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

Marie-Laure Martin-Magniette1,2,*, Julie Aubert2, Eric Cabannes3 and Jean-Jacques Daudin2
1URGV UMR INRA 1165—CNRS 8114—UEVE, 2 rue Gaston Crémieux, CP 5708, 91057 Evry Cedex, France,
2UMR INAPG/ENGREF/INRA MIA 518, 16 rue C. Bernard, 75231 Paris Cedex 05, France and
3Laboratoire 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 \( \frac{R}{G} \) 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 variability 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 \( \log_2(\frac{R_1}{G_1}) \) was wrongly calculated in place of \( \log_2(\frac{R_1}{G_1}/\frac{R_2}{G_2}) \), where \( R_i \) and \( G_i \) denote respectively the red and green intensity on the array \( i \). With the mean log-ratio \( \log_2(\frac{R_1G_2}{G_1R_2}) \), 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 \( \text{LBI} \) 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 the same 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.
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. 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 $Y_{ijk}$ be the logarithm base 2 of the measurement for array i, dye j, RNA sample k and gene g. We consider the following model:

$$Y_{ijk} = \mu + A_i + D_j + V_k + (VG)_{ik} + (DG)_{jk} + E_{ijk},$$ (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, $(VG)_{ik}$ is the g-th gene effect, and $(DG)_{jk}$ and $(VG)_{ik}$ are the corresponding interaction terms. The terms $E_{ijk}$ 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

$$Z_{ig} = V_1 - V_2 + (-1)^g(D_1 - D_2) + (VG)_{1k} - (VG)_{2k} + (DG)_{1k} - (DG)_{2k} + E_{1g},$$

where the errors $E_{ig}$ 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 order 1 dependence of the gene-specific dye bias. For the latter purpose, it suffices to test the null hypothesis $|\delta_g| = \cdot \cdot \cdot = \delta_G = 0$ against the alternative hypothesis $|\delta_g| \neq 0$. The associated test statistic can be viewed as a global index to evaluate the gene-specific dye bias.

When at least two dye-swaps are available ($p \geq 2$), we can also estimate the variance of $\delta_g$, say $\sigma^2_g$, by the empirical estimator defined by

$$\hat{\sigma}^2_g = \frac{1}{2p-2} \sum_{i=1}^{2p} (Z_{ig} - \delta_g - (-1)^g \hat{\beta}_g)^2.$$ (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 $|\beta_1| = \cdot \cdot \cdot = \beta_G = 0$ against the alternative hypothesis $|\beta_g| \neq 0$. The associated test statistic can be viewed as a global index to evaluate the gene-specific dye bias.

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:

$$Z_{ig} = (-1)^g \beta_g + E_{ig}.$$ (3)

Straightforward calculations show that $\hat{\sigma}^2_g$ is a biased estimator of $\sigma^2_g$ if $\beta_g$ differs from 0. To be precise, the bias equals $2p \beta_g^2/(2p-1)$. Therefore assuming wrongly that the $\beta_g$ are null, leads to overestimating the variance $\sigma^2_g$. Hence the power of the test for detecting a difference of expression will be lower when $\sigma^2_g$ is used in place of $\hat{\sigma}^2_g$: 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 ($\delta_g$), the gene-specific dye bias ($\beta_g$) and the variance ($\sigma^2_g$). Only two parameters per gene can be estimated. Since the major interest is the differential analysis, the parameter $\beta_g$ is usually supposed to be null. In the following section, we propose a method to assess this assumption.
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-1α 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/aminohthy.htm and http://www.microarrays.org/pdfs/aminohth.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 substate with NFU 44-571 fertilizer, BAAN SA, Vulin, 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) Aplant/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 ($\text{RegSS} = \sum_{g=1}^{G} \hat{\beta}_g^2$) and the Residuals sum of squares ($\text{RSS} = \sum_{g=1}^{G} \hat{\sigma}_g^2$). 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 genespecific 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).
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 $\beta_{g}$ 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 the phenomenon. According to the result of the differential analysis. Moreover it could help to explain the nature of the artifact. 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 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 in order to reveal its specific dye bias. This confirms that presence of gene-specific dye bias can increase the false negative rate and so decrease 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

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild type time 1</td>
<td>Control time 2</td>
<td>0.807</td>
</tr>
<tr>
<td>Wild type time 1</td>
<td>SDF time 3</td>
<td>0.762</td>
</tr>
<tr>
<td>Wild type time 1</td>
<td>Wild type time 4</td>
<td>0.745</td>
</tr>
<tr>
<td>Wild type time 1</td>
<td>Control time 5</td>
<td>0.672</td>
</tr>
<tr>
<td>Wild type time 1</td>
<td>time 6</td>
<td>0.646</td>
</tr>
<tr>
<td>Control time 2</td>
<td>SDF time 3</td>
<td>0.718</td>
</tr>
<tr>
<td>Control time 2</td>
<td>Wild type time 4</td>
<td>0.724</td>
</tr>
<tr>
<td>Control time 2</td>
<td>Control time 5</td>
<td>0.673</td>
</tr>
<tr>
<td>Control time 2</td>
<td>SDF time 6</td>
<td>0.588</td>
</tr>
<tr>
<td>SDF time 3</td>
<td>Wild type time 4</td>
<td>0.776</td>
</tr>
<tr>
<td>SDF time 3</td>
<td>Control time 5</td>
<td>0.707</td>
</tr>
<tr>
<td>SDF time 3</td>
<td>SDF time 6</td>
<td>0.657</td>
</tr>
<tr>
<td>Wild type time 4</td>
<td>Control time 5</td>
<td>0.763</td>
</tr>
<tr>
<td>Wild type time 4</td>
<td>SDF time 6</td>
<td>0.655</td>
</tr>
<tr>
<td>Control time 5</td>
<td>SDF time 6</td>
<td>0.533</td>
</tr>
</tbody>
</table>

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.

Evaluation of the gene-specific dye bias
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 $(V DG)$. 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

$$Z_{ig} = Y_{i12g} - Y_{i01g} = D_2 - D_1 + V_i - V_0 + (VG)_{i0g} - (VG)_{i1g} + (DG)_{2g} - (DG)_{1g} + (VDG)_{12g} - (VDG)_{01g} + \tilde{E}_{ig}.$$  

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

$$Z'_{ig} = (V G)_{ig} - (V G)_{0g} + (DG)_{12g} - (DG)_{10g} + (VDG)_{12g} - (VDG)_{01g} + F_{ig}.$$  

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

$$Z'_{ig} - Z'_{2g} = \delta_g + (VDG)_{12g} - (VDG)_{12g} + \tilde{E}_{ig},$$

where the errors $\tilde{E}_{ig}$ 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


