Published online 7 December 2006 *Nucleic Acids Research, 2007, Vol. 35, No. 2* ***e8***
doi:10.1093/nar/gkl871

Positional artifacts in microarrays: experimental verification and construction of COP, an automated detection tool

Haiyuan Yu[1], Katherine Nguyen[2], Tom Royce[1], Jiang Qian[3], Kenneth Nelson[2], Michael Snyder[2] and Mark Gerstein[1,4,5,*]

[1]Department of Molecular Biophysics and Biochemistry, [2]Department of Molecular, Cellular and Developmental Biology, Yale University, CT 06520, USA, [3]Wilmer Institute, Johns Hopkins School of Medicine, Baltimore, MD 21287, USA, [4]Department of Computer Science and [5]Program in Computational Biology and Bioinformatics, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA

Received January 9, 2006; Revised July 3, 2006; Accepted October 5, 2006

ABSTRACT

Microarray technology is currently one of the most widely-used technologies in biology. Many studies focus on inferring the function of an unknown gene from its co-expressed genes. Here, we are able to show that there are two types of positional artifacts in microarray data introducing spurious correlations between genes. First, we find that genes that are close on the microarray chips tend to have higher correlations between their expression profiles. We call this the 'chip artifact'. Our calculations suggest that the carry-over during the printing process is one of the major sources of this type of artifact, which is later confirmed by our experiments. Based on our experiments, the measured intensity of a microarray spot contains 0.1% (for fully-hybridized spots) to 93% (for un-hybridized ones) of noise resulting from this artifact. Secondly, we, for the first time, show that genes that are close on the microtiter plates in microarray experiments also tend to have higher correlations. We call this the 'plate artifact'. Both types of artifacts exist with different severity in all cDNA microarray experiments that we analyzed. Therefore, we develop an automated web tool—COP (*CO*rrelations by *P*ositional artifacts) to detect these artifacts in microarray experiments. COP has been integrated with the microarray data normalization tool, ExpressYourself, which is available at http://bioinfo.mbb.yale.edu/ExpressYourself/. Together, the two can eliminate most of the common noises in microarray data.

INTRODUCTION

cDNA microarray technology has enabled us to simultaneously measure expression levels of tens of thousands of genes (1,2). Many prior microarray analyses have focused on inferring functional relationships from gene expression clusters (3,4). The important assumption behind these analyses is 'guilt-by-association', i.e. co-expressed genes tend to share similar functions (5). However, we recognized that when analyzing the raw expression data, the correlation in gene expression might embody not only true biological effects, but also a significant component related to artifacts in chip design.

Printing a microarray chip

As depicted in Figure 1, in microarray experiments, the DNA samples are first prepared in 96-well (or 384-well) microtiter plates. The printing robot then transfers the DNA samples from the plates to the microarray chips using its 2 × 2 printing 2 printing tips (the organization of the printing tips on the robot might be different. For example, many robots have 4 × 4 or 4 × 12 printing tips. But the underlying principle is the same). The printing tips will then be cleaned and be used to transfer the next four DNA samples. Therefore, there are two important observations based on the printing procedure of the microarray chips:

(i) There are multiple blocks on a microarray chip. Spots within the same block are printed by the same printing tip;
(ii) Corresponding spots in different blocks are usually neighbors on the 96-well plate, even though they are far away on the chip.

Importance of our analysis

Nowadays, microarray experiments are widely used to monitor genome-wide gene expression and have recently

*To whom correspondence should be addressed. Tel: + 203 4325405; Fax: + 413 4102140; Email: mark.gerstein@yale.edu

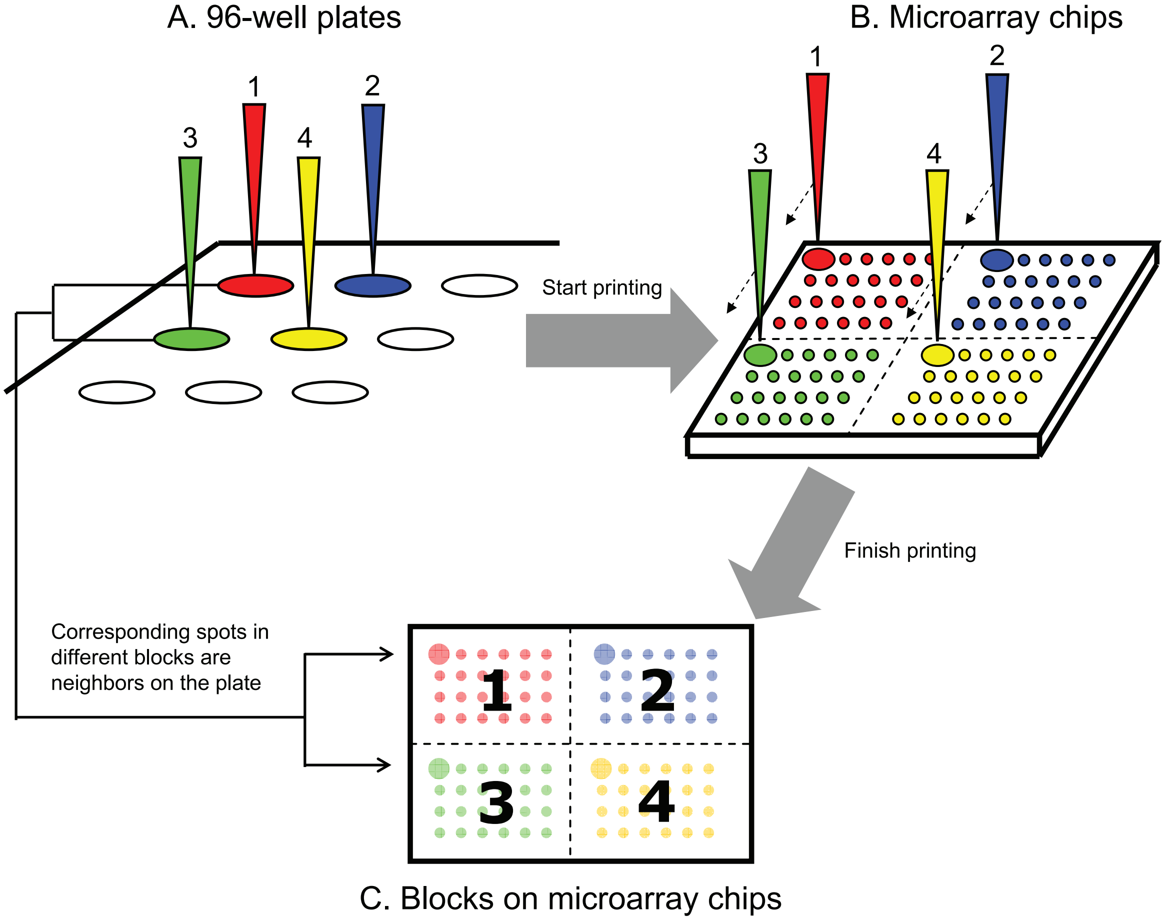

A. 96-well plates
1
2
3
4
Start printing
B. Microarray chips
1
2
3
4
Finish printing
Corresponding spots in different blocks are neighbors on the plate
1
2
3
4
C. Blocks on microarray chips

Figure 1. Illustration of experimental procedures to produce a microarray chip (**A**–**C**). Each color represents one printing tip. The spots in the same color are all transferred by the same tip of the corresponding color.

been implemented to study DNA–protein interactions and tissue-specific gene expression (6–8). It is important to draw the attention of microarray practitioners to the fact that biology and artifact contributions are intermingled in a nontrivial way. This introduces a challenge for separating them by improving the experimental design and data analysis. In particular, there is a growing interest to study the relationship between gene expression and chromosomal proximity (7,9). These studies have overlooked the fact that in many microarray chips DNA spots were printed in an order related to the gene order on the chromosomes. In this paper, we will show two types of positional artifacts creating spurious correlations between genes that simply happen to be close on the microarray chips or microtiter plates, named as chip and plate artifacts, respectively. We are also able to show that these artifacts are generic to most microarray experiments, which study different organisms and use different experimental protocols. The chip artifact only produces 0.1% noise for fully-hybridized spots, but it could be much more problematic (up to 93% noise) for partially-hybridized ones. Careful control and consideration of the artifacts in future microarray analysis is therefore necessary, before any biological conclusions can be made. To this end, we developed an automated web tool, COP (*CO*rrelations by *P*ositional artifacts), which, together with ExpressYourself, can detect and reduce these artifacts in microarray experiments.

MATERIALS AND METHODS

Microarray experiments

Experimental data were obtained from microarray slides printed in the Yale MCDB array facility. The array slides were Corning UltraGAPS and were printed using a BioRad ChipWriter Pro fitted with TeleChem SMP3 pins. The DNA printed were two oligos, A and B (see Table 1). A third oligo, named A′, is the reverse complement of A and was synthesized with a Cy5 end label. After the slides were printed with oligos A and B, they were hybridized with Cy5-labeled A′ to detect the presence of the printed oligos. Hybridizations were done overnight at 43°C in a Maui hybridization station. Hybridized slides were scanned in an Axon Instruments GenePix 4000 scanner. Both test and control chips were produced together in a single experiment to eliminate systematic differences between experiments.

Normalization of microarray data

Even though both test and control chips were produced in the same experiment, a number of variables (e.g. laser strength of the scanner to obtain intensity measurements) might lead to systematic differences in the intensity measurements of the spots on microarray chips (10,11). Therefore, the intensities between test and control chips were normalized before the

Table 1. Probes used in the microarray experiments

Name	Sequence (written 5′–3′)
A	CTGTACCATGGTCCAAGCTCAATTGGAACAACGTAATCCATACGAGTCAGATGAAGAAGCTCACGGAGGT
A′	ACCTCCGTGACGTTCTTCATCTGACTCGTATGGATTACGTTGTTCCAATTGAGCTTGGACCATGGTACAG
B	AGGAGAAGCTGCGACGCTGGAATTTCGGAATAATTAATTATCCTCCACAAGGCTCTCGTGTTTATTGTGT

comparison was performed. Our normalization procedure is based on Quackenbush's method which was reviewed in (11). Because only A spots will hybridize with Cy5-labeled A′ probes and there are the same number of A spots (80 in total) on both test and control chips, the normalization factor N is first calculated as:

$$N = \frac{\sum_{i=1}^{80} \mathrm{AT}_i}{\sum_{i=1}^{80} \mathrm{AC}_i},$$

where AT_i and AC_i represent the ith A spot on the test and control chips, respectively.

Then, each element on the control chip was normalized as:

$$C_i' = C_i \times N.$$

In this way, the mean intensities of the test and control chips are equal.

RESULTS AND DISCUSSION

Artifact related to positions on microarray chips—chip artifact

Four widely-used yeast microarray datasets were first examined:

(i) Spellman-alpha, the alpha-factor arrested cell cycle dataset from Spellman *et al*. (12);
(ii) Spellman-cdc15, the cdc15 arrested cell cycle dataset from Spellman *et al*. (12);
(iii) Zhu-alpha, the alpha-factor blocked cell cycle dataset from Zhu *et al*. (13);
(iv) Diauxic-shift, the diauxic shift dataset from DeRisi *et al*. (14);

Previous studies have shown that there are chip-related artifacts within microarray experiments (15–18). In particular, Balázsi *et al*. (16) suspected that the artifacts might be introduced by the printing tips. Here, we revisited this problem from a different angle: to explore the microarray chip artifact in a straightforward fashion, we calculated the distribution of the average expression correlation coefficient as a function of the distance of gene pairs on the chip (Figure 2A). We calculated Pearson correlation coefficients of log intensity ratios (LIRs) between different genes. The correlation was calculated using all arrays with the same layout in an experiment. All the control spots on the microarray chips were excluded from our analysis. Furthermore, because genes that are close on the chromosome tend to be co-expressed, all the gene pairs that are within 10 open reading frames (ORFs) away were also removed from our analysis to eliminate any possible biological interpretation of our results. The figure clearly shows that the closer the gene pairs on the microarray chips, the higher the average correlation coefficient, suggesting that at least a fraction of the observed correlation in microarray experiments is due to artifacts. We call this a chip artifact. Naively one could expect that chip artifacts in microarray experiments would be canceled when calculating a ratio of the sample and reference cells. This is certainly not the case if the noise at each microarray spot is not multiplicative.

Because of the nature of this chip artifact, we suspected that the artifact may stem from the fact that the printing tips are not cleaned completely; therefore, when printing a spot on the chip, it carries over some of the DNA samples from the previous one. As a result, each spot will, to some degree-depending on the severity of the carry-over, hybridize to the DNA probes complementary to the DNA samples of the previous spot. The signal of each spot thus has an artificial component related to the previous one, producing the chip artifact observed in Figure 2A. To confirm this hypothesis, we examine the chip artifact along X- and Y-directions of microarray chips, separately (the X-direction is the printing direction in our analysis). If our hypothesis holds, one would expect to see that the chip artifact is more severe along X-direction. Figure 2B shows that, within the same printing block, average correlation coefficients along the X-direction (i.e. the printing direction) are significantly higher (P-values $< 10^{-8}$) than those along the Y-direction. The results are in good agreement with our expectation. Interestingly, one might notice that the difference between X- and Y-directions becomes larger at longer distance (peaks around the distance of 30 spots), which cannot be explained by the carry-over artifact alone. However, since the effective distance of the artifact is ~30 spots as discussed below, this increasing difference might be due to some combinatorial effects of the chip artifact and other biological and/or artificial reasons. Furthermore, this phenomenon was not observed in other three panels of Figure 2B, confirming that it is not a general artifact.

Origin of the chip artifact confirmed by experiments

To further confirm the validity of our hypothesis—the chip artifact results from the carry-over during the printing process, we performed a series of microarray experiments to create spots with and without carry-over contaminations. Three different DNA probes were used: A, A′ and B. Probes A and A′ are complementary to each other, whereas probe B is not complementary to either of them (Table 1).

The basic idea here is to print the microarray chip with probes A and B; then the chips are hybridized with Cy5-labeled probe A′. In particular, we first produced a test microarray chip, on which B spots have no contamination of probe A. As illustrated in Figure 3A, these chips were produced by printing all B spots first. Then the printing tips were cleaned. All A spots were printed afterwards. Because all B spots were printed before any A spot was ever printed, carry-over contamination of B spots is not possible. Secondly, we produced a control chip in the same fashion as normal microarray experiments: each spot on the microarray chip was printed based on their chip order (see Figure 3B).

In this way, all B spots, except the first one, were printed after an A spot. If the printing tips were not cleaned completely, these B spots will contain both B probes and some carried-over A probes. Then, both chips were hybridized with probe A′. And, the intensities of the B spots on the two chips were measured and compared.

Because probe B is not labeled with Cy5 and is not complementary to Cy5-labeled probe A′, no B spot should

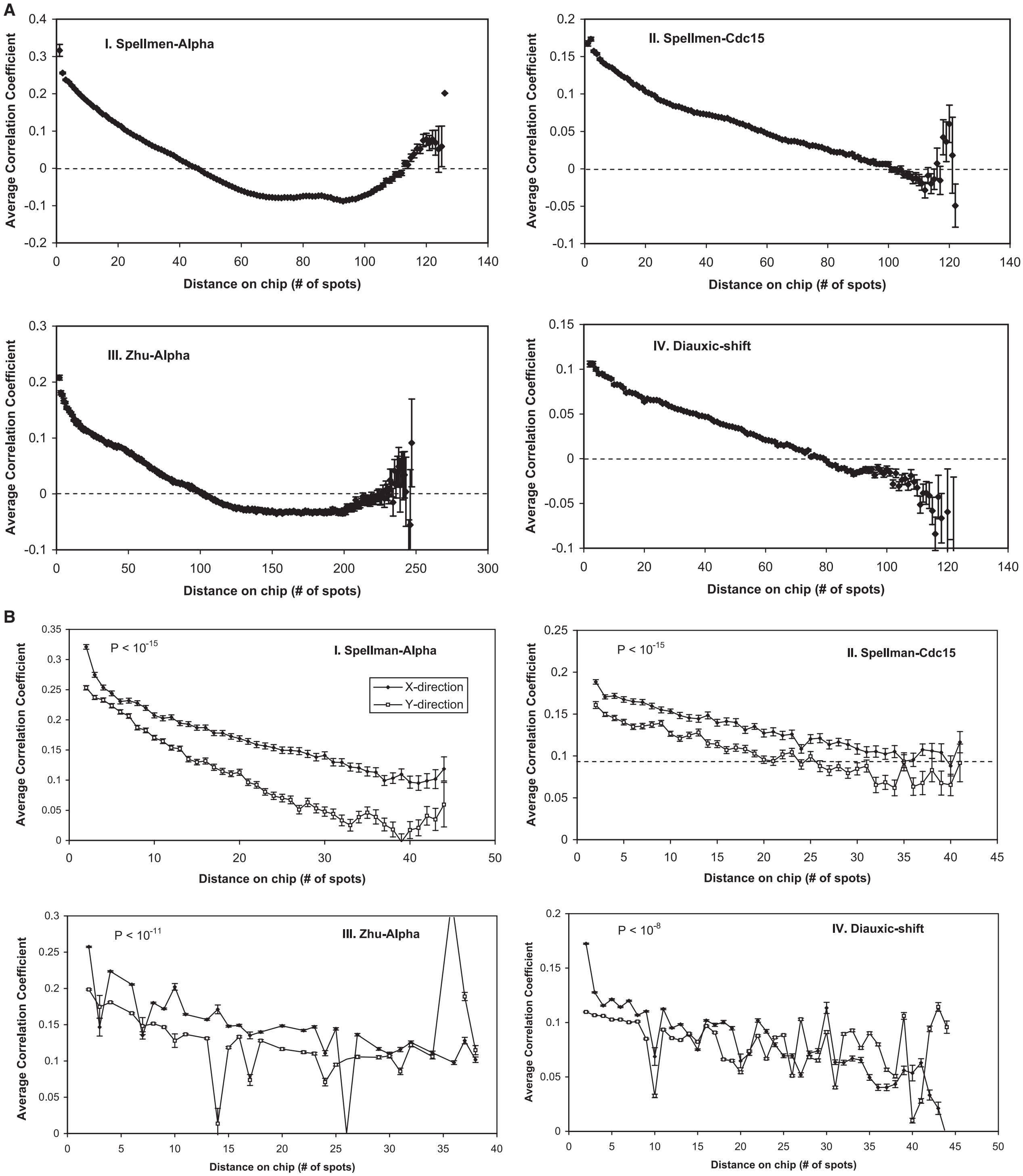


Figure 2. (**A**) Average correlation coefficient distribution as a function of the distance of gene pairs on the chip. All gene pairs on the chip are included for this analysis, except those that are close on the chromosome. (**B**) Average correlation coefficient distribution as a function of the distance of gene pairs on the chip in X- and Y-directions. *P*-values calculated by *t*-tests measure the statistical difference between the correlations of genes that are within the same printing block in X- and Y-directions. Please note that only gene pairs that are printed by the same tip (i.e. within the same printing block) are included for this analysis. The distance between two genes is measured in terms of the number of spots on the chip, i.e. the number of printed spots separating the two. All gene pairs that are close on the chromosome (within 10 ORFs) were excluded from the analysis. Error bars in all figures represent the standard errors of the data.

produce any detectable signal. However, Figure 3C clearly shows that B spots with possible carry-over contaminations on the control chip on average have significantly higher intensity (P-value $< 10^{-15}$) than those without carry-over on the test chip, whose intensities are minimal. This result indicates that B spots on the control chip are indeed contaminated by A probes, confirming that the origin of the chip artifact is possibly the carry-over during the printing process. The

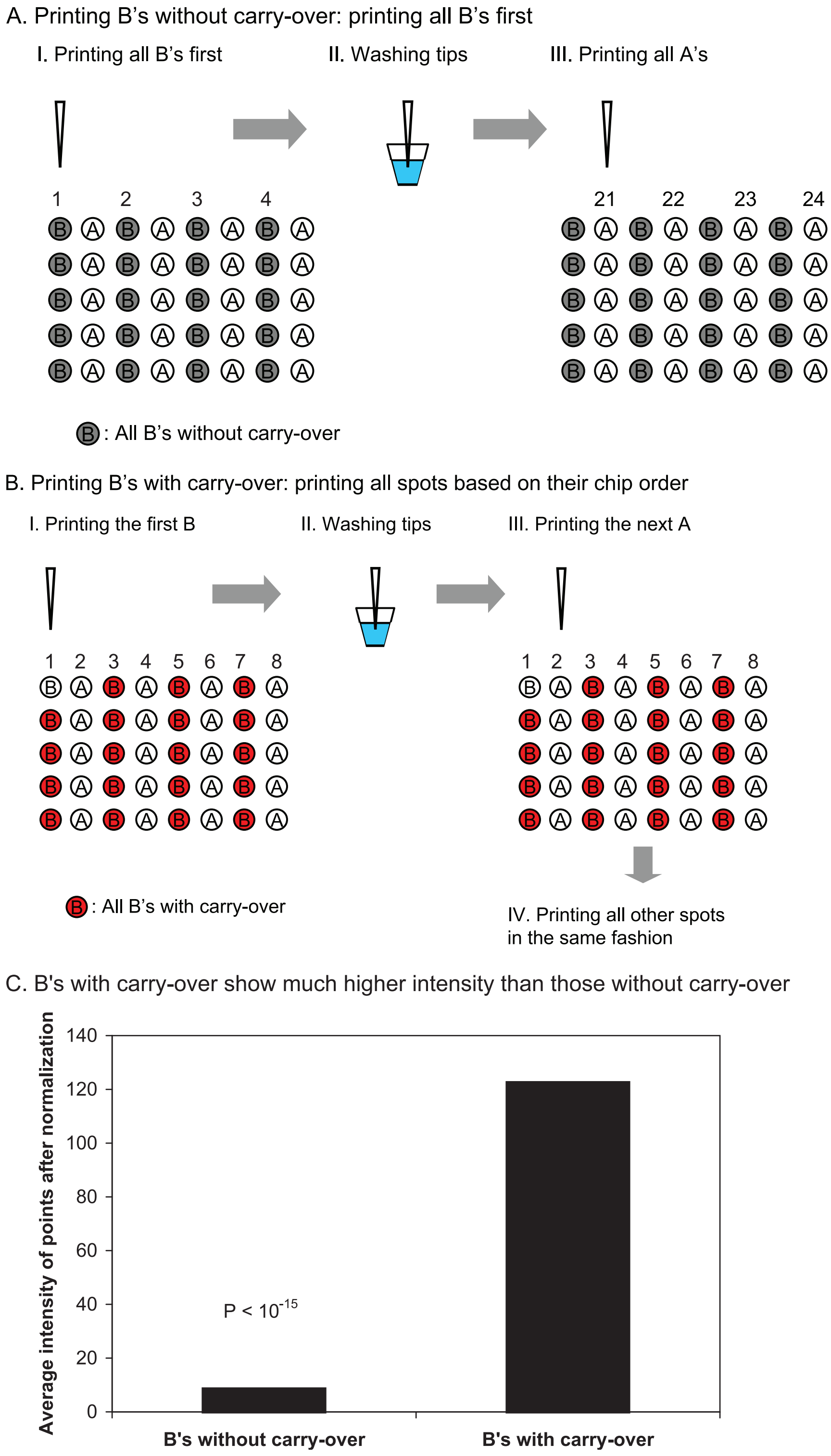


Figure 3. Illustration of the experimental design to uncover the role of carry-over in producing the chip artifact in microarray experiments. (**A**) Producing the test chip: All B’s are printed first without probe A carry-over. (**B**) Producing the control chip: Probes A and B are printed alternatively onto the chip. The numbers in both (A) and (B) indicate the order in which each spot is printed to the chip. (**C**) Comparison of the intensities of B’s with and without carry-over. All intensities were normalized against the test chip (see Methods and Materials section). P-value is calculated using the t-test.

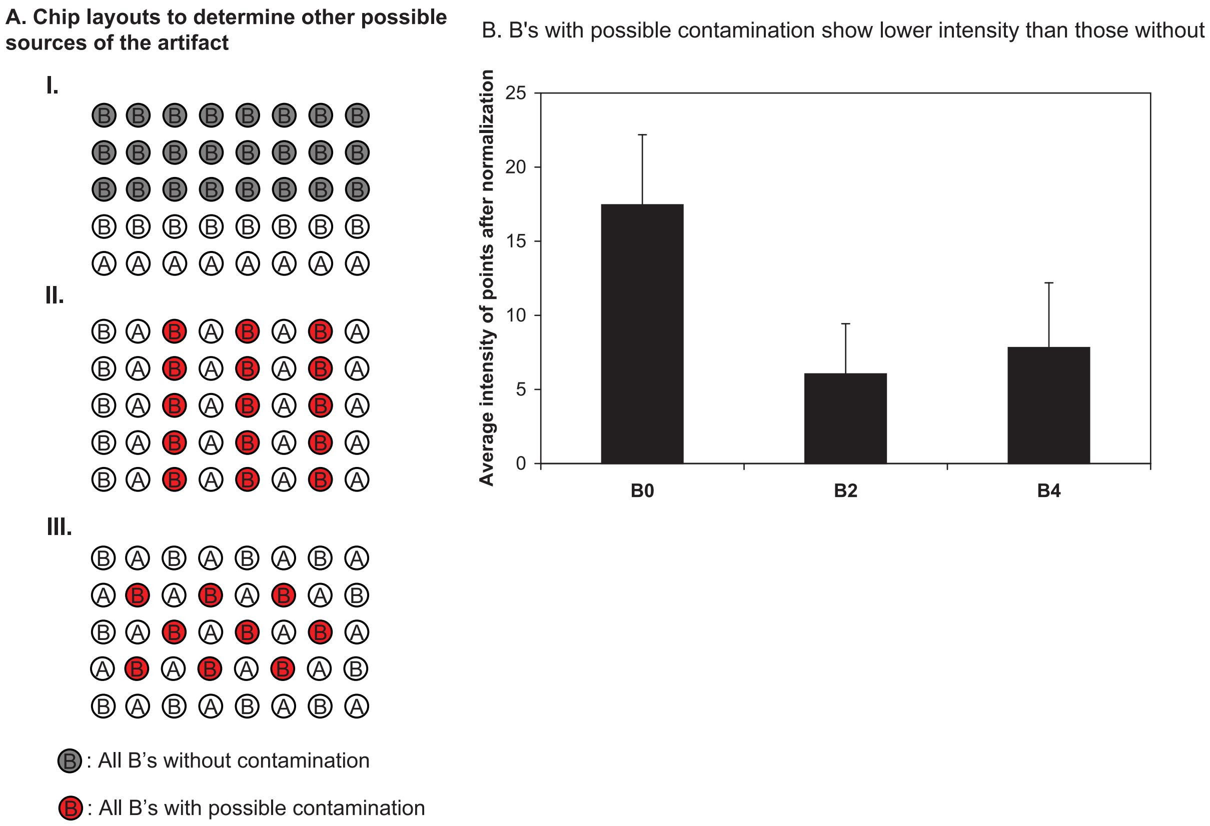


Figure 4. Illustration of the experimental design to uncover the role of other possible sources in producing the chip artifact in microarray experiments. (**A**) Producing microarray chips with different layouts. All B spots were printed first. (**B**) Comparison of the intensities of B's with and without possible contamination. All intensities were normalized against the chip with layout I (see Methods and Materials section).

microarray experiments were repeated several times and the results remain the same (see Supplementary Figure 1). In our experiments, the intensities of contaminated B spots and A spots are of the magnitudes of 100 and 10 000, respectively. Therefore, the signal of a fully-hybridized spot contains ~0.1% noise as measured by our experiments. Please note that one should not interpret the 0.1% signal-to-noise ratio as the amount of probes being carried over, because the intensity measure and the amount of probes are no longer linear when the intensity is too high or too low (19). More importantly, because most spots in microarray experiments are not saturated, the signal-to-noise ratio should be worse. For example, the un-contaminated B spots only have an average intensity of 8.5, which means the artificial carry-over intensity accounts for 93% of the measured intensities of the contaminated ones [average intensity of 122.5; (122.5 − 8.5)/122.5 = 93%]. For most of the lowly-expressed genes, this artifact could therefore become a huge problem (although extremely low intensity spots are often filtered out, making the situation a little better).

Even though the experimental results indisputably show that the carry-over during printing is a major source of the chip artifact, Figure 2B indicates that other factors might also contribute to it: if the carry-over were the only source, one would expect to see little artificial correlations between genes along the non-printing direction (i.e. the Y-direction in Figure 2B). This is not the case for all four experiments in Figure 2B genes along the Y-direction clearly have spurious correlations related to their chip distance, although it is much less severe than that along the X-direction. Other possible sources for the chip artifact include incomplete washing of cDNA after hybridization and image scanning. We also tested these possibilities by experiments: Figure 4A shows that microarray chips were printed with three different layouts. All B spots were printed first, eliminating the possibility of probe A carry-over. In the first layout, the first three rows of B spots are far away from any A spot. These B spots, called B0, will receive minimal effects from A spots, if any. In the second layout, each B spot has two neighbors that are A spots. If the two possibilities above held, these B spots, called B2, would be affected by the nearby A spots. In the third layout, each B spot is surrounded by four A spots. These B spots, called B4, would have the maximal effects from the surrounding A spots, again if the above possibilities held. However, Figure 4B clearly shows that this is not the case. B0 spots actually have higher average intensity than those of B2 and B4 spots (see also Supplementary Figure 2). There may be other possible sources. But our calculations confirm that the carry-over plays a major role in creating the chip artifact discussed here.

Effective distance of the chip artifact

Now that the cause of the chip artifact has been determined, we next investigated the effective distance of

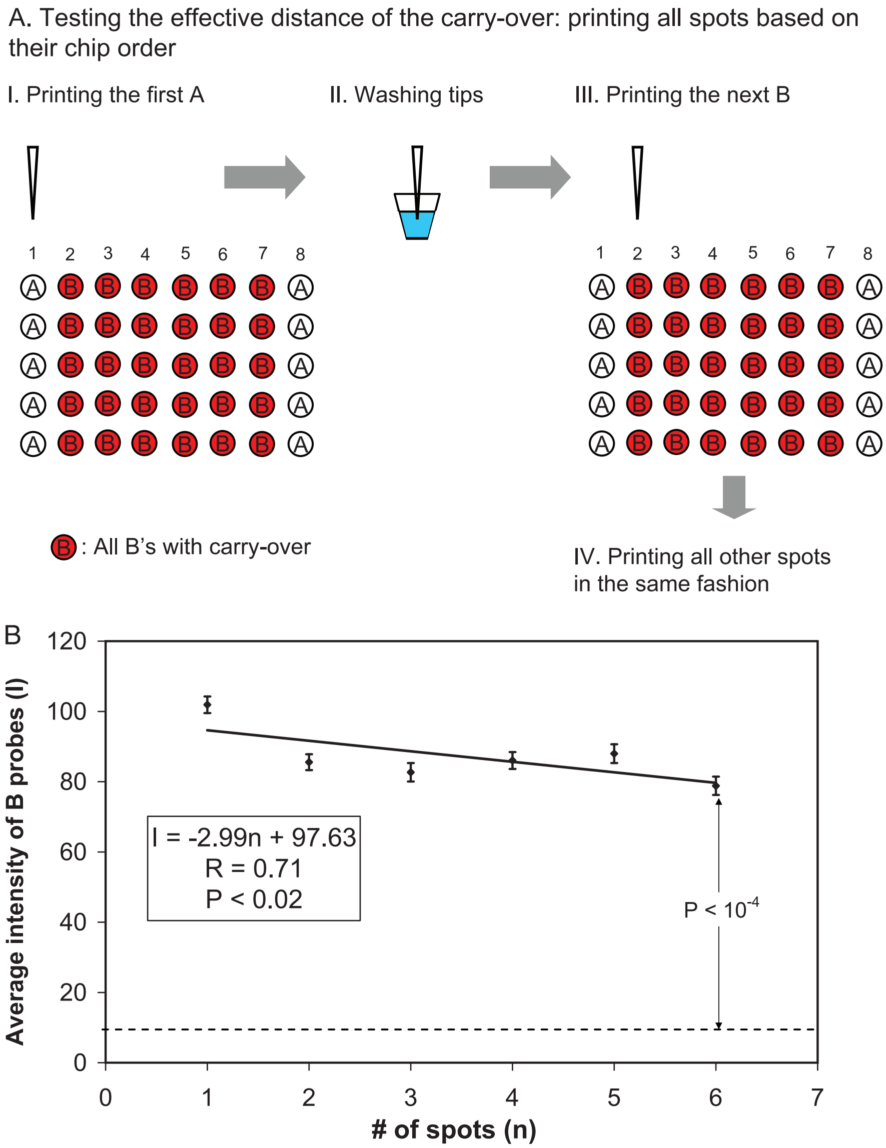


Figure 5. Illustration of the experimental design to determine the effective distance of the chip artifact. (**A**) All spots are printed based on their chip order. (**B**) Comparison of the intensities of B's printed after the A spot. All intensities were normalized against the test chip in Figure 3A (see Methods and Materials section). The P-value of the regression is calculated by the significance test for linear regression. The P-value measuring the intensity difference between B7 spots and un-contaminated spots is calculated by the t-test.

the carry-over artifact how many spots after printing a certain probe does the carry-over effect of this probe disappear? We designed a similar experiment as the ones above (see Figure 5A): first, probe A was printed (A1), followed by six B spots (B2–B6). Then, another A spot was printed (A8). The printing tips were washed after each spot. After the chip was hybridized with Cy5-labeled A′ probes, the intensity of each spot was measured. Figure 5B shows that there is a strong tendency for the carry-over intensity to decrease ($P < 0.02$), which agrees well with our intuition. However, all six B spots still have intensities much higher than those without any carry-over effect ($P < 10^{-4}$). The A8 spots serve as quality control to confirm that there is no systematic bias for spots at different positions (see Supplementary Figure 3). Based on the regression, the estimated effective distance is ~30 spots (see Supplementary Figure 4). This estimated effective distance also agrees well with the observed range of the chip artifact in Figures 2 and 6, even though it is reasonable to believe that the specific effective distance will be different in different microarray experiments. However, it should be noted that there might be other contributing factors to such a long effective distance, which is a possible direction for future analysis.

Artifact related to positions on microtiter plates—plate artifact

Since all four experiments analyzed so far were published at least five years ago, it is interesting to see whether the chip artifact has diminished with the advance of microarray technology. Therefore, we analyzed three newly-published

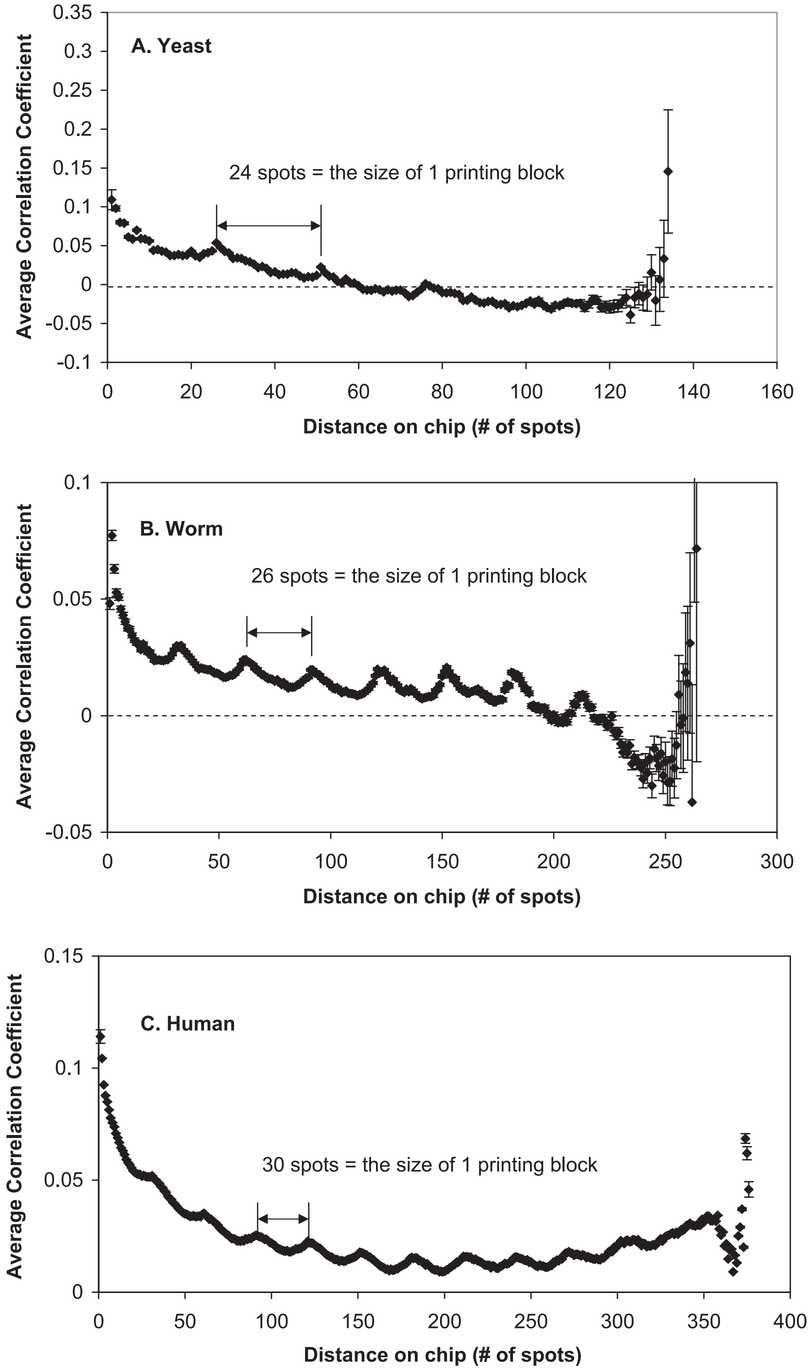


Figure 6. Average correlation coefficient distribution as a function of the distance of gene pairs on the chip. All three curves show striking periodicities, corresponding to the size of the printing block in the three experiments. All gene pairs that are close on the chromosome (within 10 ORFs) were excluded from the analysis.

microarray experiments in three different organisms performed by three independent labs:

(i) Yeast, the microarray karyotyping dataset from Dunn *et al.* (20);
(ii) Worm, the ER stress response dataset from Viswanathan *et al.* (21);
(iii) Human, the colon cancer dataset from Giacomini *et al.* (22);

Figure 5 clearly shows that the chip artifact still remains in all of the three newly-published experiments. More interestingly, all three curves show striking periodicities. When we examined the periodicities more closely, we found that the period of each curve corresponds to the size of one printing block in each experiment. As we discussed above, corresponding spots in different blocks are actually neighbors on the microtiter plates, because of the printing procedure of the microarray experiments (see Figure 1C). Therefore, our results show that genes that are close on the microtiter plates tend to have artificially higher correlations. We call this a plate artifact. The magnitude of this artifact is obviously less dominant than that of the chip artifact discussed above, but is nevertheless non-negligible. The plate artifact is clearly related to the cross-contamination during the sample-preparation and PCR processes. It could also explain the spurious correlations observed in Figure 2B along the Y-direction: the nearby spots along the Y-direction are very close on the plate as well (see Figure 1), even though they are not printed one after another.

Consequences of positional artifacts

The positional artifacts (both chip and plate artifacts) that we showed here exist with different severity in every microarray experiment that we examined. These artifacts introduce spurious correlations between genes which just happen to be close on the microtiter plates or microarray chips. They thus create substantial issues, especially because more and more people have become interested in analyzing the relationships between gene expression and chromosomal location recently (9,23). However, since a gene's position on the plates and chips in many microarray experiments are just a transformation of its chromosomal location, it can be shown that these positional artifacts could lead to false biological conclusions:

To illustrate this point, we calculated the distribution of average correlation coefficient between the expression levels of all yeast genes as a function of their chromosomal order (see Figure 7A). Here, we only examined the three yeast cell-cycle experiments (12,13), all of which show surprising periodicities. The shorter periodicity is 2 ORFs and the longer one is 22 ORFs, which is further confirmed by the Fourier

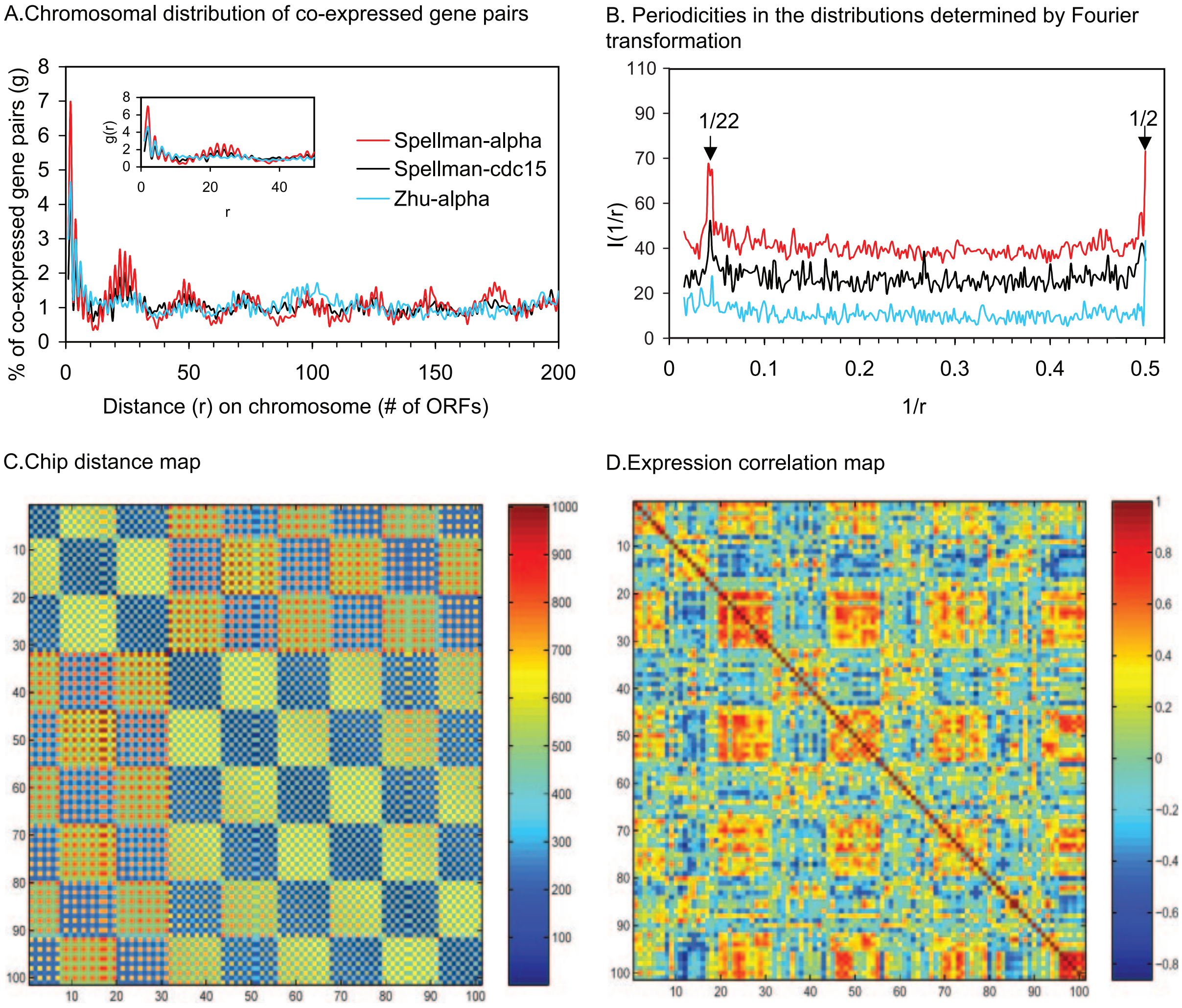


Figure 7. (**A**) Pair correlation function for Spellman-alpha-factor arrested cell cycle dataset (red), Spellman-cdc15 arrested cell cycle dataset (black), and Zhu-alpha factor blocked cell cycle dataset (light blue), the inset highlights the staggered characteristics. X-axis represents the distance between gene pairs. Y-axis represents the percentage of highly correlated pairs that have a given distance. (**B**) Power spectrum of the pair correlation function of co-expressed gene pairs determined by the Fourier transformation. Two common frequencies are indicated by the arrows. Please note that the distributions are manually shifted 20 U along the Y-axis to separate them from each other to clearly show the peaks. (**C**) Chip distance map and (**D**) Expression correlation coefficient map. Both maps are produced using Spellman-alpha-factor arrested cell cycle dataset, whose *x*- and *y*-axis represent the first 100 ORFs on chromosome IV. In the distance map, the color on each spot represents the distance between the gene on *x*-axis and the gene on *y*-axis. In the expression correlation coefficient map, the color represents the correlation coefficient between the gene pair.

transformation of the three distributions (see Figure 7B). Without considering the positional artifacts we mentioned above, one could come to the conclusion that these periodicities show the effects of the chromosomal structure on gene expression. This is particularly exciting given that the longer periodicity of 22 ORFs (∼42 kb) is of the order of the size of the chromatin loop domains, which range from 20 to 100 kb (24). However, taking into consideration the chip architecture, one could easily see that these periodicities are the results of the positional artifacts: First, in all three experiments, genes are placed on the microtiter plates based on their chromosomal order—neighboring genes on the chromosome are also neighbors on the plate. Because of the way in which the chip are produced as illustrated in Figure 1, neighboring genes on the chromosome are printed by two different tips and are far-away on the chip; whereas genes that are second neighbors on the chromosome are actually printed by the same tip and become immediate neighbors on the chip. Due to the carry-over chip artifact, neighboring spots on the chip have much higher average correlation coefficient than the far-away ones, producing the shorter periodicity of 2 ORFs. Second, we constructed a chip distance map (Figure 7C) and an expression correlation coefficient map (Figure 7D). The horizontal and vertical axes of these two maps represent the position of the genes along the same chromosome (chromosome four in this case). The colors of the distance and correlation maps represent the chip distance and expression correlation coefficient between gene pairs, respectively. Surprisingly, the patterns in both maps are very similar—the distance map also has a characteristic periodicity of 22 ORFs. Therefore, genes that are 22 ORFs away on the chromosome tend to be very close on the chip as well. Because of the carry-over chip artifact, these genes tend to have higher correlation coefficients, producing the longer periodicity of 22 ORFs. Similar results have also been observed by Balázsi *et al.* (16).

COP—detection of positional artifacts in microarray experiments

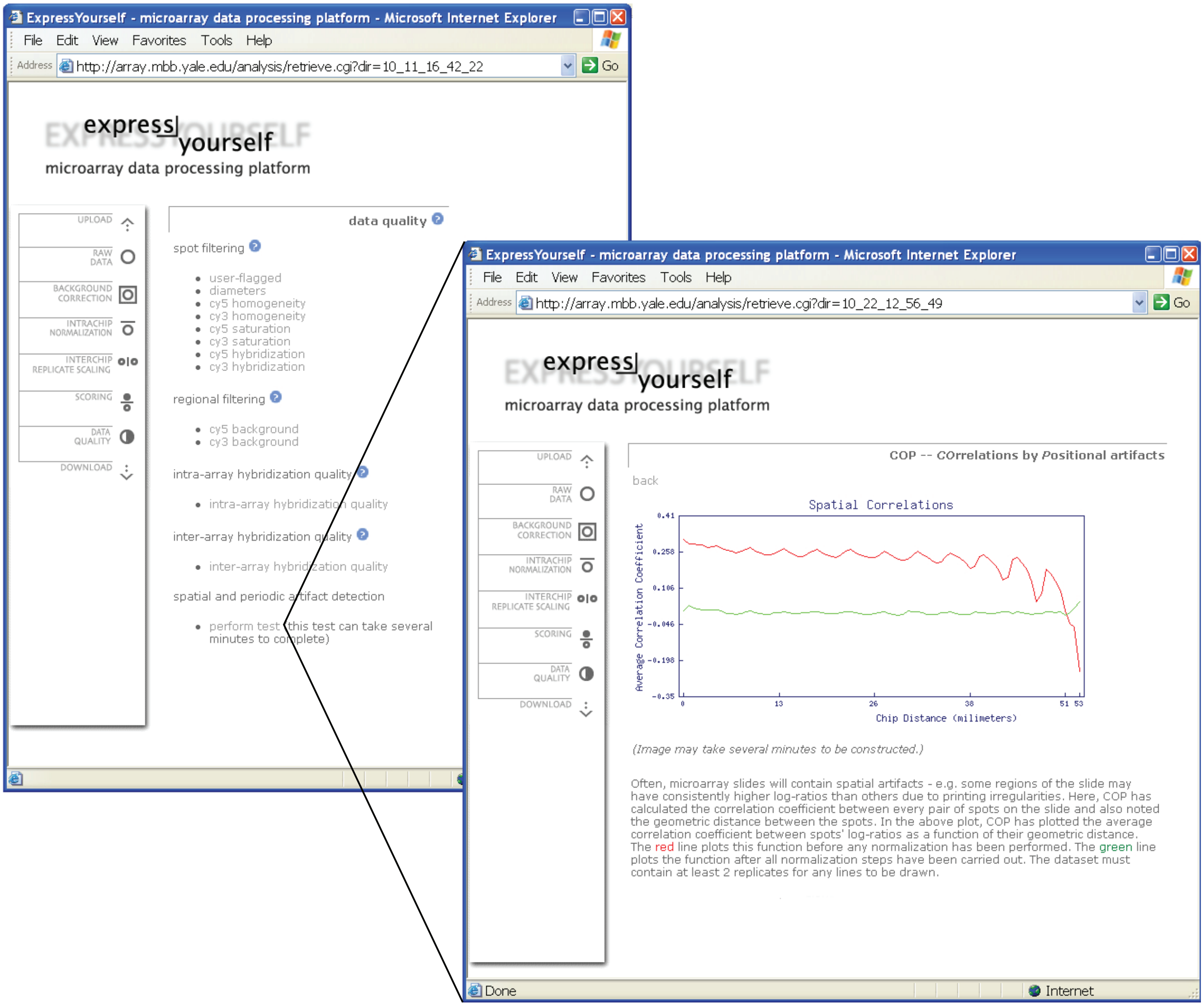
Because of the generality and severity of the positional artifacts, it is of great importance for biologists to control these artifacts in their experiments. Towards this end, we developed an automatic web tool to detect these positional artifacts in microarray data—COP. It is of course desirable to correct the artifacts upon detection. Therefore, we integrated COP with ExpressYourself, a normalization tool for microarray data previously published by Luscombe *et al.* (25). ExpressYourself assumes that the majority of the genes printed on a microarray chip do not change in the test and control samples (i.e. the Cy3 and Cy5 channels); thus, the overall mean intensity ratio between the two channels should be one, which is a common assumption in normalizing microarray data (15,26). It then removes the positional artifacts and other types of noise (systematic and random) in the microarray data by subtracting the best-fit local average LIR from the raw LIR of each local spots. In this way, the average LIR of all spots will be zero and thus the average intensity ratio is one. ExpressYourself estimates the best-fit local average LIR using the lowess regression method. Because it is known that intensity ratios may change at different intensity levels (as two dyes have different fluorescent properties) and they could also be different at different positions on the same microarray slide, ExpressYourself estimates best-fit local average LIR's at different intensity levels, as well as at different physical positions on the slide. This underlying procedure is very similar to the print-tip normalization discussed below, but, ExpressYourself does not perform the print-tip normalization.

In order to use COP, one simply uploads the raw microarray data into ExpressYourself. After the normalization is done, the user selects 'DATA QUALITY' ('perform test'. The distributions of average correlation coefficients of all spots in the experiments before and after the normalization will then be displayed. In Figure 6, we showed the results for the human colon cancer dataset from Giacomini *et al.* (22). It is clear that the normalization performed by ExpressYourself largely reduced the artifacts, but did not remove them completely. Therefore, we recommend that care must be taken if a microarray experiment contains clear artifacts after the normalization (ideally, the experiment should be repeated). Even though the threshold to discard an experiment depends on the specifics of each particular experiment, the estimated upper bound is 0.1. To estimate this upper bound, we first calculated the standard deviation of the correlation coefficients between neighboring gene pairs for each of the microarray experiments that we examined. During this process, we also noticed that the distribution of the correlation coefficients approximates a normal distribution. Assuming the distribution is normal, we found that if the average correlation is 0.1, it would be significantly >0 ($P < 10^{-3}$) using all different standard deviations observed in the examined experiments. Furthermore, because the positional artifacts are introduced during the process of manufacturing the chips, the experiment should be repeated using a different batch of chips.

Discussion and conclusions

Here, we showed, both computationally and experimentally, that two types of artifacts related to the position of the genes on the microtiter plates and microarray chips exist in microarray experiments. Genes that are close on the plate or chips tend to have spurious correlations separated from their biological functions. We therefore built an automated web tool—COP, which, together with ExpressYourself, can detect and, to some degree, correct these positional artifacts in microarray data (Figure 8). One potential problem with ExpressYourself's normalization procedure is that it might abolish some of the true correlations between the genes, especially when the chips are printed based on gene's chromosomal order. Therefore, we strongly suggest that genes on all microarray chips should be spotted randomly. Even though this does not reduce any of the artifacts in the experiment, it is much less likely that these artifacts will interfere with the subsequent biological analysis from a statistical point of view.

As discussed above, other chip-related artifacts within microarray experiments have been analyzed by previous studies, as well. For example, Yang *et al.* (15) have found a so-called 'print-tip' artifact, which was also reported by Balázsi *et al.* (16). Because of 'slight differences in the length


ExpressYourself - microarray data processing platform - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://array.mbb.yale.edu/analysis/retrieve.cgi?dir=10_11_16_42_22
Go
express yourself
microarray data processing platform
UPLOAD
RAW DATA
BACKGROUND CORRECTION
INTRACHIP NORMALIZATION
INTERCHIP REPLICATE SCALING
SCORING
DATA QUALITY
DOWNLOAD
data quality
spot filtering
user-flagged
diameters
cy5 homogeneity
cy3 homogeneity
cy5 saturation
cy3 saturation
cy5 hybridization
cy3 hybridization
regional filtering
cy5 background
cy3 background
intra-array hybridization quality
intra-array hybridization quality
inter-array hybridization quality
inter-array hybridization quality
spatial and periodic artifact detection
perform test

(this test can take several minutes to complete)

ExpressYourself - microarray data processing platform - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://array.mbb.yale.edu/analysis/retrieve.cgi?dir=10_22_12_56_49
Go
express yourself
microarray data processing platform
UPLOAD
RAW DATA
BACKGROUND CORRECTION
INTRACHIP NORMALIZATION
INTERCHIP REPLICATE SCALING
SCORING
DATA QUALITY
DOWNLOAD
COP -- COrrelations by Positional artifacts
back
Spatial Correlations
Average Correlation Coefficient
0.41
0.258
0.106
-0.046
-0.198
-0.35
0
13
26
38
51
53
Chip Distance (milimeters)
Done
Internet

(Image may take several minutes to be constructed.)

Often, microarray slides will contain spatial artifacts - e.g. some regions of the slide may have consistently higher log-ratios than others due to printing irregularities. Here, COP has calculated the correlation coefficient between every pair of spots on the slide and also noted the geometric distance between the spots. In the above plot, COP has plotted the average correlation coefficient between spots' log-ratios as a function of their geometric distance. The red line plots this function before any normalization has been performed. The green line plots the function after all normalization steps have been carried out. The dataset must contain at least 2 replicates for any lines to be drawn.

Figure 8. Screen shot of COP within ExpressYourself. The dataset used here is the human colon cancer dataset.

or in the opening of the tips and deformation after many hours of printing', there might be systematic difference between different print tips (15). Thus, the average intensity ratio of spots printed by a tip might be different from another one, creating the 'print-tip' artifact. Yang *et al*. (15) have also developed a print-tip normalization method to remove this kind of artifact by a lowess-fit procedure similar to that used in ExpressYourself. This normalization procedure is available in the 'marray' package from the Bioconductor project (27). Furthermore, Bengtsson (28) has found similar 'plate effects'—spots from different microtiter plates tend to have different average intensity ratios. And he suggested a similar normalization procedure adopted from the print-tip normalization (28). These two types of artifacts are different from the artifacts that we discussed above in that they caused systematic difference between spots printed by different tips or on different plates. More specifically, this means that spots printed by the same tips or on the same plates tend to have similar intensity ratios, but their intensity ratios are systematically different than those printed by another tip or on another plate. In our analysis, however, we found that neighboring spots tend to have higher correlations in their intensity ratios than far-away ones even though they are printed by the same tip or on the same plate. Interestingly, despite the differences, in practice the print-tip normalization procedure is effective in removing the positional artifacts in most microarray experiments.

Moreover, Smyth and Speed (26) have discovered that spots printed by the same tip on the same chip may have systematic differences in their intensity ratios resulting from the fact that different wells on a microtiter plate may contain 'different effective quantities of DNA', which they called the 'print-order' artifact. This artifact can also be removed by a similar procedure as the print-tip normalization (26). More recently, Uchida *et al*. have also detected this 'print-order' artifact in one of the microarray experiments they analyzed (29). The origin of this 'print-order' artifact is clearly different from the carry-over nature of the chip artifact we discussed above. More importantly, even though the analysis of this 'print-order' artifact has some similarity to our analysis, it is focused on the average intensity ratio of spots printed on the same chip, which does not necessarily lead to higher correlations between nearby spots across all chips in the experiment.

In spite of their differences, these different kinds of artifacts are often entangled with each other in most microarray experiments, creating a huge challenge for microarray practitioners to carefully normalize the data to remove all of these artifacts. In order to do this successfully, the causes of these artifacts have to be understood well, further highlighting the importance of our analysis.

More interestingly, even though the design and production procedure are totally different, some Affymetrix arrays show a similar artifact, as well (see Supplementary Figure 5) (30). The specific source of this artifact is still unclear. It might be related to the Y-direction correlations that we discussed in Figure 2B. The result of the Affymetrix arrays further confirms the prevalence of these positional artifacts and the importance of our analysis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
2. Shalon,D., Smith,S.J. and Brown,P.O. (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.*, **6**, 639–645.
3. Brown,P.O. and Botstein,D. (1999) Exploring the new world of the genome with DNA micrarrays. *Nature Genet.*, **21**, 33–37.
4. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
5. Altman,R.B. and Raychaudhuri,S. (2001) Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.*, **11**, 340–347.
6. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
7. Roy,P.J., Stuart,J.M., Lund,J. and Kim,S.K. (2002) Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans. Nature*, **418**, 975–979.
8. Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
9. Cohen,B., Mitra,R., Hughes,J. and Church,G. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.*, **26**, 183–186.
10. Royce,T.E., Rozowsky,J.S., Bertone,P., Samanta,M., Stolc,V., Weissman,S., Snyder,M. and Gerstein,M. (2005) Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet.*, **21**, 466–475.
11. Quackenbush,J. (2002) Microarray data normalization and transformation. *Nature Genet.*, **32** (Suppl), 496–501.
12. Spellman,P., Sherlock,G., Zhang,M., Iyer,V., Anders,K., Eisen,M., Brown,P., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
13. Zhu,G., Spellman,P.T., Volpe,T., Brown,P.O., Botstein,D., Davis,T.N. and Futcher,B. (2000) Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, **406**, 90–94.
14. DeRisi,J., Iyer,V. and Brown,P. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
15. Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.
16. Balazsi,G., Kay,K.A., Barabasi,A.L. and Oltvai,Z.N. (2003) Spurious spatial periodicity of co-expression in microarray data due to printing design. *Nucleic Acids Res.*, **31**, 4425–4433.
17. Qian,J., Kluger,Y., Yu,H. and Gerstein,M. (2003) Identification and correction of spurious spatial correlations in microarray data. *Biotechniques*, **35**, 42–44, 46, 48.
18. Kluger,Y., Yu,H., Qian,J. and Gerstein,M. (2003) Relationship between gene co-expression and probe localization on microarray slides. *BMC Genomics*, **4**, 49.
19. Dudley,A.M., Aach,J., Steffen,M.A. and Church,G.M. (2002) Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc. Natl Acad. Sci. USA*, **99**, 7554–7559.
20. Dunn,B., Levine,R.P. and Sherlock,G. (2005) Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. *BMC Genomics*, **6**, 53.
21. Viswanathan,M., Kim,S.K., Berdichevsky,A. and Guarente,L. (2005) A role for SIR-2.1 regulation of ER stress response genes in determining *C.elegans* life span. *Dev. Cell*, **9**, 605–615.
22. Giacomini,C.P., Leung,S.Y., Chen,X., Yuen,S.T., Kim,Y.H., Bair,E. and Pollack,J.R. (2005) A gene expression signature of genetic instability in colon cancer. *Cancer Res.*, **65**, 9200–9205.
23. Spellman,P.T. and Rubin,G.M. (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. *J. Biol.*, **1**, 5.
24. Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J. (1994) *Molecular Biology of the Cell, 3rd edn.* Garland Publishing, NY.
25. Luscombe,N.M., Royce,T.E., Bertone,P., Echols,N., Horak,C.E., Chang,J.T., Snyder,M. and Gerstein,M. (2003) ExpressYourself: a modular platform for processing and visualizing microarray data. *Nucleic Acids Res.*, **31**, 3477–3482.
26. Smyth,G.K. and Speed,T. (2003) Normalization of cDNA microarray data. *Methods*, **31**, 265–273.
27. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
28. Bengtsson,H. (2002) Identification and normalization of plate effect in cDNA microarray data. *Math. Sci.*, **28**.
29. Uchida,S., Nishida,Y., Satou,K., Muta,S., Tashiro,K. and Kuhara,S. (2005) Detection and normalization of biases present in spotted cDNA microarray data: a composite method addressing dye, intensity-dependent, spatially-dependent, and print-order biases. *DNA Res*, **12**, 1–7.
30. Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. *et al.* (1998) A genome-wide transcriptional analysis of the Mitotic cell cycle. *Mol. Cell*, **2**, 65–73.