nature biotechnology

Evaluation of DNA microarray results with quantitative gene expression platforms

Roger D Canales[1,10], Yuling Luo[2,10], James C Willey[3,10], Bradley Austermiller[3], Catalin C Barbacioru[1], Cecilie Boysen[4], Kathryn Hunkapiller[1], Roderick V Jensen[5], Charles R Knight[6], Kathleen Y Lee[1], Yunqing Ma[2], Botoul Maqsodi[2], Adam Papallo[5], Elizabeth Herness Peters[6], Karen Poulter[1], Patricia L Ruppel[7], Raymond R Samaha[1], Leming Shi[8], Wen Yang[2], Lu Zhang[1] & Federico M Goodsaid[9]

We have evaluated the performance characteristics of three quantitative gene expression technologies and correlated their expression measurements to those of five commercial microarray platforms, based on the MicroArray Quality Control (MAQC) data set. The limit of detection, assay range, precision, accuracy and fold-change correlations were assessed for 997 TaqMan Gene Expression Assays, 205 Standardized RT (Sta)RT-PCR assays and 244 QuantiGene assays. TaqMan is a registered trademark of Roche Molecular Systems, Inc. We observed high correlation between quantitative gene expression values and microarray platform results and found few discordant measurements among all platforms. The main cause of variability was differences in probe sequence and thus target location. A second source of variability was the limited and variable sensitivity of the different microarray platforms for detecting weakly expressed genes, which affected interplatform and intersite reproducibility of differentially expressed genes. From this analysis, we conclude that the MAQC microarray data set has been validated by alternative quantitative gene expression platforms thus supporting the use of microarray platforms for the quantitative characterization of gene expression.

To evaluate performance characteristics of gene expression measurement technologies and the data they generate, one must identify alternative quantitative platforms that can be used as references. The MAQC consortium used the TaqMan assays, Standardized (Sta)RT-PCR and QuantiGene platforms for this purpose because these platforms had been shown to have high assay specificity and detection sensitivity, broad linear dynamic range and high signal-to-analyte response[1–4]. The platforms were used to evaluate some of these performance characteristics in each commercial whole genome microarray platform investigated in the MAQC study. In addition, we report the fold-change correlation of each alternative quantitative platform relative to these microarray platforms. We observed high correlations between the quantitative platform measurements and the data derived from the microarrays and were also able to identify the sources of variability among microarray platforms relative to the quantitative platforms.

Here we define validation as a measure of the concordance and discordance of the microarray data with the quantitative reference platforms selected—we used the results of the quantitative platforms as a reference against which to evaluate the microarray platforms. We have thus not attempted to establish a 'gold standard' for expression measurements but a solid reference point to allow data validation.

Quantitative, real-time PCR has been developed over the last decade to specifically measure template molecule numbers[4,5]. The development of fluorogenic probes[6] enabled accurate quantification of PCR products through measurement of a fluorescence signal during the exponential amplification phase. TaqMan Gene Expression Assays are based on the use of the 5′ nuclease activity of *Taq* polymerase to hydrolyze a target-specific, dual-labeled, fluorogenic hybridization probe during the extension phase[7]. The number of template transcript molecules in a sample is determined by recording the amplification cycle in the exponential phase (cycle threshold or C_T), at which time the fluorescence signal can be detected above background fluorescence. Thus, the starting number of template transcript molecules is inversely related to C_T—the more template transcript molecules at the beginning, the lower the C_T[7,8]. TaqMan assays have been used in recent studies to validate microarray data[9–11].

StaRT-PCR[4,12] is a competitive PCR-based platform that enables endpoint quantification of PCR products. After RNA is converted to cDNA, the cDNA is added to a standardized mixture of internal standard (SMIS) competitive templates, aliquoted into microplate wells containing gene-specific PCR primers and amplified for 35 cycles. The individual endpoint StaRT-PCR products are then separated by size and quantified by high-throughput microfluidic electrophoresis. StaRT-PCR has also been used in studies to validate microarray data[1] and has been used to generate potential biomarkers for disease stratification[13,14].

[1]Applied Biosystems, 850 Lincoln Centre Dr., Foster City, California 94404, USA. [2]Panomics, Inc., 6519 Dumbarton Circle, Fremont, California 94555, USA. [3]University of Toledo, Toledo, Ohio 43614, USA. [4]ViaLogy Corp., 2400 Lincoln Avenue, Altadena, California 91001, USA. [5]University of Massachusetts-Boston, 100 Morrissey Blvd., Boston, Massachusetts 02125, USA. [6]Gene Express, Inc., 975 Research Drive, Toledo, Ohio 43614, USA. [7]Innovative Analytics, 7107 Elm Valley Dr., Kalamazoo, Michigan 49009, USA. [8]National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Rd., Jefferson, Arkansas 72079, USA. [9]Center for Drug Evaluation and Research, US Food and Drug Administration, 10903 New Hampshire Ave., Silver Spring, Maryland 20993, USA. [10]These authors contributed equally to this work. Correspondence should be addressed to F.M.G. (Federico.Goodsaid@fda.hhs.gov).

Published online 8 September 2006; doi:10.1038/nbt1236

The QuantiGene Reagent System[15] detects DNA and RNA directly without a reverse transcription step. It is a sandwich nucleic acid hybridization platform in which targets are captured through cooperative hybridization of multiple probes[16]. This complex is detected through signal amplification by a branched DNA amplifier and chemiluminescence signal generation. The QuantiGene assay has been used in US Food and Drug Administration–approved clinical diagnostic products for quantitative viral load determination of HIV, hepatitis C virus and hepatitis B virus with detection sensitivity of <50 transcript molecules[17–19]. Because the QuantiGene assay can measure gene expression either by measuring RNA directly without a reverse transcription step, or by measuring cDNA without PCR amplification, it provides an independent method of measurement relative to the quantitative reverse transcription (RT)-PCR and microarray platforms.

Application of these quantitative platforms in the MAQC project increased the confidence in concordance observed between the microarray platforms. In addition, the results obtained from using these platforms allowed us to explore the sources of variability among microarray platforms. With this comprehensive evaluation, we demonstrate the value of alternative quantitative platforms as tools for the independent validation of microarray data and the resolution of discordant results.

RESULTS

Assay performance of three alternative quantitative platforms

The MAQC consortium selected a list of 1,297 genes to evaluate and compare the performance of microarray and alternative quantitative platforms and to identify and analyze discordant results. TaqMan assays, StaRT-PCR and QuantiGene assays were performed on 997, 205 and 244 of the 1,297 genes, respectively. Gene lists used for analysis of selected performance metrics for quantitative platforms, and for analysis of concordance between the quantitative platforms and microarrays are shown in **Supplementary Table 1** online.

Four RNA samples A, B, C and D, provided by the MAQC consortium, were analyzed[20]. TaqMan assays were done in quadruplicate, and StaRT-PCR assays in triplicate, on cDNA generated from 10 ng total RNA (**Supplementary Methods** online). Both the TaqMan assays and StaRT-PCR were based on cDNA from a single reverse transcription reaction. QuantiGene assays were performed in triplicate directly from 500 ng of total RNA (**Table 1**). Performance metrics presented are not directly comparable because each platform assayed a different gene set, and had different assay ranges of measurements and signal-to-analyte response.

Detection sensitivity

TaqMan assay quantification is directly related to C_T. A gene is not detectable when the average $C_T > 35$ cycles. By this definition, 857 genes (86%) were detectable in both A and B. The StaRT-PCR detection limit is defined as ten transcript molecules. By this definition, 193 genes (94%) were detectable in both A and B. For QuantiGene the detection limit is defined as a signal three standard deviations (s.d.) above the background. By this standard, 223 genes (91.4%) were detectable in both A and B.

Assay range

The assay range represents the difference in signals measured on a $\log_{10}$ scale between genes with the highest and the lowest expression. The assay range for TaqMan assays was 8.1 with C_T values ranging from 8 (>10^8 transcript molecules) for 18S rRNA to 35 (~5 transcript molecules) for low expressors. For StaRT-PCR, the assay range was 6.8 with normalized transcripts of 6.4×10^7 transcript molecules for 18S rRNA to 10 transcript molecules for low expressors. For QuantiGene, the assay range was 4.1 with the highest assay range of 599 relative luminescence units (RLU) for *LDHA* and the lowest detectable signal of 0.045 RLU for *SPARCL1*.

Precision

The precision of the three alternative quantitative platforms was measured by coefficient of variance (CV) (**Fig. 1** and **Table 1**) or s.d. (**Supplementary Fig. 1** online). There were interplatform differences in the number of transcript molecules (RNA or cDNA) loaded into each assay. Because of differences in the amount of sample loaded (**Table 1**), a majority of the genes measured with QuantiGene contained >6,000 transcript molecules in the assay, whereas a majority of those measured by TaqMan assays and StaRT-PCR had less. These two platforms were used to assess the previously reported stochastic process involved in the relationship between transcript molecules loaded and CV[21]. A clear trend of increased CV with decreasing abundance of transcripts was observed for TaqMan assays and StaRT-PCR when <6,000 transcript

Table 1 Summary of platform performance metrics

Platform	Gene list	Sample processing			Detection sensitivity[a]		Dynamic range[b] (log10)	Precision[c] (median)		Accuracy[d] (median)		
Symbol	Number of genes tested	Sample input	Assay replicates	Data presentation	Both A & B above LOD	Both A & B below LOD		All data	>6,000	Linearity[e] (R2)	RA (%median)[f]	RA (%variance)[g]
TAQ	997	cDNA from 10 ng total RNA, one RT reaction	Four replicates of cDNA	Normalized against POLR2A	857 (86%)	38 (3.8%)	8.1	3.46	2.42	0.950	3.6	9.4
GEX	205	cDNA from 10 ng total RNA, one RT reaction	Three replicates of cDNA	Normalized against beta-actin	193 (94%)	4 (2.0%)	6.8	6.26	3.82	0.96[h]	0.4[h]	21.1[h]
QGN	244	500 ng total RNA	Three replicate of RNA directly	Original data	223 (91%)	5 (2.0%)	4.1	2.16	2.12	0.994	1.0	5.0

[a]Detection sensitivity: the number (percent) of detectable or undetectable genes in both sample A&B based on each platform's detection limit. [b]Assay range: based on the ratio of (highest detectable signal/lowest detectable signal) of all the genes and samples measured in each platform. [c]Precision: based on median value of CV measured either a) in all genes and all samples in each platform or b) in samples with 6,000 transcript molecules or above. [d]Based on formula C = 0.25A +0.75B and D = 0.75A + 0.25B for TaqMan assays and QuantiGene and C = 0.88A + 0.12B and D = 0.45A + 0.55B for StaRT-PCR. [e]Linearity: based on the median R^2 slope of the linear fit of assay signal from sample A, B, C, D for all the detectable genes with greater than twofold difference between A and B. 829, 125 and 223 genes are analyzed for TaqMan, StaRT-PCR, and QuantiGene, respectively. [f]RA score (% median): RA (relative accuracy) score for sample C and D for a gene is defined as (C-C′/C′) and (D-D′/D′), which represents the percent difference of experimental from the expected. Median value of % RA score for both sample C and D combined is presented here. Only detectable genes in both A & B are analyzed for each platform. [g]RA score (% variance): median value of the absolute RA scores for both sample C and D combined is presented here. [h]Based on a recalibrated data set (**Supplementary Methods**).

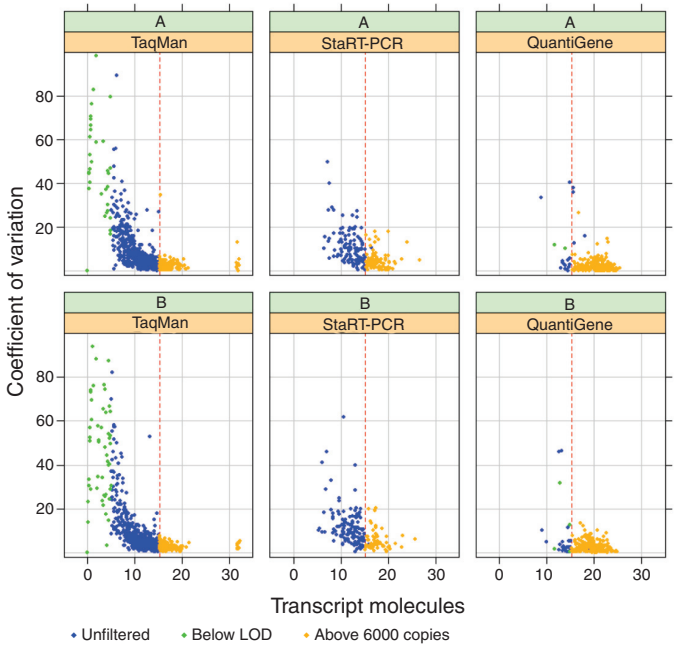

Figure 1 Effect of the number of transcript molecules on assay precision. The measured (StaRT-PCR) or estimated (TaqMan assays and QuantiGene) number of transcript molecules loaded into an assay for each gene in sample A or B was plotted against its CV. The data for the three platforms were transformed to be on the same *x*-axis scale as described in Methods. The vertical dashed line is ~6,000 transcript molecules; blue symbol, assays detecting <6,000 transcript molecules; orange, assays detecting >6,000 transcript molecules; green, assays below the limit of detection. LOD, limit of detection.

molecules (below dashed line in **Fig. 1**) were loaded as also specified in **Table 1**. For the TaqMan and StaRT-PCR platforms, each cDNA sample was split for replicate measurements, so precision measurement did not include the reverse transcription reaction. For the QuantiGene platform, replication encompassed the entire process from total RNA to chemiluminescent detection.

Relative accuracy

Relative accuracy was defined as the proximity of observed expression values for C and D to the predicted values based on measured expression values for A and B. Error handling for all platforms was on a linear scale with the exception of TaqMan assays in which errors increased exponentially because C_T is transformed to number of molecules. The percent difference between the predicted signal C′ and D′ and the actual assay signal C and D could be used as an indication of relative assay accuracy (RA). An RA score ΔC and ΔD for a target gene was defined as (C–C′/C′) and (D–D′/D′), respectively. The distribution of percent difference from expected (RA score) for each gene was presented in a box plot for each platform (**Fig. 2** and **Table 1**). The median percent difference from expected for both C and D was 3.6, 0.4, 1.0 for TaqMan assays, StaRT-PCR and QuantiGene, respectively, which are all closely centered around zero. The median distribution of the absolute value of RA scores (|ΔC| and |ΔD|) indicates the variance of percent difference between the predicted signal C′ and D′ and the actual assay range C and D. For TaqMan assays, the median variance value for 856 genes for both C and D was 9.4; for StaRT-PCR (193 genes) it was 21.1 and for QuantiGene (223) genes it was 5.0. The data for the QuantiGene platform are notable given that these values encompass the system-wide accuracy of the platform.

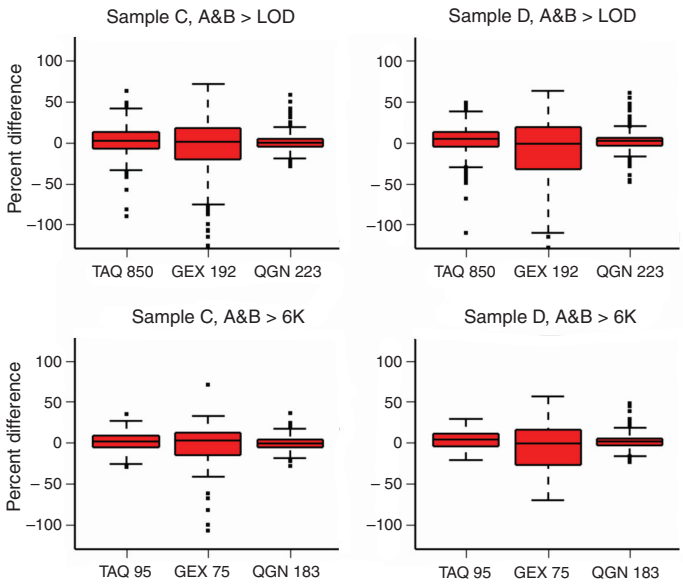

Figure 2 Analysis of assay accuracy. The values measured for C and D were compared to the values expected (% difference) based on measured A and B values. Formulas used to calculate expected C and D are provided in text. Box plot components are: horizontal line, median; box, interquartile range; whiskers, 1.5× interquartile range; black squares, outliers. TAQ, TaqMan assays; GEX, StaRT-PCR assays; QGN, QuantiGene assays; LOD, limit of detection. The number of genes for each platform is shown.

Fold-change correlation

To evaluate the concordance of fold changes between the alternative quantitative platforms, we performed regression analysis of fold differences in sample A compared to sample B. This analysis was performed using pair-wise common gene sets between platforms because the overlap between the three platforms was limited to 48 genes (**Fig. 3**). The R^2 and slope for TaqMan assays versus StaRT-PCR (92 common genes) were 0.88 and 0.93, respectively; for QuantiGene versus TaqMan assays (193 common genes), 0.81 and 0.78, respectively; and for QuantiGene versus StaRT-PCR (55 common genes), 0.85 and 0.77, respectively. Although linear regression analysis indicates good fold-change correlation across the three platforms, the respective slopes indicate compression or expansion effects between the platforms.

Concordance of microarrays with alternative quantitative platforms

We used the results of the alternative quantitative platforms as a reference to evaluate concordance with microarray platforms. For cross-platform comparison to microarrays, we evaluated four parameters (**Figs. 4** and **5**): (i) detection sensitivity, the ability of the microarrays to detect genes that were called 'present' by each alternative quantitative platform; (ii) the fold-change correlation between microarrays and each alternative quantitative platform; (iii) true positive rate (TPR), the concordance of genes called statistically differentially expressed by the TaqMan assay that are also called statistically differentially expressed in the microarrays; (iv) false discovery rate (FDR), the concordance of genes differentially expressed in microarrays that are not differentially expressed in the TaqMan assay. TaqMan assays were evaluated for all parameters, whereas StaRT-PCR and QuantiGene were evaluated only

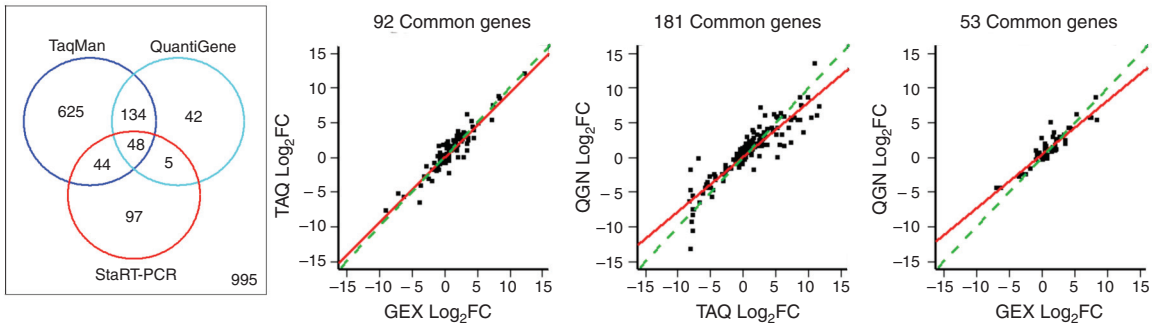

Figure 3 Correlation of fold change between alternative quantitative platforms. The sample B over sample A (B/A) fold changes ($\log_2$) for each gene common between two platforms were subjected to bivariate analysis. (**a**) TaqMan assays versus StaRT-PCR. (**b**) QuantiGene versus TaqMan assays. (**c**) QuantiGene versus StaRT-PCR. The dashed line on each graph represents the ideal slope of 1.0. The solid lines represent a linear regression fit. The overlapping gene list among the alternative quantitative platforms is represented in the Venn diagram. Linear fit: TaqMan assay versus StaRT-PCR, $Y = -0.03647 + 0.9347X$, $R^2 = 0.879$; QuantiGene versus TaqMan assay, $Y = 0.14 + 0.7825X$, $R^2 = 0.8118$; QuantiGene versus StaRT-PCR, $Y = 0.4095 + 0.7707X$, $R^2 = 0.8497$.

for parameters i and ii because fewer genes were assayed for these platforms. Detailed site-by-site analysis of genes is provided for StaRT-PCR and QuantiGene in **Supplementary Table 2** online and for TaqMan assays in **Supplementary Figure 2** online.

Detection sensitivity analysis was done for each alternative quantitative platform using the genes common to that platform and each of the microarray platforms. For this reason, assay ranges and expression characteristics of gene sets differed. There were 845, 157 and 197 genes determined to be present in sample A by TaqMan assays, StaRT-PCR and QuantiGene, respectively. At the lower ranges of gene expression, for each microarray, the fraction of genes detected decreased relative to each of the alternative quantitative platforms (**Fig. 4a–c**). In addition, detection sensitivities relative to each alternative quantitative platform varied among the microarray platforms.

A fold-change comparison between each alternative quantitative platform and each microarray platform was also performed using LOWESS smoothing (**Fig. 4d–f**, ref. 22), which does not assume a linear relationship of fold-change values between platforms. We used a total of 392, 101 and 83 genes that were present in samples A and B at each site measured by each microarray platform and shared with TaqMan assays, StaRT-PCR and QuantiGene, respectively, for comparison. Although excellent fold-change correlations were observed, varying degrees of compression of signal-to-analyte response relative to the alternative quantitative platforms were also found. These data are consistent with the analysis presented elsewhere in this issue[20]. An additional analysis was done to show that compression effects are detectable for both low and high expressors (**Supplementary Fig. 3** online).

Traditionally, analysis of accuracy is carried out by analyzing the true positive rate (TPR) and false discovery rate (FDR). In this case, the actual rates were unknown. For this reason, we compared the microarray platforms to TaqMan, which became the reference platform. Using TaqMan assay calls as the reference, we constructed contingency tables against microarray platforms, in which the concordance was determined and both the *P*-value significance of the *t*-test and fold-change directionality (up- or downregulation) were taken into consideration. Specifically, true positives (TP) are genes differentially expressed (significant *P* value for the *t*-test) in both TaqMan and microarray platforms with fold change in the same direction; true negatives (TN) are genes not differentially expressed in either platform; false positives (FP), consist of two sets of genes: (i) genes not differentially expressed in TaqMan and differentially expressed in microarrays, or (ii) genes differentially expressed in both platforms with fold change in the opposite direction; false negatives (FN), genes differentially expressed for TaqMan and not for microarrays.

For TPR analysis in TaqMan assays, microarrays were compared to genes considered differentially regulated at fold-change cut-offs of 0, 1.5 and 2.0 (**Fig. 5a–c, Supplementary Table 3** online). For microarrays, differential expression was measured using a *t*-test and controlling for FDR at a 5% level[23] for genes present in either sample A or B. For approximately half of the assay range assessed by TaqMan assays, there were consistent TPR values across array platforms. However, it is apparent that at low expression, detection percentages were directly proportional to TPR. As a result, there was also variation (up to 20%) in TPRs between array platforms (**Fig. 5a**, **Supplementary Table 3** online). FDR analysis (**Fig. 5d–f**, **Supplementary Table 3** online) using TaqMan assays as a reference also showed consistent FDRs for genes expressed at medium and high levels for the microarray platforms. As expected, alternative quantitative platforms showed ~5% discordance with arrays in agreement with the FDR cut-off used for defining differential expression in microarrays. However, genes expressed at low levels showed a variable and inverse relationship to FDR values (**Fig. 5d**, **Supplementary Table 3** online). These results support the idea that differential expression measurement depends on the detection limit for each microarray platform.

Discordant gene analysis

Alternative quantitative platforms can also be used to resolve discordance among the microarray platforms because specific assays can be designed easily to identify the source of the discordance by probing different regions. Analysis of extremely discordant results among the 997 genes shared by microarray platforms and TaqMan assays resulted in 9 genes (~1%) that exhibit twofold or greater changes in opposite directions on different platforms with $P < 0.0001$ (**Supplementary Table 4** online). Some of these genes such as *POMC*, *LTA* and *EPHA7* (**Supplementary Fig. 4** online) were considered low expressors by TaqMan assays (C_T values > 32) and, as expected, were undetected in a majority of the microarray platforms. However, some genes appeared to exhibit true discordance, of which three (*ELAVL1*, *IGFBP5*, *ABCD1*) were selected for further analysis by the three alternative quantitative platforms. To investigate the nature of the discordance, we designed probes against different regions of the three genes. For *IGFBP5* and *ABCD1*, alternative quantitative platform probes indicate consistently lower expression in sample A along the length of the transcripts (**Fig. 6**, **Supplementary Table 5** online). These results suggest that discordance between the platforms in some cases is likely to be a result of cross-hybridization of microarray probes with other sequences. For *ELAVL1*, alternative quantitative platform probes were able to evaluate differential expression characteristics of the 5′ and 3′ ends of the gene. This result is consistent with a mapping

study showing that *ELAVL1* has two alternative polyadenylation sites (unpublished observations). We also investigated some genes (*DPYD*, *PTGS2*, *FURIN*) that were discordant between the alternative quantitative platforms. *DPYD* discordant results were determined to be a result of probing different sequence locations in the gene. When probes from each alternative quantitative platform were designed to interrogate similar sequences, expression characteristics along the length of the gene were found to be in concordance. Although more 5′ probes appeared to have discrepancies in directionality of expression, these differences were found to be statistically insignificant ($P > .01$). Multiple probe locations for *PTGS2* generated expression differences in the same direction of change across all three platforms. The only gene that remained discordant after using multiple probe designs for each of the three platforms was *FURIN*. For this gene both TaqMan assays and StaRT-PCR detected differential expression in probes specific to the 5′ end of the gene. Although all platforms interrogate this region of the gene, the smaller probes (TaqMan assays; base 25–95 and StaRT-PCR; base 22–182) may be detecting a splice variant not detected by probes interrogating a longer region of the gene (QuantiGene; base 1–501). Thus, by designing probes against different regions of a gene, alternative quantitative platforms can confirm location-specific expression characteristics of genes and aid in the resolution of discordant gene expression data.

DISCUSSION

We have assessed three quantitative gene expression measurement technologies for their performance metrics, correlated the results obtained with them to DNA microarray data and then subsequently used them as a means to identify sources of discordance among microarray platforms. Our results show a good correlation between quantitative platform measurements and microarray data. This is true, regardless of whether RNA or cDNA levels were measured.

A primary focus of this study was to identify possible sources of discordance. On the basis of data reported here, we have identified specific reasons that partially explain why, as previously reported[22], groups of genes detected as differentially expressed on a particular microarray platform are occasionally not reproducible across microarray platforms.

Whereas alternative quantitative platforms could detect over 85% of the genes shared across alternative quantitative and array platforms in this study, microarray platforms were less sensitive in the detection of lower expressed genes in this set (**Fig. 4a–c**, **Supplementary Table 2** and **Supplementary Fig. 2** online). In addition, relative to the alternative quantitative platforms, detection levels varied by as much as 60% among microarray platforms for lower expressed genes in this set. Since significant differential expression in microarrays is largely dependent on the ability to reliably detect expression, intersite and interplatform variation can lead to discordant results in the gene lists.

Using TaqMan assays as a reference, TPR and FDR for the various microarray platforms differed across the assay range (**Fig. 5a,d**, **Supplementary Table 3** online). TPR was directly correlated to percent of detectable genes whereas FDR was inversely correlated, indicating that although this metric reflects the ability of each platform to detect expression, it may also be subject to the stringency defined by the array manufacturer in applying detection calls. The consequences of these varying stringencies are that whereas a relaxed stringency in detection calls can lead to better detection and differential expression concordance, there will be a higher percentage of false positives. **Supplementary Figure 2** online verifies that the discordance in differential expression is related to the intersite and interplatform variation in detection.

Using StaRT-PCR or QuantiGene as references and more stringent criteria in which a fold-change cutoff of 2.0 was applied for genes that were considered present in at least three out of five replicates in both A and B samples did not eliminate intersite or interplatform variation in detection of differentially expressed genes (**Supplementary Table 2** online). It is clear that this variation is nearly exclusively for genes expressed at low level. Even with these more stringent selection criteria, intersite variation in detection resulted in intersite and interplatform variation in lists of differentially expressed genes.

Another source of discordance in differentially expressed genes in this study was interplatform variation in compression. Using alternative quantitative platforms as a reference, interplatform variation in signal-to-analyte response was observed (**Fig. 4d–f**) and it was particularly large among genes expressed in the high or low range (**Supplementary Fig. 3** online). This platform-dependent compression was associated with discordance in differentially expressed genes (**Supplementary Table 2** online).

Whereas these results have identified specific causes of discordance in lists of detected, and/or differentially expressed genes, we found excellent fold-change correlation between each quantitative platform and each microarray platform for those genes that were detected by microarray platforms (**Fig. 4d–f**). Of the 845 genes detected in the microarray

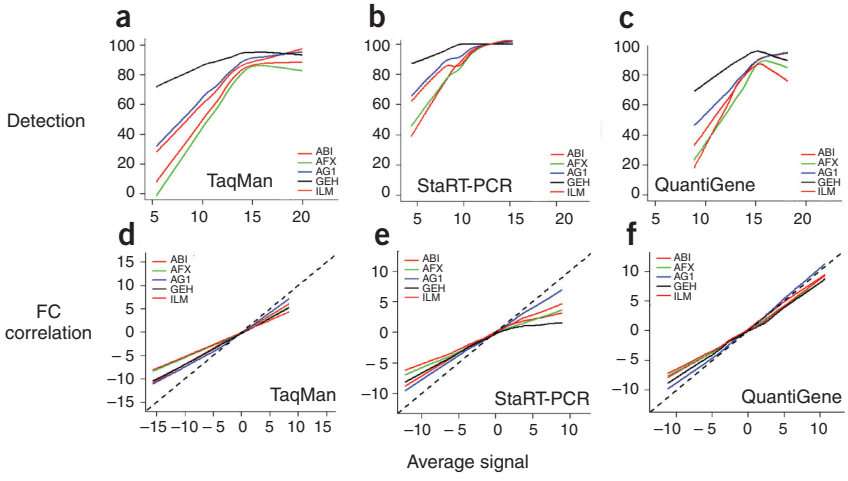

Figure 4 Performance of microarray platforms relative to alternative quantitative platforms. (**a**–**c**) Sensitivity of detection. Each microarray platform was compared to TaqMan (**a**), StaRT-PCR (**b**) or QuantiGene (**c**) for ability to detect genes expressed in sample A. Genes were analyzed based on present call criteria of being present in 3/5 replicates at one of the three microarray sites and in the majority of replicates for each alternative quantitative platform (at least 3/4 for TaqMan, 2/3 for StaRT-PCR and QuantiGene). Genes detected by each alternative quantitative platform were sorted according to their signals (scaling as described in **Fig. 1**), and the percent of genes detected by both microarray and alternative quantitative platforms from bins of 30 consecutive genes (*y* axis) were plotted against the average signal of those genes measured by the alternative quantitative platform (*x* axis). (**d**–**f**) Correlation of fold change measured by each microarray platform compared to TaqMan (**d**), StaRT-PCR (**e**) or QuantiGene (**f**). Pairwise Sample A to Sample B fold-change comparison, measured by each alternative quantitative platform (*x* axis) compared to each microarray platform (*y* axis). For each microarray platform, only genes present in both samples at each site were called present. Each line represents the Lowess smoothing fitting curve. The number of genes involved in each analysis varies with the platforms compared.

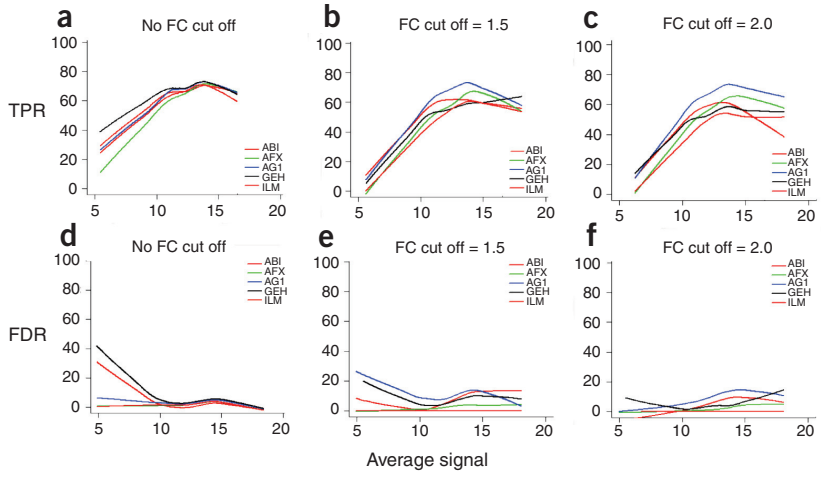

Figure 5 Assessment of true positive rates and false discovery rates using TaqMan assays. (**a–c**) True positive rate (TPR) assessment using TaqMan assays. All common genes between TaqMan assays and microarray platforms were used for the TPR analysis. TPR was defined as the percentage of differentially expressed genes in sample A compared to sample B detected by each microarray platform out of the ones detected by TaqMan assays data as truth [TPR = TP/(TP+FN)], where TP is true positive and FN is false negative in microarray. Differential expression was detected by *t*-test, where false discovery rate (FDR) was controlled at the 5% level with fold-change filters of 0 (d), 1.5 (e) and 2.0 (f). For TaqMan assays, genes were ordered according to the average signals of A and B and for bins of 50 consecutive genes, we compared the significant difference calls between each microarray platform and TaqMan assays. Concordance of differential expression was assessed for each platform. (**d–f**) False discovery rate (FDR) assessment using TaqMan assays. All common genes between TaqMan assays and microarray platforms were used for the FDR analysis. FDR was defined as FP/(TP + FP), where FP is false positive in microarrays. The FDR represents the percentage of differentially expressed genes detected only by microarray platforms out of all genes differentially expressed in microarray platforms. Notice that the FDR (relative to TaqMan assays) is slightly larger than 5%, which is expected from Benjamini Hochberg (BH) adjustment for multiple testing. Differential expression was detected by *t*-test (FDR at 5%), with fold-change level filters of 0 (**d**), 1.5 (**e**) and 2.0 (**f**).

platforms and commonly mapped to one or more of the alternative quantitative platforms, only 9 (1%) were 'extremely' discordant. A major factor contributing to these infrequent discordant results is differences in probe location. Assays designed to different locations of the discordant genes in this study demonstrated a utility of the alternative quantitative platforms (**Fig. 6**) to independently validate gene expression measurements from array platforms.

This analysis was also useful in the study of discordance observed between alternative quantitative platforms. For example, discordant expression results for *FURIN* observed in alternative quantitative platforms is consistent with a probe location difference. The limited common gene list precluded a detailed analysis of the discordance caused by low expression genes among alternative quantitative platforms. In addition, another source of potential discordance may come from the difference of measuring mRNA directly versus measuring cDNA, which were not analyzed here.

In summary, analysis of the MAQC samples by three alternative quantitative platforms revealed excellent fold-change correlation with microarray platform data while enabling identification of possible sources of intersite and interplatform discordance in lists of genes measured as differentially expressed. Advantages of the alternative quantitative platforms were partially due to assay specificity, lower detection threshold and expanded assay range. Another advantage was the ease with which they interrogated specific gene locations due to their flexible assay design. Further, analysis by these alternative quantitative technologies contributed to characterization of the MAQC samples and confirmed their value in guiding optimization of gene expression methods.

METHODS

Sample definition. Sample A was Universal Human Reference RNA (Stratagene) and sample B was human brain total RNA (Ambion). Concentrations of A and B were normalized based on total RNA as measured by OD_{260}. C was a 3:1 volumetric mixture of A and B, and D was a 1:3 volumetric mixture of A and B.

Selection of genes for validation by alternative quantitative platforms. A list of 1,297 RefSeqs was selected by the MAQC consortium. Over 90% of these genes were selected from a subset of 9,442 RefSeq common to the four platforms (Affymetrix, Agilent, GE Healthcare and Illumina) used in the MAQC Pilot-I Study (RNA Sample Pilot), based on annotation information provided by manufacturers in August 2005. This selection ensured that the genes would cover the entire intensity and fold-change ranges and include any bias due to RefSeq itself. To aid in the titration study, we included a subset of (~100) genes based on tissue-specificity (A versus B). To address cross-platform data inconsistency, we also included another subset, which showed the largest variability in log2 fold change across platforms in the Pilot-I Study. Platform vendors were queried about their 'favorite' genes (e.g., *CYP* family, *PPARA*, *HDAC* family and a small number of these were included). Consideration was also given to the inclusion of genes that were available from QuantiGene and StaRT-PCR platforms. The final list was therefore not completely unbiased.

Gene list for the MAQC study by alternative quantitative platforms. TaqMan assays: 1,000 TaqMan gene expression assays used in the study that matches with the MAQC gene list. These 1,000 assays were selected from > 200,000 available human TaqMan assays (>20,000 NCBI genes) and covered 997 genes (3 genes had more than one assay). StaRT-PCR: 103 genes were selected from the nearly 800 genes for which StaRT-PCR reagents are already available that match with the MAQC gene list. All genes that overlap with those measured by TaqMan assays and QuantiGene were included as well as an additional 102 genes for a total of 205. QuantiGene: we selected 245 QuantiGene assays (covered 244 genes) that matched with the MAQC gene list from nearly 2,600 genes for which QuantiGene probe sets are already available. All genes that overlap with those measured by TaqMan assays and StaRT-PCR were included. 55 genes were in common to all three alternative quantitative platforms.

TaqMan assays. RNA Samples: total RNA samples A (universal human reference RNA (UHRR), Stratagene), B (brain, Ambion), C (3 UHRR:1 brain) and D (1 UHRR:3 brain) as described earlier were used for all TaqMan assays. There was no additional treatment to these samples before cDNA preparation. cDNA Preparation: cDNA was prepared from total RNA Sample A, B, C and D using Applied Biosystems cDNA Archive Kit and random primers. Multiple reactions containing 10 µg total RNA per 100 µl reaction volume were run for each sample following manufacturer's recommendations. Individual reactions were pooled by sample and used for TaqMan assays analysis. TaqMan assays: each TaqMan Gene Expression Assay consists of two sequence-specific PCR primers and a TaqMan assay–FAM labeled MGB (minor groove binder) probe. Primer and probe design is described in **Supplementary Methods**. Each TaqMan assay was run in four replicates for each RNA sample. 10 ng total cDNA (as total input RNA) in a 10 µl final volume was used for each replicate assay. Assays were run with 2× Universal Master Mix without uracil-*N*-glycosylase on Applied Biosystems 7900 Fast Real-Time PCR System using universal cycling conditions (10 min at

95 °C; 15 s at 95 °C, 1 min 60 °C, 40 cycles). The assays and samples were analyzed across a total of 44–384 well plates. Robotic methods (Biomek FX) were used for plate setup and each sample and assay replicate was tracked on a per well, per plate basis. Data normalization: in QRT-PCR an endogenous control gene is used to normalize data and control for variability between samples as well as plate, instrument and pipetting differences. POLR2A was chosen as the reference gene because its C_T value was within the range of most of the genes in the study and showed the least variation across the samples (**Supplementary Fig. 5a**,**b online**). Each replicate C_T was normalized to the average C_T of POLR2A on a per plate basis by subtracting the average C_T of POLR2A from each replicate to give the ΔC_T which is equivalent to the $\log_2$ difference between endogenous control and target gene. Data analysis and filtering: the ΔC_T of each replicate for each of the 1,000 assays was presented in the final data set as the normalized data. When TaqMan gene expression assays are run on a 7900HT system in a 10 µl reaction volume, a raw C_T value of 34 represents approximately ten transcript molecules (assuming 100% amplification efficiency). At a copy number less than five, stochastic effects dominate and data generated are less reliable. Thus, a raw C_T of 35 was set as the limit of detection in this study: individual replicates which gave C_T values >35 were considered not detected and flagged as not expressed (A, absent); replicates with $C_T < 35$ were considered detectable and identified as expressed (P, present). A $C_T > 32$ and <35 (~5–40 transcript molecules) was considered a low expressing gene. For the ΔC_T calculations we used C_T of 35 for any replicate with $C_T > 35$. Fold-change calculation: the $\log_2$ fold change between two samples was calculated using ΔΔCT method[21]: the average ΔC_T of sample A was subtracted from that of samples B.

StaRT-PCR. StaRT-PCR assays were performed according to the procedures previously described in detail[4,12]. Reverse transcription: for each of the four MAQC samples, two 20 µg aliquots of RNA were reverse transcribed. Each reverse-transcription reaction took place in a 90 µl volume containing Moloney Murine Leukemia Virus (MMLV) reverse transcriptase (1,500 units), MMLV RT 5× first strand buffer (final concentrations 50 mM Tris-HCl, pH 8.3, 75 mM KCl, 3 mM MgCl2) (both from Invitrogen), oligo dT primers (1.5 µg), RNasin (70 units), and deoxynucleotide triphosphates (dNTPs) (10 mM) (all from Promega). Calibration of cDNA: After reverse transcription, the two 90 µl cDNA products for each sample were combined into a single 180 µl volume. Each sample was then calibrated. A 2 µl aliquot of undiluted, tenfold diluted, or 100-fold diluted cDNA from each sample was PCR-amplified in presence of 2 µl of SMIS. In each µl of SMIS there are 600,000 JW molecules of *ACTB* internal standard (IS). It was determined that for each MAQC cDNA sample, a 50-fold dilution would result in approximate equivalence between *ACTB* NT and IS PCR products when equivalent volumes of each were included in the PCR reaction. After 50-fold dilution, there were 4,500 µl of each cDNA sample. It was then confirmed for each sample that the amount of *ACTB* cDNA in 1 µl was approximately in balance with the 600,000 *ACTB* internal standard molecules in 1 µl of SMIS. The amount of RNA that contributed to each µl of each 50-fold diluted working solution was 4 ng. StaRT-PCR reaction conditions: for each StaRT-PCR reaction, a 20 µl reaction volume was prepared containing 2 µl of the calibrated cDNA sample, 2 µl of SMIS, 0.5 units of Taq polymerase, 2.2 µl of buffer, 0.6 ml of $MgCl_2$, 1 µl of each primer, 0.45 µl of dNTPs, and 10.65 µl of water. Range finding step: the expression level of each gene in each sample was initially unknown. Thus, to ensure that each measurement was in range of quantification (NT/IS > 1/10 and < 10/1), a range finding measurement was conducted for each gene in each sample with E SMIS. Each µl of E SMIS, contains 600 molecules of the target gene IS and 600,000 molecules of *ACTB* IS. After PCR amplification and electrophoretic separation of the PCR products, the SEM Center software then determined whether the NT/IS ratio of the PCR products was acceptable or, if not, predicted which SMIS should be used for quantification. This prediction was 95% accurate. Quantification: each 20 µl reaction volume contained 2 µl of the calibrated cDNA sample and 2 µl of the appropriate SMIS (that is, A–F), predicted to be correct in the range finding step. Triplicate measurements were made of each gene in each sample. The fold-change calculation for each gene was based on the ratio of the gene transcript in sample B over sample A.

QuantiGene. Assay procedure: the QuantiGene assays were performed according to the procedure of QuantiGene Reagent System (Panomics), which was previously described in detail[24,25]. Briefly, 10 µl of starting total RNA (500 ng) from sample A, B, C or D was mixed with 40 µl of Lysis Mixture (Panomics), 40 µl of Capture Buffer (Panomics) and 10 µl of target gene-specific probe set (CE (capture extender), 1.65 fmol/µl; LE (label extender), 6.6 fmol/µl; BL (blocker), 3.3 fmol/µl). Each sample mixture was then dispensed into an individual well of a Capture Plate (Panomics). The Capture Plate was sealed with foil tape and incubated at 53 °C for 16–20 h. The hybridization mixture was removed and the wells were washed 3× with 250 µl of wash buffer (0.1× SSC, 0.03% lithium lauryl sulfate). Residual wash buffer was removed by centrifuging the inverted Capture Plate at 1,000*g*. Signals for the bound target mRNA were developed by sequential hybridization with branched DNA (bDNA) amplifier, and alkaline phosphatase-conjugated label probe, at 46 °C for 1 h each. Two washes with wash buffer were used to remove unbound material after each hybridization step. Substrate dioxetane was added to the wells and incubated at 46 °C for 30 min. Luminescence from each well was measured using a Lmax microtiter plate luminometer (Molecular Devices). Three replicate assays measuring RNA directly (independent sampling $n = 3$) were performed for all described experiments. Genomic DNA contamination in the RNA sample, if there is any, does not affect the QuantiGene assay, since it remains doubled-stranded throughout the entire procedure and thus cannot hybridize to the probe sets at the temperature used in the assay. Data analysis and filtering: the QuantiGene assays of 244 genes are performed for MAQC samples A, B, C, D. For all samples, background signals were determined in the absence of RNA samples and subtracted from signals obtained in the presence of RNA samples. Because the QuantiGene assay measures RNA directly, no data normalization against a reference gene is required in the data analysis. The presence and absence call is determined by limit of detection (LOD) of the assay, where LOD = background + 3 s.d. of background. If at least two samples out of A, B, C, D have signals below LOD in a gene, we call the gene absence. To determine gene expression fold change in sample A versus sample B,

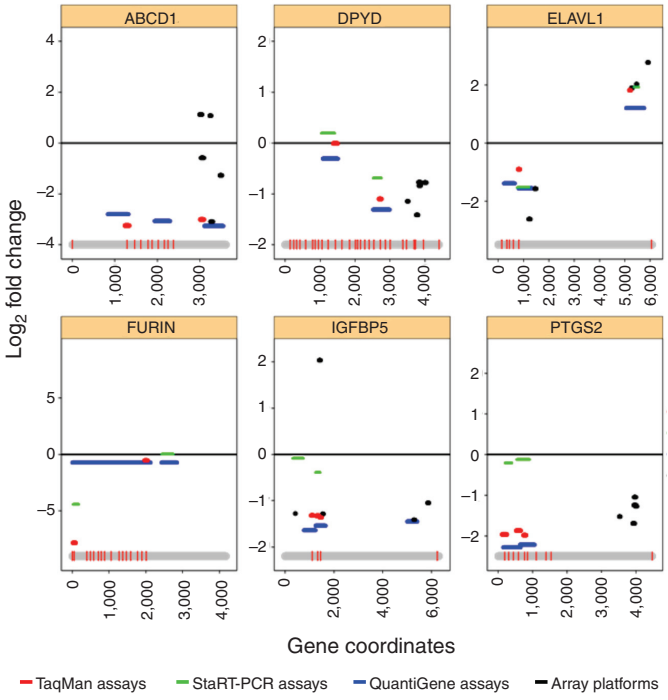

Figure 6 Resolution of fold-change discrepancy results. Fold changes were calculated for Sample B vs. Sample A in all platforms. Each panel shows expression characteristics of a discordant gene across the transcript length. *Y* axis is $\log_2$ fold change. *X* axis represents transcript length starting from the 5′ end of the transcript. Gray bar graphically illustrates the transcript and the red vertical lines represent the exon-exon junctions. Colored bars represent expression value of each probe along the length of the transcript. The length of the colored bar represents the region interrogated by the probe for each platform. Two probes for FURIN (base 1–501, and base 217–2133) produced indistinguishable fold-change value in QuantiGene assay.

we calculated the fold change (fold changes) using formula $\log_2$ fold changes = $\log_2(S_A/S_B)$, where S_A represents the assay range for a target gene in sample A and S_B represents the assay range for the target gene in sample B. A gene is considered for fold-change analysis if the signal in both sample A and sample B passes the LOD. Relative accuracy calculation: relative accuracy measures the proximity of observed expression values for C and D to the predicted values based on measured expression values for A and B. Concentrations of samples A and B were each quantified and normalized on the basis of total RNA (OD_{260}). They were then mixed on a volumetric basis to yield sample C (0.75A/0.25B) and sample D (0.25A/0.75B). If the assay range for the target mRNA is within the linear dynamic range of the assay, then the predicted assay signal for Sample C and Sample D can be calculated using the following formula: $C' = 0.75A + 0.25B$ and $D' = 0.25A + 0.75B$. TaqMan assay and QuantiGene sample input was based on total RNA. For this reason the predicted values of C and D can be calculated from the volumetric proportions of A and B based on the formula $C = 0.25A + 0.75B$ and $D = 0.75A + 0.25B$. With StaRT-PCR, as with the microarrays, each measurement was normalized to mRNA instead of the starting total RNA. As described in[26] and[27], if the fraction of mRNA is higher in sample A compared to sample B, the predicted C and D values will be different from the formula provided above. Based on analysis of optimal linearity among the MAQC samples for the StaRT-PCR data, the most likely formula was determined to be $C = 0.88A + 0.12B$ and $D = 0.45A + 0.55B$. A data set recalibrated on the basis of these assumed formulas (**Supplementary Methods**) was used to assess relative accuracy for StaRT-PCR.

Multi-platform data transformation for Figure 1. For StaRT-PCR, 6,000 transcript molecules were defined by a value of 6,000 or $\log_2$ (6,000) = 12.55. For TaqMan assays, first the C_T values were transformed from a decreasing copy number scale to an increasing copy number scale. This was accomplished by taking the absolute value of the difference of every TaqMan assay C_T value and the lowest value for TaqMan assays C_T (40). This rescaling preserves the assay range measured by TaqMan assays in the $\log_2$ space. Given that a TaqMan assay C_T value of 35 is estimated to correspond to 5 transcript molecules, the extrapolated C_T equivalent for 6,000 transcript molecules is ~24.78. This value on the transformed scale corresponds to |24.78–40| or 15.22. To scale this to the StaRT-PCR value of 6,000 transcript molecules, a rescaling value of 2.66025 was applied to all values. This factor was calculated by taking the difference between the prescaling value in TaqMan assays that corresponds to 6,000 transcript molecules (15.22) and the value of StaRT-PCR that corresponds to 6,000 transcript molecules (12.55). The same transformation was applied to QuantiGene values resulting in a rescaling factor = 13.55. This factor was generated with the estimation of 6,000 transcript molecules defined by 0.5 RLU or –1.0 on a $\log_2$ scale. These transformations result in all platforms having a post-scaling value of 12.55 on a $\log_2$ scale for an approximate threshold of 6,000 transcript molecules.

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

ACKNOWLEDGMENTS

We would like to acknowledge the contribution to this manuscript from the following members of the MAQC team: Shawn B. Baker, Anne Bergstrom Lucas, Jim Collins, Eugene Chudin, Stephanie Fulmer-Smentek, Damir Herman, Richard Shippy, Chunlin Xiao and Necip Mehmet.

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. Vondracek, M. *et al.* Transcript profiling of enzymes involved in detoxification of xenobiotics and reactive oxygen in human normal and simian virus 40 T antigen-immortalized oral keratinocytes. *Int. J. Cancer* **99**, 776–782 (2002).
2. Urdea, M. *et al.* Branched DNA amplification multimers for the sensitive, direct detection of human hepatitis virus. *Nucleic Acids Symp. Ser.* **24**, 197–200 (1991).
3. Gleaves, C.A. *et al.* Multicenter evaluation of the Bayer VERSANT HIV-1 RNA 3.0 assay: analytical and clinical performance. *J. Clin. Virol.* **25**, 205–216 (2002).
4. Bustin, S.A. (ed.). *A-Z of Quantitative PCR*. (International University Line Biotechnology Series, La Jolla, California, USA, 2004).
5. Wong, M.L. & Medrano, J.F. Real-time PCR for mRNA quantitation. *Biotechniques* **39**, 75–85 (2005).
6. Lee, L.G., Connell, C.R. & Bloch, W. Allelic discrimination by nick-translation PCR with fluorogenic probes. *Nucleic Acids Res.* **21**, 3761–3766 (1993).
7. Heid, C.A., Stevens, J., Livak, K.J. & Williams, P.M. Real time quantitative PCR. *Genome Res.* **6**, 986–994 (1996).
8. Gibson, U.E., Heid, C.A. & Williams, P.M. A novel method for real time quantitative RT-PCR. *Genome Res.* **6**, 995–1001 (1996).
9. Qin, L.X. *et al.* Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC Bioinformatics* **7**, 23 (2006).
10. Kuo, W.P. *et al.* A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat. Biotechnol.* **24**, 832–840 (2006).
11. Wang, Y. *et al.* Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. *BMC Genomics* **7**, 59 (2006).
12. Willey, J.C. *et al.* Standardized RT-PCR and the standardized expression measurement center. *Methods Mol. Biol.* **258**, 13–41 (2004).
13. Rots, M.G. *et al.* mRNA expression levels of methotrexate resistance-related proteins in childhood leukemia as determined by a standardized competitive template-based RT-PCR method. *Leukemia* **14**, 2166–2175 (2000).
14. Mullins, D.N. *et al.* CEBPG transcription factor correlates with antioxidant and DNA repair genes in normal bronchial epithelial cells but not in individuals with bronchogenic carcinoma. *BMC Cancer* **5**, 141 (2005).
15. Flagella, M. *et al.* A multiplex branched DNA assay for parallel quantitative gene expression profiling. *Anal. Biochem.* **352**, 50–60 (2006).
16. Yao, J.D. *et al.* Multicenter Evaluation of the VERSANT Hepatitis B Virus DNA 3.0 Assay. *J. Clin. Microbiol.* **42**, 800–806 (2004).
17. Elbeik, T. *et al.* Multicenter Evaluation of the Performance Characteristics of the Bayer VERSANT HCV RNA 3.0 Assay (bDNA). *J. Clin. Microbiol.* **42**, 563–569 (2004).
18. Stenman, J. & Orpana, A. Accuracy in amplification. *Nat. Biotechnol.* **19**, 1011–1012 (2001).
19. Cleveland, W. Robust locally weighted regression and smoothing scatter plots. *J. Am. Stat. Assoc.* **74**, 829–836 (1979).
20. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**, 1151–1161 (2006).
21. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B. Met.* **57**, 289–300 (1995).
22. Shippy, R. *et al.* Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics* **5**, 61 (2004).
23. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT Method. *Methods* **25**, 402–408 (2001).
24. Kern, D. *et al.* An enhanced-sensitivity branched-DNA assay for quantification of human immunodeficiency virus type 1 RNA in plasma. *J. Clin. Microbiol.* **34**, 3196–3202 (1996).
25. Wang, J. *et al.* Regulation of insulin preRNA splicing by glucose. *Proc. Natl Acad. Sci. USA* **94**, 4360–4365 (1997).
26. Shippy, R. *et al.* Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nat. Biotechnol.* **24**, 1123–1131 (2006).
27. Tong, W. *et al.* Evaluation of external RNA controls for the assessment of microarray performance. *Nat. Biotechnol.* **24**, 1132–1139 (2006).