



Statistical methodology for the analysis of dye-switch microarray experiments

Tristan Mary-Huard, Julie Aubert, Nadera Mansouri, Olivier Sandra,
Jean-Jacques Daudin

► To cite this version:

Tristan Mary-Huard, Julie Aubert, Nadera Mansouri, Olivier Sandra, Jean-Jacques Daudin. Statistical methodology for the analysis of dye-switch microarray experiments. 24. International Biometric Conference, Jul 2008, Dublin, Ireland. 17 p. hal-01197571

HAL Id: hal-01197571

<https://hal.science/hal-01197571>

Submitted on 6 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Statistical methodology for the analysis of dye-switch microarray experiments

Tristan Mary-Huard, Julie Aubert, Nadera Mansouri-Attia, Olivier
Sandra, Jean-Jacques Daudin

UMR AgroParisTech/INRA 518
UMR INRA/ENVA/CNRS 1198, Jouy en Josas



We consider a two-color microarray experiment where two treatments are compared, and where the number of slides is fixed.

We expect the experimental design to guarantee

- the balance between the two fluorochromes,
- a good compromise between technical and biological replicates.

We also need **efficient methods** for the statistical analysis of the proposed design.

Globally balanced design : dye-switch

array	1	2	3	4	5	6	7	8	9	10
Cy5	A1	B2	A3	B4	A5	B6	A7	B8	A9	B10
Cy3	B1	A2	B3	A4	B5	A6	B7	A8	B9	A10

- A_i : i^{th} biological sample in condition A
- The 10 arrays are independent
- 10 biological replicates per condition, no technical replicate.

The **dye-switch** design is often considered as optimal :

- ★ maximum number of biological replicates,
- ★ variance estimated with 9 *df*,
- ★ the statistical analysis is straightforward.

Individually-balanced design : dye-swap

array	1	2	3	4	5	6	7	8	9	10
Cy5	A1	B1	A2	B2	A3	B3	A4	B4	A5	B5
Cy3	B1	A1	B2	A2	B3	A3	B4	A4	B5	A5

- 5 pairs of biological samples
- Each pair is used two times (with reverse dyes)
- Each dye-swap is independent of the other ones

The **dye-swap** design is commonly used :

- ★ used when the technical error is high,
- ★ the statistical analysis (after averaging by swap) is straightforward,
- ★ variance estimated with only 4 *df*.

Individually-balanced : “hybrid” dye-switch

array	1	2	3	4	5	6	7	8	9	10
Cy5	A1	B1	A2	B2	A3	B3	A4	B4	A5	B5
Cy3	B1	A2	B2	A3	B3	A4	B4	A5	B5	A1

- 5 biological replicates per condition
- Each individual is used two times (one with Cy3, one with Cy5).
- Array k is correlated with $k - 1$ and $k + 1$

The **hybrid** design could be an alternative :

- ★ variance estimated with more than 4 *df*.
- ★ the statistical analysis is not straightforward,

⇒ Need of mixed models.

Statistical model for a gene

After log-transformation :

$$\begin{aligned} X_i &= \mu_A + \delta_{I(i)} + B_{j(i)} + M_i + T_i \\ Y_i &= \mu_B + \delta_{I'(i)} + B_{j'(i)} + M_i + T'_i \end{aligned}$$

$B_{j(i)} \sim N(0, \sigma_B^2)$, $j(i)$: sample number corresponding to condition A and array i .

$M_i \sim N(0, \sigma_M^2)$ effect of the spot associated to the gene under concern in microarray i

$T_i \sim N(0, \sigma_T^2)$, includes the steps of labeling, hybridization and measure of intensity of fluorescence.

Model on the difference of expressions

For slide $i = 1, \dots, 2n$:

$$\begin{aligned} D_i &= X_i - Y_i \\ &= \mu + (-1)^{i+1} \delta + BD_i + TD_i \end{aligned}$$

- $\mu = \mu_A - \mu_B$: true differential expression between A and B ,
- $BD_i = B_{j(i)} - B_{j'(i)} \sim N(0, 2\sigma_B^2)$,
- $TD_i = T_i - T'_i \sim N(0, 2\sigma_T^2)$,
- $\delta = \delta_1 - \delta_2$: gene-specific dye bias.

$$\begin{aligned} E(D_i) &= \mu + (-1)^{i+1} \delta \\ V(D_i) &= 2\sigma_B^2 + 2\sigma_T^2 \\ \text{cov}(D_i, D_j) &= 0, \text{ except } \text{cov}(D_i, D_{i+1}) = \sigma_B^2. \end{aligned}$$

Unbiased estimate of $V(\bar{D})$

Unbiased estimator of μ : $\bar{D} = \frac{1}{2n} \sum D_i$

$$V_{\bar{D}} = \frac{1}{2n} (4\sigma_B^2 + 2\sigma_T^2)$$

Naive variance estimate : $S^2 = \frac{1}{2n-2} \sum_i (D_i - \bar{D}_{(i)})^2$

$$E(S^2) = 2\sigma_B^2 + 2\sigma_T^2 < 4\sigma_B^2 + 2\sigma_T^2$$

Unbiased variance estimate $C = \frac{1}{2n-4} \sum_i (D_i - \bar{D}_{(i)})(D_{i+1} - \bar{D}_{(i+1)})$

$$E(C) = \sigma_B^2$$

$$\begin{aligned} S_c^2 &= \frac{1}{2n} (S^2 + 2C) \\ E(S_c^2) &= \frac{1}{2n} (4\sigma_B^2 + 2\sigma_T^2) \end{aligned}$$

Test statistics (for each gene)

Mixed Model with ML estimation (ML)

Mixed Model with REML estimation (REML)

Naive Method (NM)

$$T_N = \sqrt{2n} \frac{\bar{D}}{\sqrt{S^2}}$$

Unbiased Paired Method (UP)

$$T_{UP} = \sqrt{2n} \frac{\bar{D}}{\sqrt{S_c^2}}$$

Level (in %) of the test procedures (simulation of 10000 genes, nominal level= 5%)

$$\sigma_B^2 = 0.5$$

Method	$n = 5$	$n = 10$	$n = 20$	$n = 30$
NM	7.0 (0.3)	7.3 (0.3)	7.3 (0.3)	7.6 (0.3)
UP	5.2 (0.2)	5.2 (0.2)	5.3 (0.2)	5.3 (0.2)
ML	8.5 (0.3)	8.6 (0.3)	8.3 (0.3)	8.3 (0.3)
REML	4.8 (0.2)	4.2 (0.2)	4.5 (0.2)	4.9 (0.2)

$$\sigma_B^2 = 2$$

Method	$n = 5$	$n = 10$	$n = 20$	$n = 30$
NM	13.2 (0.3)	14.0 (0.3)	14.0 (0.3)	14.2 (0.3)
UP	8.2 (0.3)	6.9 (0.2)	6.0 (0.2)	5.8 (0.2)
ML	12.5 (0.4)	11.1 (0.4)	9.9 (0.3)	9.9 (0.3)
REML	14.7 (0.4)	8.6 (0.3)	5.9 (0.2)	5.5 (0.2)

Power of UP and REML tests

$\mu = 1$

n	σ_B^2	UP	REML
5	0.5	13.6	10.6
5	2	5.0	12.9
10	0.5	39.3	34.0
10	2	7.8	9.1
20	0.5	80.1	78.1
20	2	14.5	13.9
30	0.5	95.5	95.0
30	2	22.5	21.7

$\mu = 3$

n	σ_B^2	UP	REML
5	0.5	92.1	86.7
5	2	29.4	34.7
10	0.5	100	99.6
10	2	63.5	63.1
20	0.5	100	100
20	2	94.8	94.5
30	0.5	100	100
30	2	99.6	99.5

Compared CPU time

n	UP CPU	REML CPU	no REML convergence
5	2.3	787	56.9
10	2.6	212	5
20	2.8	467	0
30	3.2	1046	0.16

User CPU time of procedures UP and REML ($\sigma^2 = 0.5$).

Last column : average number of genes for which REML did not converge.

Embriogenomics experiment

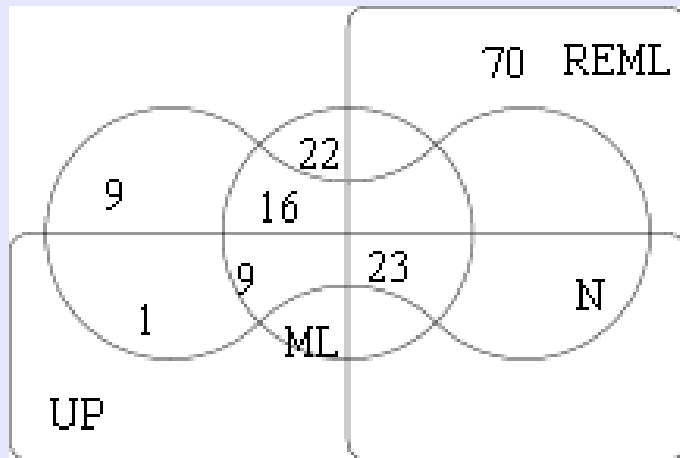
In Mammals, the implantation of the embryo is a key event in the establishment of a pregnancy.

Microarray experiment : 13300 genes in the endometrium of cows. Only 5 animals were available for each condition \Rightarrow hybrid dye-switch design.

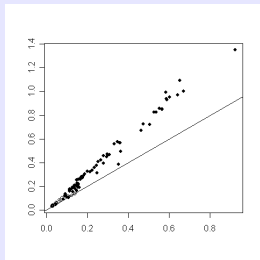
Condition A : cyclic (day 20 of cycle)

Condition B : pregnant (day 20 of pregnancy).

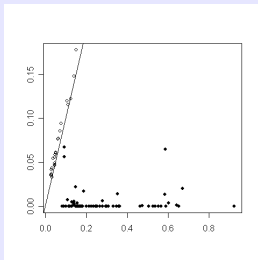
Number of genes DE found by 4 methods



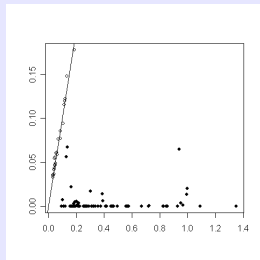
Plots of variance estimates for 93 genes



x-axis :ML
y-axis :UP



x-axis :ML
y-axis :REML



x-axis :UP
y-axis :REML

black points : classified DE only by REML
circle : classified DE by all methods.

Practical guidelines :

- ★ Correlations must be taken into account !
- ★ UP should be preferred to REML (more robust, computationally efficient).
- ★ ANOVA estimators for variances should be considered.

Extension to more complex designs :

- ★ Loop designs
- ★ Interwoven loop designs

Often proposed for their theoretical properties (Wit et al 05, Altman & Hua 06) but with no practical considerations.

T. Mary-Huard, J. Aubert , N. Mansouri-Attia, O. Sandra & J.J. Daudin (2008), *Statistical methodology for the analysis of dye-switch microarray experiments*, BMC Bioinformatics, **9**, 98.

E. Wit, A. Nobile, & R. Khanin (2005), *Near-optimal designs for dual channel microarray studies*, JRSSC, **54**(5), 817–830

N.S. Altman & J. Hua (2006), *Extending the loop design for two-channel microarray experiments*, Genet. Res., **88**, 153–163