BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 9 2005, pages 1995–2000
doi:10.1093/bioinformatics/bti302

Gene expression

Evaluation of the gene-specific dye bias in cDNA microarray experiments

Marie-Laure Martin-Magniette[1,2,*], Julie Aubert[2], Eric Cabannes[3] and Jean-Jacques Daudin[2]

[1]URGV UMR INRA 1165—CNRS 8114—UEVE, 2 rue Gaston Crémieux, CP 5708, 91057 Evry Cedex, France,
[2]UMR INAPG/ENGREF/INRA MIA 518, 16 rue C. Bernard, 75231 Paris Cedex 05, France and
[3]Laboratoire d'Immunologie Virale, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris, France

Received on July 7, 2004; revised on January 25, 2005; accepted on January 27, 2005
Advance Access publication February 2, 2005

ABSTRACT

Motivation: In cDNA microarray experiments all samples are labeled with either Cy3 or Cy5. Systematic and gene-specific dye bias effects have been observed in dual-color experiments. In contrast to systematic effects which can be corrected by a normalization method, the gene-specific dye bias is not completely suppressed and may alter the conclusions about the differentially expressed genes.

Methods: The gene-specific dye bias is taken into account using an analysis of variance model. We propose an index, named label bias index, to measure the gene-specific dye bias. It requires at least two self–self hybridization cDNA microarrays.

Results: After lowess normalization we have found that the gene-specific dye bias is the major source of experimental variability between replicates. The ratio (R/G) may exceed 2. As a consequence false positive genes may be found in direct comparison without dye-swap. The stability of this artifact and its consequences on gene variance and on direct or indirect comparisons are addressed.

Availability: http://www.inapg.inra.fr/ens_rech/mathinfo/recherche/mathematique

Contact: mlmartin@inapg.fr

INTRODUCTION

Many experimenters and statisticians (Kerr *et al*., 2002; Churchill, 2002) recommend using dye-swap design in cDNA microarray experiments to correct gene-specific dye bias. This artifact is not suppressed by normalization procedures such as the lowess (Yang *et al*., 2002). For a reference design some experimenters claim that dye-swaps are not necessary (Sterrenburg *et al*., 2002) whereas others use dye-swap design to preclude gene-specific dye bias (Pritchard *et al*., 2001; Brem *et al*., 2002). In direct comparison, even when the labeling artifact is better recognized, its consequences are often minimized. For example Yue *et al*. (2001) wrote 'Any variation observed in differential expression was likely a result of real variations in experimental mRNA levels rather than an artifact of the labeling system.' Tseng *et al*. (2001) described the gene*label interaction but concluded 'Theoretically some degree of gene-label interaction may exist. However this interaction appears to be insignificant in magnitude compared to other sources of variation in the present experiment.'

To our knowledge, few papers have investigated the influence of the gene-specific dye bias: Dombkowski *et al*. (2004) have shown that dye orientation can significantly influence results on differential analysis in a reference design. They have estimated that over 20% of the conclusions of their differential analysis may be inaccurate using an approach with single dye orientation. They did not identify the cause of the bias, but have urged the experimenters to use dye-swap until this artifact is better characterized. Rosenzweig *et al*. (2004) have investigated the nature of the gene-specific dye bias on a direct comparison experiment. Their analysis suggests that this artifact may concern the same probes. They proposed in their paper a new and less expensive design than the dye-swap, which attenuates the gene-specific dye bias but does not completely correct it.

In this paper, we propose an index to evaluate the magnitude of the gene-specific dye bias. The idea of the index comes from an analysis of two self–self hybridization slides. When we analyzed them, we were surprised to obtain many differentially expressed genes. The reason is that the mean log-ratio $\log_2(R_1R_2/G_1G_2)$ was wrongly calculated in place of $\log_2(R_1G_2/G_1R_2)$, where R_i and G_i denote respectively the red and green intensity on the array i. With the mean log-ratio $\log_2(R_1G_2/G_1R_2)$, no differentially expressed genes were obtained, as was expected. We have been amazed by the importance of the effect of a simple reverse of dye. To better understand the phenomenon we have written the corresponding statistical model, and deduced an index to estimate the magnitude of the gene-specific dye bias.

The paper is organized as follows. In the next section we present the statistical model taking gene-specific dye bias into account, and an index [label bias index (LBI)] to evaluate the magnitude of this artifact. Next the LBI is computed on experiments concerning several array types and organisms. We note that it is almost constant for each array type but varies from one to another. One array type seems to have low gene-specific dye bias. We are not able to explain the reasons, but this fact shows that it is possible to control this artifact. Finally we discuss the consequence of the gene-specific dye bias in direct and indirect comparisons, and try to give some insight into the mechanism of this bias.

[*]To whom correspondence should be addressed.

Published by Oxford University Press 2005.

METHODS

This section is devoted to the statistical model. We underline the importance of keeping the interaction between gene and dye in the model to take gene-specific dye bias into account in the differential analysis, and we evaluate the gene-specific dye bias.

Model allowing for gene-specific dye bias

A dye-swap experiment consists of two replicate microarrays where opposite dye orientations are used. Thus each RNA sample is labeled with each dye. We consider an experiment where p dye-swaps are made. To study the data, we use the analysis of variance. Our notations follow those of Kerr *et al.* (2002). Let Y_{ijkg} be the logarithm base 2 of the measurement for array i, dye j, RNA sample k and gene g. We consider the following model:

$$Y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (DG)_{jg} + E_{ijkg}, \qquad (1)$$

where A_i is the i-th array effect, D_j is the j-th dye effect, V_k is the k-th RNA sample effect, G_g is the g-th gene effect, and $(DG)_{jg}$ and $(VG)_{kg}$ are the corresponding interaction terms. The terms E_{ijkg} represent independent random errors with mean 0. If the RNA sample $k = 1$ is labeled with the dye $j = 1$ in the first array $i = 1$, then the observed difference of expression between the two RNA samples on the array i equals

$$\begin{aligned} Z_{ig} = {} & V_1 - V_2 + (-1)^i (D_1 - D_2) + (VG)_{1g} - (VG)_{2g} \\ & + (-1)^i \{(DG)_{1g} - (DG)_{2g}\} + \tilde{E}_{ig}, \end{aligned}$$

where the errors $\tilde{E}_{ig}$ are independent random variates with mean 0.

To remove systematic biases, we perform an array-by-array normalization using the lowess procedure (Yang *et al.*, 2002). It suppresses the first four constant terms, and is supposed to alleviate the DG terms and not to alter the VG terms. We refer to the work of Kerr *et al.* (2002) for an explanation. After the normalization step, the observed difference of expression between the two RNA samples on the array i equals

$$Z'_{ig} = (VG)_{1g} - (VG)_{2g} + (-1)^i \{(DG)'_{1g} - (DG)'\}_{2g} + F_{ig},$$

where the errors F_{ig} are random variates with mean 0. The normalization step implies that $\sum_{g=1}^{G} Z'_{ig} = 0$; therefore the errors F_{ig} are not independent by construction, and they verify that $\sum_{g=1}^{G} F_{ig} = 0$. It implies a weak structural dependence of order $1/G$ between the F_{ig}. In the following we assume that the F_{ig} are independent. The departure from this assumption is too weak to have any practical importance provided that $G \geq 1000$. The difference $(VG)_{1g} - (VG)_{2g}$ is the true difference of expression between the two RNA samples. It is the difference of interest for identifying differential expressed genes. When it is non-null, it states that the gene is not transcribed in the same manner in the two RNA samples. The difference $(DG)'_{1g} - (DG)'_{2g}$ represents what is called the gene-specific dye bias. When it is non-null, it states that the probe corresponding to the gene g incorporates one of the dyes preferentially. To simplify the notations, we denote the difference $(VG)_{1g} - (VG)_{2g}$ by δ_g and the difference $(DG)'_{1g} - (DG)'_{2g}$ by β_g. The observed difference of the gene g between the two RNA samples in the array i is now re-written:

$$Z'_{ig} = \delta_g + (-1)^i \beta_g + F_{ig}, \qquad (2)$$

where δ_g is the gene g differential expression and β_g the specific dye bias of the gene g. From this model we can estimate for each gene the differential expression between the two RNA samples and the gene-specific dye bias by

$$\hat{\delta_g} = \frac{1}{2p} \sum_{i=1}^{2p} Z'_{ig}$$

and

$$\hat{\beta_g} = \frac{1}{2p} \sum_{i=1}^{2p} (-1)^i Z'_{ig}.$$

When at least two dye-swaps are available ($p \geq 2$), we can also estimate the variance of F_{ig}, say σ_g^2, by the empirical estimator defined by

$$\widehat{\sigma_g^2} = \frac{1}{2p-2} \sum_{i=1}^{2p} (Z'_{ig} - \widehat{\delta}_g - (-1)^i \widehat{\beta}_g)^2.$$

It is then possible to perform a differential analysis and also an analysis of the gene-specific dye bias. For the latter purpose, it suffices to test the null hypothesis $\{\beta_1 = \cdots = \beta_G = 0\}$ against the alternative hypothesis {At least one gene is such that $\beta_g \neq 0$}. The associated test statistic can be viewed as a global index to evaluate the gene-specific dye bias. It is easily and quickly computed. We name it the LBI and it is defined by

$$\text{LBI} = \frac{\sum_{g=1}^{G} \widehat{\beta_g^2}}{\sum_{g=1}^{G} \widehat{\sigma_g^2}}. \qquad (3)$$

Under the null hypothesis and assuming that $\sum_g^2$, the LBI is distributed as a Fisher distribution with $[G - 1, (2p - 2)(G - 1)]$ degrees of freedom. The null hypothesis is rejected as soon as the test statistic is greater than $F_{G-1,(2p-2)(G-1)}(1 - \alpha)$, where $F_{a,b}(\alpha)$ denotes the α-quantile of a Fisher distribution with (a, b) degrees of freedom. Note that in practice, the null hypothesis may often be rejected since the power of the test is high. So to decide if the gene-specific dye bias is important, the LBI can be also compared with the expectation of a Fisher distribution, given by $1 + \{1/[(p - 1)(G - 1) - 1]\}$.

Although it is possible to take into account the gene-specific dye bias, in many studies the authors prefer to neglect it (e.g. Tseng *et al.*, 2001; Comander *et al.*, 2004). This leads to setting $\beta_g = 0$ for $g = 1, \ldots, G$ in the model (2). The variance of F_{ig} is thus estimated by

$$\tilde{\sigma_g^2} = \frac{1}{2p-1} \sum_{i=1}^{2p} (Z'_{ig} - \widehat{\delta}_g)^2.$$

Straightforward calculations show that $\tilde{\sigma_g^2}$ is a biased estimator of σ_g^2 if β_g differs from 0. To be precise, the bias equals $2p\beta_g^2/(2p - 1)$. Therefore assuming wrongly that the β_g are null leads to overestimating the variance σ_g^2. Hence the power of the test for detecting a difference of expression will be lower when $\tilde{\sigma_g^2}$ is used in place of $\widehat{\sigma_g^2}$: some differentially expressed genes will not be detected.

When only one dye swap is made, the model (2) is over-parametrized: the number of parameters is larger than the number of observations. It is thus impossible to estimate simultaneously the difference of expression (δ_g), the gene-specific dye bias (β_g) and the variance (σ_g^2). Only two parameters per gene can be estimated. Since the major interest is the differential analysis, the parameter β_g is usually supposed to be null. In the following section, we propose a method to assess this assumption.

Evaluation of the gene-specific dye bias from self–self hybridization slides

As noticed above, when only one dye-swap is available, the statistical model (2) is no longer usable to study the observed difference of expression between two different RNA samples. Nevertheless if we consider self–self hybridization slides where the same RNA sample is hybridized against itself, it guarantees that the true difference of expression is null ($\delta_g = 0$) and thus the model (2) becomes a one-way ANOVA model:

$$Z'_{ig} = (-1)^i \beta_g + F_{ig}.$$

It is thus possible to estimate the magnitude of the gene-specific dye bias. For that purpose we calculate the LBI, defined as previously by the statistic of Fisher to test the null hypothesis $\{\beta_1 = \cdots = \beta_G = 0\}$. If $p \neq 1$, it is defined by:

$$\text{LBI} = \frac{\sum_{g=1}^{G} \hat{\beta_g}^2}{\sum_{g=1}^{G} \hat{\sigma_g}^2}, \qquad (4)$$

with $\hat{\delta_g} = 0$, for all $g = 1, \ldots, G$. Under the null hypothesis, the LBI is distributed as a Fisher distribution with $[G - 1, (2p - 1)(G - 1)]$ degrees of

Table 1. LB1 and gene-specific dye bias from 11 self–self hybridization arrays

Organism/array	Dataset	RegSS	RSS	LBI	(a)	(b)	Mean LR	Min. LR	Max. LR
Human/array 1	Wt t1	0.158	0.034	4.64	0	120	0.87	−1.19	1.58
Human/array 1	Control t2	0.156	0.027	5.68	0	153	0.45	−1.46	1.52
Human/array 1	SDF t3	0.221	0.054	4.07	0	2	1.97	1.73	2.21
Human/array 1	Wt t4	0.227	0.047	4.86	0	113	1.19	−1.16	1.81
Human/array 1	Control t5	0.280	0.060	4.64	0	33	0.19	−1.61	1.36
Human/array 1	SDF t6	0.278	0.043	6.42	0	189	1.42	−1.26	2.95
Human/array 2	SDF t2	0.120	0.012	10.29	0	8	−1.11	−1.35	−0.98
Human/array 1	SDF t2	0.080	0.016	5.15	0	3	−2.05	−2.85	−1.45
At/CATMA	leaf	0.028	0.016	1.79	0	0	—	—	—
At/CATMA	bud	0.041	0.035	1.17	0	0	—	—	—
At/CATMA	bud	0.043	0.035	1.24	0	0	—	—	—

Wt, wild type; t, time; RegSS, regression sum of squares; RSS, residuals sum of squares; LBI, label bias index; At, *A.thaliana*; (a), number of genes differentially expressed; (b), number of genes having a significant dye bias; LR, log ratio for genes having a significant dye bias.

freedom. Consequently the null hypothesis is rejected as soon as the LBI is greater than $F_{G-1,(2p-1)(G-1)}(1-\alpha)$. It readily follows that for $p = 1$,

$$\mathrm{LBI} = \frac{\sum_{g=1}^{G}(Z'_{1g} - Z'_{2g})^2}{\sum_{g=1}^{G}(Z'_{1g} + Z'_{2g})^2}. \tag{5}$$

Under the null hypothesis, its distribution is a Fisher with $(G-1, G-1)$ degrees of freedom. The null hypothesis is thus rejected as soon as the LBI is greater than $F_{G-1,G-1}(1-\alpha)$. As previously the null hypothesis is often rejected since the number of degrees of freedom is of the magnitude of G. Consequently to decide if the gene-specific dye bias is important, the LBI can be compared with the expectation of the Fisher distribution, which is equal to $(G-1)/(G-3) \sim 1$.

The LBI gives a global overview of the gene-specific dye bias. It is also interesting to have a gene-by-gene approach. For that purpose we propose to test $\{\beta_g = 0\}$ for each gene. As in the differential analysis, it is important to model the variance suitably. We have chosen to use the mixture model of Delmar *et al*. (2004). This method identifies clusters of genes with equal variance and has the good properties of keeping a good control of false positive genes and having a good power of detection. We use the Bonferroni method (with a type I error equal to 5%) in order to keep a strong control of the false positives in a multiple comparison context (Benjamini and Hochberg, 1995).

RESULTS

Data

We calculate the LBI from several self–self hybridization arrays of human and *Arabidopsis thaliana* cells.

Experiments from human cells

Resting CD4+ T cells isolated from peripheral mononuclear blood cells of healthy donors were stimulated either by the SDF-1a chemokine (SDF), or infected by the NL4-3 wild-type strain of HIV-1 (WT) or left untreated (control). For each treatment, an aliquot was removed from the cell culture at 6 different time-points over a 24 h period (30 min, 2, 4, 8, 12 and 24 h) and RNA was extracted using the RNeasy mini kit (Qiagen) according to the manufacturer's recommendations. Samples of mRNA were submitted to the T7 amplification procedure described by Phillips and Eberwine (1996), in a very similar way as previously reported (Wang *et al*., 2000). An aliquot of 4 µg of amplified RNA from a given condition (SDF, wild type or control) at a chosen time (Table 1), was used for reverse transcription and aminoallylcoupling (for details see http://cmgm.stanford.edu/pbrown/protocols/amino-allyl.htm and http://www.microarrays.org/pdfs/amino-allyl-protocol.pdf). The two halves of each aminoallyl-cDNA were coupled to NHS-Cy3 and NHS-Cy5, then purified together and hybridized onto the same array to produce a self–self hybridization.

For the first six experiments of Table 1, duplicate experiments using cells from two independent donors (RNA from same time and condition) were performed on the same day. For the next two experiments, the procedures remained the same except that the amount of starting material was doubled in order to hybridize a couple of arrays (same sample duplication).

All samples were hybridized on the same type of array consisting of 11 520 clones except for the seventh dye-swap, which was hybridized on another array of 11 616 clones spotted in duplicate. These experiments are part of a larger study that will be published elsewhere.

The arrays were scanned on a GenePix 4000A scanner (Axon Instruments, Foster City, USA) and images were analyzed by the GenePix Pro 4.0 software (Axon Instruments, Foster City, USA). For each array, the raw data comprised the logarithm base 2 of median feature pixel intensity at wavelength 635 nm (red) and 532 nm (green). No background was subtracted. The array-by-array normalization was performed to remove systematic biases. First, we excluded spots that were considered badly formed features. Then we performed a global intensity-dependent normalization using the lowess procedure (Yang *et al*., 2002). Finally, for each block, the log-ratio median calculated over the values for the entire block was subtracted from each individual log-ratio value.

Experiments from *A.thaliana* cells

Four sets of 100 *A.thaliana* Col-0 plants were grown on horticultural potting soil (Tref substrate with NFU 44-571 fertilizer, BAAN SA, Vulaines, France) under cool white light at 100 µmol m-2 s-1 with a 16-h photoperiod at 22°C and 50% humidity. Pooled samples of the flowers or the buds were harvested. The RNA extraction and target labeling were described as in Lurin *et al*. (2004).

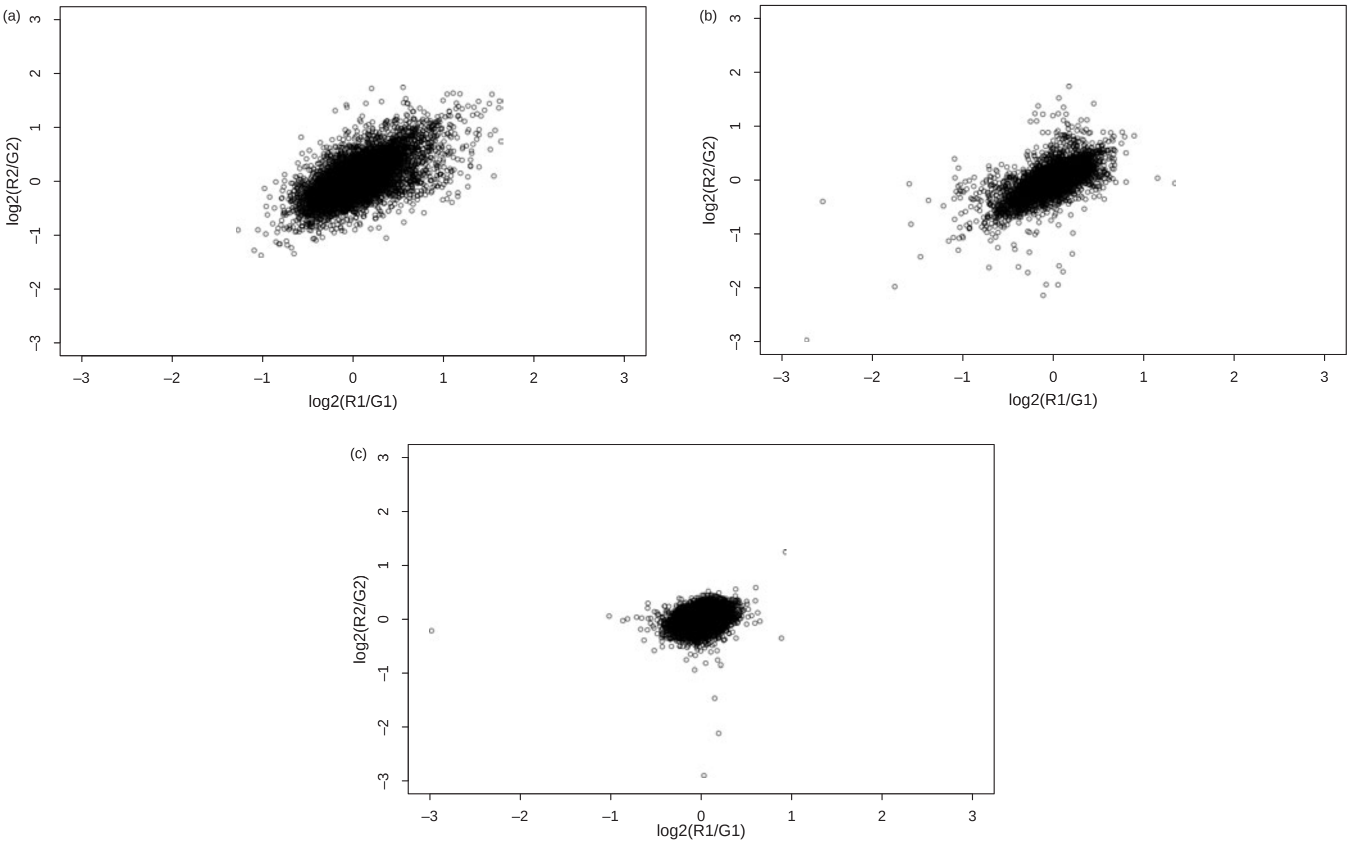

Fig. 1. Plots of the log-ratios $\log_2(R/G)$: first slide in x-axis and second slide in y-axis. (a) human/array 1, (b) human/array 2, (c) At/CATMA.

All samples were hybridized on CATMA array containing 24 576 Gene Specific Tags from *A.thaliana* (Crowe *et al*., 2003).

The arrays were scanned on a GenePix 4000A scanner (Axon Instruments, Foster City, USA) and images were analyzed by GenePix Pro 3.0 (Axon Instruments, Foster City, USA). For each array, the raw data and array-by-array normalization were respectively defined and performed as for the slides of the human cell experiments.

LBI

Table 1 summarizes the LBI computed for the 11 experiments. The LBI is the ratio between the Regression sum of squares (RegSS $= \sum_{g=1}^{G} \widehat{\beta}_g^2$) and the Residuals sum of squares (RSS $= \sum_{g=1}^{G} \widehat{\sigma}_g^2$). The RegSS, RSS and LBI values are respectively presented in the first, second and third columns of Table 1. We note that the RegSS is always $>$ RSS, so the LBI is always $>$1. The LBI shows that the RegSS is more than three times as high as the RSS in arrays 1 and 2 and less than twice as high as the RegSS in the CATMA array. So the dye bias is more important in the human experiments than in the experiments of *A.thaliana*. We recall that the ideal LBI (no gene-specific dye bias) is close to 1. In the experiments from *A.thaliana* cells, we have at our disposal four slides of CATMA, where the same sample of buds has been hybridized against itself. We use these four slides to evaluate the robustness of the LBI by calculating it on the six possible pairs of slides. The associated LBI varies between 1.12 and 1.26, which proves its robustness. We point out that the robustness has not been evaluated for arrays with a relatively high LBI because necessary data were not available.

To further illustrate the impact of the gene-specific dye bias, we plot the log-ratios $\log_2(R/G)$ for the two slides of the same dye-swap, for all the experiments (Fig. 1). As we have two replicates of self–self hybridization slides, nothing is expected to be seen. However one can see that there is a positive correlation between the two replicates. The only possible cause for such a correlation is the dye bias. Some genes have a higher intensity when labeled with one dye than with the other. Therefore the log-ratio $\log_2(R/G)$ is repeatedly higher (or lower) than it should be. This dye effect is higher on human experiments (correlation between 0.61 and 0.73) than on *A.thaliana* (correlation between 0.08 and 0.33). This confirms that the dye bias plays an important role in the experimental variability in the human experiments. In contrast, the dye bias seems to be better controlled in the *A.thaliana* experiments.

We also calculate the correlations between all the $\hat{\beta}_g$ for each human/array 1 experiment. These correlations are comprised between 0.45 and 0.81 (Table 2). As the array type is the same but experimental conditions vary, these correlations suggest that the dye bias may be attributed to the gene. Note that the possible gene effect is confounded with its position on the slide. Therefore it is impossible to separate the two possible causes of the labeling bias which are the nucleic composition of the probe and the spotting effect (Mary-Huard *et al*., 2004).

Table 2. Correlation between the dye bias for all the human/array 1 experiments

Experiment 1	Experiment 2	Correlation
Wild type time 1	Control time 2	0.807
Wild type time 1	SDF time 3	0.762
Wild type time 1	Wild type time 4	0.745
Wild type time 1	Control time 5	0.672
Wild type time 1	time 6	0.646
Control time 2	SDF time 3	0.718
Control time 2	Wild type time 4	0.724
Control time 2	Control time 5	0.673
Control time 2	SDF time 6	0.588
SDF time 3	Wild type time 4	0.776
SDF time 3	Control time 5	0.707
SDF time 3	SDF time 6	0.657
Wild type time 4	Control time 5	0.763
Wild type time 4	SDF time 6	0.655
Control time 5	SDF time 6	0.533

Identification of genes having a specific dye bias

After a global analysis of the gene-specific dye bias we identify the genes which are concerned. However to begin with, we assess the quality of the self–self hybridization slides by testing that each δ_g is null. Similar to the test of $\{\beta_g = 0\}$ for each gene, we use the mixture model of Delmar *et al*. (2004). The control of the false positives is done with the Bonferroni method at a level of 5%.

No gene is found to be regulated (column (a) in Table 1). Then, in order to identify genes with a significant dye bias, we test the labeling artifact using also the mixture model of Delmar *et al*. (see Methods section). Column (b) of Table 1 shows that between 0 and 189 genes have a significant gene-specific dye bias. This artifact is important in the human experiments and does not appear in the *A.thaliana* experiments. These results are in agreement with the LBI calculated in the previous section. Furthermore, all the genes having a significant dye bias are classified in the highest variance group from the differential analysis. This suggests that many genes from the highest variance group could not be detected as differentially expressed only because their ‘pure’ experimental variability is increased by a specific dye bias effect. This confirms that the presence of gene-specific dye bias can increase the false negative rate and so decrease the power of detection.

Table 1 contains the mean, minimal and maximal values of the $\hat{\beta}_g$ for the detected genes. One can see that the gene-specific dye bias may multiply or divide the ratio by a factor >2 which is sizeable. An analysis on the intensity level of the genes with a high specific dye bias (data not shown) shows that the intensity of these genes is in a large range between 5.5 and 15.7, with a median value between 9.5 and 10.2. Figure 2 plots the specific dye bias according to the intensity level for the first human/array 1 experiment. We can see that the magnitude of the artifact is near 0 when the intensity level is not very far from the background level. This confirms that a gene needs to be transcribed in order to reveal its specific dye bias. For higher values of the intensity level, no dependence is observed between specific dye bias and intensity level. As shown before, all expressed genes can be affected by a specific dye bias whatever their intensity level.

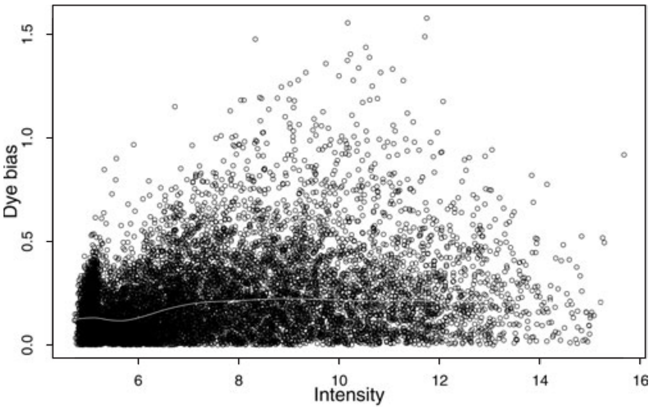

Fig. 2. Plot of the dye bias according to the intensity level in human/array 1.

DISCUSSION

Consequences of the gene-specific dye bias on direct comparison experiments

In direct comparison, two RNA samples are simultaneously hybridized on the same slide. Each sample is labeled with a dye, and it is well known that the two dyes do not have the same incorporation effectiveness. Moreover it appears that some genes are systematically badly labeled by Cy5 or Cy3 (the gene-specific dye bias). For all these reasons dye-swap design is absolutely recommended, although it is costly. Moreover in the first section we have proved that the gene*label interaction increases the experimental variability even in dye-swap experiments and thus decreases the power of the tests for detecting the differentially expressed genes.

In this paper we have proposed the LBI which is a global index to evaluate the magnitude of the gene-specific dye bias. The LBI is easily and quickly computed, and requires at least two self–self hybridization slides. After the LBI calculation we advise carrying out a gene-by-gene analysis. Even if we cannot completely describe the biochemical mechanisms of this bias, it seems that it is an artifact which involves the probes and the labeled targets, since the gene-specific dye bias can be seen only when the gene corresponding to the probe is transcribed. Consequently we advocate using a sample which hybridizes against the most possible probes. Moreover if the LBI is calculated on an array where the probes are duplicated, we think that it is better to work from the probes and not from the mean of the duplicated probes, since the gene-specific dye bias is probe-dependent. All these remarks allow us to think that the method proposed by Rosenzweig *et al*. (2004) is questionable. A condition where all genes would be transcribed simultaneously would be necessary to obtain an effective correction.

In order to investigate the gene-specific dye bias in more detail, it could be interesting for the platforms to include the LBI in their quality-control procedures, because the identification of genes which have specific dye bias is important supplementary information for the differential analysis. Moreover it could help to explain the nature of the phenomenon. According to the result of the *A.thaliana* experiment, this artifact is not an inevitability and can be well controlled. The elimination of the gene-specific dye bias could dramatically decrease the experiment cost by removing the necessity of systematic dye-swap design.

Note that the genes can be clustered either in a group without specific dye bias ($\beta_g = 0$) or in a group with specific dye bias ($\beta_g \neq 0$). The former group has a lower experimental variability than the latter in dye-swap experiments. This explains why the mixture model on gene variances is well suited to microarray experiments (Delmar *et al*., 2004).

Consequences of the gene-specific dye bias on indirect comparison experiments

In indirect comparison an RNA sample is hybridized against a control sample. The associated design is called the reference design. As we mentioned in the introduction, it is widely assumed that reference design does not require dye-swaps. The paper of Dombkowski *et al*. (2004) demonstrated from a microarray data analysis that this assumption is not reliable. By writing the statistical model, we confirm their findings. We take the notations used throughout the paper. To take into account that the gene-specific dye bias appears only when there is transcription, we include in the model (1) the interaction between the RNA sample, the dye and the gene, say (VDG). Let us assume that the dye $j = 1$ is associated with the control sample $k = 0$, then the observed difference of expression between the i-th RNA sample and the control sample is equal to

$$\begin{aligned} Z_{ig} &= Y_{i21g} - Y_{i10g} \\ &= D_2 - D_1 + V_i - V_0 + (VG)_{ig} - (VG)_{0g} + (DG)_{2g} - (DG)_{1g} \\ &\quad + (VDG)_{i2g} - (VDG)_{01g} + \tilde{E}_{ig}. \end{aligned}$$

After the normalization step the observed difference of expression between the RNA sample and the control sample equals:

$$\begin{aligned} Z'_{ig} = (VG)_{ig} - (VG)_{0g} + (DG)'_{2g} - (DG)'_{1g} \\ + (VDG)'_{i2g} - (VDG)'_{01g} + F_{ig}. \end{aligned}$$

Finally, the estimate for the differential expression of gene g between the two RNA samples is thus

$$Z'_{1g} - Z'_{2g} = \delta_g + (VDG)'_{12g} - (VDG)'_{22g} + \tilde{F}_g,$$

where the errors $\tilde{F}_g$ are random variates with mean 0. The gene*label interaction terms vanish but the interactions between the RNA sample, the dye and the gene remain. This is the reason why a dye-swap design is recommended even in indirect comparison.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Brem,R. *et al*. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Churchill,G. (2002) Fundamentals of experimental designs for cDNA microarray. *Nat. Genet.*, **32**, 490–495.

Comander,J. *et al*. (2004) Improving the statistical detection of regulated genes from microarray data using intensity-based variance estimation. *BMC Genomics*, **5**, 17.

Crowe,M. *et al*. (2003) CATMA: a complete *Arabidopsis* GST database. *Nucleic Acids Res.*, **31**, 156–158.

Delmar,P. *et al*. (2004) Varmixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics*, **21**, 502–508.

Dombkowski,A. *et al*. (2004) Gene-specific dye bias in microarray reference designs. *FEBS Lett.*, **560**, 120–124.

Kerr,M.K. *et al*. (2002) Statistical analysis of a gene expression microarray experiment with replication. *Statist. Sin.*, **12**, 203–217.

Lurin,C. *et al*. (2004) Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell*, **16**, 2089–2103.

Mary-Huard,T. *et al*. (2004) Spotting effect in microarray experiments. *BMC Bioinformatics*, **5**, 63.

Phillips,J. and Eberwine,J.H. (1996) Antisense RNA amplification: a linear amplification method for analysing the mRNA population from single living cells. *Methods*, **10**, 283–288.

Pritchard,C. *et al*. (2001) Project normal: defining normal variance in mouse gene expression. *Proc. Natl Acad. Sci. USA*, **98**, 13266–13271.

Rosenzweig,B. *et al*. (2004) Dye-bias correction in dual-labeled cDNA microarray gene expression measurements. *Environ. Health Perspect.* **112**, 480–487.

Sterrenburg,E. *et al*. (2002) A common reference for cDNA microarray hybridizations. *Nucleic Acids Res.*, **30**, e116.

Tseng,G. *et al*. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**, 2549–2557.

Wang,E. *et al*. (2000) High-fidelity amplification for gene profiling. *Nat. Biotechnol.*, **18**, 457–459.

Yang,Y. *et al*. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Yue,H. *et al*. (2001) An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res.*, **29**, E41-1.

BIOINFORMATICS LETTER TO THE EDITOR Vol. 21 no. 12 2005, pages 2803–2804
doi:10.1093/bioinformatics/bti428

Gene expression

Comment on ‘Evaluation of the gene-specific dye bias in cDNA microarray experiments’

Kevin K. Dobbin*, Joanna H. Shih and Richard M. Simon

Biometric Research Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

Received on February 16, 2005; revised and accepted on April 1, 2005
Advance Access publication April 7, 2005

In their paper in *Bioinformatics*, Martin-Magniette *et al.* (2005) recommend complete dye-swap[1] designs for both direct and indirect dual label microarray experiments. These recommendations contradict our previous recommendations (Dobbin *et al*., 2003) for designing experiments, where we suggested minimizing or eliminating the use of dye-swap arrays. We show here that the recommendations of Martin-Magniette *et al.* are fundamentally flawed, and that in most realistic situations performing extensive dye-swap arrays results in a poor experimental design.

The key error made by these authors is that they focus on over-simplified situations in which only two RNA samples are being compared. There are two problems with this approach. First, if the goal is really just to compare gene expression in two RNA samples, then obviously the best design will be to place aliquots from both samples together on each array and label each sample with each dye half the time. So there really is no design question. The second and more serious problem with this approach, however, is that comparing gene expression in two RNA samples is almost never the goal of a microarray experiment. The goal is almost always to draw conclusions that are applicable beyond the particular RNA samples being studied, and this requires independent replication (Simon *et al.*, 2002). Without independent experimental replication, either independent biological samples or independent replications of the entire experiment, depending on the context, one cannot make statistical inferences that apply beyond the RNA samples used. For example, in an experiment to evaluate the effect of different conditions on cell line gene expression, one must perform independent replicates of the experiment, in which multiple, different cell line cultures are grown up under each condition. Similarly, one cannot draw valid conclusions about differential expression in two populations of mice from an experiment that involves just two mice. One needs multiple independent mice from each population to capture the biological variation in the populations.

When multiple independent replicates from different conditions or populations are used in an experiment, then the equation Martin-Magniette *et al.* have derived, based on the model of Kerr *et al.* (2002), is no longer valid. The specific model equation[2] for the log-ratios is $Z'_{ig} = (VG)_{1g} - (VG)_{2g} + (DG)'_{1g} - (DG)'_{2g} + F_{ig}$, where Z_{ig} is the normalized log-ratio for gene g on array i, $(VG)_{1g} - (VG)_{2g}$ is the ‘variety’ effect, $(DG)'_{1g} - (DG)'_{2g}$ is gene-specific dye bias and F_{ig} is the error term. The reason the model is not valid is that it contains a single term, ‘variety,’ which represents both a sample and a condition or population. But samples are different from conditions or populations, so terms need to be added to the model to distinguish between the two, as indicated in Dobbin and Simon (2002). When such terms are added to the model, so that samples are conceptually separated from conditions or populations, the impact of taking multiple subsamples from the same batch of RNA (technical replication) becomes different from the impact of performing biologically independent replicates of the experiment. Without introducing additional terms into the model, technical replication is indistinguishable from biologically independent replication. If we let ‘variety’ represent condition or population, then a term for sample effects needs to be added to the model. Let $S(v)$ indicate a sample from condition or population V. Then the model of Martin-Magniette *et al.* (2003) needs to be changed to:

$$\begin{aligned} Z'_{igs(1)s(2)} &= (VG)_{1g} - (VG)_{2g} + (DG)'_{1g} - (DG)'_{2g} \\ &\quad + (SG)_{s(1)g} - (SG)_{s(2)g} + F_{igs(1)s(2)}. \end{aligned} \tag{1}$$

The goal of the experiment is still to make inferences about the $(VG)_{1g} - (VG)_{2g}$ term which represents differential expression between the classes of samples. But the model change is critical, because it distinguishes between different levels of replication, and results in different conclusions about the optimal experimental design. Also, conclusions about differential expression from such a model apply beyond the individual RNA samples used in the experiment, whereas conclusions based on Kerr *et al.* (2002) model do not (they apply only to the RNA samples used). Experiments with independent replicates from different classes (populations or conditions) are commonly called class comparison experiments (Simon *et al.*, 2002).

Martin-Magniette *et al.* (2005) recommend dye swapping every array in a reference[3] design. For class comparison experiments, there are situations in which a reference design may be reasonable,

*To whom correspondence should be addressed.

[1] An individual array is dye-swapped when, for each of the original batches of RNA which were tagged with Cy3 and Cy5, RNA is drawn from the same two batches and labeled in the opposite way as on the original microarray, and the two labeled samples are hybridized to a second array. When every array in an experiment is dye-swapped, this is called a complete dye-swap design.

[2] Here we follow the notation of Martin-Magniette *et al*. (2005). A simpler and reformulation of the model is presented in the supplemental material.

[3] A dual-label reference design experiment is an experiment that includes the same reference sample on each array, tagged with the same dye.

although we have shown that balanced block designs,[4] which do not use a reference, are more efficient (Dobbin and Simon, 2002). The motivation Martin-Magniette *et al*. present for recommending that reference designs always dye-swap every array is the existence of a three-way sample-by-dye-by-gene interaction, which they hypothesize exists based on a previous study by Dombkowski *et al*. (2004). In the over-simplified model Martin-Magniette *et al*. use, with just two samples and no distinction between samples on the one hand and conditions or populations on the other, the three-way interaction term introduces bias into comparisons in a reference design. The reason this bias is introduced is because the three-way interaction term can be viewed as sample-specific dye bias, and because the model fails to distinguish between samples and conditions/populations, bias related to the sample automatically becomes bias related to the condition/population. But for class comparison experiments, which allow for statistical inference beyond the particular samples studied, and which require a more sophisticated model like the one in Equation (1), dye-swap arrays are not required to remove the bias. Indeed, as we will show, a complete dye-swap reference design is clearly inferior to a reference design in this situation.

For a fixed number of arrays, a complete dye-swap reference design involves half as many independent samples as a reference design. So, dye-swapping every array in a reference design halves the effective sample size. Is such a radical reduction in sample size justified by the existence of the three-way interaction term? The answer is no. To see this, add the three-way interaction terms, $(DGS)_{dgs}$ to the model of Equation (1), and let condition $V=0$ represent the reference sample on each array,

$$\begin{aligned}Z'_{igs(v)s(0)} &= (VG)_{vg} - (VG)_{0g} + (DG)'_{1g} - (DG)'_{2g}\\ &\quad + (SG)_{s(v)g} - (SG)_{s(0)g} + (DGS)_{dgs(v)}\\ &\quad - (DGS)_{dgs(0)} + F_{igs(v)s(0)}. \qquad (2)\end{aligned}$$

Then, if we assume a reference design, the estimate of the difference in gene expression between classes 1 and 2 is: $Z_{\bullet g\bullet(1)0} - Z_{\bullet g\bullet(2)0}$, where $Z_{\bullet g\bullet(v)0}$ indicates the average of the log-ratios over the arrays with samples from class v (sample 0 indicating the reference sample on each array). The expected value of this difference is $E\lfloor Z_{\bullet g\bullet(1)0} - Z_{\bullet g\bullet(2)0}\rfloor = (VG)_{1g} - (VG)_{2g} + (DGS)_{dg\bullet(1)} - (DGS)_{dg\bullet(2)}$, where $(DGS)_{dg\bullet(v)}$ is the average of the interaction effects over samples from condition or population V. (Note that the individual SG sample effects will cancel out of the expected value, so we have omitted them.) If a random effects model is used for the three-way interaction, then $(DGS)_{dgs} \sim N(\mu,\sigma^2), E\lfloor Z_{\bullet g\bullet(1)0} - Z_{\bullet g\bullet(2)0}\rfloor = (VG)_{1g} - (VG)_{2g}$, and the reference design yields unbiased estimates of the class difference. Alternatively, if fixed effects are used for the interaction term, then under the usual model constraints, required for model identifiability, $\sum_{s\in V}(DGS)_{dgs} - (DGS)_{dg0} = 0$ for $V = 1, 2$, yielding $E\lfloor Z_{\bullet g\bullet(1)0} - Z_{\bullet g\bullet(2)0}\rfloor = (VG)_{1g} - (VG)_{2g}$, and the reference design estimates are unbiased. So, under both a fixed-effects and a random-effects model for the interaction term, the reference design yields unbiased estimates of the class distinction—without any dye-swaps. Moreover, the reference design will be more efficient than the complete dye-swap reference design. Intuitively, the reason for the improved efficiency of the reference design is that it allows twice as many samples to be used in the same number of arrays. A more detailed proof of the efficiency advantage appears in Dobbin *et al*. (2003). In conclusion, a reference design provides unbiased and more efficient estimates of differential gene expression than a complete dye-swap reference design for class comparison experiments.

Now we turn to designs that do not involve a reference sample. In this case also, Martin-Magniette *et al*. (2005) recommend a dye-swap design, but it is unclear whether by this they mean a complete dye-swap design, which dye swaps every array, or not. The motivation for recommending dye-swapping arrays in this case is somewhat different from that in the reference design case. But it is still based on the same flawed model. Their motivation is to remove the two-way dye by gene interaction, which we have shown can be done without dye-swapping arrays (Dobbin *et al*., 2003). When one properly distinguishes between samples and conditions/populations, as in our Equation (1), one finds that dye-swapping is much less efficient than independent replication of the experiment with the labeling reversed (such as in a balanced block design). And, using arguments analogous to the reference design situation above, even if sample-by-gene-by-dye interaction terms are present, dye-swapping individual arrays is not necessary to remove the bias from the class comparisons. So, systematically dye-swapping individual arrays in a non-reference design is inadvisable when the goal is class comparison.

Finally, while we have shown that neither the existence of interactions between gene and dye, nor interactions between gene and dye and sample, justify systematically dye-swapping individual arrays, one might wonder if interactions between gene and dye and population/condition would change the situation. These interaction terms would appear as $(DGV)_{dgv} - (DGV)_{dg0}$ in Equation (2). Such an interaction term has not to our knowledge been empirically evaluated. But, for the sake of argument, suppose it did exist. In the case of non-reference designs for class comparison, the bias would cancel out of comparisons between the populations/conditions in a balanced block design, so this design would remain optimal. No dye-swaps would be required. Hence, even under this fairly unlikely scenario, dye-swapping is not a good idea.

In conclusion, the findings of Martin-Magniette *et al*. (2005) must be carefully interpreted within their very limited context, and in practice dye-swap arrays should be used sparingly if at all, particularly in class comparison experiments.

REFERENCES

Cochran,W.G. and Cox,G.M. (1999) *Experimental Designs*, 2nd ed. John Wiley and Sons, New York, NY.

Dobbin,K. and Simon,R. (2002) Comparison of microarray designs for class comparison and class discovery. *Bioinformatics*, **18**, 1438–1445.

Dobbin,K. *et al*. (2003) Statistical design of reverse dye microarrays. *Bioinformatics*, **19**, 803–810.

Dombkowski,A.A. *et al*. (2004) Gene-specific dye bias in microarray reference designs. *FEBS Lett.*, **560**, 120–124.

Kerr,M.K. *et al*. (2002) Statistical analysis of gene expression microarray experiment with replication. *Statist. Sinica*, **12**, 203–217.

Martin-Magniette,M. *et al*. (2005) Evaluation of gene-specific dye bias in cDNA microarray experiments. *Bioinformatics*, **9**, 1995–2000.

Simon,R. *et al*. (2002) Design of studies using DNA microarrays. *Genet. Epidemiol.*, **23**, 21–36.

[4] A balanced block design for two classes pairs a samples from one class with a sample from the other class on each array, balancing the labels used for each class but using each biologically independent sample only once. Balanced block designs generalize to multiple classes, and have a long history in statistical literature (see, for example, Cochran and Cox, 1992).

BIOINFORMATICS

Vol. 21 no. 14 2005, page 3065
doi:10.1093/bioinformatics/bti479

Letter to the Editor

Answer to the comments of K. Dobbin, J. Shih and R. Simon on the paper ‘Evaluation of the gene-specific dye-bias in cDNA microarray experiments’

M.-L. Martin-Magniette, J. Aubert, E. Cabannes and J.-J. Daudin

Received and accepted on April 19, 2005
Advance Access publication June 7, 2005

We would like to thank K. Dobbin, J. Shih and R. Simon for their comments about Martin-Magniette *et al*. (2005). Their remarks relate to the design of microarray experiments and notably about the use of dye-swaps. We, however, want to make it clear that our manuscript primarily focuses on the detection, quantification and correction of the gene-specific dye-bias introduced in dual-color microarray experiments.

The most important point in Martin-Magniette *et al*. (2005) is the following: by dint of studying technical errors, we will be able to identify and then to remove most of them. For the first time, in Martin-Magniette *et al*. (2005), we were able to quantify the gene-specific dye bias by calculating the Labelling Bias Index, (LBI). The LBI measured on different array types shows that this artifact seems to be very low in some cases and could be thus neglected. However, it is high in other cases and it cannot thus be neglected for data issued from these array types. This measure is of crucial importance as it allows users to evaluate the impact of this bias on their data. Moreover, we think that each platform should at least know in what class it belongs for each array types. Although this artifact is not lowered nor understood, it is simply dangerous to underestimate it. The gene-specific dye bias is not an inevitability and can be well controlled, as we point out in our paper.

Recently, in another paper, Dobbin *et al*. (2005) studied the gene-specific dye bias. Although they reached the same conclusions, some of their remarks are explained in Martin-Magniette *et al*. (2005): e.g. Dobbin *et al*. (2005) have found that ‘(the gene-specific) dye bias appears to have masked the true differential expression’. This is explained in Martin-Magniette *et al*. (2005): the variance associated to a gene is overestimated by the dye bias effect in model (2) of Martin-Magniette *et al*.

Dye-swaps constitute a simple and effective design to remove gene-specific dye bias when it is high. Nevertheless, we agree with Dobbin *et al*. (2005), that a balanced block design may be better than dye-swaps in some situations. As the former designs allow the use of more biological samples, the estimation of the biological variability will be more precise. Even if balanced block designs are statistically more efficient, the following considerations should be taken into account before choosing the experimental design:

- Even with rigorous experimental procedures some sources of variability remain (quality and yield of target purification, labelling efficiency, …). Performing dye-swaps will allow to differentiate biological from technical variability.
- It is often difficult to balance the dye for every treatment in complex designs, when samples are hardly available. For instance, such situations are encountered in sex-balanced medical studies.
- Moreover, some redundant experimental procedures (quality control of mRNA, preparation of targets for indirect dye-labelling) used in dye-swap experiments, decrease the financial cost to <2-fold the cost of a single slide hybridization, thus rendering this design much more attractive.

The experimental design must be adapted not only to the research question but also to the amount of biological material available. Finally, class prediction or differential expression, e.g. the question of interest, do not necessarily imply the same experimental design.

REFERENCES

Dobbin,K.K. *et al*. (2005) Characterizing dye bias in microarray experiments. *Bioinformatics*. 2005 May 15; **21**(10): 2430–2437.

Martin-Magniette,M.L. *et al*. (2005) Evaluation of gene-specific dye bias in CDNAS microarray experiments. *Bioinformatics*, **21**, 1995–2000. 2005 May 1; **21**(9): 1995–2000.