dyebias: a Bioconductor package to correct gene-specific dye bias using the GASSCO method

Philip Lijnzaad[1,2,*], Thanasis Margaritis[2,*], Dik van Leenen[2], Patrick Kemmeren[2] and Frank Holstege[2]

1: Netherlands Bioinformatics Center (NBIC), P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

2: Department of Physiological Chemistry, University Medical Center Utrecht, Universiteitsweg, Utrecht, The Netherlands

*: These authors contributed equally to this work

ABSTRACT

Summary: Gene-specific dye bias (GSDB) is an artifact of two-colour microarrays that has hampered the field for many years. It manifests itself as genes showing a strong influence of the dye orientation. The artifact reduces statistical power, or in the worst case, invalidates conclusions drawn from microarray studies. We recently introduced the GASSCO method (Margaritis *et al.*, Mol. Sys. Biol. **5**:266) that can detect and correct gene specific dye bias in a robust, general and efficient manner.

Availability: The `dyebias` package implementing the GASSCO method is part of Bioconductor. A reference manual and worked examples are included.

Contact: p.lijnzaad@umcutrecht.nl

1 INTRODUCTION

Two-colour microarrays are an important tool in the study of many aspects of transcription. Samples obtained through RNA extraction or Chromatin Immunoprecipitaton (ChIP) enrichment are labeled with fluorescent dye (*e.g.,* Cy5), and assayed on microarray together with the appropriate reference material labeled differently (*e.g.* Cy3). After scanning and quantification the data are normalized to correct both for imbalance between the two channels as well as for intensity-dependent dye bias.

A source of bias not addressed by the usual normalization procedures such as LOESS (Yang *et al.*, 2002) is gene specific dye bias (GSDB). This is a gene-, or rather probe-specific dye effect readily visible in dye-swapped hybridizations. Frequently such pairs of hybridizations show genes that are strongly influenced by the dye orientation; in the worst case, this bias overshadows real, biological changes.

2 DESCRIPTION

In a recent paper (Margaritis *et al.*, 2009), we have shown that GSDB is a term adding to the unbiased M-value (that is, $\log_2$[Cy5/Cy3]), and depending on both the individual hybridization and the probe:

$$M_{ij}^{*} = M_{ij} + GSDB_{ij} = M_{ij} + iGSDB_i \cdot F_j \qquad [1]$$

That is, the biased M^* of gene i in hybridization j is the sum of the unbiased M_{ij} and the dye bias term $GSDB_{ij}$. The latter is the product of the so-called intrinsic gene specific dye bias (iGSDB) of gene i and the slide bias F of hybridization j. The iGSDB depends on the probe sequence and can even predicted from it with a reasonable degree of accuracy. The slide bias depends largely on the labeling percentage (Margaritis *et al.* 2009). The hybridization-dependence implies that using dye swaps may not fully correct the artifact, as the two hybridizations in such a pair may have different slide biases. Each hybridization is corrected separately, making the method very flexible.

The iGSDBs of a set of probes is estimated from a number of self-self hybridizations, or from a number of dye-swapped hybridizations. If samples and conditions are sufficiently homogeneous, one set of iGSDB estimates can be reused for many different hybridizations. Alternatively, the iGSDBs can be re-estimated for each set of hybridizations. The slide bias is estimated from the total dye bias of the most strongly biased probes relative to the their iGSDB.

3 SOFTWARE

The `dyebias` package is part of Bioconductor (Gentleman *et al*., 2004) since Bioconductor release 2.4. It contains functions to estimate the iGSDBs to correct a set of hybridizations; and to plot the results of the correction in order to judge the correction process. The inputs and outputs use the `marray` package (Yang *et al.*, 2007). If the dye swaps in the hybridizations used to estimate the iGSDBs are unbalanced, the `LIMMA` package (Smyth, 2005) is used for estimating the iGSDBs, rather than the simple (and much faster) averaging of the data. The `dyebias` package comes with full documentation and working examples, including an extensive 'vignette' showing, amongst others, how data deposited in GEO (Barrett *et al.*, 2007) can be dye bias corrected *post hoc*.

The steps taken to do the dye bias correction are as follows: (1) normalized data is loaded; (2) the iGSDBs are either estimated from this data set, or loaded if determined previously; (3) probes deemed good estimators of the slide bias are marked. For example, probes targeting genes known to have high biological variability are excluded. In addition, genes for which the average expression is too low or too high are usually excluded. From the remaining probes, (4) those having an iGSDB below the 5^{th} percentile (greenest) or above the 95^{th} percentile (reddest) are used to estimate the slide bias; (5) the total dyebias is calculated and subtracted from the measured M^*-value.

The assumption behind the GASSCO method is that the total dye bias of a probe is the product of its iGSDB and the slide bias. Imagine a set of self-self hybridizations with linearly increasing slide bias. If one would plot the M^*-value for each gene, the result

would be a 'fan' of lines that diverge with increasing slide bias. The slope of the line for each probe would be proportional to its iGSDB. Since the number of lines is too overwhelming to give a meaningful plot, we instead bin the probes by their iGSDBs, and plot the median per bin. This is depicted in Figs. 1a and b for 70 hybridizations from a classification study of head and neck squamous cell carcinomas (Roepman *et al.*, 2005). Fig. 1a is the uncorrected data showing the predicted 'fan', which disappears after correction (Fig. 1b), demonstrating that the procedure is able to eliminate the dye bias.

To see how the correction performed for an individual hybridization, another routine is provided. It produces the familiar MA-plot (or alternatively, RG-plot), but with additional colouring. To highlight the changes, genes/probes with the 5% reddest and greenest iGSDB are coloured accordingly. The correction moves the genes towards the *M=0* axis, which can be seen clearly in Figs. 1c and d (the green points are now mostly covered by red points). The total variance of *M* goes down, sometimes dramatically (here, around 45%).

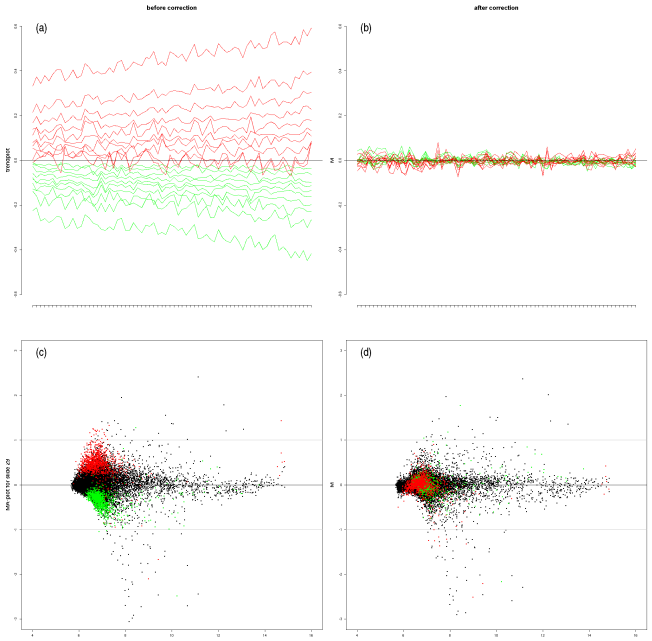

Fig. 1. (a,b): Median M-value per iGSDB-bin before (a) and after correction, of 70 hybridizations sorted by slide bias.
(c,d): M-values of an individual hybridization (with average slide bias) before (c) and after (d) correction using GASSCO. All data from (Roepman et al., 2005).

4 CONCLUDING REMARKS

A few caveats are worth pointing out. Data has to be normalized conventionally (*e.g.*, using LOESS) prior to the dye bias correction. The set of hybridizations used to estimate the iGSDBs should be large enough (say, at least 8), and preferably balanced. Data should be presented as Cy5 *vs.* Cy3, not as sample *vs.* reference. GSDB depends on the total amount of label present, and therefore also on transcript abundancies. Consequently, if these are likely to differ greatly between studies (e.g., when studying different tissue types or cell lines), separate iGSDB estimates are needed for each.

The `dyebias` package has been in production in our laboratory for nearly two years, and is very robust. As described in Margaritis *et al.* (2009), it is very general and able to correct many different kinds of data, including cDNA data and ChIP-on-chip data. The procedure is very efficient; for the above data set (70 slides, 25392 spots) it took around 20 m on a laptop with an Intel Core™2 Duo CPU (model T7200) running a 32-bits Linux 2.6.24 kernel at 2 GHz with 2 GB of memory.

ACKNOWLEDGEMENTS

We would like to thank Marian Groot Koerkamp for helpful discussions.

Funding: This work was partly (PK) funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) grant 863.07.007.

REFERENCES

Barret, T. *et al*. (2007) NCBI GEO: mining tens of millions of expression profiles -- database and tools update. *Nucleic Acids Research* **35**, D760-D765.

Gentleman, R.C. *et al*. (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80.

Margaritis *et al.* (2009) Adaptable gene-specific dye bias correction for two-channel DNA microarrays. *Mol. Sys. Biol.* **5**, 266.

Roepman, P. et al. (2005) An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. Nat. Gen. **37**, 182-186.

Smyth, G. K. (2005) Limma: linear models for microarray data. In Gentleman, R., *et al.* (eds.), *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, pp. 397-420.

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

Yang Y.H. *et al*. (2007) marray: Exploratory analysis for two-color spotted microarray data. http://www.maths.usyd.edu.au/u/jeany