Identification and correction of spurious spatial correlations in microarray data

Jiang Qian, Yuval Kluger, Haiyuan Yu, and Mark Gerstein
Yale University, New Haven, CT, USA

BioTechniques 35:42-48 (July 2003)

Microarray experiments are providing a huge amount of genome-wide data on gene expression. Many prior expression analyses have focused on inferring functional relationships (1–7); however, the quality control and normalization of the raw data that result from microarrays have received less attention. Here we address a systematic error that arises from microarrays and discuss current methods to resolve the problem.

It is well known that the data from high-throughput experiments embody a significant component of measurement error that must be removed before any analysis can be applied to the data. An intuitive idea is to repeat the experiments and decrease the noise by averaging the measurements from replicates (8). Unfortunately, microarrays are still difficult to repeat; in most cases, researchers do not have many replicates for analysis. A Bayesian probabilistic approach has been proposed to address the problem of the small repetition number for microarray experiments (9). While random error can be canceled by replicate experiments, systematic error will not diminish by averaging replicates. For example, a notorious systematic error in microarray experiments is that the expression ratio of a particular gene at different conditions is a function of its absolute expression levels. If one uses a simple fold-change cut off, then the genes with low expression levels tend to numerically meet the given cut off, even though they are not truly differentially expressed. Different methods have been proposed to deal with this problem (10–15).

In this review, we want to direct attention toward a type of systematic error that is manifested by the strong interaction between neighboring spots on the array. If the replicate experiments are performed on the arrays with same-chip geometry, then these interactions will not be canceled by the replicates. We will first demonstrate this noise via a case study, and then we will discuss the possible source of these artifacts. Finally, we will discuss current methods to solve the problem, in particular, a local averaging approach called standardization and normalization of microarray data (SNOMAD) (16). We examined several different yeast microarray data sets: diauxic shift, α-factor-arrested cell cycle, cdc15-arrested cell cycle, and cdc28-arrested cell cycle (17–19).

To demonstrate the systematic error in the microarray data, we offer the following evidence. The relationship between gene expression and physical chip distance can be revealed by comparing the chip distance map (Figure 1A) to an expression correlation coefficient map (Figure 1B). The horizontal and vertical axes of these two maps represent the positions of the genes along a chromosome. The colors on the distance and correlation maps represent the chip distance and expression correlation coefficient between gene pairs, respectively. Interestingly, the highly correlated gene expression regions (Figure 1B, red blocks) always correspond to the short chip distance regions (Figure 1A, red blocks), indicating that the observation of two genes to be coexpressed could be mainly due to their short physical distance on the chip.

We also calculated the average correlation coefficient of gene expression profiles as a function of the physical chip distance between two genes. Figure 2 shows the result for a microarray data set of the yeast α-arrested cell cycle. Without an artifact, the average correlation coefficient should be independent of the chip distance. However, Figure 2 shows that the closer two genes are on the chip, the higher their average correlation coefficient is. This indicates that this data set contains a large proportion of artifacts. Actually, this phenomenon is not unique to

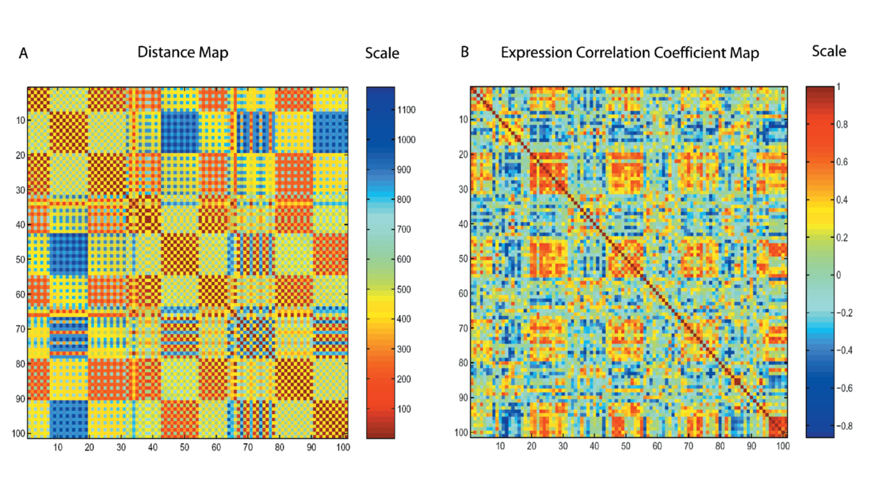

Figure 1. Distance map and expression correlation coefficient map. Both maps are produced using the yeast α-factor-arrested cell-cycle data set, whose x-axis and y-axis represent the first 100 open reading frames on chromosome IV. (A) Distance map. The color on each spot represents the distance between the gene on the x-axis and the gene on the y-axis. (B) Expression correlation coefficient map. The color represents the correlation coefficient between the gene pair. The color codes are to the right of the maps. For a detailed analysis, please refer to Yu et al. (manuscript in preparation).

microarrays; we performed the same calculation on cell-cycle data from the GeneChip® (Affymetrix, Santa Clara, CA, USA) (1) and obtained similar results with respect to the artifact.

For microarray experiments, there is yet additional evidence for an artifact. In the cell-cycle experiments, researchers measured the gene expression ratio between the different cell-cycle stages, using asynchronous cultures of the same cells as a control sample. This control sample is labeled by Cy™3 (green). Ideally, the expression profiles of green signal for all genes should be proportional. Thus, the average correlation coefficient for the green signal should be 1. However, according to Figure 3, we found a pattern similar to that of Figure 2. This is a clear manifestation of an artifact.

From this analysis, we can see that the artifact phenomenon is significant and exists in many chips. Thus, the artifact must be taken into account before any conclusion can be drawn based on the raw, uncorrected expression data. The following is an example of what we just stated. A naïve analysis of α-factor-arrested yeast cell-cycle data suggests that chromosomal spatial organization affects gene expression in a systematic way, as displayed in the distribution of highly correlated gene pairs as a function of the relative pair chromosomal distance. The figure shows that (*i*) adjacent gene pairs tend to have high correlation coefficients, which is consistent with findings by Cohen et al. (20), and (*ii*) genes that are not in the same vicinity on the chromosome are more likely to be co-expressed if their spacing is a multiple of 22 open reading frames (ORFs) in microarray experiments. Given the fact that many chips, including this particular microarray, are printed according to a simple transformation of the gene order on the chromosome, the observed long-range correlation could be associated with an inherent chip artifact.

The source of the artifact is unknown, but it might be related to the following processes or their combinations: (*i*) the spotting of DNA probes on the chips; (*ii*) plate effects; (*iii*) the washing of cDNA after hybridization; (*iv*) cross hybridization; or (*v*) image scanning. The effect of spotting of DNA probes on the chip is also called print tip effect. A systematic difference may exist between the print tips and lead to spatial bias between the sectors on the chip. The analysis of variance (ANOVA) method (11) or MA-plot (19) allows for the detection of this spatial bias, and a lowess normalization approach was proposed to correct the systematic bias (15). The plate effect originates from PCR amplification bias between different plates. This effect would also introduce further variability in the measurement of gene expression. Nonspecific probes were used to correct the effect, based on the assumption that DNA concentration results in this bias. All these effects are not easy to separate. Furthermore, some of the assumptions, such as equal variance between different sectors, may be invalid. This makes the detection and correction of these effects even more difficult.

It would be most desirable, of course, to completely correct for the artifact after determining its source. However, in practice, it is difficult to correct for all the spatial bias. For example, Yang's normalization method is able to correct the spatial artifact due to print tips (15). However, the unit corrected here is the chip "sector" (or block), and this is a rather coarse division; the spatial bias from other sources may still exist within sectors after Yang's sector-based normalization. We believe that even without fully understanding the source of the problem, researchers are still able to improve the signal quality.

Here we discuss a popular method of "spatial lowess" in detail (16). This local normalization method allows for the detection and/or correction of spatially systematic artifacts in microarray data without exactly attributing the artifacts to certain sources. We applied the method from Colantuoni et al. (also called SNOMAD) (16) to the α-factor-arrested yeast cell-cycle data and found that their method reduced the artifact effect but failed to remove it complete-

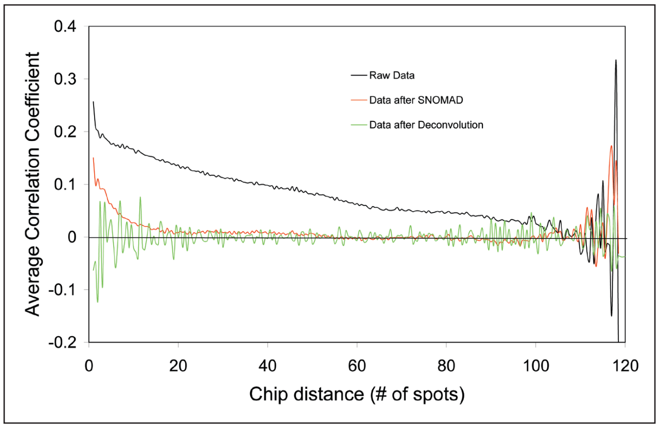

Figure 2. Average correlation coefficient distribution as a function of the distance of gene pairs on the chip. The distance between genes is measured in terms of the number of spots on the chip. This distribution is calculated using the α-factor-arrested cell-cycle data set. The black line is the distribution for the raw data. The red line is the distribution for the data after the standardization and normalization of microarray data (SNOMAD). The green line is the distribution for the data after deconvolution.

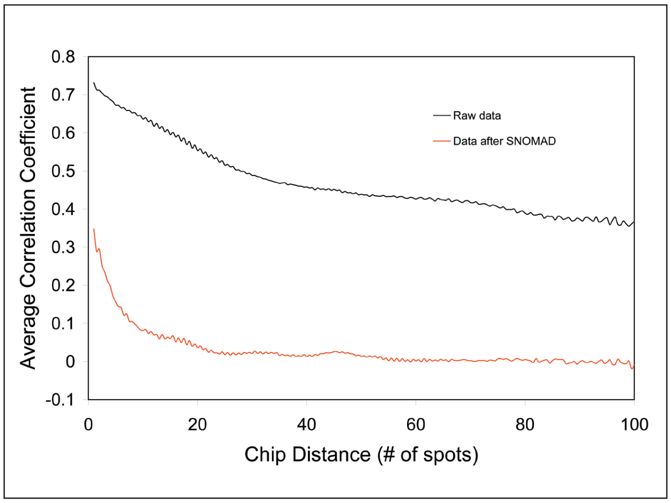

Figure 3. Average correlation coefficient distribution for green signals. The black line is the distribution for the raw data. The red line is the distribution for the data after the standardization and normalization of microarray data (SNOMAD).

ly. The red line in Figure 2 shows the average correlation coefficient as a function of the chip distance after local normalization. Apparently, the problem is diminished when we compare the situation before the normalization, as shown by the black line in Figure 2. Figure 4 shows the self-correlation of genes that were printed twice on the chip. Without the normalization procedure, the mode of the distribution of all self-correlations is approximately 0.1, whereas, in an ideal situation, it should be 1. The application of SNOMAD drives this distribution to the right, with a slightly higher mode at 0.15, which is evidence that SNOMAD improves data quality. The red line in Figure 3 shows that after local normalization, the pair correlation function of the green signals is still not homogeneous, which means the method cannot remove the artifact completely. Surprisingly, it even produces correlation coefficients between green signals close to 0. These correlations should ideally be 1 because we only use local normalization when we processed the green signals using SNOMAD. The results are similar for several normalization methods (data not shown). An important assumption of SNOMAD is that the artifact is isotropic on the chip, which is actually untrue in most cases. For example, we calculated the distribution of the average correlation coefficient for all the gene pairs in the same rows and a similar distribution for those pairs in the same columns. Figure 5 shows the results for α-factor-arrested yeast cell cycle, and it is clear that the artifact along the x-axis is quite different from that along the y-axis.

We propose a deconvolution approach to address the artifact problem in chip experiments. This is actually more of a general approach than local normalization and may be able to take into account anisotropic effects. We assume that for each sample indexed by the letter t, the measured signal ϕ^t at a chip location $\vec{x} \equiv (x,y)$ can be expressed as a convolution of the true signals

$$\psi^t\text{: } \phi^t(\vec{x}) = \sum_{\vec{u}} c(\vec{u})\psi^t(\vec{x} - \vec{u}),$$

where the deviation of c from a δ function represents the extent of the chip artifact. (Note that ψ^t is the ratio of the red and green channels.) Thus, neighboring and non-neighboring spots affect the signal measured at the point $\vec{x}$. According to the convolution theorem, the Fourier transform of $\phi^t(\vec{\boldsymbol{x}})$ is given by $\phi^t(\vec{k}) = c(\vec{k})\psi^t(\vec{k})$. Because the true signals ψ^t and the envelope $c(\vec{k})$ that represent the artifact are unknown, we inspect the artifact-free ratios $\mathrm{R}^t(\vec{k}) \equiv \phi^t(\vec{k})/\phi^*(\vec{k}) = \psi^t(\vec{k})/\psi^*(\vec{k})$ for all samples t, where $\phi^*(\vec{k})$ is some reference measurement, such as the average of $\phi^*(\vec{k})$ across all samples. A preliminary result from this idea is illustrated in Figure 2, where we show average correlation coefficients as a function of the physical distance of gene pairs on the chip. These distributions were calculated using an α-factor-arrested cell-cycle data set. Clearly, SNOMAD fails to remove all the artificial components in the expression profiles. On the contrary, the distribution after our proposed deconvolution method is no longer distance-dependent. The inverse Fourier transforms of these ratios, denoted by $R^t(\vec{x})$, have no straightforward biological interpretation. Nevertheless, under the assumption that the convolution model is adequate, substitution of $\phi^t(\vec{x})$ for these ratios and application of the above three pieces of evidence (designed to reveal the artifact) allow us to diminish chip artifact effects.

To understand the effect of the chip artifact, we propose printing a unique sequence in each spot of the array. Hybridization with the corresponding cDNA will allow us to study the spatial variations of the red and green signals and their ratios, where the dyes represent two distinct or equivalent cDNA samples. Replicating this experiment using different spot spacing will allow us to study the spot-to-spot interaction effect and the associated correlation length. To remove the spatial variation effect, we propose a variant of this approach. Instead of printing a unique sequence at each spot, one can link it to each of the DNA targets. The red channel can then be assigned to the mRNA samples, and the green channel can be

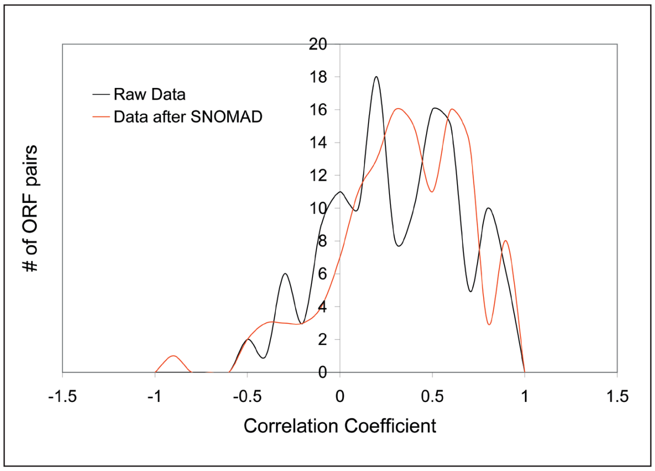

Figure 4. Distribution of correlation coefficient for the duplicated genes. The x-axis represents the correlation coefficient between a duplicated gene pair. The y-axis represents the number of duplicated gene pairs. The black line is the distribution for the raw data, and the red line is the distribution for the data after the standardization and normalization of microarray data (SNOMAD).

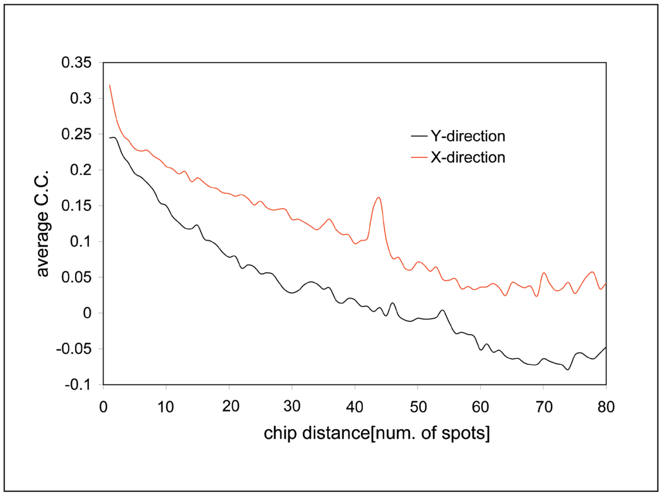

Figure 5. Average correlation coefficient distributions along the x-axis and y-axis on the chip. The distributions are calculated using the raw α-factor-arrested cell-cycle data set. The black line is the distribution along the y-axis, and the red line is the distribution along the x-axis. C.C., correlation coefficient.

assigned to the unique sequence. By applying the same amounts of the single-sequence cDNA (green dye) and sample cDNA (red dye) and normalizing the signal of the red channel by the green channel signal, one can partially remove the chip artifact (because this normalization also transforms the green signal to a constant value at all spots). Moreover, these ratios are proportional to mRNA concentration. Thus, one can compare expression levels between different genes, as with the GeneChips, and not simply compare the relative variation of the expression of a gene across experimental conditions. Note that placing the different probes that correspond to a single gene in random locations on the chip [as is done in the new U133 Affymetrix chips (1)] and estimating their average intensity do not wipe out this artifact.

In summary, we demonstrate a systematic spatial artifact that arises from microarray experiments. The source of the artifact is not fully understood. We show that a local mean normalization method is useful but cannot completely solve the problem. Finally, we propose experimental and analytical procedures to quantify and manage this artifact.

ACKNOWLEDGMENTS

J.Q., Y.K., and H.Y. contributed equally to this work. The authors are grateful to Dr. Nicholas Luscombe and Dr. Mike Snyder for useful discussion and acknoledge support from the National Institutes of Health (grant no. NP50 HG02357-01).

REFERENCES

1. **Alter, O., P.O. Brown, and D. Botstein.** 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA *97*:10101-10106.
2. **Ben-Dor, A., R. Shamir, and Z. Yakhini.** 1999. Clustering gene expression patterns. J. Comput. Biol. *6*:281-297.
3. **Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein.** 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*:14863-14868.
4. **Heyer, L.J., S. Kruglyak, and S. Yooseph.** 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. *9*:1106-1115.
5. **Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub.** 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA *96*:2907-2912.
6. **Tavazoie, S., J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church.** 1999. Systematic determination of genetic network architecture. Nat. Genet. *22*:281-285.
7. **Toronen, P., M. Kolehmainen, G. Wong, and E. Castren.** 1999. Analysis of gene expression data using self-organizing maps. FEBS Lett. *451*:142-146.
8. **Dudoit, S., Y.H. Yang, M. Callow, and T. Speed.** 2000. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical report no. 578 (http://stat-www.berkeley.edu/users/terry/zarray/Html/matt.html).
9. **Baldi, P. and A.D. Long.** 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics *17*:509-519.
10. **Fielden, M.R., R.G. Halgren, E. Dere, and T.R. Zacharewski.** 2002. GP3: GenePix post-processing program for automated analysis of raw microarray data. Bioinformatics (Oxford) *18*:771-773.
11. **Finkelstein, D., R. Ewing, J. Gollub, F. Sterky, J.M. Cherry, and S. Somerville.** 2002. Microarray data quality analysis: lessons from the AFGC project. Arabidopsis Functional Genomics Consortium. Plant Mol. Biol. *48*:119-131.
12. **Mutch, D.M., A. Berger, R. Mansourian, A. Rytz, and M.A. Roberts.** 2002. The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. BMC Bioinfomatics *3*:17.
13. **Tran, P.H., D.A. Peiffer, Y. Shin, L.M. Meek, J.P. Brody, and K.W. Cho.** 2002. Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. Nucleic Acids Res. *30*:e54.
14. **Tseng, G.C., M.K. Oh, L. Rohlin, J.C. Liao, and W.H. Wong.** 2001. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res. *29*:2549-2557.
15. **Yang, Y.H., S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed.** 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. *30*:e15.
16. **Colantuoni, C., G. Henry, S. Zeger, and J. Pevsner.** 2002. Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. BioTechniques *32*:1316-1320.
17. **DeRisi, J., V. Iyer, and P. Brown.** 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science *278*:680-686.
18. **Spellman, P., G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher.** 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell *9*:3273-3297.
19. **Zhu, G., P.T. Spellman, T. Volpe, P.O. Brown, D. Botstein, T.N. Davis, and B. Futcher.** 2000. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. Nature *406*:90-94.
20. **Cohen, B., R. Mitra, J. Hughes, and G. Church.** 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat. Genet. *26*:183-186.

Received 12 November 2002; accepted 11 April 2003.

Address correspondence to Mark Gerstein, MB&B Department, Bass 432A, 266 Whitney Avenue, Yale University, New Haven, CT 06520, USA. e-mail: mark.gerstein@yale.edu