Statistical Applications in Genetics and Molecular Biology

Volume 7, *Issue* 1 2008 *Article* 29

Reducing Spatial Flaws in Oligonucleotide Arrays by Using Neighborhood Information

Jose Manuel Arteaga-Salas* Andrew P. Harrison†

Graham J. G. Upton‡

*University of Essex, jmarte@essex.ac.uk

†University of Essex, harry@essex.ac.uk

‡University of Essex, gupton@essex.ac.uk

Reducing Spatial Flaws in Oligonucleotide Arrays by Using Neighborhood Information*

Jose Manuel Arteaga-Salas, Andrew P. Harrison, and Graham J. G. Upton

Abstract

We address the problem of detection and correction of spatial flaws in oligonucleotide microarrays. We present two similar procedures, of which one is intended solely for use with replicates and the other has wider applicability. By constructing a set of replicates, with one realistically flawed, we are able to examine the extent to which our procedures are capable of repairing the flaw. We find that, for this purpose, our procedures are superior to the existing 'Harshlight' procedure.

KEYWORDS: flaws, oligonucleotide arrays, visualisation

*J.M.A-S. is supported by a CONACYT scholarship (178596). We thank the referees of a previous version for their constructive suggestions.

1 Introduction

Oligonucleotide microarrays are widely used to provide information about gene expression. Each microarray provides information in the form of measured light intensities at upwards of half a million locations (arranged on a square grid) on the array surface. Each location corresponds to a probe (a sequence of 25 bases) providing information about a designated gene. The design comprises pairs of "perfect match" (PM) and "mismatch" (MM) probes that differ only in their central base. The intention of the chip manufacturers was that each MM probe would provide a measure of the background value resulting from the outer 24 bases, so that the difference between the PM and MM expression values would provide information concerning the extent of the occurrence of a particular gene. For each gene there is at least one probe set, with (typically) from 11 to 20 probes in a set. The individual values in a probe set are then converted into a single expression value for the gene using one of a number of algorithms (e.g. RMA; Irizarry *et al*. (2003b)). Although these algorithms use robust statistical procedures, and are generally very effective in recovering the signal (Cope *et al.*, 2004), they make no reference to values in other probe sets.

We believe that, to some degree, every oligonucleotide array contains spatial flaws. The most common flaws appear to be the consequence of trapped bubbles and are manifested as rings or arcs (Suárez-Farinas *et al*., 2005a). These are usually seen towards a side of an array (Langdon *et al*., 2008). Irregular shaped blobs may also be found, as are occasional 'scratches'(Upton and Lloyd, 2005). However, standardization algorithms such as quantile scaling do not consider the spatial locations of the values that they process.

In this paper we show that the spatial compactness of flaws can be exploited to provide the basis for corrections that are quite independent of any subsequent chip-wide correction procedure (such as quantile scaling). The methods that we propose involve the comparison of the expression levels of a probe across a number of arrays. The methods will be most effective when the expression levels should be identical (as in technical replicates). They should be nearly as effective when the expression levels are affected only by slight individual variations (as in biological replicates). However, one method can be used when comparing unrelated arrays.

In Section 2 we illustrate the flaws found in examples of replicate arrays and in unrelated arrays. In Section 3 we briefly describe some previously suggested correction routines while, in Section 4, we present two new correction procedures, one suitable for any group of arrays and one suitable only for replicates. In Section 5 we create artificial replicates and present results for these and other arrays in Section 6.

2 Visualising spatial flaws

With three or more arrays (which need not be replicates), visualization of defects is a simple matter. With L_{klr} denoting the logarithm of the signal intensity at location (k,l) in array r, an effective procedure involves calculating and plotting the values of d_{klr} given by:

$$d_{klr} = L_{klr} - M_{kl} \tag{1}$$

where M_{kl} is the logarithm of a reference value for location (k,l). Comparing eight arrays from the Affymetrix HG-U95 Spike-in dataset, Cheng and Li (2005) choose the values in one arbitrarily chosen array to be the set of reference values. Studying two-color arrays Reimers and Weinstein (2005) use the average values (across replicates) in each pixel as the reference values. We recommend using a more robust comparator such as the median value, as used by Suárez-Farinas *et al*. (2005a).

To show the potential sizes and forms of spatial defects, we have chosen three different array types downloaded from the Gene Expression Omnibus (GEO) (available at http://www.ncbi.nlm.nih.gov/geo/). To visualize the flaws we have identified cells in the array for which $|e^{d_{klr}} - 1| > 0.25$ — in other words, those cells that differ by more than 25% of the median array value in that location. Not all cells so identified will correspond to flaws, since there will be many cells whose unusual magnitude is a result of a biological signal. However, these genuinely interesting cells will not be spatially clustered.

Figure 1 shows the flaws found in four biological replicates (GSM149276-9) of the Affymetrix GeneChip Drosophila Genome Array DrosGenome1 (GEO number GSE6515) (see Magalhães *et al*. (2007) for a description of the data context). There are a number of irregular dark blobs (concentrations of low values) visible in nearly every bottom row, while GSM149276 has a small region of high values in the top right corner and GSM149277 has many high values.

DOI: 10.2202/1544-6115.1383

Figure 1: Spatial flaws in four replicates of the GSE6515 experiment hybridized in DrosGenome1 arrays. The upper row shows the locations of unusually large values and the lower row the locations of unusually small values. In each case 'unusually' implies cells that differ by more than 25% of the median array value in that location.

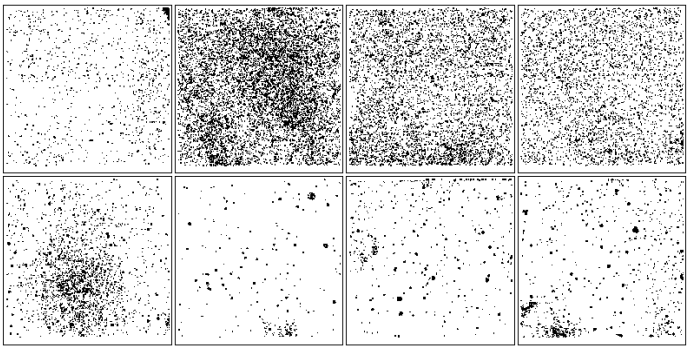


Figure 2 refers to three biological replicates (GSM29929, 29930 and 29932) of the Affymetrix GeneChip Yeast Genome S98 Array YG-S98 (GEO number GSE1723) (see Tai *et al*. (2005) for a description of the data context). Replicate GSM29929 has an extensive region of low values to the left of center as well as other prominent regions of unusually low values. GSM29930 has diffuse regions of high values. GSM29932 has unusually low values towards the bottom of the array.

Figure 2: Three replicates of the GSE1723 experiment hybridized in YG-S98 arrays; interpretation as for Figure 1.

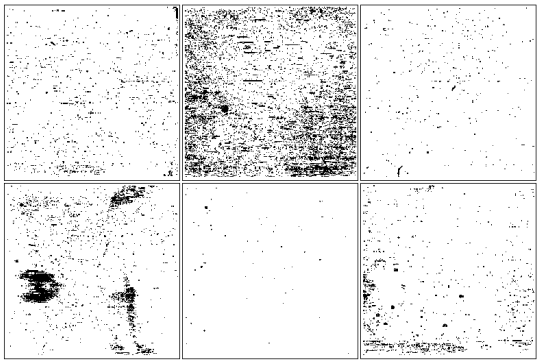
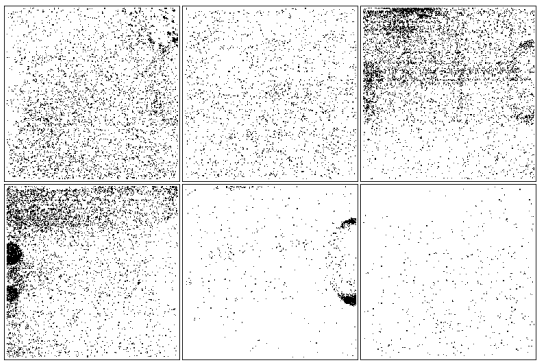


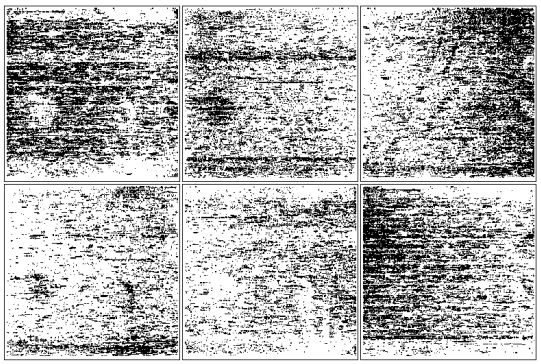
Figure 3 displays results for the three technical replicates of the second of the fourteen hybridizations of the Affymetrix Latin Square experiment (see http://www.affymetrix.com) utilizing the GeneChip Human Genome HG-U133A.

Figure 3: Spatial flaws in three technical replicates of the spiked GeneChip Human Genome U133A arrays; interpretation as for Figure 1.



The first replicate shows "finger prints" corresponding to compact regions of low values, whilst the low values in the second replicate resemble "coffee rings". Isolated high regions appear in the top right of the first replicate and there are left-right high-value linear features in the third replicate.

Figure 4: Spatial flaws uncovered by comparison of three unrelated arrays; interpretation as for Figure 1.



The same approach can be used without replicates, providing that information from several arrays of the same type is available. As an example Figure 4 shows the result of comparing the yeast array GSM29929 (see Figure 2) with yeast arrays from two entirely unrelated experiments: GSM34758 from experiment GSE1938 (see Pitkänen *et al*. (2004) for details) and GSM67551 from experiment GSE3076 (see

DOI: 10.2202/1544-6115.1383

Guan,*et al.* (2006)). The flaws in GSM29929 that were very obvious in Figure 2 are still apparent (though they are much less apparent). The two comparator arrays were chosen essentially randomly, and it is apparent that both have features involving high values (a single part row in GSM34758) and 'tyre marks' in GSM67551.

3 Existing adjustment procedures

We believe that nearly all arrays contain some spatial defects (Langdon *et al*., 2008), though these are often confined to very small regions. We now give very brief descriptions of some procedures that have been suggested for dealing with defects.

3.1 Reimers-Weinstein

Reimers and Weinstein (2005) detected spatial flaws by comparing the values in an array with the corresponding trimmed mean values from a collection of arrays. Having detected flaws their counsel was to either reject the array, or to reject the affected probes.

3.2 Upton-Lloyd

Upton and Lloyd (2005) suggested subtracting from each cell in the array the smallest neighbouring value found in a $(2m+1) \times (2m+1)$ window. With small m this procedure corrects the background intensity levels and removes the majority of spatial flaws. However, the procedure fails because it often distorts the data by choosing as a reference value a neighbour with a genuinely large value.

3.3 Harshlight

The Harshlight package (Suárez-Farinas *et al*. (2005b)) is a tool (freely available in Bioconductor) for the identification and correction of spatial biases in microarrays. The procedure works with the logarithms of the ratio of the value in a chip to the corresponding median value. The variability of these ratios is then decomposed into expected background noise and a probe-specific contribution. By examining the locations where large probe-specific contributions exist, the Harshlight procedure provides summaries of the proportions of an array that contain so-called Compact, Diffuse, and Extended defects. The user specifies whether flawed values should be replaced with 'N/A' (Not Available) or with the median value of the replicates at that location.

3.4 caCORRECT

The caCORRECT scheme introduced by Stokes *et al* (2007) is intended for use when there is data available from many arrays. Each probe is normalized by reference to the variance and standard deviation of the other probes in that position, and the normalized values are visualised as 'heatmaps'. Users are invited to make their own choices of probes to be corrected, or to use an inbuilt 'batch mode'. The corrections to be made are left to the user, with substitution of the mean or median probe value suggested as possibilities. Their own choice would be "to not replace data in any way" but simply to ignore data beieved to be incorrect.

4 The new adjustment routines

We now introduce two related adjustment procedures that seek to use information about the extent and magnitude of a flaw to produce a revised value that remains unrelated to the other values at that array location.

A seriously flawed value will lead to an extreme d-value, and Figures 1 to 3 indicated that d is a useful statistic. However, since the magnitudes of the PM-values are often very different to those of the MM-values, the magnitudes of the d-values for the PM-probes may have a different scale to those of the d-values for the MM-probes. The new adjustment routines use *standardized* d-values, d^*_{klr}, given by:

$$d^*_{klr} = \frac{d_{klr}}{S_{kl}} \tag{2}$$

where S_{kl} is the standard deviation of the L_{klr} values.

We propose two different adjustment routines which can be used separately or in sequence. Both routines work by comparing values across three or more arrays and assume that, at any specific location, at most one of the arrays being compared will be flawed. This is reasonable since Langdon *et al*. (2008) report that, for most array types spatial flaws affect between 1% and 3% of cells, while studies using replicates rarely involve more than three replicate arrays. However, in the (hopefully) rare instance where an instrument malfunction results in identical errors across all arrays, the methods described below would fail to detect the problem.

4.1 Local probe effect (LPE) adjustment

This adjustment can be used whenever multiple arrays (which need not be any type of replicates) are available. It uses the entire spatial structure in the region immediately surrounding a probe to decide whether adjustment should take place. If the

DOI: 10.2202/1544-6115.1383

decision is that a cell should be scaled, then the procedure begins with the calculation of the d-values using Equation (1). To explain the LPE procedure we begin by defining the quantities I_{kl} and G_{kl} as follows:

I_{kl} The identifier of the array corresponding to the case where d_{klr} takes its largest absolute value.

G_{kl} This takes the value 1 if the d-value with the largest magnitude (at this location) was positive, and is otherwise equal to -1.

Using these two quantities, define the code E_{kl} by:

$$E_{kl} = I_{kl} \times G_{kl} \tag{3}$$

so that, with R arrays, E_{kl} takes one of the values $\{-R, -(R-1), ..., -2, -1, 1, 2, ..., (R-1), R\}$. In an area with no spatial flaws, a location will be equally likely to be associated with any of these $2R$ possible codes.

To determine whether a flaw affects location (k,l), we consider the codes of the informative locations in the 5×5 window centered on this location. Here, we use the term *informative locations* to mean locations containing either a PM or MM probe (as opposed to locations used for quality control or locations outside the array). If the window contains an unusually large number of locations having the same code, then we conclude that this is due to a spatial flaw. If a majority of the informative locations display the same E-value, (corresponding to array r, say) then we consider adjusting the value in cell (k,l,r). Note that the E-value for the central location need not be the majority value. The probability of making the adjustment as a result of a chance arrangement of E-values depends on the number of informative cells, N, and the number of arrays, R. This is a binomial situation with parameters N and $1/2R$, with the probability of interest being the probability of a value greater than $N/2$. For the case $N = 24$ (omitting the central square of the window),with $R = 3$, this probability is about 3×10^{-5}.

With array r identified for correction, let Z be the subset of the N informative locations within the window. For each location in Z calculate the d^*-values for array r using Equations (1) and (2), and let $\bar{d}^*$ be their average. This value will be used to correct the original value in location (k,l,r) providing the signs of $\bar{d}^*$ and d^*_{klr} are the same. If this is not the case then the value will be left unadjusted. Working with logarithms, the adjusted value, L^a_{klr}, is given by

$$L^a_{klr} = L_{klr} - S_{kl}\bar{d}^* \tag{4}$$

where, as before, S_{kl} is the standard deviation of the L_{klr} values.

The procedure will work best with biological or technical replicates, but, if it were suspected that all replicates were affected by a similar error, then it could be used (albeit with reduced efficiency, because of the presence of different biological signals in the different arrays) with unrelated arrays.

4.2 Complementary probe pair (CPP) adjustment

This adjustment is only suitable for use with replicates and relies upon the fact that each PM probe is situated directly beside its matching MM probe. The assumption is that, because the probes are adjacent, any spatial flaw that affects one of the pair will also affect the other member of the pair. For this adjustment routine each (PM, MM) pair is treated separately and, with R arrays being compared, it results in at most one of the R PM-values being altered, and at most one of the R MM-values being altered.

Suppose that cells (k,l) and $(k,l+1)$ are the locations of the PM and MM probes, respectively. Using M_{kl}, the median of the R PM-values at cell (k,l), the values of $d^*_{kl1}, d^*_{kl2}, \ldots, d^*_{klR}$ are determined using Equations (1) and (2). The d^*-values are also calculated for the neighboring MM probes. The array in need of correction is indicated by the d^*-value with the greatest absolute magnitude, though correction only takes place if both d^*-values for that array have the same sign (indicating that both members of the (PM, MM) pair are unusually large, or unusually small, compared to the corresponding values in the other arrays).

Suppose that the values in array r have been identified for correction. Denote the logarithms of the adjusted values in locations (k,l) and $(k,l+1)$ by L^a_{klr} and $L^a_{k(l+1)r}$, respectively. The adjustment at cell (k,l) is calculated from the R initial values at location $(k,l+1)$ and the adjustment at cell $(k,l+1)$ is calculated from the initial values at location (k,l). The revised values are given by:

$$L^a_{klr} = L_{klr} + \frac{S_{kl}}{S_{k(l+1)}}(M_{k(l+1)} - L_{k(l+1)r}) \quad (5)$$

$$L^a_{k(l+1)r} = L_{k(l+1)r} + \frac{S_{k(l+1)}}{S_{kl}}(M_{kl} - L_{klr}) \quad (6)$$

In each equation the bracketed term quantifies (on a logarithmic scale) the difference between the value in array r and the typical (i.e. median) value. Even on a logarithmic scale the variability of PM-values at a location is not the same as that of the corresponding MM-values, so that the difference is scaled up (or down) to match the scale of the data to which it is being applied.

4.3 Example of the adjustments

Table 1(a) shows a typical set of paired PM and MM values for a case where $R = 3$. Table 1(b) shows their natural logarithms from which we determine the median values $M_{kl} = 4.93$ and $M_{k(l+1)} = 4.61$, and the standard deviations $S_{kl} = 0.18$, and $S_{k(l+1)} = 0.10$. Using Equations (1) and (2) leads to the d^*-values in Table 1(c); for example, $d^*_{kl1} = (5.11 - 4.93)/0.18 = 1.01$.

DOI: 10.2202/1544-6115.1383

Table 1: Example of the adjustment of PM and MM values. (a) The observed values ($R = 3$). (b) The logarithm of the observed values L_{klr}. (c) The corresponding d^*-values. (d) The adjusted values.

	Array			Array		
	1	2	3	1	2	3
	(a) Observed values			(b) L_{klr} values		
PM	165	116	138	5.11	4.75	4.93
MM	114	94	101	4.74	4.54	4.61
	(c) d^*-values			(d) Adjusted values		
PM	1.01	-0.99	0	**132**	116	138
MM	1.26	-0.72	0	**103**	94	101

The d^*-value with the greatest absolute magnitude is in array 1. Since both d^* values in this array have the same sign, an adjustment is made. Since the d^*-values are positive, the adjustment results in reduced values for array 1. These are shown in Table 1(d) where, for example,

$$132 = \exp\left\{5.11 + \frac{0.18}{0.10}(4.61 - 4.74)\right\}$$

Table 2(a) gives the E-values for the 5×5 array centered on the probe considered in Table 1. Table 1 (c) demonstrated that the largest absolute d^* value was in array 1, and was positive, so that the central value is shown as 1. However, it is not the value in replicate 1 that is adjusted downwards, but the value in replicate 2 that is adjusted upwards. This is because the majority of E-values in the table are "-2", indicating low values in array 2. Table 2 (b) shows the corresponding d^* values.

The average of the 19 d^*-values, $\bar{d}^*$, is -1.45. We have seen that the d^*-value for the central cell in array 2 has the same sign, so that the adjustment can be made. The revised value is:

$$L^a_{kl2} = \ln(116) - (0.18)(-1.45) = 5.01.$$

Exponentiating gives a revised value of 150 for the value in array 2 (compared with the original 116 and the values in the other arrays of 165 and 138).

4.4 Combining adjustment routines

Note that, since the transformation has the effect of moving an extreme towards the median, it is always the case that the adjusted values are less variable than the original values.

Table 2: Example of the adjustment of a probe. (a) The E-values for a 5×5 window. (b) The d^*-values for the locations affected by the spatial bias.

(a) E-values	**-2**	**-2**	**-2**	**-2**	**-2**
	-2	**-2**	**-2**	**-2**	**-2**
	-1	**-2**	1	1	1
	2	**-2**	**-2**	**-2**	**-2**
	3	**-2**	**-2**	**-2**	**-2**
(b) d^*-values	-1.41	-1.48	-1.53	-1.49	-1.42
	-1.69	-1.57	-1.62	-1.14	-1.15
		-1.72			
		-1.38	-1.65	-1.32	-1.04
		-1.27	-1.49	-1.61	-1.59

At every pair of (PM, MM) locations, the values in one of the R arrays will be adjusted providing the sign of the d^* value for the PM cell in the identified array is the same as the sign for the MM cell in that array. If there are no spatial flaws, then the proportion of cells adjusted is approximately $(R-1)/2R$.

Although each adjustment routine uses values in the local neighborhood, their definitions of neighborhoods are very different and this suggests that using the two procedures in succession could be beneficial. In this section the arrays compared are replicates since CPP is not appropriate for other arrays.

5 Creation of test replicates

In order to demonstrate that the methods correctly identify flawed cells and make appropriate corrections, we create a realistic set of replicate arrays, in which one has known spatial flaws. To obtain realistic flaws, we used the results from the analysis (reported later) of the Human Genome U133 arrays from the Latin Square Experiment. Writing L as the logarithm of the original value in the first (left-hand) array of those illustrated in Figure 3 and with L^a as the logarithm of the adjusted value, we calculated error array **E**, with elements $\{E_{kl}\}$ defined by

$$E_{kl} = L_{kl} - L^a_{kl}. \tag{7}$$

There are four stages in our construction of a set of test replicates:

1. We chose (arbitrarily) the first replicate of the third of the hybridization sets of Human Genome U133A Latin Square experiment as our base array, **A**, and

DOI: 10.2202/1544-6115.1383

denote the log-expression measurements for this array by $\{A_{kl}\}$. We work with three copies of this array, denoted as $\mathbf{B_1}$, $\mathbf{B_2}$ and $\mathbf{B_3}$.

2. We now incorporated both additive and multiplicative random errors to the arrays in a manner suggested by the model of Durbin and Rocke (2003):

$$L_{kl} = \log(b + \mu e^{\eta_{kl}} + \varepsilon_{kl}) \tag{8}$$

where b is the background intensity, μ is an expression intensity, and the $\{\eta\}$ and $\{\varepsilon\}$ are random errors. The additive error $\varepsilon \sim N(0, \sigma_\varepsilon)$ reflects the variability of the background, and the multiplicative error $\eta \sim N(0, \sigma_\eta)$ represents the proportional errors most noticeable with highly expressed genes.

Durbin and Rocke (2003) provide a method for estimating σ_η and σ_ε using replicate arrays. Using this method we estimated the values of σ_η and σ_ε for five sets of Human Genome U133A replicates from the same Latin Square experiment (chosen at random, but excluding the set analyzed in Section 3). The medians of these five estimates were $\sigma_\eta = 0.076$ and $\sigma_\varepsilon = 3.235$ and it is these median values that we used to generate the normally distributed random errors for each cell of the arrays $\mathbf{B_1}$, $\mathbf{B_2}$ and $\mathbf{B_3}$.

3. Replicates vary somewhat in their average intensities. Based on our analysis of the five sets of U133A replicates studies at stage 2, we revised the values in $\mathbf{B_2}$ by multiplying each by 1.027 and the values in $\mathbf{B_3}$ by multiplying each by 0.973.

4. As a final stage the error matrix $\mathbf{E}$ was added to $\mathbf{B_1}$. This new array containing the spatial flaws will be denoted as $\mathbf{C_1}$. We create arrays $\mathbf{C_2}$ and $\mathbf{C_3}$ by copying arrays $\mathbf{B_2}$ and $\mathbf{B_3}$, respectively.

6 Results

6.1 The Harshlight test

The Harshlight package analyzes a set of replicate arrays and produces a report with details about the types of flaws found (described as 'Extended', 'Compact' or 'Diffuse'). In the tables that follow 'HMS' refers to the substitution of flawed values by the median value of the replicates at that location.

Table 3 shows the percentages of spatial defects reported by Harshlight for the four arrays of the GSE6515 experiment used in Figure 1 (where, for example, 'CPP+LPE' means one application of the CPP procedure is followed by one

Table 3: Percentages of defects reported by the Harshlight package for four Drosophila arrays (see Figure 1).

Array	1			2			3			4		
	E	C	D	E	C	D	E	C	D	E	C	D
Original	15.6	0.6	0.3	4.5	0.7	2.3	4.3	0.4	1.0	1.2	0.5	1.2
CPP	0.7	0	0	2.4	0	1.5	2.9	0	0.4	0.3	0	0
LPE	1.2	0	0	3.8	0.1	1.8	4.0	0	0.8	0.3	0.1	0.9
CPP+LPE	0.7	0	0	2.3	0	1.5	2.8	0	0.4	0.3	0	0
LPE+CPP	0.5	0	0	2.1	0	0.3	2.6	0	0.4	0.1	0	0
HMS	18.2	0	0	6.6	0	0	4.3	0	0.8	0.8	0.1	0.4
HMS twice	18.2	0	0	6.7	0	0	4.4	0	0	0.8	0	0
HMS thrice	18.2	0	0	6.7	0	0	4.4	0	0	0.8	0	0

of the LPE procedure). The Table shows the percentages for the original data, for the spatial normalized data using the procedures presented in this work, and for the data normalized with HMS (applied once, twice, or thrice). Originally the replicates have as many as 16.5% of locations that are subject to spatial flaws. This is reduced to less than 6% by the use of LPE and to less than 3% when both procedures are used. The HMS procedure is much less effective: in three cases it results in an increase in the proportion of cells that are judged to be part of an Extended defect. It is also noteworthy that using HMS repeatedly does not help matters.

The most effective combination is LPE followed by CPP. This is logistically convenient, since, while LPE can be used with any group of arrays, CPP can only safely be used with replicates. Our LPE+CPP algorithm is therefore informed by the user as to whether it is appropriate to carry out the CPP procedure.

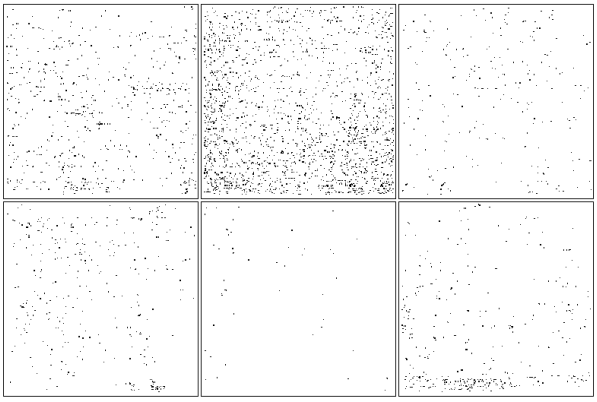
Table 4: Percentages of defects as reported by the Harshlight package for: (a) three Yeast arrays (see Figure 2); and (b) three spiked HGU133A arrays (see Figure 3).

	Array	1			2			3		
		E	C	D	E	C	D	E	C	D
(a)	Original	15.1	0.2	9.0	5.8	0.4	5.1	2.3	0.3	6.7
	LPE+CPP	0.6	0.2	1.0	0.4	0.1	1.9	0.1	0.1	3.7
	HMS thrice	6.8	0	0	6.9	0	0	3.0	0	0
(b)	Original	5.0	0	10.0	5.0	0.3	1.6	2.6	0	3.1
	LPE+CPP	0	0	1.0	0	0	0.6	0	0	0
	HMS thrice	0.2	0	0	1.7	0	0	2.0	0	0

DOI: 10.2202/1544-6115.1383

The analysis presented in Table 3 was repeated for the arrays illustrated in Figures 2 and 3 with confirmatory results (see Table 4). We also applied LPE to the three (non-replicate) yeast arrays with the results shown in Figure 5: little trace remains of the distinctive features seen previously.

Figure 5: Spatial flaws remaining in the three yeast arrays. The layout is as in Figure 2.



6.2 Performance with known flaws

Table 5: Results of applying the LPE+CPP or HMS (twice) corrections to three simulated arrays, with array 1 containing flaws.

Cell type	Revision	% of these cells adjusted	Initial($\mathbf{C}-\mathbf{B}$) Mean	Initial($\mathbf{C}-\mathbf{B}$) RMSE	Final($\mathbf{D}-\mathbf{B}$) Mean	Final($\mathbf{D}-\mathbf{B}$) RMSE
Array $\mathbf{C}_1$						
Flawed	LPE+CPP	100%	-0.056	0.159	-0.023	0.117
	HMS(twice)	100%			-0.035	0.137
Not Flawed	LPE+CPP	10%	0	0	0.004	0.120
	HMS(twice)	4%			0.002	0.113
Arrays $\mathbf{C}_2$ and $\mathbf{C}_3$						
Not Flawed	LPE+CPP	13 %	0	0	-0.005	0.133
	HMS(twice)	0.3 %			-0.034	0.134

We applied the LPE+CPP procedure and the HMS procedure twice (since LPE+CPP is a combined adjustment) to the arrays $\mathbf{C_1}$, $\mathbf{C_2}$ and $\mathbf{C_3}$ to obtain corrected ar-

rays $\mathbf{D_1}$, $\mathbf{D_2}$ and $\mathbf{D_3}$. If the corrections were perfect then each of these would equal the initial array $\mathbf{A}$, with $\mathbf{D_1} - \mathbf{C_1} = \mathbf{E}$. However, since both additive and multiplicative errors have been introduced, exact equality will not be achieved. Table 5 summarizes the results of applying both adjustments. To estimate the amount of bias remaining in the arrays after any correction method the Root Mean Squared Error (RMSE) was used.

In array $\mathbf{C_1}$ 24% of the locations were contaminated with spatial flaws, all of which were adjusted by both procedures. The table shows that using LPE+CPP resulted in the mean bias being reduced by 58% with the RMSE being reduced by 26%. Applying HMS(twice) was less effective (the reductions were by 38% and 14%, respectively).

Table 5 also shows that LPE+CPP adjusted far more of the cells affected only by random noise (see, especially, the results for arrays $\mathbf{C_2}$ and $\mathbf{C_3}$). That this is not a bad thing is born out by the final column of the table which demonstrates that the net effect was a set of values that better resembled the values prior to the addition of noise.

6.3 Testing using Affycomp

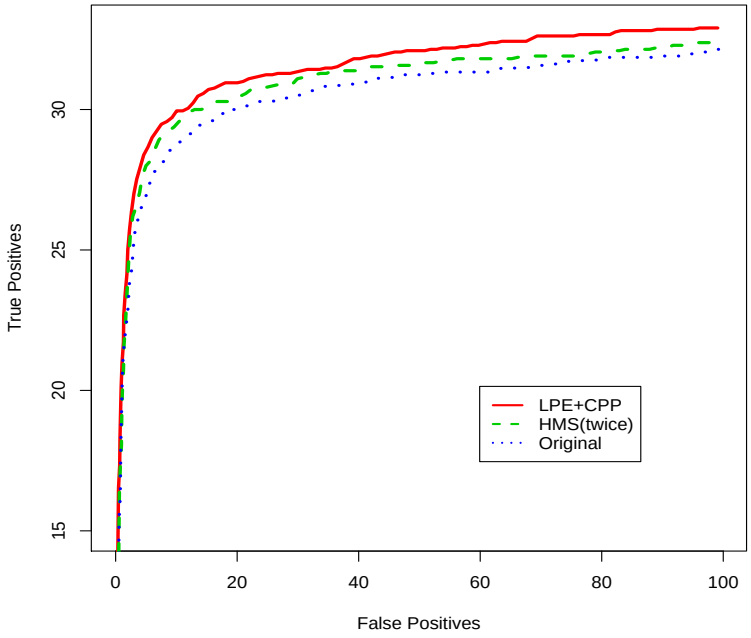
To show the effect of our normalization procedure in gene expression measurements and that the effects of probe level artifact correction persist through probeset summary, the comparison tool Affycomp (Cope *et al.* (2004)) (available as part of Bioconductor) will be used. This package contains a set of graphical tools for summaries of Affymetrix probe level data. Using the plots available in Affycomp, a comparison of expression measurements with different summarization methods can be obtained.

The microarray data used for these comparisons were the spiked-in arrays that formed part of the Latin Square experiment using Human Genome U133A chips made publicly available by Affymetrix. The fourteen hybridization sets were separately pre-processed with HMS(twice) or with LPE+CPP and then summarized with the popular algorithm RMA (Irizarry *et al*. (2003b)) (using the suggested default setting).

Figure 6 shows the Receiver Operator Curves (ROC) generated by Affycomp for the SpikeIn133A dataset (with 2-fold changes) summarizing the true positive/false positive behavior of the data summarized with RMA. By reducing the noise in the system, the fold-changes between the pairs of fold-change replicates should become more apparent. Thus, when all the genes are arranged in order of apparent fold change their expression level, we would expect the genuine changes due to the spike-ins to become more apparent (i.e. to appear higher in the ordered list), whereas the artificial changes of magnitude due to random variation and spatial

DOI: 10.2202/1544-6115.1383

Figure 6: ROC Curves for the SpikeIn133A dataset at 2-fold changes using RMA.



artifacts should be reduced in size and should move down the ordered list.

Table 6: Changes in ranking of the genes in the spike-in experiment, following correction with either Harshlight or LPE+CPP

	Harshlight (twice)		LPE+CPP	
	Increased rank	Decreased rank	Increased rank	Decreased rank
Spike-ins	36.5%	33.3%	45.7%	38.6%
Others	50.5%	49.4%	50.3%	49.7%

Table 6 compares the results of using Harshlight (twice) with those obtained using LPE+CPP. With both procedures the effect of removing spatial artifacts reduces the ranks of some genes very considerably with the result that more genes have an increased ranking than have a decreased ranking. This is most marked for the LPE+CPP procedure and the spike-in genes, reflecting the improved performance seen in Figure 6.

7 Summary and discussion

We have presented two procedures that are very effective at identifying and correcting spatial flaws in oligonucleotide microarrays. The CPP procedure is only appropriate with replicates, but the LPE procedure can be used with any group of arrays (though, of course, it will be most effective with replicates). We employed the LPE procedure on the trio of unrelated arrays shown in Figure 4. In all 22% of cells received an adjustment. As one would anticipate, the flawed areas were largely removed, though the picture remained very noisy (reflecting the genuine differences between the experimental situations).

In correcting any particular cell, neither adjustment uses information from that cell in the comparator arrays. We have compared the performance of our procedures with that of the Harshlight package (Suárez-Farinas *et al*. (2005b)) and found our new procedures to be more effective.

References

Cheng, C. and Li, L. M. (2005) Sub-array normalization subject to differentiation. *Nucleic Acids Res.*, **33**, 5565-5573.

Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z., Speed, T. P. (2004). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**(3), 323-331

Durbin, B. and Rocke, D. M. (2003) Estimation of transformation parameters for microarray data. *Bioinformatics*, **19**, 1360-1367.

Guan, Q., Zheng, W., Tang, S., Liu, X., Zinkel, R., Tsui, K-W., Yandell, B. S., Culbertson, M. R. (2006) Impact of Nonsense-Mediated mRNA Decay on the Global Expression Profile of Budding Yeast. PLoS Genet.; **2** (11), e203.

Ekstrøm, C. T., Bak, S. and Rudemo, M. (2005) Pixel-level signal modelling with spatial correlation for two-colour microarrays. *Stat. Applic. Gen. Mol. Biol.*, **4**, 1-14.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. and Speed, T. P. (2003b) Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Res.*, **31**, e15.

Kim, K., Page, G. P., Beasley, T. M., Barnes, S., Scheirer, K. E. and Allison, D. B. (2006) A proposed metric for assessing the measurement quality of individual microarrays. *BMC Bioinformatics*, **7**, 35.

DOI: 10.2202/1544-6115.1383

Langdon, W. B., Upton, G. J. G., da Silva Camargo, R. and Harrison, A. P. (2008) A Survey of Spatial Defects in Homo Sapiens Affymetrix GeneChips, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, submitted.

Li, C. and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* **98**, 31-36.

Pitkänen, J-P, Törmä, Alff, S., A., Huopaniemi, L., Mattila, P., Renkonen, R. (2004) Excess mannose limits the growth of phosphomannose isomerase PMI40 deletion strain of Saccharomyces cerevisiae. *J. Biol. Chem.* **279**, 55737-43.

Magalhães, T., Palmer, J., Tomancak, P., Pollard, K. S. (2007) Transcriptional control in embryonic Drosophila midline guidance assessed through a whole genome approach. *BMC Neuroscience*, **8**:59

Reimers, M. and Weinstein, J. N. (2005) Quality assessment of microarrays: Visualisation of spatial artifacts and quantitation of regional biases. *BMC Bionformatics*, **6**, Art. 166.

Stokes, T. H., Moffitt, R. A., Phan, J. H. and Wang, M. D. (2007) *c*hip *a*rtifact *CORRECT*ion (caCORRECT): A bioinformatics system for quality assurance of genomics and proteomics array data. *Ann. Biomed. Eng.*, **35**, 1068-1080.

Suárez-Fariñas, M., Haider, A. and Wittkowski, K. M. (2005) "Harshlighting" small blemishes on microarrays. *BMC Bioinformatics*, **6**, Art. 65.

Suárez-Fariñas, M., Pellegrino, M., Wittkowski, K. M. and Magnasco, M. (2005) Harshlight: a "corrective make-up" program for microarray chips. *BMC Bioinformatics*, **6**, Art. 294.

Tai, S. L., Boer, V. M., Daran-Lapujade, P., Walsh, M. C. *et al*. (2005) Two-dimensional transcriptome analysis in chemostat cultures. Combinatorial effects of oxygen availability and macronutrient limitation in Saccharomyces cerevisiae. *J. Biol. Chem.* **280**(1), 437-447.

Upton, G. J. G. and Lloyd, J. C. (2005). Oligonucleotide arrays: Information from replication and spatial structure. *Bioinformatics*, **21**, 4162-4168.

Yuan, D. D. and Irizarry, R. A. (2006) High-resolution spatial normalization for microarrays containing embedded technocal replicates. *Bioinformatics*, **22**, 3054-3060.