

Gene expression

# A modified LOESS normalization applied to microRNA arrays: a comparative evaluation

Davide Risso<sup>1</sup>, Maria Sofia Massa<sup>1</sup>, Monica Chiogna<sup>1</sup> and Chiara Romualdi<sup>2,\*</sup><sup>1</sup>Department of Statistical Sciences, University of Padova, via C. Battisti 241 and<sup>2</sup>Department of Biology, University of Padova, via U. Bassi 58/B, 35121 Padova, Italy

Received on November 10, 2008; revised on June 18, 2009; accepted on July 15, 2009

Advance Access publication July 23, 2009

Associate Editor: Ivo Hofacker

## ABSTRACT

**Motivation:** Microarray normalization is a fundamental step in removing systematic bias and noise variability caused by technical and experimental artefacts. Several approaches, suitable for large-scale genome arrays, have been proposed and shown to be effective in the reduction of systematic errors. Most of these methodologies are based on specific assumptions that are reasonable for whole-genome arrays, but possibly unsuitable for small microRNA (miRNA) platforms. In this work, we propose a novel normalization (loessM), and we investigate, through simulated and real datasets, the influence that normalizations for two-colour miRNA arrays have on the identification of differentially expressed genes.

**Results:** We show that normalizations usually applied to large-scale arrays, in several cases, modify the actual structure of miRNA data, leading to large portions of false positives and false negatives. Nevertheless, loessM is able to outperform other techniques in most experimental scenarios. Moreover, when usual assumptions on differential expression distribution are missed, channel effect has a strikingly negative influence on small arrays, bias that cannot be removed by normalizations but rather by an appropriate experimental design. We find that the combination of loessM with eCADS, an experimental design based on biological replicates dye-swap recently proposed for channel-effect reduction, gives better results in most of the experimental conditions in terms of specificity/sensitivity both on simulated and real data.

**Availability:** LoessM R function is freely available at <http://gef.ucrib.unipd.it/papers/miRNA-simulation/>

**Contact:** chiara.romualdi@unipd.it

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

MicroRNAs (miRNAs) are small, non-coding RNAs that regulate the expression of target mRNAs. Although thousands of miRNAs have been identified, only few have been functionally linked to specific biological pathways. Microarray-based expression analysis is an ideal strategy for identifying candidate miRNAs correlating with biological pathways and for generating molecular signatures of disease states. For systematic investigation of miRNA expression, oligonucleotide-based microarrays for miRNAs have been recently

developed (Liu *et al.*, 2004) and several commercial platforms are now available.

Data normalization has been revealed as a crucial step to reduce possible systematic errors that, if ignored, will bias final results. Several approaches have shown to be effective in the reduction of systematic errors, for a comprehensive review see Quackenbush (2001), Bilban *et al.* (2002), Smyth and Speed (2003) and Leung and Cavalieri (2003). Many methodologies currently used for mRNA microarrays are directly applicable to miRNA arrays, in particular, loess and global normalizations (but sometimes preprocessing is missed) seem to be preferred among others. However, there are specific characteristics of miRNA data that should be considered when choosing the appropriate method of analysis.

Most of these methods are based on two assumptions: (i) only a small portion of spots is differentially expressed, and (ii) differentially expressed spots are homogeneously distributed between over- and under-expressed. These assumptions, reasonable for large-scale genome arrays, could fail for miRNA platforms that, like custom arrays, are printed with a relatively small number of selected probes. Experiments on two samples with most of miRNA differentially expressed, predominately in one direction, are not unusual. In such situations, it is still difficult to decide which method uniformly outperforms in different experimental conditions.

Recently, Rao *et al.* (2008) compare the performance of several normalizations on miRNA microarrays based on single channel technology, showing a better performance of quantile normalization. They use technical replicates in order to evaluate the reduction of the difference in miRNA expression values between duplicate samples. However, normalization techniques could have a profound effect on subsequent analyses, such as differential expression testing, as demonstrated by Hoffmann *et al.* (2002), for single channel and by Chiogna *et al.* (2009), for dual channel mRNA microarray technology. Based on this consideration, Pradervand *et al.* (2009) compare specificity and sensitivity of miRNA normalizations for single channel technology using a set of true positive genes, confirming the goodness of quantile and proposing a glog transformation based on invariant set of miRNAs. Similarly, the work of Hua *et al.* (2008) is focused on the comparison of preprocessing techniques for two-channel technology taking into consideration the identification of differentially expressed genes (DEGs). They use RT-PCR expression quantification of eight genes to compare normalizations: the higher the correlation with RT-PCR values, the better the normalization. They demonstrate that print-tip

\*To whom correspondence should be addressed.

loess (P-loess) performs slightly better than the others. However, in a similar study, Sarkar *et al.* (2009) do not find significant differences between P-loess and other normalizations, proposing an alternative transformation based on  $\log$ .

In this work, we propose a novel normalization procedure, called loessM, based on the loess algorithm. LoessM scales expression data on the global median expression rather than on zero, as usually done for whole-genome arrays. This modification relaxes the assumption of symmetry among up- and down-regulated genes. Moreover, we compare and evaluate the impact that different normalizations for two-channel microarray have on the identification of DEGs. We find, on simulated and real data, that loessM, in case of absence of channel effect (different efficiency in the fluorophores incorporation in the labelling protocol), shows better performances, especially in the most extreme experimental cases.

Moreover, our analysis underlines the profound negative influence that channel-effect bias has on miRNA arrays. When usual assumptions on differential expression distribution fail, channel bias can be removed not by normalizations, rather by a specific experimental design. Here, we show that the combination of loessM and eCADS—a recently proposed biological replicates dye-swap design (Dabney and Storey, 2007)—is able to maintain the correct structure of expression data, showing better performance among most of the compared preprocessing procedures.

## 2 METHODS

### 2.1 Simulation models

Hereafter, we will note the  $\log$  transformation of Cy3/Cy5 as M and the  $\log$  transformation of the squared root of Cy3\*Cy5 as A (as used in the MA-plot).

**2.1.1 Hierarchical models** We considered the two mixture models proposed by Kendziorski *et al.* (2003) (see Section 1 in Supplementary Material). In the first model, Gamma–Gamma (GG), the intensities for the replicates in both conditions (Cy3 and Cy5) are assumed to be independently generated from Gamma distributions with a constant shape parameter  $\alpha$  and gene-specific random scales  $\lambda_g$ , assumed to have a Gamma distribution with shape hyperparameter  $\alpha_0$  and scale hyperparameter  $\nu$ . In the second model, lognormal-normal (LNN), the  $\log$  intensities are assumed to be normally distributed, with constant variance  $\sigma^2$  and gene-specific random means  $\mu_g$ , that are themselves normally distributed with hyperparameters  $\mu_0$  and  $\tau$ .

Hierarchical models GG and LNN simulate datasets without intensity-dependent systematic bias, a common situation in microarray data. Thus, we decided to introduce a systematic bias effect obtained through the addition (to the  $\log$ -ratio simulated by GG and LNN) of a component inversely proportional to A, opportunely scaled.

**2.1.2 Balagurunathan's model** Balagurunathan *et al.* (2002) procedure (BAM, hereafter) models two distinct fluorescent intensities, one for each channel, based on the same true gene-specific signal.

The true intensity signal ( $I_g$ ) for each gene  $g$  is assumed to be generated from an exponential distribution with a constant parameter  $\lambda$ . The fluorescence of each channel is then generated from a Normal density with gene-specific mean  $I_g$  and gene-channel-specific SD  $\alpha_{g,k} I_g$ , where  $\alpha_{g,k}$  is the predetermined coefficient of variation of gene  $g$  in channel  $k$ . The subsets of genes assumed to be differentially expressed are then transformed by adding to the  $\log$ -intensity ratio a random quantity  $b_g$ , generated from a Beta distribution with constant parameters.

The authors introduced a flexible family of functions that models most of the possible channel biases and in our work, we considered two sets of parameters, corresponding to an experiment with no channel effect and to an

experiment with strong unbalanced channel intensity, following suggestions in Balagurunathan *et al.* (2002).

**2.1.3 Simulation Plan** For each model and set of parameters, we simulated 10 matrices with 1000 gene expression levels on 15 experiments, separately for the Cy5 and Cy3 channels. Thus, each simulated matrix consisted of 1000×30 values. Simulated matrices were generated in order to take in consideration different types of expression scenarios: (i) we increased the proportion of DEGs; (ii) we increased the asymmetry of the expression distribution; and (iii) we combined asymmetry and large amount of DEGs. Then, we considered the following situations (the first number is referred to the percentage of down-regulated and the second to the percentage up-regulated genes): 0–6 %, 0–20 %, 0–40 %, 10–30%, 10–40% and 20–20%.

When channel effect was introduced in BAM model, matrices were generated in order to consider two different scenarios: (i) channel effect in the same direction of differential expression, e.g. Cy3 more efficient in the labelling step jointly with 20% up-regulated and 0% down-regulated genes, and (ii) channel effect in the opposite direction of differential expression, e.g. Cy5 more efficient in the labelling step jointly with 20% up-regulated and 0% down-regulated genes. In the first condition, channel effect overdraws differential expression, while in the second one it hides gene deregulation. However, changing the direction of channel effect is equivalent to fix channel effect and change differential expression direction. Then, in case of BAM model, we considered Cy3 as more efficient and simulated the above differential situations and their opposite.

At the end of the preprocessing, for each simulation we had 10 different matrices, on each of which SAM analysis was performed (Tusher *et al.*, 2001). Receiver operating characteristic (ROC) curves were obtained calculating sensitivity and specificity for top ranking genes from 10 to 600 steps by 20. Then, average sensitivity and specificity were calculated.

### 2.2 Real data

Normalization performances on real datasets have been compared using: (i) correlation with RT-PCR, spike-in and control data, whenever available and, (ii) specificity and sensitivity of the statistical test using true positives features.

We used the following datasets: (i) three large-scale cDNA datasets (Baird *et al.*, 2005; De Pitta *et al.*, 2005, 2006), datasets A, B and C, appropriately modified in order to know a priori the true positives, (ii) one large-scale small oligos dataset (dataset D) with true positives obtained through RT-PCR (Patterson *et al.*, 2006), (iii) one miRNA (dataset E) (Sarkar *et al.*, 2009) with RT-PCR validation and spike-in controls and (iv) one miRNA (dataset F) (Wang *et al.*, 2007), where true positives were unknown.

Except for dataset E, kindly provided by the authors, all the remaining datasets are publicly available at the GEO database. For more details on these datasets, see Section 2 in Supplementary Material.

**2.2.1 Evaluation criteria for datasets A, B and C** Modification of real datasets A, B and C is essentially based on a selection of genes from a large-scale array, so as to reproduce miRNA arrays characteristics, maintaining the complex expression data structure typical of microarrays. The procedure for data modification is summarized in the following steps: (i) normalize each dataset (A, B and C) with loess procedure; (ii) mark the features found to be differentially expressed after SAM test (False Discovery Rate, FDR = 0) as deregulated; (iii) randomly sample 1000 features from the whole dataset in order to have the desired percentage of up- and down-regulated features (see Section 2.1.3 for details), and consider their raw expression values for subsequent analyses; (iv) repeat Step (iii) 10 times. On these matrices, we proceed as described for simulated data (See Section 2.1.3).

The use of loess normalization previous to SAM analysis (in Step i) should not advantage loess-type normalization performances. It is used just to avoid the selection of a higher number of false positives, that surely would be

identified if raw data were considered; the subsequent random selection of genes from the whole dataset should balance the possible advantage.

As in real datasets channel-effect direction is fixed and unknown, we use the following conditions : 0–6%, 0–20%, 0–40%, 10–30%, 10–40% and 20–20% and vice versa.

**2.2.2 Evaluation criteria for dataset D** Dataset D provides a list of true positive genes obtained by means of TaqMan RT-PCR. Agilent platform has been selected because of the similarity between oligonucleotides and miRNA probes length. This similarity should give expression values more reliable than in the previous cases. RT-PCR defined our true positives: 481 true over-expressed and 350 true under-expressed. Thus, we simulated asymmetrical differential distribution selecting separately true over- or under-expressed genes and then randomly selecting from the remaining expression values in order to reach a total number of 1000 probes. Thus, we had three different scenarios: (i) asymmetrical over-expression (48%), (ii) asymmetrical under-expression (35%) and (iii) symmetrical differential expression distribution. Finally, we used dye-swap experiments to evaluate eCADS design. ROC curves and correlation between expression values obtained through microarrays and RT-PCR (Patterson *et al.*, 2006) were used to compare normalizations performances.

**2.2.3 Evaluation criteria for dataset E** Dataset E is characterized by an experimental plan specifically designed for quality control check. The authors constructed a synthetic universal miRNA reference (pool of 480 chemically synthesized miRNAs corresponding to all known human miRNA probes present on the microarray), optimized to provide relatively uniform level of intensity in the green channel. The expected real median Cy3 intensity is 12 (log transformation) [see Supplementary file 1 of Sarkar *et al.* (2009) for details]. Then, Cy3 intensity distribution across experiments, correlation coefficients with spike-in controls and proportion of true positive genes (by RT-PCR) identified were used to compare normalization performances.

**2.2.4 Evaluation criteria for dataset F** Dataset F (Wang *et al.*, 2007) does not have external gene validation, thus we decided to use two different strategies to compare normalizations: (i) intensity distribution of negative controls, empty spots, that are supposed to have an expression distribution homogeneously distributed around zero and characterized by small variability and (ii) literature evidences on the list of DEGs identified after normalizations.

Wang *et al.* (2007) performed filtering and global normalization on raw data; then, they identified as DEGs those genes resulted to be significant with both SAM and *t*-test. We considered only genes differentially expressed across all samples, without group distinction [as reported in supplementary table in Wang *et al.* (2007)]. Then, in order to compare our results with those of Wang *et al.*, we applied their filtering procedure before the preprocessing technique and we adopted both SAM and *t*-test for the gene list comparison.

### 2.3 Biological replicates dye-swap: eCADS

Dabney and Storey (2007) developed a procedure, named eCADS, that accounts for dye bias through the combination of experimental design and statistical test. Their experimental design requires only a single array per sample pair instead of technical replicates. In this work, we used only eCADS experimental design without taking into account the statistical test, fANOVA, proposed in Dabney and Storey (2007).

For BAM model, we evaluated eCADS by applying a channel effect in one direction (i.e. Cy5 more efficient) for the first eight experiments (columns), and in the other direction (i.e. Cy3 more efficient) for the other seven experiments.

De Pitta *et al.* (2005, 2006) and Patterson *et al.* (2006) applied a technical replicates dye-swap design in their study. Then, for datasets A, B and D, we were able to evaluate eCADS design: half of the experiments were chosen

with reference marked with Cy5 and sample with Cy3 and half with inverted labelling.

Unfortunately, datasets C, E and F did not perform dye-swap replicates. In these cases, in the DEGs list comparison, we assumed the absence of a channel effect.

### 2.4 A new normalization technique

Loess normalization and its variants (Futschik and Crompton, 2004, 2005; Yang *et al.*, 2002) rely on the use of a non-linear regression technique (the widely used loess, LOcally WEighted Scatterplot Smoothing) based on robust local regression of the log ratios Cy3/Cy5 on overall spot intensity Cy3\*Cy5 (the loess smoother for the so called MA-plots). If (R,G) are the fluorescence intensity pairs for each gene on each array (where R = red for Cy5 and G = green for Cy3), loess normalization assumes that the red and green intensities are related by a factor dependent from A:

$$\log_2(R/G) \rightarrow \log_2(R/G) - c(A) \rightarrow \log_2(R/k(A)G),$$

where  $c(A)$  is the loess fit to the MA-plot. Then, loess and its modification assume that (i) the majority of genes are equally expressed and (ii) the distribution of the log-intensity ratio of deregulated genes is roughly symmetric about zero (Yang *et al.*, 2002). If assumptions (i) and (ii) do not hold, e.g. most of the features are over-expressed, the factor  $c(A)$  erroneously scales expression data towards zero, increasing the false positive rate among under-expressed and the false negatives among over-expressed.

Here, we propose a modification of the factor  $c(A)$ ,  $c^*(A)$ , which scales expression values towards the general median expression values. That is:

$$c^*(A) = c(A) - \text{Med}(M),$$

$$\log_2(R/G) \rightarrow \log_2(R/G) - c^*(A),$$

where  $\text{Med}(M)$  is the median of  $M$  on the microarray experiment. We call this modification loessM.

Then, we compare sensitivity and specificity performance of SAM test after global (Quackenbush, 2001), loess and P-loess, (applied only on the real datasets) (Yang *et al.*, 2002), glog (Huber *et al.*, 2002; Rocke and Durbin, 2003), q-spline (Workman *et al.*, 2002), generalized procrustes analysis (GPA); (Xiong *et al.*, 2008), loessM and the combination of loessM and GPA. Since OLIN and OSLIN normalizations (Futschik and Crompton, 2005) performances have been demonstrated to be highly similar to that of loess (Chiogna *et al.*, 2009), we decide to exclude them from the analysis.

The glog normalization is known to be affected by a large number of DEGs all in the same direction (Huber *et al.*, 2003). Huber and colleagues (2003), in fact, introduced a parameter ( $q_{lts}$ ) that, excluding outliers from parameter estimation, should make glog a more robust procedure. In this study, we consider the most robust case ( $q_{lts} = 0.5$ ), as it gave the best results (data not shown).

Statistical analyses have been performed with R using packages Biobase, limma, marray, affy, vsn, vegan and samr. For the GPA, we used the original script kindly provided by the authors.

## 3 RESULTS AND DISCUSSION

In this study, we evaluate the impact that normalizations have on the identification of DEGs in miRNA arrays with two-channel technology. The widely used techniques for genome-wide arrays are typically based on at least two main assumptions: (1) small proportion of DEGs and (2) symmetrical distribution of DEGs between over- and under-expressed. In small arrays these two assumptions can be easily missed. Here, we present a novel intra-array normalization, loessM (free of assumptions 1 and 2) that combined with GPA—an assumption-free inter-array normalization

(Xiong *et al.*, 2008)—shows better results when these assumptions are relaxed. However, in these cases, channel bias cannot be removed by normalization, but rather by an appropriate experimental design. To this aim, we show how assumption-free normalizations with eCADs design (Dabney and Storey, 2007) allow an effective channel-bias reduction. In the following, we report the performance of the selected normalizations among which raw data (no normalization) are included. In simulated data, raw data performance is often comparable with those derived by the better techniques, however, this result is not unexpected. Simulated data have been generated in order to have channel bias highly exceeding intensity-dependent bias (see Supplementary Fig. S11). Thus, in case that assumptions 1 and 2 are relaxed, even raw data are able to correctly identify most of DEGs. Confirming our consideration, in real datasets, raw data perform well only in cases where channel effect is in the same direction of differential expression. However, raw data are not the point of our study, real data are usually noisier than simulated ones; the point is that, in these cases, the usual normalizations strongly alter data structure hiding differential expression.

*Simulation results—data without channel bias:* normalization comparison for GG, LNN and BAM model without channel effect inclusion are quite similar and reported in Supplementary Figures S1 and S2. Normalizations seem to perform similarly in case of (i) small portion of DEGs with totally asymmetric distribution (0–6%, Supplementary Figs S1D, J and S2A) and (ii) large portion of DEGs, symmetrically distributed between up- and down-regulated (20–20%, Supplementary Figs S1E, K and S2F). However, with the increasing degree of asymmetry, differences among normalizations become stronger (Figs S1B, H and S2B–E): q-spline, global and loess perform worse than the other methodologies. In the most extreme asymmetrical cases (Supplementary Figs S1F and S2B–E), loessM, GPA and loessM with GPA show the best performance. In LNN model (Fig. S1L), GPA and glog transformation perform at best. This result is not unexpected, since LNN model is basically the model behind the glog normalization (see Section 3 in Supplementary Material).

*Simulation results—data with channel bias:* dye bias is a systematic difference between the incorporation rates of the fluorescent dyes used for labelling targets. In particular, up- or down-regulated miRNAs are systematically increased or decreased according to the Cy5 or Cy3 differential incorporation. When usual assumptions on differential expression distribution are missed, channel effect cannot be removed by the normalizations and differential expression can be hidden or overdrawn. In fact, when channel effect is introduced in BAM model (Fig. 1 and Supplementary Fig. S3) normalization performances markedly change according to the direction of both effects, as expected. When differential expression is in the same direction of channel bias (differential expression amplification), loessM, GPA and loessM combined with GPA strikingly outperform other normalizations (Fig. 1 A, C and Supplementary Fig. S3); otherwise, when differential expression is in the opposite direction of channel bias (Fig. 1 E, G and Supplementary Fig. S3) (differential expression masked), loess and global increase their performances. In the case of channel effect in the same direction of differential expression, in fact, normalizations assuming differential expression distribution centred around zero, erroneously shift data towards zero, missing most of the up-regulated features and introducing several false positives

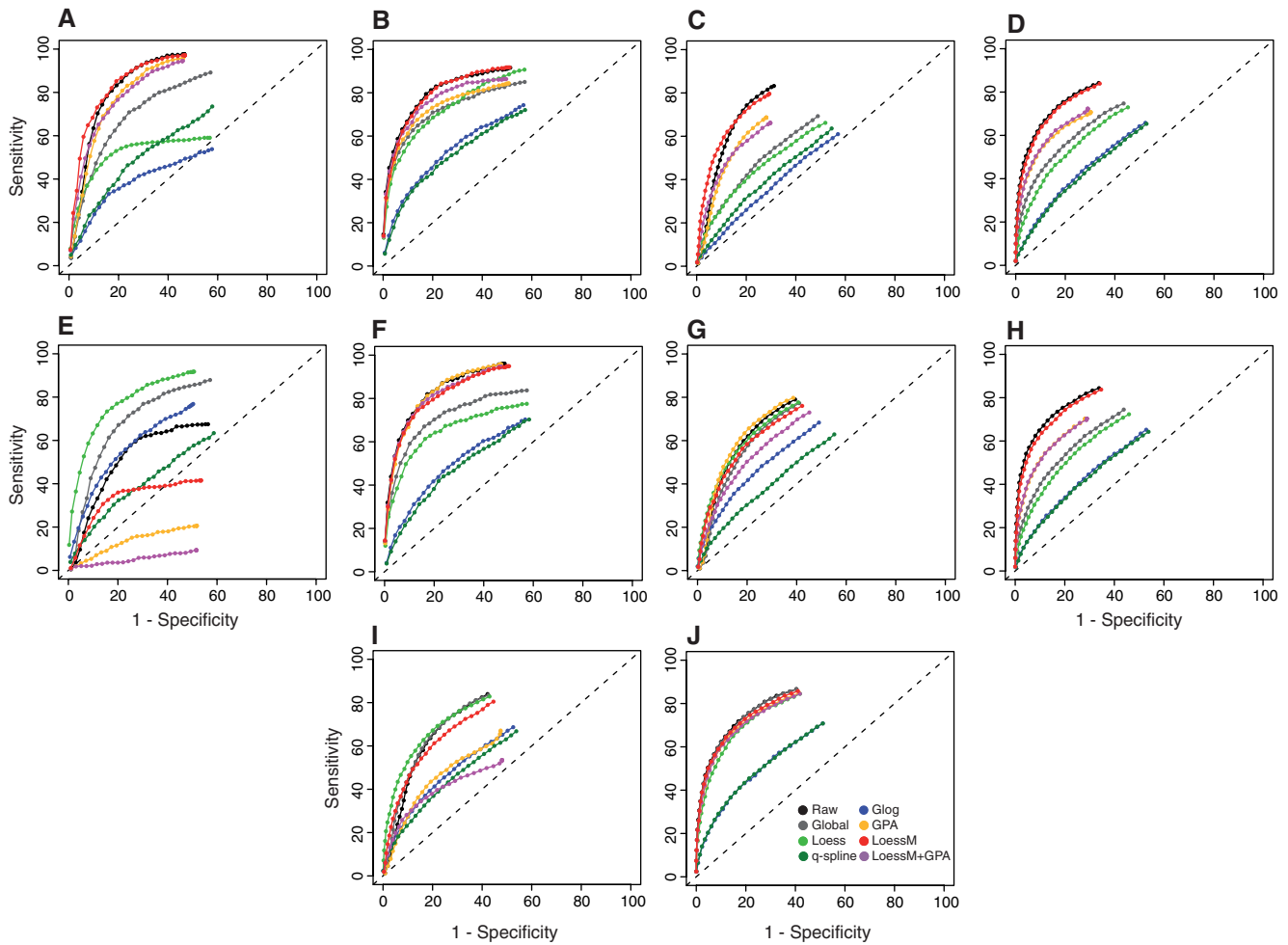
among down-regulated ones. On the other hand, if channel effect and differential expression are in the opposite directions with a small portion of deregulation, loessM and GPA erroneously capture channel-dependent effects instead of differential expression. Note that channel effect and differential expression directions are a priori unknown and cannot be distinguished by normalizations, but rather by a specific experimental design.

*Simulation results—data with channel bias and eCADs design:* Dabney and Storey (2007) develop a procedure, named eCADs, which needs only a single array per sample pair, using biological replicates, instead of technical replicates to account for dye bias. Taking advantage of their results, we simulate data following eCADs design (see Section 2 for details). Figure 1 and Supplementary Figure S3 show the results. It is immediately evident how eCADs leads to more concordant results among normalization performances in the two scenarios: channel effect and differential expression in the same and in the opposite direction. In particular, with eCADs design, loessM improves performance becoming the best normalization across most of the simulated conditions (Fig. 1B, D, H, J and Supplementary Fig. S3).

*Real data results—dataset A, B and C:* results obtained using modification of large-scale real datasets are in agreement with those obtained through simulated ones (Fig. 2 and Supplementary Figs S4–S9). The presence of channel effects in these datasets is even more evident than in simulated data, especially for dataset B. While in datasets A and B, GPA and glog outperform the other techniques, in case of channel effect and deregulation in the opposite direction (Fig. 2A and Supplementary Figs S4 and S6), loessM and loessM plus GPA are the best normalizations in the opposite scenario (Fig. 2B, Supplementary Figs S4J and S6). In dataset C, on the contrary, the difference between both conditions is much less evident (Supplementary Fig. S8), suggesting a reduced channel effect in dataset C, and loessM and loessM plus GPA always outperform the other normalizations.

Fortunately, for datasets A and B, the authors applied a classical dye-swap experimental design; thus, for these datasets, we try to evaluate the impact that eCADs design has on the reduction of channel effect. Figures 2, Supplementary Figs S5 and S7 show the results obtained through eCADs design for datasets A and B. Using eCADs design, dataset A shows more homogeneous results; loessM and GPA improve performance becoming the best normalizations in most of the experimental conditions (Fig. 2C, Supplementary Fig. S5). In particular, in case of strong differential unbalance (Supplementary Fig. S5), loessM plus GPA outperforms the other techniques, while in cases of 20–20% and 0–6% there seem to be no significant differences among normalizations. However, results for dataset B remain slightly different (Supplementary Fig. S7). This result confirms the strong channel effect that characterizes these experiments, effect that probably cannot be removed, but only partially reduced by eCADs.

*Real data results—dataset D:* dataset D, characterized by short oligo lengths, shows the same specificity and sensitivity trend (Fig. 2). When an asymmetric distribution is simulated (35% of under-expressed genes), we observe different normalization performances according to the direction of the bias (Fig. 2D, E and Supplementary Fig. S9). However, using eCADs design, GPA, loessM and the combination of both show performances superior than those of the other normalizations (Fig. 2F and Supplementary Fig. S9). This is in agreement with previous datasets results. In



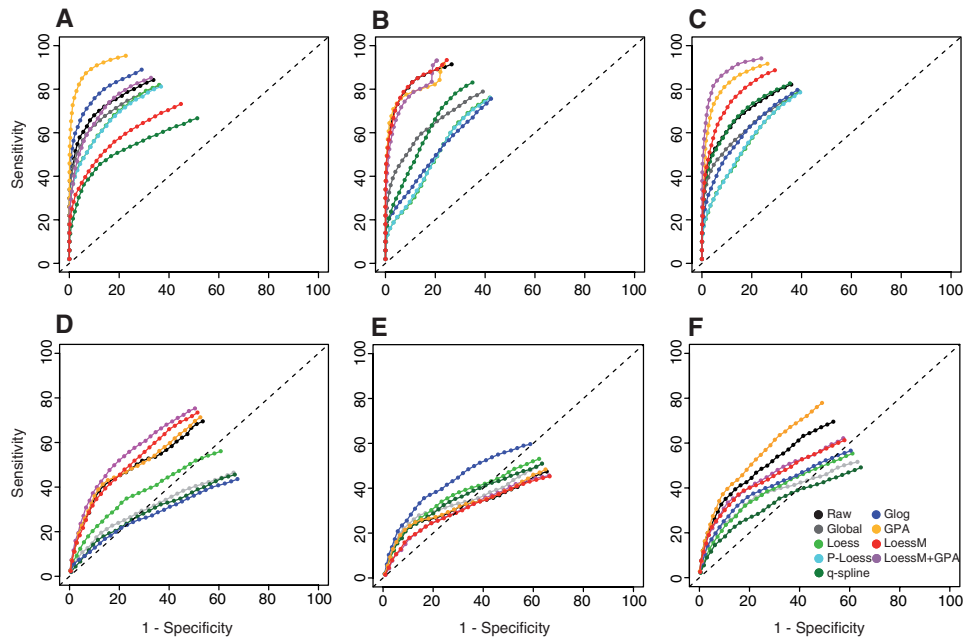
**Fig. 1.** Specificity and sensitivity trend for BAM model with channel-effect bias with and without eCADs design. Simulated matrices were generated considering channel effect and differential expression effect in the same (A–D) and in the opposite direction (E–H). Hereafter we indicate with the symbol ‘+’ when both effect are in the same direction and ‘-’ otherwise. Matrices are characterized by (numbers represent, respectively, proportion of down- and up-regulated features): 0–6% [(A) for ‘+’ (E) for ‘-’], 10–40% [(C) for ‘+’ (G) for ‘-’], 20–20% (I). With eCADs design: 0–6% [(B) for ‘+’ (F) for ‘-’], 10–40% [(D) for ‘+’ (H) for ‘-’], 20–20% (J). See Supplementary Figure S3 for the remaining 0–20%, 0–40%, 10–30% panels.

addition, we calculate the correlation between microarray and RT-PCR values (see Supplementary Table S1). The combination of loessM and GPA shows the highest correlation coefficient ( $r = 0.81$ ).

**Real data results—dataset E:** miRNA dataset E was specifically designed by the Authors to have green channel intensity to be used as quality control check. Thus, to verify the normalization performances, we compare the distribution of the median green channel intensities: the closer to the expected real median (given by the composition of the reference pool), the better the normalization. Figure 3A shows the median distribution after all the normalizations across the 29 experiments. It is evident that GPA, loessM and loessM+GPA show the closest green channel median to the expected one (black dotted line). Dataset E comprises also a series of spike-in spots, whose intensities were used to compare normalizations: values in each array are plotted against the corresponding average across all arrays. Ideally, if no normalization is required, all points fall on the diagonal. Supplementary Figure S10 shows that after

loessM+GPA normalization, we obtain regression lines closest to the diagonal.

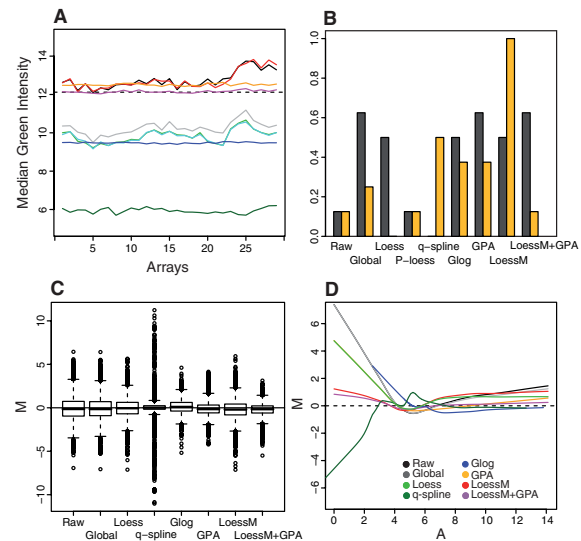
Using RT-PCR results as true positive genes, we estimate the number of true positive rate after each normalization. Figure 3B shows that loessM is able to reach the highest rate of true positives (especially down-regulated genes), followed by GPA, but the combination of both, in this case, shows a poor performance. However, it should be noted that using Cy3 as a quality check, the experimental design becomes more similar to a single channel design. Thus, for the identification of DEGs careful consideration should be paid to the reference used for GPA normalization and to the values used for groups comparison. In this particular case, we suggest to use two separate GPA references (one for each group) and to use only Cy5 values (after normalization) for groups comparison. The presence of Cy3 value as quality control has the effect of overestimating gene variance and FDR. Using this second approach, we reach better results for loessM+GPA (Supplementary Table S2).



**Fig. 2.** Specificity and sensitivity for modified real dataset A pertaining to 10–40% scenario and dataset D. Hereafter we indicate with symbol ‘+’ when channel effect and differential expression are in the same direction and ‘-’ otherwise. Dataset A [(A) for ‘+’ (B) for ‘-’ and (C) for eCADs], dataset D in the case of down-regulation [(D) for ‘+’ (E) for ‘-’ and (F) for eCADs]. For the complete panels see Supplementary Figures S4–S9.

*Real data results—dataset F:* dataset F does not have a dye-swap design, therefore, we are unable to use eCADs design. Hence, in this comparative evaluation, we assume the absence of channel effect. First of all, to evaluate the performance of all the normalizations, we select only empty spots, assuming that, in the absence of systematic bias, empty spots should show an homogeneous distribution with small variance around zero. Figure 3C and D shows the expression distribution and the loess fit of all the 4590 (90 empty spots × 51 experiments) empty spots after all the normalizations. It is worth to note that loessM, GPA, loessM+GPA, glog, qspline and loess show the thinnest boxplot (Figure 3C). However, only loessM, GPA and their combination show the best loess fit, without any systematic bias for spots with small and high intensities (Fig. 3D).

The authors, using global normalization, find 45 miRNAs differentially expressed, 24 (53%) up-regulated and 21 (46%) down-regulated. Using loessM and GPA, we identify 89 differentially expressed miRNAs [19 in common with the 45 found by Wang *et al.* (2007)], with a strong unbalance between up-regulation (2%) and down-regulation (98%). A global down-regulation of miRNAs in tumour versus normal samples has been widely observed in several studies (Bottoni *et al.*, 2007; Chang *et al.*, 2008; Lu *et al.*, 2005). One hypothesis is that several miRNAs are up-regulated during differentiation and then down-regulated during proliferation (Lu *et al.*, 2005). The comparison between loessM and the original results leads to 19 genes recognized as DEGs, two with non-concordant expression (we found *miR-145* and *miR-26a-1* under-expressed, while Wang and colleagues found them over-expressed). However, our results are in agreement with Chang *et al.* (2008), Ozen *et al.* (2008) and Lu *et al.* (2005), who found *miR-26a-1* under-expressed in several human cancers. In particular, Chang *et al.* (2008) demonstrated that oncogenic transcription factor,



**Fig. 3.** Normalization comparison for datasets E (A and B) and F (C and D). (A) Green channel median distribution after normalizations (dotted line represents the expected real median value). (B) Proportion of true positives (over-expressed: black; under-expressed: yellow) identified after normalizations. (C) Log ratio distribution of blank spots after normalizations. (D) Residual systematic bias (after normalizations) in the MA-plot of only blank spots.

Myc, activated in many human malignancies, represses a broader set of miRNAs among which *miR-26a-1*. Furthermore, *miR-145* has been validated through RT-PCR to be under-expressed in prostate cancer (Ozen *et al.*, 2008). In addition, among those identified only

using loessM, 21 (30%) have been found down-regulated by Lu *et al.* (2005), while among those obtained only by Wang *et al.* (2007) using global, 13 (32%) have been found by Lu *et al.* (2005), but only one with concordant sign.

#### 4 CONCLUSIONS

Normalization is crucial for the reduction of systematic bias that characterizes array expression data. Many preprocessing techniques currently used for mRNA microarrays are directly applicable to miRNA arrays, even if there are specific characteristics of a miRNA array that should be considered when choosing the appropriate method of analysis. Several studies have demonstrated how miRNA expression could be characterized by a large number of DEGs, often in one direction. Here, we show that, in these cases, assumption-free normalizations such as loessM (intra-array), jointly with GPA (inter-array), outperform the other normalizations in terms of sensitivity and specificity. However, in case of marked channel effects, the eCADS experimental design is necessary to reduce dye biases, and improve the loessM plus GPA performances.

*Funding:* University of Padova (grants CPDR070805 and CPDA075919); Fondazione CARIPARO under the grant ‘A computational approach to the study of skeletal muscle genomic expression in health and disease’.

*Conflict of Interest:* none declared.

#### REFERENCES

- Baird, K. *et al.* (2005) Gene expression profiling of human sarcomas: insights into sarcoma biology. *Cancer Res.*, **65**, 9226–9235.
- Balagurunathan, Y. *et al.* (2002) Simulation of cDNA microarrays via a parameterized random signal model. *J. Biomed. Opt.*, **7**, 507–523.
- Bilban, M. *et al.* (2002) Normalizing DNA microarray data. *Curr. Issues Mol. Biol.*, **4**, 57–64.
- Bottoni, A. *et al.* (2007) Identification of differentially expressed microRNAs by microarray: a possible role for microRNA genes in pituitary adenomas. *J. Cell Physiol.*, **210**, 370–377.
- Chang, T. *et al.* (2008) Widespread microRNA repression by Myc contributes to tumorigenesis. *Nat. Genet.*, **40**, 43–50.
- Chiogna, M. *et al.* (2009) A comparison on effects of normalisations in the detection of differentially expressed genes. *BMC Bioinformatics*, **10**, 61.
- Dabney, A. and Storey, J. (2007) Normalization of two-channel microarrays accounting for experimental design and intensity-dependent relationships. *Genome Biol.*, **8**, R44.
- De Pitta, C. *et al.* (2005) A leukemia-enriched cDNA microarray platform identifies new transcripts with relevance to the biology of pediatric acute lymphoblastic leukemia. *Haematologica*, **90**, 890–898.
- De Pitta, C. *et al.* (2006) Gene expression profiling identifies potential relevant genes in alveolar rhabdomyosarcoma pathogenesis and discriminates PAX3-FKHR positive and negative tumors. *Int. J. Cancer*, **118**, 2772–2781.
- Futschik, M. and Crompton, T. (2004) Model selection and efficiency testing for normalization of cDNA microarray data. *Genome Biol.*, **5**, R60.
- Futschik, M. E. and Crompton, T. (2005) OLIN: optimized normalization, visualization and quality testing of two-channel microarray data. *Bioinformatics*, **21**, 1724–1726.
- Hoffmann, R. *et al.* (2002) Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol.*, **3**, research0033.1–research0033.11.
- Hua, Y. *et al.* (2008) Comparison of normalization methods with microRNA microarray. *Genomics*, **92**, 122–128.
- Huber, W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**(Suppl. 1), S96–S104.
- Huber, W. *et al.* (2003) Parameter estimation for the calibration and variance stabilization of microarray data. *Stat. Appl. Genet. Mol. Biol.*, **2**.
- Kendzioriski, C. *et al.* (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.*, **22**, 3899–3914.
- Leung, Y. F. and Cavalieri, D. (2003) Fundamentals of cDNA microarray data analysis. *Trends Genet.*, **19**, 649–659.
- Liu, C. *et al.* (2004) An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. *Proc. Natl Acad. Sci. USA*, **101**, 9740–9744.
- Lu, J. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 843–848.
- Ozen, M. *et al.* (2008) Widespread deregulation of microRNA expression in human prostate cancer. *Oncogene*, **27**, 1788–1793.
- Patterson, T. *et al.* (2006) Performance comparison of one-color and two-color platforms within the microarray quality control (MAQC) project. *Nat. Biotechnol.*, **24**, 1140–1150.
- Pradervand, S. *et al.* (2009) Impact of normalization on miRNA microarray expression profiling. *RNA*, **15**, 493–501.
- Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.*, **2**, 418–427.
- Rao, Y. *et al.* (2008) A comparison of normalization techniques for microRNA microarray data. *Stat. Appl. Genet. Mol. Biol.*, **7**.
- Rocke, D. M. and Durbin, B. (2003) Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics*, **19**, 966–972.
- Sarkar, D. *et al.* (2009) Quality assessment and data analysis for microRNA expression arrays. *Nucleic Acids Res.*, **37**, e17.
- Smyth, G. and Speed, T. (2003) Normalization of CDNA microarray data. *Methods*, **31**, 265–273.
- Tusher, V. G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
- Wang, T. *et al.* (2007) A micro-RNA signature associated with race, tumor size, and target gene activity in human uterine leiomyomas. *Genes Chromosomes Cancer*, **46**, 336–347.
- Workman, C. *et al.* (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.*, **3**, research0048.1–research0048.16.
- Xiong, H. *et al.* (2008) Using generalized procrustes analysis (GPA) for normalization of CDNA microarray data. *BMC Bioinformatics*, **9**, 25.
- Yang, Y. H. *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.