Selecting differentially expressed genes in samples subgroups on microarray data

Carina Silva

Escola Superior de Tecnologia da Saúde de Lisboa - IPL
Centro de Estatística e Aplicações da Universidade de Lisboa - CEAUL

15th November 2018

Content

Content

Content

Some genetics

- Each cell contains a complete copy of the organism's genome.

Some genetics

- Each cell contains a complete copy of the organism's genome.
- Cells are of many different types and states: blood, nerve, skin cells, dividing cells, cancerous cells, etc.

Some genetics

- Each cell contains a complete copy of the organism's genome.
- Cells are of many different types and states: blood, nerve, skin cells, dividing cells, cancerous cells, etc.
- What makes the cells different?

Some genetics

- Each cell contains a complete copy of the organism's genome.
- Cells are of many different types and states: blood, nerve, skin cells, dividing cells, cancerous cells, etc.
- What makes the cells different?
- **Differential gene expression**, i.e., when, where, and in what quantity each gene is expressed.

Some genetics

- Each cell contains a complete copy of the organism's genome.
- Cells are of many different types and states: blood, nerve, skin cells, dividing cells, cancerous cells, etc.
- What makes the cells different?
- **Differential gene expression**, i.e., when, where, and in what quantity each gene is expressed.
- On average, 40% of our genes are expressed at any given time.

Some genetics

- All cells have the same DNA.

Some genetics

- All cells have the same DNA.
- However different cells synthesize different proteins.

Some genetics

- All cells have the same DNA.
- However different cells synthesize different proteins.
- If a gene is transcribed into mRNA, then it is assumed that the gene is being expressed.

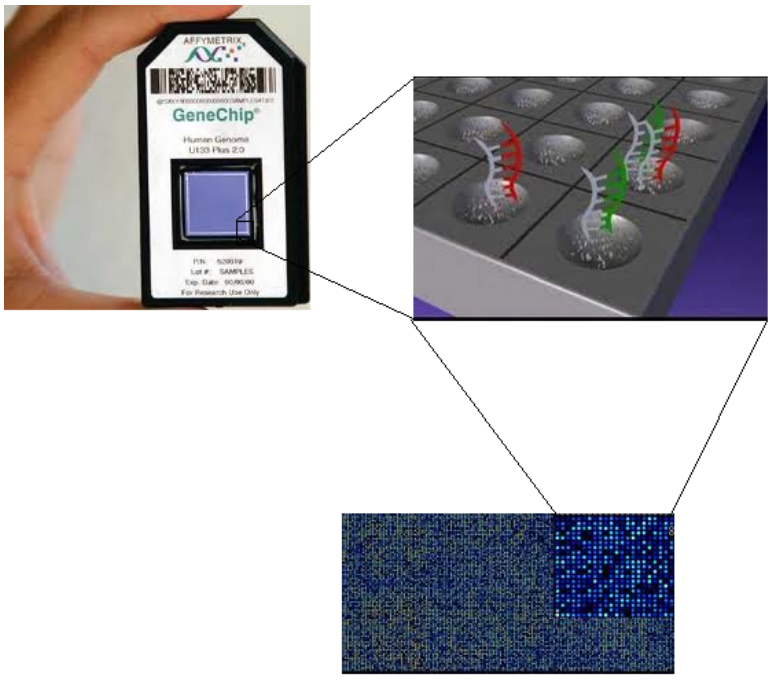
Some genetics

- All cells have the same DNA.
- However different cells synthesize different proteins.
- If a gene is transcribed into mRNA, then it is assumed that the gene is being expressed.
- The concentration of RNA in a cell defines its “biological state”.

Some genetics

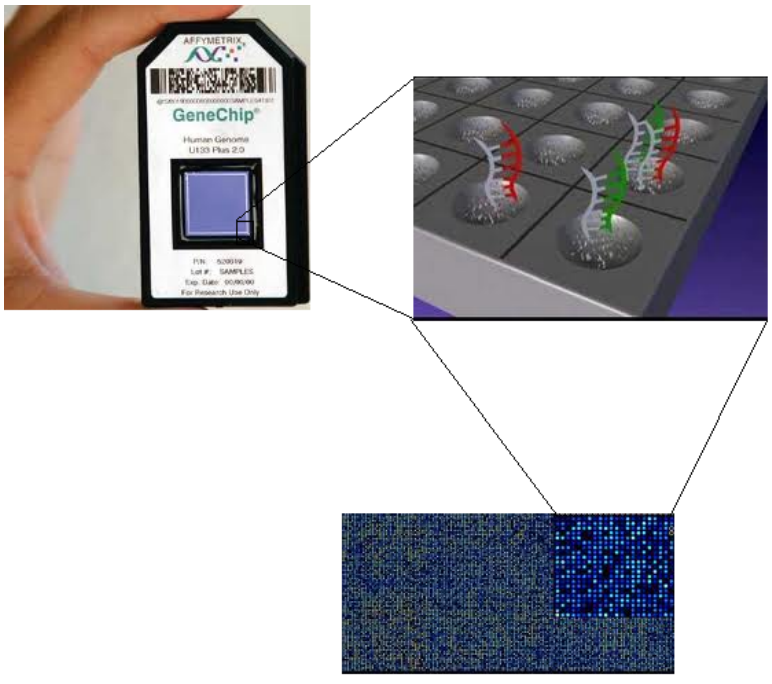
- All cells have the same DNA.
- However different cells synthesize different proteins.
- If a gene is transcribed into mRNA, then it is assumed that the gene is being expressed.
- The concentration of RNA in a cell defines its “biological state”.
- Genes differential expression is a reflection of the concentration of mRNA in the cell exposed to different experimental conditions.

Microarrays



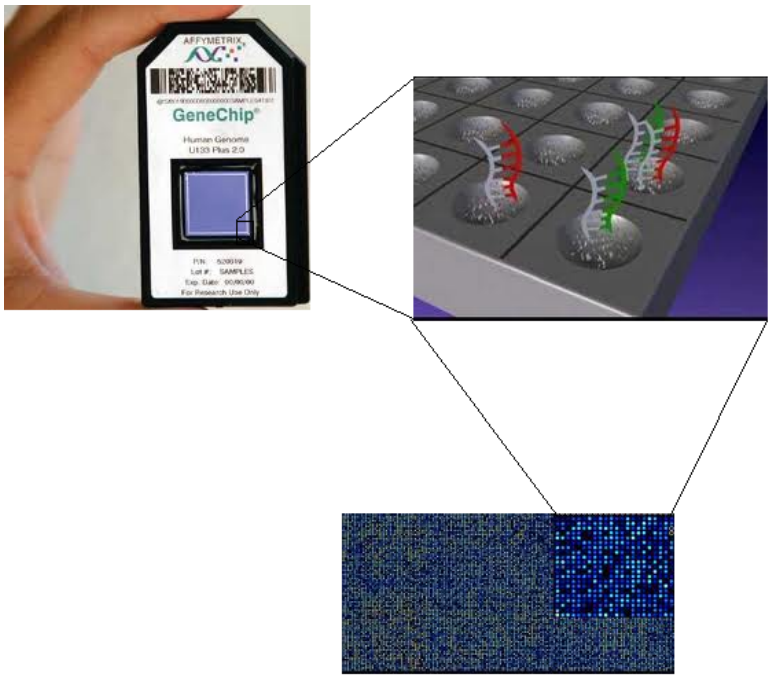
- Microarrays are a technology that allows to analyse the expression of thousands of genes simultaneously.

Microarrays



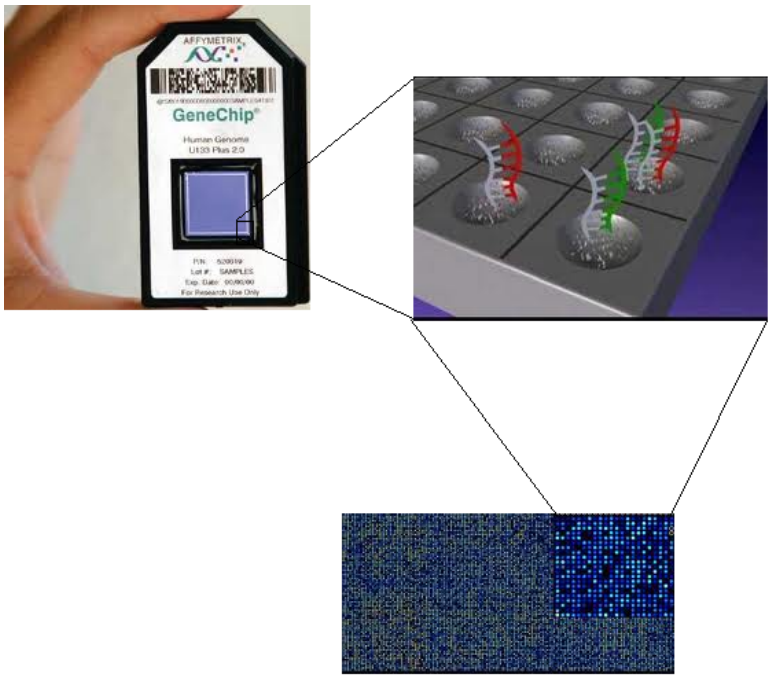
- Microarrays are a technology that allows to analyse the expression of thousands of genes simultaneously.
- The sample (target) hybridizes to the array.

Microarrays



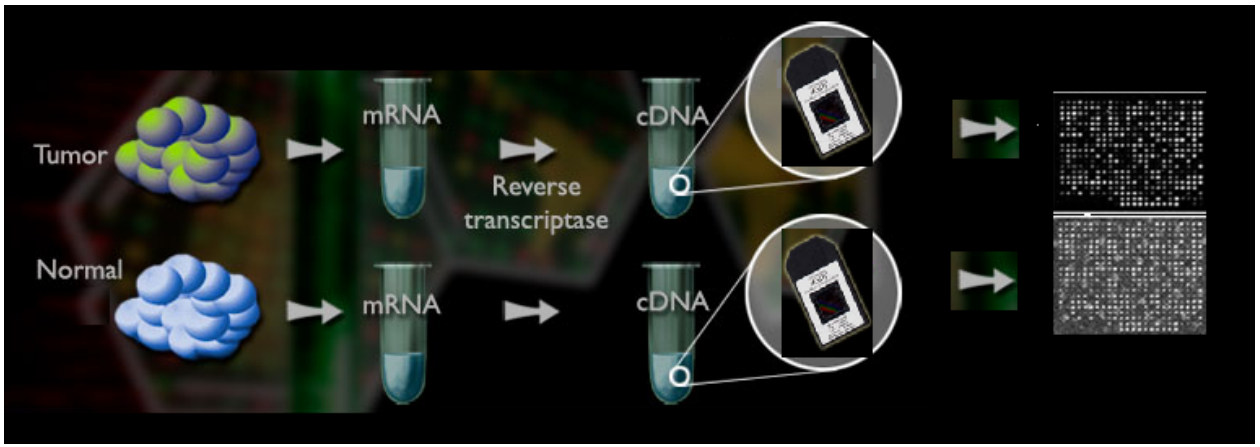
- Microarrays are a technology that allows to analyse the expression of thousands of genes simultaneously.
- The sample (target) hybridizes to the array.
- If the gene is active, the complementary target hybridizes to array probe (matching of complementary bases: A-T, G-C).

Microarrays

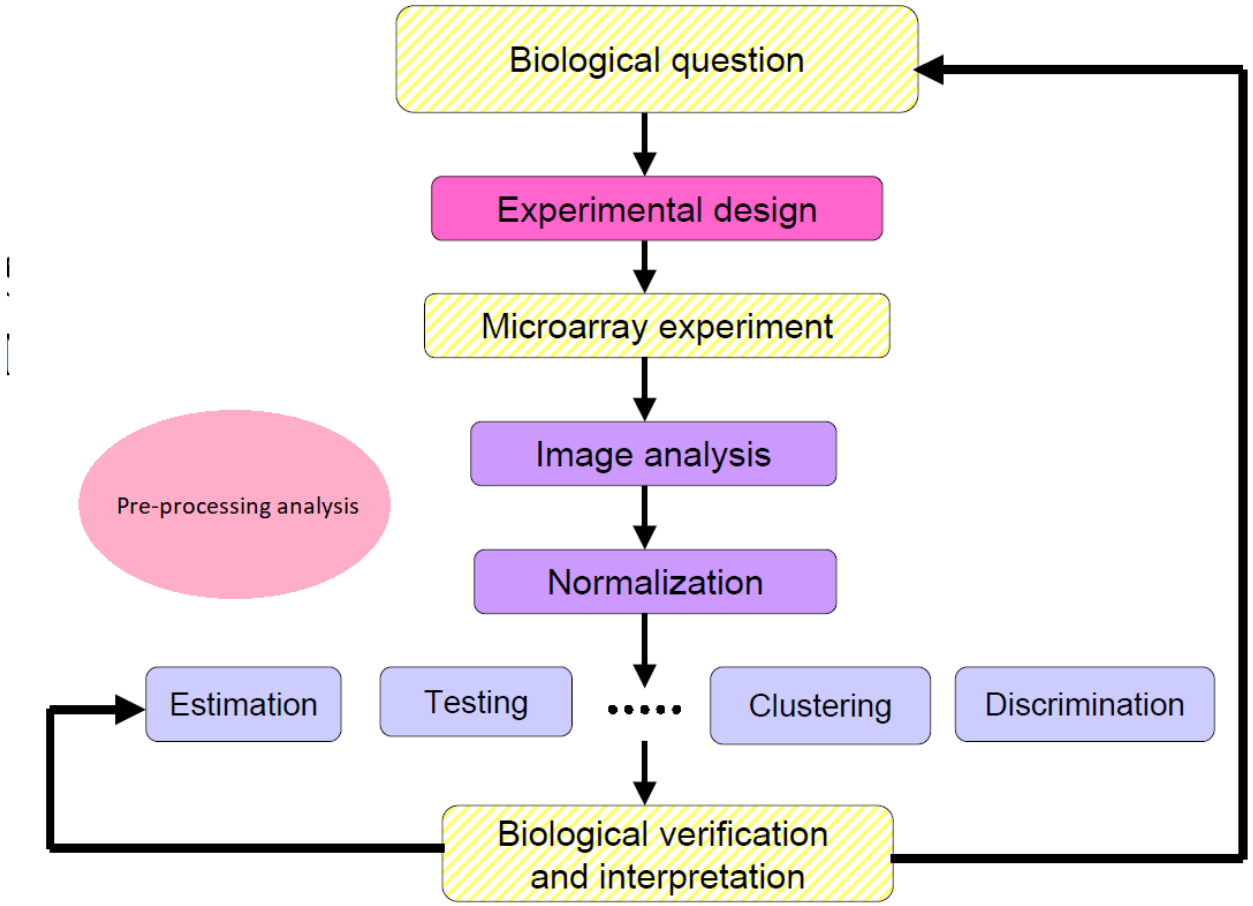


- Microarrays are a technology that allows to analyse the expression of thousands of genes simultaneously.
- The sample (target) hybridizes to the array.
- If the gene is active, the complementary target hybridizes to array probe (matching of complementary bases: A-T, G-C).
- When there is hybridization between the target and the probe, this is represented by a fluorescence.

Affymetrix Microarrays



Microarrays and statistics



Microarray Data

- The input of this initial process is the *gene expression matrix*, whose rows (1000-50000) represent genes and whose columns represent the samples (from 2 to several).

	*Group*1			*Group* 2		
	Chip_1	$\cdots$	Chip_k	Chip_{k+1}	$\cdots$	Chip_n
Gene_1	$x_{1,1}$	$\cdots$	$x_{1,k}$	$x_{1,k+1}$	$\cdots$	$x_{1,n}$
Gene_2	$x_{2,1}$	$\cdots$	$x_{1,k}$	$x_{2,k+1}$	$\cdots$	$x_{1,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Gene_p	$x_{p,1}$	$\cdots$	$x_{p,k}$	$x_{p,k+1}$	$\cdots$	$x_{p,n}$

- x_{ik}, correspond to the expression level of the probeset i of chip k after a pre-prossessing analysis and usually on log_2 scale.

Biological problem

- **Identification of subtypes of cancer on microarrays experiments**

Biological problem

- **Identification of subtypes of cancer on microarrays experiments**
- In cancer research, a common approach for prioritizing cancer-related genes is to compare gene expression profiles between cancer and normal samples, selecting genes with consistently higher expression levels in cancer samples.

Biological problem

- **Identification of subtypes of cancer on microarrays experiments**
- In cancer research, a common approach for prioritizing cancer-related genes is to compare gene expression profiles between cancer and normal samples, selecting genes with consistently higher expression levels in cancer samples.
- Such an approach ignores tumor heterogeneity and is not suitable for finding cancer genes that are overexpressed in only a subgroup of a patient population.

Biological problem

- **Identification of subtypes of cancer on microarrays experiments**
- In cancer research, a common approach for prioritizing cancer-related genes is to compare gene expression profiles between cancer and normal samples, selecting genes with consistently higher expression levels in cancer samples.
- Such an approach ignores tumor heterogeneity and is not suitable for finding cancer genes that are overexpressed in only a subgroup of a patient population.
- As a result, important genes differentially expressed in a subset of samples can be missed by gene selection criteria based on the difference of sample means.

Differentially expressed genes

Assume:

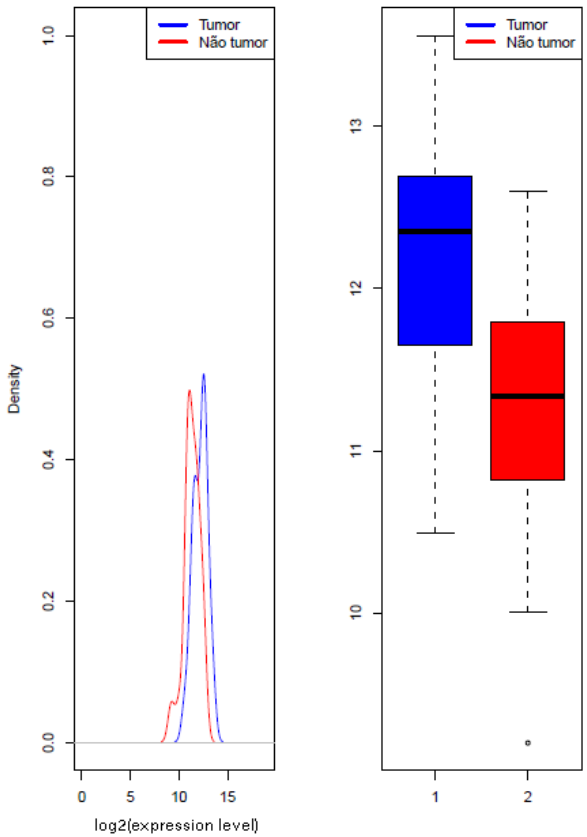
- X a random variable (r.v.) which represents the expression level of the controls and $F_X(c)$ its correspondent C.D.F.;

Differentially expressed genes

Assume:

- X a random variable (r.v.) which represents the expression level of the controls and $F_X(c)$ its correspondent C.D.F.;
- Y a r.v. which represents the expression levels of the experimental condition and $F_Y(c)$ its C.D.F..

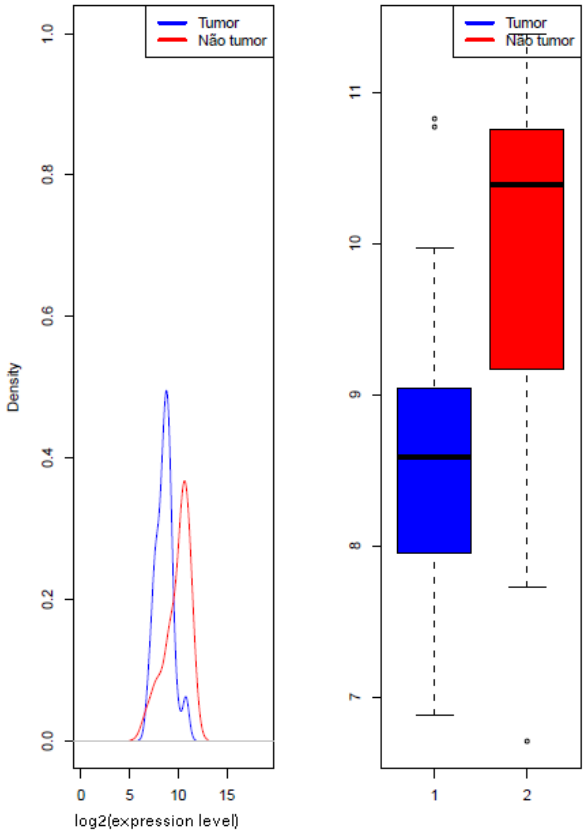
Up-regulated genes



Up-regulated gene: is a gene which has been observed to have higher expression (higher mRNA levels) in the experimental sample (Y) compared to the control one (X).

$F_X(c) > F_Y(c), \forall c.$

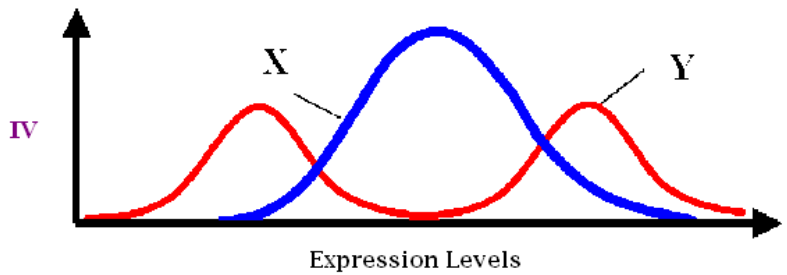
Down-regulated genes



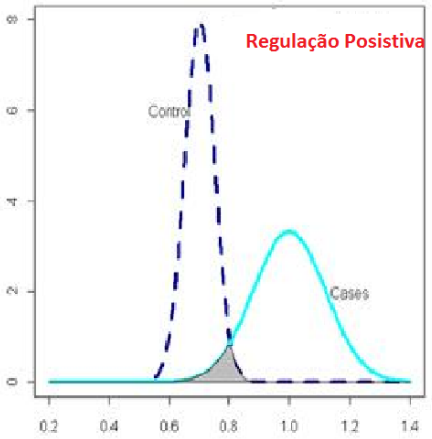
Down-regulated gene: is a gene which has been observed to have higher expression (higher mRNA levels) in the control sample (X) compared to the experimental one (Y). $F_X(c) < F_Y(c), \forall c.$

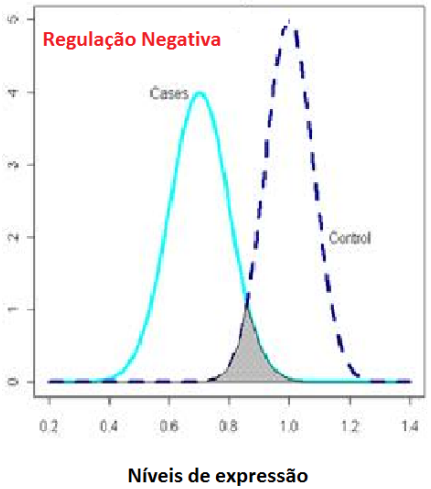
Special genes

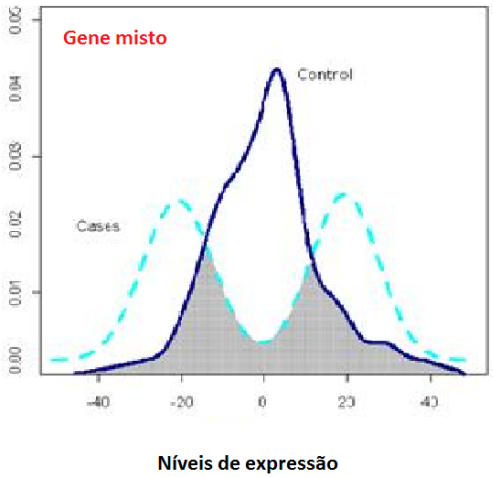
Genes with a bimodal or a multimodal distribution within a class (considering a binary study) may indicate the presence of unknown subclasses with different expression values, meaning that there are two separate peaks in the distribution; one peak due to a subclass clustered around a low expression level, and a second peak due to a subclass clustered around a higher expression level.

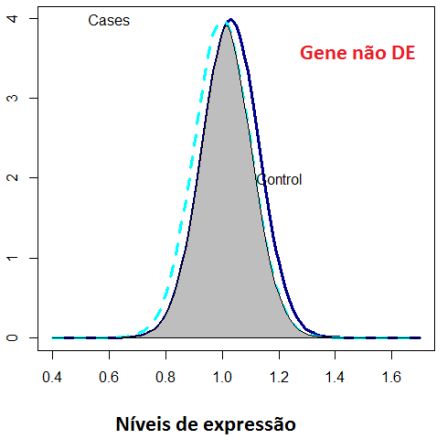


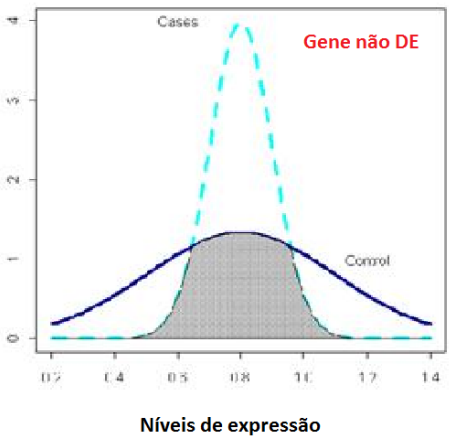
Possible scenarios



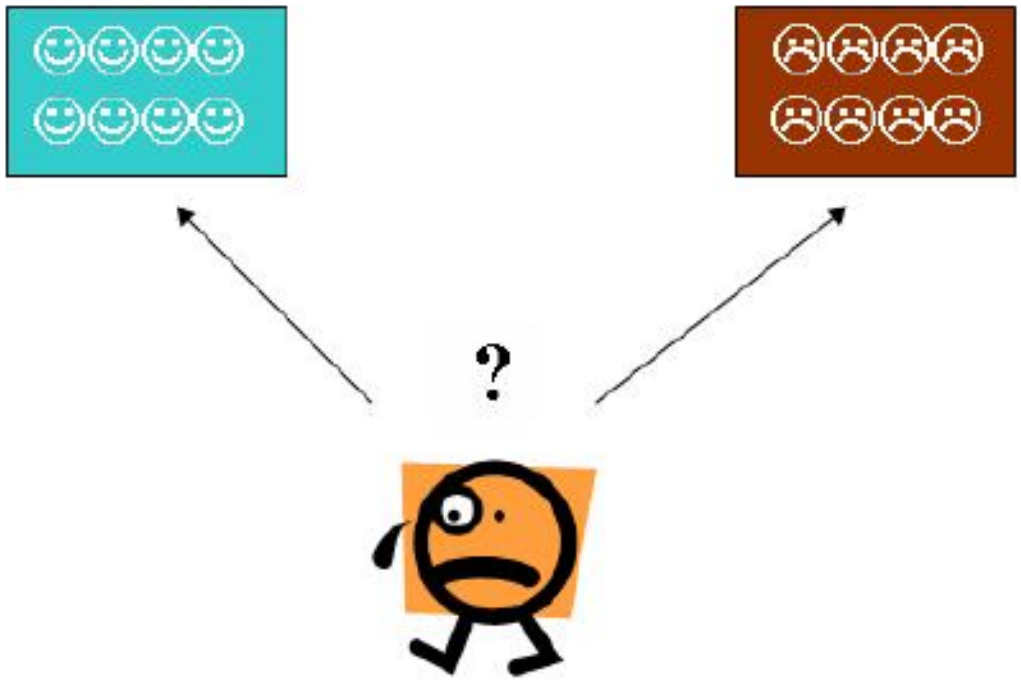




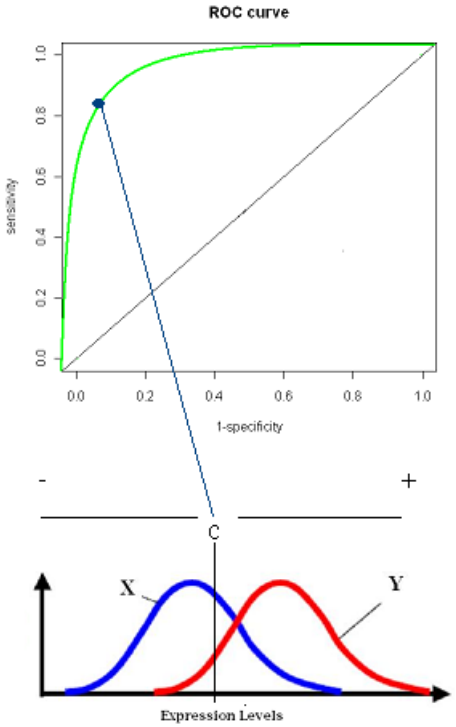




ROC curves in the selection of DE genes

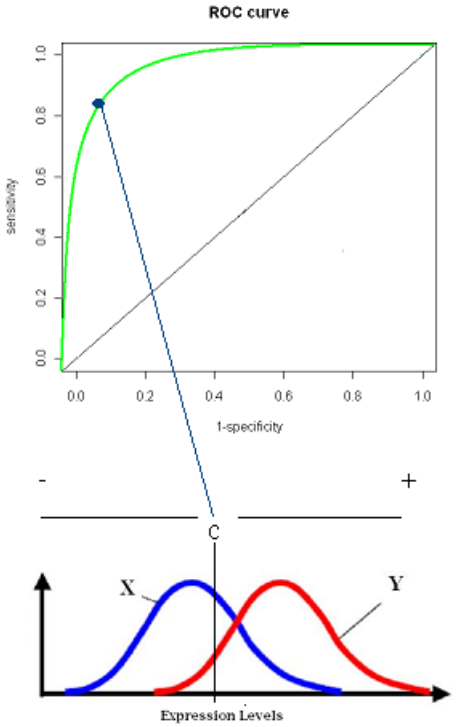


ROC curve properties



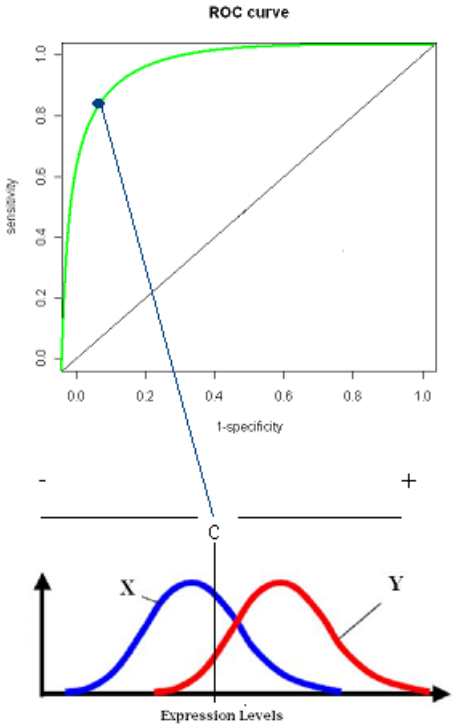
- The ROC curve results from the relationship between the proportion of true positives (sensitivity) and proportion of false positives (1-specificity) obtained for each cut-off point of the variable of decision.

ROC curve properties



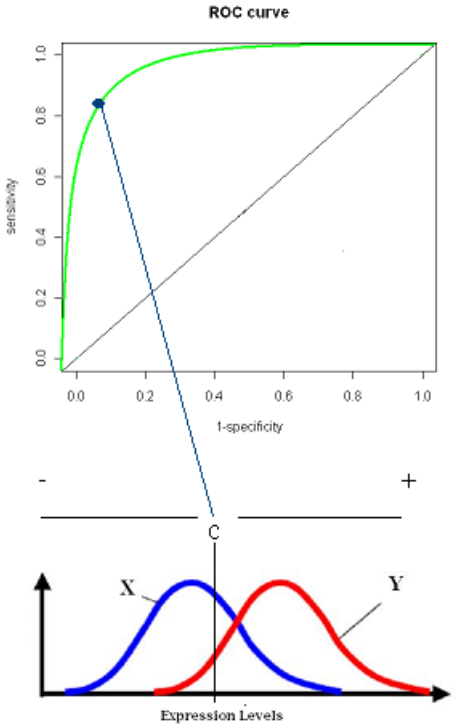
- The ROC curve results from the relationship between the proportion of true positives (sensitivity) and proportion of false positives (1-specificity) obtained for each cut-off point of the variable of decision.
- The ROC curve always starts at (0,0) and ends at (1,1).

ROC curve properties



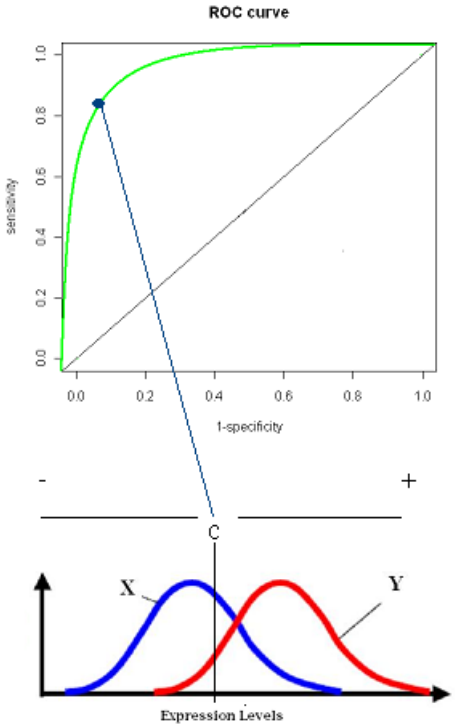
- The ROC curve results from the relationship between the proportion of true positives (sensitivity) and proportion of false positives (1-specificity) obtained for each cut-off point of the variable of decision.
- The ROC curve always starts at (0,0) and ends at (1,1).
- The ROC curve is always above the reference line.

ROC curve properties



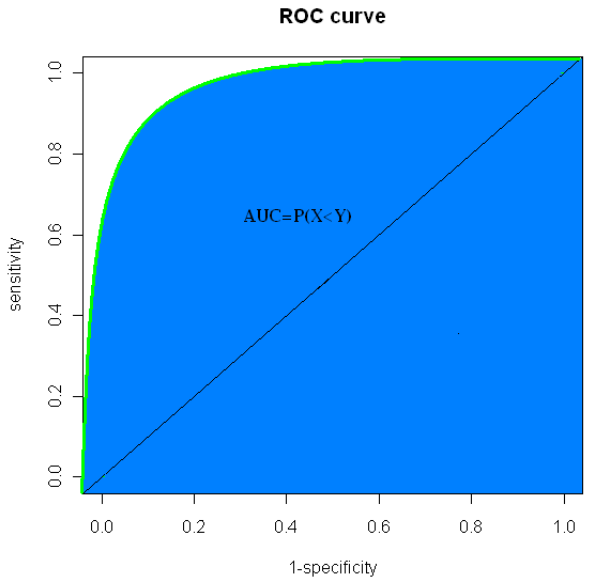
- The ROC curve results from the relationship between the proportion of true positives (sensitivity) and proportion of false positives (1-specificity) obtained for each cut-off point of the variable of decision.
- The ROC curve always starts at (0,0) and ends at (1,1).
- The ROC curve is always above the reference line.
- A ROC curve that is a diagonal line (sensitivity = 1 - specificity) corresponds to a uninformative test.

ROC curve properties



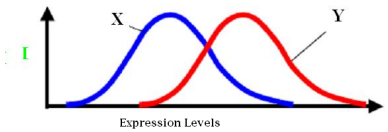
- The ROC curve results from the relationship between the proportion of true positives (sensitivity) and proportion of false positives (1-specificity) obtained for each cut-off point of the variable of decision.
- The ROC curve always starts at (0,0) and ends at (1,1).
- The ROC curve is always above the reference line.
- A ROC curve that is a diagonal line (sensitivity = 1 - specificity) corresponds to a uninformative test.
- Traditionally high values of the decision variable, correspond to the presence of the artifact of interest.

Global index of overall performance - AUC

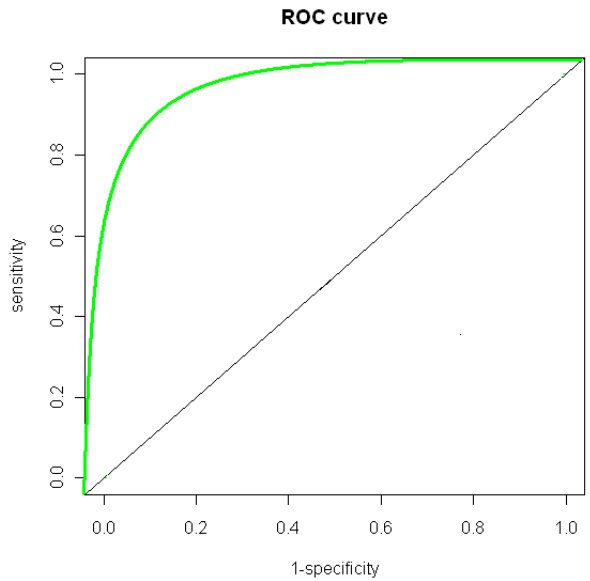


- Comparison of ROC curves often follows by comparing their areas under the curve (AUC).
- The largest possible AUC is 1, the smallest (for an informative test) is 0.5.

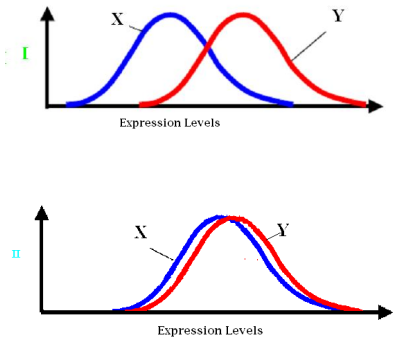
Gene expression densities and ROC curves



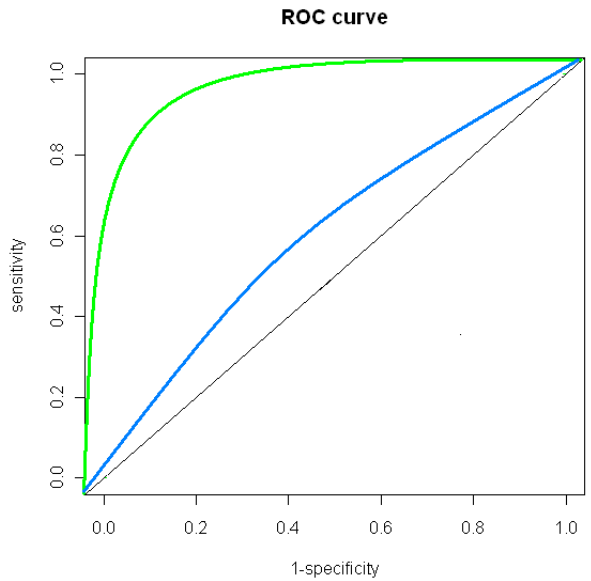
AUC=P(X<Y)



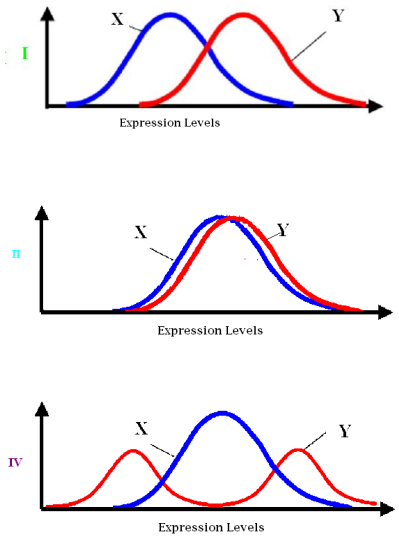
Gene expression densities and ROC curves



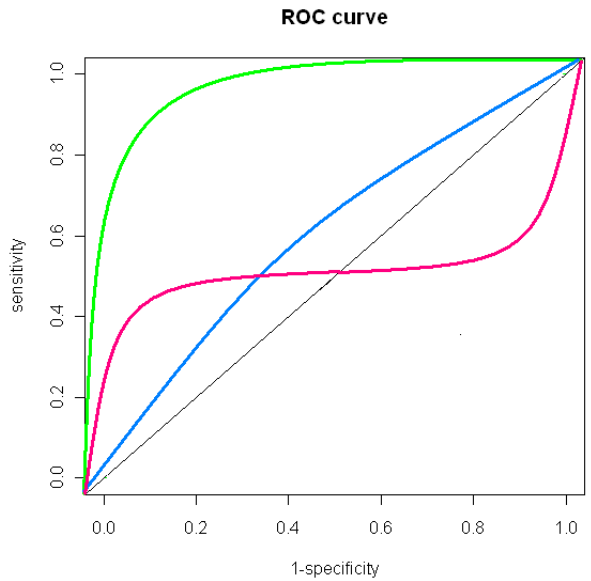
AUC=P(X<Y)



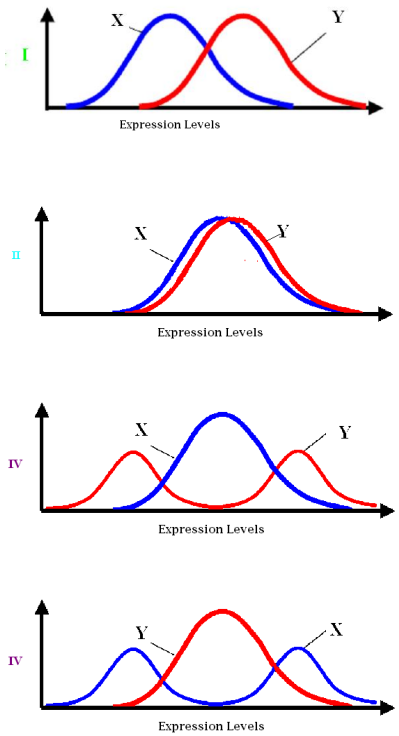
Gene expression densities and ROC curves



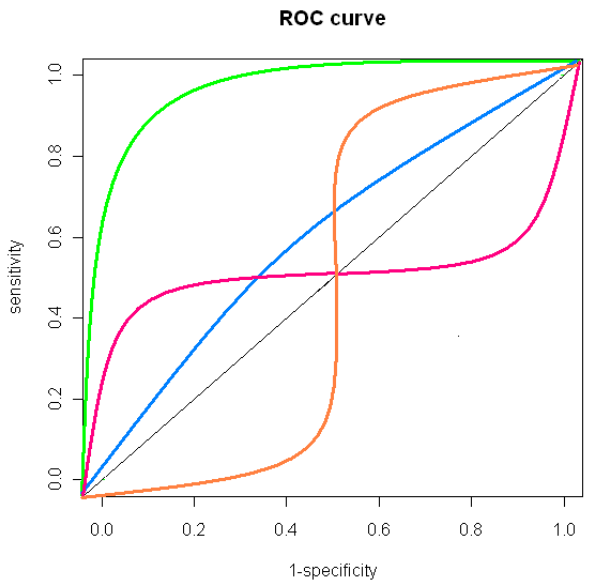
AUC=P(X<Y)



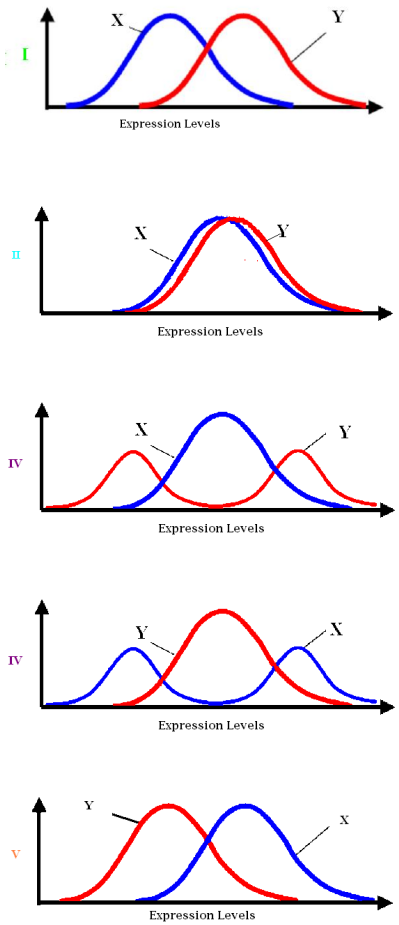
Gene expression densities and ROC curves



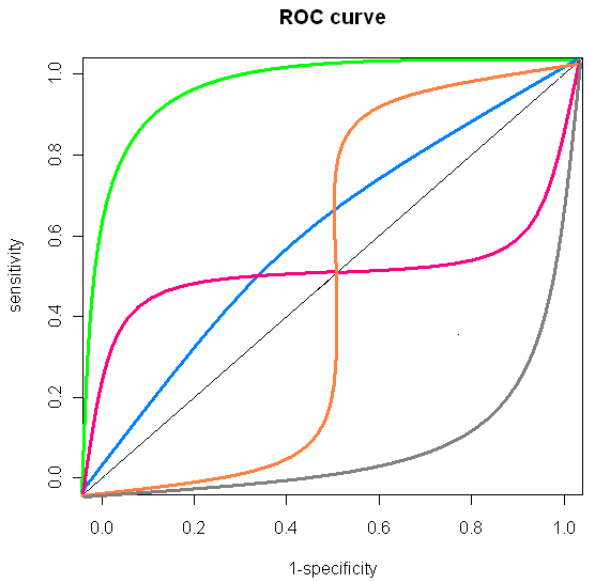
$$\mathbf{AUC = P(X < Y)}$$



Gene expression densities and ROC curves



AUC=P(X<Y)



Not proper ROC curves in the selection of DE genes

- AUC $\in [0, 1]$

Not proper ROC curves in the selection of DE genes

- AUC $\in [0, 1]$
- AUC values ≈ 1 - up-regulated genes.

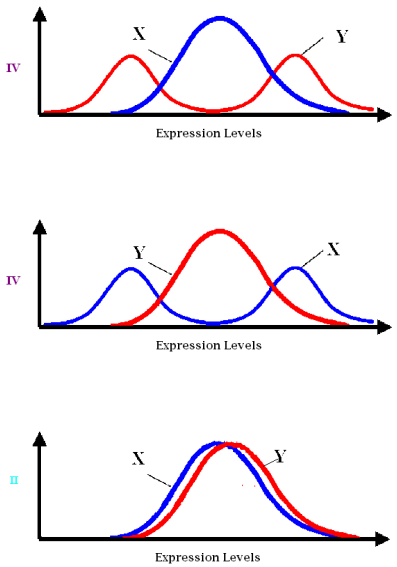
Not proper ROC curves in the selection of DE genes

- AUC $\in [0, 1]$
- AUC values ≈ 1 - up-regulated genes.
- AUC values ≈ 0 - down-regulated genes.

Not proper ROC curves in the selection of DE genes

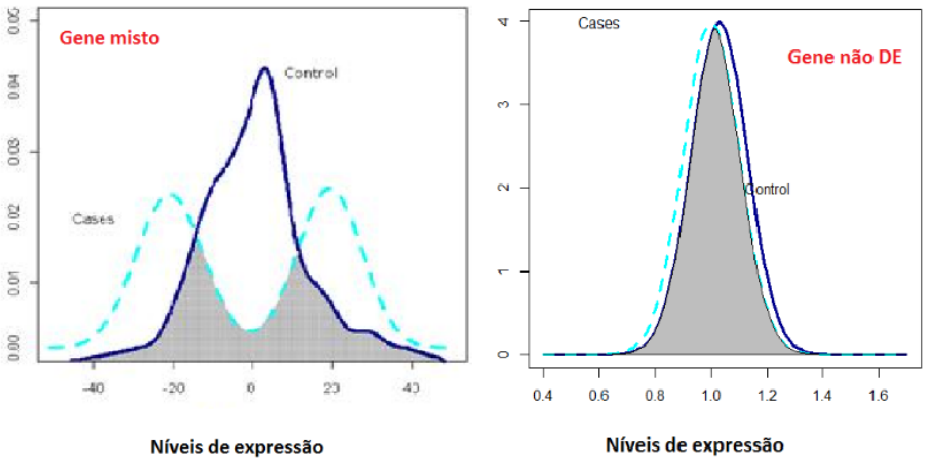
- AUC $\in [0, 1]$
- AUC values ≈ 1 - up-regulated genes.
- AUC values ≈ 0 - down-regulated genes.
- AUC values $\approx$ around 0.5 - special genes.

Not proper ROC curves in the selection of DE genes



AUC$\approx$ 0.5$\Longrightarrow$ this kind of genes will never be selected using traditional ROC analysis, neither with a method based on mean differences.

ROC curves and DE genes

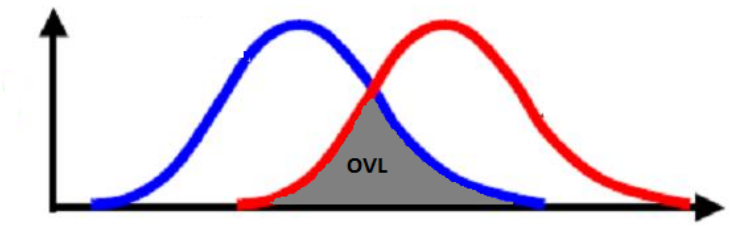


Overlapping Coefficient - OVL

OVL is the common area shared by the two densities.

$$OVL(X, Y) = \int_{-\infty}^{+\infty} min[f_X(c), g_Y(c)]dc$$

OVL $\in [0, 1]$
$OVL = 1$ if and only if $f_X(c) = f_Y(c)$
$OVL = 0$ if and only if $f_X(c) * f_Y(c) = 0$



Non-parametric OVL

It was proposed an algorithm to estimate OVL based on kernel densities of the underlying conditions.

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \forall x \in S,$$

where K is the kernel function, h bandwidth and S the support(Rosenblatt, 1956).

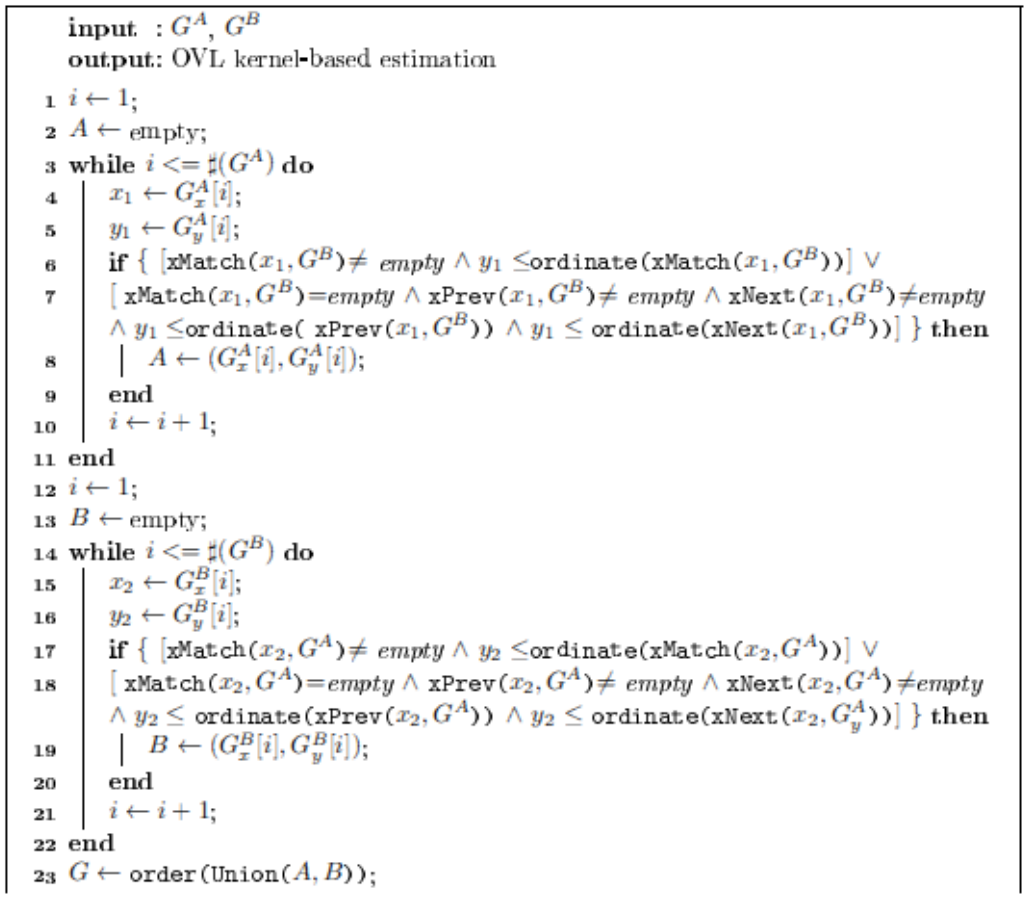
Non-parametric OVL

- In this work it is used the gaussian kernel: $(2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}u^2\right)$.
- Silverman (1986) proposed h:

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} \min\left(s, \frac{R}{1.34}\right) n^{-\frac{1}{5}},$$

where R is the interquartile range and s the empirical standard deviation.

The algorithm



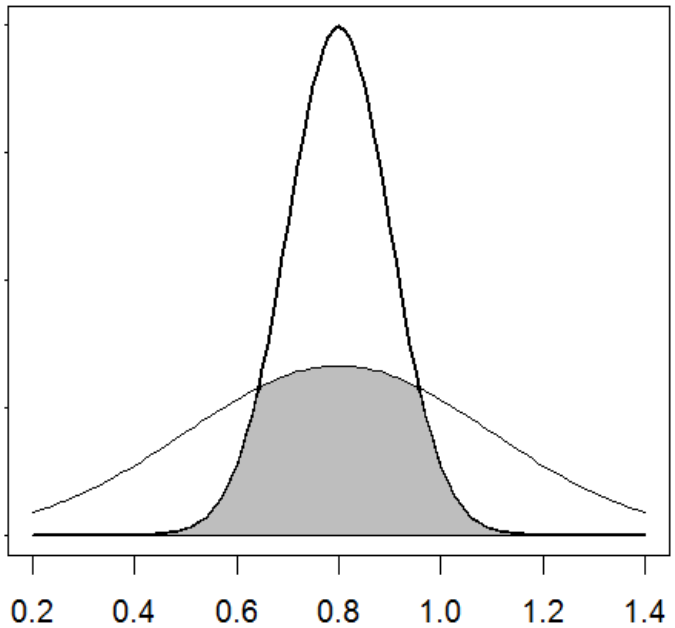
input : G^A, G^B
output: OVL kernel-based estimation

$i \leftarrow 1$;
$A \leftarrow$ empty;
while $i <= \sharp(G^A)$ **do**
  $x_1 \leftarrow G_x^A[i]$;
  $y_1 \leftarrow G_y^A[i]$;
  if { $[\texttt{xMatch}(x_1, G^B) \neq \mathit{empty} \wedge y_1 \leq \texttt{ordinate}(\texttt{xMatch}(x_1, G^B))] \vee$
  $[\texttt{xMatch}(x_1, G^B) = \mathit{empty} \wedge \texttt{xPrev}(x_1, G^B) \neq \mathit{empty} \wedge \texttt{xNext}(x_1, G^B) \neq \mathit{empty}$
  $\wedge\, y_1 \leq \texttt{ordinate}(\texttt{xPrev}(x_1, G^B)) \wedge y_1 \leq \texttt{ordinate}(\texttt{xNext}(x_1, G^B))]$ } **then**
    $A \leftarrow (G_x^A[i], G_y^A[i])$;
  end
  $i \leftarrow i + 1$;
end
$i \leftarrow 1$;
$B \leftarrow$ empty;
while $i <= \sharp(G^B)$ **do**
  $x_2 \leftarrow G_x^B[i]$;
  $y_2 \leftarrow G_y^B[i]$;
  if { $[\texttt{xMatch}(x_2, G^A) \neq \mathit{empty} \wedge y_2 \leq \texttt{ordinate}(\texttt{xMatch}(x_2, G^A))] \vee$
  $[\texttt{xMatch}(x_2, G^A) = \mathit{empty} \wedge \texttt{xPrev}(x_2, G^A) \neq \mathit{empty} \wedge \texttt{xNext}(x_2, G^A) \neq \mathit{empty}$
  $\wedge\, y_2 \leq \texttt{ordinate}(\texttt{xPrev}(x_2, G^A)) \wedge y_2 \leq \texttt{ordinate}(\texttt{xNext}(x_2, G_y^A))]$ } **then**
    $B \leftarrow (G_x^B[i], G_y^B[i])$;
  end
  $i \leftarrow i + 1$;
end
$G \leftarrow \texttt{order}(\texttt{Union}(A, B))$;

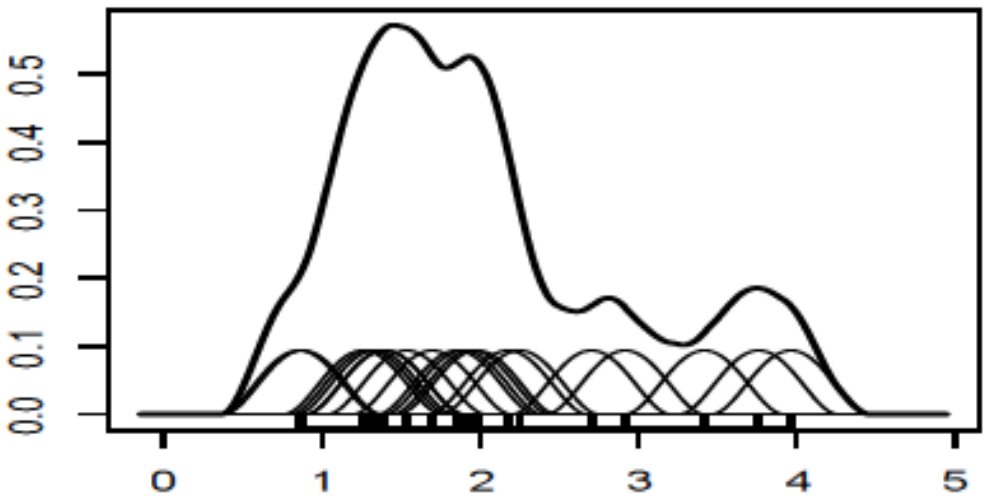
Figure: Pseudocode. Source: Silva-Fortes *et al.* (2012), *BMC Bioinformatics*

The algorithm in a picture

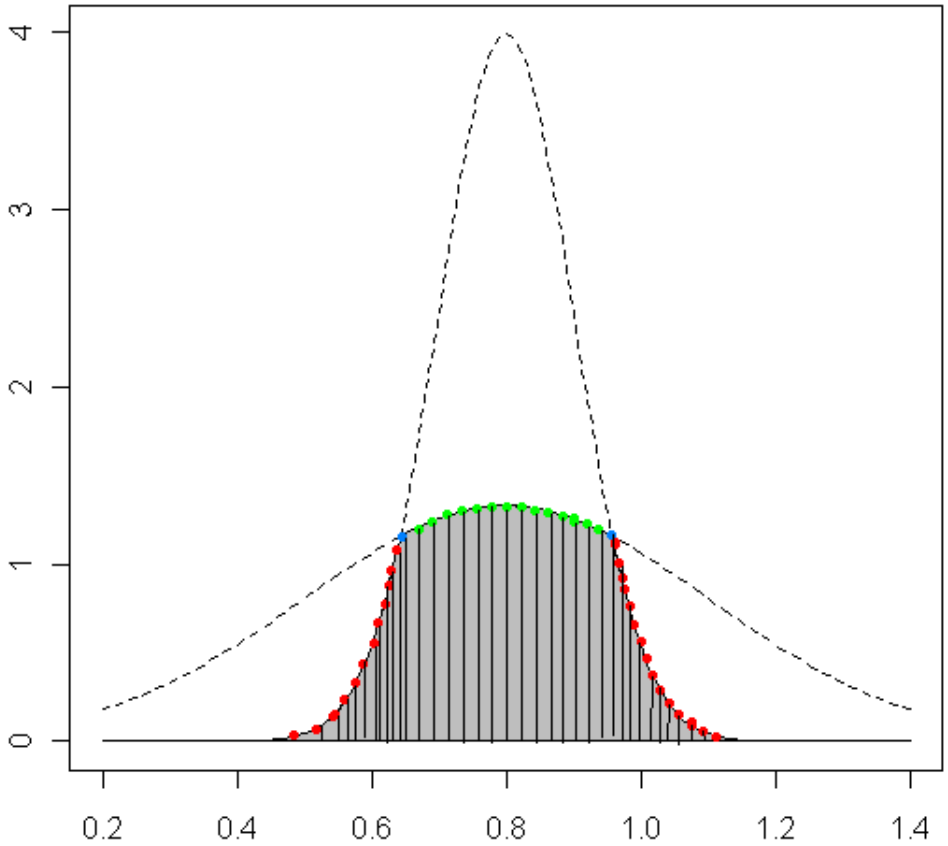
In R: `>plot(density(x)) >lines(density(y))`



The algorithm in a picture



The algorithm in a picture

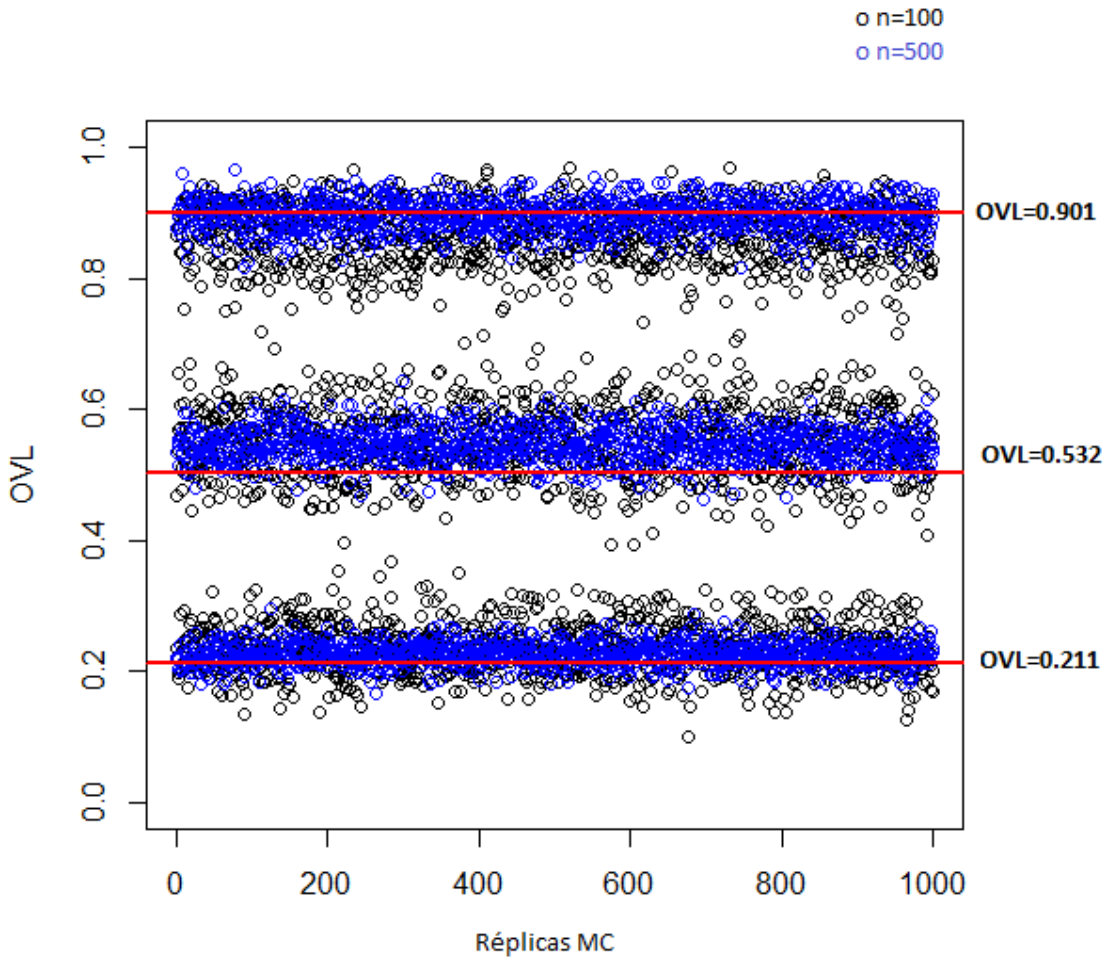


OVL algorithm - a MC simulation study

Table: Estimatives of the MC mean, MC standard error and relative bias of the OVL estimated by the proposed algorithm.

	$n_1 = n_2 = 100$			$n_1 = n_2 = 500$		
OVL	MC mean	MC Standard Error	Relative Bias	MC Mean	MC Standard Error	Relative Bias
$X_1 \sim N(20,4)$ $X_2 \sim N(10,4)$ OVL=0.2112	0.2342	0.0012	0.1088	0.2257	5.58E-04	0.0687
$X_1 \sim N(20,4)$ $X_2 \sim N(15,4)$ OVL= 0.532	0.5512	0.0017	0.0362	0.5449	8.3E-04	0.0244
$X_1 \sim N(20,4)$ $X_2 \sim N(19,4)$ OVL=0.901	0.8687	0.0014	−0.0359	0.8953	8.1E-04	−0.0063

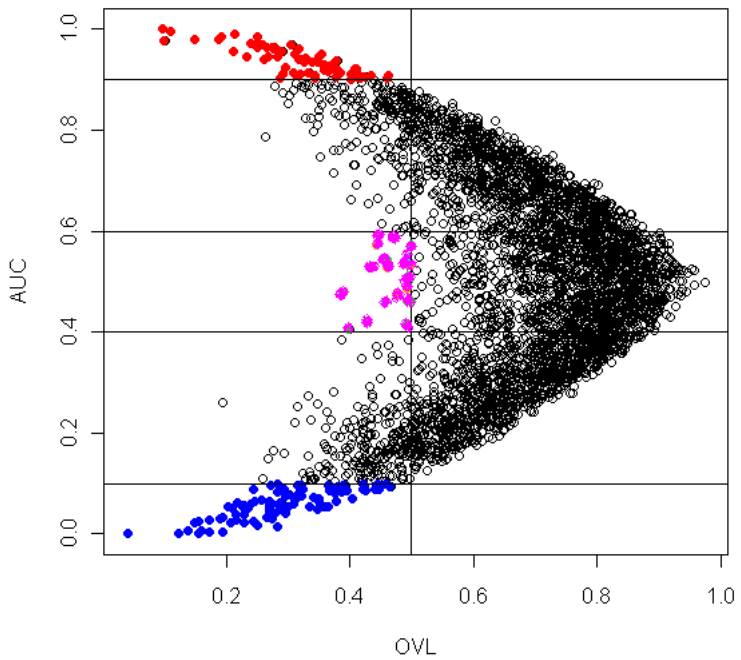
OVL algorithm - a MC simulation study



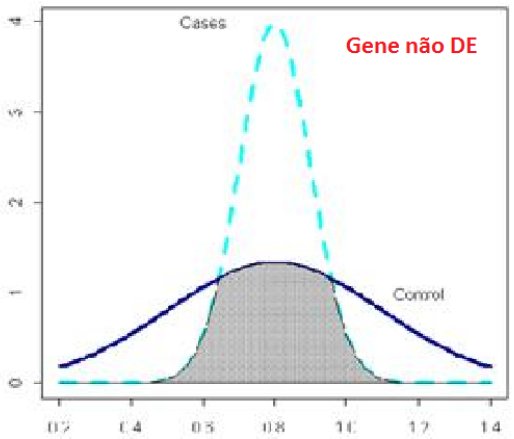
OVL algorithm - a MC simulation study

- There are no restrictions concerning the number of intersections of the pdf.
- It admits any support of the variables.
- Simulation study revealed that the algorithm gives OVL estimates with lower bias.
- Implementation time concerning a data base with 10000 genes was near by 60 minutes.

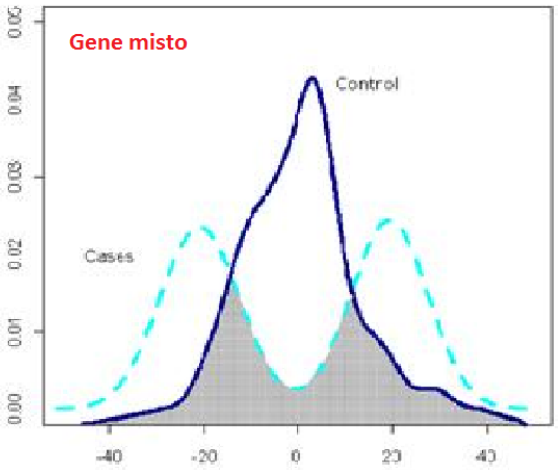
Arrow plot - AUC and OVL



OVL not sufficient!!



Níveis de expressão



Arrow plot

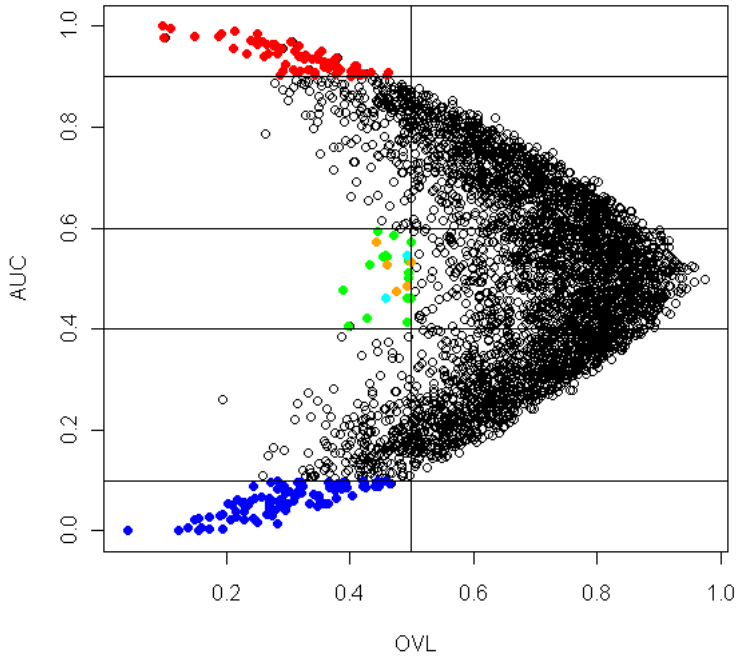


Figure: Fonte: Silva-Fortes, C., Amaral Turkman, M. A. e Sousa, L. (2012). Arrow Plot: a new graphical tool for selecting up and down-regulated genes and genes differential expressed on subsamples. *BMC Bioinformatics*, 13:147.

Concluding remarks

- Nonparametric approach.

Concluding remarks

- Nonparametric approach.
- No restrictions on the OVL estimation, considering the number of intersections of the densities and on x-axis coordinates for both classes.

Concluding remarks

- Nonparametric approach.
- No restrictions on the OVL estimation, considering the number of intersections of the densities and on x-axis coordinates for both classes.
- Based on measuring the distance between two distributions.

Concluding remarks

- Nonparametric approach.
- No restrictions on the OVL estimation, considering the number of intersections of the densities and on x-axis coordinates for both classes.
- Based on measuring the distance between two distributions.
- "Arrow plot" is an exploratory graphical method for microarray experiments to identify genes with different expression levels between two types of samples (up- and down-regulated) and also to identify genes with a special behavior that could lead to find subclasses that may provide useful insights about biological mechanisms underlying physiologic or pathologic conditions.

Concluding remarks

- Nonparametric approach.
- No restrictions on the OVL estimation, considering the number of intersections of the densities and on x-axis coordinates for both classes.
- Based on measuring the distance between two distributions.
- "Arrow plot" is an exploratory graphical method for microarray experiments to identify genes with different expression levels between two types of samples (up- and down-regulated) and also to identify genes with a special behavior that could lead to find subclasses that may provide useful insights about biological mechanisms underlying physiologic or pathologic conditions.
- For these reasons, our results indicate that "Arrow plot" represents a new flexible and useful tool for the analysis of gene expression profiles from microarray experiments.

Next steps

- R package.
- Extension of Arrow plot in the Next Generation Sequencing (NGS) experiments.

Thank you