Nucleic Acids Research, 2005, Vol. 33, No. 19 ***e172***
doi:10.1093/nar/gni167

Protocols for the assurance of microarray data quality and process control

L. D. Burgoon[1,3,4], J. E. Eckel-Passow[5], C. Gennings[6], D. R. Boverhof[2,3,4], J. W. Burt[2,3,4], C. J. Fong[2,3,4] and T. R. Zacharewski[2,3,4,*]

[1]Department of Pharmacology and Toxicology, [2]Department of Biochemistry and Molecular Biology, [3]National Food Safety and Toxicology Center, [4]Center for Integrative Toxicology, Michigan State University, East Lansing, MI 48824, USA, [5]Department of Health Sciences Research, Mayo Clinic Cancer Center, Rochester, MN 55905, USA and [6]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA 23298, USA

Received March 23, 2005; Revised June 9, 2005; Accepted October 4, 2005

ABSTRACT

Microarrays represent a powerful technology that provides the ability to simultaneously measure the expression of thousands of genes. However, it is a multi-step process with numerous potential sources of variation that can compromise data analysis and interpretation if left uncontrolled, necessitating the development of quality control protocols to ensure assay consistency and high-quality data. In response to emerging standards, such as the minimum information about a microarray experiment standard, tools are required to ascertain the quality and reproducibility of results within and across studies. To this end, an intralaboratory quality control protocol for two color, spotted microarrays was developed using cDNA microarrays from *in vivo* and *in vitro* dose-response and time-course studies. The protocol combines: (i) diagnostic plots monitoring the degree of feature saturation, global feature and background intensities, and feature misalignments with (ii) plots monitoring the intensity distributions within arrays with (iii) a support vector machine (SVM) model. The protocol is applicable to any laboratory with sufficient datasets to establish historical high- and low-quality data.

INTRODUCTION

Microarray technology provides the ability to simultaneously measure the expression of thousands of genes in a cell, tissue or model of interest. However, numerous potential sources of experimental variation (1,2) have raised concerns regarding assay consistency, and data quality which confounds the ability to compare datasets between independent investigators and undermines the utility of intralaboratory (i.e. local), interlaboratory (i.e. collaborative center) or global scale (i.e. public repository) data sharing and exchange efforts (3,4). Consequently, quality assurance and control protocols that assess the reproducibility of data by identifying deviations or abnormal trends in assay performance and data quality are required.

A quality assurance plan (QAP) is a standard operating procedure (SOP) that describes the necessary steps to ensure the process of array production, hybridization and analysis are of high quality. QAPs include control methods which are used to test and monitor the quality of the entire process. Whereas quality control methods seek to identify low quality products, QAPs integrate information to determine why low quality products were produced, and to establish best practices to prevent future low quality events. The success of a QAP should be measured in terms of the ability to identify low quality products, and to improve the production process to lower the rate of low quality occurrences. However, these are inherently functions of the production processes, and thus subject to human error.

Although several quality assurance and control methods have been proposed, criteria for differentiating high- from low-quality microarrays is lacking, leaving assessment open to interpretation. Many methods attempt to address this impediment through a variance-based statistical method, however they suffer from a lack of training, as the method solely tests the hypothesis of deviation from the rest of the population, and fail to judge data based on prior knowledge. Therefore, arrays that are technically of low quality (i.e. high background, low feature signal intensity, misaligned features or inappropriately distributed feature intensity values) can still be labeled as high quality, if they belong to a larger population of low-quality arrays.

*To whom correspondence should be addressed at Department of Biochemistry and Molecular Biology, Michigan State University, 223 Biochemistry Building, Wilson Road, East Lansing, MI 48824-1319, USA. Tel: +517 355 1607; Fax: +517 353 9334; Email: tzachare@msu.edu

In lieu of these more complicated quality assurance and control methods, data quality has been reported in terms of sample clustering by assessing whether biological replicates cluster together (5). Although this methodology determines whether or not biological replicates exhibit similar behavior, it provides minimal insight into the technical quality of the assay (i.e. these are microarrays of high quality). For example, similarly treated biological replicates may cluster together, or yield similar patterns, in light of poor technical quality (e.g. high background and narrow dynamic range). Moreover, this method may yield false-negative results in a background of extensive biological variation.

In addition, quality assessments can be stratified to the feature (6,7), subgrid or block (8) or microarray (9,10) level. Although examination of each stratum is crucial, a comprehensive analysis strategy based on all strata would be advantageous. Thus, the most robust, comprehensive quality assurance and control protocol would incorporate aspects of training by using historical datasets (HDS) of known quality, provide analysis at all microarray quality strata, and diagnose possible sources of poor quality data that could be corrected and addressed to minimize future problems (i.e. quality assurance).

In this report, a three step intralaboratory quality control protocol is proposed to assess spotted microarray data quality as a first step towards ensuring publicly accessible data is of high quality. Global feature and background signal intensities as well as signal-to-noise ratios (SNRs) are first assessed to identify problems with raw microarray data quality (Division 1). The feature identification process, commonly referred to as gridding, is then computationally examined to identify potentially misaligned features, which can be corrected to minimize potential downstream errors in normalization and functional assignment (Division 2). Finally, a more in-depth assessment of raw and normalized data distributions is utilized to ensure that a sufficient dynamic range has been achieved for subsequent analyses (Division 3). A total of 388 time-course and dose-response two color cDNA microarray datasets are used to establish high- and low-quality HDS and to demonstrate the utility of the protocol.

MATERIALS AND METHODS

Creation of the HDS, test and validation sets

A 388 datasets, derived from *in vivo* and *in vitro* dose-response and time-course experiments using sequence verified cDNA microarrays were used to create both high- and low-quality HDS. Further details on microarray assay procedures are available at http://dbzach.fst.msu.edu/. Microarrays were scanned using an Affymetrix 428 scanner, and images were quantified using GenePix v5.0 or v5.1.

Global statistics are calculated as:

$$\bar{x}_d = \frac{1}{n}\sum_{i=1}^{n} x_{di}$$

where d represents the dye (Cy3 or Cy5), n represents the number of features on the array, and x_{di} represents the median feature intensity (either feature signal or background from the image analysis software) for the dth dye and the ith feature.

The HDS consists of 155 microarrays that were further classified as high- (87 microarrays) or low- (68 microarrays) quality based on corroboration by quantitative real-time PCR (QRT–PCR) ($P < 0.05$ for the correlation of the gene expression pattern of selected genes), low feature background intensity, congruent distributions of data points and detection of comparable numbers of features. The background feature intensity does not have a threshold *per se*, rather it is based on visual inspection for high overall signal and anomalies such as smears, waves and excessive dust, the ratio of signal to background being >20, the number of identified features, where at least 95% of the features are detectable, and the distribution of intensity values must be comparable across the experiment. Examples of high and low quality images for each criteria are provided as Supplementary Data to further assist in defining the thresholds we initially used to establish our historical training set (HDS). Arrays not found to have the desired characteristics were categorized as low quality. Quality assignments are not a weighted vote approach, but rather an all or nothing voting scheme, where high-quality arrays must meet all of the qualifications listed, and are specific to our HDS. The training set was derived from a random sampling of both high- and low-quality datasets to form a high- (44 microarrays) and low- (40 microarrays) quality training sets.

The validation dataset consisted of the 233 arrays not included in the HDS. The quality of these arrays was assessed in the same manner as the HDS, resulting in 174 high- and 59 low-quality arrays (Figure 1).

Division 1 analysis

Predictive variables include any parameter of interest to the investigator that may be indicative of quality. For example, these variables may include (i) the mean feature intensity across the array for each dye, (ii) the mean background intensity across the array for each dye, (iii) the mean ratio of the feature and background intensities, (iv) atmospheric ozone concentration, (v) laser intensity, (vi) the interquartile range (see Division 3), (vii) percent saturated features and (viii) percent undetected features.

To automate microarray quality classification, a support vector machine (SVM) model is trained on a set of data (i.e. the training set), and this model is validated against a larger, independent set of data (i.e. the validation set). To find the most optimum set of features, or variables, for the SVM, an expected cost of misclassification function (ECM) was created. The best SVM model is determined to be the one that minimizes the ECM function,

$$\begin{aligned}\mathrm{ECM} = {} & c(\mathrm{HQ}|\mathrm{LQ})P(\mathrm{HQ}|\mathrm{LQ})p(\mathrm{LQ}) \\ & + c(\mathrm{LQ}|\mathrm{HQ})P(\mathrm{LQ}|\mathrm{HQ})p(\mathrm{HQ})\end{aligned}$$

where c(HQ|LQ) is the cost of classifying a low-quality (LQ) array as high-quality (HQ), P(HQ|LQ) is the probability of classifying a LQ array as HQ, p(LQ) is the a priori probability of being LQ, c(LQ|HQ) is the cost of classifying a HQ array as LQ, P(LQ|HQ) is the probability of classifying a HQ array as LQ, p(HQ) is the *a priori* probability of being HQ, and given the constraint that $p(\mathrm{LQ}) + p(\mathrm{HQ}) = 1$. The cost function represents a real cost, of misclassification, and will be specific to each laboratory. For example, in our laboratory we estimate that the c(LQ|HQ) is ~$100, the estimated cost of repeating a

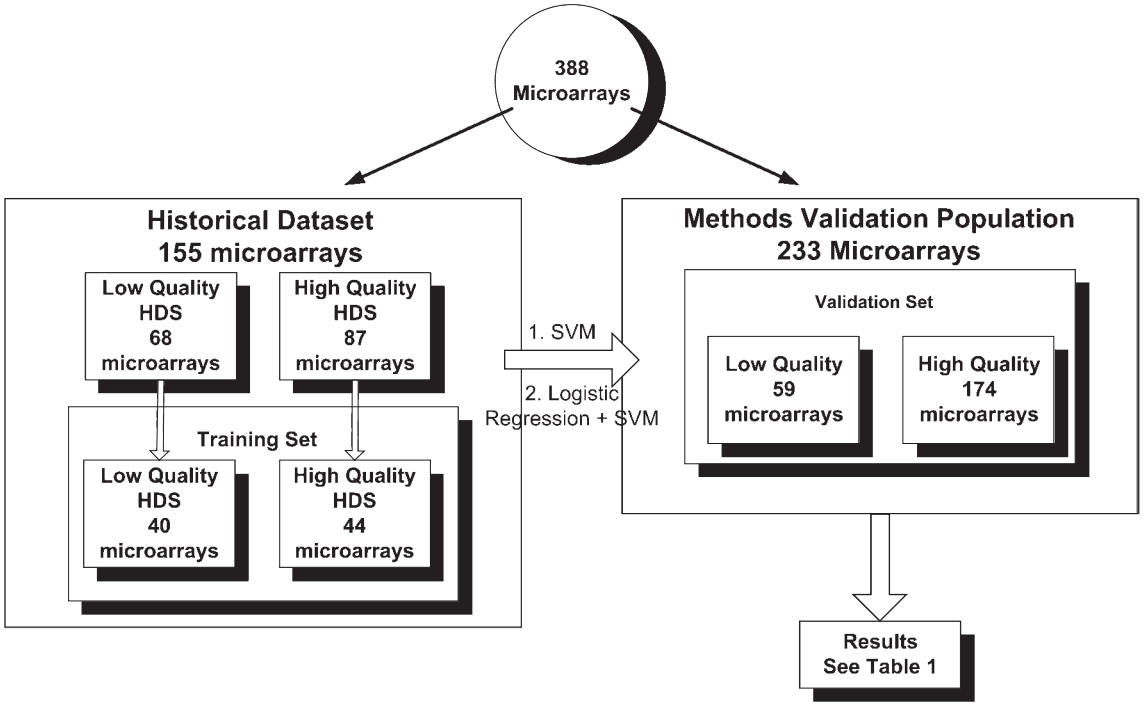

Figure 1. Historical, training and validation datasets. The complete dataset of 388 microarrays was divided into two sets, the HDS ($n = 155$) and the validation set ($n = 233$). Both of these datasets were further subdivided into high- and low-quality datasets. The high- and low-quality HDS were further subdivided into their respective training sets by random sampling. The SVM models (SVM and logistic regression + SVM) were trained on the same training set data, and validated against the same validation datasets. The results of the validations are summarized in Table 1.

microarray within our laboratory, whereas the c(HQ|LQ) is ~$500, the estimated cost of verifying the expression of false positive genes through QRT–PCR, and following false leads. The P(HQ|LQ) and P(LQ|HQ) are functions of the SVM model, while the a priori probabilities are qualities of the laboratory. In our laboratory, the a priori probability of an array being HQ is ~0.80.

The SVM model that minimizes the ECM is used for all future classification purposes. As microarray data becomes available (i.e. scanned and quantified) the resultant SVM model was used to classify microarrays as either high or low quality. High-quality microarrays continue through the protocol, while low-quality microarrays were flagged for repeat experiments. All data was stored for future inclusion into the HDS.

The SVM training and analysis were performed using the e1071 package in R v1.8.1 using a radial basis kernel. Details of SVM implementation are given in its documentation.

Division 2 analysis

Feature alignment was assessed using a loess non-parametric regression procedure that was originally developed as a normalization method to estimate bias on a per array, print-tip or subgrid, and channel basis, and is visualized by MA-plots. Feature alignment is analyzed using a variant of the standard MA-plot (11), referred to as a modified MA-plot (12). With respect to the modified MA-plot the true signal intensity for the ith feature is either estimated as the average signal intensity across all arrays, dyes and treatments ($\hat{\mu}_i$) or as the signal intensity across all arrays and dyes for each of the j treatment groups separately ($\hat{\mu}_{ij}$) for a particular experiment. The choice between using $\hat{\mu}_i$ versus $\hat{\mu}_{ij}$ is discussed in detail in (12). Thus, the estimated true signal intensity is a substitute for the A-term in the modified MA-plot. The M-term estimates the bias associated with using $\hat{\mu}_i$ or $\hat{\mu}_{ij}$ to estimate the true signal intensity such that M is equal to the difference between each signal intensity with its corresponding estimated true signal intensity. After computing the estimated true signal intensity and the bias, a modified MA-plot is constructed separately for every array and a non-parametric regression smoother is fit to each print-tip on the corresponding array individually. If the non-parametric regression smoother for a particular print-tip, or for a subset of print-tips, is an obvious outlier, feature alignment is investigated. All procedures were performed in SAS v8.2.

Division 3 analysis

Intensity distribution was assessed using box-and-whisker plots on a per array basis. Line plots demonstrating trends in global mean feature intensity, global mean background intensity and the count of saturated features were created depicting upper control limits (UCLs) and lower control limits (LCLs) for each metric. Acceptable numbers of saturated features have been historically established in this laboratory to be 1–2% of the total number of features. To assist in quality analysis it is generally useful to group microarrays performed on the same date together when plotting to identify temporal trends. All procedures were performed in SAS v8.2.

RESULTS

Figure 2 provides an overview of the microarray data quality control protocol which is divided into General Quality Metrics (Division 1), Feature Alignment (Division 2), and Distributional Alignment (Division 3). Two additional divisions are included to place the protocol into context within the overall data management scheme.

Establishment of high- and low-quality HDS

High- and low-quality HDS were created to anchor quality assessments to arrays of known quality to prevent inappropriate assessment of arrays as high quality due simply to low variance within the study. High quality was defined empirically based on corroboration by a complementary technology (e.g. QRT–PCR), low feature background intensity, congruent distribution of data points and detection of a comparable number of identified features. For example, among high-quality arrays, QRT–PCR corroborates >80% of the gene expression trends exhibited by arrays (13–15). Arrays not found to have the desired characteristics in all of the above categories were labeled as low quality.

The HQ-HDS is based on a random sampling of the high-quality microarrays from all investigators within our laboratory (HQ-HDS: $n = 87$), and a LQ-HDS similar to the HQ-HDS, but representing a random sampling of the low-quality microarrays (LQ-HDS: $n = 68$) from an overall total of 388 time-course and dose-response two color cDNA microarrays. Each HDS consists of the Cy3 and Cy5 global mean feature signal intensity (where global refers to the entire microarray), Cy3 and Cy5 global mean background signal intensity, and the Cy3 and Cy5 global SNR of the global mean feature signal intensity to the global mean background signal intensity) for each array in the dataset.

Division 1: SVMs predict microarray quality

Division 1 analysis utilizes the HQ- and LQ-HDSs to develop and train a SVM model that discriminates best quality classes utilizing all six classification variables present within the HDS. The SVM model accurately classified (100%) a random

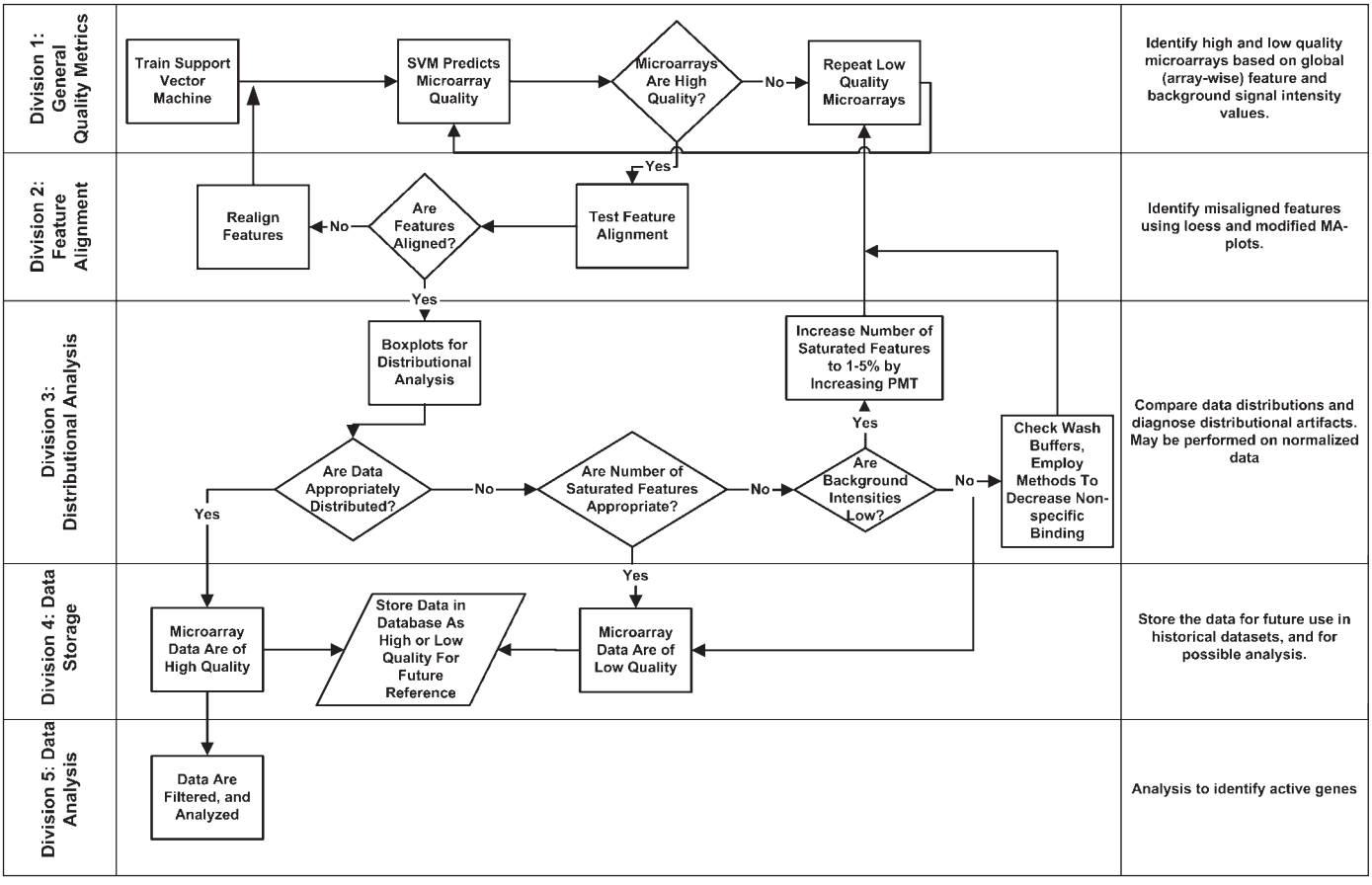

Figure 2. Microarray quality control protocol. General Quality Metrics, Feature Alignment, and Distributional Alignment divisions are depicted with two additional divisions that place the protocol into context with the overall data management infrastructure. General Quality Metrics (Division 1) analysis uses a SVM model trained on the HDS, which includes a combination of the high- and low-quality datasets (HQ-HDS and LQ-HDS, respectively). The predictor variables are filtered through a step forward logistic regression to identify the most discriminatory and predictive variables for use in training the SVM, and further analysis. Feature Alignment (Division 2) conducts a loess analysis based on treatment, dye and microarray variables using the raw intensity values from each array to determine if a subgrid has been misaligned during the quantification process. Distributional Analysis (Division 3) combines box-and-whisker plots with standard line plots to identify trends in data distributions and the number of saturated and unidentified features. Quality control output is stored within the database for further HDS refinement (Division 4) and the data is forwarded for analysis (Division 5).

Table 1. Comparison of the predictive accuracy of SVM models for microarray quality predictions

	Sensitivity	Specificity	PPV	NPV
All predictor variables[a]	0.93	0.70	0.89	0.79
EMC minimized model[b]	0.96	0.92	0.97	0.90

[a]All predictor variables includes six variables: Cy3 and Cy5 mean global feature intensities; Cy3 and Cy5 mean global background intensities, Cy3 and Cy5 SNRs (ratio of the two above listed values).
[b]The ECM minimized model variables are those from the model that showed the lowest expected cost of misclassification. The lowest ECM model included the following three variables: Cy5 mean global feature intensity, Cy3 mean global background intensity, and the Cy5 SNR.

sampling of low- ($n = 40$) and high-quality ($n = 44$) datasets from the HDS, here after referred to as the training set. Since this is a binary system the term positive is used to denote high-quality microarrays, while negative is used to denote low-quality microarrays. The positive predictive value (PPV) is the proportion of predicted high-quality arrays relative to the number of true high-quality arrays. The negative predictive value (NPV) is similar to the PPV except it is calculated with respect to low-quality arrays. The SVM model accurately predicts high-quality microarrays when using a validation set (a randomly selected subset of the HDS, not including arrays from the training set) of 59 low-quality and 174 high-quality datasets, with a PPV of 89%, but performed less effectively when predicting low-quality microarrays, with a NPV of 79% (Table 1). In other words, 89% of the true high-quality arrays were accurately predicted to be of high quality, while only 79% of the true low-quality arrays were accurately predicted to be of low quality.

Expected cost of misclassification function improves the predictive accuracy of the SVM

To identify the combination of most predictive features (i.e. the most predictive variables of the six used in the first SVM), a series of 50 models with different combinations of features were constructed, including the model with all features and models with only one feature. The most optimum model was chosen by minimizing the ECM function based on the classification of the validation set. The ECM values ranged from 10.93 to 112.3, with the best model using three features, and the worst model using only one feature, the Cy3 global mean background.

The optimum model consisted of the Cy5 global (whole array) mean feature signal intensity, Cy3 global mean background, Cy5 global SNR. This model exhibited modest improvement in the sensitivity of the SVM (96% as opposed to 93%), and a vast improvement in the specificity of the SVM (92% as opposed to 70%). Similar increases were also exhibited in the PPV (97% compared to 89%) and NPV (90% compared to 79%) as well. These results suggest that assessment using all available variables to train the SVM model contributes to noise that compromise array quality predictions made on the validation set.

An SVM is necessary for classification as the data exhibit a non-linear separating margin (Figure 3). This precludes the use of linear models, such as logistic regression models, for efficient classification of data quality.

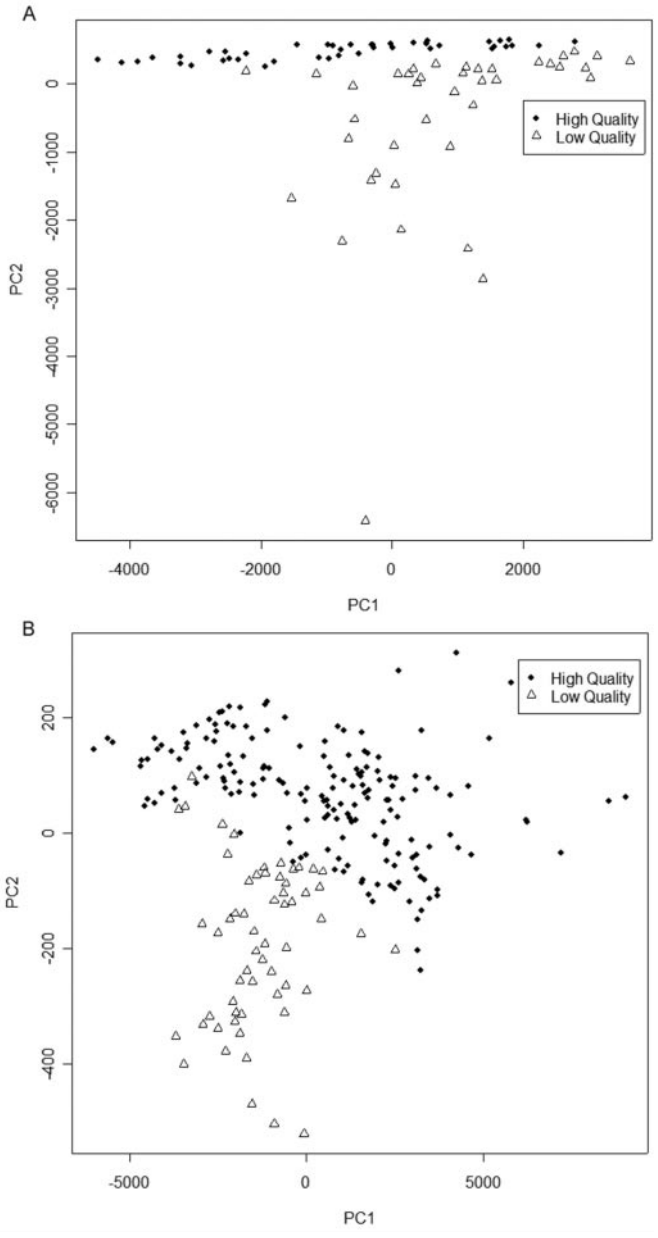

Figure 3. Principal components plots of the training and validation sets. The (**A**) training and (**B**) validation sets were subjected to PCA, and the first three principal components were plotted. It is clear that a non-linear margin provides separation between the high- and low-quality populations.

Division 2: non-parametric regression methods detect grid misalignments

A non-parametric regression procedure is utilized for detecting grid misalignments. MA-plots have been used to visualize microarray normalization implemented on the print-tip level (12,16). In addition to aiding in normalization, MA-plots assist with the identification of misaligned grids. Non-parametric regression methods, initially introduced to estimate bias, are also capable of identifying misaligned microarray quantification grids on a per array basis provided that most of the microarrays under study are correctly aligned, and that misalignment is an infrequent and aberrant event (12). Whereas

most of the microarray grid blocks (a geographical region on the microarray where all features are printed by the same print-tip) have a slight non-linear relationship, misaligned blocks will exhibit a significantly greater slope than correctly aligned blocks such that they appear as obvious outliers in the MA-plot (Figure 4A and B).

Arrays demonstrating misaligned features are identified for follow-up and realignment. The realignment of the block will result in the alteration of the global intensity values for that array and as a result are resubmitted for Division 1 analysis. During the realignment process, it may be possible to diagnose possible causes of the misalignment, such as high background,

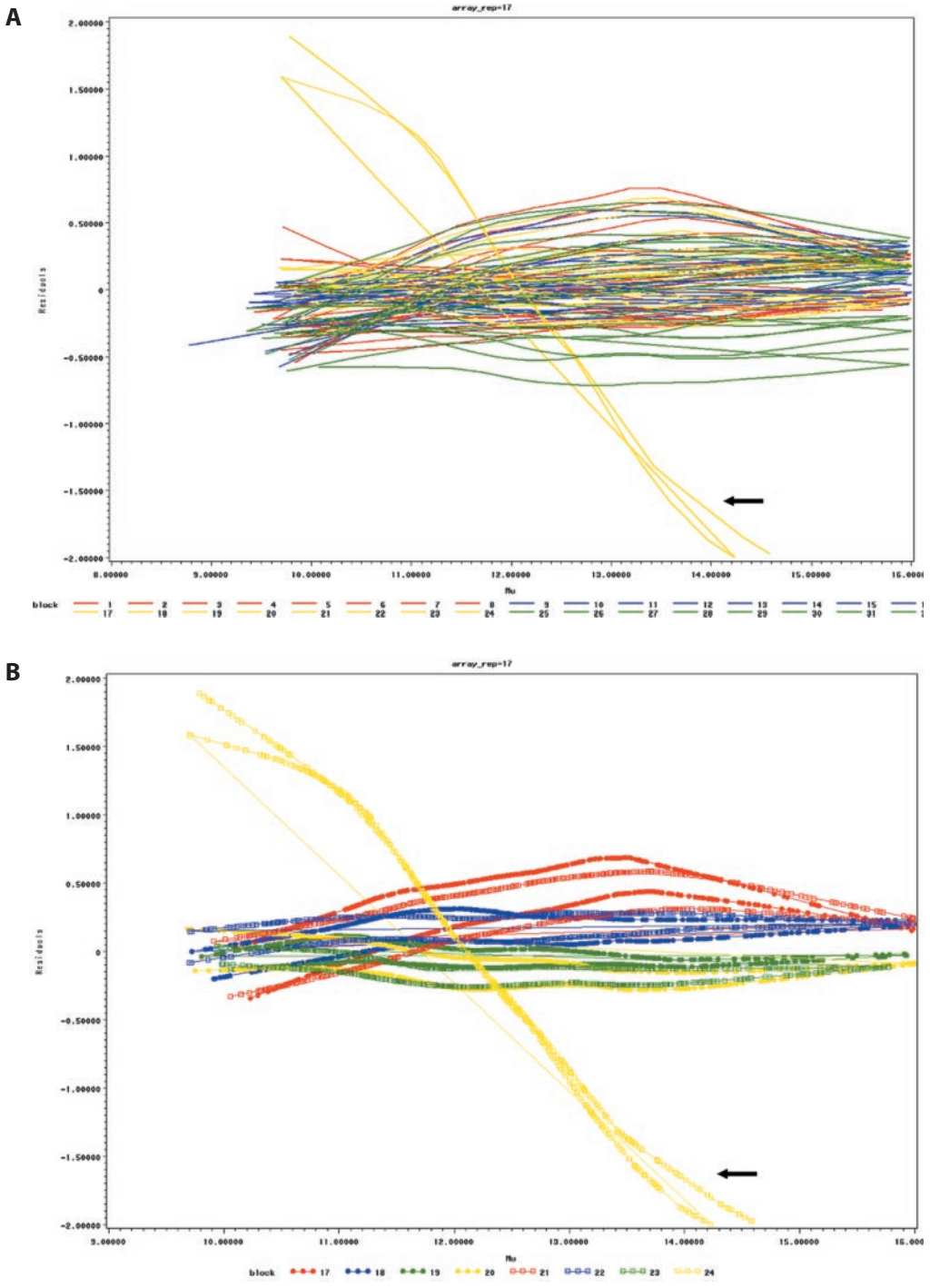

Figure 4. Loess analysis of microarray data identifies microarrays with misaligned grids. (**A**) Loess analysis of the raw intensity values from each array identified one misaligned subgrid on this microarray as evidenced by the lines with large and sharp slopes (arrow). Each subgrid is represented by two lines, one for each dye. (**B**) Subgrids 17–24 were identified as possibly problematic in A, and plotted in B for better resolution, identifying subgrid #24 as the putatively misaligned subgrid. The investigator verified the misalignment using the quantification software and corrected it before further analysis.

dust contamination, or robotic printing error, facilitating corrective action to minimize future occurrences thus improving assay performance and consistency.

Division 3: identifying compressed and similar data distributions in microarray data

Division 3 identifies microarrays with compressed or non-uniform dynamic range. Box-and-whisker plots were used to analyze feature intensity distributions on a per-microarray basis (Figure 5). Based on empirical observations, optimal distributions have the following characteristics: (i) a 25th percentile of ~700–2000 U, (ii) a 75th percentile of ~7000–10 000 U (i.e. interquartile range spanning intensities of 5000–9300 U), (iii) a median of ~3000–6000 U and 4 a mean within the interquartile range defined by the boxed region in Figure 5. The distribution of mean Cy5 median feature intensity values for the HQ-HDS is shown in Figure 5 (array_rep = 0). Based on these criteria, microarrays 19–24 fail to show appropriate distributions because the 75th percentile is lower than the recommended range of 7000–10 000 U (array_reps: 19–24). Microarrays 13–18 approach appropriate distributions, since the 75th percentile of the feature intensity distribution is closer to the recommended 75th percentile (i.e. 7000–10 000 U) which is more consistent with the empirically defined recommendations based on the HQ-HDS (Figure 5A). Boxplots generated on the $\log_2$ transformed data (Figure 5B) further illustrate the data compression in arrays 13–18, and demonstrate that the distributions for arrays 19–24 also exhibit a downward shift in the IQR.

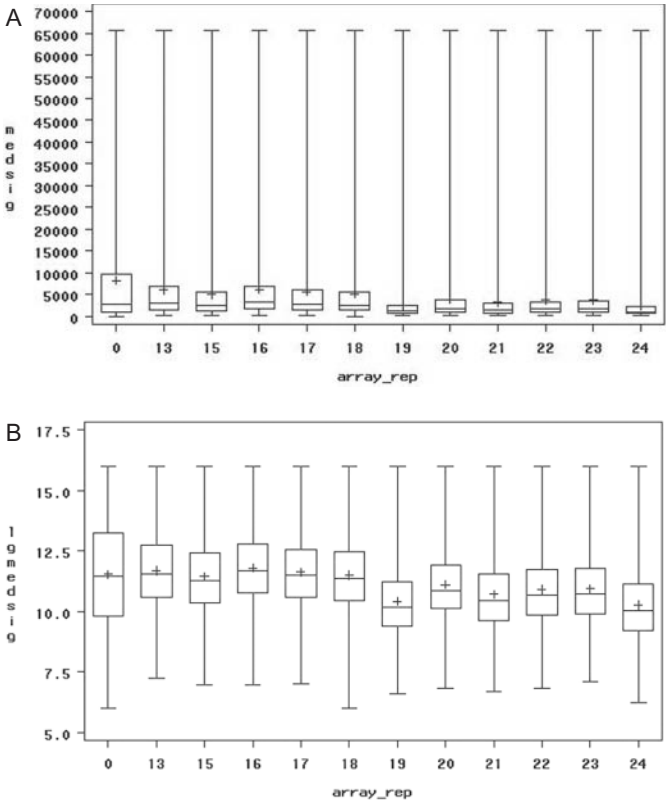

Figure 5. Illustration of the box-and-whisker plot to examine the distribution of feature intensities. Boxes represent the interquartile range, with the 75th percentile at the top and the 25th percentile at the bottom. The line in the middle of the box represents the 50th percentile, or median, while the plus represents the mean. The pluses for arrays 20–24 lie on the 75th percentile line of the box. Whiskers represent the rest of the distribution, with their terminations representing the lowest and highest feature intensity values. The *x*-axis represents the individual microarray, while the *y*-axis represents the feature intensity values. The boxplot of the HQ-HDS population of median Cy5 signals (array_code = 0), illustrating a broad range of values, from eight randomly selected HQ-HDS arrays. (**A**) Ideally the 75th percentile would be in the range of 7000–13 000 U, with an interquartile range of ~5000–9500 U. The arrays under study (array_code > 0) exhibit some compression (Cy5 channel shown here), as indicated by compressed interquartile ranges (i.e. boxes), with microarrays 19–24 exhibiting the greatest compression issues. (**B**) Visualization of the same data using boxplots of the $\log_2$-transformed data. Here, the high quality distribution appears centered ~11.51 U, with an interquartile range spanning 9.86–13.27. It is evident that the remainder of these distributions are compressed compared to the HQ-HDS, and the distributions for 19–24 are shifted downwards, a feature not evident when using normal space boxplots.

As the interquartile range and number of saturated features are positively correlated (Figure 6), the number of saturated features serves as a useful surrogate marker to ensure comparable data distributions are achieved during array scanning. Figure 7 shows the number of saturated features per array for the microarrays shown in Figure 5 (array_reps > 0). Typically this plot includes the upper and lower control limits (empirically defined to be 2 and 1%, respectively). These control limits set the boundaries for acceptable data (i.e. data must lie between the control limits to be acceptable). However, on this plot all of the microarrays (15–24) are well below the LCL (in the range of 0.1–0.5% of the features). Consequently, microarrays 19–24 have severely compressed dynamic range, as reflected by the low number of saturated features.

Implementation

The protocol is an initial step to provide a non-biased data quality assessment tool that facilitates the sharing of high-quality data, albeit on a per lab basis. It is meant to be implemented locally, with a focus on intralaboratory or collaborative project quality assessments as opposed to broad quality assessments of datasets within public repositories. It is assumed that at the very least, some form of feature quality control, such as that found in image quantification software [e.g. GenePix (Axon Instruments), AnalyzerDG (Molecularware)] or which can be implemented separately (6,7), before implementation of these methods. A more detailed listing of assumptions is provided in Table 2. The primary goal is to ensure arrays are of comparable quality, and to minimize unnecessary technical variation that may skew future results. As such, these techniques are platform independent, but do not support cross-platform quality comparisons within a study or across a public repository.

To implement the full protocol, an internally established HDS of high- and low-quality microarrays must be available in order to assess quality metrics of interest for Division 1 analysis. The predictive variables presented in this study are specific to our HDS; implementations of the general method by other groups may identify additional variables, although overlaps between laboratories are likely. Use of the ECM function facilitates the identification of SVM models that are more predictive than others. Investigators may wish to explore other methods of feature selection, or to utilize the first several principal components from a principal components analysis (PCA), to create SVM models that

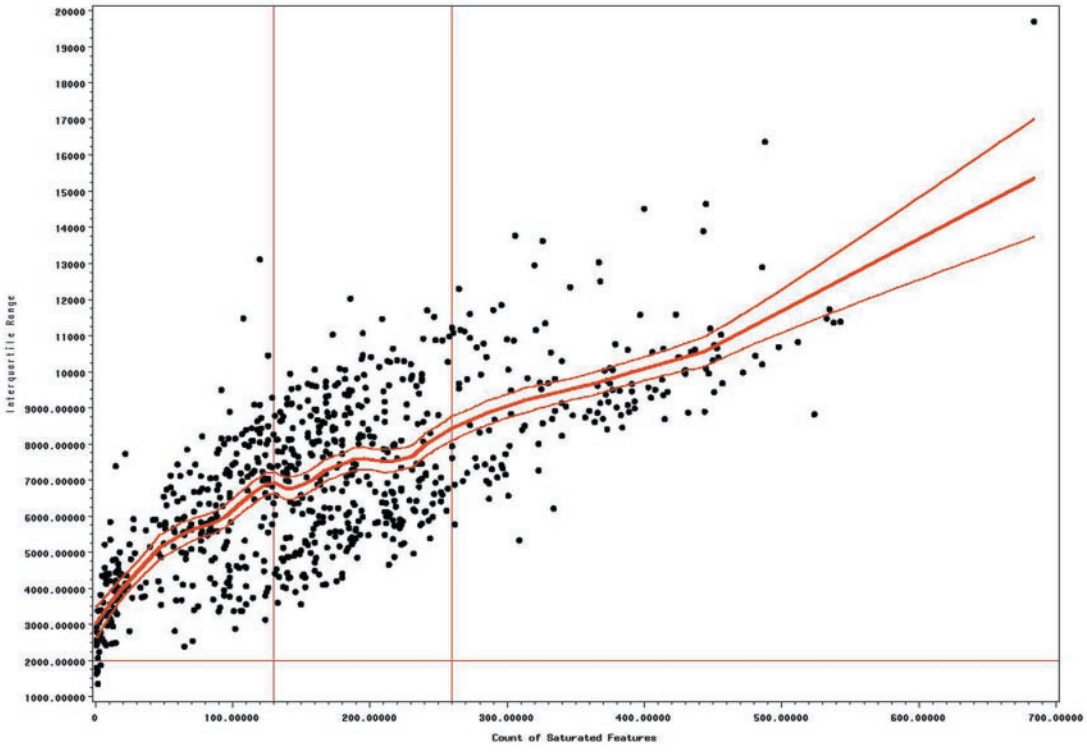

Figure 6. Interquartile range increases as a function of the number of saturated spots. The interquartile range is a measure of data spread, calculated as the difference between the 75th and 25th percentiles. The interquartile range increases with increasing number of saturated features, suggesting lower numbers of saturated features contribute to compressed ranges. By increasing the number of saturated spots compression is minimized. The lines on the plot represent the loess best fit line and the 95% confidence intervals.

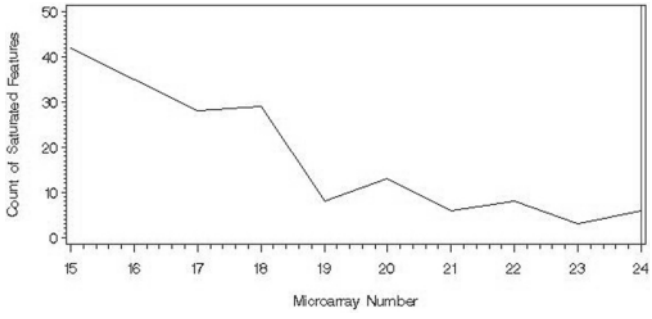

Figure 7. Saturated features correlate with compressed distributions. The microarrays depicted are the same shown in the box-and-whisker plot in Figure 5. The largest degree of distributional compression in Figure 5 corresponds to microarrays 19–24, the ones with the lowest number of saturated features.

minimize the ECM. Ultimately, the investigator must decide which variables are most predictive when used in the SVM. Investigators may also be required to use an alternative kernel in the SVM procedure to ensure optimal discrimination.

Division 2 and 3 analyses may be implemented without the use of the HDS, may be implemented independent of Division 1, and each other, and may be implemented in any order. Division 2 and 3 analyses may use any statistical software that supports LOESS and boxplot creation, such as R or SAS. Although the algorithm for Division 2 is commonly used for normalization, our use in this manner is not for its normalization properties, but rather its visualization properties. Thus, use of Division 2 does not preclude the use of other normalization techniques.

Table 2. Applied assumptions for intralaboratory quality control and assurance protocol[a]

1. Test and training datasets were obtained using the same, pre-agreed SOP
2. Test and training datasets used the same microarray platform
3. Microarray scanning is performed using the same equipment
4. Image analysis (including segmentation and background calculation methods) used the same approach for test and training datasets
5. Same normalization methods were used for test and training datasets (Division 3 analyses)

[a]The datasets available for this manuscript were insufficient to test the necessity of each assumption.

DISCUSSION

Quality control measures are performed to ensure that extreme or unusual variation and other technical issues do not overshadow biological and treatment variance. Although the goal of normalization is to minimize technical variation across samples, most normalization techniques will be more successful if less technical variation is present before normalization. Therefore, quality control techniques are used to identify technical variation arising from assignable causes due to the process. If the variability exceeds a chosen threshold, low-quality datasets can be identified and eliminated or corrected before further analysis while addressing sources of undesirable

variation in future studies, thus improving assay performance and consistency. Normalization on the other hand corrects for variability that arises from assignable causes.

By controlling the quality of the data, assurances can be made that the results from these studies are due more to biological variation, and less to technical variation. Furthermore, by decreasing the technical variation, more accurate estimates of gene expression may be made, while making more power available for significance testing. This has direct impacts on knowledge that is exchanged through data sharing via scientific publications and public data repositories.

A streamlined and standardized process of microarray quality control has been developed that encompasses several complementary techniques. The protocol combines a trained SVM model and non-parametric regression model with more classical techniques such as box-and-whisker and line plots. Although, it is possible to approach the line plot using a Shewhart plot, where control limits are defined based on the variance (NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, accessed on April 5, 2004), for our purposes empirically defined control limits are preferred. Several different variables, including the feature signal and background intensity levels, SNRs, grid alignment, data distribution and dynamic range, and the number of saturated and undetected features are used to assess data quality on a per array basis, thus providing a streamlined, high-throughput analysis method to identify quality assurance issues that require intervention.

Quality assignments by the SVM improved only when the most predictive variables, as determined by minimizing the ECM function, were used (Table 1). By using the most predictive variables the SVM improves in sensitivity (i.e. true positive rate), specificity (i.e. true negative rate), PPV and NPV, suggesting that the SVM model that minimized the ECM function does not contain as much noise compared to the model that uses all of the predictive variables, and that the ECM can be used as an objective function for evaluating SVM performance. Also, collinearity between the Cy5 SNR and the Cy5 background was not exhibited in the most predictive model (Figure 8). With respect to the protocol, microarrays that are of high quality progress to Division 2 analysis while the samples from the low-quality microarrays are flagged to repeat the hybridization.

Predictive variables may differ among labs, and are expected to be technology/platform and protocol dependent. In this study it is not surprising that Cy5 mean global feature intensity and Cy5 SNR are both included in the model as it is reported to be more susceptible to environmental factors, such as ambient ozone levels, than Cy3 (17).

Division 2 analyses focuses on grid alignment using MA-plots, and plotting the data on a per-block or subgrid basis to identify block misalignments. This streamlines the process of realignment which can be reassessed in Divisions 1 and 2, and minimizes the need to conduct costly, time consuming, and potentially unnecessary repeat hybridizations.

Division 3 analyses are concerned with data distributions, and ensuring proper dynamic range. Appropriately and similarly distributed data are considered to be of high-technical quality and are forwarded for further analysis. Data distributions are assessed using box-and-whisker plots, where the highest intensity value should be at saturation

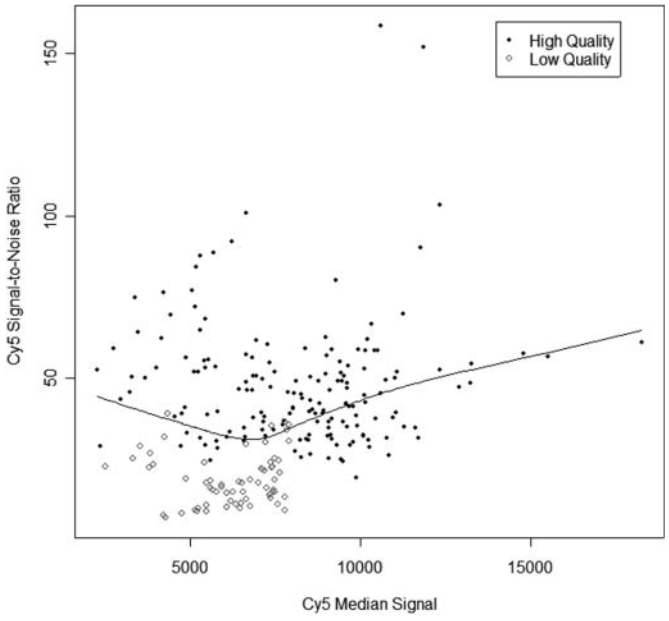

Figure 8. Cy5 mean global feature intensity and Cy5 SNR are not correlated. The scatterplot of mean global feature intensity and the SNR illustrates that no relationship exists, *per se*, between the two variables. The line in the plot represents the loess best fit line through the data. The dip in the line is a result of the concentration of low-quality arrays between 5000 and 7500 U on the *x*-axis.

(65 535 U). Data exhibiting appropriate distributions have yielded comparable results to those verified by QRT–PCR (13). Most problems with compressed interquartile range and distributions are linked to inappropriate photomultiplier tube (PMT) gain settings. The PMT gain should be set to obtain a comparable number of saturated features (our experience is that 1–2% is appropriate) in order to achieve similarly shaped data distributions across all arrays (i.e. 75th percentile of ~7000–10 000 U and 25th percentile of ~700–2000 U, with a mean within the interquartile range).

The most reliable indicator of obtaining appropriate dynamic ranges during the scanning process is the number of saturated features, and not the PMT value. We advocate shifting the PMT value in order to obtain a proper data distribution, and sacrificing the overall background intensity. Ideally, the background signal intensity will be low enough so that shifts in PMT will not adversely affect the number of identifiable features. Thus, it is not advisable to standardize the PMT gain value for an entire microarray experiment, as it is expected that optimal PMT gain values will vary by microarray. Changing the gain values following a microarray scan repeatedly may result in photobleaching of the dyes, especially when smaller amounts of labeled samples (e.g. <15 μg of starting material) are hybridized to the array. Following scanning, diagnostic plots can be used to determine if the number of saturated features meet the criteria (1 and 2% as the LCL and UCL, respectively, are typically used). Abbreviated and compressed data distributions can manifest problems in downstream analysis and normalization, and may compromise subsequent statistical analysis of gene expression changes.

For example, arrays 19–24 exhibit the greatest degree of data compression in addition to shifted IQRs (Figure 5) and

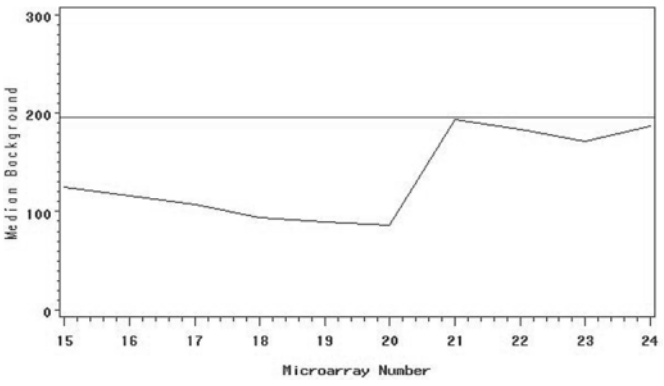

Figure 9. Background signal should be sacrificed for more saturated features. The microarrays depicted are the same shown in the box-and-whisker plot in Figure 5. The arrays with the largest Cy3 background are arrays 21–24. The reference line represents the mean Cy3 background for the HQ-HDS. In this case, the investigator was more concerned with obtaining a low Cy3 background than an optimal number of saturated features. Cy3 background should be sacrificed to increase the number of saturated features as the mean background for those arrays is below the mean for the HQ-HDS.

highlight the correlation between the number of saturated features and the compressed distribution (Figure 7). The low background levels for these microarrays (Figure 9) is a likely contributing factor since the PMT gain was purposefully set low to minimize background intensity, resulting in the constricted interquartile range. Instead, PMT levels should have been increased to achieve 1–2% feature saturation to increase the probability of obtaining an appropriate and uniform distribution (dynamic range) across all microarrays within the study.

Some may question the necessity of a large-scale QAP, especially with regard to experiments employing the reference design, where a common reference sample is present on all of the microarrays which could serve as indicator of quality. However, in a well-controlled experiment, the reference sample will be independent of the treatment samples, especially when using dye-swaps and technical replicates. Thus, with the samples being independent, they will also be uncorrelated; meaning signal intensities from the reference should not reflect signal intensities from the treatments. Consequently, variance in the two sample classes will also be independent, and the quality of the reference signal will have no relationship to the quality of the treatment signal. This also would preclude the use of the reference sample based method based from those designs that do not incorporate a reference sample, such as the loop family of designs.

Following these quality control methods, only high-quality data should proceed to normalization and higher-order analyses. However, all microarray data should be stored in an appropriate database, including low-quality microarray data, for future refinement of the HDSs. This ensures the quality of work being generated within a laboratory to be of their highest quality. However, it does not facilitate comparisons to the general body of publicly available data. By ensuring data being produced at the laboratory level is of the best local quality, investigators ensure the reproducibility of their results. However, the burden of quality assessment by the public user and peer reviewers still remains a challenge that is beyond the scope of these methods.

CONCLUSIONS

This protocol serves as an initial step to assess intralaboratory or collaborative group data quality for studies conducted using the same spotted microarray platform. Quality control ensures data integrity and is essential to facilitate subsequent analysis and meaningful interpretation that support conclusions, future hypotheses and knowledge-based decision making. It provides complementary QA/QC methods that include automated, high-throughput quality assessment using SVMs. Combining this protocol with other methods such as biological replicate clustering (5), and spot quality control assessments provides a more complete quality control protocol that ensures the integrity of cDNA and oligonucleotide microarray data. The adoption of such measures is necessary to instill confidence in data uploaded to public repositories, an emerging requirement for a growing number of prestigious journals. However, the development of an enterprise solution that assesses data quality across platforms and between independent groups available within public repositories is needed in order to realize comprehensive knowledge extraction from publicly available complex datasets.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Gary Jahns for providing constructive comments on this work. This work has been supported by NIH grants ES11271, ES12245 and Superfund P42 ES04911. Support for L.D.B. was provided by T32 ES07255. Support for J.E.E. was provided by R25 CA92049 and T32 ES007334. T.R.Z. is partially supported by the Michigan Agriculture Experiment Station. Funding to pay the Open Access publication charges for this article was provided by The National Institute of Environmental Health Sciences.

Conflict of interest statement. None declared.

REFERENCES

1. Hessner,M.J., Meyer,L., Tackes,J., Muheisen,S. and Wang,X. (2004) Immobilized probe and glass surface chemistry as variables in microarray fabrication. *BMC Genomics*, **5**, 53.
2. Jarvinen,A.K., Hautaniemi,S., Edgren,H., Auvinen,P., Saarela,J., Kallioniemi,O.P. and Monni,O. (2004) Are data from different gene expression microarray platforms comparable? *Genomics*, **83**, 1164–1168.
3. Ulrich,R.G., Rockett,J.C., Gibson,G.G. and Pettit,S.D. (2004) Overview of an interlaboratory collaboration on evaluating the effects of model hepatotoxicants on hepatic gene expression. *Environ. Health Perspect.*, **112**, 423–427.
4. Miles,M.F. (2001) Microarrays: lost in a storm of data? *Nature Rev. Neurosci.*, **2**, 441–443.
5. Grant,G.R., Manduchi,E., Pizarro,A. and Stoeckert,C.J.,Jr (2003) Maintaining data integrity in microarray data management. *Biotechnol. Bioeng.*, **84**, 795–800.
6. Wang,X., Ghosh,S. and Guo,S.W. (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res.*, **29**, E75–5.
7. Hautaniemi,S., Edgren,H., Vesanen,P., Wolf,M., Jarvinen,A.K., Yli-Harja,O., Astola,J., Kallioniemi,O. and Monni,O. (2003) A novel

strategy for microarray quality control using Bayesian networks. *Bioinformatics*, **19**, 2031–2038.
8. Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. *et al*. (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res*., **31**, 94–96.
9. Petri,A., Fleckner,J. and Matthiessen,M.W. (2004) Array-A-Lizer: a serial DNA microarray quality analyzer. *BMC Bioinformatics*, **5**, 12.
10. Model,F., Konig,T., Piepenbrock,C. and Adorjan,P. (2002) Statistical process control for large scale microarray experiments. *Bioinformatics*, **18**, S155–S163.
11. 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.
12. Eckel,J.E., Gennings,C., Therneau,T.M., Burgoon,L.D., Boverhof,D.R. and Zacharewski,T.R. (2005) Normalization of two-channel microarray experiments: a semiparametric approach. *Bioinformatics*, **21**, 1078–1083.
13. Boverhof,D.R., Fertuck,K.C., Burgoon,L.D., Eckel,J.E., Gennings,C. and Zacharewski,T.R. (2004) Temporal- and dose-dependent hepatic gene expression changes in immature ovariectomized mice following exposure to ethynyl estradiol. *Carcinogenesis*, **25**, 1277–1291.
14. Sun,Y.V., Boverhof,D.R., Burgoon,L.D., Fielden,M.R. and Zacharewski,T.R. (2004) Comparative analysis of dioxin response elements in human, mouse and rat genomic sequences. *Nucleic Acids Res*., **32**, 4512–4523.
15. Boverhof,D.R., Tam,E., Harney,A.S., Crawford,R.B., Kaminski,N.E. and Zacharewski,T.R. (2004) 2,3,7,8-Tetrachlorodibenzo-p-dioxin induces suppressor of cytokine signaling 2 in murine B cells. *Mol. Pharmacol*., **66**, 1662–1670.
16. Dudoit,S., Yang,H.Y., Callow,M.J. and Speed,T. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–139.
17. Fare,T.L., Coffey,E.M., Dai,H., He,Y.D., Kessler,D.A., Kilian,K.A., Koch,J.E., LeProust,E., Marton,M.J., Meyer,M.R. *et al*. (2003) Effects of atmospheric ozone on microarray data quality. *Anal. Chem*., **75**, 4672–4675.