Consolidated strategy for the analysis of microarray spike-in data

Matthew N. McCall and Rafael A. Irizarry*

Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205, USA

Received April 17, 2008; Revised June 19, 2008; Accepted June 20, 2008

ABSTRACT

As the number of users of microarray technology continues to grow, so does the importance of platform assessments and comparisons. Spike-in experiments have been successfully used for internal technology assessments by microarray manufacturers and for comparisons of competing data analysis approaches. The microarray literature is saturated with statistical assessments based on spike-in experiment data. Unfortunately, the statistical assessments vary widely and are applicable only in specific cases. This has introduced confusion into the debate over best practices with regards to which platform, protocols and data analysis tools are best. Furthermore, cross-platform comparisons have proven difficult because reported concentrations are not comparable. In this article, we introduce two new spike-in experiments, present a novel statistical solution that enables cross-platform comparisons, and propose a comprehensive procedure for assessments based on spike-in experiments. The ideas are implemented in a user friendly Bioconductor package: spkTools. We demonstrated the utility of our tools by presenting the first spike-in-based comparison of the three major platforms—Affymetrix, Agilent and Illumina.

INTRODUCTION

Assessing sensitivity presents a challenge for microarray technology because one needs experimental designs in which the correct outcome for a given measurement is known a priori. Spike-ins provide a way to do this and are therefore used extensively for assessment purposes (1–8). In fact, spike-in experiments will be integral to government-led projects that will help determine the practicality of microarrays in clinical applications. An example is the External RNA Control Consortium (ERCC) (9), led by the National Institute for Standards and Technology (NIST). Creating well-characterized, tested RNA spike-in controls is the first goal of the ERCC. Proper statistical analysis strategies for the data generated by these experiments are indispensable. Unfortunately, the statistical assessments for spike-in experiments presented in the literature vary widely and are not generally applicable. For reasons explained in detail below, comparisons across platforms are particularly problematic. In this article, we propose a consolidated strategy that builds on a widely used benchmark methodology (10): we assessed specificity and sensitivity in a way that can be easily related to practical performance. We propose a solution to the cross-platform problem and demonstrate its use by analyzing data from spike-in experiments performed by Affymetrix, Agilent and Illumina. The data were preprocessed with the most commonly used procedures, as described in the next section. We refer to the processed data (in log2 scale) as expression values.

An important fact that has been overlooked by previous assessments is that microarray performance largely depends on concentration levels (11). Assessments based on experiments for which spike-in concentrations lead to unusually high expression measurements have resulted in misleading conclusions (12). For this reason, it is essential that the distribution of observed expression for the spike-in transcripts reflects the distributions seen in typical experiments. Figure 1 shows the typical distribution of expression values for the background RNA for the three studied data sets. The tick marks on the x-axis represent the average expression at each reported spike-in level. This figure illustrates that the spike-in transcripts resulted in higher expression measurements, on average, than the background RNA transcripts. Furthermore, we see that, relative to their respective background RNA distributions, the Agilent and Illumina spike-ins have higher observed expression than those in the Affymetrix experiment. Previous work (11) suggests that comparing platform performance without correcting for this leaves Affymetrix at a disadvantage.

For spike-in experiments to be useful in a cross-study assessment, we need to understand how the reported nominal concentrations relate across data sets. The reported concentrations attempt to quantify the amount of spike-in RNA in a sample relative to the total amount of RNA. However, in our experience, the reported values are...
impossible to relate between the different manufacturers. One reason is that two approaches have been used: (i) adding prelabeled spike-ins to the target solution just before hybridization and (ii) adding spike-ins to the total RNA at the beginning of amplification. We prefer the second approach because it better imitates a real experiment (13). Specifically, it reproduces the technical variation due to cDNA synthesis, fragmentation, labeling and hybridization. However, the Affymetrix and Illumina data presented in this article are from experiments that followed the first approach, while the Agilent experiment followed the second approach. A molarity calculation is easy to perform in the first approach but difficult in the second because the only true known values in the experiment are the mass of the spike-in and the mass of total RNA. For this reason, Agilent reports nominal values in different units (relative concentration) than Affymetrix and Illumina [picomolar (pM)]. However, even when picomolar concentrations are reported we find that nominal concentrations do not map well across experiments (Table 1). We hope that the ERCC will solve this problem by standardizing protocols. Here, we propose a data-driven solution that permits cross-platform comparison with existing data sets.

## MATERIALS AND METHODS

### Experimental protocols

The platforms used were Affymetrix’s HGU133A GeneChip, Agilent’s 4x44K Whole Human Genome Oligo Microarray and Illumina’s Human-6 v2 Beadchip. The experiments for each platform were performed by the respective manufacturer. Each manufacturer followed different experimental procedures (Table 2). The raw data were preprocessed with the default procedures: Affymetrix was preprocessed using RMA (14); Agilent used background subtraction and normalized to the 75th percentile; Illumina used local background subtraction and quantile normalization (15).

### Relating nominal concentrations across data sets

Our solution to the problem of mapping was to replace each nominal concentration with the average log expression across arrays (ALE) for genes spiked in at that concentration. This approach assures that performance across the different manufacturers.

### Table 1. Nominal concentration to ALE mapping

| Reported nominal concentration | Average expression value | Proportion of genes below | ALE strata | SD |
|-------------------------------|--------------------------|----------------------------|------------|----|
| **Affymetrix (pM)**           |                          |                            |            |    |
| 0.125                         | 5.1                      | 0.35                       | Low        | 0.87 |
| 0.250                         | 5.2                      | 0.38                       | Low        | 0.90 |
| 0.500                         | 5.3                      | 0.40                       | Low        | 0.74 |
| 1.000                         | 5.7                      | 0.48                       | Low        | 0.72 |
| 2.000                         | 6.4                      | 0.62                       | Med        | 0.82 |
| 4.000                         | 7.1                      | 0.73                       | Med        | 0.79 |
| 8.000                         | 7.8                      | 0.82                       | Med        | 0.68 |
| 16.000                        | 8.4                      | 0.89                       | Med        | 0.67 |
| 32.000                        | 9.3                      | 0.94                       | Med        | 0.79 |
| 64.000                        | 10.2                     | 0.97                       | Med        | 0.72 |
| 128.000                       | 11.2                     | 0.99                       | Med        | 0.54 |
| 256.000                       | 12.0                     |                           | High       | 0.49 |
| 512.000                       | 12.6                     |                           | High       | 0.51 |
| **Agilent (relative concentration)** |                          |                            |            |    |
| 0.20                          | 4.6                      | 0.34                       | Low        | 1.35 |
| 2.00                          | 4.8                      | 0.36                       | Low        | 1.40 |
| 20.00                         | 6.3                      | 0.49                       | Low        | 0.63 |
| 200.00                        | 9.1                      | 0.73                       | Med        | 0.47 |
| 666.67                        | 10.9                     | 0.86                       | Med        | 0.36 |
| 2000.00                       | 12.5                     | 0.94                       | Med        | 0.3  |
| 6666.67                       | 14.2                     | 0.97                       | Med        | 0.21 |
| 20000.00                      | 15.6                     | 0.99                       | Med        | 0.16 |
| 66666.67                      | 17.1                     | 1.00                       | High       | 0.31 |
| 200000.00                     | 18.0                     | 1.00                       | High       | 0.25 |

| **Illumina (pM)**             |                          |                            |            |    |
| 0.01                          | 2.9                      | 0.38                       | Low        | 0.39 |
| 0.03                          | 3.0                      | 0.49                       | Low        | 0.46 |
| 0.10                          | 3.9                      | 0.68                       | Med        | 0.83 |
| 0.30                          | 5.4                      | 0.76                       | Med        | 1.51 |
| 1.00                          | 7.5                      | 0.85                       | Med        | 1.26 |
| 3.00                          | 9.5                      | 0.94                       | Med        | 1.05 |
| 10.00                         | 11.4                     | 0.98                       | Med        | 1.01 |
| 30.00                         | 12.8                     | 1.00                       | High       | 0.81 |
| 100.00                        | 14.1                     | 1.00                       | High       | 0.65 |
| 300.00                        | 14.8                     | 1.00                       | High       | 0.34 |
| 1000.00                       | 15.0                     | 1.00                       | High       | 0.26 |

This table contains summary measures specific to each nominal spike-in level. The first column shows the nominal concentrations as originally reported. The second column shows the average of all observed expression values associated with the row’s nominal concentration. The third column shows the proportion of background RNA with expression values less than the average expression value. The fourth column shows the ALE strata (defined in text) associated with the row’s nominal concentration. Finally, the fifth column shows the SD of all observed expression values associated with the row’s nominal concentration.
assessments based on spike-in data are related to expression measurements that are defined consistently across platforms: low, medium and high ALE values correspond to low, medium and high observed expression values, respectively (Table 1).

**Accuracy assessment**

With the ALE values in place, we were ready to adapt some of the existing statistical assessments to cross-platform comparisons. We started with a basic assessment of accuracy: the *signal detection slope* (10). Microarrays are designed to measure the abundance of sample RNA. In principle, we expect a doubling of nominal concentration to result in a doubling of observed intensity. In other words, on the log 2 scale, the slope from the regression of expression on nominal concentration can be interpreted as the expected observed difference when the true difference is a fold change of 2. Thus, an optimal result is a slope of one, and values higher and lower than one are associated with over and under estimation, respectively (Figure 2).

**ALE strata**

It has been noted that at very high and very low concentrations one typically observes lower slopes compared to those seen at medium concentrations (11). To address this, we consider the signal detection slopes for genes spiked-in at low, medium and high ALE values (Figure 2). We implemented a data-driven approach to selecting these two cut-offs.

We defined $f$ to be the function that maps nominal log concentration $x$ to expected observed concentration $f(x)$. Using a cubic spline, fitted to the observed data, we obtained a parametric representation of $f$. We then looked for concentrations for which clear changes in sensitivity occurred, i.e. values of $x$ with large slope changes. Note that large changes in slope result in local maxima in the absolute value of the second derivative of $f$. For each platform, the absolute value of the second derivative $f''$ showed two clear local maxima (Supplementary Figure 1). For each platform, we mapped each concentration $x$ to its corresponding empirical percentile $\Phi(x)$ and plotted $|f''(x)|$ against $\Phi(x)$ (Supplementary Figure 2). The percentiles that maximized the slope change were similar across platforms. The modes for the average curve were 0.615 and 0.993. Therefore, for the purpose of this comparison, we assigned as low, ALE values less than the 60th percentile of the distribution of background RNA. Similarly, we defined as high ALE values above the 99th percentile. The remaining ALE values, between the 60th and 99th percentile, were denoted as medium. Our choice of cut-points was further motivated by observing that for the Affymetrix data the 60th percentile provided a good cut-off for distinguishing genes called present from genes called absent (Supplementary Figure 3).

**Precision assessments**

To complete our comparison, we needed to assess specificity. Because the majority of microarray studies rely on

| Table 2. Description of data sets |
|----------------------------------|
| **Affymetrix** | **Agilent** | **Illumina** |
| Background RNA | HeLa complex cRNA | Human Osteosarcoma (MG-63) purchased from Ambion (Cat. # 7868) | Human Liver purchased from Ambion (Cat. # 7960) |
| Spike-in production | 30 cDNA clones isolated from a lymphoblast cell line, eight artificially engineered, four eukaryotic controls from the polyA spike control kit | In vitro synthesized, polyadenylated transcripts derived from the Adenovirus E1A gene | In vitro transcription of cloned bacterial and viral genes |
| Background correction | RMA | Background subtraction and spatial-detrending | Local background subtraction |
| Normalization | RMA | Normalized to the 75th percentile on each microarray | Quantile normalization |

This table gives a brief comparison of the three data sets used in this analysis.

**Figure 2.** Observed versus nominal values. For each of the three platforms, expression values are plotted against the log (base 2) of the reported nominal concentration. The regression slope obtained utilizing all the data and the regression slopes obtain within each ALE value strata are shown. The slope of each line is reported in the legend. The vertical lines divide the ALE strata.
relative measures (e.g. fold change) as opposed to absolute ones, we focused on the precision of the basic unit of relative expression: log-ratios. We adapted the precision assessment of Cope et al. (10) that focused on the variability of log-ratios generated by comparisons expected to produce log-ratios of 0. Our set of comparisons was created by making all possible comparisons between spiked-in transcripts across arrays in which they had the same nominal concentration and from all possible comparisons within the background RNA. We referred to this group of comparisons as the Null set. The SD of these log-ratios served as a basic assessment of precision and has a useful interpretation: it is the expected range of observed log-ratios for genes that are not differentially expressed. Table 3 and Figure 3 show results for the three platforms.

Because specificity varies with nominal concentration (11), we stratified these comparisons into low, medium and high ALE values. In Figure 3, many outliers were observed on each platform. This was expected given the documented problem of cross-hybridization. Because a platform with larger SD and small outliers might be preferable to the one with a smaller SD but large outliers, we included the 99.5th percentile of the null distribution as a second summary assessment of specificity. Note that in a typical experiment close to 0.5% of null genes are expected to exceed this value, which translates to approximately 100 genes on whole-genome arrays. Figure 3 also includes comparisons of spike-ins expected to yield a certain fold change. These serve to further demonstrate the variability of relative expression across ALE strata. They also serve as a rough illustration of the accuracy of log-ratios for each ALE strata and each platform.

**Performance assessments**

Precision and accuracy assessments on their own may not be of much practical use. However, the summary statistics described earlier (Table 3) can be easily combined to answer any practical question, as long as it can be posed in a statistical context. We focus on two summaries related to the common problem of detecting differentially expressed genes. Note that we purposely developed summaries that do not directly penalize for a lack of accuracy and precision as long as the real differences are detected. However, as expected, detection ability was highly dependent on accuracy and precision.

For the first example, we computed the chance that, when comparing two samples, a gene with true log fold change \( \Delta = 1 \) will appear in a list of the top 100 genes.
(highest log-ratios). We refer to this quantity as the probability of being at the top (POT) and recommend computing it separately in each ALE strata. Specifically, we assume that the log-ratios in each ALE strata follow a normal distribution with mean and variance estimated from the data (accuracy slope and SD in Table 3) and compute the probability that a random variable from that distribution exceeds the 99.5th percentile of the null distribution.

As a second example, we computed the expected size of a gene list one would have to consider to find $n$ genes that have a true log fold change $\Delta$. To perform this calculation, we assumed $m_1$ genes were differentially expressed and $m_0$ were not. Note that $m_1 + m_0$ is the number of genes on the array. Furthermore, we assumed that the true log-ratios in each ALE strata followed a normal distribution with mean and variance estimated from the data (accuracy slope and SD in Table 3). The empirical distribution was used for the null genes. With these assumptions in place, we computed the gene list size for $n = 10$, $m_1 = 100$ and $m_0 = 10,000$, we calculate the gene list size, $N$, required to obtain $n = 10$ true fold changes (Table 3). We refer to this quantity as the gene-list needed to detect $n$ true-positives (GNN). Again, we recommend computing it separately in each ALE strata.

### Imbalance measure

Those interested in taking advantage of our methodology should know that an important requirement is a spike-in experimental design that does not confound nominal concentrations and genes. A large source of variability in microarray data is the probe-effect (16) and these vary across platforms. We fitted an analysis of variance (ANOVA) model to describe the probe effect for each platform (Table 4). Note that if nominal concentrations are confounded with genes, it becomes impossible to separate differences due to signal detection from differences in probe affinities. Many of the previously published spike-in experiments suffer from this confounding effect. To quantify design imbalance, we used the following measure of imbalance developed by Wu (17):

$$IB = \sum_{i=1}^{p} \lambda_i \sum_{u_i} \left( \sum_{t=1}^{T} n_i^t(u_i) - \frac{1}{T} n_i^2(u_i) \right)$$

where $i$ denotes each covariate, $\lambda_i$ an optional weight associated with each covariate, $u_i$ are the possible levels for covariate $i$, $t$ represents the treatment levels, $n_i^t(u_i)$ is the number of units with its $i$-th covariate at level $u_i$ receiving treatment $t$, and $n(u_i)$ is the total number of units with its $i$-th covariate at level $u_i$ (17). In our case, the two covariates are probe and array, and the treatment is nominal concentration. Since imbalance is defined as a weighted sum of the imbalance due to each covariate, we chose to report the probe and array imbalance separately to give a better understanding of the source of the imbalance in each design. In order to not penalize large designs, we divided the probe imbalance by the number of probes and the array imbalance by the number of arrays. These results are included in Table 4.

### Table 4. ANOVA results

| Platform       | Affymetrix (RMA) | Affymetrix (MAS5) | Agilent | Illumina |
|----------------|------------------|-------------------|---------|----------|
| Concentration effect | 2.48             | 2.77              | 4.53    | 2.19     |
| Probe effect    | 0.54             | 0.55              | 0.44    | 0.38     |
| Array effect    | 0.17             | 0.17              | 0.19    | NA       |
| Measurement error| 0.47             | 0.72              | 0.69    | 0.54     |
| Probe imbalance | 0                | 0                 | 3.60    | 0        |
| Array imbalance | 0                | 0                 | 0       | 1059.67  |

To understand the variability contributed by differences in nominal concentrations, probe effect and array, we fitted a three-way ANOVA model containing only main effects to the expression values from the spike-in transcripts. The estimated SD of each effect is shown in the first three rows. The fourth row shows the SD of the error term. Finally, a measure of the amount of confounding between nominal concentration and the other two effects is included in rows five and six. We use the measure presented by Wu (17). An optimal design, such as a Latin Square, will have a measure of 0 for each imbalance. The more confounding the larger these values. Note, the large imbalance due to array in the Illumina design. In this experiment array and nominal concentration were completely confounded. However, because the array effect is small (the arrays are normalized) this was not as much of a problem. In the Agilent experiment, there is a small amount of confounding between probe and concentration because a Latin Square design was used with a single concentration/gene combination missing.

We have developed a software package that permits quick and easy creation of plots and tables such as those presented here. The software is freely available as the spkTools package from the Bioconductor Project (18). This package defines a new S4 class that extends the ExpressionSet class (18) to include a matrix of nominal concentrations; this new class is called SpikeInExpressionSet. The functions implemented in this package take an object of this type as their input and automatically produce the tables and plots presented in this paper. Of particular interest is a function named spkAll, which is a wrapper function for all the functions contained in this package. When run on a SpikeInExpressionSet object, it produces the full complement of tables and plots shown in this article and saves them with easily recognizable file names. Although this package was designed with the intent of producing the full array of results for each experiment, the functions can also be applied separately with a few exceptions where the output of one function is required as the input of another. Further details and examples outlining the use of these functions can be found in the help files accompanying the package.

### RESULTS

The ANOVA analysis (Table 4) revealed that all platforms have similar sized probe effects. This underscores the importance of balancing genes and concentration levels. The Agilent experiment had a small imbalance (Table 4). This was because Agilent used one less concentration mixture than the number of spike-in probes. The Illumina array had a very large array imbalance because array and concentration were completely confounded.
However, because all data sets were normalized we expected the array effect for Illumina to be small, as with Affymetrix and Agilent. This type of confounding is therefore less problematic.

Figure 2 demonstrated that Agilent performed best with regard to accuracy in all concentration bins. While Illumina performed better than Affymetrix in the medium concentrations, Affymetrix performed better in the low and high concentrations. Affymetrix was most consistent across all bins. If we had looked at only the overall slope, Illumina would have appeared to perform best because fold changes are overestimated in the medium concentrations. Figure 2 also shows the changing relationship between expression and nominal concentration. For all three platforms, the slope is small at low nominal concentrations, larger at high concentrations, and largest at medium concentrations. However, the difference between these slopes and the nominal concentrations at which the shift between bins occurs varies across platforms. It is this fact that illustrates why it is crucial to view nominal concentration and expression as platform dependent measures.

Figure 3 highlights two important findings: (i) precision depends strongly on concentration with higher variability observed for low concentrations, (ii) Affymetrix, which had the worst accuracy, has the best precision, especially for low concentrations where the difference was substantial.

In terms of the POT and GNN assessment, Affymetrix outperforms Agilent which outperforms Illumina: the gene list size in the low/medium/high strata for Affymetrix were 37/34/25, for Agilent they were 682/38/26 and for Illumina they were 1489/60/46 (Table 3). To provide a graphical version of this summary, we included boxplots of observed log-ratios for comparison with nonzero nominal log-fold changes $\Delta > 0$ (Figure 3). Due to the different designs, using the same expected log-fold change, $\Delta$, for all platforms was not possible. We used the closest possible values instead: $\log_2(4)$ for Affymetrix and $\log_2(3)$ for Agilent and Illumina. Log ratio ($M$) versus average intensity ($A$) plots also depict both accuracy and precision and are included as Supplementary Figure 4.

DISCUSSION

We have described a general assessment procedure for microarray data based on spike-in experiments and demonstrated how the procedure can be used to compare across different experiments and microarray platforms. A novel aspect of the approach is that we independently assess performance at low, medium, and high concentrations using ALE values: an empirically constructed mapping between nominal concentration and observed expression. This mapping is important because nominal concentrations can not be interpreted in the same way in all experiments. In our approach, measurements are interpreted relative to the distribution of background RNA expression.

Our results demonstrate that while Agilent and Illumina had better overall accuracy, Affymetrix has better precision. In the medium and high strata, Affymetrix and Agilent performed similarly, and better than Illumina, according to the POT and GNN measures. In the low strata, Affymetrix greatly outperformed Agilent and Illumina. Affymetrix’s advantage was due to the smaller number of outliers (Table 3 and Supplementary Figure 1). Note that to keep the article focused, we considered a basic analysis approach based on fold change. However, the spkTools package can be used for more elaborate platform comparisons. For example, to help reduce outliers, we filtered genes called absent or undetected by the manufacturer’s software. Because various noisy comparisons are no longer considered, this approach improves specificity in the low strata. However, sensitivity is made worse because true differences are accidentally filtered away (Supplementary Table 1). Problems with these detection calls have been documented (19).

As we previously described the 60th and 99th percentiles provided the best cut-points for this analysis; however, that need not be the case in future analyses. For this reason, the spkTools package permits the user to choose any two percentiles to use as the cut-points. We recommend that the optimal cut-points for a future data set be determined in a manner similar to what we described.

It is important to note that the microarray products compared here are of different generations, with Affymetrix’s the oldest and Agilent’s the newest. Also, the spike-in targets and background RNA vary between platforms (Table 2). Finally, different preprocessing algorithms will result in differences in performance (16). To illustrate this, we ran Affymetrix data processed with the manufacturers default MAS 5.0 through our assessment (Table 3 and Supplementary Figure 6). An interesting finding was that with MAS 5.0, instead of RMA, Affymetrix no longer had an advantage in the low strata. We expect results for Agilent and Illumina to improve with the development of novel preprocessing algorithms for these technologies by the scientific community. Because we expect a large increase in spike-in experiments and preprocessing algorithms, we developed the spkTools package to permit quick and easy creation of plots such as those presented here.

Spike-in experiments have been criticized for producing artificial data with little resemblance to real data produced by a typical experiment. A particular limitation of the data shown here is the use of technical replicates: the data fails to incorporate the biological variation present in most experiments. However, acceptable sensitivity and specificity measures, determined by spike-in experiments such as those presented here, are a minimal requirement for a microarray platform. A technology not performing well in our assessment will not perform well in the more complicated setting of real experiments. A strength of our spike-in data is that they permit a focused assessment based on the most basic attributes of this technology. Furthermore, we expect vastly improved spike-in experiments, e.g. using biological replicates, to emerge in large numbers once the ERCC makes it first formal recommendation. We have therefore developed a general tool that can be readily applied to data from these experiments. Furthermore, the strategy we described and used successfully to compare the three main microarray platforms using Latin-Square spike-in...
experiments for the first time, can serve as a blueprint for future methods and analyses.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Online. Raw data and annotation files are available from http://rafalab.org.

ACKNOWLEDGEMENTS

We thank Agilent, and especially Anne Bergstrom Lucas, for providing the Agilent spike-in data and insightful discussions on spike-in experiments. We thank Illumina, and especially Shawn Baker, for sharing the Illumina spike-in data. We also thank Affymetrix for making their spike-in experiments public. This work was supported by National Institutes of Health (1R01GM083084-01 to R.I., 1R01RR021967-01A2 to R.I., T32GM074906 to M.M.).

Conflict of interest statement. None declared.

REFERENCES

1. Fodor, S., Read, J., Pirrung, M., Stryer, L., Lu, A. and Solas, D. (1991) Light-directed, spatially addressable parallel chemical synthesis. Science, 251, 767-773.
2. Fodor, S., Rava, R., Huang, X., Pease, A., Holmes, C. and Adams, C. (1993) Multiplexed biochemical assays with biological chips. Nature, 364, 555-556.
3. Schena, M., Shalon, D., Davis, R. and Brown, P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270, 467-470.
4. Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chec, M., Mittmann, M., Want, C., Kobayashi, M., Horton, H., et al. (1996) DNA expression monitoring by hybridization of high density oligonucleotide arrays. Nat. Biotechnol., 14, 1675-1680.
5. Higuchi, R., Fockler, C., Dollinger, G., Watson, R. and Kinetic, P. (1993) Analysis: real-time monitoring of DNA amplification reactions. Biotechnology, 11, 1026-1030.
6. Heid, C., Stevens, J., Livak, K. and Williams, P. (1996) Real time quantitative PCR. Genome Res., 6, 986-994.
7. Wittwer, C., Herrmann, M. G., Moss, A. A. and Rasmussen, R. P. (1997) Continuous fluorescence monitoring of rapid cycle DNA amplification. Biotechniques, 22, 130-139.
8. Cronin, M., Ghosh, K., Sistare, F., Quackenbush, J., Vilk, V. and O’Connell, C. (2004) Universal RNA reference materials for gene expression. Clin. Chem., 50, 1464-1471.
9. Baker, S., Bauer, S., Beyer, R., Brenton, J., Bromley, B., Burrill, J., Causton, H., Conley, M., Elespuru, R., Fero, M., et al. (2005) The external RNA controls consortium: a progress report. Nat. Methods, 2, 731-734.
10. Cope, L., Irizarry, R., Jaffe, H., Wu, Z. and Speed, T. (2004) A benchmark for Affymetrix GeneChip expression measures. Bioinformatics, 20, 323-331.
11. Irizarry, R., Wu, Z. and Jaffe, H. (2006) Comparison of Affymetrix GeneChip expression measures. Bioinformatics, 22, 789-794.
12. Irizarry, R., Cope, L. and Wu, Z. (2006) Feature-level exploration of a published Affymetrix GeneChip control dataset. Genome Biol., 7, 204.
13. Tong, W., Lucas, A., Shippy, R., Fan, X., Fang, H., Hong, H., Orr, M., Chu, T., Guo, X., Collins, P. et al. (2006) Evaluation of external RNA controls for the assessment of microarray performance. Nat. Biotechnol., 24, 1132-1139.
14. Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs, B. and Speed, T. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res., 31, e15.
15. Kuhn, K., Baker, S., Chudin, E., Lieu, M., Oester, S., Bennett, H., Rigault, P., Barker, D., McDaniel, T. and Chee, M. (2004) A novel, high-performance random array platform for quantitative gene expression profiling. Genome Res., 14, 2347-2356.
16. Irizarry, R., Warren, D., Spencer, F., Kim, L., Biswal, S., Frank, B., Gabrielson, E., Garcia, J., Geoghagan, J., Germino, G. et al. (2005) Multiple-laboratory comparison of microarray platforms. Nat. Methods, 2, 345-350.
17. Wu, C. (1981) Iterative construction of nearly balanced assignments I: categorical covariates. Technometrics, 23(1), 37-44.
18. 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.
19. Zilliox, M. and Irizarry, R. (2007) A gene expression bar code for microarray data. Nat. Methods, 4, 11-913.