nature biotechnology

Evaluation of external RNA controls for the assessment of microarray performance

Weida Tong[1], Anne Bergstrom Lucas [2], Richard Shippy[3], Xiaohui Fan[1,4], Hong Fang[5], Huixiao Hong[5], Michael S Orr[6], Tzu-Ming Chu[7], Xu Guo[8], Patrick J Collins[2], Yongming Andrew Sun[9], Sue-Jane Wang[6], Wenjun Bao[7], Russell D Wolfinger[7], Svetlana Shchegrova[2], Lei Guo[1], Janet A Warrington[8] & Leming Shi[1]

External RNA controls (ERCs), although important for microarray assay performance assessment, have yet to be fully implemented in the research community. As part of the MicroArray Quality Control (MAQC) study, two types of ERCs were implemented and evaluated; one was added to the total RNA in the samples before amplification and labeling; the other was added to the copyRNAs (cRNAs) before hybridization. ERC concentration-response curves were used across multiple commercial microarray platforms to identify problematic assays and potential sources of variation in the analytical process. In addition, the behavior of different ERC types was investigated, resulting in several important observations, such as the sample-dependent attributes of performance and the potential of using these control RNAs in a combinatorial fashion. This multiplatform investigation of the behavior and utility of ERCs provides a basis for articulating specific recommendations for their future use in evaluating assay performance across multiple platforms.

ERCs are synthetic or naturally occurring RNA species that are added to an RNA sample for the purpose of quality control of the assay. Most commercial microarray platforms contain probes specifically designed for interrogating ERC transcripts. These probes have been extensively prototyped and optimized for performance on each microarray platform. To provide an enhanced assessment of the analytical performance of the system during data collection, a variety of ERCs can be added to the sample in a range of concentrations spanning high to low abundance by evaluating assay performance across the expected range of concentrations in the sample[1]. A well-constructed concentration-response series of ERCs is useful in many ways for assessing assay performance. Depending on the point in the assay the ERCs are added, they can be used to identify potentially failed steps during the assay process. Realizing the potential importance of ERCs for analytical performance assessment, the External RNA Control Consortium (ERCC) was established in 2003 with the objective of developing a set of ERC transcripts that could be used with various gene expression profiling technologies, including microarray platforms[2].

ERCs can also be useful for evaluating different data analysis methods[3]. The cRNA data set from Affymetrix, known as the Latin square data set (http://www.affymetrix.com/support/technical/sample_data/datasets.affx), consists of data from 42 cRNAs, which were prelabeled and added to a hybridization solution at various known concentrations. A similar data set is also provided by GeneLogic (http://www.genelogic.com/newsroom/studies/index.cfm). Both data sets are freely available and have been widely used in the research community for comparative performance analysis of GeneChip-specific normalization and gene selection methods[4–7]. Recently, Choe *et al.*[8,9] demonstrated the value of using a large number of cRNA transcripts at concentration ratios varying from one- to fourfold to compare the performance of different data analysis scenarios.

The MAQC study[10] provides a rich data resource to investigate various issues associated with DNA microarray platforms, including the performance of ERCs across various platforms. In this project, the probes for the ERC transcripts (**Supplementary Methods** online) are unique nonmammalian sequences selected to minimize cross-hybridization with transcripts from mammalian species such as human, mouse and rat. Seven microarray platforms were evaluated and ERCs were used in the following platforms: Applied Biosystems Genome Survey Microarray, Affymetrix GeneChip, both Agilent's One-Color and Two-Color platforms, GE Healthcare CodeLink and Eppendorf (data not shown). With these data sets, the following questions were asked: (i) Do the ERCs behave in the expected manner? (ii) Can outlying assays be identified using ERCs? (iii) Can ERCs assess the accuracy of ratios between different samples? (iv) Can ERCs provide information other than assay quality? (v) How does the choice of normalization and data processing methods affect the ERCs data?

RESULTS

The utility and performance behavior of ERCs were investigated using two independent sets of data; the MAQC data set[10] and rat toxicogenomics

[1]National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Rd., Jefferson, Arkansas 72079, USA. [2]Agilent Technologies, Inc., 5301 Stevens Creek Blvd., Santa Clara, California 95051, USA. [3]GE Healthcare, 7700 S. River Pkwy., Suite #2603, Tempe, Arizona 85284, USA.[4]Pharmaceutical Informatics Institute, Zhejiang University, Hangzhou 310027, China. [5]Z-Tech Corporation, National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Rd., Jefferson, Arkansas 72079, USA. [6]Center for Drug Evaluation and Research, US Food and Drug Administration, 10903 New Hampshire Ave., Silver Spring, Maryland 20993, USA. [7]SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513, USA. [8]Affymetrix, Inc., 3420 Central Expressway, Santa Clara, California 95051, USA. [9]Applied Biosystems, 850 Lincoln Centre Dr., Foster City, California 94404, USA. Correspondence should be addressed to W.T. (weida.tong@fda.hhs.gov).

Published online 8 September 2006; doi:10.1038/nbt1237

(TGx) data set[11]. Because the results in this paper are derived from two independent experiments the following nomenclature is used to provide clarity.

The subset of the MAQC data set used for the present analysis corresponds to four genome-wide commercial microarray platforms, Affymetrix GeneChip (AFX), Applied Biosystems Genome Survey Microarray (ABI) and Agilent One-Color (AG1) and Agilent Two-Color (AGL) microarrays. Data were generated for each of these platforms by three different test sites with five technical replicates for each of the four RNA samples (A, B, C and D[10,12]). Each data set is denoted by platform_site_replicate; for example, AG1_2_A1 denotes Agilent One-Color platform, test site 2, sample A and replicate 1.

The rat TGx data set that is denoted by platform_RAT contains data from Affymetrix (AFX_Rat), Agilent One-Color microarray (AG1_Rat), Applied Biosystems (ABI_Rat) and GE Healthcare (GEH_Rat). This experiment was performed at one test site with six biological replicates for each of six different treatments. The nomenclature for the site, mentioned above, is therefore not applicable, yet it's necessary to make a distinction between samples and that is provided in Methods and within the figures.

Two types of ERCs were investigated. One type is added to the total RNA (called tERC hereafter) before initiating the cDNA synthesis and *in vitro* transcription steps of the RNA labeling procedure. When added in this manner, the tERC generally assesses the efficiency of the target preparation as well as the performance of the hybridization and scanner. The other type of ERC is added to the cRNA (called cERC hereafter) immediately before hybridization, which allows assessment of the assay performance from the hybridization onward. Applied Biosystems and Affymetrix platforms used both types of ERCs in their respective protocols, whereas Agilent used tERC and GE Healthcare used cERC only (**Fig. 1**).

The concentration-response behavior of both tERCs and cERCs was evaluated using a linear regression analysis in an effort to identify microarray assays that show outlier behavior. This is a favorable approach as the analysis is self-contained within each microarray, and therefore, does not require replicates to assess outliers. The behavior of both ERC types was investigated further to determine if additional ERC-specific analysis methods could be useful for analytical performance assessment.

External RNA control concentration-response curves

The ERC transcripts span a range of concentrations in the Affymetrix, Agilent and GE Healthcare microarray platforms, making them suitable for concentration-response analyses. The Agilent One-Color platform has ten tERCs that span six logs of concentration and interrogate the lower and upper limits of assay signal detection (**Supplementary Table 1** online). The Affymetrix platform has four tERCs that span one and a half logs of concentration and the GE Healthcare platform has six ERCs that span three logs of concentration. For the Applied Biosystems microarray platform, ERC controls are spiked at a single fixed concentration, rendering them unsuitable for a concentration-response analysis.

Figure 2 depicts the concentration-response curves for AG1, AFX and GEH_RAT. In general, all platforms exhibited accurate concentration-response patterns. In addition, performance differences are observed for tERCs relative to cERCs as seen in the data from AFX where the tERCs show decreased linear correlations compared to the cERC plots (**Fig. 2,** comparing the second and third rows of graphs for the AFX platform). This result is somewhat expected as the tERCs are introduced earlier in the assay process and are subject to multiple sources of variation introduced during sample amplification and labeling, more closely approximating the analytic manipulation. In contrast, the cERCs are added just before hybridization, and their more stable performance reflects fewer sample manipulations after these controls are added.

Two assays generated by AG1 site 2 (AG1_2_D2 and AG1_2_A3) have noticeably higher signals for tERCs at the lowest concentrations, indicating potential assay outliers. However, the specific problematic step of the assay for these two data sets cannot be identified because the behavior of tERC reflects the performance associated with multiple steps of the experiment. The benefit of using both tERCs and cERCs is demonstrated with the AFX platform, where the combination was used to elucidate procedural problems in the assay. In this example the AFX cERC performance is stable and consistent across all three test sites, but tERCs in site 1 have lower y-intercepts as compared to the other two sites, indicating that for site 1 the target preparation yield or labeling efficiency differed from the other sites (**Fig. 2**).

Concentration-response curves in one-color microarray assays

In addition to visually inspecting the concentration-response curves to interrogate the performance over the dynamic range of an assay, we calculated the linear regression statistics of the linear portion of the curves for outlier identification, including R^2 correlations and slopes. **Figure 3** (**Supplementary Table 2** online) plots the linear regression slope versus R^2 correlations for AG1, AFX and GEH_Rat. Three outlying assays were identified for AG1 site 2 (**Fig. 3a**); AG1_2_D1 has a normal R^2 with a low slope, whereas AG1_2_D2 has a normal slope with a low R^2 and AG1_2_A3 has both low slope and R^2.

An assay with a concentration-response slope of one indicates no compression of the signal because values of x and y are identical across the regression fit. By inspecting the slopes in **Figure 3**, different degrees of compression in gene expression data are observed between three

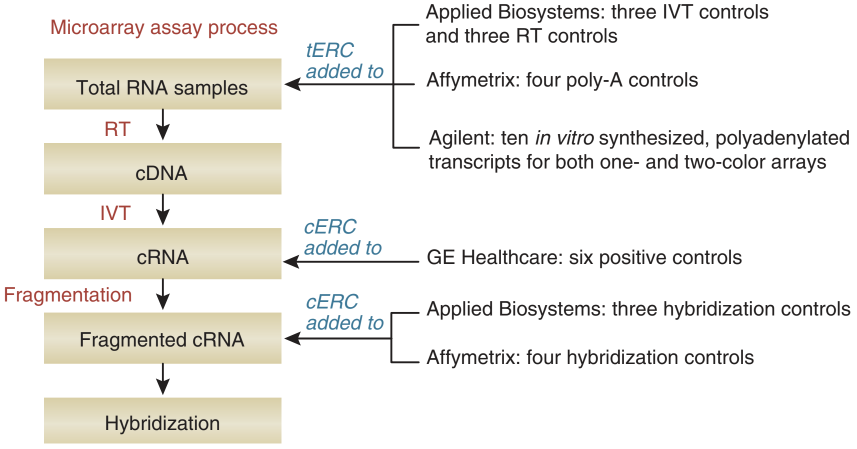

Microarray assay process
Total RNA samples
RT
cDNA
IVT
cRNA
Fragmentation
Fragmented cRNA
Hybridization
tERC added to
cERC added to
cERC added to
Applied Biosystems: three IVT controls and three RT controls
Affymetrix: four poly-A controls
Agilent: ten in vitro synthesized, polyadenylated transcripts for both one- and two-color arrays
GE Healthcare: six positive controls
Applied Biosystems: three hybridization controls
Affymetrix: four hybridization controls

Figure 1 Overview of external RNA controls (ERCs) implemented in Affymetrix, Agilent, Applied Biosystems and GE Healthcare platforms. Two types of ERCs are implemented in these four commercial microarray platforms. The first type of ERC is added to the total RNA (tERC) before initiating the cDNA synthesis and IVT (*in vitro* transcription) steps of the RNA labeling procedure. The second type of ERC is added to the cRNA (cERC) just before the cRNA is placed into the hybridization mixture. Applied Biosystems and Affymetrix platforms use both types of ERCs in their respective protocols, whereas Agilent uses the tERC and GE Healthcare uses cERC in this study. RT, reverse transcription.

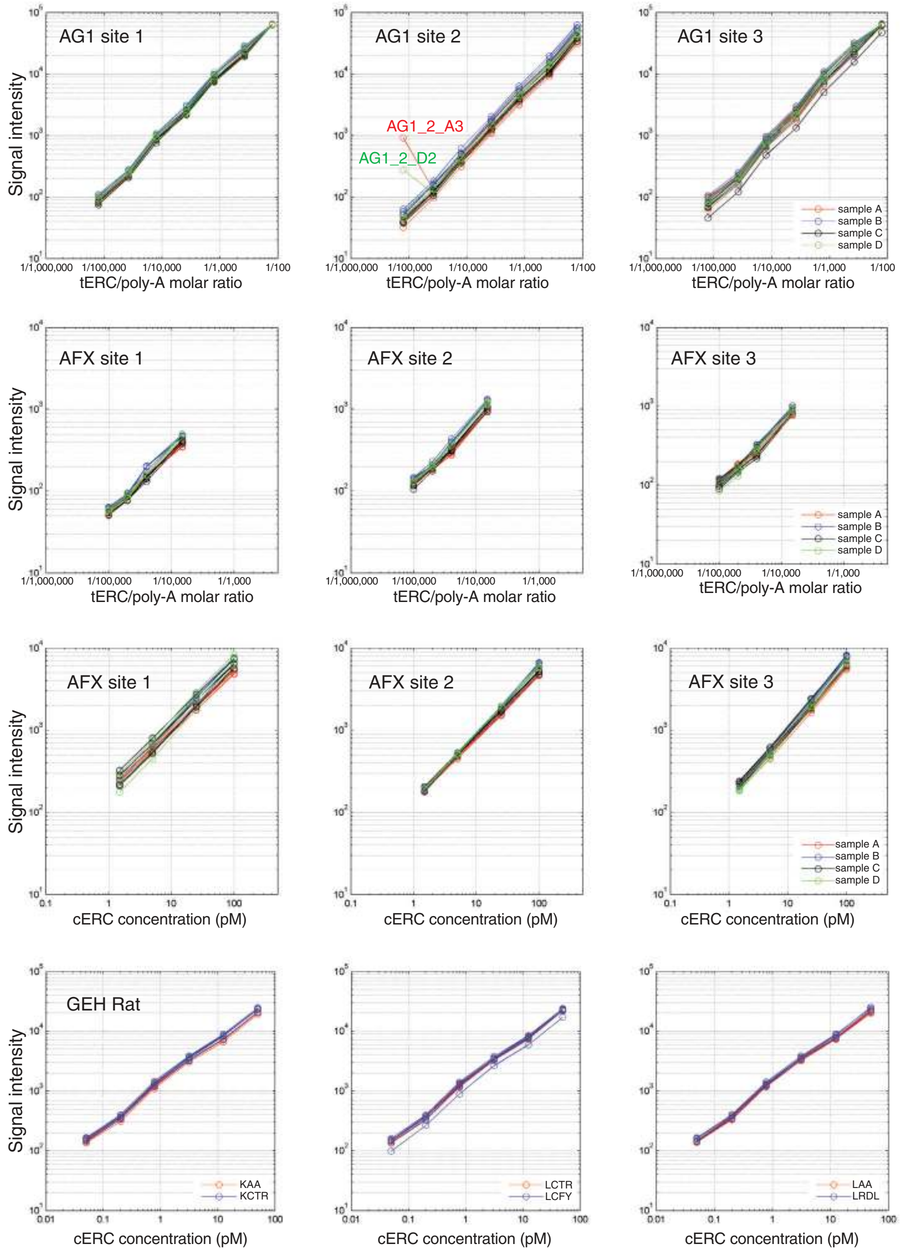


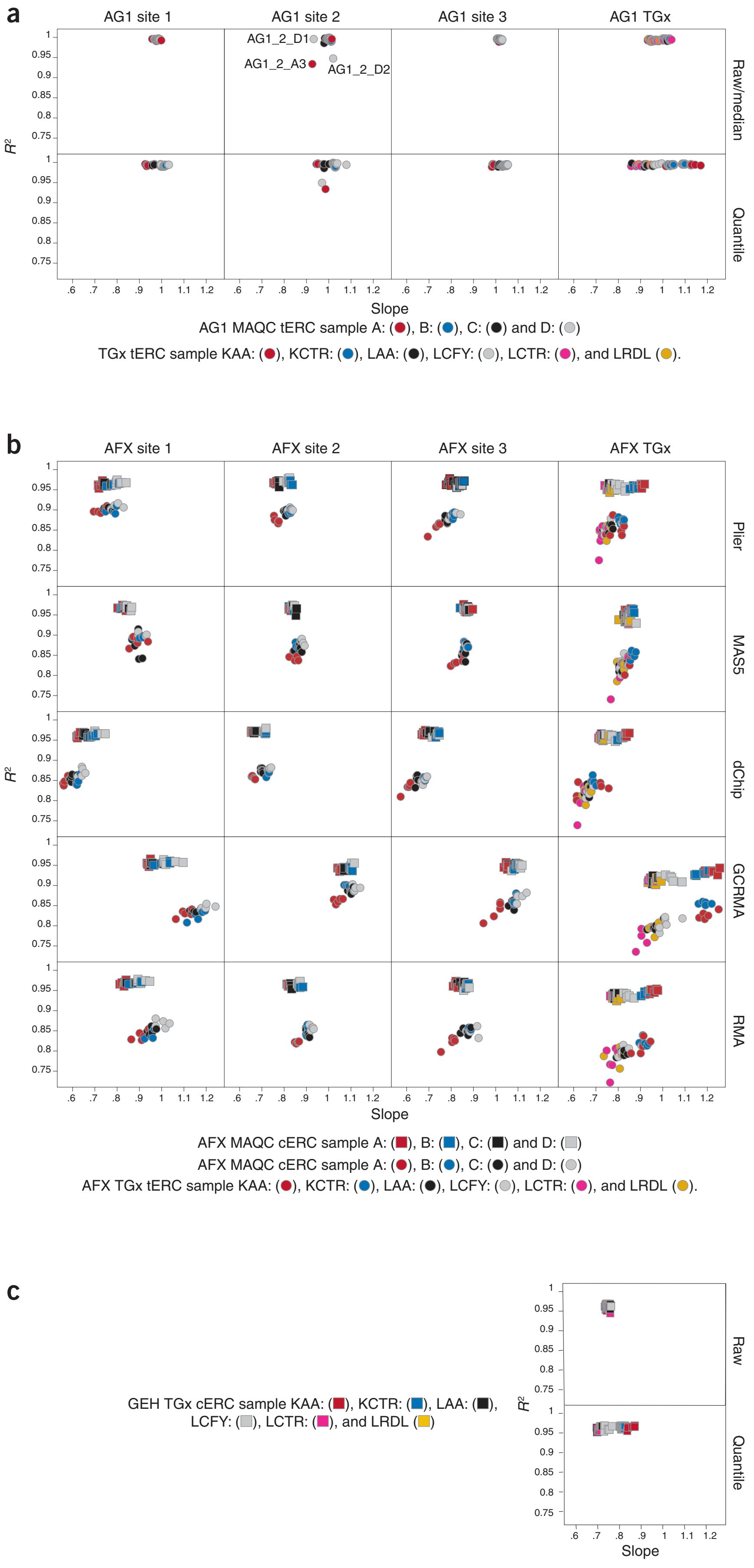
Figure 2 Concentration-response curves for ERCs on the Agilent, Affymetrix and GE Healthcare microarray platforms. Each concentration-response curve is generated from an individual microarray data set and represents the concentration of either the tERC (spiked poly-A molar ratio) or of cERC (spiked concentration in pM) on the *x*-axis as a function of normalized signal intensity on the *y*-axis. The amount of cERC added to the hybridization mixture is expressed in molar concentration based on the mass of the cERC transcript added to a specific volume of the hybridization mixture. The assumptions used to calculate the poly-A mass ratio for the different tERCs were that the average percentage of mRNA in total RNA is 2%, the average transcript length is 2,000 bases and the average molecular weight of a single base is 330 g/mol. The cERC concentration and tERC poly-A molar ratio used for this figure are summarized in **Supplementary Table 1** online. The Agilent platform is presented in the first row where seven of the ten tERCs with the highest concentrations are plotted to better compare scales to the other platforms (the full concentration-response curve is presented in **Supplementary Fig. 9** online). The Affymetrix platform is presented in the second and third rows and illustrates the combinatorial approach of using both tERCs (second row) and cERCs (third row). The GE Healthcare platform is presented in the fourth row illustrating the cERC concentration-response from the rat toxicogenomics study. This figure illustrates the different approaches each manufacturer employs for either tERC, cERC or both, when assessing assay quality using ERCs. Two microarrays from AG1 site 2 (AG1_2_D2 and AG1_2_A3) exhibit higher than expected signals for the tERCs with the lowest concentrations, indicating that these could be outlying assays. AA, aristolochic acid; RDL, riddelliine; CFY, comfrey. 'L' indicates samples isolated from livers and 'K' samples isolated from kidneys of treated rats. CTR, control (liver or kidney from untreated rats).

platforms. The AG1 platform has very little compression with a slope close to 1 for tERCs. However, ERC data for AFX and GEH_RAT experiments appear compressed to a similar extent with slopes that are detectably <1. The effect of normalization methods on the compression was also investigated. For AFX, PLIER[13], MAS5[14], RMA[6], GCRMA[15] and dChip[16] algorithms were used, whereas median scaling and quantile normalization methods were applied for both AG1 and GEH. For AFX, dChip compresses the gene expression data more than other methods whereas GCRMA has little compression or even a small degree of expansion (**Fig. 3b**)[15]. For AG1, the quantile normalization tends to separate A and C from D and B by slope in accordance with the mRNA abundance of the samples as shown in the "Performance of External RNA Controls" section (**Fig. 3a**). This sample-dependent behavior associated with the quantile normalization is also observed for AFX (PLIER, RMA and GC-RMA) (**Fig. 3b**) and GEH_RAT (**Fig. 3c**).

External RNA controls in two-color microarray assays

Agilent was the only two-color platform to use ERCs in the MAQC study. Agilent formulates two-color ERCs into two different mixtures that span 2.3 logs of concentration and are mixed with different concentrations to give the following expected ratios: 1:10, 1:3, 1:1, 3:1 and 10:1 (**Supplementary Table 3** online). This type of ERC formulation adds an additional dimension to the typical one-color concentration-response analysis, because not only should the tERCs generate signals proportional to the concentrations within each of the samples, but the two-color assays should also generate observed ratios equal to the expected ratios (or $\log_{10}$ ratios) when the data sets are dye normalized and analyzed. This accuracy assessment is contained within each probe interrogating a specific ERC transcript.

The observed versus expected ERC $\log_{10}$ ratio plots for AGL are presented in **Figure 4**. There were two outlying assays at AGL site 1 with major assay performance failures generating $\log_{10}$ ratios close to zero across the assay. AG1_1_01 had to be rescanned three weeks after the initial experiment and had faded significantly and AG1_1_85 had the same ERC control mixture added to the samples resulting in $\log_{10}$ ratios of 0 for all ten of the ERC transcripts. Two assays in AGL site 2 were also determined to be outliers. These outliers were found to have increased



within-feature noise, which might result from sample contamination from the reagents used to purify the labeled cRNA. A similar observation is obtained when comparing the linear regression correlation coefficients from the observed versus expected ratios, where the outliers are determined based on R^2 correlations for the linear fit beyond two s.d. below the mean for that site (**Supplementary Fig. 1** online).

In the MAQC study, the two-color microarray assays used only samples A and B with a dye-swap experimental design. The *y*-intercept was >0 (shifted up) for Cy5(B)/Cy3(A) in all three sites and the *y*-intercept was <0 (shifted down) for Cy5(A)/Cy3(B) at all three sites (**Fig. 4**). This shift indicates differences in mRNA abundance between sample A and sample B, which will be further analyzed in the following section.

Performance of external RNA controls

The MAQC data sets were generated from four RNA samples with an incremental increase of brain total RNA across the samples: A (0%), C (25%), D (75%) and B (100%). Because the relative mRNA abundance is not expected to be the same between Stratagene Universal Human Reference RNA (UHRR or sample A) and Ambion Human Brain Reference RNA (brain or sample B), the effect of the mRNA abundance on the ERC behavior was investigated in terms of the signal intensity with the objective of developing other ERC-specific analysis methods for assay assessment.

The tERC signal intensity increases in proportion to increasing concentrations of brain mRNA in the sample mixture, whereas the signal intensity from the biological probes exhibits reverse trends (**Supplementary Fig. 2** online). The general trend was conserved

Figure 3 Concentration-response linear regression results for the Agilent, Affymetrix and GE Healthcare microarray platforms. (**a**–**c**) The R^2 correlation coefficients (*y*-axis) versus slope (*x*-axis) from a regression analysis based on the linear portion of the concentration-response curves for AG1 (**a**); AFX (**b**) and GEH_Rat (**c**). Data used in creating this figure are in **Supplementary Table 2** online. Abbreviations are as defined in **Figure 2**. For AG1, two types of data normalization methods are presented for both MAQC and TGx data sets: raw/median scaling and quantile normalization. For AFX five types of data normalization methods are presented for both the MAQC and TGx data sets: PLIER, MAS5, dChip, RMA and GCRMA. For GEH_Rat, the raw/median data are presented for the TGx data set. This analysis indicates that (i) a degree of compression in signal is evident with the slope <1 for Affymetrix and GE Healthcare platforms, (ii) the quantile normalization method causes the data to separate by sample type and (iii) three outlying assays are identified in AG1 site 2.

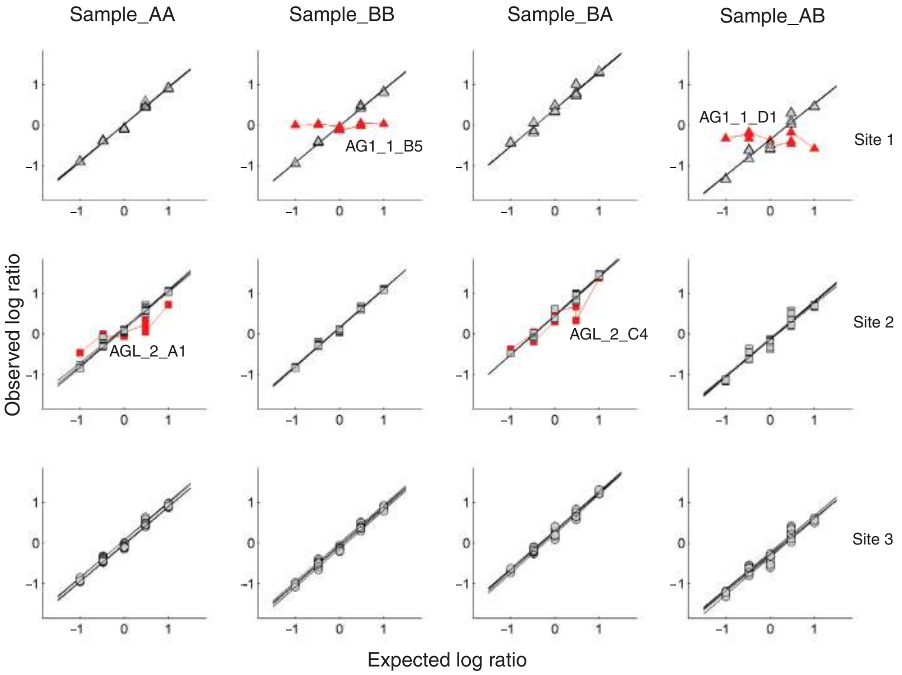


Figure 4 Expected versus observed $\log_{10}$ ratio comparison for Agilent Two-Color ERC data. The expected $\log_{10}$ ratios on the *x*-axis were based on the quantity of each tERC transcript spiked into the total RNA (**Supplementary Table 3** online). The dye-normalized signal ratios obtained from the AGL Feature Extraction software are plotted as observed $\log_{10}$ ratios on the *y*-axis. These are grouped by site name and ordered by sample combination. In the Two-Color assay, four pairs of RNA samples were generated by using only samples A and B. The samples are named AA, BB, AB and BA where the letters represent the RNA sample type with the first letter denoting the sample labeled with Cy5 and the second letter the sample labeled with Cy3. Four outlying assays are highlighted as red, two from site 1 and two from site 2.

across the different normalization methods when PLIER, MAS5, RMA and dChip were examined for AFX and the median scaling and quantile normalizations were applied to AG1 (**Supplementary Fig. 3** online). This behavior was more pronounced when the ratio of the median tERC signal intensity was divided by the corresponding median biological probe intensity and plotted against the percentage of brain in the biological target sample as depicted in **Figure 5** (**Supplementary Table 4** online), where a positive linear correlation was observed across three different one-color platforms (ABI, AFX and AG1) with slopes >0 and high correlation coefficients ($R^2 > 0.8$). Two titration points (sample C and sample D) were plotted based on the amount of brain in the sample based on the volumetric mixing of samples A and B where C = 75%A + 25%B and D = 25%A + 75%B (**Fig. 5**). This plot is accurate if the percentage of mRNA is equal between sample A and sample B. However, the Agilent two-color tERC data indicate that the percentage of mRNA was higher in sample A compared to that in sample B (**Fig. 4** and **Supplementary Fig. 4** online). If we assume that sample A has 1.5-fold more mRNA as compared to sample B[12], the percentage of brain RNA in sample C becomes 18% and for sample D becomes 67%. When these values are used in the *x*-axis of **Supplementary Fig. 5** online, the correlation coefficients improve for all of the samples at all of the sites for three different microarray platforms, further supporting the hypothesis that the samples have different percentages of mRNA.

The effect of the mRNA abundance differences between the four samples on cERC signal intensities was also investigated. Unlike tERC signal intensities, the cERC signal intensities across the four RNA samples for the ABI and AFX exhibited no significant difference (**Supplementary Fig. 6** online), indicating that the cERCs added before hybridization are unaffected by the differences in the relative abundance of the sample mRNA tested in this set of experiments. The observation is also not affected by the choice of normalization (**Supplementary Fig. 7** online). This result further supports the hypothesis that the differences between the biological samples occur at an earlier stage of target preparation.

Additional analyses using external RNA controls

For most assays identified as problematic, one or several ERCs behave differently from the others, which should be captured by an intensity-based unsupervised analysis, such as principal component analysis (PCA)[17] or hierarchical cluster analysis (HCA)[17]. PCA based on tERC signal intensity identified AG1_1_D2, AG1_1_A1 and AG1_3_B3 as outliers, consistent with the PCA plot based on the entire microarray (**Fig. 6a**). Agilent's Feature Extraction QC Report uses a different algorithm: the concentration-response curve fit to the linear portion is performed on a log-log plot after a parameterized sigmoidal curve fit of the data. The R^2 correlations and slopes from the AG1 QC report are shown in **Figure 6b**. This type of sigmoidal curve fitting ignores the differences seen in the tERCs outside the linear range and results in identification of a different set of outlying assays than in the analysis shown in **Fig. 3a**, but with the same assays as identified in the PCA analysis (**Fig. 6a**). Results similar to those in **Fig. 6a** are also observed using HCA (**Supplementary Fig. 8** online). These analyses, as well as approaches based on the concentration-response curve (**Figs. 2** and **3**) demonstrate the value of combining various ERC-specific approaches to enhance the capability of assay assessment.

DISCUSSION

A number of microarray manufacturers use ERCs to assess the technical performance of their gene expression assays. This study investigated the utility of ERCs, with emphasis on cERCs and tERCs, for assay assessment across five commercial microarray platforms using the MAQC data set[10] and a rat toxicogenomic data set[11].

This study explores several different uses of ERCs for assay assessment. First, the observed ERC signal intensities were examined against the expected concentrations to visually detect potential outlying assays, which tend to deviate from the expected concentration-response curve trend. Second, the concentration-response curves were modeled for identification of potential outlying assays using output variables from linear regression analysis. These two approaches take advantage of the unique characteristic of ERCs spiked across a wide range of differing concentrations. However, for some platforms such as Applied Biosystems, ERCs are spiked in at a constant concentration, requiring analysis methods other than the concentration-response curve analysis. Thus, PCA and HCA were conducted based on the ERC signal intensity, and the ERC-identified outlying data sets are consistent with the analysis results based on the biological whole-microarray data. These approaches are complimentary to each other and could be used in conjunction to enhance the discrimination of outlier identification.

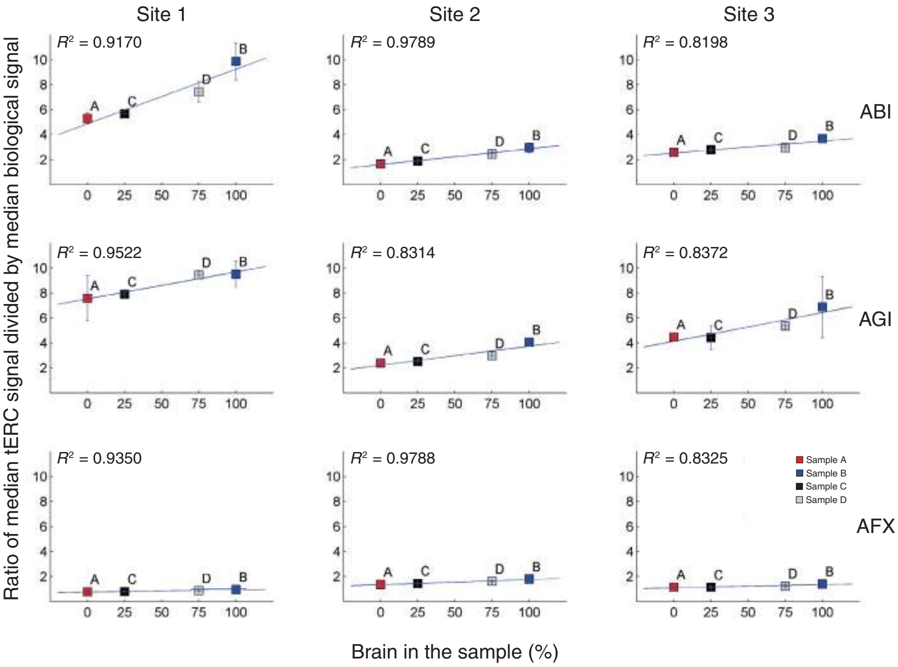


Figure 5 Illustration of the sample-dependent behavior of tERC signal across the MAQC samples. The ratio of the median tERC signal to the median biological signal is plotted against the percentages of brain RNA in the different samples (0%, 25%, 75%, and 100% for A, C, D and B, respectively). In all nine groupings (three sites for each of three platforms), the slope was greater than zero with high correlation coefficients, indicating that the tERC signal intensity is dependent on the abundance of mRNA or biological differences of the samples. Data used in creating this figure, along with the statistical assessment, are summarized in **Supplementary Table 4** online.

ERCs added at different steps in the assay offer a quality control for different steps of the assay process. cERCs are tolerant to differences in the mRNA abundance in the total RNA samples and provide the advantage of being able to assess assay performance independent of the total RNA sample complexity (**Fig. 2**). A limitation of the cERCs is the inability to detect variability that may occur during target preparation. Because tERCs are added into the assay process at a very early stage, they can reveal failures during sample collection, storage, labeling and amplification as well as hybridization, scanning and data collection. As poor target quality is a common reason for aberrant assay results, there is value in being able to use tERC to assess this independently, while using cERCs to differentiate post-labeling sources of variation. Therefore, these two types of ERCs are most valuable when used in combination. This utility was demonstrated through the analysis of the AFX site 1 data. The combination of tERC and cERC information assisted in the determination of sample amplification and labeling yields that differed from other sites and underlies the spread in the variability data.

Our key findings can be summarized as follows. The cERCs exhibit stable and consistent performance across both samples and sites. tERC signal intensities increased and the biological probe signal intensity decreased in proportion to increasing amounts of brain RNA in the samples. When the tERC is added to total RNA samples, it is assumed that the tERC transcripts are at different relative proportions to the pool of biological RNA transcripts. As the abundance of mRNA is relatively higher in sample A as compared to the brain (sample B), the median signal of biological probes was found to be higher in sample A than in sample B, whereas the median tERC signal had the inverse relationship. We further determined that different levels of compression in gene expression exist across commercial platforms, indicating that care must be taken when conducting a cross-platform comparison with respect to making absolute fold-change assessments. And, finally, we also determined that quantile-based normalization approaches, such as those used in PLIER, RMA and GCRMA for Affymetrix and for the Agilent One-Color and GE Healthcare platforms, reveal the variability of the concentration-response slope estimates. This increase in variability may result from the differences in percentage mRNA between samples A and B. Although the median-normalized signals of the tERCs and cERCs are relatively consistent, their relative ranks within samples A and B are different. Quantile normalization forces the distributions of all data sets to be identical, moving the signals for the tERCs and cERCs away from their original raw expression values.

Because no single common standard set of external RNA controls using extended concentration range and a Latin square design are in place for use across platforms in the microarray community, it is not

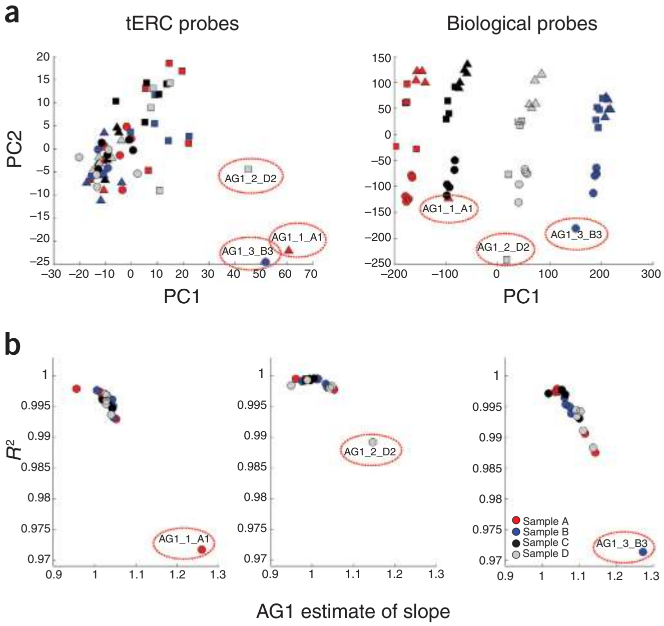


Figure 6 Alternative analysis using ERCs. (**a**) The Principal Component Analysis (PCA) of the Agilent tERC signal intensity is compared with the Agilent biological signal intensities. The graphs are colored by sample and shaped by site (site 1, triangle; site 2, square; site 3, circle). The same three assays (AG1_1_A1, AG1_2_D2, and AG1_3_B3) are potential outliers based on their shift in both the tERC and the biological signal. (**b**) Similar to **Figure 3a**, except that the parameterized sigmoidal curve–fitted linear regression data from the Agilent QC Report concentration-response curves was used to compare R^2 correlation data (*y*-axis) and slope data (*y*-axis). The same three outlying assays identified in the PCA are shown as potential outliers in this analysis (circled in red) demonstrating identification of outlier agreement between two fairly different analyses.

Box 1 Recommendations for the implementation of external RNA controls

- One key benefit of external RNA controls (ERCs) is the ability to get a qualitative assessment of assay performance. This benefit will be more fully realized when an extensive set of ERCs is available.
- A comprehensive study is needed for modeling concentration-response behavior based on large data sets to determine the tolerance ranges for linear fit, slope and *y*-intercept for assay assessment, specifically in the context of false positives and false negatives.
- The development of ERC-specific analysis approaches is encouraged.
- ERCs that are added at both the total RNA level and cRNA level are valuable as they enable failure analysis for different steps of the assay. Using both types of ERCs in the same assay is beneficial for monitoring quality at multiple steps in the process.

yet possible to run the ideal set of external controls for a study of this nature[1]. Thus, the intent of this study was to identify key attributes of ERC performance that should be considered for designing better ERCs and associated analysis approaches in the future, which is one of the many important ERCC endeavors[1]. Based on the findings of this study, several points of consideration are summarized in **Box 1**.

METHODS

MAQC and TGx data sets. There are two types of data sets considered in this study; both are generated from the MAQC project. The difference between these two data sets is the nature of RNA samples used for generating the gene expression data. The MAQC data set used two calibrated RNA samples (A-Stratagene Universal Reference RNA and B-Ambion Brain reference RNA) and their two mixtures (C- 75%A/25%B and D-25%A/75%B). Applied Biosystems data (ABI), Affymetrix GeneChip data (AFX), and Agilent's One-Color platform data (AG1) were generated using these four RNA samples. Each platform comprises a total of 60 microarrays, five technical replicates for each of four samples (A, B, C and D) for one test site (20 microarrays) and data from three test sites were used. In addition, Agilent Two-Color platform data (AGL) were also generated, but using only samples A and B. For AGL, four sets of assays were conducted with five replicates for each set, two dye swap experiments using brain-Cy5/UHRR-Cy3 (sample BA) and UHRR-Cy5/brain-Cy3 (sample AB) along two types of self-self hybridizations with brain-Cy5/brain-Cy3 (sample BB) and UHRR-Cy5/UHRR-Cy3 (sample AA), resulting in a total of 20 assays. The toxicogenomics (TGx) data set applied the RNA samples from rat livers in a TGx study. The detailed experimental protocol is described elsewhere[11]. Briefly, six-week-old Big Blue rats were treated with three compounds for 12 weeks and then killed. The compounds were aristolochic acid, a potent nephrotoxin and carcinogen that is present in plants used in herbal medicines, riddelliine, a carcinogenic pyrrolizidine alkaloid that contaminates various plants, and comfrey, a plant consumed by humans that is a rat liver carcinogen. RNA samples were isolated from livers of the rats treated with three compounds along with a liver control. In addition, RNA samples were also isolated from kidneys associated with treatment of aristolochic acid and a kidney control. Thus, there were a total of six types of rat RNA samples (four from liver and two from kidney). Six biological replicates (rats) were generated for each type of six RNA samples. The gene expression data were generated from four microarray platforms, Applied Biosystems (ABI_Rat), Affymetrix GeneChip (AFX_Rat), Agilent One-Color microarray (AG1_Rat), and GE Healthcare CodeLink (GEH_Rat). For each platform, 36 microarrays were generated, six for each of six groups.

Applied Biosystems external RNA controls. These controls contains a suite of controls (>1,592 control probes) that can be used to check the quality of many aspects of an expression profiling experiment. These controls include the following: blank features, control ladders, hybridization controls, *in vitro* transcription (IVT) labeling controls, reverse transcription labeling controls, negative controls, spatial calibration controls and manufacturing quality controls. Among these controls, we used only the IVT and reverse transcription labeling controls and the hybridization controls, which are spiked at a single fixed concentration. For the hybridization controls, three unlabeled probes are spotted on the microarray: HYB_Control_1_Cp (60 replicates), HYB_Control_2_Cp (60 replicates) and HYB_Control_3_Cp (115 replicates). The hybridization cERCs consist of three digoxigenin-labeled 60-mer oligo control targets supplied with the chemiluminescence detection kit HYB_Control_1_Ct, HYB_Control_2_Ct and HYB_Control_3_Ct. The digoxigenin-labeled oligo targets (cDNA or cRNA) are added to the hybridization mixture. Presence of signal indicates hybridization occurrence and signal strength indicates hybridization stringency. IVT controls consist of three synthetic double-stranded cDNA with a T7 promoter and bacterial control gene sequences: *bioB*, 1,000-nt ds-cDNA; *bioC*, 750-nt ds-cDNA; *bioD*, 600-nt ds-cDNA. Five probes were used for each of three bacterial control genes, *bioB*, *bioC* and *bioD* targeting different regions of the control genes. This resulted in 15 probes and each probe is spotted eight times. Reverse transcription controls consist of three synthetic mRNAs with bacterial control gene sequences: *lys*, 1000-nt mRNA with poly(A) tail; *phe*, 1,400-nt mRNA with poly(A) tail; and *dap*, 1,900-nt mRNA with poly(A) tail. The synthetic mRNAs are added to the reverse transcription reaction with the RNA sample when using the reverse transcription labeling kit or the RT-IVT labeling kit. There are five control probes for each reverse transcription control gene targeting different regions on the gene, and each probe is spotted eight times with a total of 120 reverse transcription control probes. More detail on these controls can be found in http://docs.appliedbiosystems.com/pebiodocs/00113259.pdf and http://docs.appliedbiosystems.com/pebiodocs/04338853.pdf.

Affymetrix external RNA controls. ERCs on GeneChip eukaryotic microarrays include poly-A controls (*lys*, *phe*, *thr* and *dap*) and hybridization controls (*bioB*, *bioC*, *bioD* and *cre*). Poly-A controls are *Bacillus subtilis* genes that are modified by the addition of poly-A tails, and then cloned into pBluescript vectors. The GeneChip Poly-A RNA Control Kit (P/N 900433) contains a presynthesized mixture of *lys*, *phe*, *thr* and *dap*. These poly A–tailed sense RNA samples can be spiked into isolated RNA samples as controls for the labeling and hybridization processes. Hybridization controls consists of *bioB*, *bioC*, *bioD* and *cre*. *BioB*, *bioC* and *bioD* represent genes in the biotin synthesis pathway of *Escherichia coli*; *Cre* is the recombinase gene from P1 bacteriophage. The GeneChip Eukaryotic Hybridization Control Kit (P/N 900299 and 900362) contains a mixture of biotin-labeled cRNA transcripts of *bioB*, *bioC*, *bioD* and *cre*. They can be spiked into the hybridization mixture, independent of RNA sample preparation, and used to evaluate sample hybridization efficiency. More detail can be found in GeneChip Expression Analysis Technical Manual (http://www.affymetrix.com/support/technical/manual/expression_manual.affx) and GeneChip Expression Analysis Data Analysis Fundamentals (http://www.affymetrix.com/support/downloads/manuals/data_analysis_fundamentals_manual.pdf).

Agilent external RNA controls. The Agilent One-Color ERC Kit contains a mixture of ten *in vitro* synthesized, polyadenylated transcripts derived from the Adenovirus E1A gene. These transcripts are premixed at concentrations that span six logs and differ by one log or half-log increments (**Supplementary Table 1** online). The ERC mixture is added to the total RNA, amplified and labeled with Cy3-dye. When the ERCs are used in processing Agilent One-Color microarray assays, the Agilent Feature Extraction (version 8.5) QC Report contains a number of tables and graphs providing information on system performance. These include an indication of the linear portion of the dynamic range of the microarray experiment, the high and low detection limits of the experiment and the reproducibility of the controls with coefficient of variation (CV) percentage calculations across the replicate probes for each of the ten ERCs. For more details, see http://www.chem.agilent.com/scripts/literaturePDF.asp?iWHID=42629. The Agilent Two-Color ERC Kit contains the same ten tERC transcripts as used in the Agilent One-Color platform. Each transcript is premixed into two different ERC mixtures at known concentrations such that the ten transcripts are present in mass equivalents extending across 2.3 logs of concentration and represent ratios spanning from 1:10 to 10:1 (**Supplementary Table 3** online). These two mixtures are spiked into either the Cy3 or Cy5 labeling reactions and colabeled with the total RNA. The Agilent Feature Extraction (version 8.5) QC Report contains a number of tables and graphs providing information on system performance. These include a measure of the expected versus observed log ratios that provide an indication of system accuracy, as well as a determination of the reproducibility of the controls with CV percentage calculations across the replicate probes for each of the ten ERCs. For more details, see http://www.chem.agilent.com/scripts/literaturePDF.asp?iWHID=40485.

GE Healthcare external RNA controls. Each CodeLink Whole Genome bioarray, from GE Healthcare, contains a set of positive-control probes designed against six *E. coli* genes. For each of the six bacterial genes there are five unique probe sequences represented in an 8 redundancy per rat bioarray. Therefore, there are a total of 240 positive-control probes within each bioarray, which are used to assess microarray quality by reporting dynamic range and sensitivity. Each of the six bacterial transcripts is supplied individually as poly-A(+) mRNA, ranging in size from 1,000 to 1,300 ribonucleotides. These control RNAs can be spiked at different concentrations into the total RNA starting material or labeled individually with biotin and spiked into the cRNA before hybridization. The cRNA spiking method, as used in this study, is the manufacturer's recommendation for independently measuring bioarray quality because effects due to sample integrity and purity are circumvented. The positive-control poly-A(+) mRNAs supplied with the CodeLink Expression Assay Reagent Kit are *araB*, *entF*, *fixB*, *hisB*, *gnd* and *leuB*. These transcripts are reverse transcribed and amplified individually, incorporating biotin, and arranged in a dilution series from 50 fM to 50 pM, in fourfold concentration increments. The final hybridization concentrations of biotinylated spikes in the hybridization solution are *araB* (51.2pM), *entF* (12.8pM), *fixB* (3.2pM), *hisB* 0.80pM, *gnd* (0.20fM) and *leuB* (50.0fM). For more details, see http://www4.amershambiosciences.com/APTRIX/upp00919.nsf/Content/WD%3AExternal+RNA+co%28274354027-B500%29?OpenDocument&hometitle=WebDocs.

Microarray data preprocessing and normalization. Data preprocessing and normalization were performed in ArrayTrack, an FDA microarray data management, analysis and interpretation software[18,19]. For Affymetrix GeneChip, five different sets of normalized data were used, PLIER, MAS5, dChip, RMA and GCRMA. Present and Absent Calls were generated for each probe set. For the Agilent One-Color microarray, the raw data (gProcessedSignal data), Median Scaling data and Quantile normalized data were used. Negative values and ERCs were not included in the normalization. For the Two-Color microarray, only the dye-normalized Log Ratio data was used, without any further normalization. For the Applied Biosystems Microarray, signal intensity is associated with two measurements, signal/noise ratio and detection call (or flag). The spots having a ratio >3 and flag <8,191 were considered Present. For GE Healthcare CodeLink, the raw data and quantile-normalized data were used.

Concentration-response curve analysis. An ERC commonly has multiple replicates placed in different positions of a microarray. In the concentration-response analysis, the ERC signal is the mean intensity over the replicates for AG1 and AGL. For Affymetrix, Applied Biosystems and GE Healthcare platforms, an ERC gene consists of multiple probes targeting different regions of the ERC gene. Thus the ERC signal is calculated by first averaging the signals from different probes of the same gene and then the mean value is calculated over multiple replicates.

The concentration-response curve shown in **Figure 2** was generated by plotting the concentration of either tERC (spiked poly-A molar ratio) or cERC (spiked concentration in pM) on the *x*-axis as a function of signal intensity on the *y*-axis. The amount of cERC added to the hybridization mixture can be expressed in molar amounts based on the mass of the cERC transcript added to a specific volume of the hybridization mixture. Determining the final molar amount of tERCs in the final hybridization mixture is more difficult. One method is to express the ERC as a mass fraction of the total RNA used in the experiment, which has been recommended by ERCC[1]. A second method is to use a number of assumptions to determine the poly-A mass ratios. The assumptions used for this paper are that the average percentage of mRNA in total RNA is 2%, the average transcript length is 2,000 bases and the average molecular weight of a single base is 330 g/mol. Using these assumptions and the known length of the individual tERCs, the poly-A mass ratio for the different tERCs was calculated. Both cERC concentration and tERC poly-A molar ratio used for analysis are summarized in **Supplementary Table 1**.

The linear regression analysis of the concentration-response curve was based on the linear portion of the curve (**Fig. 3**), which were generated in JMP Genomics (http://www.jmp.com/). All ERCs were used in analysis for both AFX and GEH_Rat but only six of ten tERCs were applied for AG1 by removing one top tERC at the signal saturation range and three bottom tERCs at the noise level. Agilent's Feature Extraction QC Report uses a similar algorithm for the same analysis. In this method, the concentration-response curve fit to the linear portion was performed on a log-log plot after a parameterized sigmoidal curve fit of the data.

Note: Supplementary information is available on the Nature Biotechnology website.

DISCLAIMER

This work includes contributions from, and was reviewed by, the FDA. The FDA has approved this work for publication, but it does not necessarily reflect official Agency policy. Certain commercial materials and equipment are identified in order to adequately specify experimental procedures. In no case does such identification imply recommendation or endorsement by the FDA, nor does it imply that the items identified are necessarily the best available for the purpose.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details).

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. ERCC. Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics* **6**, 150 (2005).
2. ERCC. The External RNA Controls Consortium: a progress report. *Nat. Methods* **2**, 731–734 (2005).
3. Hill, A.A. *et al.* Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol* **2**, RESEARCH0055 (2001).
4. Rajagopalan, D. A comparison of statistical methods for analysis of high density oligonucleotide array data. *Bioinformatics* **19**, 1469–1476 (2003).
5. Irizarry, R.A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
6. Irizarry, R.A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
7. Freudenberg, J., Boriss, H. & Hasenclever, D. Comparison of preprocessing procedures for oligo-nucleotide micro-arrays by parametric bootstrap simulation of spike-in experiments. *Methods Inf. Med.* **43**, 434–438 (2004).
8. Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M. & Halfon, M.S. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol.* **6**, R16 (2005).
9. Dabney, A.R. & Storey, J.D. A reanalysis of a published Affymetrix GeneChip control dataset. *Genome Biol.* **7**, 401 (2006).
10. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**, 1151–1161 (2006).
11. Guo, L. *et al.* Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol.* **24**, 1162–1169 (2006).
12. Shippy, R. *et al.* Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nat. Biotechnol.* **24**, 1123–1131 (2006).
13. "Guide to Probe Logarithmic Intensity Error (PLIER) Estimation", Affymetrix Technical Note, http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf
14. Microarray Suite User's Guide, Version 5.0, http://www.affymetrix.com/support/technical/manuals.affx
15. Wu, Z., Irizarry, R.A., Gentleman, R., Murillo, F.M. & Spencer, F.A model based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* **99**, 909–917 (2004).
16. Li, C. & Wong, W. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31–36 (2001).
17. Fang, H., Xie, Q., Boneva, R., Fostel, J., Perkins, R. & Tong, W. Gene expression profile exploration of a large dataset on chronic fatigue syndrome. *Pharmacogenomics*, **7**, 429–440, (2006).
18. Tong, W. *et al.* ArrayTrack–supporting toxicogenomic research at the US Food and Drug Administration National Center for Toxicological Research. *Environ. Health Perspect.* **111**, 1819–1826 (2003).
19. Tong, W. *et al.* Development of public toxicogenomics software for microarray data management and analysis. *Mutat. Res.* **549**, 241–253 (2004).