Biochip informatics-(I) :

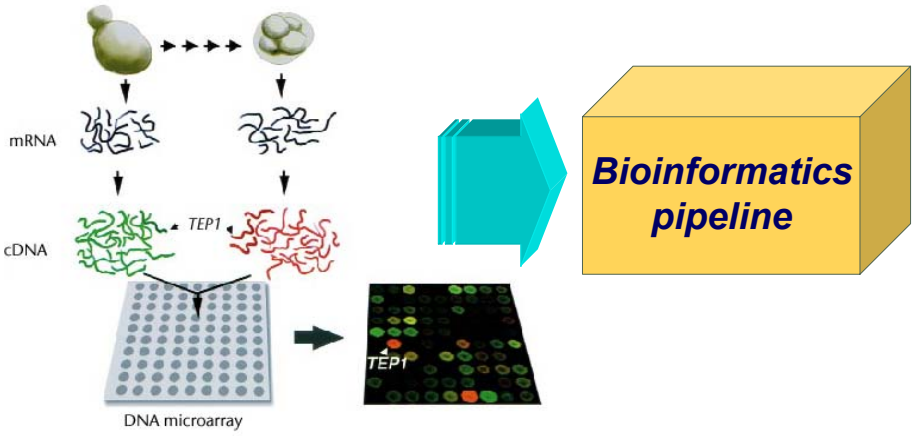
biochip normalization & differential expression

Ju Han Kim, M.D., Ph.D.
SNUBI: SNUBiomedical Informatics
http://www.snubi.org/

Biochip Informatics - (I)

- *Biochip basics*
- *Preprocessing*
- *Episodes 1 and 2*
- *Global normalization*
- *Intensity dependent normalization*
- *Controlling regional variation*
- *Alternatives*
- *Differential expression*
- *Multiple hypothesis testing*
- *Classification*

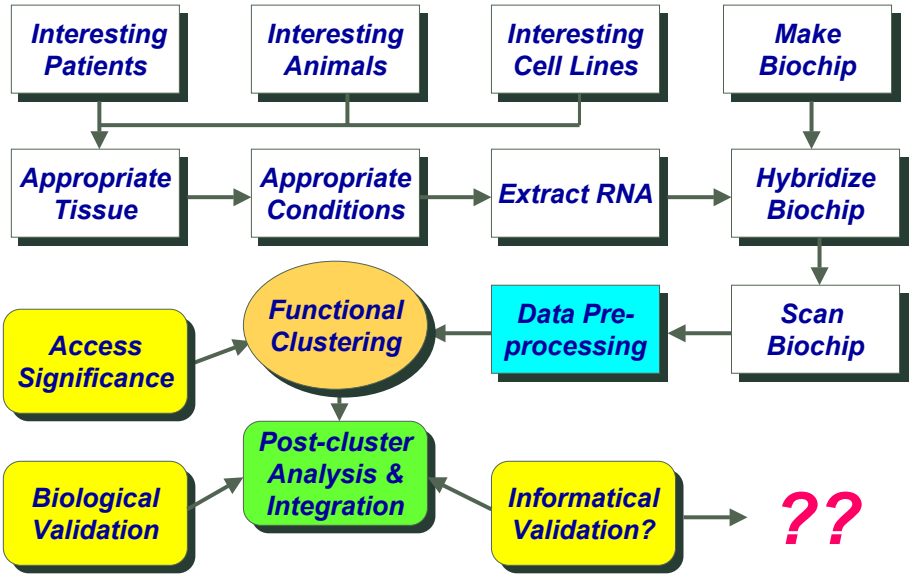
Biochip basics



Terminology

- ***Sample (Target):*** *RNA (cDNA) hybridized to the array, aka target, mobile substrate.*
- ***Probe:*** *DNA spotted on the array, aka spot, immobile substrate.*
- ***Sector (Block):*** *rectangular matrix of spots printed using the same print-tip (or pin), aka print-tip-group*
- ***Slide, Array:*** *printed microarray*
- ***Batch:*** *collection of microarrays with the same probe layout.*
- ***Channel:*** *data from one color (**Cy3 = cyanine 3 = green, Cy5 = cyanine 5 = red**).*

A Biochip Informatics Strategy



Clinical relevance of Biochip informatics

Dx.

Discovery

Tx.

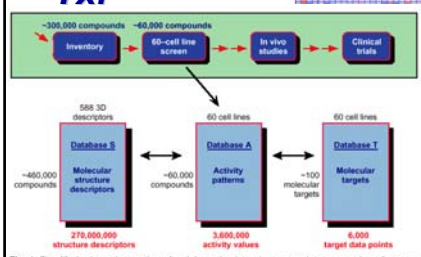
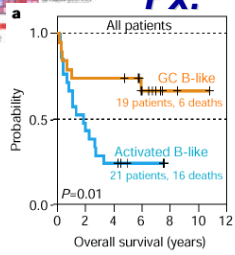
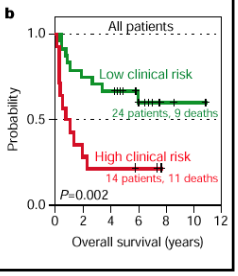


Fig. 1. Simplified schematic overview of an information-intensive approach to cancer drug discovery and molecular pharmacology at the NCI. Each row of the activity (**A**) database represents the pattern of activity of a particular compound across the 60 cell lines. As described in the text, the **A** database can be related to a structure (**S**) database containing 2D or 3D chemical structure characteristics of the compounds and a target (**T**) database containing information on possible molecular targets or modulators of activity within the cells.

Px.

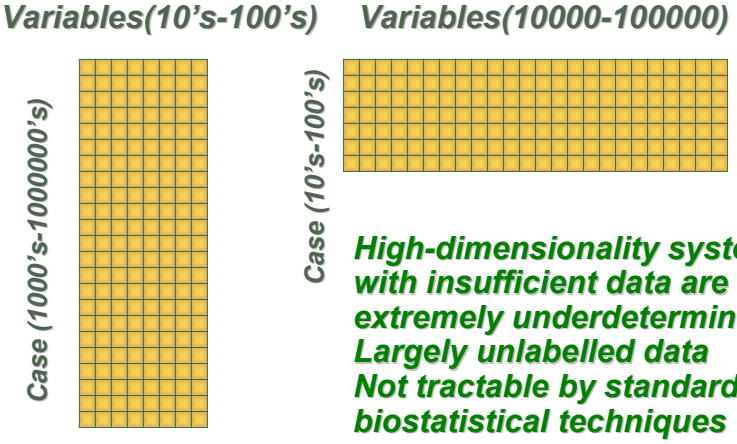




Biochip informatics: challenges

- **Pre-processing:**
  - ✓ technology variation
  - ✓ noise & data filtering
  - ✓ missing / negative values / P & A calls
  - ✓ data scaling
  - ✓ Can I assume normality?
  - ✓ chip quality, other artifacts
- **Functional Clusters:**
  - ✓ clustering quality, consistency, & robustness
- **Statistical Issues:**
  - ✓ study design / # of replicates / multiple testing
- **Integrative Biochip Informatics**
  - ✓ Can we get more out of it?

Why integrative approach?



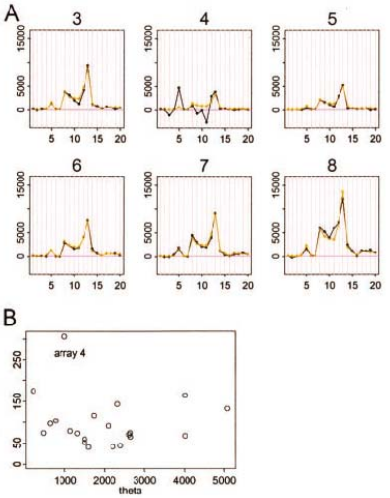
High-dimensionality systems with insufficient data are extremely underdetermined
Largely unlabelled data
Not tractable by standard biostatistical techniques

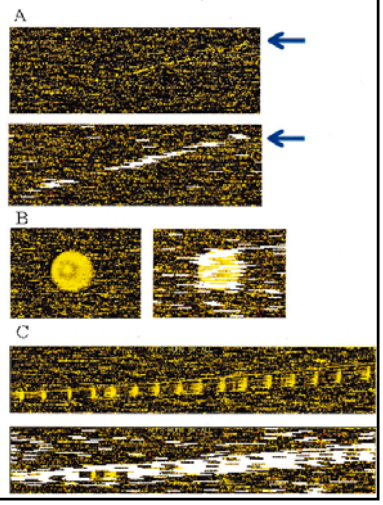
Preprocessing

- technology variation
- noise & outlier detection
- missing / negative values / P & A calls
- data scaling
- Can I assume normality?
- chip quality, other artifacts

Preprocessing: Noise

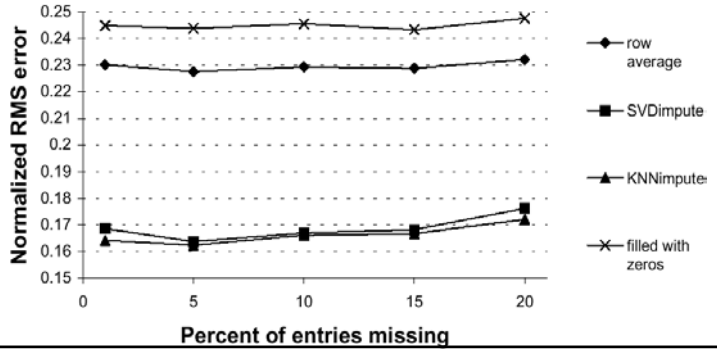
Weighted average difference $\tilde{\theta} = \frac{\sum_N (PM_n - MM_n)\phi_n}{N}$





Preprocessing: missing values

- put in zeros
- row average values
- weighted KNN
- SVD (requires iteration)



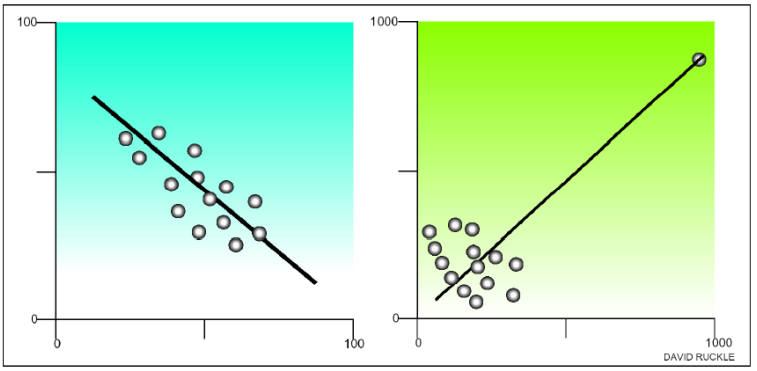
Preprocessing: Filtering

- intensity-based filter
  - ✓ floors or cut-offs: min. expression level
  - ✓ 2 s.d above (local) background
  - ✓ percentage-based cut-offs
  - ✓ also consider saturation
- low variance filter (across conditions)
- statistical filter (with replicates)
- low entropy filter / jackknife clustering

$$H(x) = -\sum_i p(x_i)\log_2(p(x_i))$$

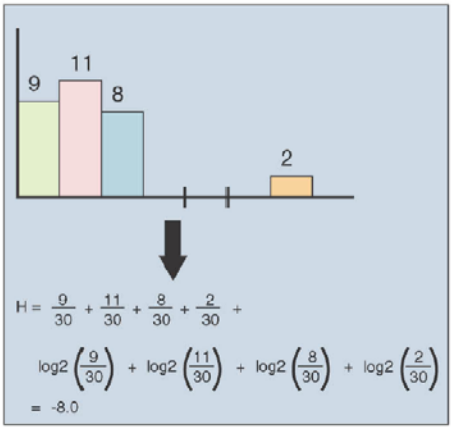
- target ambiguity filter (databases)

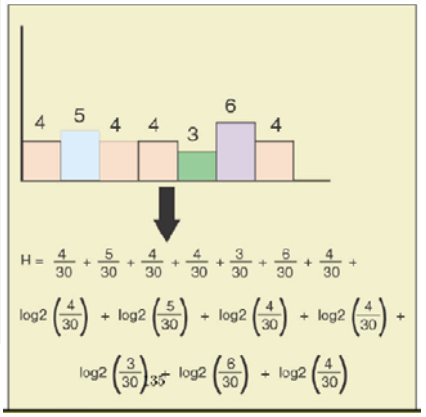
Filtering: Low entropy filter



* Jackknife clustering

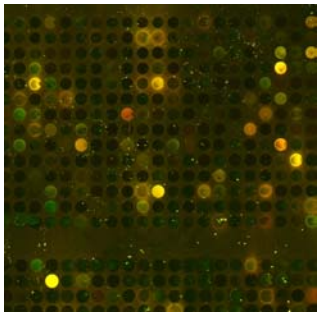
Filtering: Low entropy filter

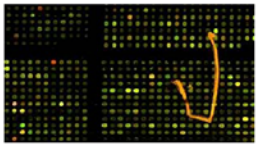




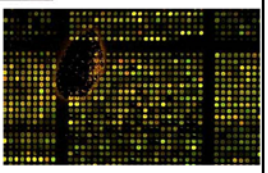
Chip quality

Debris

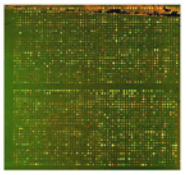




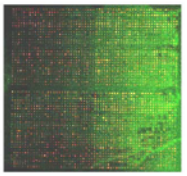
Scratch?



Bubble



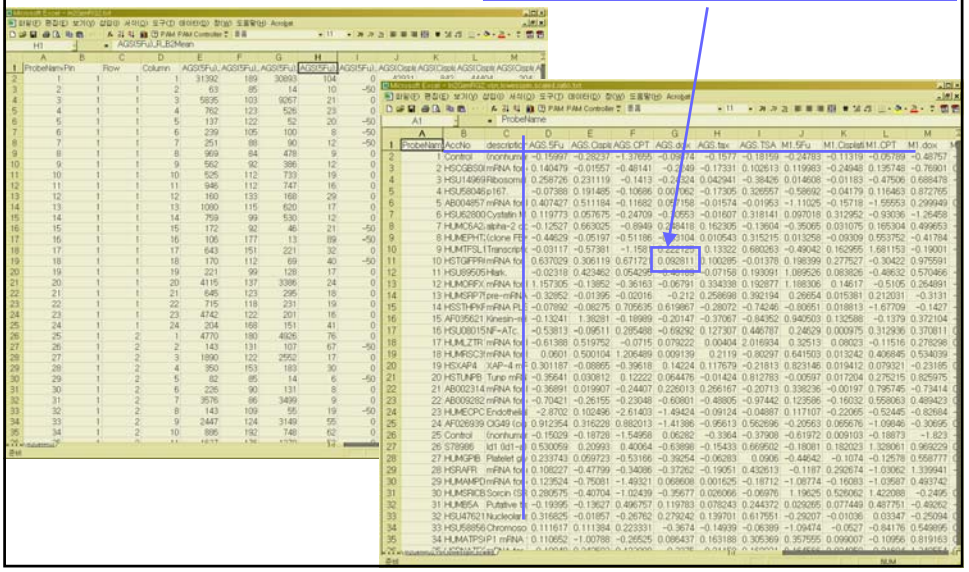
Edge effect



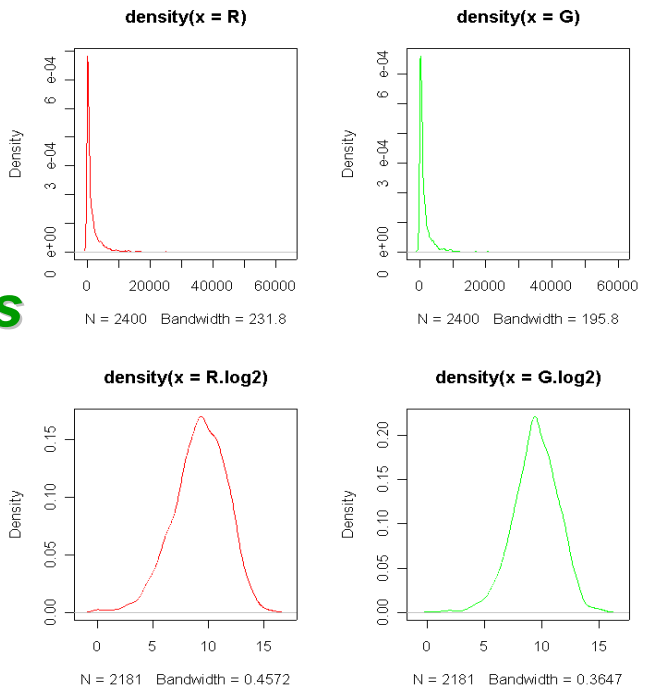
Background haze

Data structure

Gene expression level of gene 10 on slide 4
= Log2(Red intensity / Green intensity)



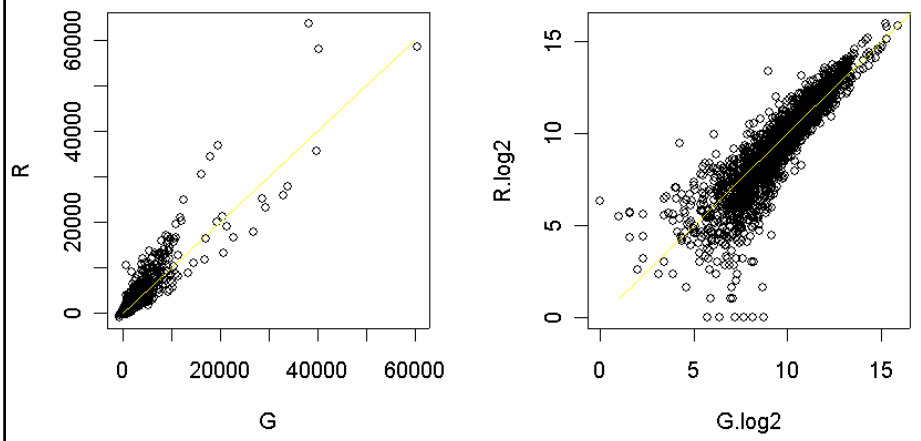
density plots of intensities



Intensity vs. intensity plot

For the same, equally-treated samples,

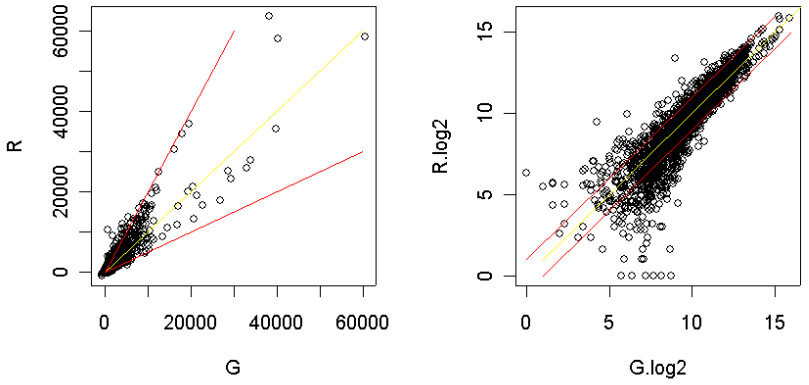
$I(R_i) = I(G_i)$



Intensity vs. intensity plot

For different sample, what about R/G ratios?

What possibly are the problems?

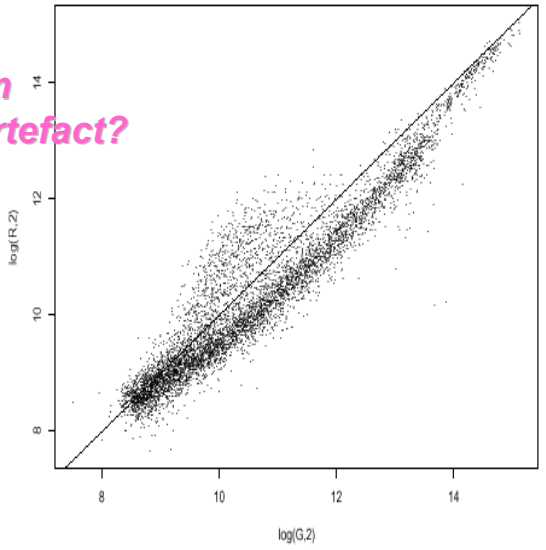


Log transformations

- **Pros**
  - ✓ constant c.v., i.e., mean ∝ s.d.
  - ✓ useful way of handling ratio values
  - ✓ symmetry for up or down
- **Cons**
  - ✓ level of expressions & significance, low < high
  - ✓ non-positive values
  - ✓ only linear calibration transformations
- **Alternatives**

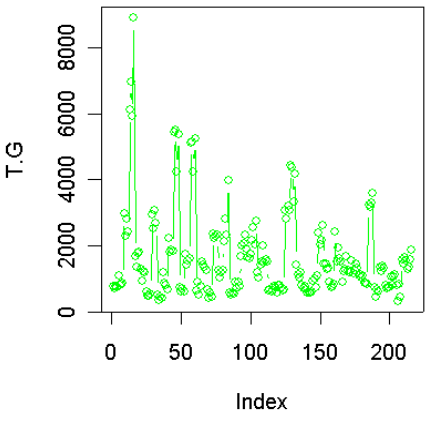
Episode - I

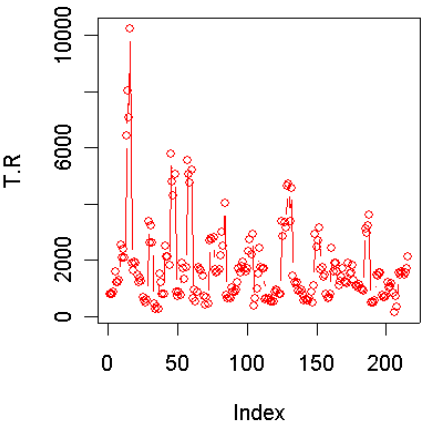
a frequency pattern
Is it typical or an artefact?



Episode - II

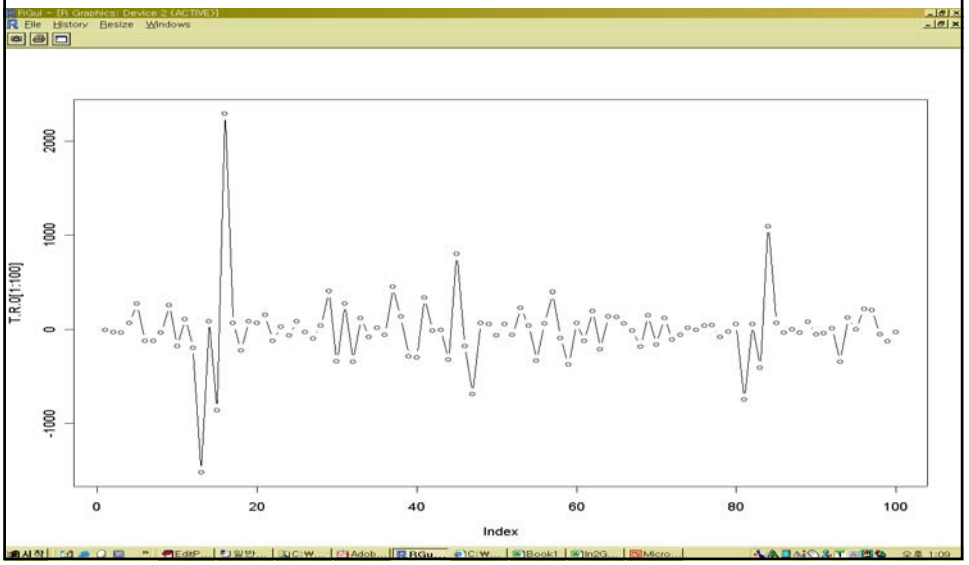
a frequency pattern
Is it typical or an artifact?



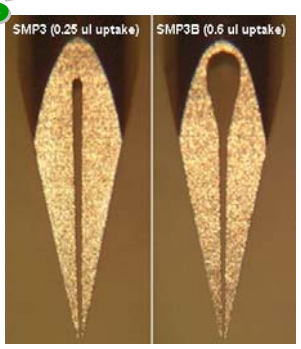


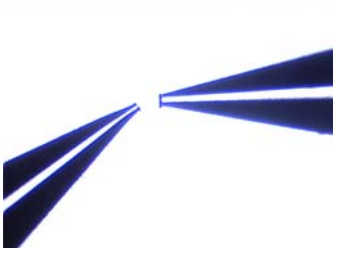
Episode - II

a frequency pattern
Is it typical or an artifact?



Pin tips

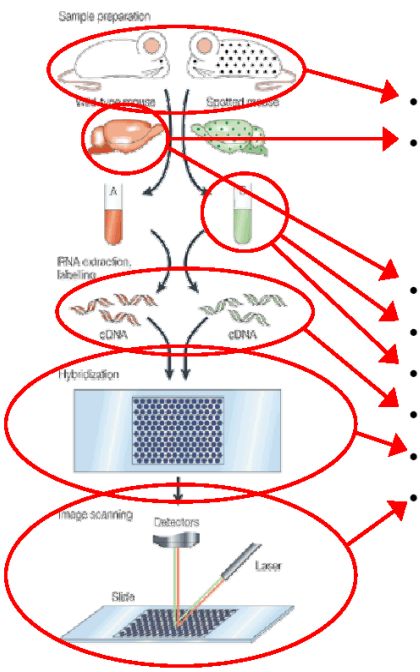




Normalization

Correcting systematic variation

- **Simple additive and multiplicative**
- **Linear vs. non-linear**
- **Sources of error**
- **Kinds?**
  - ✓ **dye**
  - ✓ **spotting**
  - ✓ **experiment, slide**
  - ✓ **scale, scanning**



Sources of Variability

(cDNA Array Example)

- Biological Heterogeneity in Population
- Specimen Collection/ Handling Effects
  - Tumor: surgical bx, FNA
  - Cell Line: culture condition, confluence level
- Biological Heterogeneity in Specimen
- RNA extraction
- RNA amplification
- Fluor labeling
- Hybridization
- Scanning
  - PMT voltage
  - laser power

(Geschwind, *Nature Reviews Neuroscience*, 2001)

Sources of error

- *tissue contamination*
- *mRNA preparation and RNA degradation*
- *amplification efficiency*
- *reverse transcription efficiency*
- *hybridization efficiency and specificity*
- *clone identification and mapping*
- *PCR yield, contamination*
- *spotting efficiency*
- *Pin geometry*
- *DNA-support binding*
- *Dye labeling*
- *Slide variation*
- *other array manufacturing-related issues*
- *scanning*
- *image segmentation*
- *signal quantification*
- *'background' correction*

Why normalize?

To correct systematic variations such as

- *To balance the fluorescence intensities of the two dyes (green Cy3 and red Cy5 dye)*
- *To allow the comparison of expression levels across experiments (slides)*
- *To adjust scale of the relative gene expression levels (as measured by log ratios) across replicate experiments*

Normalization issues, cDNA chips

- Within-slide
  - ✓ What genes to use
  - ✓ Location
  - ✓ Scale
- Paired-slides (dye swap)
  - ✓ Self-normalization
- Between slides

Which genes to use?

- All genes
- Constantly expressed genes
- Controls
  - ✓ Spiked controls
  - ✓ Genomic DNA titration series
- Other ‘useful’ set of genes

Global normalization

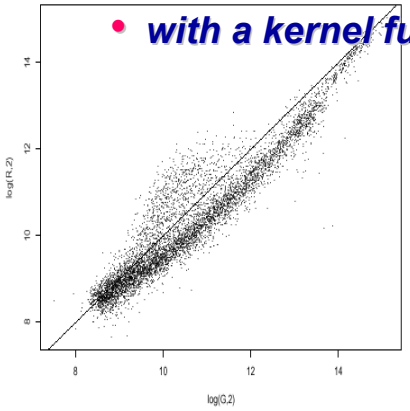
Total amount of mRNA measured is constant

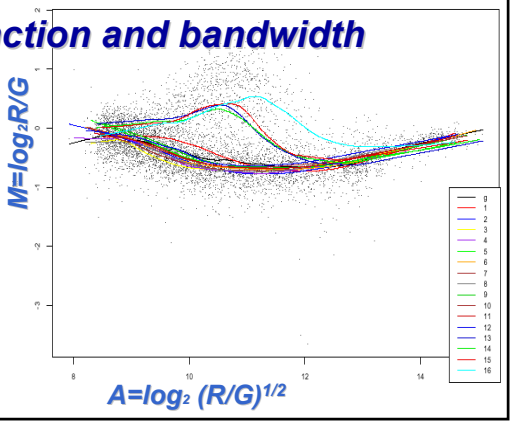
$\log_2 R/G \rightarrow \log_2 R/G - c = \log_2 R/(kG)$

- Set mean or median to zero.
- Assumes R and G are related by a constant factor
- Ignoring intensity-and-space-dependent variations

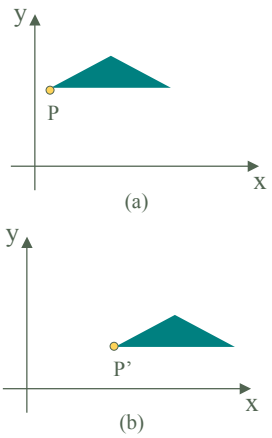
Non-parametric smoothing: lowess

- Spread is a function of intensity
- Regression without parametric assumption
- on a X-Y plot
- with a kernel function and bandwidth





Geometric transformations: translation



$$x' = x + T_x,$$
$$y' = y + T_y$$

$$P = \begin{bmatrix} x \\ y \end{bmatrix}, \quad P' = \begin{bmatrix} x' \\ y' \end{bmatrix}, \quad T = \begin{bmatrix} T_x \\ T_y \end{bmatrix}$$

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & 0 & T_x \\ 0 & 1 & T_y \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

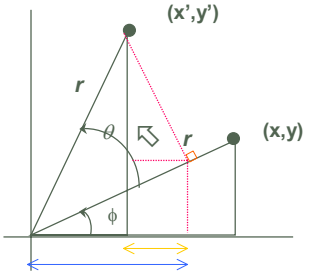
Geometric transformations: rotation

$x = r \cos \phi, \quad y = r \sin \phi$

$x' = r \cos (\phi + \theta) = r \cos \phi \cos \theta - r \sin \phi \sin \theta$

$y' = r \sin (\phi + \theta) = r \cos \phi \sin \theta + r \sin \phi \cos \theta$

$\therefore \; x' = x \cos \theta - y \sin \theta, \quad y' = x \sin \theta + y \cos \theta$



$$P' = R \cdot P$$

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

Geometric transformations: scaling

$x' = x \cdot s_x$, $y' = y \cdot s_y$

Scaling factor : s_x (x축으로 크기 조정),

s_y (y축으로 크기 조정)

$P' = S \cdot P$

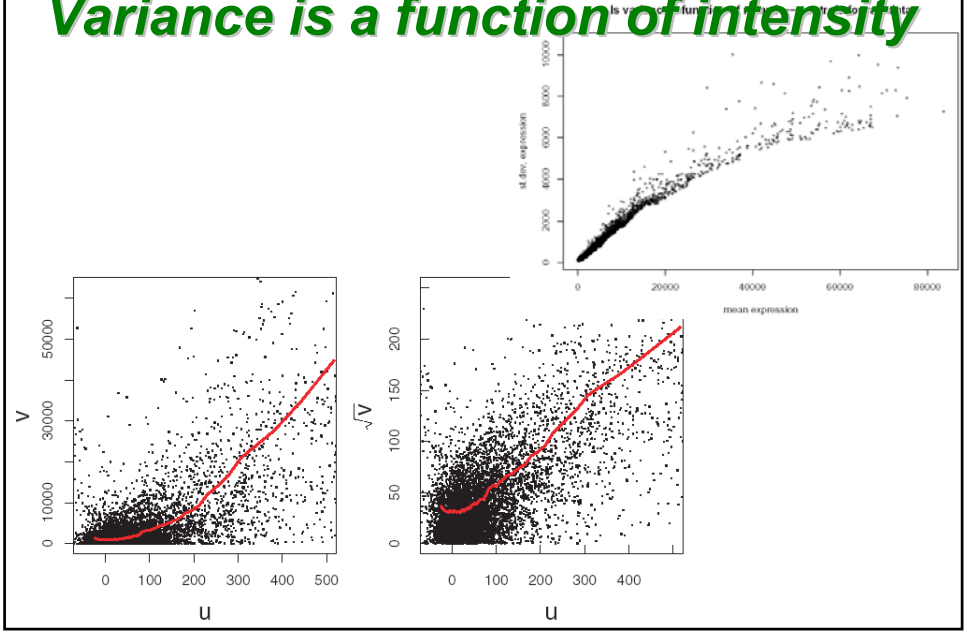
Uniform Scaling: $s_x = s_y$

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$$P' = T + P$$
$$P' = S \cdot P$$
$$P' = R \cdot P$$

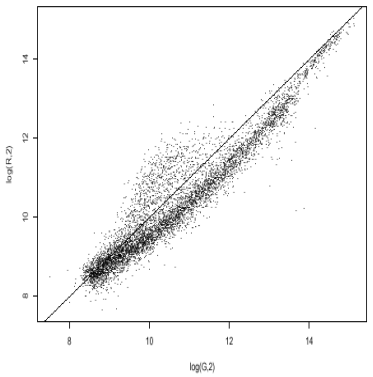
$$P' = \underbrace{(T \cdot R \cdot S \cdot T \ldots)}_{M} \cdot P$$

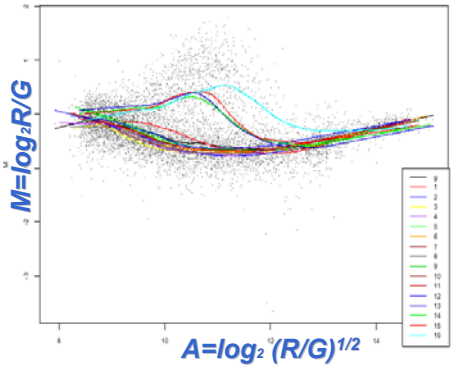
Variance is a function of intensity



The M vs. A plot

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(-\pi/2) & -\sin(-\pi/2) \\ \sin(-\pi/2) & \cos(-\pi/2) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$





M = $\log_2(R/G)$: log intensity-ratio
A = $\log_2(R*G)/2$: mean log-intensity

Intensity-dependent normalization

Dye bias is dependent on spot intensity!

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

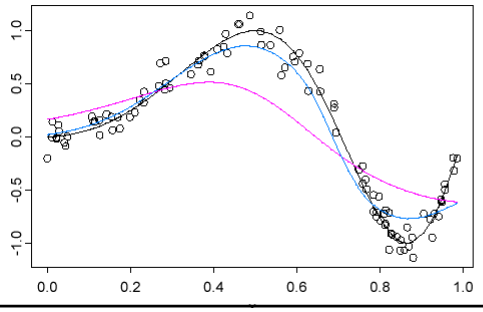
- ***Apply a robust scatter-plot smoother, lowess***
- ***The lowess() function in R with f=20%***
- ***Assumes roughly symmetric up- and downs at all intensity levels***

Nonparametric smoothing

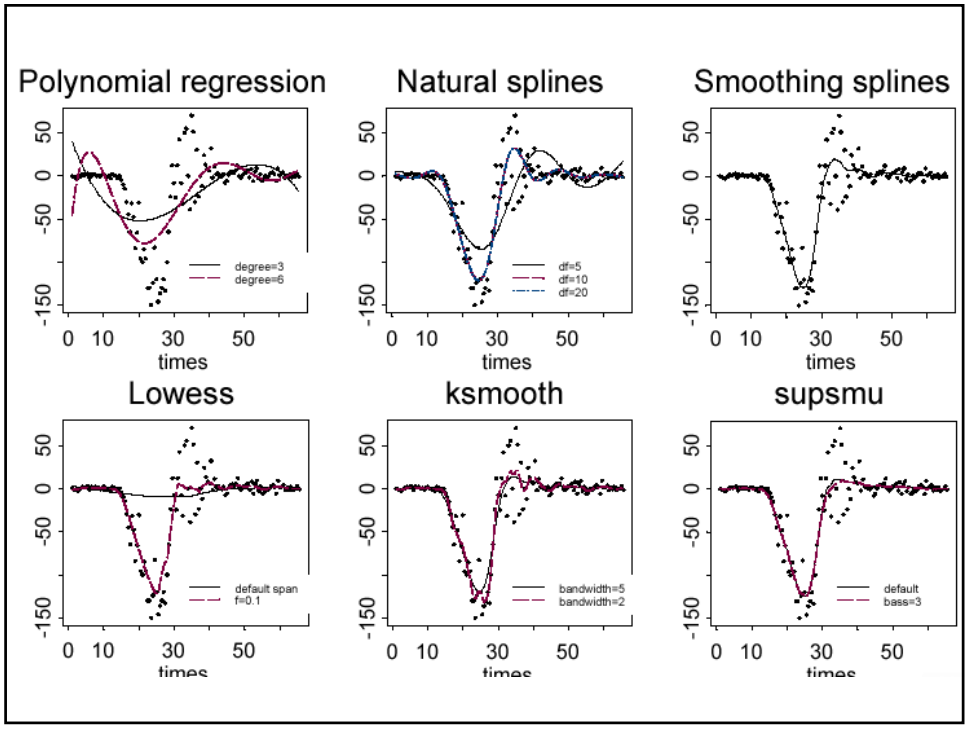
- *Smoothing*
  - ✓ *Consider X Y plot*
  - ✓ *Draw a regression line which requires no parametric assumptions*
  - ✓ *The regression line is not linear*
  - ✓ *The regression line is totally dependent on the data*
- *Two components of smoothing*
  - ✓ *Kernel function, calculating weighted mean*
  - ✓ *Bandwidth, a window span determining smoothness of the regression line*

Nonparametric smoothing

- *Types of kernel functions*
  - ✓ *Uniform, Triangular, Normal, Others*
- *Bandwidth*
  - ✓ *The wider, the smoother*
  - ✓ *Bigger impact than the kernel function*

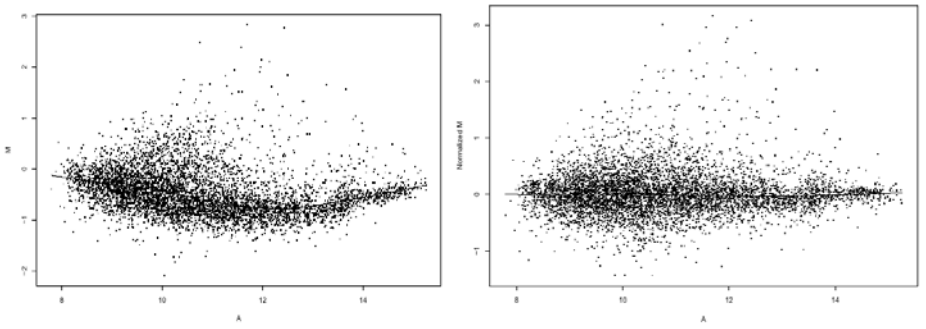


Black = true
Blue = small bandwidth
Red = too large bandwidth



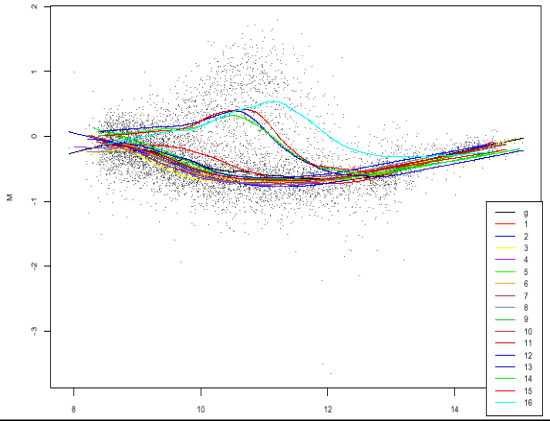
Global vs. print-tip-group lowess

Roughly equal number of genes are up- or down-regulated at all intensity levels (or only few genes are expected to change)



Global vs. print-tip-group lowess

For every print-tip group, changes roughly symmetric at all intensity levels



Within print-tip-group box plots for print-tip-group normalized M

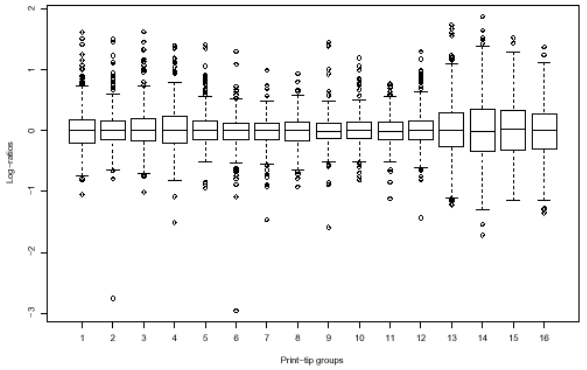


Figure 3. *Within-slide normalization.* Boxplot displaying the log-ratio distribution after within-print-tip-group location normalization for each of the 16 print-tip groups. The array was printed using a 4 by 4 print-head and the print-tip groups are numbered first from left to right then from top to bottom starting from the top left corner (data from apo A1 knock-out mouse #8 in experiment (A)).

Taking scale into account

Assumptions: all print-tip-group have the same spread

True ratio is μ_{ij}.

The observed ratio is $M_{ij} = a_i \mu_{ij}$.

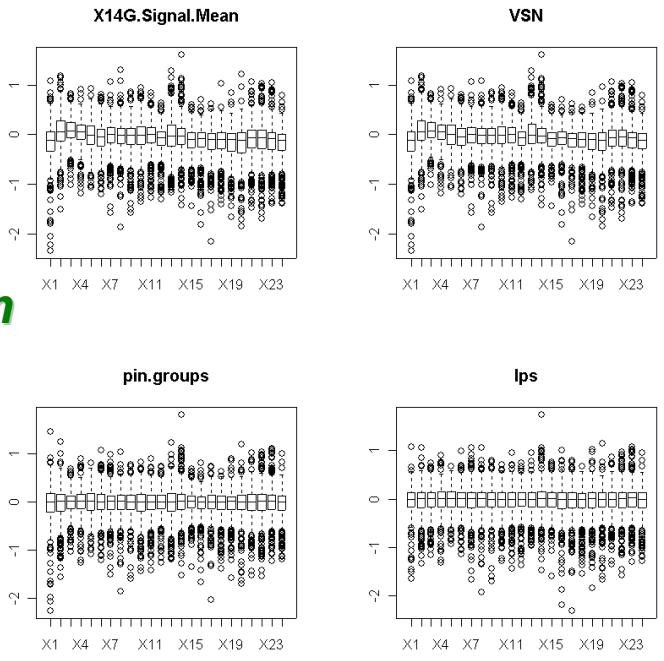
A robust alternatives to the latter estimator is

$$\hat{a}_i = \frac{MAD_i}{\sqrt[I]{\prod_{i=1}^{I} MAD_i}}$$

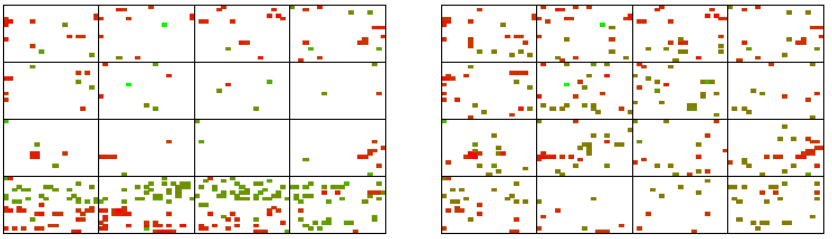
where the MAD(Median Absolute Deviation) is defined by

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

Location / scale calibration



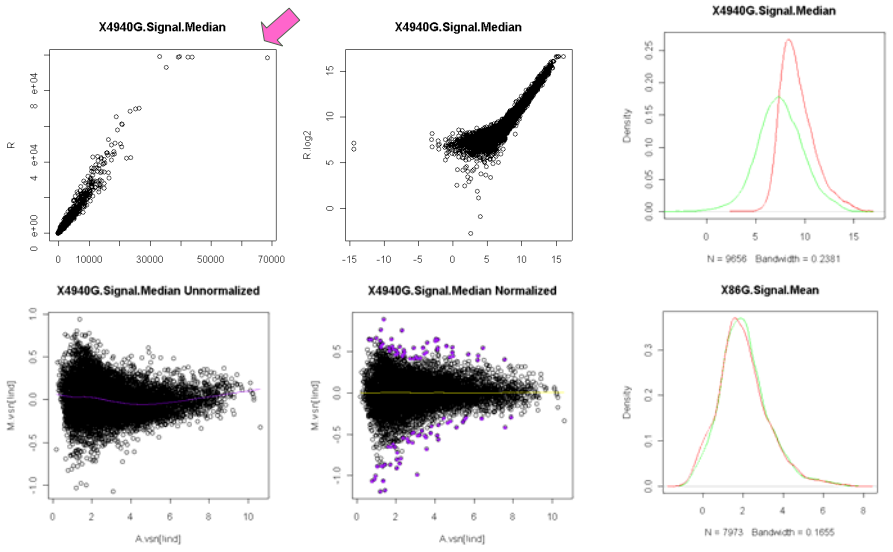
Location + scale normalization



(a) (b)

Figure 2. *Within-slide normalization.* Spatial plot of the array highlighting the spots with the largest 5% absolute log-ratios. The different shades of red represent positive log-ratios and the different shades of green represent negative log-ratios. The plot is divided into 16 grids representing the 16 different print-tip groups. Each small rectangular cell represents the log-ratio of a spot on the array. (a) Extreme log-ratios after within-print-tip-group location normalization but before scale adjustment. (b) Extreme log-ratios after within-print-tip-group location and scale normalization (data from apo A1 knock-out mouse #8 in experiment (A)).

Effect of normalization



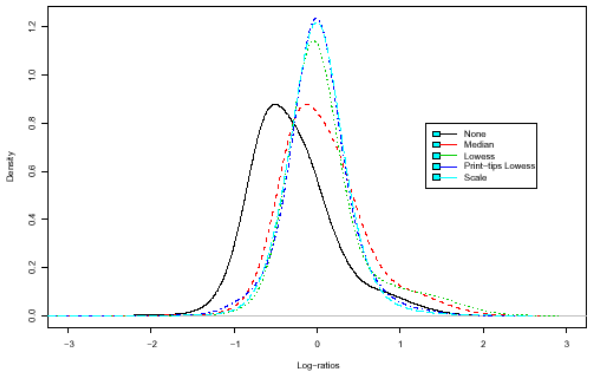
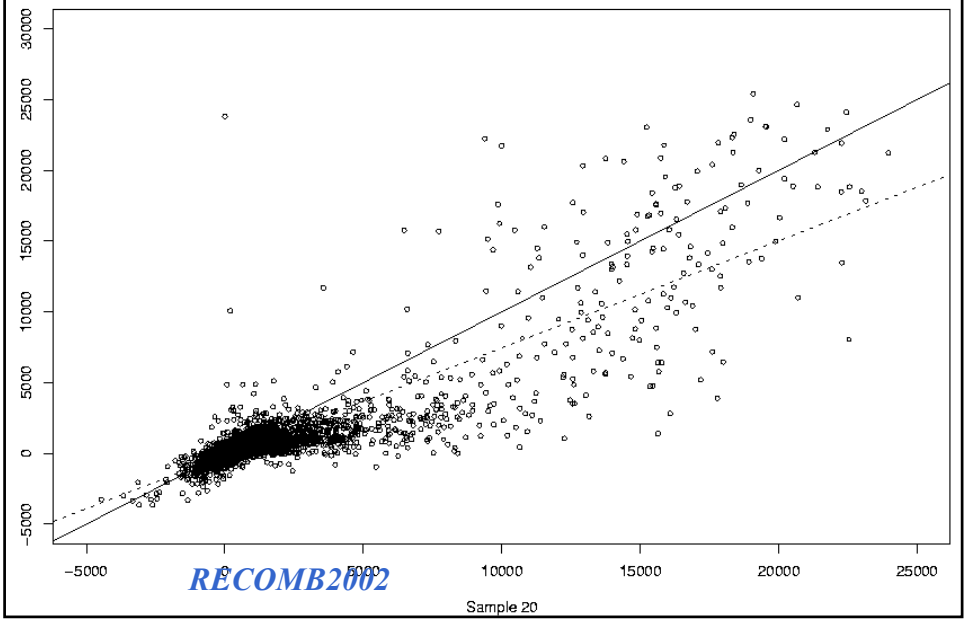


Figure 7. *Within-slide normalization.* Density plots of the log-ratios before and after different normalization procedures. The solid black curve represents the density of the log-ratios before normalization. The red, green, blue, and cyan curves represent the densities after global median normalization, intensity dependent location normalization, within-print-tip-group location normalization, and within-print-tip-group scale normalization, respectively (data from apo A1 knock-out mouse #8 in experiment (A)).

Biochip Informatics - (I)

- Biochip basics
- Preprocessing
- Episodes 1 and 2
- Global normalization
- Intensity dependent normalization
- Controlling regional variation
- **Alternatives**
- Differential expression
- Multiple hypothesis testing
- Classification

Non-linear normalization



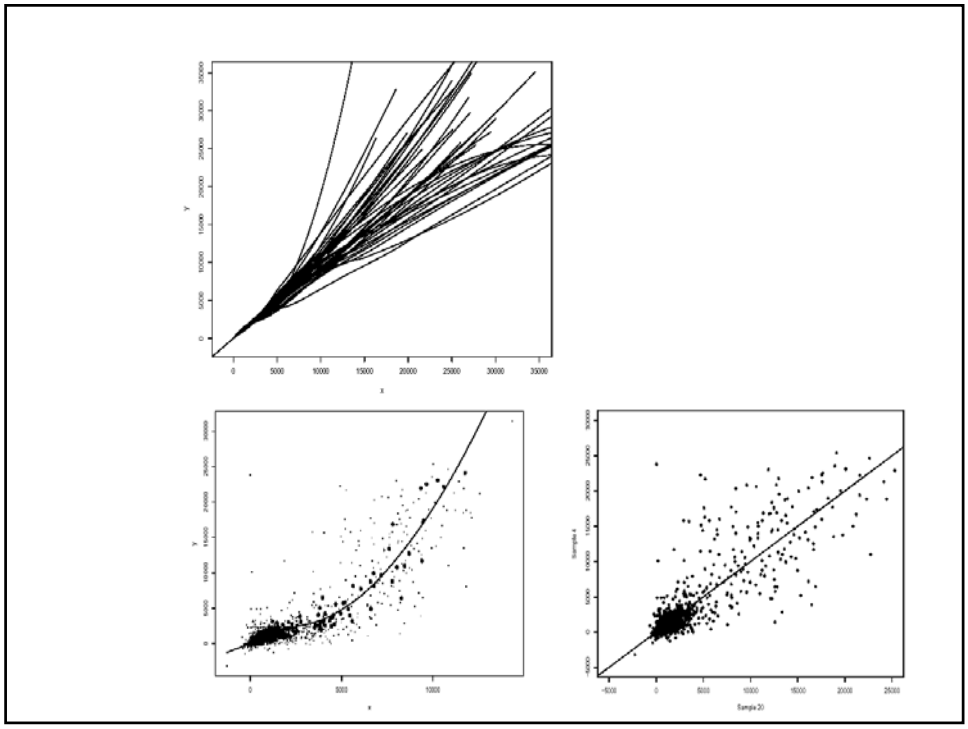
RECOMB2002

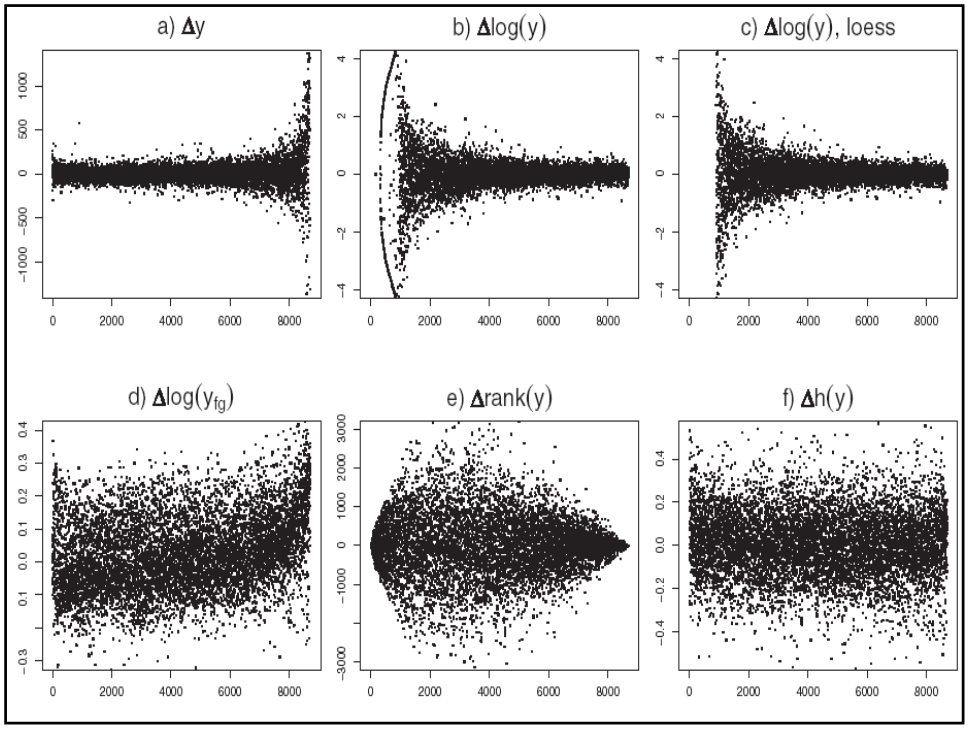
Robust Non-linear Normalization

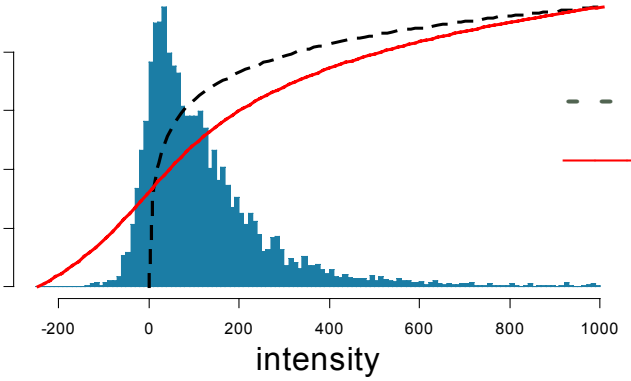
- *A rank score for more robust identification of non-differentially expressed genes*

$$R_k = \sum_{\substack{i,j=1 \\ i \neq j}}^{p} \left(x_{ki} - x_{kj} \right)^2$$

Assumption: there are non-differentially expressed genes at all range of median expression level




a) Δy
b) Δlog(y)
c) Δlog(y), loess
d) Δlog(y_fg)
e) Δrank(y)
f) Δh(y)

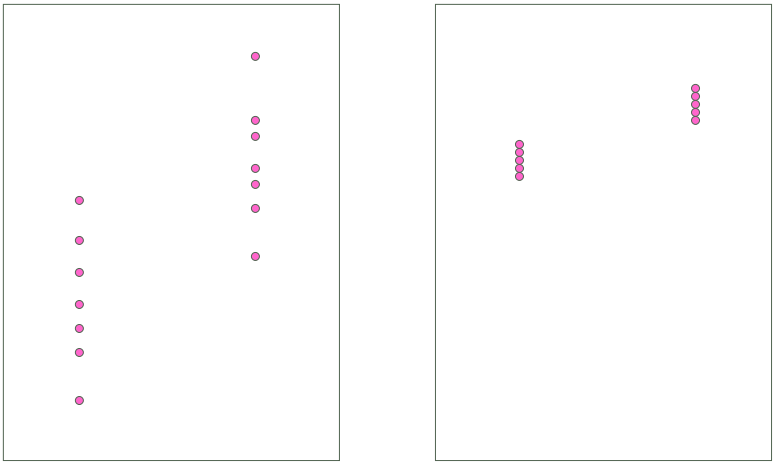


$$\operatorname{arsinh}(x) = \log\left(x + \sqrt{x^2 + 1}\right)$$

$$\lim_{x \to \infty} \left(\operatorname{arsinh} x - \log x - \log 2\right) = 0$$

Differential expression

Form a statistic from the central value and spread.



deviation, sum of deviation, variance, standard deviation

Statistical testing

- ***Form a statistic (such as T) (for each gene) from the data***
- ***Calculate the null distribution(s) for the statistic***
- ***Choose the rejection region***
- ***Compare the statistic to the null distribution(s) of the statistic***
- ***Assigning a score***

Form a statistic

$u_{jk}(i)$: the $\log_2$R/G value of the j-th gene on the k-th array in the two groups(i=1,2)

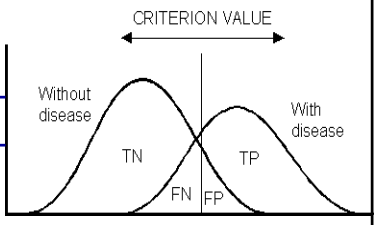
- ***log fold-change:*** $\bar{u}_j(2) - \bar{u}_j(1)$
- ***T:*** $(\bar{u}_j(2) - \bar{u}_j(1))/s_j$
- ***Wilcoxon*** *(rank sum):* $r_j = \Sigma r_{jk}$

- ***SAM*** *(shrunken centroid):* $(\bar{u}_j(2) - \bar{u}_j(1))/(s_j+s_0)$
- ***Baldi's*** *(Bayesian):* $(\bar{u}_j(2) - \bar{u}_j(1))/sqrt((1-w)s_j^2+ws_0^2)$

**** Depends on the way how the pop. variability is accounted and how to borrow strength across genes.***

Diagnostic test characteristics

	D +	D –	
T +	a	b (α)	a+b
T -	c (β)	d	c+d
	a+c	b+d	a+b+c+d



- **Sensitivity = a / (a+c) = p(T+|D+)**
- **Specificity = d / (b+d) = p(T-|D-)**
- **Positive Predictive Value = a / (a+b) = p(D+|T+) = 1- FDR**
- **Negative Predictive Value = d / (c+d) = p(D-|T-) = power**
- *(α) rejected a true null, (β) fail to reject a false null*

Diagnostic test characteristics

예제) ***Sensitivity=99.99%, specificity=99.9%*** *인 최신의 에이즈 검사가 개발되었다. 철수는 이 검사에 양성반응을 보였다. 철수가 에이즈에 감염되었을 확률은 얼마인가? (현재 한국인의 에이즈 유병율은* ***0.0001*** *이라고 한다.)*

1. **99%**
2. **95%**
3. **80%**
4. **50%**
5. **10%**

Diagnostic test characteristics

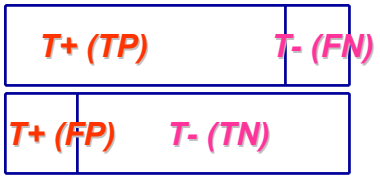
- Sensitivity=99.99%
- Specificity=99.9%

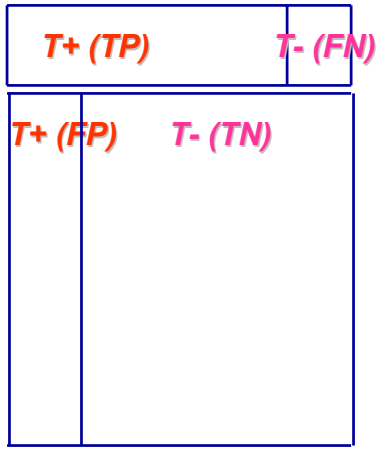
Prevalence=0.0001

	AIDS	no AIDS	
T +	9,999	1 00,000	109,999
T -	1	999 00,000	99,900,001
	10,000	1000 00,000	100,010,000

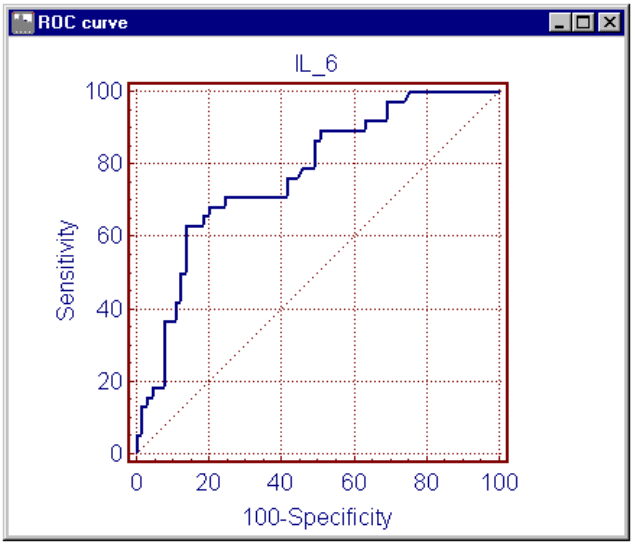
- Positive predictive value = 9,999 / 109,999 < 10%
- Negative predictive value = 1.0

Diagnostic test characteristics





ROC curve



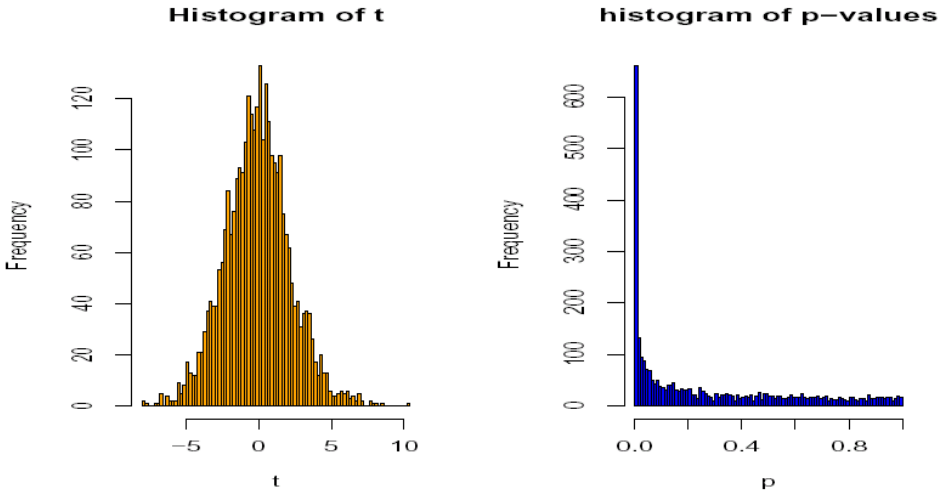
Multiple testing problem

- ***Thousands of hypotheses are tested simultaneously***
- ***The gene is not differentially expressed vs. there is no gene that is differentially expressed.***
- ***increased chance of false positives*** *(α, Type-I)*
- ***should adjust your p-values***

- ***FWER (Family-wise error rate): prob. of at least one false positive***
  - ✓ ***FWER =*** *Pr(α > 0)*

- ***FDR (False discovery rate): expected proportion of false positives among the rejected hypotheses ('95)***
  - ✓ ***FDR =*** *E(α / rejected)·Pr(rejected > 0)*
  - ✓ ***positive FDR, pFDR =*** *E(α / rejected | rejected >0)*

Multiple hypothesis testing



Biochip Informatics - (I)

- Biochip basics
- Preprocessing
- Episodes 1 and 2
- Global normalization
- Intensity dependent normalization
- Controlling regional variation
- Alternatives
- Differential expression
- **Multiple hypothesis testing**
- Classification

Multiple testing in microarray

- *Microarray experiments are large and exploratory*
- *5% of FDR says that, among the 100 genes said significant about 95 may be truly significant.*
- *What dose 5% of FWER in microarray experiments?*
- *… and compared to the top (arbitrary) 100 list?*
- *What about symmetric vs. asymmetric rejection regions?*
- *Robustness against the kind of dependence?*
- *….*

Control of the FWER

Bonferroni correction

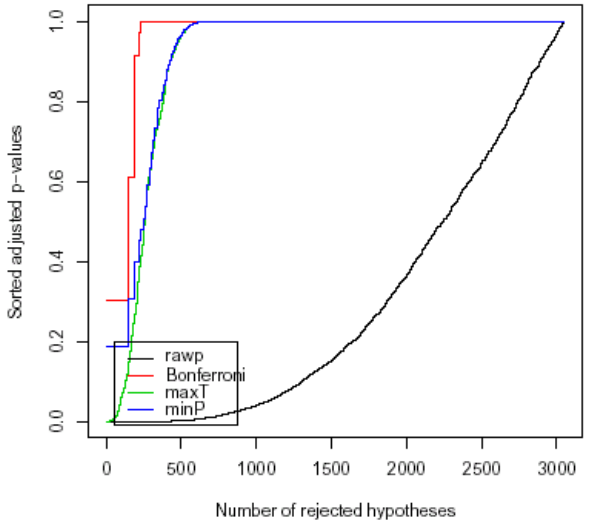
- *complete null hypothesis that there is no gene that is differentially expressed (i.e., a+c = a+b+c+d)*
- *weak control of type-I error*
  - ✓ *FWER =*
    $Pr(\alpha > 0) = Pr(\text{at least one adjusted } p_g < c | H_0)$
- *? dependence structure*

Westfall/Young's minP adjusted p-values

- *by re-sampling*
- *strong control of type-I error (i.e., regardless of a+c)*
- *step-down procedure*
- *maxT*

Control of the FWER

Golub's leukemia data



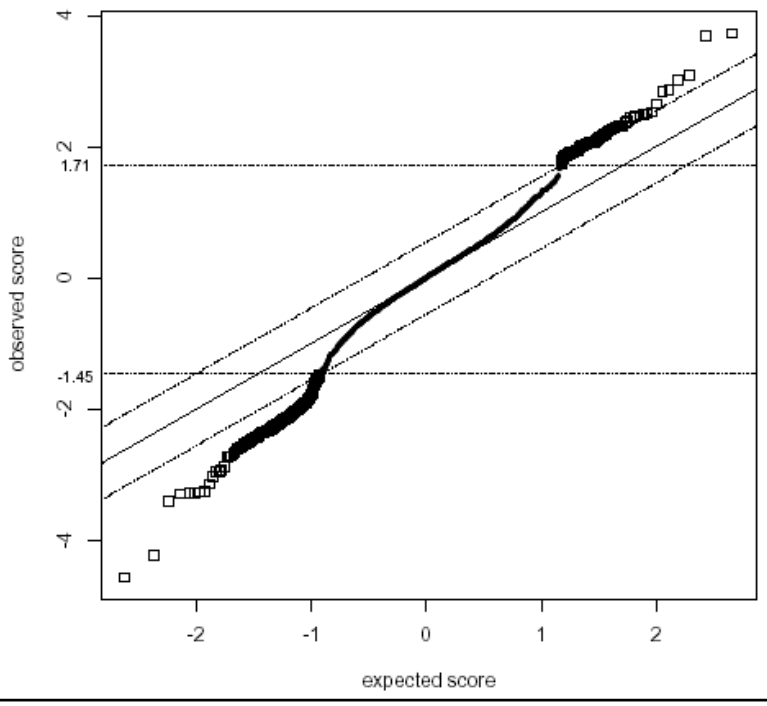
FDR and SAM

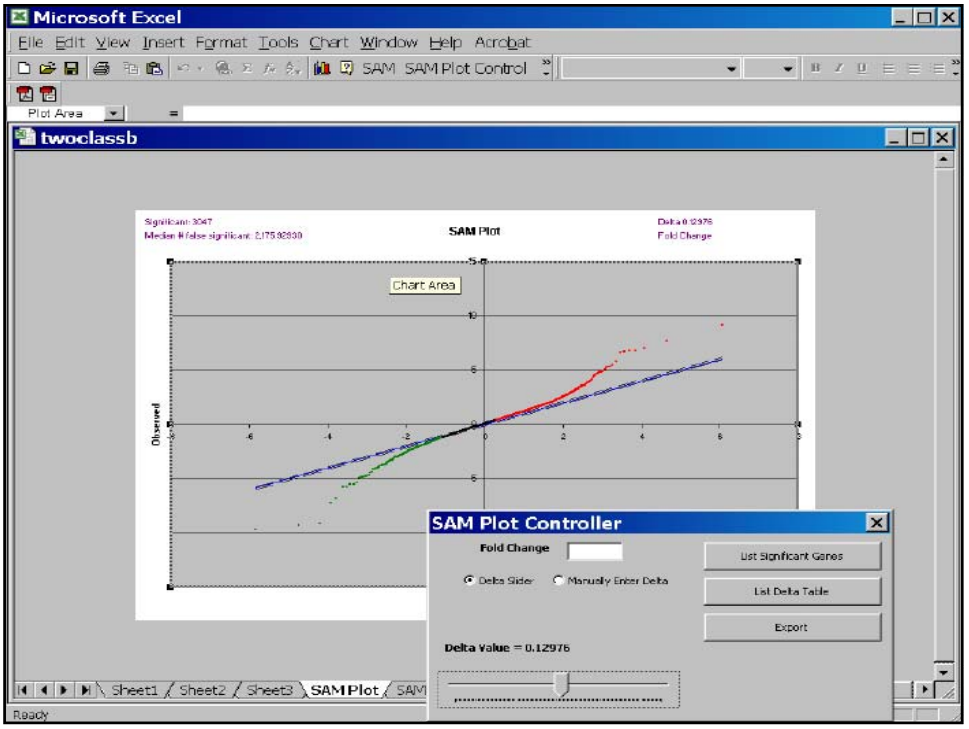
- **Statistic:** $d_j = (\bar{u}_j(2) - \bar{u}_j(1))/(s_j+s_0)$
- s_0 **: fudge factor**
  - ✓ ***a small positive number choosen as the percentile of the*** s_i ***values that makes the C.V. of*** d_j ***approximately constant as a function of*** s_i***.***
  - ✓ ***Also dampens large values of that arise from genes with low expression***
- ***Estimate the number*** m_0 ***of invariant genes.***
- ***Taking B permutations, compute the mean number of significant genes in B.***
- $\hat{E}(V) = \#\{d_j$***: gene***$_j$ ***unchanged and*** $d_j \geq t_2$ ***or*** $d_j \leq t_1\} \bullet m^*_0/m$
- ***Estimated FDR =*** $\hat{E}(V)/R$

The SAM procedure

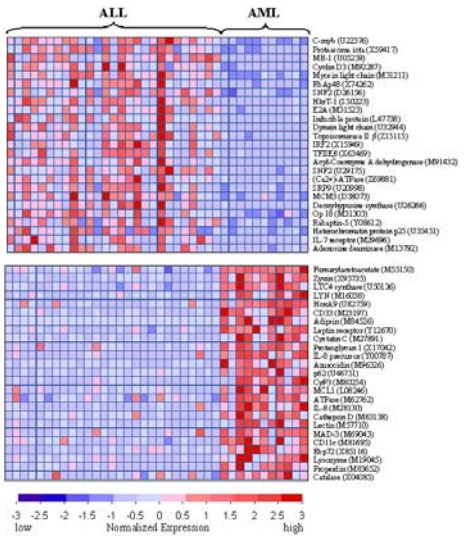
- **Compute the ordered statistics** $d_{(1)} \leq d_{(2)} \ldots \leq d_{(J)}$
- **For b=1,…, B (randomly) permute the class labels, compute test statistic** d_j^{*b} **and the corresponding order statistics** $d_{(1)}^{*b} \leq d_{(2)}^{*b} \ldots \leq d_{(J)}^{*b}$

  From the set of B permutations, estimate the expected order statistics by $\hat{d}_{(j)} = (1/B)\Sigma_b d_{(j)}^{*b}$ **for j=1,…,J**
- **Plot the** $d_{(J)}$ **values versus the** $\hat{d}_{(j)}$ **. For fixed threshold** Δ **determine** $t_1(\Delta)$ **and** $t_2(\Delta)$**.**
- **Estimate the FDR**





Molecular Classification of Cancer

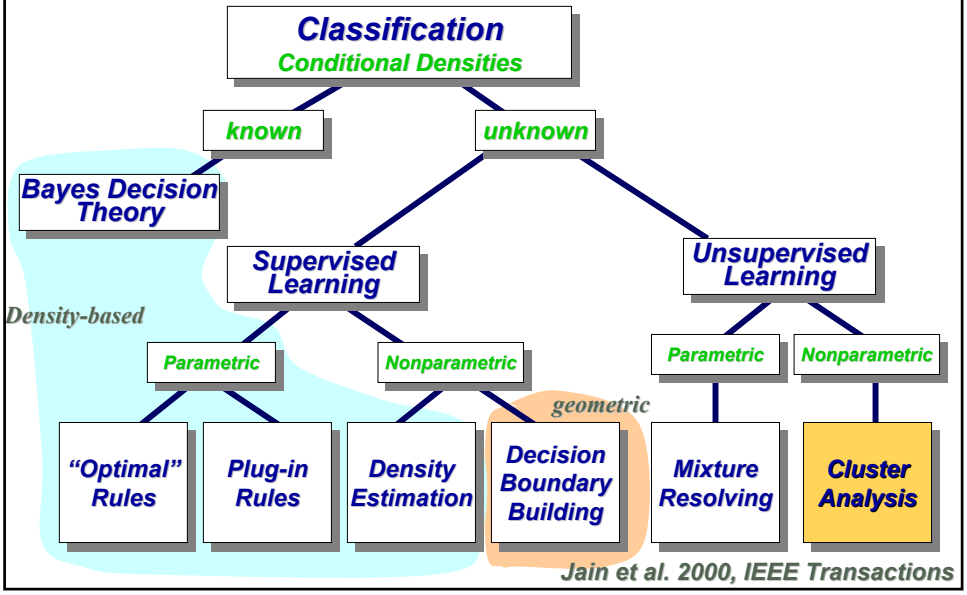


Golbe, et al., Science 1999;403:503-511

Machine Learning Approach in Bioinformatics

- **Supervised Machine Learning**
  - Linear Discriminant Analysis / Logistic Regression / PCA
  - Classification Tree
  - Artificial Neural Network
  - Support Vector Machine
  - Rough Sets
  - Reinforcement Learning
  - Hidden Markov Model
- **Unsupervised Machine Learning**
  - Hierarchical Tree Clustering
  - Partitional Clustering
  - Self-Organizing Feature Maps
  - Matrix Incision Algorithms

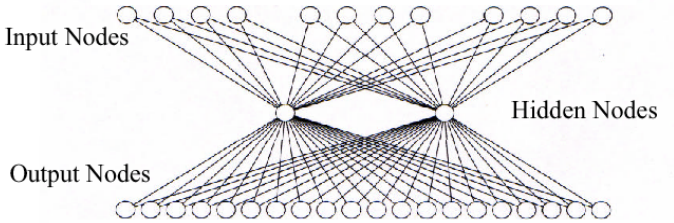
Supervised vs unsupervised classifications



Jain et al. 2000, IEEE Transactions

Artificial Neural Network

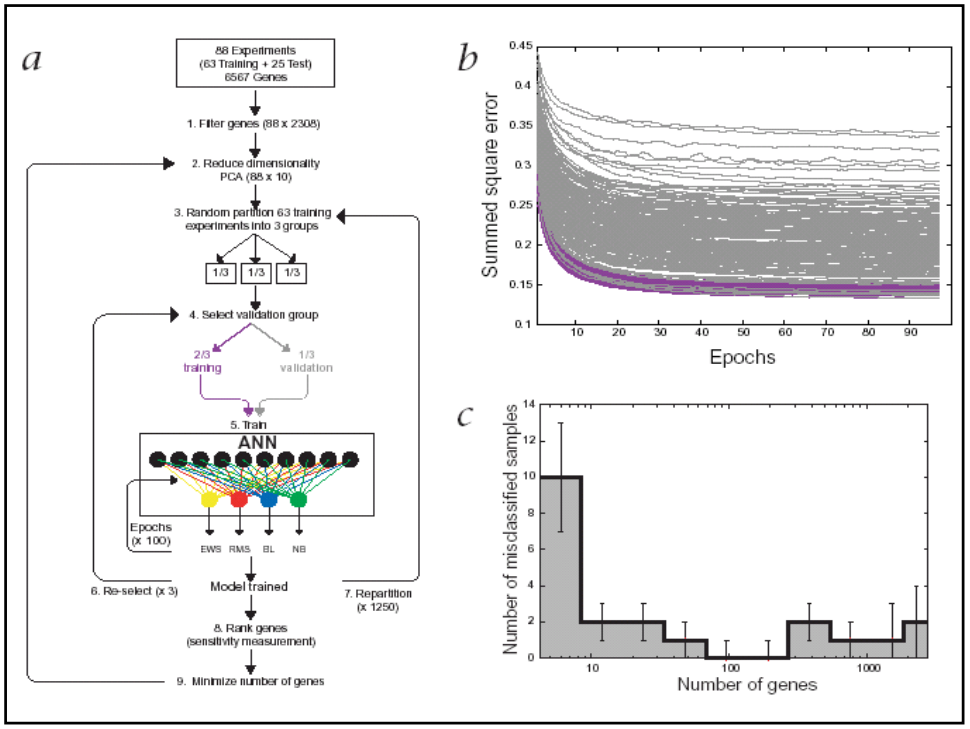
A Universal Function Approximator



$$y_i = f_i(x_i)$$

$$x_i = \sum_{j \in N(i)} w_{ij}\, y_j + w_j$$

$$y_i = f_i(x_i) = f_i\{ \sum_{j \in N(i)} w_{ij}\, y_j + w_j \}$$



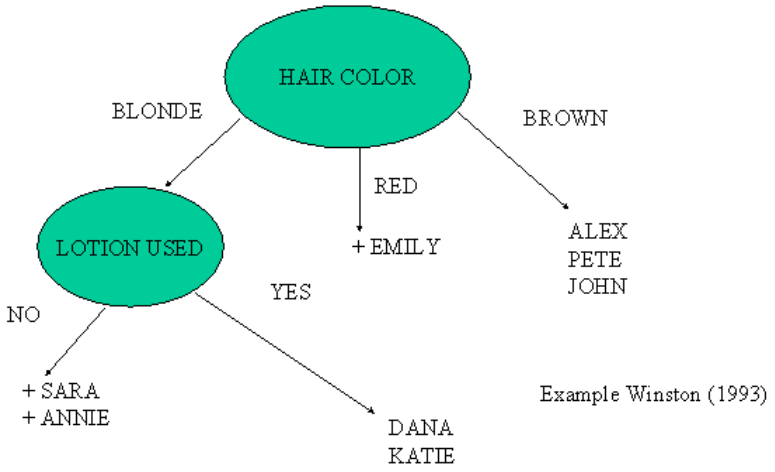
Classification tree as an ex. of classification

Sunburn at the Beach

NAME	HAIR	HEIGHT	WEIGHT	LOTION	RESULT
Sarah	Blonde	Average	Light	No	sunburn
Dana	Blonde	Tall	Average	Yes	None
Alex	Brown	Short	Average	Yes	None
Annie	Blonde	Short	Average	No	sunburn
Emily	Red	Average	Heavy	No	sunburn
Pete	Brown	Tall	Heavy	No	None
John	Brown	Average	Heavy	No	None
Katie	Blonde	Short	Light	Yes	None

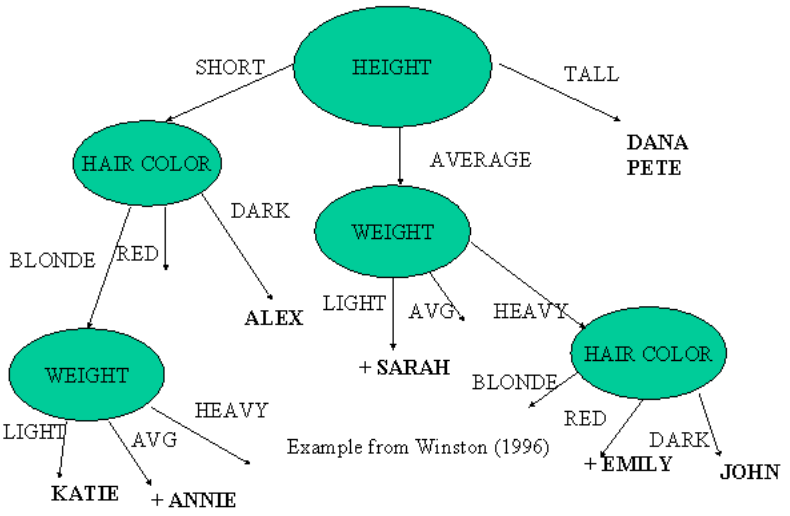
Example from Winston (1996)

Classification tree as an ex. of classification



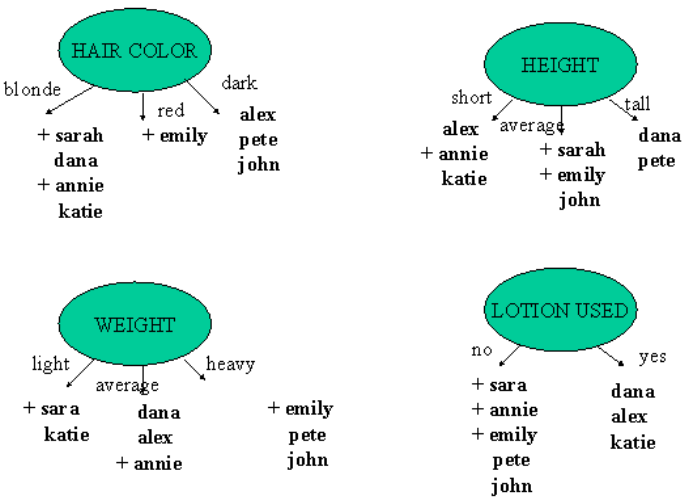
Example Winston (1993)

Classification tree as an ex. of classification



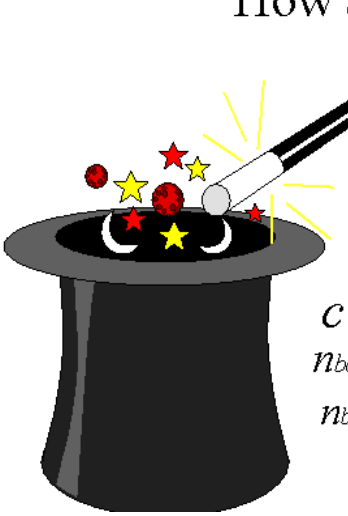
Example from Winston (1996)

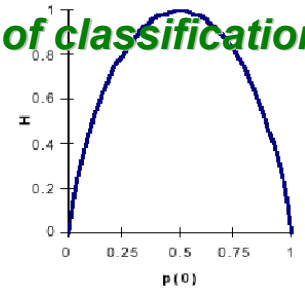
Classification tree as an ex. of classification



Example from Winston (1993)

Classification tree as an ex. of classification





$$\sum_{c} -\left(n_{bc} / n_b\right) * \log 2\left(n_{bc} / n_b\right)$$

c	Represents classes in group b
n_{bc}	Number cases of class c in group b
n_b	Number of cases in group b

Classification by the shrunken centroids

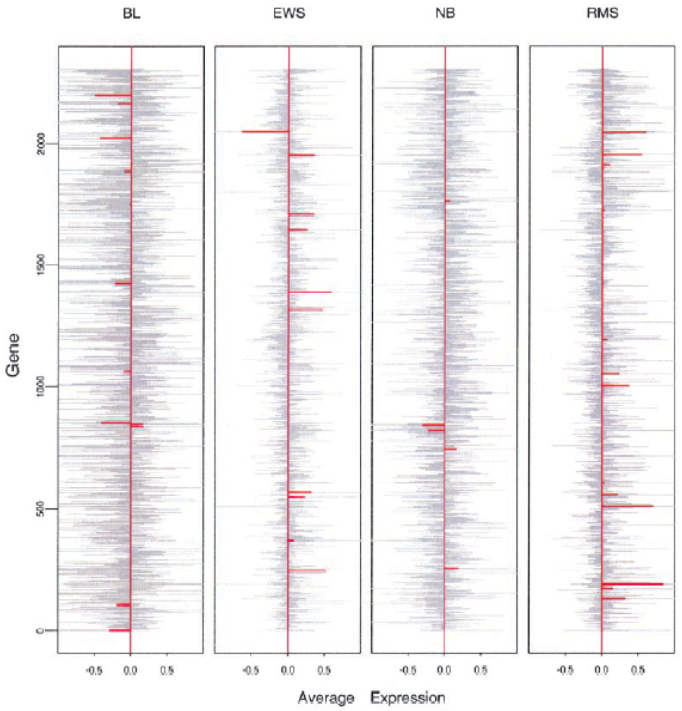
For the K class problem, i genes and j slides
- n_k: number of samples in class k
- C_k : indices of the n_k samples
The mean expression value of gene i in class k

Let

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \cdot (s_i + s_0)}, \qquad [1]$$

where s_i is the pooled within-class standard deviation for gene i:

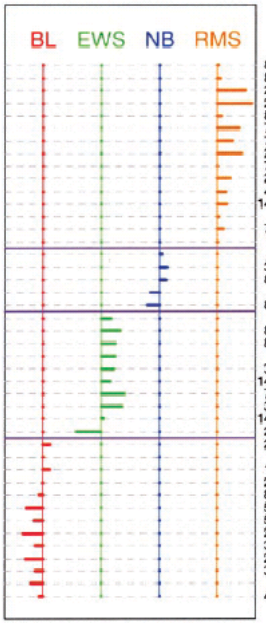
$$s_i^2 = \frac{1}{n - K} \sum_k \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2, \qquad [2]$$



shrunken centroids

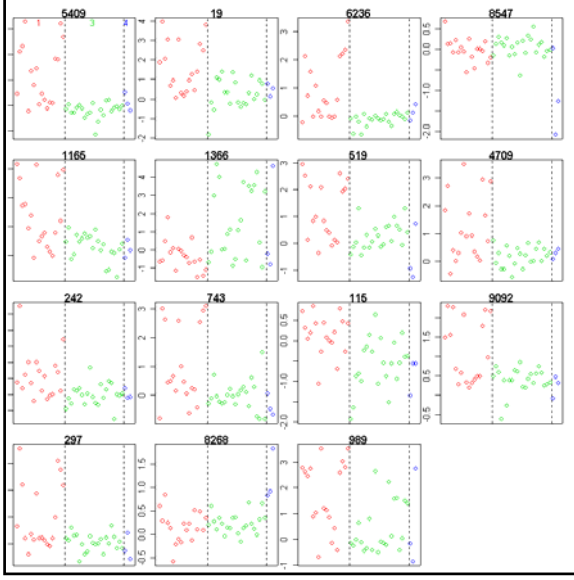
Fig. 1. Centroids (grey) and shrunken centroids (red) for the SRBCT dataset. The overall centroid has been subtracted from the centroid from each class. The horizontal units are log ratios of expression. From left to right, the numbers of training samples for each class are 8, 23, 12, and 20. The order of the genes is arbitrary.

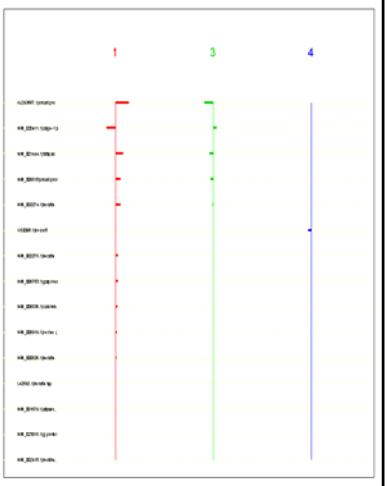
classification exercise



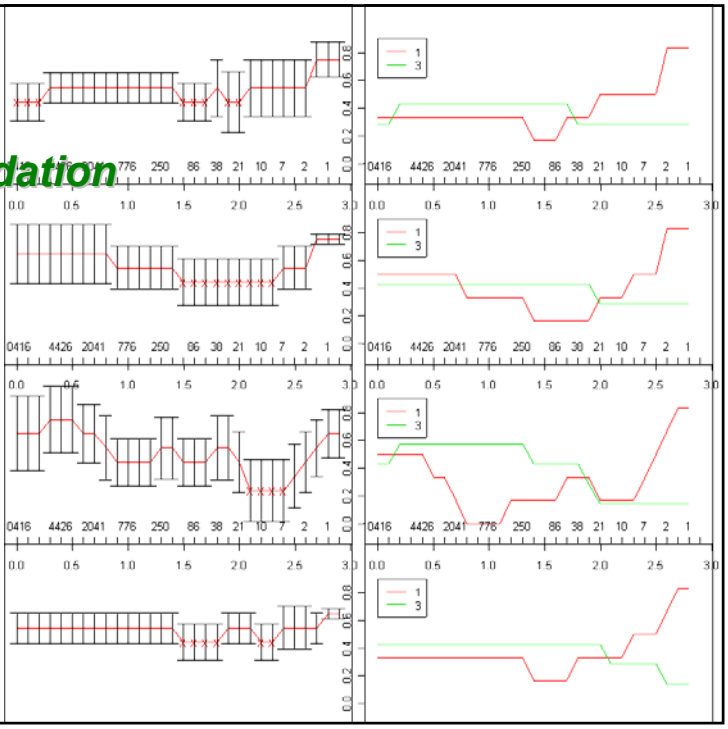
813841 tissue plasminogen activator
859359 quinone oxidoreductase homolog
207274 insulin-like growth factor 2
296448 insulin-like growth factor 2 (somatomedin A)
898219 homolog of mouse mesoderm specific transcript
784224 fibroblast growth factor receptor 4
796258 sarcoglycan alpha (dystrophin-associated glycoprotein)
244618 EST
769253 presenilin 2 (Alzheimer disease 4)
298062 troponin T2, cardiac muscle isoforms
461425 myosin MYL4
1409509 troponin T1, slow skeletal muscle isoforms
42558 L-arginine:glycine amidinotransferase
769716 neurofibromin 2 (mutated in neurofibromatosis type 2)
25725 farnesyl-diphosphate farnesyltransferase 1
44563 growth associated protein 43 (GAP43)
325182 N-cadherin (neuronal)
812105 ALL1-fused gene from chromosome 1q
41591 meningioma 1 (disrupted in balanced translocation)
810057 cold shock domain protein A
52076 neuroblastoma protein (NOE1)
866702 Fas-associated protein tyrosine phosphatase 1
814260 follicular lymphoma variant translocation protein 1
43733 glycogenin 2
357031 tumor necrosis factor alpha-induced protein 6
1435862 MIC2 surface antigen (CD99)
770394 IgG Fc fragment receptor transporter, alpha chain
377461 caveolin 1 (caveolae protein)
1473131 transducin-like enhancer of split 2
295985 EST
241412 E74-like factor 1 (ets domain transcription factor)
80109 major histocompatibility complex, class II, DQ alpha 1
183337 major histocompatibility complex, class II, DM alpha
233721 insulin-like growth factor binding protein 2
897788 receptor type protein tyrosine phosphatase F
563673 antiquitin 1
504791 glutathione S-transferase A4
212542 cDNA DKFZp586J2110
365826 growth arrest-specific protein 1
204545 EST
308163 EST
21652 alpha 1 catenin (cadherin-associated protein)
486110 profilin 2

Feature selection





K-fold cross-validation



Cross-validation probabilities