.79		0.03
obs 010	-0.58	-0.40	0.13		-0.45
obs 011	-0.50	-0.42	0.66		0.01
obs 012	-0.86	-0.29	0.42		-0.63
obs 013	-0.16	0.29	0.17		-0.04
obs •••					
obs n	-1.79	0.94	2.13		-0.66



input node

Incrementally decrease the learning rate and the neighborhood size, and repeat

Algorithm of SOM

Step 0: Initialize weights $\mathbf{w}_i(t)$.
Set topological neighborhood parameters $N_c(t)$.
Set learning rate parameters $\alpha(t)$ and $h_{ci}(t)$.

Step 1: For each input vector $\mathbf{x}(t)$, do

a. Finding a BMU: $\|\mathbf{x}(t) - \mathbf{w}_c(t)\| = \min_i \|\mathbf{x}(t) - \mathbf{w}_i(t)\|$

b. Learning process:

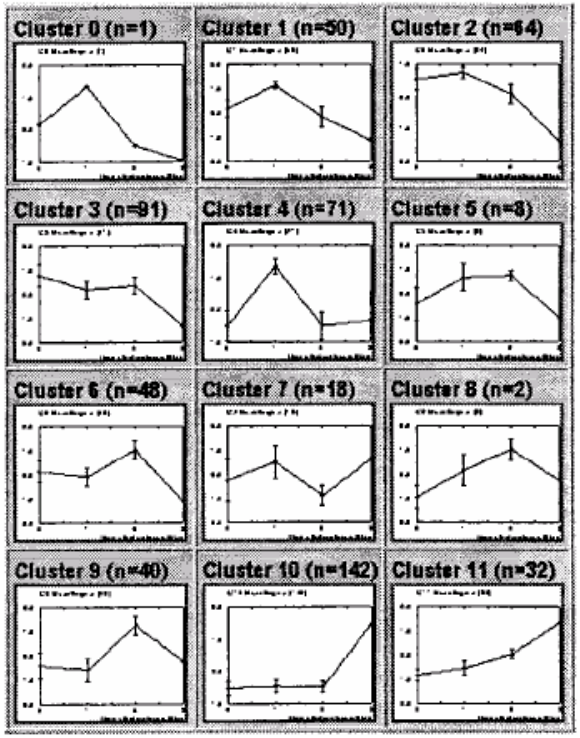
$$\mathbf{w}_i(t+1) = \begin{cases} \mathbf{w}_i(t) + h_{ci}(t)[\,\mathbf{x}(t) - \mathbf{w}_i(t)\,], & i \in N_c(t) \\ \mathbf{w}_i(t), & \text{o.w.} \end{cases}$$

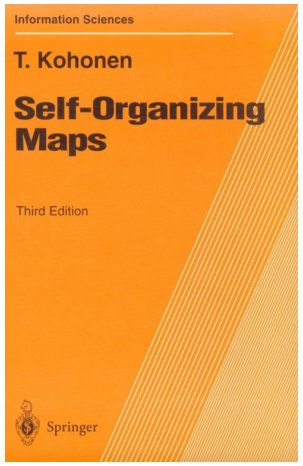
c. Go to the next unvisited input vector. If there are no unvisited input vector left then go back to the very first one and go to Step 2.

Step 2: Incrementally decrease the learning rate and the neighborhood size, and repeat Step 1.

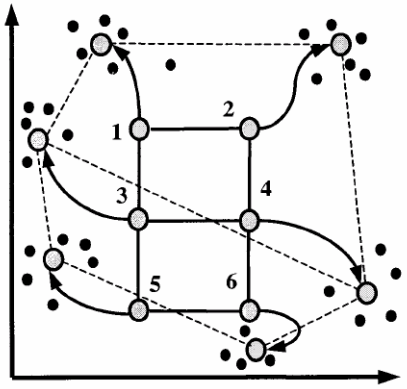
Step 3: Keep doing Steps 1 and 2 for a sufficient number of iterations.

HL-60 4×3 SOM 567 genes





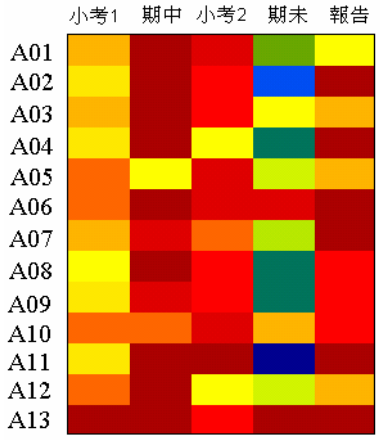
1995, 1997, 2001

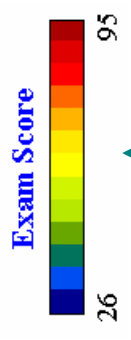


Macrophage Differentiation in HL-60 cells

Tamayo, P. et al. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc Natl Acad Sci* 96:2907-2912.

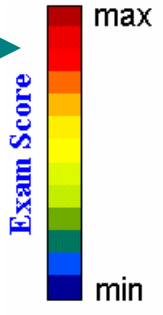
Heat Map: Data Image, Matrix Visualization

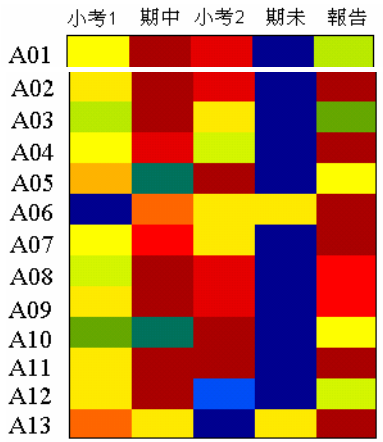




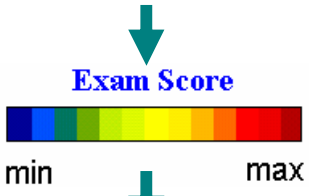
Range Matrix Condition

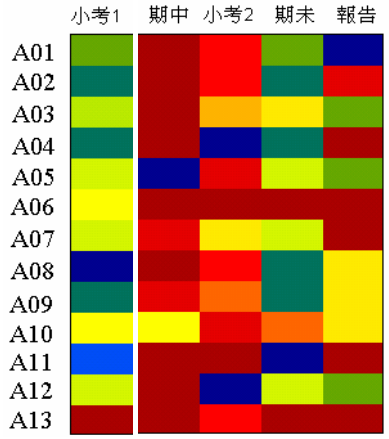
	A	B	C	D	E	F
1	學號	小考1	期中考	小考2	期末考	報告
2	A01	69	92	85	45	62
3	A02	66	90	83	36	90
4	A03	72	92	80	62	70
5	A04	68	90	60	37	95
6	A05	74	60	86	54	70
7	A06	77	90	88	88	95
8	A07	73	88	77	51	95
9	A08	61	90	84	40	82
10	A09	66	88	82	39	80
11	A10	76	75	87	72	80
12	A11	64	90	90	26	95
13	A12	75	90	60	55	70
14	A13	92	90	83	90	95





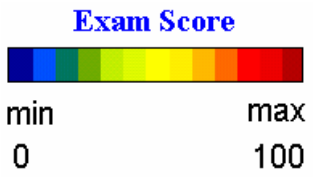
Range Raw Condition





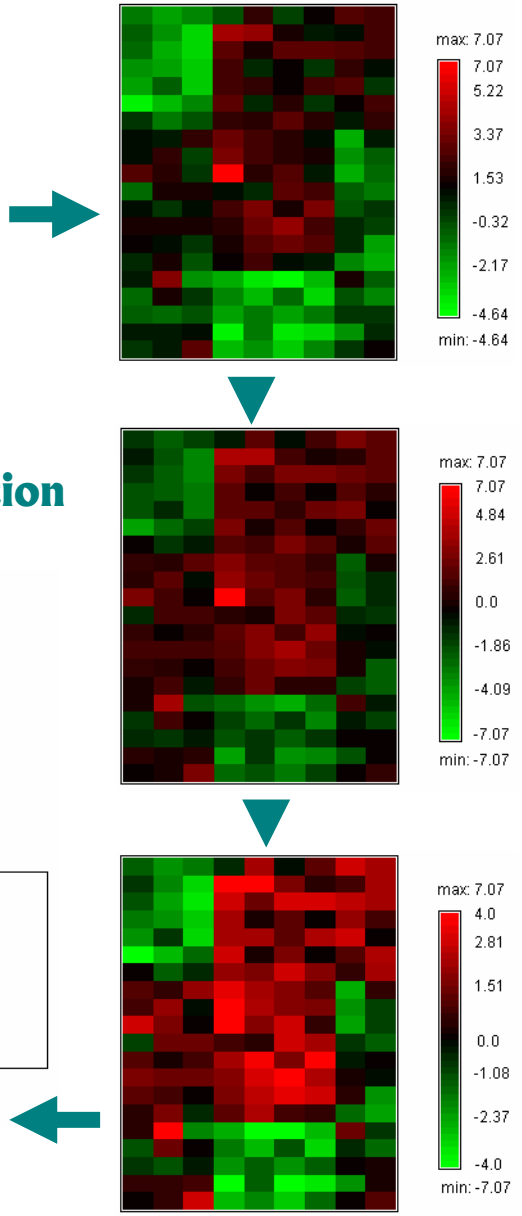
Range Column Condition

What about this one?

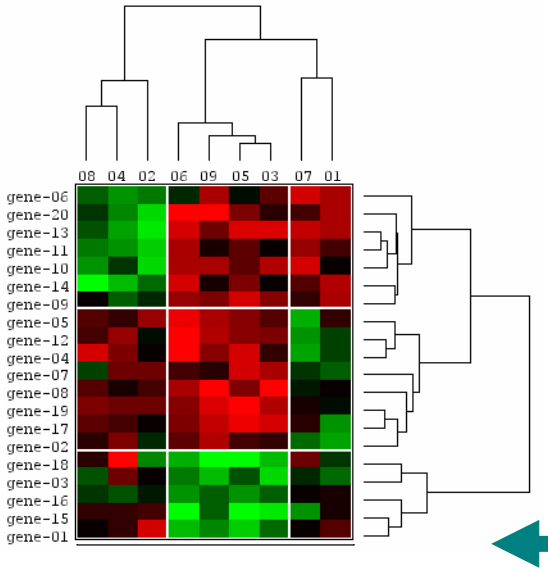


Heat Map: Display Conditions

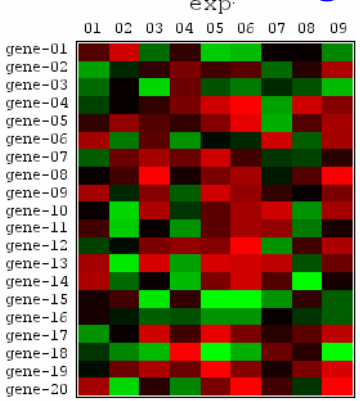
	A	B	C	D	E	F	G	H	I
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03	2.48
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92	2.52
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92	2.48
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20	1.03
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.68	3.03	0.19
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	0.14	1.27	2.67
7	0.21	-1.36	-0.44	2.22	1.85	3.11	2.03	0.67	2.40
8	1.13	0.79	2.25	3.65	2.52	2.09	1.13	-2.59	0.67
9	0.95	2.33	-0.07	3.89	2.72	2.13	1.75	-2.17	-0.90
10	3.04	1.85	0.21	7.07	2.01	3.05	0.76	-2.58	-1.04
11	-1.02	1.65	1.53	0.95	0.60	3.12	2.52	-0.77	-1.40
12	1.21	0.24	1.04	2.50	3.69	1.81	3.98	-0.33	0.11
13	1.74	1.60	1.70	2.02	3.45	4.46	2.69	0.41	-0.09
14	1.34	1.06	0.06	1.81	2.90	3.64	3.04	0.49	-2.33
15	0.57	1.81	-0.47	1.40	2.70	0.99	0.82	-1.61	-2.56
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64	-2.92	1.55	-0.71
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24	-3.41	-0.59	-1.64
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37	-1.41	0.08	0.25
19	0.75	0.66	1.04	-4.26	-1.41	-3.99	-3.53	-2.17	0.34
20	0.15	0.68	3.18	-2.86	-2.01	-3.18	-1.58	0.10	1.28



Center Matrix Condition

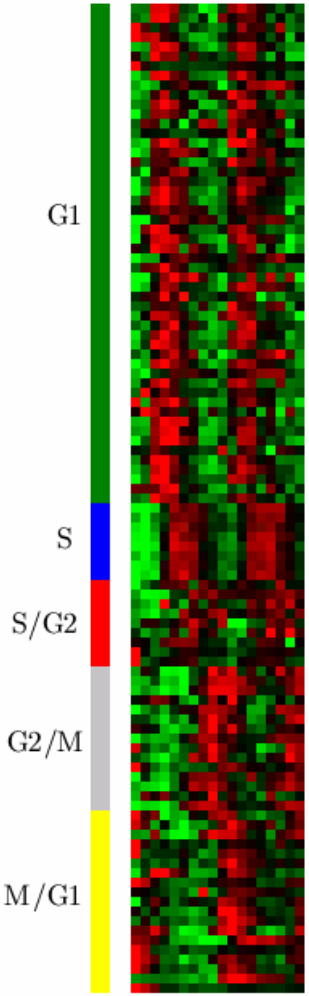


Without ordering



Microarray Data of Yeast Cell Cycle

- Synchronized by alpha factor arrest method (Spellman et al. 1998; Chu et al. 1998)
- 103 known genes: every 7 minutes and totally 18 time points.



K-Means Clustering

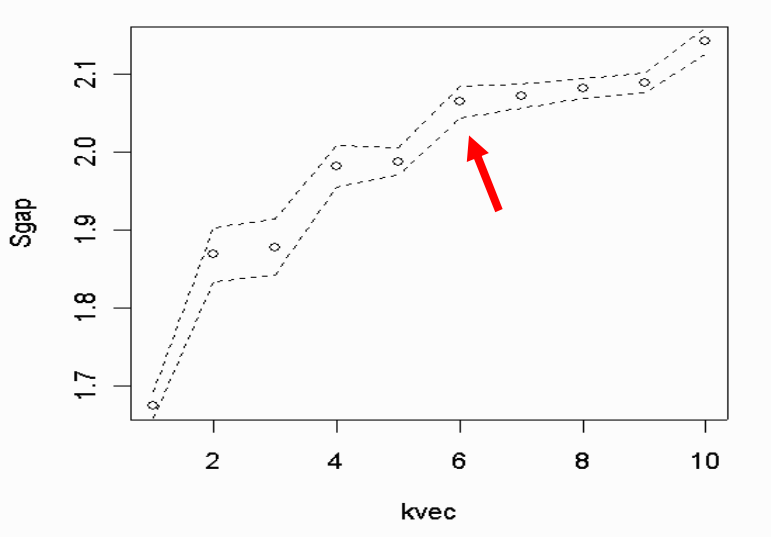
- Data

Baseline: Culture Medium (CM-00h)

OH-04h, OH-12h, OH-24h

CA-04h, CA-24h

SO-04h, SO-24h

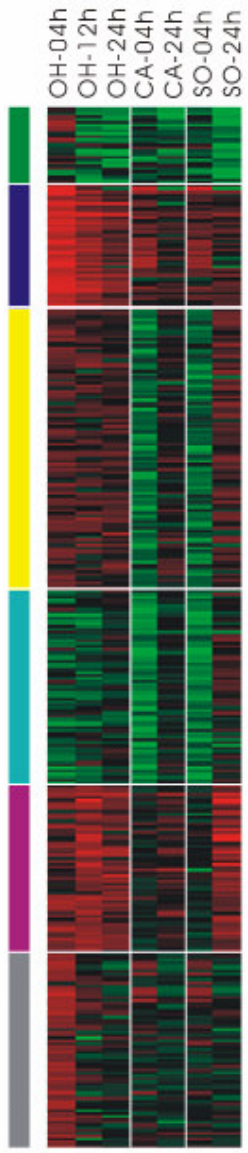


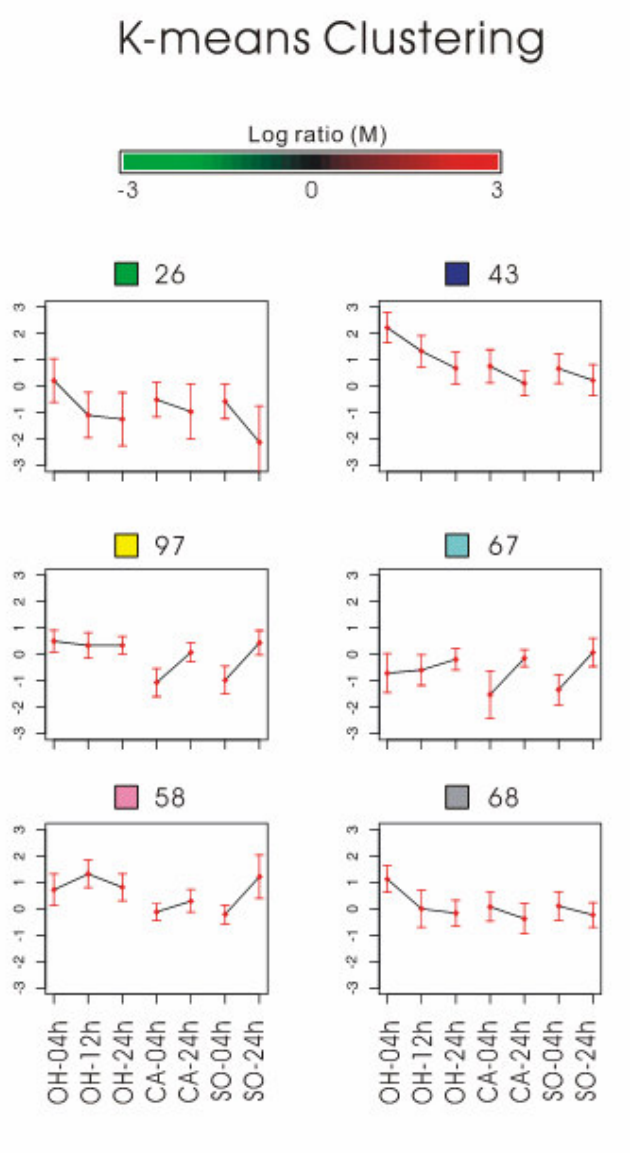
J. R. Statist. Soc. B (2001)
63, *Part* 2, *pp.* 411–423

Estimating the number of clusters in a data set via the gap statistic

Robert Tibshirani, Guenther Walther and Trevor Hastie

Stanford University, USA





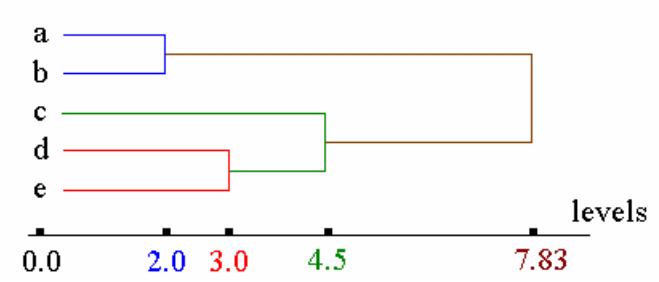
Hierarchical Clustering and Dendrogram

(Kaufman and Rousseeuw, 1990)

Example: Average-Linkage

distance matrix

	a	b	c	d	e
a	0	2	6	10	9
b		0	5	9	8
c			0	4	5
d				0	3
e					0



	{a, b}	c	d	e
{a, b}	0	5.5	9.5	8.5
c		0	4	5
d			0	3
e				0

$$D(\{a,b\},\{c\}) = \frac{1}{2}[D(a,c) + D(b,c)]$$

$$= \frac{1}{2}(6+5) = 5.5$$

	{a, b}	c	{d, e}
{a, b}	0	5.5	9.0
c		0	4.5
{d, e}			0

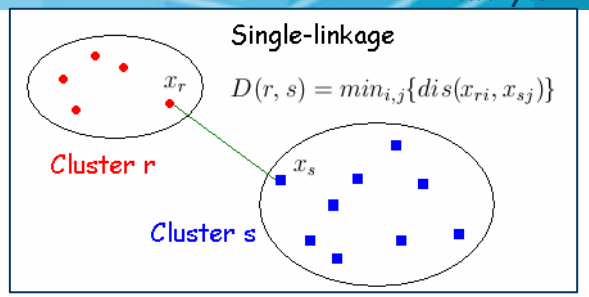
$$D(\{a,b\},\{d,e\})$$

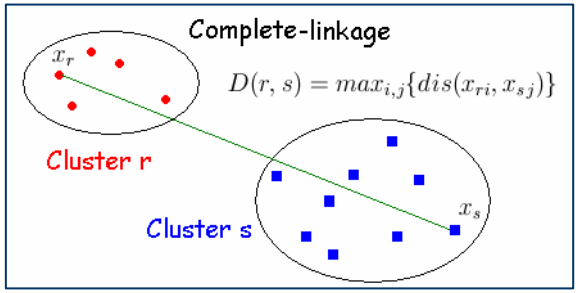
$$= \frac{1}{4}[D(a,d)+D(a,e)+D(b,d)+D(b,e)]$$

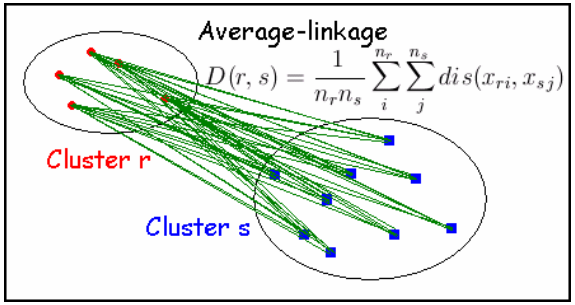
$$= \frac{1}{4}(10+9+9+8) = 9$$

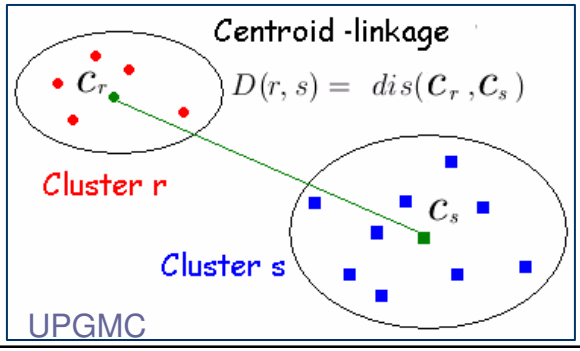
	{a, b}	{c, d, e}
{a, b}	0	7.83
{c, d, e}		0

UPGMA (Unweighted Pair-Groups Method Average)

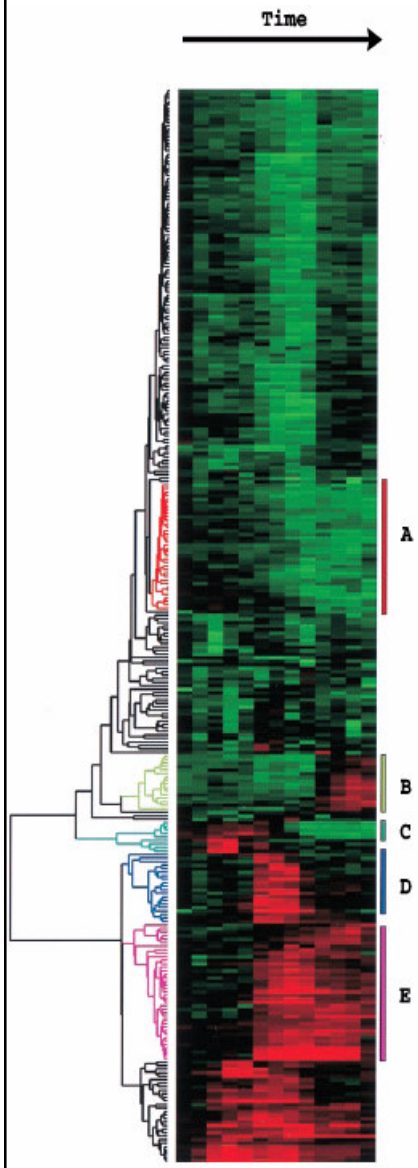








Display of Genome-Wide Expression Patterns



Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

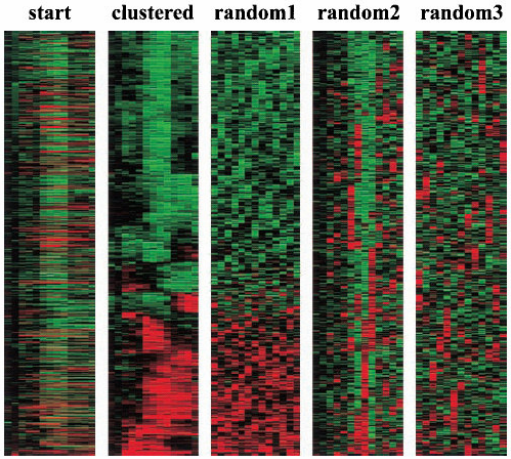
Cluster analysis and display of genome-wide expression patterns

MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

FIG. 1. Clustered display of data from time course of serum stimulation of primary human fibroblasts. Experimental details are described elsewhere (11). Briefly, foreskin fibroblasts were grown in culture and were deprived of serum for 48 hr. Serum was added back and samples taken at time 0, 15 min, 30 min, 1 hr, 2 hr, 3 hr, 4 hr, 8 hr, 12 hr, 16 hr, 20 hr, 24 hr. The final datapoint was from a separate unsynchronized sample. Data were measured by using a cDNA microarray with elements representing approximately 8,600 distinct human genes. All measurements are relative to time 0. Genes were selected for this analysis if their expression level deviated from time 0 by at least a factor of 3.0 in at least 2 time points. The dendrogram and colored image were produced as described in the text; the color scale ranges from saturated green for log ratios −3.0 and below to saturated red for log ratios 3.0 and above. Each gene is represented by a single row of colored boxes; each time point is represented by a single column. Five separate clusters are indicated by colored bars and by identical coloring of the corresponding region of the dendrogram. As described in detail in ref. 11, the sequence-verified named genes in these clusters contain multiple genes involved in (*A*) cholesterol biosynthesis, (*B*) the cell cycle, (*C*) the immediate–early response, (*D*) signaling and angiogenesis, and (*E*) wound healing and tissue remodeling. These clusters also contain named genes not involved in these processes and numerous uncharacterized genes. A larger version of this image, with gene names, is available at http://rana.stanford.edu/clustering/serum.html.

Software:
Cluster and TreeView

FIG. 3. To demonstrate the biological origins of patterns seen in Figs. 1 and 2, data from Fig. 1 were clustered by using methods described here before and after random permutation within rows (random 1), within columns (random 2), and both (random 3).



Cluster Validation

Cluster Validation

Assess the quality and reliability of the cluster sets.

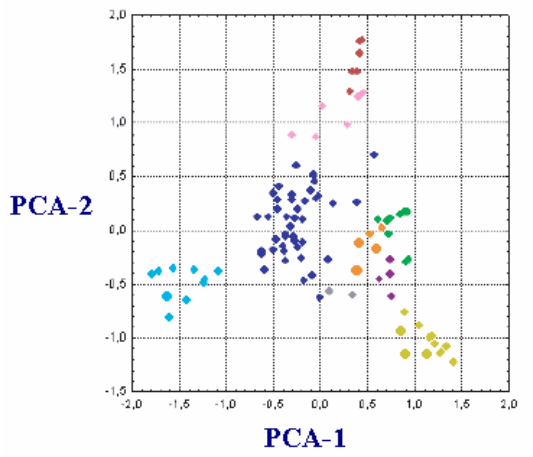
- **Quality: clusters can be measured in terms of homogeneity and separation.**
- **Reliability: cluster structure is not formed by chance.**
- **Ground Truth: from domain knowledge.**

NOTE:
Help to decide the number of clusters in the data.

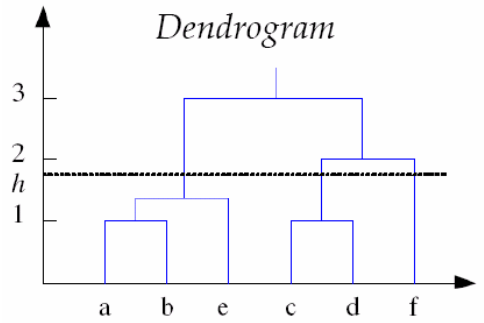
Choosing the Number of Clusters

(1) K is defined by the application.

(2) Plot the data in two PAC dimensions.

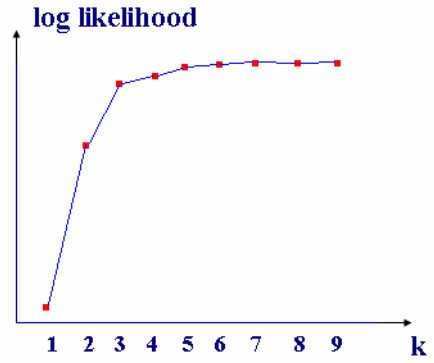


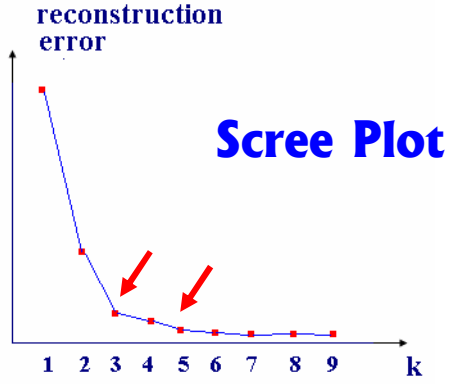
(4) Hierarchical clustering: look at the difference between levels in the tree.



(e.g., k-means: within-cluster sum of squares)

(3) Plot the **reconstruction error** or log likelihood as a function of k, and look for the elbow.





Scree Plot

Calinski and Harabasz (1974): CH(k)
Hartigan (1975): $H(k)$
Krzanowski and Lai (1985): KL(k)
Kaufman and Rousseeuw (1990): $s(i)$

J. R. Statist. Soc. B (2001)
63, *Part* 2, *pp.* 411–423

Estimating the number of clusters in a data set via the gap statistic

Robert Tibshirani, Guenther Walther and Trevor Hastie
Stanford University, USA

Literatures on Cluster Validation

2007

- Marcel Brun, Chao Sima, Jianping Hua, James Lowey, Brent Carroll, Edward Suh and Edward R. Dougherty, (2007), Model-based evaluation of clustering validation measures, Pattern Recognition 40(3), 807-824.
- Francisco R. Pinto, João A. Carriço, Mário Ramirez and Jonas S Almeida, (2007), Ranked Adjusted Rand: integrating distance and partition information in a measure of clustering agreement, BMC Bioinformatics, 8:44.

2006

- Susmita Datta and Somnath Datta, (2006), Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes, BMC Bioinformatics 2006, 7:397. [web]
- Anbupalam Thalamuthu, Indranil Mukhopadhyay, Xiaojing Zheng and George C. Tseng, (2006), Evaluation and comparison of gene clustering methods in microarray analysis, Bioinformatics 22(19), 2405-2412.
- Giorgio Valentini , (2006), Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data, Bioinformatics, 22(3), 369-370.
- Susmita Datta and Somnath Datta, (2006), Evaluation of clustering algorithms for gene expression data, BMC Bioinformatics 2006, 7(Suppl 4):S17. [web]

2005

- Tibshirani, Robert; Walther, Guenther (2005), Cluster Validation by Prediction Strength, Journal of Computational & Graphical Statistics 14(3), pp. 511-528(18)
- Julia Handl, Joshua Knowles and Douglas B. Kell, (2005), Computational cluster validation in post-genomic data analysis, Bioinformatics 21(15), 3201-3212. [web] [supp]
- Nadia B,Francisco A,Padraig C. (2005), An integrated tool for microarray data clustering and cluster validity assessment, Bioinformatics 21:451. [Web]
- Julia Handl and Joshua Knowles, (2005) Exploiting the trade-off -- the benefits of multiple objectives in data clustering, Proceedings of the Third International Conference on E
- Nikhil R Garge, C ithier? BMC Bioinformatics 2

More than 30 papers for Microarray!

2004

- Daxin Jiang, Chun Tang and Aidong Zhang, (2004), Cluster analysis for gene expression data: a survey, IEEE Transactions on Knowledge and Data Engineering 16(11), 1370- 1386. [web]
- Kimberly D. Siegmund, Peter W. Laird and Ite A. Laird-Offringa, (2004), A comparison of cluster analysis methods using DNA methylation data, Bioinformatics 20(12), 1896-1904.
- Tilman Lange, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann, Stability-Based Validation of Clustering Solutions, Neural Comp. 2004 16: 1299-1323.

2003

- Datta S, Datta S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics. 2003 Mar 1;19(4):459-66.
- N. Bolshakova and F. Azuaje, (2003), Cluster validation techniques for genome expression data, Signal Processing 83(4), 825-833.

2001

- K. Y. Yeung, D. R. Haynor and W. L. Ruzzo, (2001), Validating clustering for gene expression data, Bioinformatics 17(4), 309-318. [web]
- Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis,(2001), On Clustering Validation Techniques, Journal of Intelligent Information Systems, 17(2), 107 - 145.
- Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci U S A. 2001 Jul 31;98(16):8961-5.
- Levine E, Domany E. Resampling method for unsupervised estimation of cluster validity. Neural Comput. 2001 Nov;13(11):2573-93.
- Maria Halkidi, Michalis Vazirgiannis, Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set, icdm, p. 187, First IEEE International Conference on Data Mining (ICDM'01), 2001

~2000

- Zhang K, Zhao H. Assessing reliability of gene clusters from gene expression data. Funct Integr Genomics. 2000 Nov;1(3):156-73.
- Xie, X.L. Beni, G. (1991), A validity measure for fuzzy clustering, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 13(8), 841-847.
- Peter Rousseeuw, (1987), Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20(1), 53-65.
- Lawrence Hubert and Phipps Arabie (1985), Comparing partitions, Journal of Classification 2(1), 193-218.
- Wallace, D. L. 1983. A method for comparing two hierarchical clusterings: comment. Journal of the American Statistical Association 78:569-576.
- E. B. Fowlkes; C. L. Mallows, (1983), A Method for Comparing Two Hierarchical Clusterings, Journal of the American Statistical Association, 78(383), 553-569.
- William M. Rand, (1971), Objective Criteria for the Evaluation of Clustering Methods, Journal of the American Statistical Association 66(336), 846-850.

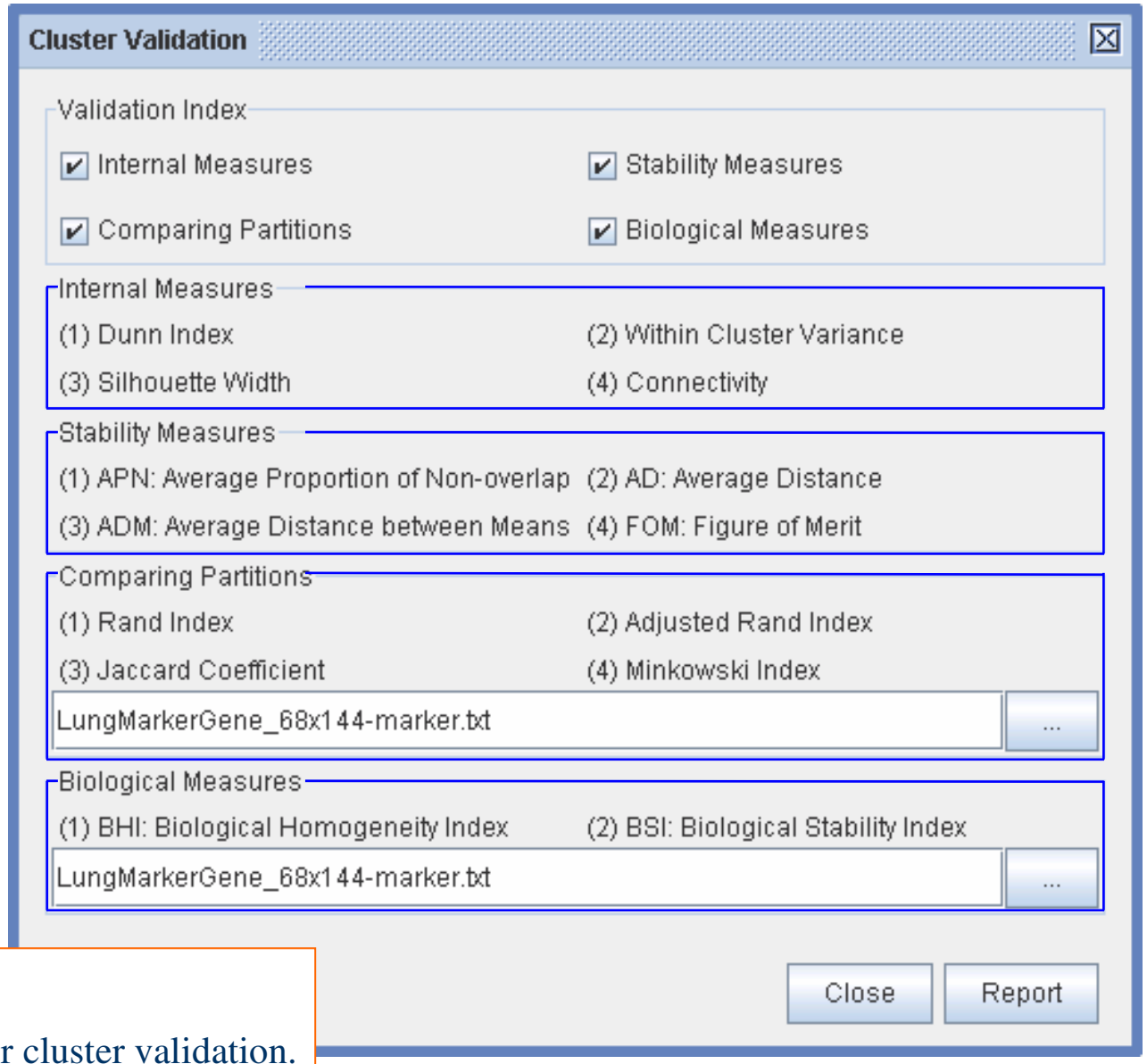
Cluster Validation Index

Internal Measures

Stability Measures

Comparing Partitions

Biological Measures



See also

clValid: an R package for cluster validation.

Biological Evaluation

- Biological Homogeneity Index (BHI)
- Biological Stability Index (BSI)

ProbeSet	Clustering	GO-BP Category
38389_at	1	0
1662_r_at	1	0
32607_at	1	0
1582_at	1	0
34699_at	1	0
37890_at	2	0
36008_at	2	1 2 3
36591_at	2	1 2 3 8 10
32081_at	2	1 2 3 4 5 6 7 9 10
668_s_at	2	1 2 3
41535_at	2	1 2 3 4
37666_at	2	1 2 3
40310_at	2	1 2 3 4 5 8 9
34256_at	3	1 2 3
38790_at	3	1
39175_at	3	1 2 3
35819_at	3	1 8
37639_at	3	1 2 3
31508_at	3	1 9
31505_at	4	1 2 3
1882_g_at	4	1 2 3 4 6
33154_at	4	1 2 3
837_s_at	4	1 2 3
35194_at	4	1
38422_s_at	4	1 2 3 4 5
33131_at	4	1 2 3 4 6 7

Example:
GO (Gene Ontology)
Multiple Functional Categories

Susmita Datta and Somnath Datta, (2006), Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes, BMC Bioinformatics 7:397.

Biological Evaluation: Homogeneity

Biological Homogeneity Index (BHI)

- $\mathcal{B} = \{B_1, \ldots, B_F\}$: a set of F functional classes, not necessarily disjoint,
- B^i: the functional class containing gene i (with possibly more than one functional class containing i).
- B^j: the function class containing gene j,
- $\mathrm{I}(B^i = B^j) = \begin{cases} 1, & \text{if } B^i \text{ and } B^j \text{ match}, \\ 0, & \text{otherwise.} \end{cases}$
- Given statistical clustering partition $\mathcal{C} = \{C_1, \ldots, C_K\}$ and set of biological classes $\mathcal{B} = \{B_1, \ldots, B_F\}$, the BHI is defined as

$$BHI(\mathcal{C}, \mathcal{B}) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k(n_k - 1)} \sum_{i \neq j; i,j \in C_k} \mathrm{I}\left(B^i = B^j\right).$$

- $n_k = n(C_k \cap \mathcal{B})$: the number of annotated genes in statistical cluster C_k.
- Range: $[0, 1]$, maximum.

Biological Evaluation: Stability

Biological Stability Index (BSI)

- The BSI is defined as

$$BSI(\mathcal{C},\mathcal{B}) = \frac{1}{F}\sum_{k=1}^{F}\frac{1}{n(B_k)(n(B_k)-1)}\frac{1}{M}\sum_{\ell=1}^{M}\sum_{i\neq j; i,j\in B_k}\frac{n(C^{i,0}\cap C^{j,\ell})}{n(C^{i,0})},$$

- $C^{i,0}$: the statistical cluster containing observation i based on all the data.
- $C^{j,\ell}$: the statistical cluster containing observation j when column ℓ is removed.
- Range $[0,1]$: maximum.

	A	B	C	D	E	F	G	H	I
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03	2.48
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92	2.52
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92	2.48
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20	1.03
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.68	3.03	0.13
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	0.14	1.27	2.67
7	0.21	-1.36	-0.44	2.22	1.85	3.11	2.03	0.67	2.40
8	1.13	0.79	2.25	3.65	2.52	2.09	1.13	-2.59	0.67
9	0.95	2.33	-0.07	3.89	2.72	2.13	1.75	-2.17	-0.90
10	3.04	1.85	0.21	7.07	2.01	3.05	0.76	-2.58	-1.04
11	-1.02	1.65	1.53	0.95	0.60	3.12	2.52	-0.77	-1.40
12	1.21	0.24	1.04	2.50	3.69	1.81	3.98	-0.33	0.11
13	1.74	1.60	1.70	2.02	3.45	4.46	2.69	0.41	-0.09
14	1.34	1.06	0.06	1.81	2.90	3.64	3.04	0.49	-2.33
15	0.57	1.81	-0.47	1.40	2.70	0.99	0.82	-1.61	-2.56
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64	-2.92	1.55	-0.71
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24	-3.41	-0.59	-1.64
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37	-1.41	0.08	0.25
19	0.75	0.66	1.04	-4.26	-1.41	-3.99	-3.53	-2.17	0.34
20	0.15	0.68	3.18	-2.86	-2.01	-3.18	-1.58	0.10	1.28

Full data (nxp)

Compare two clusterings

Repeat: 1,…p

	A	B	C	D	E	F	G	H
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.68	3.03
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	0.14	1.27
7	0.21	-1.36	-0.44	2.22	1.85	3.11	2.03	0.67
8	1.13	0.79	2.25	3.65	2.52	2.09	1.13	-2.59
9	0.95	2.33	-0.07	3.89	2.72	2.13	1.75	-2.17
10	3.04	1.85	0.21	7.07	2.01	3.05	0.76	-2.58
11	-1.02	1.65	1.53	0.95	0.60	3.12	2.52	-0.77
12	1.21	0.24	1.04	2.50	3.69	1.81	3.98	-0.33
13	1.74	1.60	1.70	2.02	3.45	4.46	2.69	0.41
14	1.34	1.06	0.06	1.81	2.90	3.64	3.04	0.49
15	0.57	1.81	-0.47	1.40	2.70	0.99	0.82	-1.61
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64	-2.92	1.55
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24	-3.41	-0.59
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37	-1.41	0.08
19	0.75	0.66	1.04	-4.26	-1.41	-3.99	-3.53	-2.17
20	0.15	0.68	3.18	-2.86	-2.01	-3.18	-1.58	0.10

left-out column ℓ

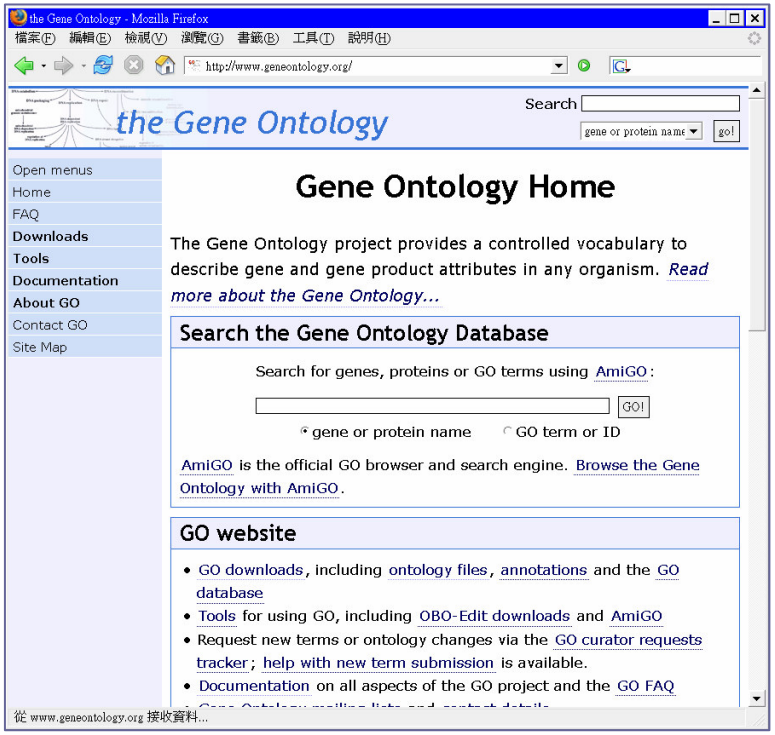
sample

Remaining data (nx(p-1))

Obtain Functional Categories (Annotation)

MIPS: the Munich Information Center for Protein Sequences

- http://mips.gsf.de/
- MIPS: a database for protein sequences and complete genomes, Nucleic Acids Research, 27:44-48, 1999



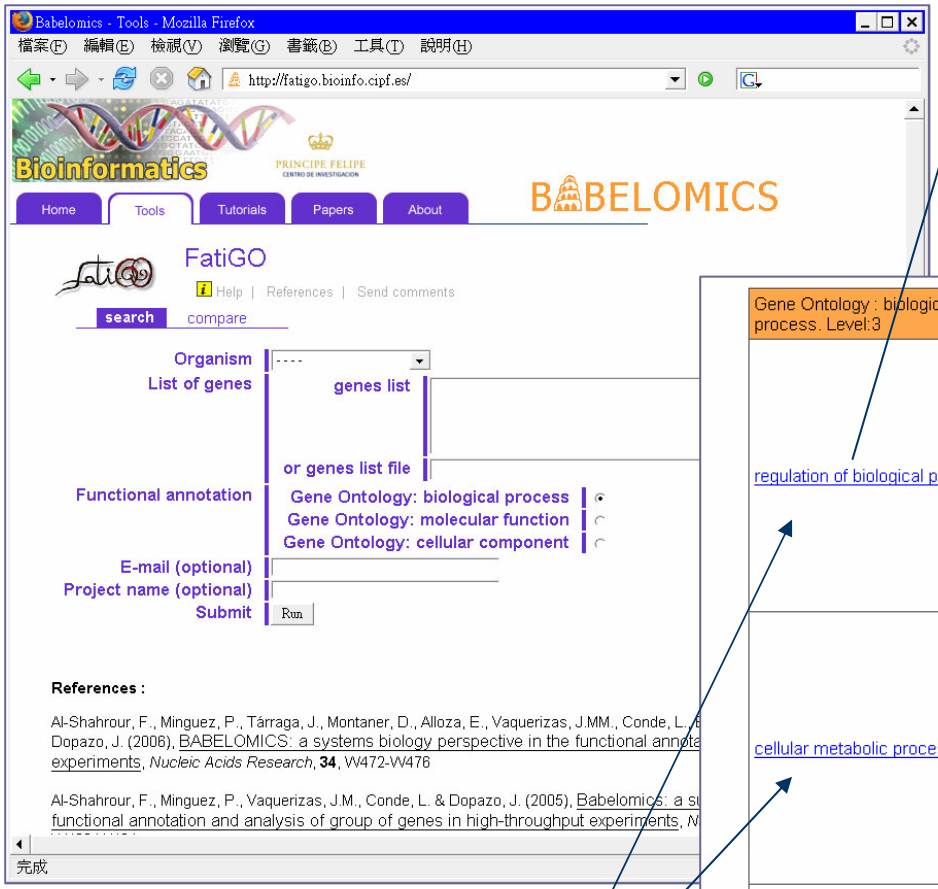
GO: Gene Ontology

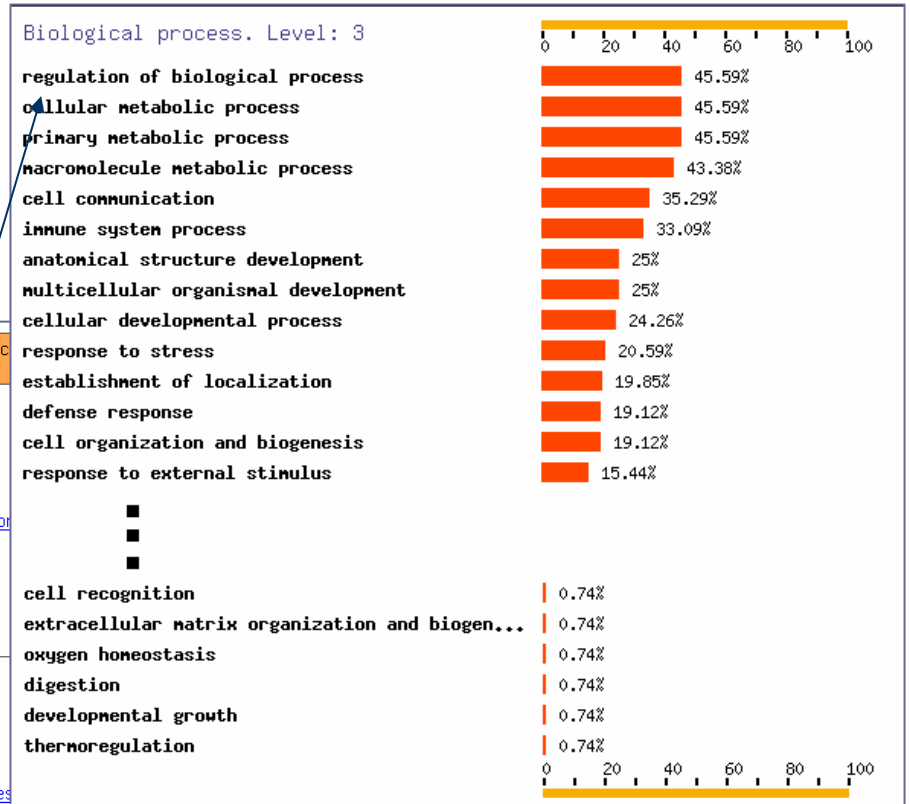
- A GO annotation is a Gene Ontology term associated with a gene product.
- http://www.geneontology.org/
- The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet. (2000) 25: 25-29.

- FatiGO (Al-Shahrour et al., 2004)
- FunCat (Ruepp et al., 2004)

FatiGO

http://babelomics.bioinfo.cipf.es/index.html



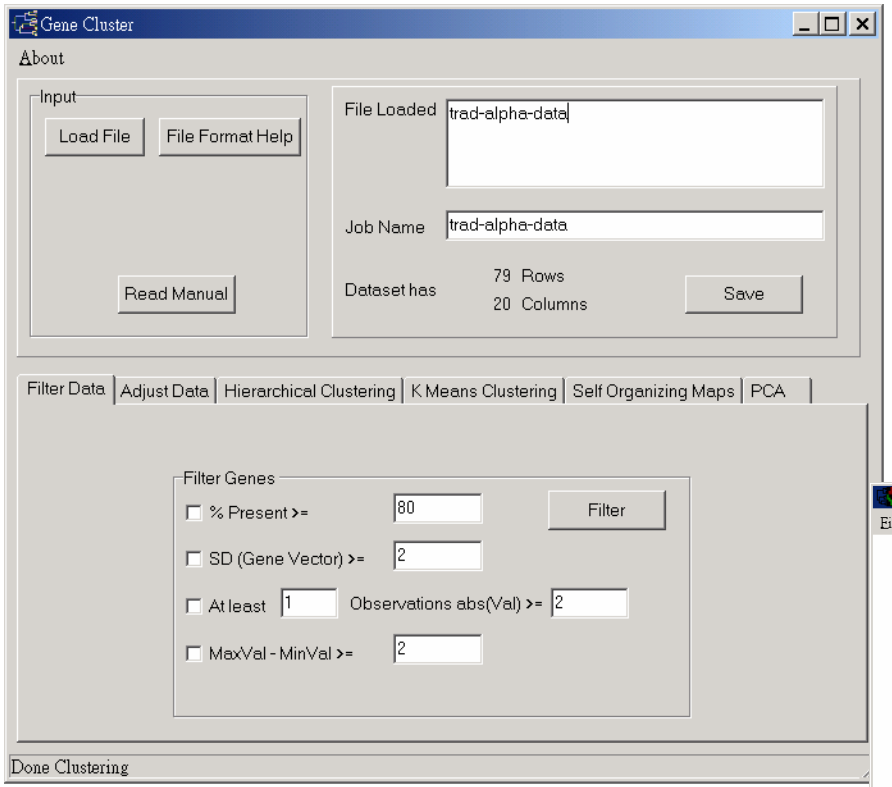


The ontologies are used to categorize gene products.

- **Biological process ontology**
- **Molecular function ontology**
- **Cellular component ontology**

Software for Clustering

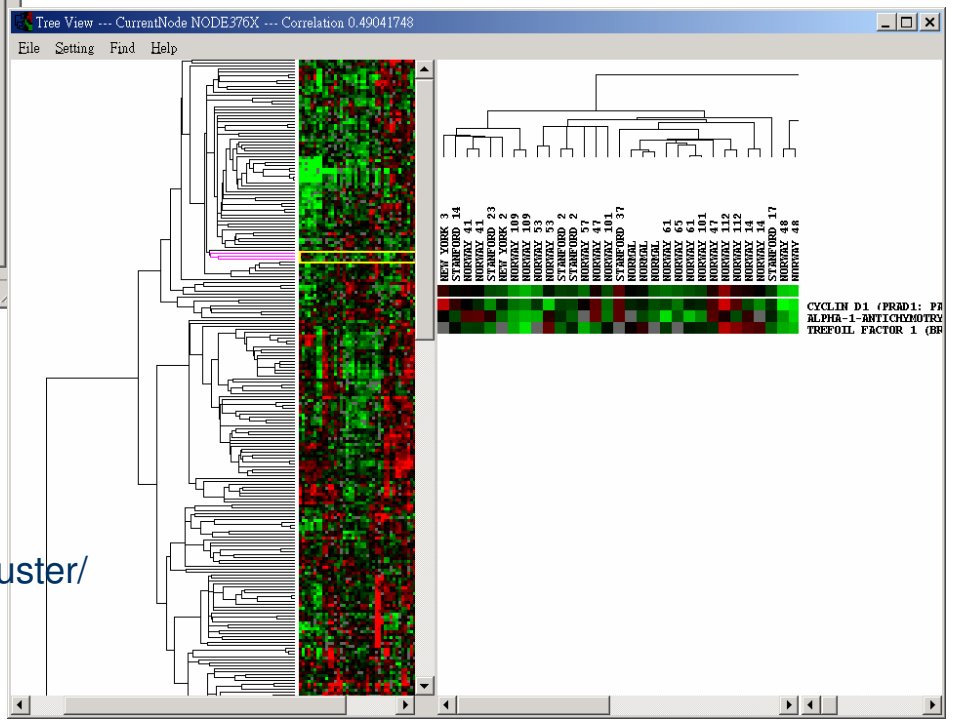
Cluster and TreeView



Eisen Lab

http://rana.lbl.gov/EisenSoftware.htm

Eisen MB, Spellman PT, Brown PO, Botstein D. (1998) **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci.* 95(25):14863-8.



De Hoon, M.J.L.; Imoto, S.; Nolan, J.; Miyano, S.; "**Open source clustering software**". Bioinformatics, 20 (9): 1453--1454 (2004)

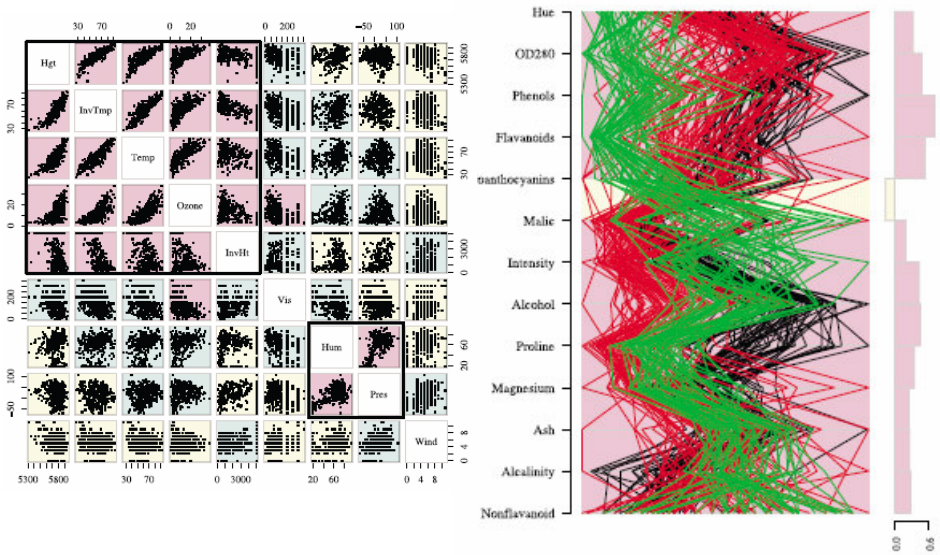
http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/

Gclus, PermutMatrix

gclus: Clustering Graphics

(R package)
http://cran.r-project.org/src/contrib/Descriptions/gclus.html

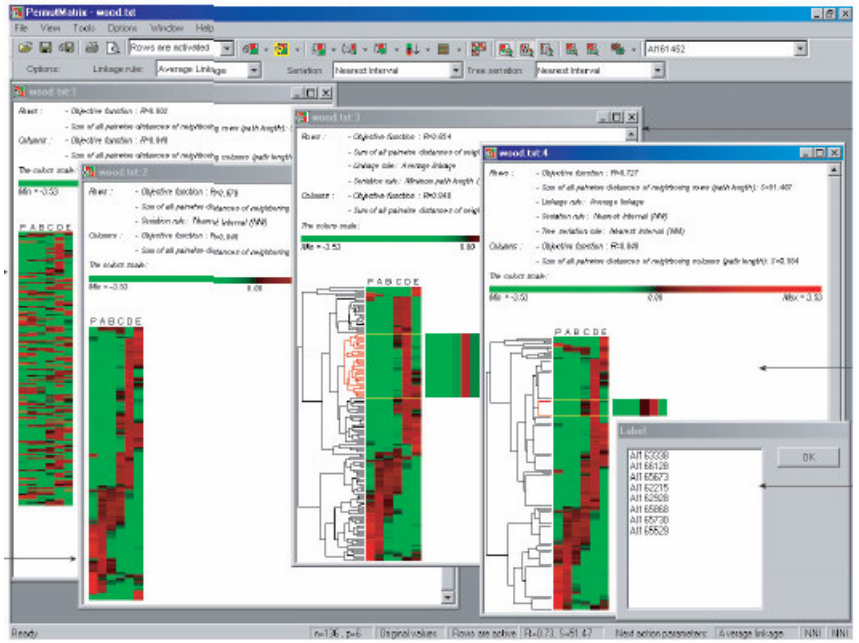
Catherine B. Hurley, (2004), Clustering Visualizations of Multidimensional Data, Journal of Computational & Graphical Statistics, Vol. 13, No. 4, pp.788-806



PermutMatrix

http://www.lirmm.fr/~caraux/PermutMatrix

Caraux, G., and Pinloche, S. (2005), "Permutmatrix: A Graphical Environment to Arrange Gene Expression Profiles in Optimal Linear Order," Bioinformatics, 21, 1280-1281.



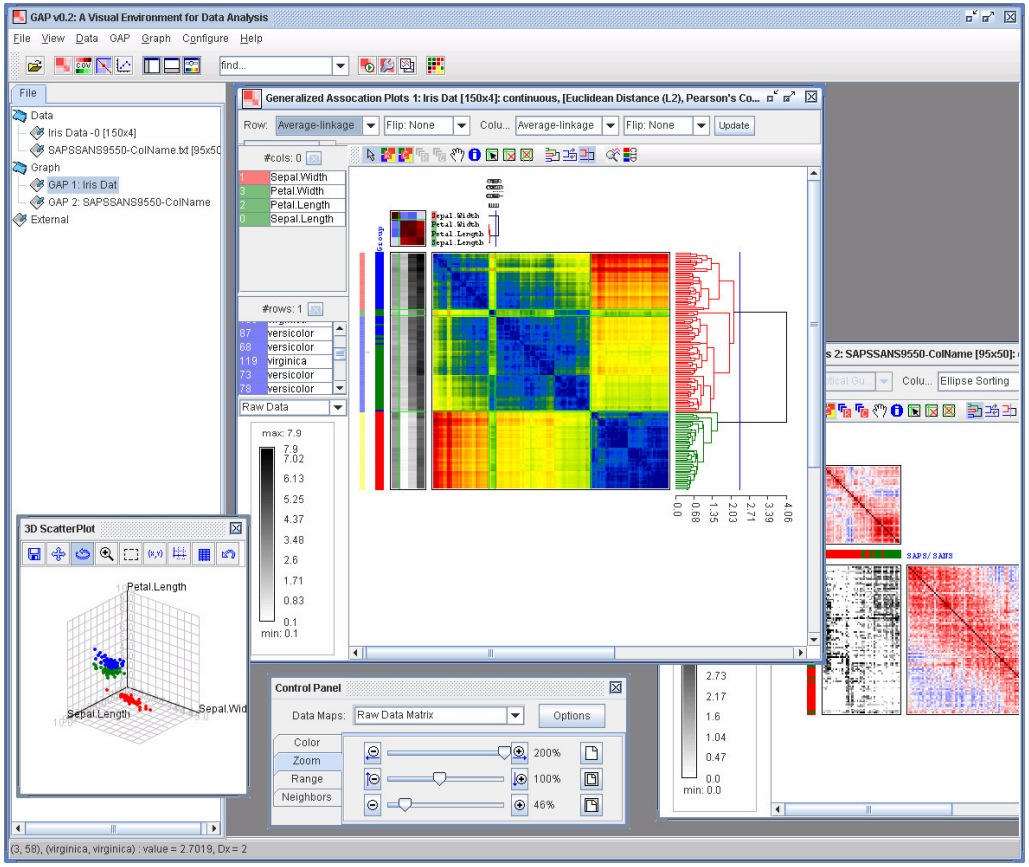
GAP Software verison 0.2

Generalized Association Plots

- Input Data Type: continuous or binary.
- Various seriation algorithms and **clustering analysis**.
- Various display conditions.
- Modules:
  GAP with Covaraite Adjusted,
  Nonlinear Association Analysis,
  Missing Value Imputation.

Statistical Plots

- 2D Scatterplot, 3D Scatterplot (Rotatable)



Chen, C. H. (2002). Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices. Statistica Sinica 12, 7-29.
Wu, H. M., Tien, Y. J. and Chen, C. H. (2006). GAP: a Graphical Environment for Matrix Visualization and Information Mining.

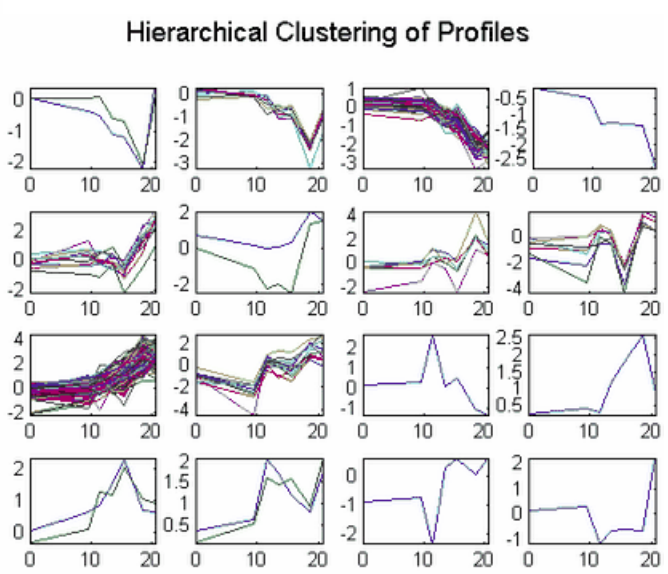
http://gap.stat.sinica.edu.tw/Software/GAP

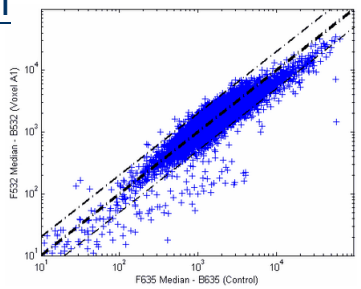
Matlab: Bioinformatics ToolBox

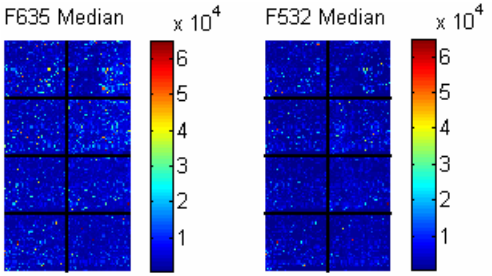
The MathWorks
Bioinformatics Toolbox

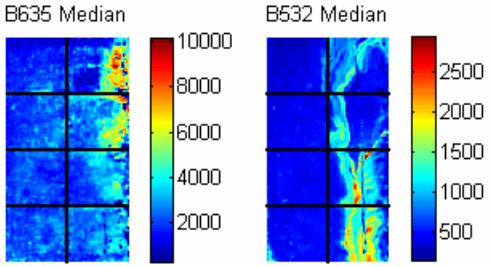
http://www.mathworks.com/access/helpdesk/help/toolbox/bioinfo/index.html

- Data Formats and Databases — Access online databases, read and write to files with standard genome and proteome formats such as FASTA and PDB.
- Sequence Alignments — Compare nucleotide or amino acid sequences using pairwise and multiple sequence alignment functions.
- Sequence Utilities and Statistics — Manipulate sequences and determine physical, chemical, and biological characteristics.
- Microarray Analysis — Read, filter, normalize, and visualize microarray data.
- Protein Structure Analysis — Determine protein characteristics and simulate enzyme cleavage reactions.
- Prototype and Development Environment — Create new algorithms, try new ideas, and compare alternatives.
- Share Algorithms and Deploy Applications — Create GUIs and stand-alone applications.



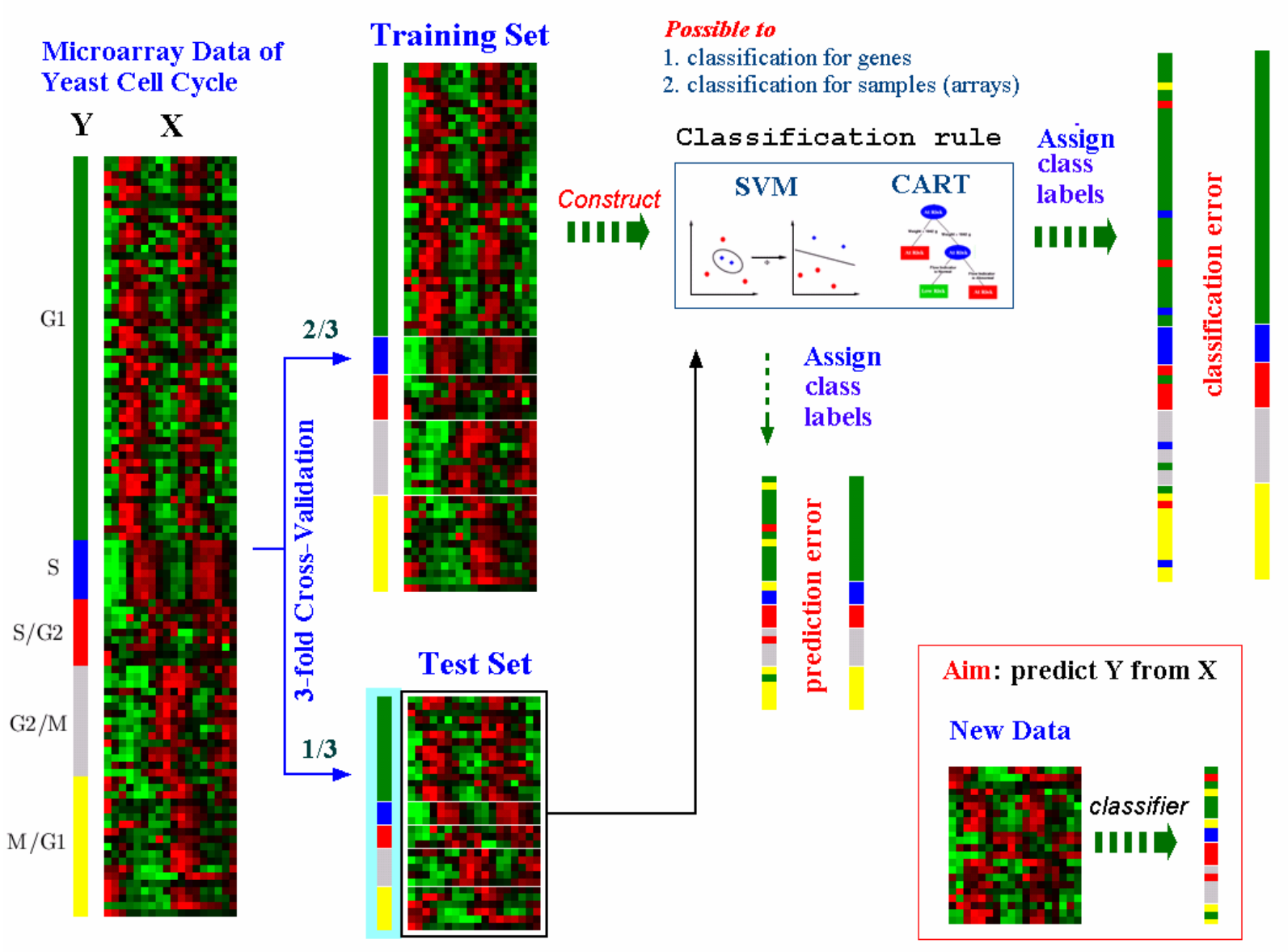






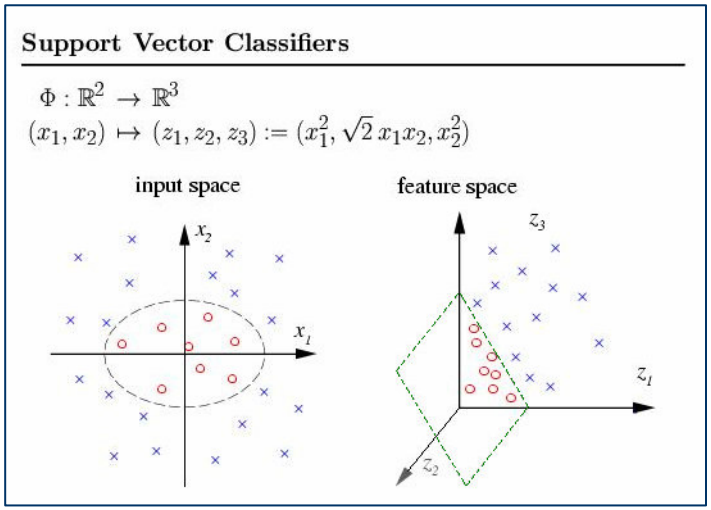
Classification of Genes, Tissues or Samples

Supervised Learning

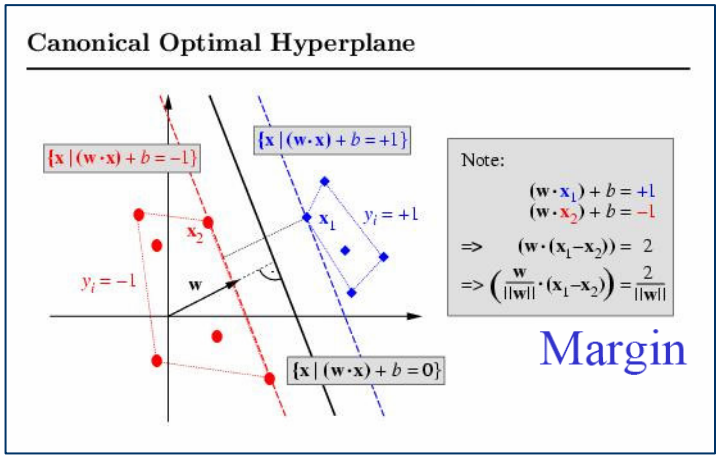


Support Vector Machine (SVM)

SVMs (Vapnik, 1995) map the data (input space) into high dimensional space (feature space) through a kernel function ϕ and then find a hyperplane $\mathbf{w}$ to separate two groups (binary classification).



Kernel Machines



Quadratic Optimization Problem

- To find the optimal hyperplane
  (solve the quadratic optimization problem)
  To minimize the quadratic form $|W|^2 = (W * W)$ subject to the linear constraints $y_i((x_i * W) + b_0) \geq 1$

decision function
$f(\mathbf{X}) = \text{sign}\,((\mathbf{X} * W) + b_0)$

Multi-class problem

Two approaches for multi-class classification:

- **one-against-others**: The kth SVM model is constructed with all of the samples in the kth class with one group, and all other samples with the other group.
- **one-against-one**: The SVM trained model is constructed by using any two of classes. Therefore, there are total $K(K-1)/2$ classifiers.

Software

SVMTorch, Collobert and Bengio, 2001
LIBSVM, Chang and Lin, 2002

Brown et al. (2000). Knowledge-based Analysis of Microarray Gene Expression Data Using Support Vector Machines, PNAS 97(1), 262-267.

Assume: Genes of similar function yield similar expression pattern.

Data

Yeast Gene Expression
[2467x 80] out of
[6,221x 80] has
accurate functional
annotations.

Tricarboxylic acid
Respiration
Ribosome
Proteasome
Histone
Helix-turn-helix

Table 1. Comparison of error rates for various classification methods

Class	Method	FP	FN	TP	TN	S(M)
TCA	D-p 1 SVM	18	5	12	2,432	6
	D-p 2 SVM	7	9	8	2,443	9
	D-p 3 SVM	4	9	8	2,446	12
	Radial SVM	5	9	8	2,445	11
	Parzen	4	12	5	2,446	6
	FLD	9	10	7	2,441	5
	C4.5	7	17	0	2,443	−7
	MOC1	3	16	1	2,446	−1
Resp	D-p 1 SVM	15	7	23	2,422	31
	D-p 2 SVM	7	7	23	2,430	39
	D-p 3 SVM	6	8	22	2,431	38

Table 3. Predicted functional classifications for previously unannotated genes

Class	Gene	Locus	Comments
TCA	YHR188C		Conserved in worm, *Schizosaccharomyces pombe*, human
	YKL039W	PTM1	Major transport facilitator family; likely integral membrane protein; similar YHL017w not co-regulated.
Resp	YKR016W		Not highly conserved, possible homolog in *S. pombe*
	YKR046C		No convincing homologs
	YPR020W	ATP20	Subsequently annotated: subunit of mitochondrial ATP synthase complex
	YLR248W	CLK1/RCK2	Cytoplasmic protein kinase of unknown function
Ribo	YKL056C		Homolog of translationally controlled tumor protein, abundant, conserved and ubiquitous protein of unknown function

⋮

Kernel Machines:
http://www.kernel-machines.org
Support Vector Machines:
http://www.support-vector.net
MATLAB Support Vector Toolbox:
http://www.isis.ecs.soton.ac.uk/resources/svminfo
SVM Application List:
http://www.clopinet.com/isabelle/Projects/SVM/applist.html

Useful Links and Reference

Y.F.Leung's FUNCTIONAL GENOMICS

Home | Functional Genomics | Microarray | Bioinformatics | Proteomics | Genome mapping Complex disease mapping Linkage analysis | About Y.F.Leung

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

AFFYMETRIX

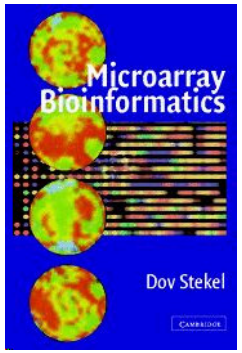
http://www.affymetrix.com

Bibliography on Microarray Data Analysis

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

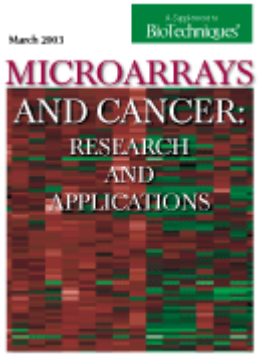
BIOINFORMATICS

http://bioinformatics.oupjournals.org

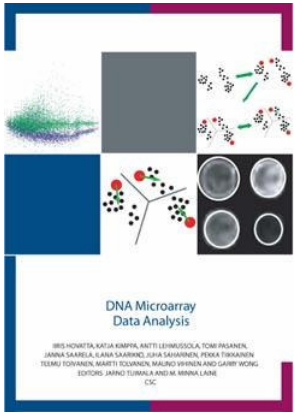


Stekel, D. (2003). Microarray bioinformatics, New York : Cambridge University Press.

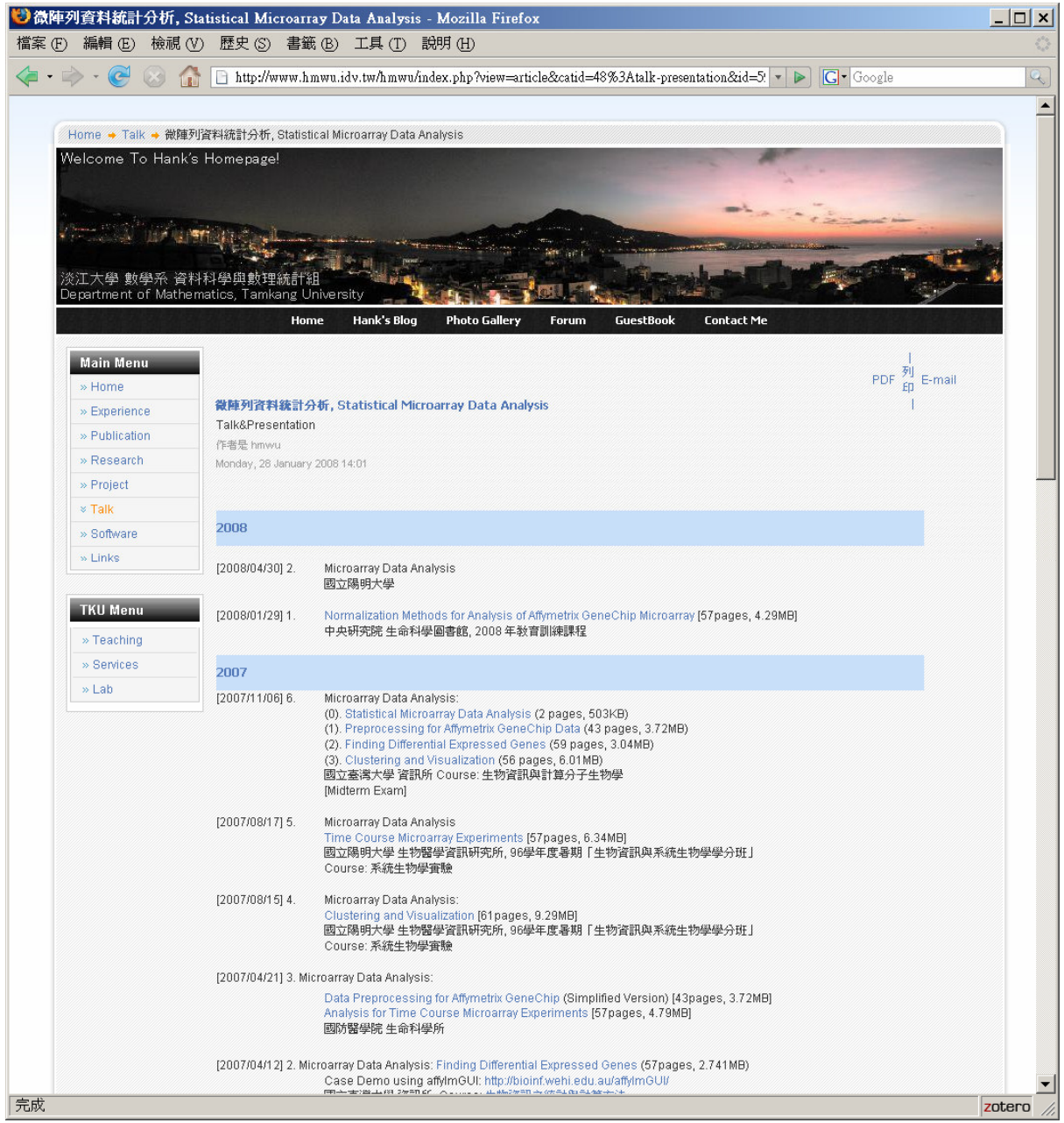
- Speed Group Microarray Page: Affymetrix data analysis
  http://www.stat.berkeley.edu/users/terry/zarray/Affy/affy_index.html
- Statistics and Genomics Short Course, Department of Biostatistics Harvard School of Public Health.
  http://www.biostat.harvard.edu/~rgentlem/Wshop/harvard02.html
- Statistics for Gene Expression
  http://www.biostat.jhsph.edu/~ririzarr/Teaching/688/
- Bioconductor Short Courses
  http://www.bioconductor.org/workshop.htm



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



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



Thank You!

淡江大學 數學系

吳漢銘 助理教授

hmwu@mail.tku.edu.tw
http://www.hmwu.idv.tw