



# Improving type II error rates of multiple testing procedures by use of auxiliary variables

## Application to microarray data

Maela Kloareg    David Causeur

*Laboratoire de Mathématiques Appliquées  
Agrocampus Rennes  
IRMAR CNRS UMR 6625*

ASMDA June 2007

- 1 Introduction
- 2 The gene-by-gene test statistics : using an auxiliary variable
- 3 Impact on the Benjamini-Hochberg procedure results
- 4 Future work



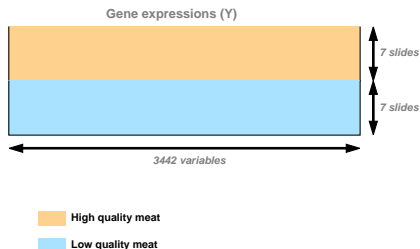
## Biological context

- Aim : find genes involved in pork quality



## Biological context

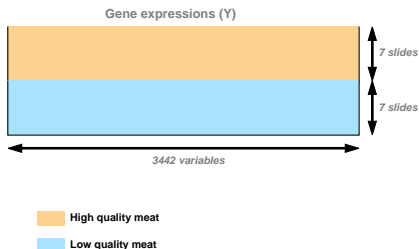
- Aim : find genes involved in pork quality
- Microarray data
  - 14 pigs : 7 high quality, 7 low quality
  - 3442 genes (radioactivity : 1 membrane per pig):





## Biological context

- Aim : find genes involved in pork quality
- Microarray data
  - 14 pigs : 7 high quality, 7 low quality
  - 3442 genes (radioactivity : 1 membrane per pig):



- Main concern: dealing with high dimensionality



## Multiple tests (Benjamini and Hochberg, 1995)

- Gene-by-gene tests : Student t-test (for instance)



## Multiple tests (Benjamini and Hochberg, 1995)

- Gene-by-gene tests : Student t-test (for instance)
- Cut-off  $t$  on p-values

	Declared non DE	Declared DE	Total
non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$



## Multiple tests (Benjamini and Hochberg, 1995)

- Gene-by-gene tests : Student t-test (for instance)
- Cut-off  $t$  on p-values

	Declared non DE	Declared DE	Total
non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

$$FDR_t = \mathbb{E} \left[ \frac{V_t}{R_t} | R_t > 0 \right] \mathbb{P}(R_t > 0)$$



## Multiple tests (Benjamini and Hochberg, 1995)

- Gene-by-gene tests : Student t-test (for instance)
- Cut-off  $t$  on p-values

	Declared non DE	Declared DE	Total
non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

$$FDR_t = \mathbb{E} \left[ \frac{V_t}{R_t} | R_t > 0 \right] \mathbb{P}(R_t > 0)$$

Which cut-off ensures  $FDR_t \leq \alpha$  ?



## Multiple tests (Benjamini and Hochberg, 1995)

- Gene-by-gene tests : Student t-test (for instance)
- Cut-off  $t$  on p-values

	Declared non DE	Declared DE	Total
non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

$$FDR_t = \mathbb{E} \left[ \frac{V_t}{R_t} \mid R_t > 0 \right] \mathbb{P}(R_t > 0)$$

Which cut-off ensures  $FDR_t \leq \alpha$  ?

$$\text{BH} : t = p_{k^*} : k^* = \arg \max_k \left\{ k, m_0 \frac{p_k}{k} \leq \alpha \right\}$$



## Multiple tests (Benjamini and Hochberg, 1995)

- Gene-by-gene tests : Student t-test (for instance)
- Cut-off  $t$  on p-values

	Declared non DE	Declared DE	Total
non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

$$FDR_t = \mathbb{E} \left[ \frac{V_t}{R_t} \mid R_t > 0 \right] \mathbb{P}(R_t > 0)$$

Which cut-off ensures  $FDR_t \leq \alpha$  ?

$$\text{BH} : t = p_{k^*} : k^* = \arg \max_k \{k, m_0 \frac{p_k}{k} \leq \alpha\}$$

- Number of declared DE (positiv genes) : 10



## Multiple tests (Benjamini and Hochberg, 1995)

- Gene-by-gene tests : Student t-test (for instance)
- Cut-off  $t$  on p-values

	Declared non DE	Declared DE	Total
non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

$$FDR_t = \mathbb{E} \left[ \frac{V_t}{R_t} | R_t > 0 \right] \mathbb{P}(R_t > 0)$$

Which cut-off ensures  $FDR_t \leq \alpha$  ?

$$\text{BH} : t = p_{k^*} : k^* = \arg \max_k \{k, m_0 \frac{p_k}{k} \leq \alpha\}$$

- Number of declared DE (positiv genes) : 10

... and what about the type II error rate?

$$FNR = \mathbb{E} \left[ \frac{T_t}{W_t} | W_t > 0 \right] \mathbb{P}(W_t > 0)$$



## Multiple tests (Benjamini and Hochberg, 1995)

- Gene-by-gene tests : Student t-test (for instance)
- Cut-off  $t$  on p-values

	Declared non DE	Declared DE	Total
non DE	$U_t$	$V_t$	$m_0$
DE	$T_t$	$S_t$	$m_1$
Total	$W_t$	$R_t$	$m$

$$FDR_t = \mathbb{E} \left[ \frac{V_t}{R_t} \mid R_t > 0 \right] \mathbb{P}(R_t > 0)$$

Which cut-off ensures  $FDR_t \leq \alpha$  ?

$$\text{BH} : t = p_{k^*} : k^* = \arg \max_k \{k, m_0 \frac{p_k}{k} \leq \alpha\}$$

- Number of declared DE (positiv genes) : 10

... and what about the type II error rate?

$$NDR = \mathbb{E} \left[ \frac{T_t}{m_1} \right]$$



# Taking into account the shared information

The approximation of the joint distribution can improve the power



## Taking into account the shared information

The approximation of the joint distribution can improve the power

- Control of the FDR under dependency (Benjamini and Yekutieli, 2001)



## Taking into account the shared information

The approximation of the joint distribution can improve the power

- Control of the FDR under dependency (Benjamini and Yekutieli, 2001)
- Bootstrap or permutation methods (Dudoit *et al.*, 2003)



## Taking into account the shared information

The approximation of the joint distribution can improve the power

- Control of the FDR under dependency (Benjamini and Yekutieli, 2001)
- Bootstrap or permutation methods (Dudoit *et al.*, 2003)
- "Moderated" t-statistics (Lönstedt and Speed, 2002)



## Taking into account the shared information

The approximation of the joint distribution can improve the power

- Control of the FDR under dependency (Benjamini and Yekutieli, 2001)
- Bootstrap or permutation methods (Dudoit *et al.*, 2003)
- "Moderated" t-statistics (Lönstedt and Speed, 2002)
- Empirical Bayes approaches (Kendzioriski *et al.*, 2003 ; Efron, 2006)



# Integrating external information

Improving the power by using prior information such as Gene Ontology or Pathways



## Integrating external information

### Improving the power by using prior information such as Gene Ontology or Pathways

- Filtering of genes before statistical analysis (von Heydebreck *et al.*, 2004)



## Integrating external information

### Improving the power by using prior information such as Gene Ontology or Pathways

- Filtering of genes before statistical analysis (von Heydebreck *et al.*, 2004)
- Analysing data in terms of gene sets (Goeman and Bühlmann, 2007 ; Efron and Tibshirani, 2007)



## Integrating external information

### Improving the power by using prior information such as Gene Ontology or Pathways

- Filtering of genes before statistical analysis (von Heydebreck *et al.*, 2004)
- Analysing data in terms of gene sets (Goeman and Bühlmann, 2007 ; Efron and Tibshirani, 2007)
- Weighting groups of genes (Roeder *et al.*, 2007)



## Integrating external information

### Improving the power by using prior information such as Gene Ontology or Pathways

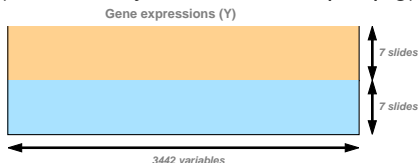
- Filtering of genes before statistical analysis (von Heydebreck *et al.*, 2004)
- Analysing data in terms of gene sets (Goeman and Bühlmann, 2007 ; Efron and Tibshirani, 2007)
- Weighting groups of genes (Roeder *et al.*, 2007)

Improving the power by using covariate adjustment (Tsiatis *et al.*, 2006)



## Additional information in our biological context

- Aim : find genes involved in pork quality
- Microarray data
  - 14 pigs : 7 high quality, 7 low quality
  - 3442 genes (radioactivity : 1 membrane per pig):

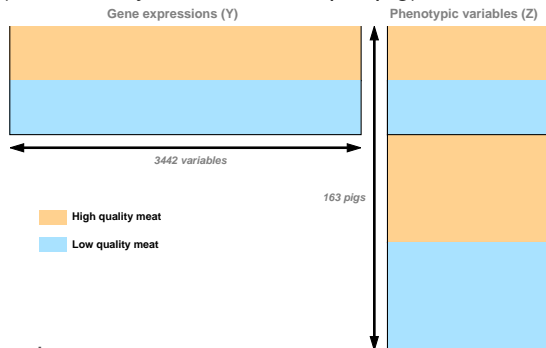


- High quality meat
- Low quality meat



## Additional information in our biological context

- Aim : find genes involved in pork quality
- Microarray data
  - 14 pigs : 7 high quality, 7 low quality
  - 3442 genes (radioactivity : 1 membrane per pig):



- Additional information
  - physiological measurements : pH of the meat
  - for 163 pigs (87 high quality, 76 low quality)



## Double-sampling in other areas of statistics

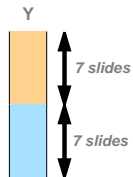
- Origins : Sample survey, estimation of a mean (Cochran, 1963)
- Parametric regression issues : Econometrics (Conniffe, 1985)
- Nonparametric regression issues (Breslow *et al.*, 2003)
- Multivariate monotone designs (Causeur, 2005)
- Testing issues (Causeur and Husson, 2007)



- 1 Introduction
- 2 The gene-by-gene test statistics : using an auxiliary variable**
- 3 Impact on the Benjamini-Hochberg procedure results
- 4 Future work



## The gene-by-gene test statistics



High quality meat

Low quality meat

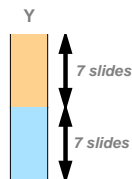


## The gene-by-gene test statistics

Model

$$\mathbb{E} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \mu_1^{(y)} \\ \mu_2^{(y)} \end{bmatrix}$$

$$\text{Var} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \sigma_y^2 I_n$$



High quality meat

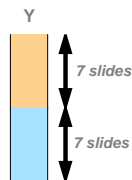
Low quality meat

# The gene-by-gene test statistics

Model

$$\mathbb{E} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \mu_1^{(y)} \\ \mu_2^{(y)} \end{bmatrix}$$

$$\text{Var} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \sigma_y^2 I_n$$



High quality meat

Low quality meat

Student test statistic

$$T = \frac{\Delta \bar{Y}}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

# The gene-by-gene test statistics

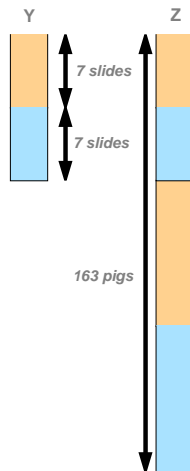
Model

$$\mathbb{E} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \mu_1^{(y)} \\ \mu_2^{(y)} \end{bmatrix}$$

$$\text{Var} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \sigma_y^2 I_n$$

Student test statistic

$$T = \frac{\Delta \bar{Y}}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

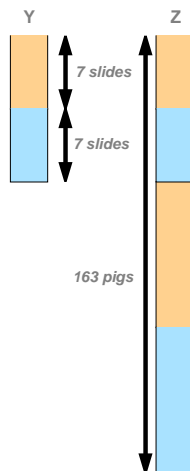


# The gene-by-gene test statistics

Model

$$\mathbb{E} \begin{bmatrix} Y_1 \\ Y_2 \\ Z \end{bmatrix} = \begin{bmatrix} \mu_1^{(y)} \\ \mu_2^{(y)} \\ \mu^{(z)} \end{bmatrix}$$

$$\text{Var} \begin{bmatrix} Y_1 \\ Y_2 \\ Z \end{bmatrix} = \begin{bmatrix} \sigma_y^2 I_n & \sigma_{yz} [I_n \ 0] \\ \sigma_{yz} \begin{bmatrix} I_n \\ 0 \end{bmatrix} & \sigma_z^2 I_N \end{bmatrix}$$

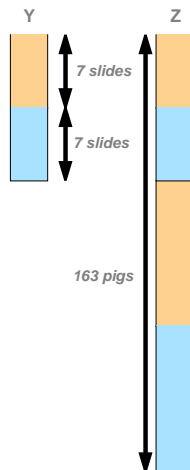


# The gene-by-gene test statistics

Model

$$\mathbb{E} \begin{bmatrix} Y_1 \\ Y_2 \\ Z \end{bmatrix} = \begin{bmatrix} \mu_1^{(y)} \\ \mu_2^{(y)} \\ \mu^{(z)} \end{bmatrix}$$

$$\text{Var} \begin{bmatrix} Y_1 \\ Y_2 \\ Z \end{bmatrix} = \begin{bmatrix} \sigma_y^2 I_n & \sigma_{yz} [I_n \ 0] \\ \sigma_{yz} \begin{bmatrix} I_n \\ 0 \end{bmatrix} & \sigma_z^2 I_N \end{bmatrix}$$



Moderated test statistic

$$T(\Sigma) = \frac{\Delta \bar{Y} + \rho \frac{\sigma_y}{\sigma_z} [\Delta \bar{Z} - \Delta \bar{z}]}{\sigma_y \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{1 + \rho^2 \left[ \frac{f_1 f_2}{f} - 1 \right]}}$$



## Comparative study of the power functions

Power functions bounds

Under  $H_1 : \mu_1 - \mu_2 = \delta\sigma_y$ ,  $T(\Sigma) \sim \mathcal{N}[\delta_{n,N}(\rho); \mathbf{1}]$  with

$$\delta_{n,N}^{-2}(\rho) = (1 - \rho^2)\delta_{n,N}^{-2}(\rho = 0) + \rho^2\delta_{n,N}^{-2}(\rho = 1).$$



## Comparative study of the power functions

Power functions bounds

Under  $H_1 : \mu_1 - \mu_2 = \delta\sigma_y$ ,  $T(\Sigma) \sim \mathcal{N}[\delta_{n,N}(\rho); 1]$  with

$$\delta_{n,N}^{-2}(\rho) = (1 - \rho^2)\delta_{n,N}^{-2}(\rho = 0) + \rho^2\delta_{n,N}^{-2}(\rho = 1).$$



## Comparative study of the power functions

Power functions bounds

Under  $H_1 : \mu_1 - \mu_2 = \delta\sigma_y$ ,  $T(\Sigma) \sim \mathcal{N}[\delta_{n,N}(\rho); \mathbf{1}]$  with

$$\delta_{n,N}^{-2}(\rho) = (1 - \rho^2) \delta_n^{-2} + \rho^2 \delta_N^{-2}$$



## Comparative study of the power functions

Power functions bounds

Under  $H_1 : \mu_1 - \mu_2 = \delta\sigma_y$ ,  $T(\Sigma) \sim \mathcal{N}[\delta_{n,N}(\rho); 1]$  with

$$\delta_{n,N}^{-2}(\rho) = (1 - \rho^2) \delta_n^{-2} + \rho^2 \delta_N^{-2}$$

Power<sub>n</sub> ≤ Power<sub>n,N</sub>(ρ) ≤ Power<sub>N</sub>

Equality holds if  $\rho = 0$

Equality holds if  $\rho = 1$



## Small sample distribution when $\Sigma$ is estimated

ML estimators of the variance parameters (Causeur, 2005)

Under  $H_1 : \mu_1 - \mu_2 = \delta\sigma_y$ ,

$$T(\hat{\Sigma}) \approx \frac{\sqrt{n_1 + n_2 - 2} \sqrt{1 + \frac{\rho^2}{1 - \rho^2} \frac{f_1 f_2}{f}} T_1}{\sqrt{T_2 + \frac{n_1 + n_2 - 2}{N_1 + N_2 - 2} \frac{f_1 f_2}{f} T_3}}$$



## Small sample distribution when $\Sigma$ is estimated

ML estimators of the variance parameters (Causeur, 2005)

Under  $H_1 : \mu_1 - \mu_2 = \delta \sigma_y$ ,

$$T(\hat{\Sigma}) \approx \underbrace{\sqrt{n_1 + n_2 - 2} \sqrt{1 + \frac{\rho^2}{1 - \rho^2} \frac{f_1 f_2}{f}}}_{T(\rho^2)} \frac{T_1}{\sqrt{T_2 + \frac{n_1 + n_2 - 2}{N_1 + N_2 - 2} \frac{f_1 f_2}{f} T_3}}$$



## Small sample distribution when $\Sigma$ is estimated

ML estimators of the variance parameters (Causeur, 2005)

Under  $H_1 : \mu_1 - \mu_2 = \delta\sigma_y$ ,

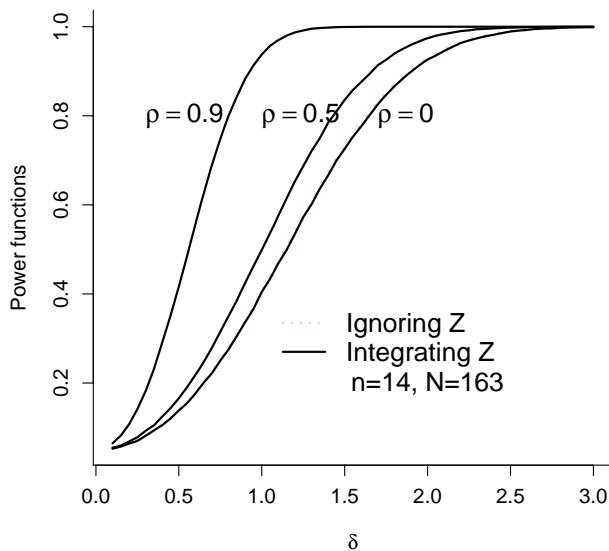
$$T(\hat{\Sigma}) \approx \underbrace{\sqrt{n_1 + n_2 - 2} \sqrt{1 + \frac{\rho^2}{1 - \rho^2} \frac{f_1 f_2}{f}}}_{T(\rho^2)} \frac{T_1}{\sqrt{T_2 + \frac{n_1 + n_2 - 2}{N_1 + N_2 - 2} \frac{f_1 f_2}{f} T_3}}$$

If  $\rho$  is not too small,

$$\text{Power}_n \leq \text{Power}_{n,N}(\rho) \leq \text{Power}_N$$



## Small sample distribution when $\Sigma$ is estimated



- 1 Introduction
- 2 The gene-by-gene test statistics : using an auxiliary variable
- 3 Impact on the Benjamini-Hochberg procedure results**
- 4 Future work



## Distribution of the type II error rate

### Simulations study

- 100 variables  $Y$ ,  $n_1 = n_2 = 10$



## Distribution of the type II error rate

### Simulations study

- 100 variables  $Y$ ,  $n_1 = n_2 = 10$
- 50 variables under  $H_1$  :  $\Delta\mu = 1.25\sigma_y$



## Distribution of the type II error rate

### Simulations study

- 100 variables  $Y$ ,  $n_1 = n_2 = 10$
- 50 variables under  $H_1$  :  $\Delta\mu = 1.25\sigma_y$
- Auxiliary variable  $Z$  with  $N_1 = N_2 = 75$



## Distribution of the type II error rate

### Simulations study

- 100 variables  $Y$ ,  $n_1 = n_2 = 10$
- 50 variables under  $H_1$  :  $\Delta\mu = 1.25\sigma_y$
- Auxiliary variable  $Z$  with  $N_1 = N_2 = 75$
- $Z$  under  $H_1$  :  $\Delta\mu = 2\sigma_z$



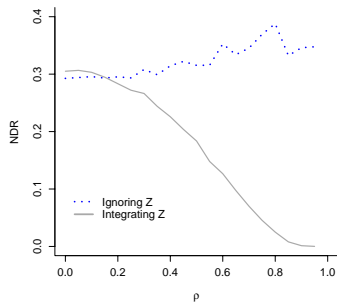
## Distribution of the type II error rate

### Simulations study

- 100 variables  $Y$ ,  $n_1 = n_2 = 10$
- 50 variables under  $H_1$  :  $\Delta\mu = 1.25\sigma_y$
- Auxiliary variable  $Z$  with  $N_1 = N_2 = 75$
- $Z$  under  $H_1$  :  $\Delta\mu = 2\sigma_z$
- Intra-group correlation between  $Y$  and  $Z$ :  $\rho$

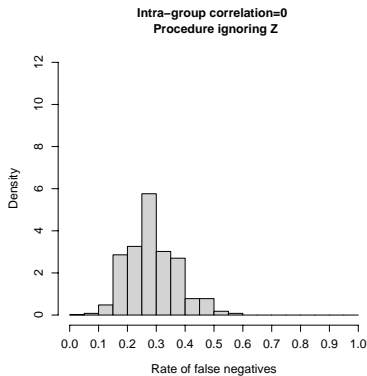
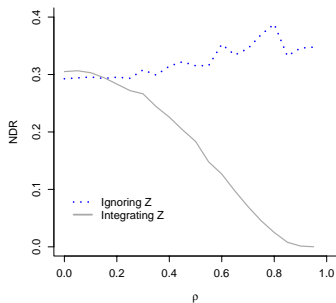


## Distribution of the type II error rate



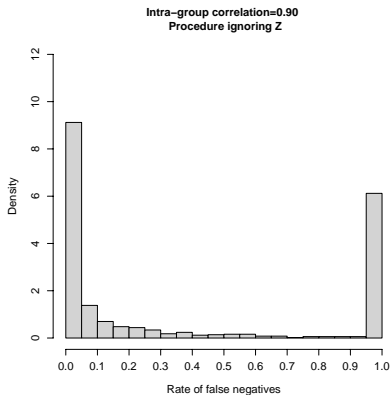
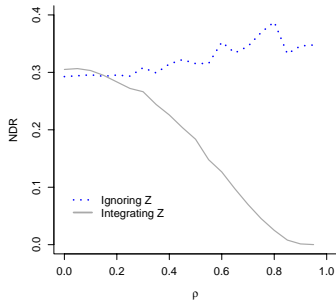


## Distribution of the type II error rate



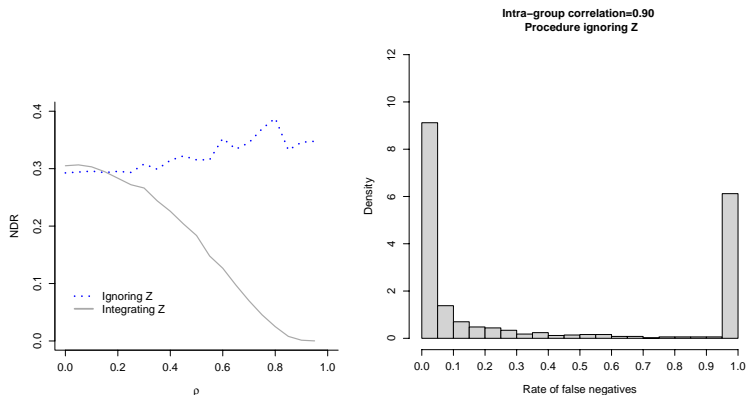


## Distribution of the type II error rate





## Distribution of the type II error rate

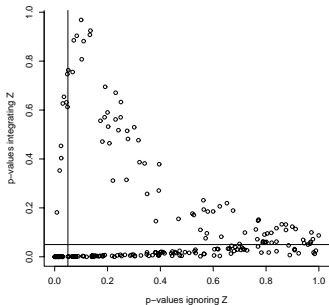


- Owen, 2005 : Variance of the number of false discovery that takes account of the correlations between test statistics



## Back to the biological data

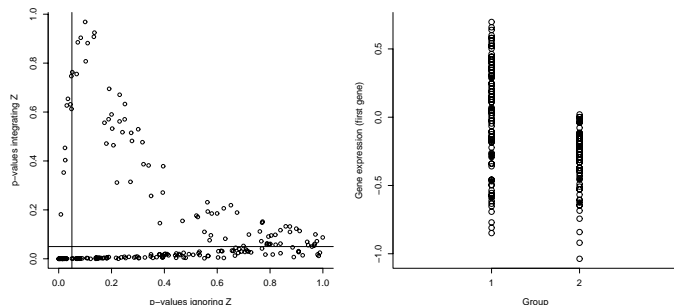
202 gene expressions are well correlated to pH





## Back to the biological data

202 gene expressions are well correlated to pH



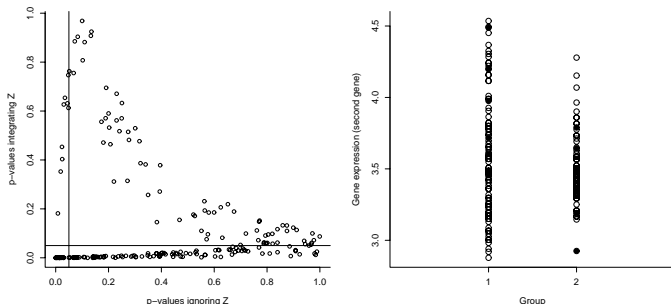
Two examples of big differences between both strategies

- Example 1: p-value = 0.890 (without pH) vs 0.007 (with pH)



## Back to the biological data

202 gene expressions are well correlated to pH



Two examples of big differences between both strategies

- Example 1: p-value = 0.890 (without pH) vs 0.007 (with pH)
- Example 2: p-value = 0.035 (without pH) vs 0.650 (with pH)



## Back to the biological data

Global results after BH procedure (only 202 moderated p-values)

With pH	Ignoring pH		Sum
	Positive	Negative	
Positive	10	56	66
Negative	0	3376	3376
Sum	10	3432	3442

- 1 Introduction
- 2 The gene-by-gene test statistics : using an auxiliary variable
- 3 Impact on the Benjamini-Hochberg procedure results
- 4 Future work



## Future work

- Mathematical assessments of some properties
  - Control of FDR
  - Impact of  $\rho$  on the *NDR*
  - Impact on FDR and NDR variance
  - Relaxing the hypothesis of an equal within-group correlation ?
  - How to deal with multiple Z variables ?
  - Adaptive step-up procedures involving auxiliary variables




## Future work

- Mathematical assessments of some properties
  - Control of FDR
  - Impact of  $\rho$  on the *NDR*
  - Impact on FDR and NDR variance
  - Relaxing the hypothesis of an equal within-group correlation ?
  - How to deal with multiple Z variables ?
  - Adaptive step-up procedures involving auxiliary variables
- Deriving optimal sampling designs for microarray experiments
  - Accounting for experimental costs
  - Optimal allocation for a given equivalent number of slides



## Future work

- Mathematical assessments of some properties
  - Control of FDR
  - Impact of  $\rho$  on the *NDR*
  - Impact on FDR and NDR variance
  - Relaxing the hypothesis of an equal within-group correlation ?
  - How to deal with multiple Z variables ?
  - Adaptive step-up procedures involving auxiliary variables
- Deriving optimal sampling designs for microarray experiments
  - Accounting for experimental costs
  - Optimal allocation for a given equivalent number of slides
- Development of a  package