STATISTICAL CHALLENGES IN GENE DISCOVERY THROUGH MICROARRAY DATA ANALYSIS

Sreekumar. J[1] and Jose. K. K.[2]

[1]Central Tuber Crops Research Institute,Kerala, India
[2]Dept. of Statistics, St. Thomas College, Pala, Kerala, India

email:sreejyothi_in@yahoo.com

Introduction

- In 1953, Francis Crick and James Watson: double-helical structure of DNA Figure
- A gene is a fragment of DNA: (start codon, e.g. AUG or GUG) , (stop codon, e.g. UAG etc)
- A genome can be comprised of thousands of genes.
- A protein is a sequence of 20 different types of amino acids
- The process of protein synthesis: Central dogma of molecular biology Figure
  - transcription
  - translation

Introduction....

- DNA microarrays are devices that measure the gene expression of many thousands of genes simultaneously
- The new data promise to enhance fundamental understanding of life on a molecular level and may prove useful in medical diagnosis, treatment and drug design
- Novel computational tools and reliable data processing are essential for the meaningful and accurate interpretation of microarray data
- In the context of expression-based classification based on a large number of genes, the primary interest is to select a small set of genes, which have a good prediction performance

Introduction....

- DNA microarrays are devices that measure the gene expression of many thousands of genes simultaneously
- The new data promise to enhance fundamental understanding of life on a molecular level and may prove useful in medical diagnosis, treatment and drug design
- Novel computational tools and reliable data processing are essential for the meaningful and accurate interpretation of microarray data
- In the context of expression-based classification based on a large number of genes, the primary interest is to select a small set of genes, which have a good prediction performance

Introduction....

- DNA microarrays are devices that measure the gene expression of many thousands of genes simultaneously
- The new data promise to enhance fundamental understanding of life on a molecular level and may prove useful in medical diagnosis, treatment and drug design
- Novel computational tools and reliable data processing are essential for the meaningful and accurate interpretation of microarray data
- In the context of expression-based classification based on a large number of genes, the primary interest is to select a small set of genes, which have a good prediction performance

Introduction....

- DNA microarrays are devices that measure the gene expression of many thousands of genes simultaneously
- The new data promise to enhance fundamental understanding of life on a molecular level and may prove useful in medical diagnosis, treatment and drug design
- Novel computational tools and reliable data processing are essential for the meaningful and accurate interpretation of microarray data
- In the context of expression-based classification based on a large number of genes, the primary interest is to select a small set of genes, which have a good prediction performance

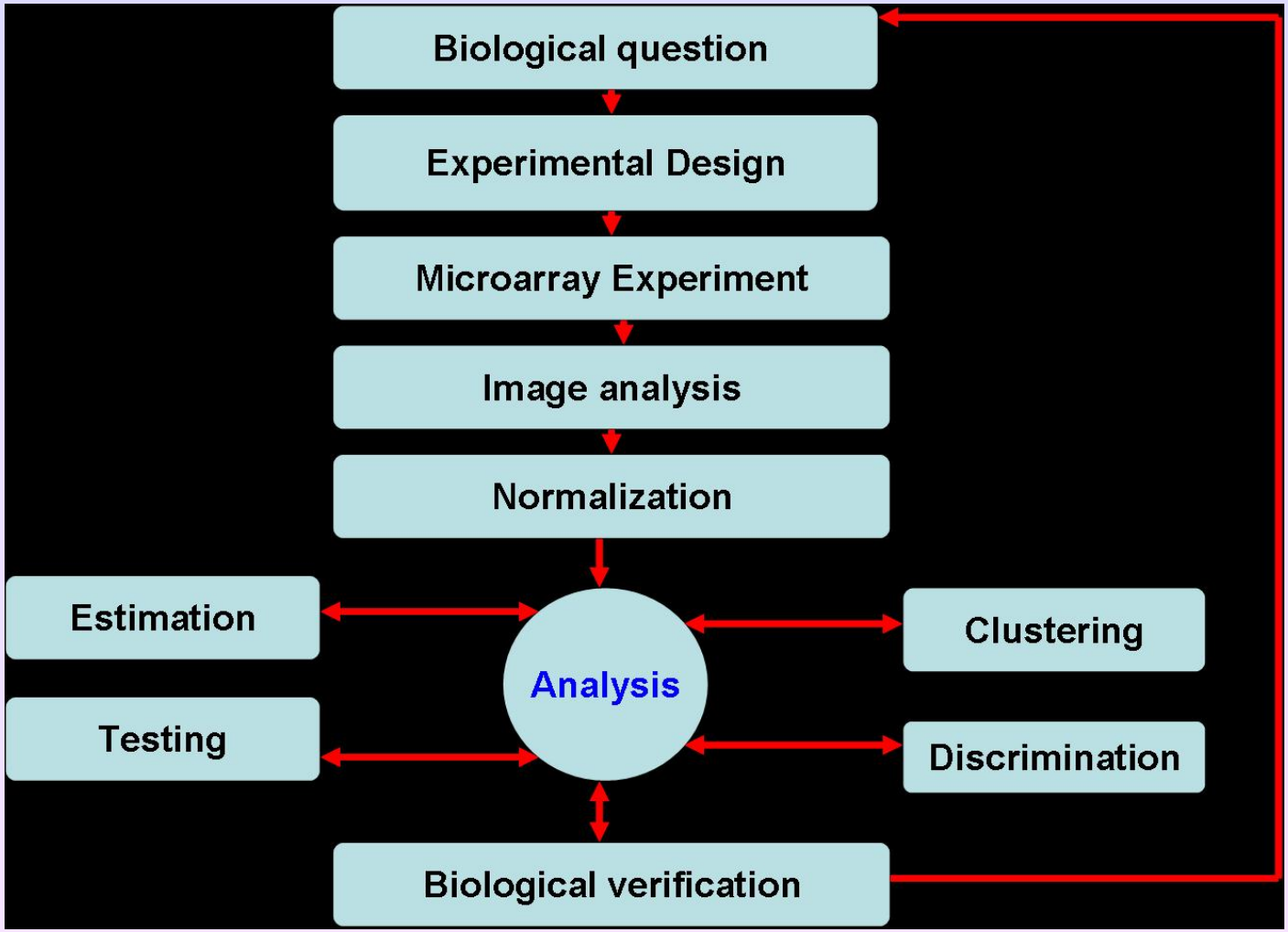


Figure: Steps in the analysis of Microarray data

cDNA Microarray Experimental Procedure

- The process consists of four steps:
  - printing
  - preparation of the biological material
  - hybridization
  - image processing and data analysis
- Selection of a set of genes of interest
- DNA clones of interest (ESTs) are then amplified by PCR to generate a sufficient amount to allow printing onto a glass microslide(Probes)

- The printing is made by an arrayer
- An arrayer is a robot with a certain number of pins programmed to deposit EST aliquots in an array configuration Figure
- sample preparation,labelling: isolating RNA (e.g., samples from plants under normal condition and plants under drought stress),(Targets)
- The samples are then mixed and hybridized to the arrayed DNAs on the glass slide and any unhybridized cDNA is washed out

- Slides are imaged using a laser scanner
- Typically, the two measurements (one for red and the other for green) for each spot indicate the relative abundance of the corresponding mRNA in the two tissue samples Figure
- The scanned images are analyzed using image analysis software, which evaluates the expression of a gene by quantifying the ratio of the fluorescence intensities of a spot

Representation of gene expression data

- Gene expression data can be represented as a real matrix

$$\begin{matrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{21n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{matrix}$$

- Each row in the matrix contains the expression data regarding a specified gene gene expression profile
- The vector in the j^{th} column of the matrix X, lists the genome wide fluorescence ratios measured by the j^{th} array
- Each array represents a condition, different time periods or tissue profile

Statistical issues...

- The expression levels are noisy
- Noise creaps into the data in all stages
- accidental / controllable (Preparation of cells, environmental)
- systematic source of variation
  - array
  - spotter
  - dye
  - imaging

Preprocessing

- normalization and filtering
  - Discard genes where experiments went visibly wrong
  - Compute ratio between red and green to account for spot effect
  - Take logarithm
  - normalize to zero mean
- Variance stabilization(Transformation)
- whether the data need transformation and which?

Visualization tools

Helps in interpreting the results of microarray experiments. The most commonly used of these are illustrated

- Heatmaps : small cells, each consisting of a colour, which represent relative expression values
- Box plots: present various statistics for a given data set. The plots consist of boxes with a central line and two tails
- MA plots: They are often used as an aid when normalizing two-colour cDNA microarrays, where $M = \log_2 \frac{R}{G}$ and $A = \log_2 \sqrt{(RG)}$
- Volcano plots: are used to look at fold change and statistical significance simultaneously
- p-value histograms: the p-values for a test of differential expression for each gene Figure

Error Distribution of Gene Expression data

- Gene expression data contains a large number of low expressed genes whose expression level tends to follow a skewed and definitely non normal distribution
- The lognormal distribution is frequently used to model positively skewed gene expression distributions.
- we applied the asymmetric Laplace distribution Figure
- The distribution of error helps in further interpretation and normalization techniques

Gene selection

- One of the important problems to be addressed in analysis of microarray data is the identification of differentially expressed gene for further investigation.
- Fold change is the simplest method for identifying differentially expressed genes
- It is known to be unreliable because statistical variability was not taken into account
- Fold change method is subject to bias if the data have not been properly normalized

t-test

- Classic statistical approaches used for detecting differences between two groups include the parametric t-test

$$t_{n_1+n_2-2} = \frac{\bar{X_1} - \bar{X_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_1}}} \quad (1)$$

- t-test utilize the variance
- However the small sample sizes : can affect the variance estimates.

- If the two groups have different variances, then the two-sample unequal variance t-statistic will be more appropriate.
- Welch method proposes an elaborate correction for degrees of freedom under unequal variances of samples

$$\nu = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \Big/ \frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2 - 1)} \tag{2}$$

where s_1^2, s_2^2, n_1 and n_2 are the sample variances and the number of observations in each category.

To address the shortcomings of the t-test

$$t^* = \frac{\bar{M}}{(s + a)/\sqrt{n}} \tag{3}$$

- Efron *et al.*, percentile of the distribution
- The SAM t-test (S-test) : gene variances
- The regularized t-test : gene-specific and global average variance: weighted average of the two
- Broberg : Minimizing a combination of estimated false positive and false negative rates over a grid of significance levels and varinace

- Baldi and Long proposed : t-statistic with a Bayesian adjusted denominator cyber T program , moderated t.

$$t^* = \frac{\bar{M}}{\tilde{s}/\sqrt{n}} \tag{4}$$

where

$$\tilde{s}^2 = \frac{\nu_0 \sigma_0^2 + (n-1)s^2}{\nu_0 + n - 2}$$

is the regularized standard deviation. ν_0 is the strength of the prior and σ_0^2 is the background variance estimated from all genes or a set of subset genes.

- Smyth extended this to linear model set up

Generalized p value technique

Let $X_{gij}, i = 1, 2; j = 1, 2..., n_i$ and $g = 1, 2..., n_g$ denote the random samples of gene expression data assumed to follow lognormal distribution. Let $Y_{gij} = \ln(X_{gij})$, Define,

$$\bar{Y_{gi}} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad \text{and} \quad S_{gi}^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(Y_{gij} - \bar{Y_{gi}}\right)^2, i = 1, 2. \tag{5}$$

for each gene $g, g = 1, 2, ..., n_g$. Further let $\bar{y_{g1}}$, $\bar{y_{g2}}$, s_{g1}^2 and s_{g2}^2 denote the observed values of $\bar{Y_{g1}}$, $\bar{Y_{g2}}$, S_{g1}^2 and S_{g2}^2 respectively.

- Let

$$T_{gi} = \bar{y_{gi}} - \frac{\bar{Y_{gi}} - \mu_{gi}}{S_{gi}/\sqrt{n_{gi}}} s_{gi}/\sqrt{n_{gi}} + \frac{1}{2}\frac{\sigma_{gi}^2}{S_{gi}^2} s_{gi}^2 \quad (6)$$

$$= \bar{y_{gi}} - \frac{Z_{gi}}{U_{gi}/\sqrt{n_i - 1}} \frac{s_{gi}}{\sqrt{n_i}} + \frac{1}{2}\frac{s_{gi}^2}{U_{gi}^2/(n_i - 1)}, \quad (7)$$

for $i = 1, 2$ and $g = 1, 2, ..., n_g$,
where $Z_{gi} = \sqrt{n_i}(\bar{Y_{gi}} - \mu_{gi})/\sigma_{gi} \sim N(0, 1)$
and $U_{gi}^2 = (n_i - 1)S_{gi}^2/\sigma_{gi}^2 \sim \chi_{n_i-1}^2$

- Define generalized pivotal quantity $T_g = T_g^* - (\eta_{g1} - \eta_{g2})$, where $T_g^* = T_{g1} - T_{g2}$
- Hence The generalized p–value for the two sided test can be obtained as:

$$2 \times \min\left\{P(T_g \leq 0), P(T_g \geq 0)\right\} \quad (8)$$

Other apparoaches

- Thomas *et al.,* : Regression modelling
- Non parametric (eg:Wilcoxon rank sum test)
- B statistic (log odds ratio) by Lonnstedt and Speed
- Bootstrap t–test

A typical ANOVA model under multiple conditions is

$$y_{ijkg} = \mu + A_i + D_j + AD_{ij} + G_g + AG_{ig} + VG_{kg} + DG_{jg} + \epsilon_{ijkg} \quad (9)$$

where y_{ijkg} is the measured intensity from array i, dye j, variety k and gene g

$$F_1 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{rss1/df_1} \quad (10)$$

Other F-like statistics (F_2 and F_3)

$$F_2 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{(rss1/df1 + \sigma^2_{pool})/2} \quad (11)$$

$$F_3 = \frac{(rss_0 - rss_1)/(df_0 - df_1)}{\sigma^2_{pool}} \quad (12)$$

- The mixed model ANOVA treats some of the factors in the experimental design as random samples from a population and there are multiple levels of variance (biological, array, spot and residual)
- Constructing an appropriate F-statistics using the mixed model is tricky
- We can also apply random effects models which use BLUP (best linear unbiased prediction) for the estimation of the gene expression effects

Multiple hypothesis testing

Table: Type I and Type II errors in multiple hypothesis testing.

	Null hypotheses not rejected	Null hypotheses rejected	
true (non-diff.genes)	$h_0 - V_n$	V_n (Type I errors)	h_0
false (diff.genes)	U_n (Type II errors)	$h_1 - U_n$	h_1
	$m\text{-}R_n$	R_n	m

- p-value: The p-value or observed significance level p is the chance of getting a test statistic as or more extreme than the observed one, under the null hypothesis H_0 of no differential expression.

The commonly used Type I error rates in multiple hypotheses testing are:

- Per comparison error rate (PCER): the expected value of the number of Type I errors over the number of hypotheses,

$$PCER = \frac{E(V_n)}{m}$$

- Per-family error rate (PFER): the expected number of Type I errors,

$$PFE = E(V_n)$$

- Family-wise error rate: the probability of at least one Type I error

$$FEWR = Pr(V_n = 1)$$

- False discovery rate (FDR)is the expected proportion of Type I errors among the rejected hypotheses

$$FDR = E(\frac{V_n}{R_n}; R_n > 0) = E(\frac{V_n}{R_n}|R_n > 0) \times Pr(R_n > 0)$$

- Positive false discovery rate (pFDR): the rate that discoveries are false

$$pFDR = E(\frac{V_n}{R_n}|R_n > 0)$$

There are single step procedures for controlling FWER

- Bonferroni single step adjusted *p*-values
- Sidak single-step adjusted *p*-values
- Westfall and Young single-step minP adjusted *p*-values

False Discovery Rate

FDR is defined to be the expected value of the ratio of the number of incorrectly rejected hypotheses and the total of number of rejected hypotheses. Controlling FDR proves more powerful than FWER and has become increasingly adopted for genomic studies.

Bayesian variable selection

- The literature on microarray data is mainly based on two distributions: the log-normal and the gamma distributions
- That often appear to be effective when used in a Bayesian hierarchical framework
- In the case of empirical Bayes studies, inference is usually made on some quantities related to the posterior distribution of the parameter of interest,or of a certain type of hypothesis.
- Lee *et al.,* proposed a hierarchical bayesian model and employed latent variables to specialize the model to a regression setting and applied variable selection to select differentially expressed genes

Conclusion and Future recommendations

- Multiple methods can produce list of differentially expressed genes
- Beware of multiple hypothesis testing problems
- In many areas (normalization algorithms and false-discovery-rate estimation procedures), the need for thoroughly evaluating existing techniques currently seems to outweigh the need to develop new techniques.
- There may be no single best way to represent microarray data
- More research is needed on how to best examine intersections between sets of findings and evaluate complex multi-component hypotheses. Bayesian approaches might be especially advantageous here
- For all statistical procedures, the fact that transcripts are not necessarily independent should be considered. The potential impact of this on the performance of procedures should be assessed, and ways to accommodate this are needed.

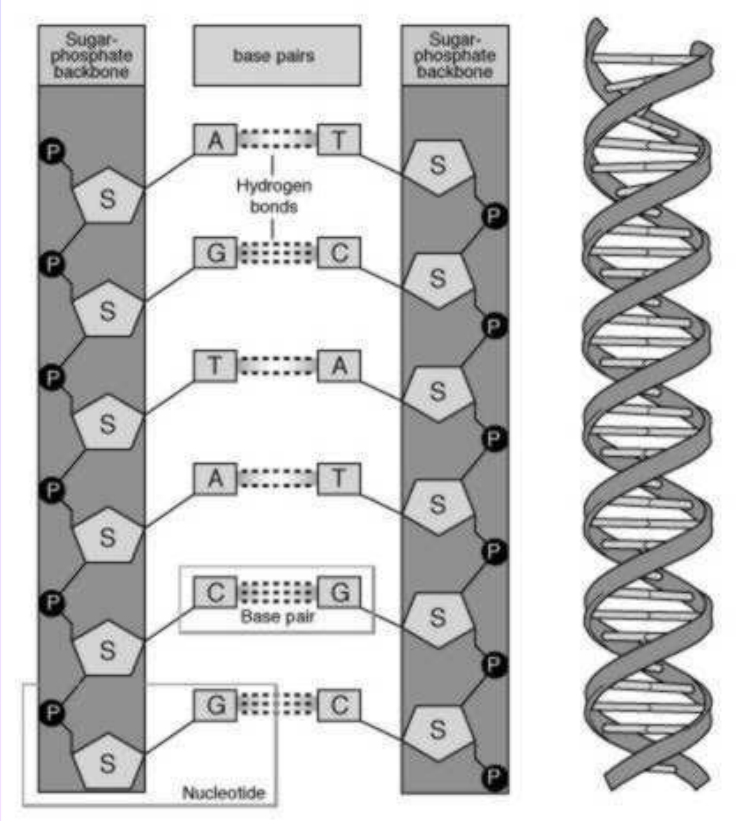
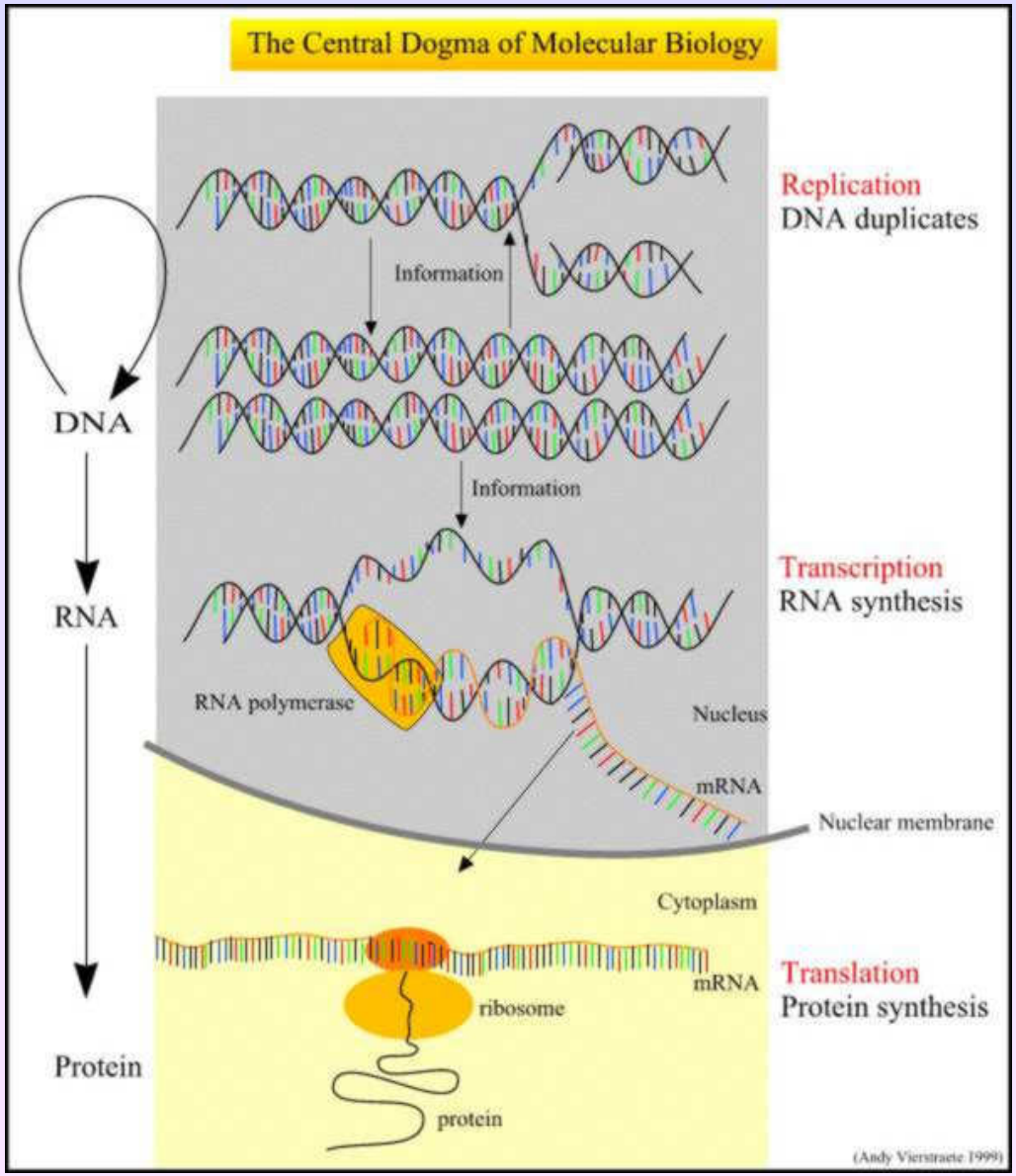


Figure: The structure of DNA

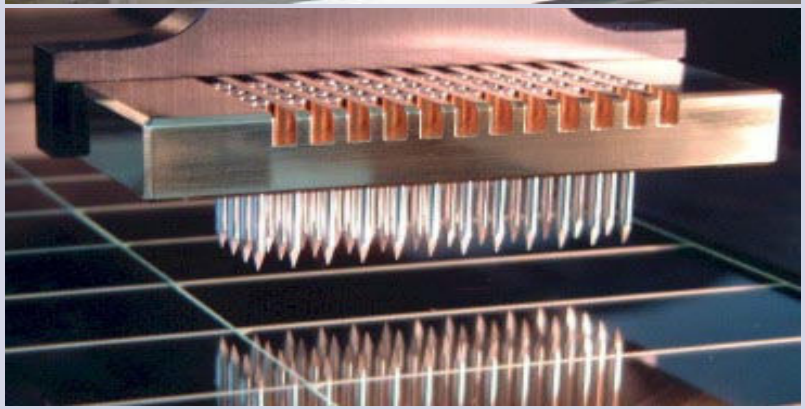
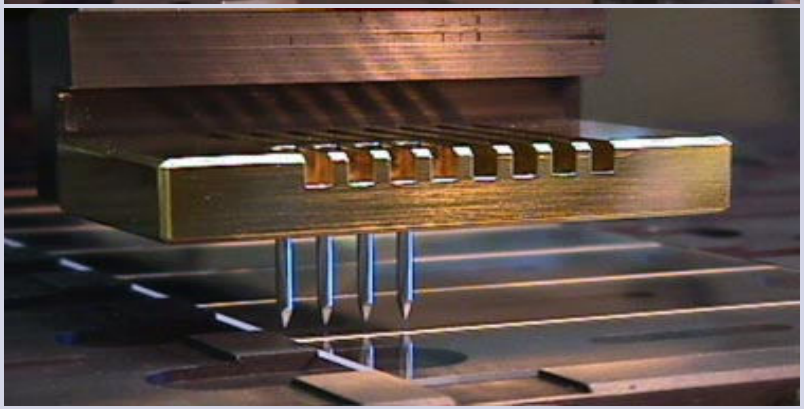
Back to presentation



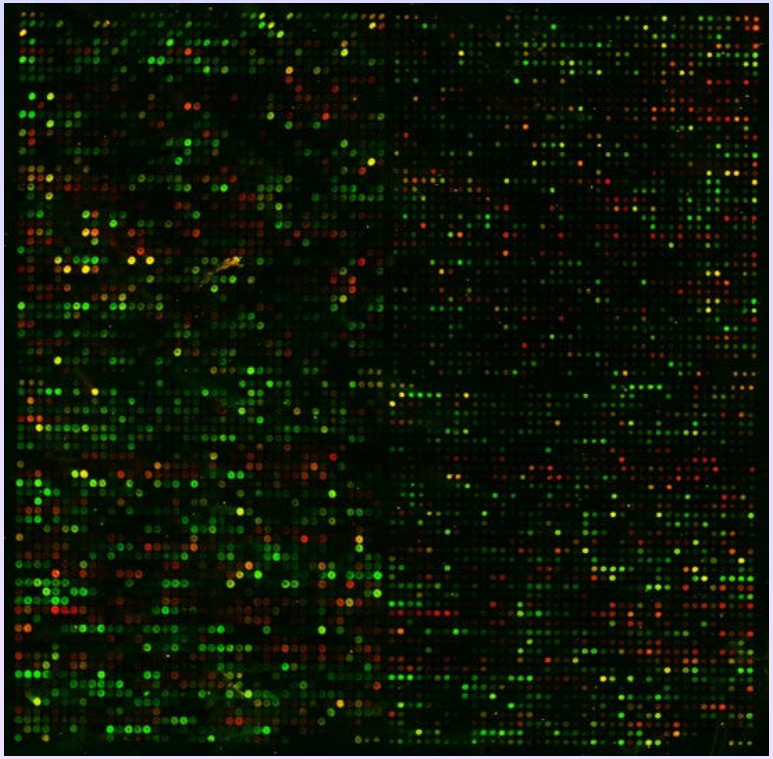
Back to presentation

Microarray spotting device and examples of commonly used print heads



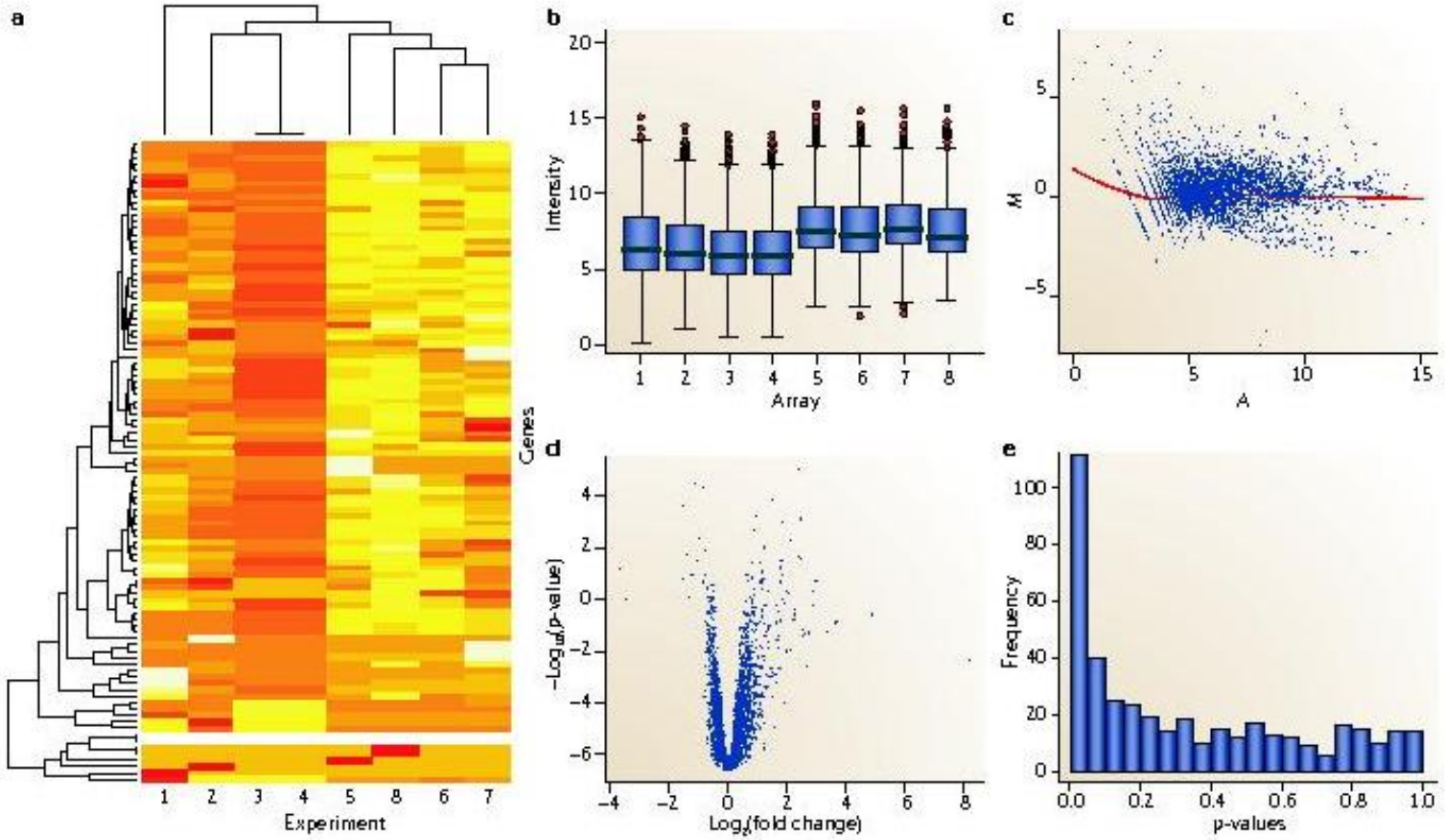



Back to presentation



- red : more expressed in test sample
- green : more expressed in reference sample
- yellow : equally expressed
- black : no expression

Back to presentation



Back to presentation

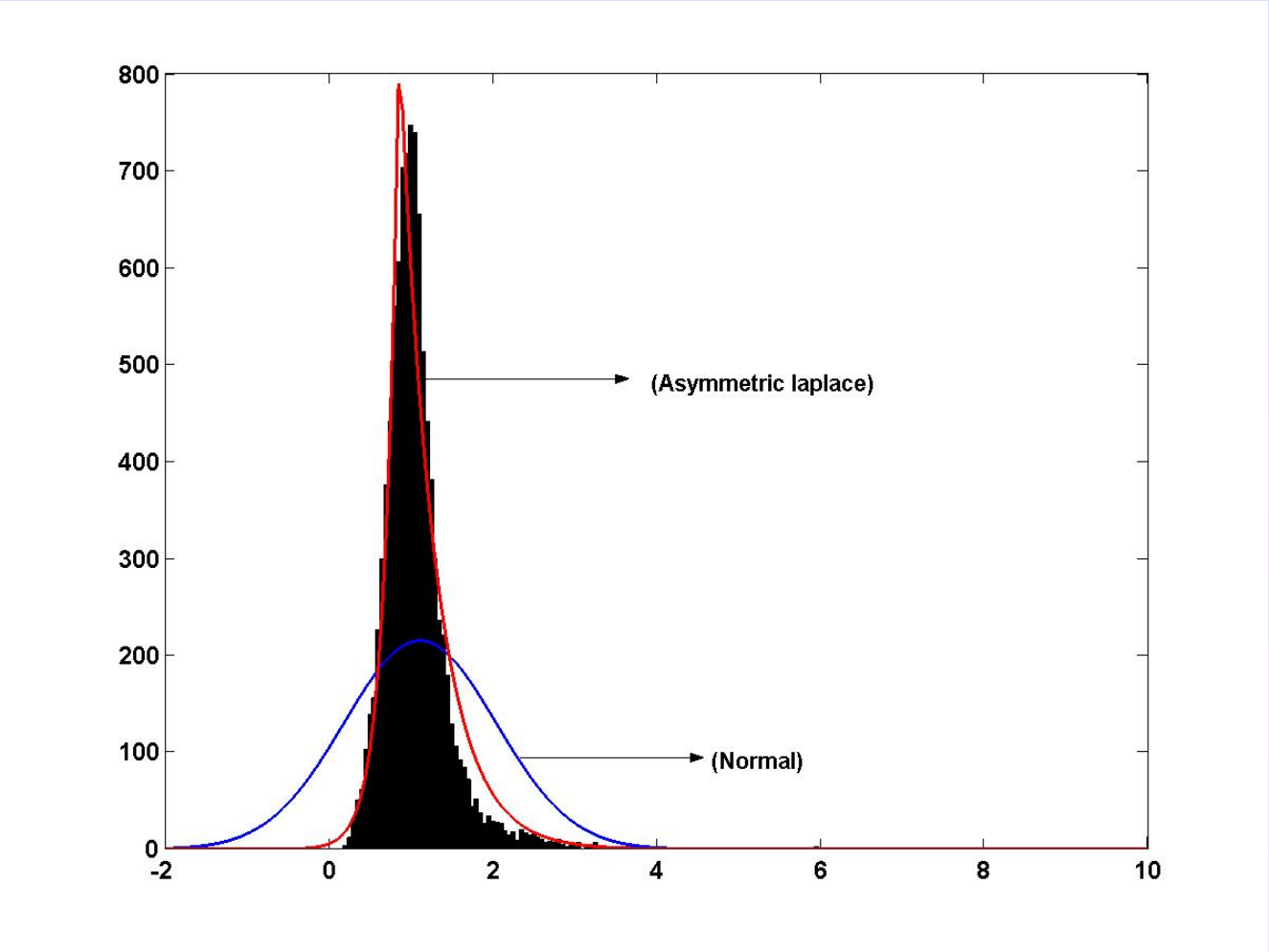


Figure: Histogram

Back to presentation