

IDENTIFICATION AND NORMALIZATION OF PLATE EFFECTS IN cDNA MICROARRAY DATA

HENRIK BENGTSSON

Preprints in Mathematical Sciences
2002:28



LUND UNIVERSITY

Lund Institute of Technology
Centre for Mathematical Sciences
Mathematical Statistics

Identification and normalization of plate effects in cDNA microarray data

Henrik Bengtsson

Mathematical Statistics, Centre for Mathematical Sciences,
Lund University, Box 118, SE-221 00 Lund, Sweden.
hb@maths.lth.se

December 12, 2002 (revised)

Abstract

Introducing a new way of visualizing cDNA microarray data we have identified a new type of systematic variation, which we refer to as *plate effects*. We believe that plate effects are due to non-biological differences in the cDNA clones products spotted onto the microarray slides. By comparing the consistency of all replicates (both within and between slides) after performing 42 different normalization strategies we claim that the plate effects are non-negligible and should be corrected for in addition to the already known intensity dependent effects. For comparison between normalization strategies we introduce a novel robust genewise variability measure, which we call *measure of reproducibility*. In our case, we also found that not doing background correction improves the reproducibility significantly.

Keywords: cDNA microarrays; systematic variation; artifacts; plate effects; clone effects; normalization.

Contents

1	Introduction	3
2	Data	4
3	Artifacts in data	4
3.1	Systematic variation between plates	5
3.2	Possible sources for different artifacts	7
4	Normalizing data	8
4.1	Intensity dependent normalization	8
4.2	Platewise median normalization	9
4.3	Intensity dependent normalization performed platewise	9
4.4	Subsequential normalization strategies	10
4.5	About scale normalization	11
4.6	Alternative ways to group the data	11
5	Comparison between different strategies	11
5.1	Measure of reproducibility	12
6	Results	13
7	Discussion	15
8	Acknowledgements	15
	References	18
	List of Figures	19
	List of Tables	20

1 Introduction

The cDNA microarray technology is a relatively new technique for simultaneously measuring (relative) gene expression levels in two different cell lines. The technique covers a huge spectrum of applications such as identification of oncogenes (genes up regulated in cancer), time-serie studies of cell lines under different environmental stress conditions, clustering of expression profiles for identifying related genes etc. The cDNA microarray technique is also started to being used in clinic trial for making diagnosis on patient-to-patient basis. With the vast amount of biological data produced an increasing need for new statistical and computational methods is seen.

Robotically, using a so called *arrayer*, up to 40000 unique *complementary deoxyribonucleic acid* (cDNA) spots are printed onto a microscope glass slide. RNA from the two samples are purified separately, amplified and then turned into cDNA by reverse transcription. During the reverse transcription the two solutions of cDNA are labeled with two different dyes, commonly the fluorescent dyes Cy3 and Cy5. Equal amount of the two labeled cDNA solutions are cross-hybridized to the cDNA spots on the glass slide. The numerous amount of cDNA strands and the competitive nature of the hybridization makes it possible to approximate the relative gene expression of a certain gene with the relative amount of Cy3 and Cy5 in the corresponding spot. The amount of Cy3 and Cy5 in each spot is obtained by scanning the hybridized microarray slide at wavelengths 632 nm and 535 nm, which excite the Cy3 and Cy5 fluorescent dyes, respectively. The foreground and the background signals of the spots are extracted from the two images to be used in the downstream analysis.

In the above process, there are several sources of error that introduce artifacts in data. In addition to problems of such steps as selecting sample tissues and extracting RNA, transcribing it into cDNA and so on, there are several other quality related problem. Dust sticking to the slides could produce false signal peaks inside some spots, non-linearity between the two channels could favor one of the cell lines, spatial variation across a slide or systematic variation between replicated slides could additionally result in false conclusions. The list of artifacts is long where some are well understood whereas some are still to be explained or identified.

In this paper we will discuss what we believe is an artifact due to systematic variation between the cDNA clone libraries and/or systematic variation between spots printed during the several hours long printing process. We refer to this artifact as *plate effects*. As explained later, it is not obvious how to distinguish, a priori, between plate effects and effects due to differences between the two fluorescent labels. To identify the most dominant sources of systematic variations, we will apply a technique where different sequences of normalization methods, referred to as *normalization strategies*, are applied to data. Introducing a measure of reproducibility, we are then able to find the sequences of normalization steps that gives the most consistent signals across replicated slides.

All methods discussed in this paper have been implemented using the R software [IG96] and the com.braju.sma package [Ben01a], which is an object-oriented extension to the sma package [BDL⁺01] and which relies on the R.classes bundle for object-oriented programming with reference [Ben01b].

The data analyzed is described in section 2, followed by section 3, which lists possible artifacts in data. Previously known systematic variations are discussed, but also a not commonly discussed variation, referred to as *plate effects*, is introduced. In section 4 different methods to normalize for artifacts are explained. Different methods for normalizing for plate effects are

given in detail. Section 5 defines the 42 different normalization strategies to be compared and the *Measure of Reproducibility* (M.O.R.) used to compare the different strategies. The results are discussed in section 6 and section 7 concludes with a discussion of underlying reason for plate effects and suggestion of continued investigations.

2 Data

Data for the eight cDNA microarray slides analyzed in this paper comes from the Matt Callow Lab at Lawrence Berkeley National Laboratories [CDG⁺00]. The experimental setup was to compare the gene expression between apolipoprotein AI (apo AI) knockout mice and a reference of C57Bl/6 mice. Two main experiments were done. The first one compared eight different apo AI knockout mice with a reference pool of eight different C57Bl/6 mice on eight cDNA microarray slides. The second experiment compared the individual C57Bl/6 mice with the pool of them (same pool as above). In total 16 slides were hybridized. In this paper we will only present our results using the eight slides from the second experiment, but the same analysis on the eight apo AI slides shows very similar results (and indeed almost identical numbers). However, the first set of slides is believed to contain fewer differentially expressed genes compared to the other set. Hence, it is easier to argue that some of the underlying assumptions used in the analysis are true. Part of the extracted RNA from the livers of the eight individual mice in each slide set was synthesized separately into cDNA together with Cy3-dUTP. A mix of the same extracted RNA was then used to produce the Cy5-dUTP labeled reference cDNA. A similar setup was used for the six unique apo AI knockout mice.

For the production of the microarrays, approximately 5600 expressed sequence tags (ESTs) from the I.M.A.G.E. consortium database [LAPS96] were selected. 257 of these ESTs represent genes with possible and confirmed associations with lipid metabolism. The slides were spotted using a 4-by-4 print-tip arrayer with a total of 7056 spots on each slide. After hybridization the slides were scanned and the foreground and background signals were extracted using the Spot software [YBDS00]. Of the 21 rows in each print-tip group the last two were excluded resulting in a total of 399 spots per print-tip group or in total 6384 spots per slide. Among these spots there were in total 5357 different genes and 840 blank spots. Among the 5357 genes, six of them were spotted trice, 175 of them duplicated and the rest were spotted only once per slide. In figure 1, the log gene expression ratios on slide 6 are shown. To the left, the log gene expression ratios, $M = \log_2(R/G)$, are plotted against the log gene intensities, $A = 1/2 \cdot \log_2(R \cdot G)$, where R and G in this case are the non-background subtracted signals for the Cy5 (red) and Cy3 (green) channels, respectively. To the right, a spatial plot is depicting how the gene expressions are positioned on the slide.

3 Artifacts in data

As seen in the left plot in figure 1, the log ratios are strongly dependent on the log intensities. A low intensity spot tends to be redder (up regulated) than a high intensity spot and this intensity dependency is non-linear. All slides show the same systematic effects. Because of the way the experiment was set up, most genes are expected to be non-differentially expressed and therefore it is believable that this non-linearity is an artifact and should be corrected for. There is a variety of methods available for correcting for this effect, but throughout this paper we will use the intensity dependent normalization methods introduced by [YDLS]. Furthermore, looking at the spatial distribution of log ratios, cf. the right plot in figure 1, we find a repeated

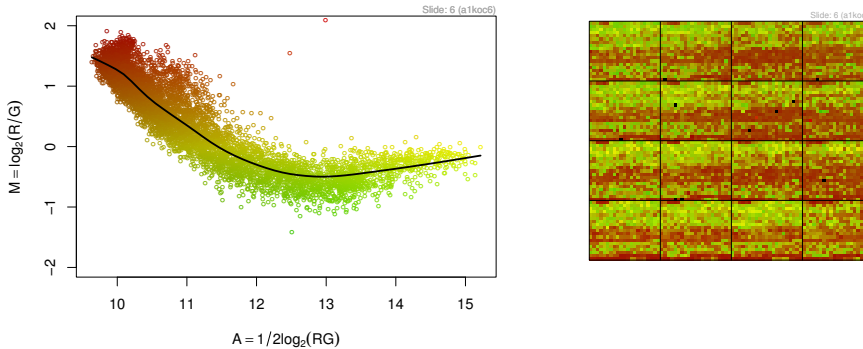


Figure 1: The log ratios, M , of gene expression for the Cy3 and the Cy5 channels of the 6384 spots on slide 6. *Left*: The log ratios, M , versus the log intensities, A , of the data. The solid line is the fitted robust local regression line (lowess) and it shows the intensity dependent bias effect of slide 6. *Right*: The same data plotted as laid out on the sixth cDNA microarray slide. Green colors correspond to a negative log ratio and red colors to a positive log ratio. The 4-by-4 grid separates the regions that were spotted by each individual print-tip.

pattern of horizontal stripes, not only for this particular slide but for all slides in the data set. Comparing this plot with the M vs. A plot and keeping the intensity dependent artifacts in mind, we notice that these stripes might be due to different intensities between the stripes. The red areas with apparent up-regulated genes are also darker and the down-regulated and the non-differentially expressed genes tend to be located in brighter areas. We will return to this in the next section. As shown in section 4.1, doing an intensity dependent normalization will indeed remove most of the spatial artifacts, but the final results will indicate that the non-linearity between the Cy3 and Cy5 channels is not the only contributor to such systematic variations. Even if it would, it still has to be explained why the intensity varies spatially across the slides in such a regular pattern.

3.1 Systematic variation between plates

At each so called *dip* the print-tips of the arrayer spot a number of (different) cDNA clones onto the glass slide. In our case 16 spots are spotted at each dip. Depending on how the arrayer is programmed it can then either move on to the *rest of the slides* and spot the same clones over and over again with possible intermediate refilling of the print-tips, or it can get a new set of clones from the *micro-titer plate*, we say that the arrayer does a *source visit*, and finish the first glass slide before moving onto the next. The most commonly used approach is to spot all glass slides in the arrayer before spotting the next set of clones. The positions of all the spots spotted at each dip are fully determined by the layout of the print-tips (here 4-by-4 print-tips approximately $200 \mu\text{m}$ apart) and the way we have chosen to program the arrayer. Most commonly is that the next set of 16 clones is spotted to the right of the ones from the previous source visit. This is repeated until one row is filled when the arrayer continues with the next row (starting over at the very left). Most modern arrayers allow user to spot the clones in any order and at any position, e.g. column wise or randomly, but in most cases (also in this current setup) the arrayer moves from left to right row by row. With the knowledge of the exact printing of the slides one can create a plot where the data, e.g. the log ratios, of the 6384 spots are plotted in the order of when they were printed onto the slides. Such a plot is shown

in figure 2, which has the log ratio on the y-axis and the time of printing on the x-axis. From this plot it is clear that the log ratios vary with the time point *when* the clones were spotted onto the slides. The systematic variation of the log ratios over time is exactly that found to be repeated over the rows within each print-tip group. The horizontal and vertical lines and the box plot to the right will be explained below. At a first glance it looks like there are four to

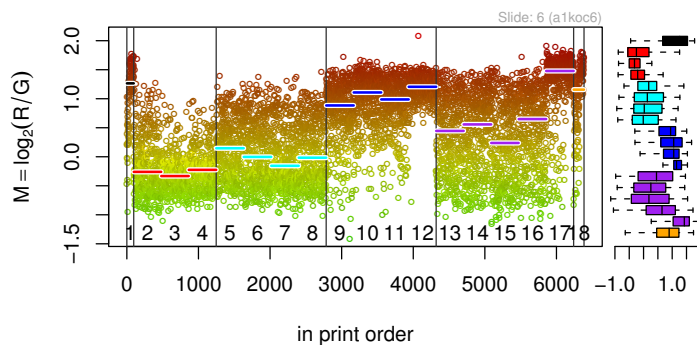


Figure 2: *Left*: The log ratios of the 6384 spots on slide 6 in the order of when they were printed onto the slide. At each print dip 16 spots were spotted simultaneously. The vertical lines mark out the different clusters of plates. The horizontal lines are the biases of each plate group, which are colored according to which plate cluster (source) they belong to. *Right*: Box plot showing the median and the variability for each plate using the same ordering and coloring as in the left figure. Plate 1 is at the top and plate 18 at the bottom.

six printing periods (separated by vertical lines for clarity) with different properties. It turns out that the clones can be grouped into these groups depending on what lab produced them. The slides used here were spotted using 18 different 384-well micro-titer plates and it turns out that each of the vertical lines coincides exactly with a plate *and* lab switch. In figure 2 the plate numbers are shown along the x-axis. Instead of a *print-order plot* we could also call it a *plate-order plot*. In order not to focus too much on different lab procedures we will refer to these different groups of plates as *plate clusters*. Plate 1 consists mostly of water (empty spots), plates 2-4 come from the I.M.A.G.E. clone library, plates 5-8 were clones selected by the Kingley lab at Stanford University, plates 9-12 were brain-tissue clones obtained by the Cheng Lab at Lawrence Berkeley National Laboratories, plates 13-17 were purchased from Research Genetics and plate 18 comes from the Vulpe Lab. For plate 1, only 96 clones were spotted and for plate 18 144 clones were spotted. For all other plates all 384 clones were spotted. Further more, 95 out of the 96 spots (99%) from plate 1 are blanks, 336 out of 384 spots (88%) on plate 12, 326 out of 384 (85%) on plate 17 and 64 out of the 144 spots (44%) from plate 18 are blanks. In total there are 840 empty spots. In figure 2 the plate clusters have been clarified by vertical lines. The horizontal lines, which are colored according to which plate cluster they belong to, are the *median log ratio* for each plate. The box plot to the right depicts the distribution of the log ratios for each plate. Plate 1 is at the top and plate 18 at the bottom. The same coloring scheme will be used throughout the paper.

As suggested before, the systematic variations in data, especially the spatial effects seen in figure 1 could be explained by the combination of systematic variation in the spot intensities

across the slides and an intensity dependent log-ratio effect. The print-order plot of the log intensities in figure 3 together with the print-order plot of the log ratios in figure 2 confirms this idea.

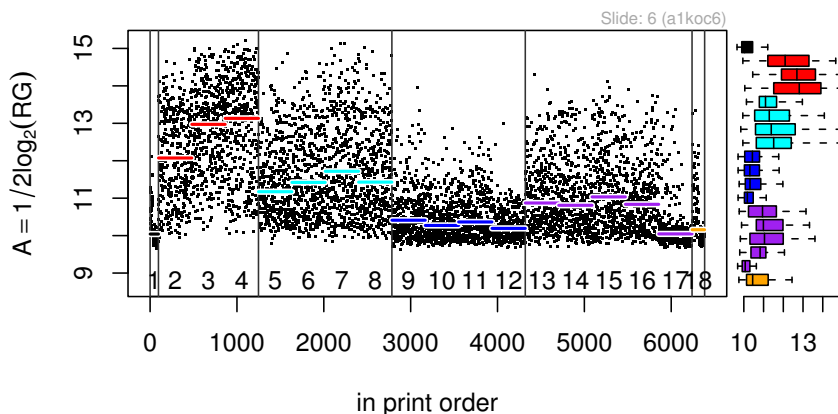


Figure 3: The log intensities on slide 6. The plates with a lot of blanks (1,12,17 & 18) do have low intensities. Furthermore, the clones from brain tissue on plate 9-12 are, as expected, not very responsive to the test and control samples, which are from liver RNA. *Left*: The log intensities in print order. *Right*: Box plot displaying the median and the variability of the log intensities for each plate using the same ordering and coloring as in the left figure.

Although the variation of the signal strengths (intensities) between the plates explains most of the spatial effects we will find that there is also a systematic variation in log ratios between the plates that can not be explained by the intensity effects.

3.2 Possible sources for different artifacts

The source of plate effects comes from the clones and the plates, i.e. from the step before spotting the slides. In this sense, all slides should approximately have the same plate effects. The intensity dependent effects, however, results from a non-linearity between the Cy3 and the Cy5 channels. There are at least two different possible sources for this non-linearity. First is the fact that the incorporation rate of Cy5-dUTP is about 60% (in moles) of that of Cy3-dUTP [HQA⁺00]. The second source of intensity dependent effects occurs during the scanning of the slides. Due to *quenching*, the Cy3 and the Cy5 channels are not linearly responsive over the full intensity range [RCB⁺01]. Since the quenching effect is added after the plate effect and the labeling effect it is natural to remove it first. However, since the quenching effect and the labeling effect both have the same symptom it is hard to distinguish between them. Even if it would be possible to identify the exact amount of quenching a remaining problem is to separate the plate effect from the labeling effect. This is especially hard since these two effects are most likely combined in the hybridization step. Exactly how the two effects are combined is unknown. We will attack the problem of cleaning up data by first assuming that the quenching effect is minimal and negligible. Since we can not distinguish between the plate

and labeling effects, we will try different sequences of normalization methods, referred to as *normalization strategies* and then use a measure of reproducibility (section 5.1) to find the best sequence of normalization steps and let it guide us to which artifact is the most dominant one. The above discussion could, technically speaking, be formulated as a variance component model, from which we could find existing identifiability problems. From such a probabilistic model it would be possible to address our problem using maximum likelihood methods. Such models have been suggested, but they tend to be very specific and ad hoc, especially if they are aiming to be general models. Here we will keep such model thinking in mind, but we will only use algorithms to reflect our model in mind.

4 Normalizing data

There are several different approaches to use for taking the plate effects into account when normalizing the data. One can either try to find a physical and biological model to explain the effects or one could take a blindfolded black box model for normalizing the data. Here we will use the latter approach, mainly because at the moment of writing we do not know at what step(s) or combination of steps in the cDNA microarray process the plate effects actually occurs and also because it is somewhat more objective approach for arguing for existence of plate dependent effects. Next we will present a few alternatives and in section 5.1 we define a *measure of reproducibility* for comparison between them.

4.1 Intensity dependent normalization

In figure 4 the log ratios after scaled print-tip intensity normalization (cf. [YDLS]) are plotted. The spatial artifacts are less significant after the normalization, but the variation between the plates are still there. One possible strategy is to remove the plate biases after doing the intensity normalization, cf. section 4.4. For details on intensity dependent normalization methods see [YDLS].

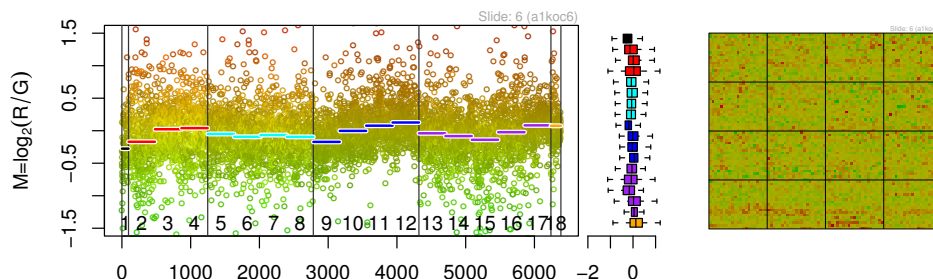


Figure 4: Gene expressions (for slide 6) *after* scaled print-tip intensity normalization. *Left*: Print-order plot of the log ratios with a horizontal box plot showing the distribution of the log ratios within each plate group. *Right*: The spatial location of the log ratios. Normalizing for intensity dependent artifacts does indeed remove a lot of the spatial and some of plate effects.

4.2 Platewise median normalization

Another way to normalize the data can be found by assuming that the average log ratio on each plate or plate cluster should be zero. This is similar to the assumption used for the intensity normalization where one either assumes zero shift for the whole slide or for each print-tip group, but instead of print-tip groups we here focus on plates or plate clusters. Even if the two assumptions look similar, there is an intrinsic difference between them. The clones on the plates might be selected in such a way that some plates may contain *many* clones expected to be differentially expressed, whereas such an *a priori* selection of clones will be less likely to affect the print-tip groups since they contain clones from *all* plates and therefore are more random. The assumption of zero shift for each plate can be violated if the expression levels of the clones on a plate biologically differ significantly between the test and the control samples, which could be the case with the apo AI vs. control pool dataset. We do not believe this is the case in mouse experiment studies here, especially since we are looking at eight control mice versus the pool of them. Hence, we assume that all plates should have an average log ratio of zero, which will also fulfill the assumption that the overall shift on the slide is zero. In figure 2, the median log ratio of each plate was depicted by a colored horizontal line. The results from

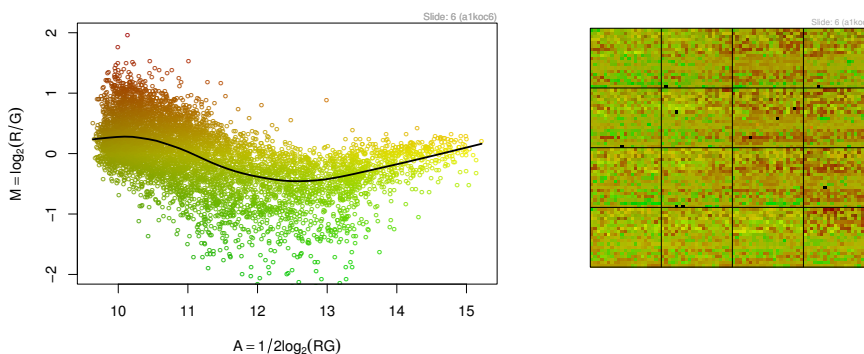


Figure 5: Gene expressions (for slide 6) *after* platewise bias normalization. *Left*: The log ratios versus the log intensities. *Right*: The spatial location of the log ratios. Removing the constant bias within each plate group will remove a lot of the spatial and some of intensity dependent effects.

normalizing the data platewise assuming zero bias, i.e. by shifting all the horizontal lines in figure 2 to zero is shown in the M vs. A plot and the spatial plot in figure 5. Comparing these with the corresponding plots for the non-normalized data (figure 1) one sees that the intensity dependent effects are somewhat removed by doing a plate-bias normalization and the previously so strong spatial effect is also decreased. The remaining intensity dependent effects are studied in section 4.4.

4.3 Intensity dependent normalization performed platewise

An alternative method is to do intensity normalization on each plate individually. This idea is also very similar to the print-tip intensity normalization method, but with the focus on the plates. In figure 6 the lowess estimated intensity curves for each plate are shown. Here one has to be even more careful since the number of data points might be too small to fit a smooth curve robustly over a large intensity interval. In such cases we have to look for larger groups, see section 4.6, or find a clever way to combine different fitted curves, cf. [YDL⁺02]. In

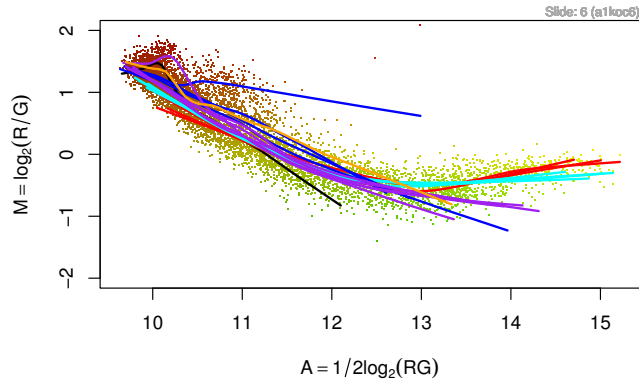


Figure 6: Intensity dependency for each plate (on slide 6). The top most curve (blue) of plate 12 is shifted upwards because it has only one data point in the upper intensity range, $A = [12, 13]$, with a high value of the log ratio.

our data this problem arises for plate 1 and plate 18, but for simplicity we will neglect such problems in this study. We do not believe it will affect our general results. However, if we were about to present a list of differentially expressed genes we would have to consider this problem more carefully. The platewise intensity normalized data for slide 6 is depicted in the print-order and spatial plot in figure 7.

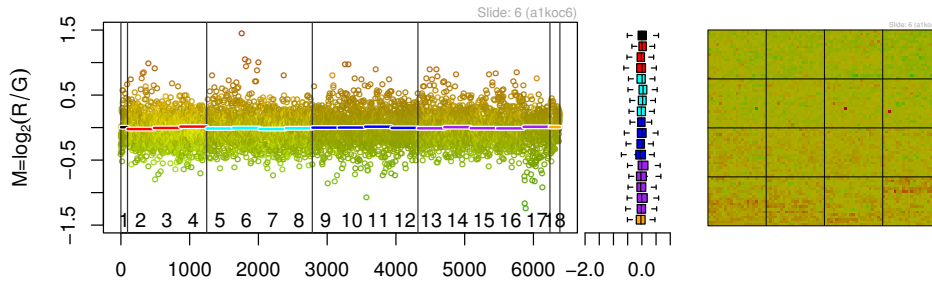


Figure 7: Gene expressions (for slide 6) *after* platewise intensity dependent normalization. *Left*: Print-order plot of the log ratios with a horizontal box plot show the distribution of the log ratios within each plate group. *Right*: The spatial location of the log ratios.

4.4 Subsequential normalization strategies

When doing a platewise constant-shift normalization some of the intensity effect will be removed, but not completely as seen in figure 5. The remaining effect can be removed by performing a regular intensity normalization. However, care has to be taken since such a normalization will correct plate normalized data *too much* and reintroduce “new” plate effects. A naive approach is to do another round of plate dependent bias normalization. Similar strategies can of course be applied to correct for remaining plate effects after an initial intensity dependent normalization or any other normalization methods. More importantly, the appearance of *both* plate effects and intensity effects raises the question of which one of the two effects is

the most dominant and which one should be corrected for first or if both effects should be corrected for simultaneously. We leave the problem of doing simultaneous normalization for the future and focus on simpler cases, which is most likely the ones that will be used in practice. Recall section 3.2 for a short discussion about this problem.

4.5 About scale normalization

Since some of the plates come from different sources with clones from different animals and different tissues and since they have been prepared by different people it is likely that their variability differ. For this reason we do *not* assume nor believe that all the plates should have equal variability. Therefore a normalization of the scale is not reasonable. However, for similar reason as given in the previous paragraph, it is reasonable to assume equal variability across print-tip groups. It is *neither* correct to assume that all plates should have about the same average *intensity*. The noise and signal levels in both the Cy3 and Cy5 channels differs between plates (or sources) and therefore not only the variance of log ratios, but also the variance of the log intensities are expected to differ. Note that even with these arguments it is still correct to assume that the log ratios are zero on average.

4.6 Alternative ways to group the data

Instead of focusing on plates we could create other groups, which are normalized separately. The main thing to keep in mind is, as for almost all normalization methods, that we must be confident that it is possible to assume that the average log ratio for each group is zero. For instance, in our case we could have grouped the data into one group containing all the empty spots and then the rest of the clones grouped according to which lab it was that produced the clones. Another possible way is to estimate the average plate median shifts or plate intensity dependent shifts for all (standardized) slides and use these platewise constants or smooth curves for normalizing each individual slide. In this paper, we will not follow these tracks, but only perform platewise normalizations.

5 Comparison between different strategies

We will compare seven major approaches for normalizing data with each other and with the non-normalized case. In one group we normalize the data for (slide, print-tip and scaled print-tip) intensity dependent effects (labeled SI(A), Pr(A) and sPr(A) in the tables and the figures), in another group we normalize the data for (constant) plate dependent effects (labeled PI) and in a third group we will do intensity base normalization for each plate group (PI(A)). Each of these normalizations will be done on both non-background subtracted and background subtracted data (labeled bg). In addition to these, we will test different sequences of them. In one group we first unshift the plate biases and then do standard intensity normalizations (labeled PI-SI(A), PI-Pr(A) and PI-sPr(A)). The reverse order is also tested (labeled SI(A)-PI, Pr(A)-PI and sPr(A)-PI). We do also test plate-intensity-plate normalization (PI-SI(A)-PI, PI-Pr(A)-PI and PI-sPr(A)-PI). Finally we also combine the intensity normalization over plates with the standard intensity normalization methods in that order (PI(A)-SI(A), PI(A)-Pr(A) and PI(A)-sPr(A)) and in the reverse order (SI(A)-PI(A), Pr(A)-PI(A) and sPr(A)-PI(A)). As for the other cases, all these strategies are performed on both non-background subtracted and background subtracted data. In total we compare $2 \cdot (1 + 3 + 1 + 1 + 3 + 3 + 3 + 3 + 3) =$

42 different strategies. Not mentioned yet, as a very final step, for each of the strategies (independently), we rescale the log ratios so that all slides have the same MAD (within each normalization strategy, *not* between). In detail, the rescaling is done as

$$\begin{aligned} Z_l &\leftarrow \text{MAD}_{k=1:K} M_{k,l}, \forall l \\ \bar{Z} &\leftarrow \left(\prod_{l=1}^L Z_l \right)^{1/L} \\ M_{k,l} &\leftarrow \frac{\bar{Z}}{Z_l} \cdot M_{k,l}, \forall k, l \end{aligned}$$

where $k = 1, \dots, K$, $l = 1, \dots, L$ and K is the number of spots on each slide and L is the number of slides (not to be confused with the number of genes and the number of replicates).

5.1 Measure of reproducibility

To be able to decide on the optimal *normalization strategy* we will measure the *variabilities of the gene-by-gene residuals*. The average variability and the variability of the variabilities can be seen as a measure of reproducibility of the data from the biological system. Ignoring the bias, the smaller the variabilities are, the closer to the correct values we are. Since we do not know the true values of the expression ratios we can not check if the different normalization methods introduce bias to data or not and for this reason we can neither give a rigorous proof that the normalized data is more *consistent*. However, the normalization methods described about have all the property of normalizing the expression ratios for the genes to a common level and therefore also removing part of the biases. Since the normalization methods, except for scale normalization, are only using additive operations and not multiplicative operations the risk for over-fitting data by chance is expected to be low. Further more, if we are looking for differentially expressed genes, i.e. outliers, small biases in the average expression ratios will not that much affect which genes we find in the end.

We define the variability, d_i , of gene i as the median absolute deviation of the gene residuals:

$$d_i = 1.4826 \cdot \text{median}_{j=1:J} |r_{i,j}|,$$

where 1.4826 is a constant making the MAD a consistent estimator of the standard deviation for normally distributed data. The residuals are defined as

$$r_{i,j} = M_{i,j} - \text{median}_{j=1:J} M_{i,j},$$

where $M_{i,j}$ is the log ratios of gene i and replicate j . In figure 8 the log ratios for the 24 replicates of the triplicated gene 1540 are plotted for two different strategies. We see that for this specific gene the print-tip followed by plate intensity normalization (red discs) makes the measurements (of this gene) much *more consistent* compared to the non-normalized approach (blue circles). The former has a genewise MAD of 0.115 (sample variance of 0.00988) whereas the latter has a MAD of 0.190 (sample variance of 0.0697). Obviously this might only be true for this specific gene and therefore we will look at the *overall distribution of the gene-by-gene variabilities*. This can be done looking at some upper quantile, e.g. the 95% or 99% quantile of the genewise variabilities, or one could look at the mean variability, which is

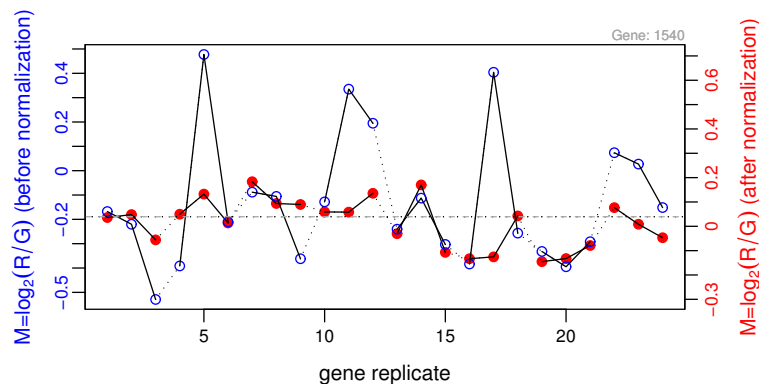


Figure 8: The log ratios for all the 24 replicates of gene 1540 found on the eight slides after performing intensity normalization first for each print-tip group and then for each plate group (red discs) compared to not doing any normalization at all (blue circles). The left scale is for the normalized log ratios and the right one for the non-normalized ones. Note that the unit lengths of the two vertical axes are equal, but their origins have been aligned for simplifying visual comparison.

not only a measures of the most common variability but it is also sensitive to genes with high variabilities;

$$M.O.R. = \frac{1}{N} \sum_{i=1}^N d_i$$

where N is the number of genes. Note that the *smaller* $M.O.R.$ is the *more* reproducible the data is. Finally, the rescaling of the log ratios across all slides, which was discussed in previous section, will in general increases (worsen) the $M.O.R.$ measure. However, the rescaling is commonly small and therefore its effect on the $M.O.R.$ and hence its effect on the overall results in this paper is negligible.

6 Results

Using the measure of reproducibility it turns out that the best way of normalizing our data is to correct for both the intensity effect within each print-tip group, $Pr(A)$ or $sPr(A)$, and for the intensity effect within each plate group, $Pl(A)$. The overall optimal strategy, for this data set, is to start with the scaled print-tip intensity normalization, $sPr(A)$, *followed* by the platewise intensity normalization, $Pl(A)$. Doing so, the first step reduces the genewise variability by approximately 55%, whereas the second step reduces it another 5% relative to the variability of the non-normalized data (or approximately 10% compared to the previous step). See table 1 and the box plot in figure 9, which illustrate this.

This is an argument that the systematic variation seen in the data can not be explained solely by intensity dependent bias effects, but also by some other effects, which we refer to as plate effects.

Using the measure of reproducibility, we also found that *not* doing background correction resulted in more coherent measurements across slides and across replicates. This was the

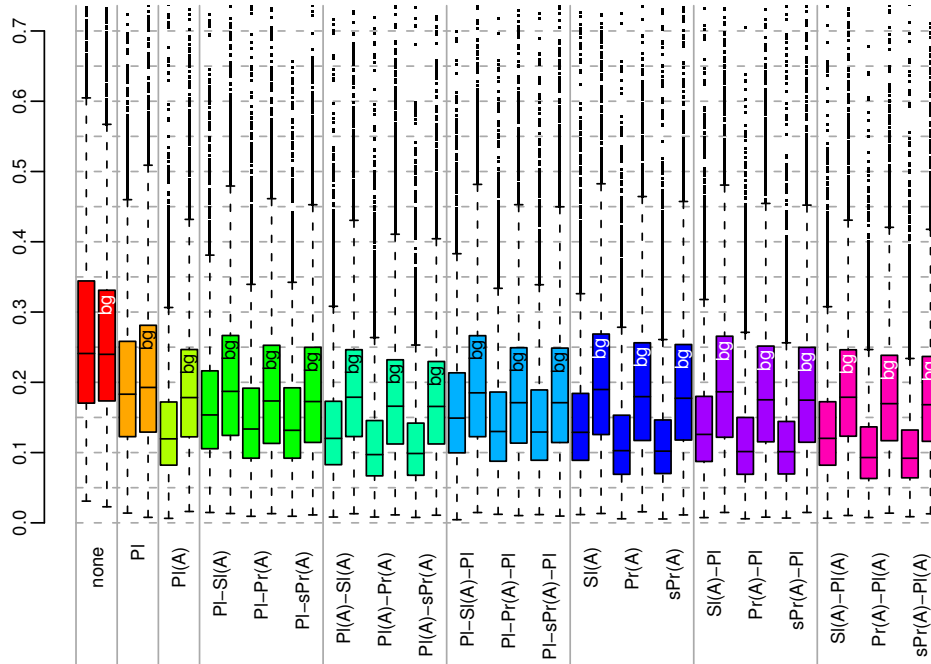


Figure 9: Comparison between the sample distributions of the gene log-ratio variabilities after (2×21) different normalization strategies have been applied. Note that the tails of the sample distributions have been cut of at 0.75 and that they do extend (approximately equally far) beyond this level. Legend: none - No normalization, PI - Constant plate bias removed, PI(A) - Intensity dependent plate bias removed, Pr(A) - Print-tip intensity dependent bias removed, sPr(A) - Scaled print-tip intensity dependent bias removed, SI(A) - Slide intensity dependent bias removed. Normalization strategies that start out with background subtracted signals are in addition labeled with 'bg'. For example, PI-sPr(A) refers to a normalization sequence followed by a plate normalization and ended by a scaled print-tip normalization.

case for all normalization strategies tested. The only time background subtraction actually improved the reproducibility was in the case when no normalization was performed (labeled 'none' in the table and in the box plot).

Furthermore, we see that doing intensity normalization for each print-tip individually, Pr(A) or sPr(A), is always better than doing it for the whole slide at once, SI(A). We believe that is because normalization print-tip by print-tip will correct for spatially dependent variation across the slide, whereas slide-by-slide normalization will not.

It should be mentioned that we get almost identical results if a non-robust measure of reproducibility, e.g. the sample variance, is used instead of the sample median absolute deviation in the calculation of the variability of the genewise residuals. Finally, performing the same analysis using the apo AI knockout mice dataset, which expect to have a few differentially expressed genes, also gives that a print-tip followed by a plate intensity normalization performs better than just a print-tip intensity normalization. In [YDLS] the authors identify eight genes (or ESTs) to be the most differentially expressed genes. The top ten genes (the eight listed

plus two others) were also confirmed by real-time quantitative polymerase chain reaction (RT-PCR) to be differentially expressed. All the normalization strategies that we found to be the best do all rank these ten genes as the most differentially expressed genes. However, these genes are so extreme that almost all normalization strategies applied will identify them to be the most differentially expressed. Hence, we can only conclude that our best methods will not remove the signal for these top ten genes.

7 Discussion

The actual source of the plate effects is unknown. It could be due to different salt concentrations of the spots, which in turn are due to different levels of evaporation between the plates. This would explain how the plate effect is propagated to the signals, but it still does not explain why it occurs in the first place. We believe this is an important research topic, which hopefully will gain more knowledge about the cDNA microarray technology.

The large difference between not performing background correction and doing it is interesting. When subtracting the background from the foreground signal more noise (variation) *is* indeed introduced, but we do not believe it is of the order of magnitude as seen for almost all normalization strategies tested. We believe that the background estimates, which are based on measurements for the surrounding regions of the spots, are not the same as the background noise affecting the internals of the spots. A physical explanation for this would be that the background noise found in the regions around the spots are not sticking to the spots themselves, because where the spots are the glass is already covered by cDNA and nothing but the matching cDNA sequences can hybridize to the foreground areas. Furthermore, some (not all) of the background estimates from the Spot image analysis software shows clear print-order effect too. This was surprising and has to be investigated further.

Finally, an argument against our measure of reproducibility is that it is only including the variance. The reason for this is that we do not know the true gene expression levels for any genes, except maybe for the top ten most differentially expressed genes. However, recently it has become common to add control spots to cDNA microarray slides, e.g. spike-ins, negative and positive controls etc. We are currently working on methods for comparing normalization strategies using both the bias and variance information of such control spots.

8 Acknowledgements

I am in big debt to professor Terry Speed at the Statistics Department, University of California, Berkeley, for supervising me while visiting the Statistics Department at UC Berkeley in 2000/2001. I am also very thankful to Jean Yang and the rest of the Speed Group for taking the time to explain the cDNA microarray technique to me and for fruitful discussions about it. I would also like to thank Matt Callow at the Lawrence Berkeley National Laboratory for providing the data and answering all my questions about details in the experimental setup. Also, if it would not be for a small chat in July 2001 that I had with Karen Berger at the Monica Moore's Lab, Ernest Gallo Clinic & Research Center in Emerville, maybe I would never have found out about the plate effects. I would also like to thank my supervisors Ola Hössjer and Jan Holst for giving me great feedback, support and also for proof reading this paper. Finally

a big thanks to all the great contributors of the R project, which provides an important foundation for my applied research.

This work was indirectly supported by the following grants: The Swedish Foundation for International Cooperation in Research and Higher Education (STINT), The Foundation BLANCEFLOR Boncompagni-Ludovisi, née Bildt, The Fulbright Commission Grant, The Ernhold Lundströms Stiftelse and The Royal Swedish Academy of Sciences Research Grant.

Method	0%	1%	50%	<i>M.O.R.</i>	99%	100%	% <i>M.O.R.</i>
none	0.0309	0.0615	0.241	0.270	0.708	1.17	1
bg-none	0.0228	0.0631	0.240	0.264	0.676	1.23	0.977
PI	0.0139	0.0375	0.183	0.204	0.588	1.59	0.756
bg-PI	0.0078	0.0393	0.193	0.218	0.623	1.08	0.806
PI(A)	0.00637	0.0261	0.120	0.139	0.438	1.22	0.514
bg-PI(A)	0.0161	0.0440	0.178	0.199	0.597	1.11	0.739
PI-SI(A)	0.0147	0.0389	0.154	0.173	0.515	1.36	0.641
bg-PI-SI(A)	0.0133	0.0401	0.187	0.209	0.600	0.929	0.776
PI-Pr(A)	0.00933	0.0299	0.134	0.154	0.495	1.29	0.569
bg-PI-Pr(A)	0.0131	0.0330	0.173	0.196	0.588	0.953	0.726
PI-sPr(A)	0.00939	0.0285	0.132	0.154	0.500	1.34	0.570
bg-PI-sPr(A)	0.0113	0.0329	0.173	0.195	0.595	1.04	0.724
PI(A)-SI(A)	0.00829	0.0252	0.120	0.139	0.440	1.22	0.516
bg-PI(A)-SI(A)	0.0130	0.0443	0.179	0.200	0.598	1.11	0.740
PI(A)-Pr(A)	0.00827	0.0223	0.0971	0.119	0.428	1.17	0.441
bg-PI(A)-Pr(A)	0.0113	0.0362	0.166	0.185	0.575	1.00	0.686
PI(A)-sPr(A)	0.00771	0.0231	0.0987	0.118	0.445	1.41	0.438
bg-PI(A)-sPr(A)	0.0109	0.0370	0.166	0.185	0.576	1.21	0.684
PI-SI(A)-PI	0.00444	0.0353	0.149	0.168	0.503	1.08	0.621
bg-PI-SI(A)-PI	0.0147	0.0382	0.185	0.208	0.594	0.940	0.769
PI-Pr(A)-PI	0.0121	0.0273	0.130	0.149	0.479	1.09	0.553
bg-PI-Pr(A)-PI	0.0103	0.0313	0.171	0.194	0.583	0.988	0.718
PI-sPr(A)-PI	0.0117	0.0279	0.129	0.150	0.486	1.06	0.556
bg-PI-sPr(A)-PI	0.00964	0.0320	0.171	0.194	0.599	1.08	0.718
SI(A)	0.0118	0.0310	0.129	0.149	0.499	0.907	0.552
bg-SI(A)	0.0133	0.0417	0.190	0.212	0.631	1.00	0.786
Pr(A)	0.00575	0.0234	0.103	0.124	0.446	0.924	0.460
bg-Pr(A)	0.0157	0.0349	0.180	0.200	0.592	1.05	0.741
sPr(A)	0.00549	0.0228	0.102	0.123	0.480	1.27	0.455
bg-sPr(A)	0.0113	0.0349	0.177	0.199	0.601	1.26	0.737
SI(A)-PI	0.00746	0.0297	0.126	0.147	0.499	0.900	0.543
bg-SI(A)-PI	0.0148	0.0365	0.186	0.209	0.618	0.944	0.775
Pr(A)-PI	0.00575	0.0220	0.101	0.122	0.446	0.911	0.453
bg-Pr(A)-PI	0.0081	0.0331	0.175	0.196	0.587	0.977	0.727
sPr(A)-PI	0.00682	0.0231	0.101	0.121	0.476	1.26	0.449
bg-sPr(A)-PI	0.0148	0.0348	0.175	0.196	0.606	1.18	0.726
SI(A)-PI(A)	0.00661	0.0253	0.120	0.139	0.447	0.983	0.514
bg-SI(A)-PI(A)	0.0104	0.0436	0.179	0.199	0.599	1.12	0.739
Pr(A)-PI(A)	0.0076	0.0217	0.0929	0.111	0.394	0.848	0.412
bg-Pr(A)-PI(A)	0.0138	0.0369	0.170	0.189	0.553	0.950	0.702
sPr(A)-PI(A)	0.00872	0.0220	0.0919	0.110	0.424	1.01	0.407
bg-sPr(A)-PI(A)	0.0129	0.0372	0.168	0.188	0.562	1.15	0.696

Table 1: Comparison between different normalization strategies. Legend: none - No normalization, PI - Constant plate bias removed, PI(A) - Intensity dependent plate bias removed, Pr(A) - Print-tip intensity dependent bias removed, sPr(A) - Scaled print-tip intensity dependent bias removed, SI(A) - Slide intensity dependent bias removed. Normalization strategies that start out with background subtracted signals are in addition labeled with 'bg'.

References

- [BDL⁺01] Ben Bolstad, Sandrine Dudoit, Ingrid Lonnstedt, Natalie Roberts, and Jean Yee Hwa Yang. R package: Statistics for microarray analysis (sma). <http://www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html>, 2001. Statistics Department, University of California at Berkeley.
- [Ben01a] Henrik Bengtsson. com.braju.sma - an object oriented package for microarray analysis in [R]. <http://www.braju.com/R/>, 2001. Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Sweden.
- [Ben01b] Henrik Bengtsson. R.classes - a bundle for object-oriented programming with reference in [R]. <http://www.braju.com/R/>, 2001. Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Sweden.
- [CDG⁺00] M. Callow, S. Dudoit, E. Gong, T. Speed, and E. Rubin. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research*, 10(12):2022–9, December 2000.
- [HQA⁺00] Priti Hegde, Rong Qi, Kristie Abernathy, Cheryl Gay, Sonia Dharap, Renee Gaspard, Julie Earle-Hughes, Erik Snedrud, Norman Lee, and John Quackenbush. A concise guide to cDNA microarray analysis. *Biotechniques*, 3(29):548–562, September 2000.
- [IG96] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.
- [LAPS96] Greg Lennon, Charles Auffray, Mihael Polymeropoulos, and Marcelo Bento Soares. The I.M.A.G.E. Consortium: An integrated molecular analysis of genomes and their expression. *Genomics*, 33:151–152, April 1996.
- [RCB⁺01] Latha Ramdas, Kevin R Coombes, Keith Baggerly, Lynne Abruzzo, W Edward Highsmith, Tammy Krogmann, Stanley R Hamilton, and Wei Zhang. Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biology*, 2(11):research0047.1–0047.7, 2001.
- [YBDS00] Yee Hwa Yang, Michael Buckley, Sandrine Dudoit, and Terry Speed. Comparison of methods for image analysis on cDNA microarray data. Technical Report 584, Statistics Department, University of California at Berkeley, Nov 2000.
- [YDL⁺02] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M. Lin, Vivian Peng, John Ngai, and Terry Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4), Apr 2002.
- [YDLS] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, and Terry Speed. Normalization for cDNA microarray data (manuscript in preparation).

List of Figures

- 1 The log ratios, M , of gene expression for the Cy3 and the Cy5 channels of the 6384 spots on slide 6. *Left*: The log ratios, M , versus the log intensities, A , of the data. The solid line is the fitted robust local regression line (lowess) and it shows the intensity dependent bias effect of slide 6. *Right*: The same data plotted as laid out on the sixth cDNA microarray slide. Green colors correspond to a negative log ratio and red colors to a positive log ratio. The 4-by-4 grid separates the regions that were spotted by each individual print-tip. 5
- 2 *Left*: The log ratios of the 6384 spots on slide 6 in the order of when they were printed onto the slide. At each print dip 16 spots were spotted simultaneously. The vertical lines mark out the different clusters of plates. The horizontal lines are the biases of each plate group, which are colored according to which plate cluster (source) they belong to. *Right*: Box plot showing the median and the variability for each plate using the same ordering and coloring as in the left figure. Plate 1 is at the top and plate 18 at the bottom. 6
- 3 The log intensities on slide 6. The plates with a lot of blanks (1,12,17 & 18) do have low intensities. Furthermore, the clones from brain tissue on plate 9-12 are, as expected, not very responsive to the test and control samples, which are from liver RNA. *Left*: The log intensities in print order. *Right*: Box plot displaying the median and the variability of the log intensities for each plate using the same ordering and coloring as in the left figure. 7
- 4 Gene expressions (for slide 6) *after* scaled print-tip intensity normalization. *Left*: Print-order plot of the log ratios with a horizontal box plot showing the distribution of the log ratios within each plate group. *Right*: The spatial location of the log ratios. Normalizing for intensity dependent artifacts does indeed remove a lot of the spatial and some of plate effects. 8
- 5 Gene expressions (for slide 6) *after* platewise bias normalization. *Left*: The log ratios versus the log intensities. *Right*: The spatial location of the log ratios. Removing the constant bias within each plate group will remove a lot of the spatial and some of intensity dependent effects. 9
- 6 Intensity dependency for each plate (on slide 6). The top most curve (blue) of plate 12 is shifted upwards because it has only one data point in the upper intensity range, $A = [12, 13]$, with a high value of the log ratio. 10
- 7 Gene expressions (for slide 6) *after* platewise intensity dependent normalization. *Left*: Print-order plot of the log ratios with a horizontal box plot showing the distribution of the log ratios within each plate group. *Right*: The spatial location of the log ratios. 10
- 8 The log ratios for all the 24 replicates of gene 1540 found on the eight slides after performing intensity normalization first for each print-tip group and then for each plate group (red discs) compared to not doing any normalization at all (blue circles). The left scale is for the normalized log ratios and the right one for the non-normalized ones. Note that the unit lengths of the two vertical axes are equal, but their origins have been aligned for simplifying visual comparison. 13

9 Comparison between the sample distributions of the gene log-ratio variabilities after (2×21) different normalization strategies have been applied. Note that the tails of the sample distributions have been cut of at 0.75 and that they do extend (approximately equally far) beyond this level. Legend: none - No normalization, PI - Constant plate bias removed, PI(A) - Intensity dependent plate bias removed, Pr(A) - Print-tip intensity dependent bias removed, sPr(A) - Scaled print-tip intensity dependent bias removed, SI(A) - Slide intensity dependent bias removed. Normalization strategies that start out with background subtracted signals are in addition labeled with 'bg'. For example, PI-sPr(A) refers to a normalization sequence followed by a plate normalization and ended by a scaled print-tip normalization. 14

List of Tables

1 Comparison between different normalization strategies. Legend: none - No normalization, PI - Constant plate bias removed, PI(A) - Intensity dependent plate bias removed, Pr(A) - Print-tip intensity dependent bias removed, sPr(A) - Scaled print-tip intensity dependent bias removed, SI(A) - Slide intensity dependent bias removed. Normalization strategies that start out with background subtracted signals are in addition labeled with 'bg'. 17

Preprints in Mathematical Sciences 2002:28
ISSN 1403-9338
LUTFMS-5027-2002
Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden
<http://www.maths.lth.se/>