Nucleic Acids Research, 2005, Vol. 33, No. 22 ***e187***
doi:10.1093/nar/gni186

Use of a mixed tissue RNA design for performance assessments on multiple microarray formats

Karol L. Thompson*, Barry A. Rosenzweig, P. Scott Pine, Jacques Retief[1], Yaron Turpaz[1], Cynthia A. Afshari[2], Hisham K. Hamadeh[2], Michael A. Damore[2], Michael Boedigheimer[2], Eric Blomme[3], Rita Ciurlionis[3], Jeffrey F. Waring[3], James C. Fuscoe[4], Richard Paules[5], Charles J. Tucker[5], Thomas Fare[6], Ernest M. Coffey[6], Yudong He[6], Patrick J. Collins[7], Kurt Jarnagin[8], Susan Fujimoto[8], Brigitte Ganter[8], Gretchen Kiser[9], Tamma Kaysser-Kranich[9], Joseph Sina[10] and Frank D. Sistare[10]

Center for Drug Evaluation and Research, US FDA, Silver Spring, MD 20993, USA, [1]Affymetrix Inc., Santa Clara, CA 95051, USA, [2]Amgen Inc., Thousand Oaks, CA 91320, USA, [3]Abbott Laboratories, Abbott Park, IL 60064, USA, [4]National Center for Toxicological Research, US FDA, Jefferson, AR 72079, USA, [5]National Center for Toxicogenomics, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA, [6]Rosetta Inpharmatics LLC, Seattle, WA 98109, USA, [7]Agilent Technologies, Palo Alto, CA 94304, USA, [8]Iconix Pharmaceuticals Inc., Mountain View, CA 94043, USA, [9]GE Healthcare, Chandler, AZ 85248, USA and [10]Merck & Co. Inc., West Point, PA 19486, USA

Received August 15, 2005; Revised October 20, 2005; Accepted November 21, 2005

ABSTRACT

The comparability and reliability of data generated using microarray technology would be enhanced by use of a common set of standards that allow accuracy, reproducibility and dynamic range assessments on multiple formats. We designed and tested a complex biological reagent for performance measurements on three commercial oligonucleotide array formats that differ in probe design and signal measurement methodology. The reagent is a set of two mixtures with different proportions of RNA for each of four rat tissues (brain, liver, kidney and testes). The design provides four known ratio measurements of >200 reference probes, which were chosen for their tissue-selectivity, dynamic range coverage and alignment to the same exemplar transcript sequence across all three platforms. The data generated from testing three biological replicates of the reagent at eight laboratories on three array formats provides a benchmark set for both laboratory and data processing performance assessments. Close agreement with target ratios adjusted for sample complexity was achieved on all platforms and low variance was observed among platforms, replicates and sites. The mixed tissue design produces a reagent with known gene expression changes within a complex sample and can serve as a paradigm for performance standards for microarrays that target other species.

INTRODUCTION

Genome-scale gene expression technologies are increasingly being used for safety and efficacy assessments of pharmaceuticals, in disease diagnosis and in risk assessment of environmental contaminants. However, there is currently concern about the comparability of microarray data (1). Variable results have been reported on data reproducibility across different microarray platform formats and across laboratories in multisite studies (2–8). Factors contributing to discrepant results include differences in probe specificity (8), low technical reproducibility (9) and the use of inappropriate statistical methods. Microarray data comparability can be increased with the use of standard protocols and reagents (6), by careful alignment of probe sequences between platforms (10) and by filtering genes below a noise threshold (11). Other assay

*To whom correspondence should be addressed at US Food and Drug Administration, White Oak Life Sciences Building 64, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA. Tel: +1 301 796 0126; Fax: +1 301 796 9818; Email: karol.thompson@fda.hhs.gov
Present address:
Jacques Retief, Iconix Pharmaceuticals Inc., Mountain View, CA 94043, USA

parameters, like ozone levels, may also need to be tightly controlled (12). More recent publications that take these effects into account report reproducible results between platforms and laboratories (6,8,13). The value of and confidence in genomic data will be greatly increased if there can be reliable comparison and integration of results across experiments, users and platforms.

To ensure the data reliability needed to move microarray technology from the realm of research to an integrated use in regulatory decision making, there is a need to establish 'best practices' in sample generation, sample processing and data analysis (14). There are many options in target labeling, hybridization, washing, signal extraction and data normalization, each of which can affect data accuracy, precision and comparability. As the technology continues to evolve and improve there is a need for a common, objective approach to the assessment and optimization of laboratory and data processing protocols. A critical part in this process is the adoption and use of universal standards for gene expression technologies (15). The currently available quantitative standards for microarrays are limited to sets of control RNA provided by array manufacturers. This RNA is designed to hybridize to specific probes on the array, but not to the queried genomic sequences. This RNA is 'spiked' into samples at different concentrations and ratios to make precision and accuracy estimates. Although different platform formats typically use different sets of control sequences, there is an effort underway to design a common set to be adopted as a standard across platforms (15). Additional standards that are currently available include benchmark datasets that have designed-in changes in gene expression from sample dilution or spiked-in RNA (http://www.affymetrix.com/support/technical/sample_data/datasets.affx). These datasets can be used to compare and improve analytical and statistical methods (16,17).

To generate a fuller understanding of the biological response of an organism, it is important to be able to integrate genomic data from different studies and sources (18). This process requires an assurance of data comparability that could be provided by a reference material (RM) that measures data precision, accuracy and dynamic range on multiple microarray formats. A reference material could also provide the basis for benchmarking laboratory proficiency and assessing the performance of reagents and methods. To provide a performance standard for laboratories engaged in toxicogenomic studies, a collaborative project was initiated to design and test a complex biological RM for use on three commercial rat expression microarrays (http://www.cstl.nist.gov/biotech/UniversalRNAStds;/Thompson.pdf). The arrays used in this study, from Affymetrix Inc., GE Healthcare and Agilent Technologies, are oligonucleotide arrays of different design and that use different measurement methodologies. The Affymetrix GeneChip® RAE230A array is composed of 25mer *in situ* synthesized oligonucleotides organized in sets of 11 pairs of perfect match (PM) and mismatch (MM) probes per gene for ~16 000 rat transcripts. The MM probe has a single base substitution in the middle base of the corresponding PM probe sequence and is used as a measure of local background signal. The GE Healthcare CodeLink® UniSet Rat I array uses a one probe per gene design composed of presynthesized 30mer oligonucleotides for ~10 000 rat transcripts covalently linked to a 3D matrix. The Affymetrix and CodeLink systems both involve one-color sample labeling with one sample hybridized per array. The Agilent G4130A array contains one 60mer *in situ* synthesized oligonucleotide for each of ~20 000 rat transcripts. This array uses a two-color labeling system, with experimental and reference samples hybridized to the same slide.

For highly parallel assays of gene expression, it is not practical to design a RM that can assess the specificity and sensitivity of all probes. However, RMs can potentially be designed for microarrays that have other properties in common with single analyte RMs: measurement of accuracy, precision and linear range (15), close resemblance to the test agent (19), stability and reproducibility across lots, and similar output on multiple array formats. For measurement of assay accuracy, RM analytes need to be present in known quantities. Gene expression is usually evaluated on microarrays relative to a matched control or reference sample, so accurate and precise measurement of real fold change differences between two samples is an important performance characteristic for microarray experiments (20). In this study, a RM for microarrays was designed that has defined fold change differences in transcript levels. It is composed of two samples, each containing different proportions of RNA from different tissues. Gene transcripts that are predominantly expressed in only one of the tissues in the mixture should produce signals directly proportional to the relative amount of that particular tissue RNA in the mixture. The probes that measure these tissue-selective transcripts on the three array platforms in this study are the proposed reference probes for the RM.

MATERIALS AND METHODS

Microarray platforms and collaborating laboratories

Eight sites were involved in testing the mixed tissue RNA reference material (MTRRM); the results are reported anonymized. The samples were run on Affymetrix GeneChip® RAE230A arrays at three sites (sites 1, 2 and 3), on GE Healthcare CodeLink UniSet Rat I arrays at two sites (sites 4 and 5), and on Agilent G4130A arrays at 4 sites (sites 3, 7, 8 and 9). Site 6 used the RM to help calibrate their in-house platform (data not shown). Data from this study are available at EBI ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) under accession number E-TABM-16.

Animal studies

Animal care and procedures were approved by the Institutional Animal Care and Use Committee at the US FDA. Male Sprague-Dawley (SD) rats were ordered from Charles River Laboratories (Product no. Crl:CD(SD)IGS) or from Harlan Laboratories (Product no. Hsd:SD) at 6 weeks of age. Three separate shipments of eight rats were ordered and received for this study within a 2 months time frame. Each shipment was processed and pooled separately to create three biological replicate sets. Shipments 1 and 2 were made from rats ordered from Charles River; Harlan was the source of shipment 3. The rats received certified rodent diet #5002C (Purina Mills Inc.) *ad lib* and drinking water purified by reverse osmosis. The animals were acclimated for 6 days

before euthanasia. The animals were on a 12 h light/dark cycle and euthanasia was performed consistently within 4 to 6 h after the start of the light cycle. The average weight at sacrifice (7 weeks of age) was 223 ± 9 g.

RNA isolation

After euthanasia in a slow-charged CO_2 chamber, the rats were immediately decapitated to allow for rapid access to the brain. Four organs were quickly removed in the following order: brain, liver, kidneys and testes. The whole brain, including brain stem but excluding pituitary gland, was collected. Tissues were quickly dissected into 0.5 cm sections while submersed in RNA*later* (Ambion) in sterile petri plates and placed into 50 ml tubes containing RNA*later* at a ratio of 10 ml per 1g of tissue. Tissues were stored at 4°C for a minimum of 24 h and a maximum of 72 h. All tissue RNA was isolated using a Tempest rotor-stator homogenizer (VirTis) and QIAGEN RNA isolation kits, following the manufacturer's protocol. Brain RNA was isolated using QIAzol reagent, followed by a clean-up step with an RNeasy Maxi kit. Kidney, liver and testes were homogenized in 15 ml QIAGEN RNeasy Lysis buffer (RLT) per mg of tissue, diluted to 30 ml RLT per mg, and RNA was isolated on RNeasy Maxi columns using 15 ml homogenate per column. After an additional clean-up step on RNeasy Maxi columns, RNA was aliquoted and stored at −70°C. The integrity of each RNA sample was assessed on an Agilent 2100 Bioanalyzer (Agilent Technologies). Total RNA was quantitated by ultraviolet (UV)/visible wavelength spectrophotometry in TNE [40 mM Tris–HCl (pH 7.5), 1 mM EDTA (pH 8.0), 150 mM NaCl]. For each tissue, equal amounts of RNA were pooled from each of the eight animals in the same shipment to create tissue shipment specific pools.

SD rat RNA was also tested from two commercial RNA sources. Total RNA isolated from rat brain (Catalog no. 7912), kidney (Catalog no. 7926), liver (Catalog no. 7910) and testicle (Catalog no. 7934) were obtained from Ambion. Total RNA isolated from brain (Catalog no. 737001), kidney (Catalog no. 737007), liver (Catalog no. 737009) and testes (Catalog no. 737023) were obtained from Stratagene. One lot of Stratagene brain RNA (Lot no. 0610696) could not be used to make the MTRRM because it behaved on microarrays like RNA from a tissue different from brain. MTRRM were prepared from the commercial RNA in proportions based on the RNA concentrations provided by the supplier to make MTRRM batches 4 and 5 from Ambion and Stratagene RNA, respectively. An independent batch (batch 6) of the MTRRM was prepared at site 3 under the same protocol used to prepare batches 1–3 from a set of eight rats.

MTRRM batch preparation

After pooling, the RNA was quantitated by measuring OD_{260} on a UV/visible wavelength spectrophotometer in TNE, checked for purity by OD_{260}/OD_{280} ratio, and checked for RNA integrity on the Agilent 2100 bioanalyzer. RNA quantitation was confirmed on a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies). Accurate RNA quantitation in TNE was found to be important for replication of the results with the MTRRM. The average OD_{260}/OD_{280} ratio for each tissue pool across the 3 shipments was the following: brain, 2.09 ± 0.01; kidney, 2.03 ± 0.04; liver, 2.07 ± 0.13 and testis, 2.09 ± 0.03. Each pool of same-tissue RNA from each shipment was run on Affymetrix RAE230A arrays (Affymetrix, Inc.) using the protocols associated with site 1 below. Two mg each of two mixtures (Mix1 and Mix2) were prepared for each shipment from the same-tissue RNA pools to make MTRRM batches 1–3 from shipments 1–3, respectively. Mix1 consisted of 200 μg testis RNA, 600 μg liver RNA, 800 μg brain RNA and 400 μg kidney RNA. Mix2 consisted of 800 μg testis RNA, 400 μg liver RNA, 400 μg brain RNA and 400 μg kidney RNA. A total of 50 μg aliquots of each mixture for each of the 3 batches were frozen at −70°C. The 3 batches of Mix1 and Mix2 RNA samples were run on microarrays at eight anonymized sites. For site 7, 100 μg of Mix1 and Mix2 were treated with deoxyribonuclease I (E.C. 3.1.21.1) (DNA- *free*, Ambion), and diluted to a concentration of 0.2 μg/μl.

Gene expression measurement on Affymetrix RAE230A arrays

Sites 1–3 ran 3 batches of the MTRRM on Affymetrix RAE230A arrays using either the standard or the alternate protocol for labeling and processing specified by the manufacturer (http://www.affymetrix.com/support/technical/manual/expression_manual.affx). Standardized amounts of input total RNA per labeling reaction (5 μg), labeled cRNA target per array (15 μg) and hybridization volume per array (200 μl) were used by the 3 sites. Labeling and processing conditions at sites 1 and 3 included the use of the T7-Oligo(dT) promoter primer kit (Affymetrix Part no. 900375), reagents for cDNA synthesis from Invitrogen, cDNA and cRNA Clean-up using the Sample Clean-up Module (Affymetrix Part no. 900371), and synthesis of biotin-labeled cRNA using an Enzo kit (Affymetrix Part no. 900182). Site 2 used the alternate protocols for cDNA clean-up that includes phenol/chloroform extraction with Phase-Lock gels and for cRNA clean-up that use the RNeasy kit (QIAGEN Part no. 74103). At sites 1 and 2, microarrays were stained and washed on an Affymetrix GeneChip® Fluidics Station 400 using the EukGE-WS2 protocol. The arrays were scanned on an Affymetrix GeneChip® Scanner 2500 using default settings. Site 3 used an Affymetrix GeneChip® Fluidics Station 450 with the EukGE-WS2v4 protocol and an Affymetrix GeneChip® Scanner 3000 with default settings. Affymetrix MAS5 software was used to calculate signal values for tissue-selectivity determinations and for intra-site and cross-site comparisons. All data was globally scaled to a target intensity of 500. For some applications, Mix1 array data were normalized to a selected probe set on the Mix2 array (the 10% trimmed mean of kidney-selective probes (listed in Supplementary Table 1). Gene summary values were also calculated from CEL files using the Probe Logarithmic Intensity Error Estimation (PLIER) algorithm. In contrast to the MAS 5.0 algorithm, which applies a one-step Tukey's biweight estimate to produce a robust weighted mean signal for each probe set the PLIER algorithm uses maximum likelihood type estimates in a model-based framework for finding probe expression estimates. PLIER signal calculations for body map data were performed using default settings (quantile normalization, mismatch background estimation, perfect match minus mismatch and full optimization). An affinity model for the PLIER analysis was constructed from the

four same-tissue RNA pools from each shipment that were run on 12 Affymetrix RAE230A arrays. Normalization between Mix1 and Mix2 was performed either on signals or ratios, where indicated. If PLIER signal estimates were not quantile normalized, Mix1 signal data was normalized by the 10% trimmed mean Mix2 signal of kidney-selective analytes. If PLIER signals were quantile normalized, Mix1:Mix2 ratios were normalized by dividing each ratio by the 10% trimmed mean ratio of the subset of kidney-selective reference probes.

Gene expression measurements on CodeLink UniSet Rat I arrays

At sites 4 and 5, the MTRRM was run on CodeLink RU1 arrays using a standardized amount of input total RNA per labeling reaction (2 μg), labeled cRNA target per array (10 μg) and hybridization volume per array (250 μl) (http://www5.amershambiosciences.com/aptrix/upp01077.nsf/Content/codelink_user_protocols). At site 4, target labeling was performed using the manufacturer's manual labeling cDNA target preparation protocol; site 5 used the manufacturer's automated target preparation protocol. Site 4 used the manufacturer's recommended hybridization and detection protocols. A few modifications to this protocol were made at site 5 (21). As a secondary label, site 4 used Cy5-Streptavidin and site 5 used Alexa fluor 647-Streptavidin. Both sites used the Axon 4000B scanners at settings defined in the user manual. Version 2.3 of the CodeLink Expression analysis software was used for feature extraction. A global normalization of each array by the median normalized intensity was performed. For some applications, this step was followed by normalization of Mix1 to Mix2 using the 10% trimmed mean signal of the kidney-selective probe subset. Alternatively, Mix1:Mix2 ratios were normalized by dividing each ratio by the 10% trimmed mean ratio of the subset of kidney-selective reference probes.

Gene expression measurements on Agilent G4130A arrays

The four sites running the MTRRM on Agilent arrays used either standardized protocols for sample labeling and hybridization or a propriety in-house method. Sites 3, 8 and 9 used the Agilent Low Input RNA Fluorescent Linear Amplification Kit (Part no. 5184-3523) for target labeling and the Agilent 60mer microarray processing protocol version 2.0 (Part no.G4140-90030). All sites used the Agilent DNA microarray scanner (Part no. G2565BA). Site 7 used DNase-treated MTRRM and proprietary protocols for target labeling, hybridization and washing. Mix1 and Mix2 samples were run on the same array, in dye-swap replicate experiments. For this phase of the project, the data was extracted and processed using a standard method. The TIFF images from all four sites were processed with Agilent Feature Extraction software version A.7.4.47 (a prerelease version that is algorithmically identical to v 7.5.1) using default settings. Adjustment for local variations in background signal was performed using a spatial detrending algorithm. To normalize and correct for dye bias, a combined method of linear scaling in each channel followed by LOWESS curve fitting ('Linear&LOWESS' option in version A.7.5) was used. Mix1 and Mix2 signals were calculated from the average of dye-swap replicates. For some applications, this step was followed by normalization of Mix1 to Mix2 using the 10% trimmed mean signal of the kidney-selective probe subset. Alternatively, Mix1: Mix2 ratios were normalized by dividing each ratio by the 10% trimmed mean ratio of the subset of kidney-selective reference probes. Features flagged as outliers by the feature extraction software were not removed from the analysis for this study.

Platform intersection

To identify the probes on three commercial rat expression array platforms (Affymetrix RAE230A, Agilent G4130A and CodeLink UniSet Rat I) that were potential reporters of expression levels for the same gene transcripts, probe annotation data were intersected by GenBank accession number and/or UniGene identifier using the annotation files supplied by the manufacturer that were available in June 2003. Approximately, 6300 probes were identified that could be intersected by annotation on all three platforms. This number includes duplicate listings when a probe on one platform could be linked to multiple probes on a second platform.

Tissue-selectivity index

Tissue-selectivity was determined using body map data, i.e. signal values averaged across multiple control animal samples for the individual tissues in the MTRRM. For each probe on each of the three platforms, a tissue-selective index (TSI) was determined as follows: the average signal value in a selected tissue was divided by the maximum average signal value among the other three non-selected tissues.

Body map data on Affymetrix RAE230A arrays was generated from the pooled tissue samples that are components of the MTRRM, as described above. Each sample was composed of RNA pooled from brain, kidney, liver or testes samples across eight male SD rats that were in the same shipment cohort. Using these samples from three biological replicate experiments, an average signal value was determined for each probe in each of the four tissues.

Body map data on CodeLink UniSet Rat I arrays was derived from individual control animal data from vehicle-treated male SD rats (vehicle not specified) provided by Iconix Pharmaceuticals. Data was excluded for probes which showed identified associations with process drift due to array protocol changes over time. An average signal value for each of 8565 probes in brain, kidney and liver RNA was calculated across 25 control animal datasets. Six control animal datasets were available to calculate an average signal in testes RNA.

Body map data on Agilent G4130A arrays was derived from individual control animal data generated through a collaboration between NIEHS and Iconix. Brain, kidney, liver and testis RNA samples from three SD rats, that received a 0.5% carboxymethyl cellulose (CMC) vehicle treatment for 5 days, were run on Agilent G4130A arrays. Each tissue sample was run once on each of the Cy3 and Cy5 channels in a dye-swap with the Iconix Reference RNA on the other channel. The Iconix Reference RNA is a pooled RNA extracted from an equal tissue mixture of 7 rat tissues taken from 10 male SD rats, vehicle-treated with 0.5% CMC for 3 days. The signal channel corresponding to the control tissue sample was separated from the reference sample channel for each dye-swap pair, resulting in one Cy3 and one Cy5 signal value for each

probe for each of three control animal samples per tissue. An average signal value was calculated for each probe in each tissue from these six signal values.

Relative signal intensity in the MTRRM

Using data generated by participating labs, an average signal value in the MTRRM was calculated for each probe on each of the three microarray platforms. For each site, the average signal value across three replicate sets of Mix1 and Mix2 experiments was determined for each probe, expressed as a percent of the average maximal signal (%Max) in the same experimental set, and then averaged across all sites using the same microarray platform in the study ($n = 3$ for RAE230A, $n = 2$ for RU1, $n = 4$ for G4130A). The average %Max was used to sort probes into nine exponentially distributed bins as follows: <0.4; 0.4–0.8; 0.8–1.6; 1.6–3.2; 3.2–6.4; 6.4–12.8; 12.8–25.6; 25.6–51.2 and >51.2%Max, respectively.

MTRRM reference probe selection

For probes that could be linked across three platforms by annotation, tissue-selectivity and relative signal intensity were weighed together to derive a list of candidate probes for the MTRRM. Probes were first sorted into nine exponentially spaced bins based upon their %Max on the Affymetrix platform. From each bin, 5–8 probes with the highest combined TSI values for each platform were chosen in order to select ~200 probes in total (~50 per tissue). The selected probes were then binned according to their %Max on the Agilent and CodeLink arrays and reselected, if necessary, to achieve a similar distribution on each platform. Probes for tissue-selective gene transcripts that did not receive a MAS5.0 present call on all of the selective tissue samples run on Affymetrix RAE230A for this study were not chosen for the analyte list. Testes-selective genes with signal intensities <0.8%MAX were also excluded because, in this intensity range, the contribution of non-selective signal to selective signal greatly attenuated the observed ratio for these probes.

To confirm that the candidate MTRRM analytes measured the same gene transcripts on each platform, probes were sequenced mapped to a common exemplar. Using annotations from UniGene Build 135, probes were aligned against the corresponding NCBI Reference Sequence database (RefSeq) sequence (22) or, if not available, a common mRNA or EST sequence. For a few exemplars that were not RefSeqs, probes aligned to the reverse complemented strand of the GenBank sequence. Cross-platform intersected probes that could not be mapped to a common exemplar sequence were filtered from the list. For one of the analytes, a single exemplar sequence was not found that contained the probe sequences for all three platforms, so two overlapping exemplars are listed in Supplementary Table 1. Gene symbol and RefSeq status were updated for all exemplars using the information available in the NCBI public databases in November 2005.

In silico modeling of microarray ratio measurements

An average signal intensity (I) for each probe in each of the four tissues was calculated from body map data available for each platform and used to calculate a modeled ratio (R) for each analyte based upon tissue RNA proportions in Mix1 and Mix2 using the following formula:

$$R_{\text{Analyte}} = \frac{I_{\text{Mix1}}}{I_{\text{Mix2}}} = \frac{I_{\text{Brain}}(0.4) + I_{\text{Liver}}(0.3) + I_{\text{Kidney}}(0.2) + I_{\text{Testis}}(0.1)}{I_{\text{Brain}}(0.2) + I_{\text{Liver}}(0.2) + I_{\text{Kidney}}(0.2) + I_{\text{Testis}}(0.4)}$$

An average ratio for each set of tissue-selective analytes was calculated from 46–55 individual R_{Analytes}.

Reverse transcription polymerase chain reaction (qRT–PCR)

Relative gene transcript levels between Mix1 and Mix2 were determined using qRT–PCR. cDNA was generated from total RNA using random hexamer primers and Superscript II reverse transcriptase (Invitrogen). qRT–PCR was performed using SYBR Green PCR Master Mix reagents (Applied Biosystems) on the ABI Prism 7900HT sequence detection system as described in User Bulletin #2 (updated 10/2001). Seven 2-fold serial dilutions were used to prepare relative standard curves for each of the targets and their endogenous reference (18s rRNA). Gene expression data were normalized by dividing the amount of target mRNA by the endogenous reference. Relative changes in gene expression were calculated by dividing the amount of target mRNA in Mix1 by the amount in Mix2.

High-performance liquid chromatography (HPLC)-purified oligonucleotide primers were obtained from BioServe Biotechnologies. The UniGene name, symbol, sequence accession number and primer sequences for qRT–PCR for each target transcript are provided below.

The three brain-selective targets were chromogranin B (*Chgb*, NM_012526.1, forward primer: GGAAAAGTTCAGCCAGCGG, reverse primer: CAGCGAATGGCTCGTCTCTC), neurofilament 3, medium (*Nef3*, NM_017029.1, forward primer: TGTACCTAGGGAATTTGCCAGTTT, reverse primer: CGAGTGCCCCTCTTTCAACA), and neurofilament, light polypeptide (*Nfl*, NM_031783, forward primer: GACCTCCTCAATGTCAAGATGG, reverse primer: TCGCCTTCCAAGAGTTTCCT). The kidney-selective targets were kidney-specific membrane protein (*Tmem27*, NM_020976.1, forward primer: GAAATTTCCCACGTCCTGCTTT, reverse primer: GCACTGTTGATCCGTTTCCTGT) and trefoil factor 3 (*Tff3*, NM_013042.1, forward primer: AGCTCCACACCCTGGACTCTT, reverse primer: TGAGTGTTACCCTGGGCCAC). The two liver-selective targets were hepatic lipase (*Lipc*, NM_012597, forward primer: GCTCCCATCCACTTGTCATGA, reverse primer: TTTCTAGCAAGCCATCCACCG) and complement component 9 (*C9*, NM_057146.1, forward primer: CATGTCAAAACGGAGGCACA, reverse primer: TGCACTGTTGATCCGTTTCCT). The three testis-selective targets were phosphorylase kinase, gamma 2 (*Phkg2*, M73808, forward primer: AACTGTGCCTTCCGGCTCTA, reverse primer: CTGCTGCTCCCCCTTCTTC), A kinase anchor protein 4 (*Akap4*, NM_024402, forward primer: AAACAAGACCAGCCTAAGACGG, reverse primer: GAGGAGCCAGTTGAGGACACTT), and ATPase, Na^+/K^+ transporting, alpha 4 polypeptide (*Atp1a4*, NM_022848, forward primer: TGGATGAGCTGAGTGCCAAGT, reverse primer: CGTCTGTGACGCTAAGACCCTT). *Nfl*, *Lipc*, *Akap4* and

Atp1a4 were not selected to be on the final cross-platform list of MTRRM analytes.

Statistical analysis

Two ANOVA models were applied to identify the major sources of variability within the MTRRM data collected on three platforms from eight sites using three biological replicate batches. A two-way ANOVA, using S-PLUS®, was used to study the tissue and gene effects as well as their interactions. A mixed-model three-way ANOVA was performed using Partek® software (Copyright, Partek Inc.). Partek and all other Partek Inc. product or service names are registered trademarks or trademarks of Partek Inc. This model was used to determine the contribution to variation by the platform, batch and site effects. The input data for these models are the batch-specific ratios for 199 tissue-selective analytes without the Mix1:Mix2 normalization step. Affymetrix signal estimates were calculated using PLIER. Four probes were excluded from the analysis because they were masked as suboptimal probes in the Manufacturing Slide Report files for the lots of CodeLink arrays used in this study.

One sample *t*-tests comparing replicate Mix1:Mix2 ratios (for single analytes or across sets of analytes) to a theoretical mean were performed using GraphPad Prism version 3.03 for Windows (GraphPad Software, San Diego, CA). Two sample *t*-tests were performed in Excel and Partek software. ANOVA within body map data was performed using Partek Genomics Solutions software version 6.1. Multiple comparison corrections were performed using the Benjamini-Hochberg False Discovery Rate procedure within Partek software.

RESULTS

Design of a rat mixed tissue RNA reference material

The use of a mixed tissue RNA design for a RM depends on identification of tissues with highly dissimilar gene expression profiles. Large uniformly-populated in-house databases containing samples from many different rat tissues were mined to identify tissue pairings with the least similar gene expression profiles. In an analysis of expression data from 10 tissues harvested from vehicle-treated male SD rats and run on CodeLink UniSet Rat I arrays, it was determined that liver and brain was the most divergent tissue pair. Brain, liver, testis, intestine and bone marrow were identified as providing an optimal representation of low, medium and highly expressed genes with diverse expression ratios. Similar results were seen in a second analysis that involved single channel data from five tissues (liver, brain, muscle, kidney and heart) from untreated control rats that had been run on Agilent IJS oligonucleotide arrays. The tissue pairs that had the greatest number of genes differentially expressed at or above a statistical confidence level were brain and either heart, liver, kidney or muscle. The next most divergent pairings were of kidney and either heart, liver or muscle. Similar results were found in other species (mouse, human and dog). The tissue composition of the RM was based on the tissue dissimilarity analysis, relevance to toxicology and reproducibility and quantity of source material. Based on these criteria, liver, kidney, brain and testis were selected to be the four rat tissue RNA components of the RM.

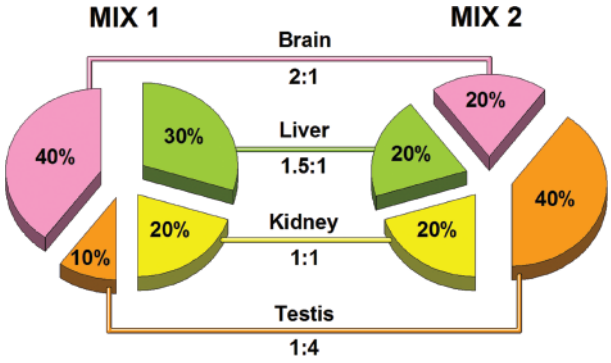

Figure 1. MTRRM Composition. The relative input proportions of total RNA from four rat tissues are shown for Mix1 and Mix2.

A mixed tissue RNA reference material was designed to consist of two mixtures (Mix1 and Mix2) of four tissues in four different proportions (Figure 1). Mix1 contains total RNA from four tissues in the proportions of 40% brain, 30% liver, 20% kidney and 10% testes. Mix2 contains 40% testes RNA and 20% each of brain, kidney and liver RNA. The tissue ratios between the two mixtures were designed to provide potential measurement of 1-, 1.5-, 2- and 4-fold changes. The 2- or 1.5-fold changes are common threshold fold change cutoffs for the prioritization of significantly changed genes on microarrays. A larger fold change ratio can potentially be a more sensitive indicator of data compression or signal saturation. In addition, components designed to be at a 1:1 ratio between Mix1 and Mix2 could be used for normalization between arrays. The proportions of tissue RNA in each mixture that would produce these ratios and minimize the dilution of any one tissue RNA were determined empirically and by *in silico* modeling. A limitation of this design is that accuracy of measurement of tissue-selective analyte ratios will be affected by contributing signals from non-selective tissue RNA in the RM that could arise from either gene-specific or cross-hybridizing transcripts, or from background. These additions to the selective signal may constitute a significant proportion of the signal at low expression levels.

Additionally, the protocol for MTRRM formulation was designed to prepare a complex RNA mixture that will be relatively invariant in composition when remade over time and in different laboratories. Each tissue RNA in the MTRRM was prepared from a whole-organ homogenate and pooled from eight male SD rats. Three biological replicate sets of mixtures (batches) were made from eight male SD rats that had been received in three different shipments from two different suppliers. The age of the animals when received and at sacrifice was kept constant, as was the time and method of sacrifice. To maximize recovery, brain RNA was prepared by a method designed for recovery of RNA from tissue of high lipid content. Further testing is needed to determine how deviations from this protocol affect the properties of the RM.

MTRRM analytes will be limited to transcripts that can be measured on multiple array formats and that produce signals proportional to the tissue RNA input. To select these analytes, the first step was to identify the probes on Affymetrix RAE230A, Agilent G4130A and CodeLink UniSet Rat I arrays that were potential reporters of expression levels for the same gene transcripts. Approximately 6300 probes were identified

that could be intersected by GenBank accession number and/or UniGene identifier across the three platforms.

Probes were next identified for gene transcripts that behaved as 'tissue-selective' on each platform, i.e. had an average expression level that was significantly higher in a given tissue than in the other tissues in the MTRRM. Body map data, defined as microarray data generated from multiple examples of brain, kidney, liver or testes RNA samples from control male SD rats, were collected on each platform. Using signal values averaged across multiple control animal samples, a tissue-selectivity index was calculated for each probe on all three platforms for each of the four tissues in the RM. For example, a brain TSI is calculated for a given probe from its average signal in control brain RNA samples divided by its highest average signal among control kidney, liver or testis RNA samples. About 2900 probes, matched by annotation across all three platforms, were minimally selective for the same-tissue (TSI >1) on each of the three platforms.

Defining the limits of measurement linearity is a recommended performance evaluation for the use of multiplex assays for diagnostic applications (http://www.fda.gov/cdrh/oivd/guidance/1210.pdf). Guidelines for linearity evaluations of quantitative measurement procedures in clinical laboratories recommend at least five different concentration measurements be run in duplicate (19). On microarrays, expression level measurements typically span a 2–3 $\log_{10}$ range (23,24), but the reliable range of linear signal measurement can be limited by background and signal saturation. Based on expression level and combined TSI, ~50 analytes per tissue were chosen (203 total) that produced signals in the MTRRM that spanned the dynamic range and were reliably tissue-selective on three platform formats (see Materials and Methods). As a final step, oligonucleotide probe sequences for each analyte, which were originally intersected across platforms by UniGene or GenBank identifier, were mapped to a consensus gene transcript. About 90% of the identified exemplar sequences were annotated as RefSeq sequences. Sequence mapping of probes should provide a cross-platform intersection key that is more stable over time than one matched solely by UniGene cluster assignments, which are subject to change. The probes selected for measurement of tissue-selective analytes within the MTRRM and their corresponding exemplar transcript sequences are listed in Supplementary Table 1. For each reference probe, the TSI and relative signal intensity on each of the three rat expression array formats are reported in Supplementary Table 2. A statistical measure of tissue-selectivity was generated by ANOVA comparison of multiple examples of control expression levels in brain, kidney, liver and testes, followed by multiple comparison correction. For most of the reference probes, a statistically significant difference in signal between the four tissues was observed. The few exceptions tended to be probes with signals in the lowest intensity bin, which were included for dynamic range determinations.

Performance characteristics of a rat MTRRM

The rat MTRRM described above was designed to allow measurement of precision, accuracy and linear range as a function of signal strength on three different commercial rat expression arrays. To help assess whether these criteria were met, three biological replicate sets of the MTRRM (batches 1–3) were run on Affymetrix RAE230A, Agilent G4130A or CodeLink RU1 arrays. The study was performed at eight different laboratories highly experienced in running microarrays in order to reduce variability due to different levels of laboratory proficiency. The generated datasets allowed for comparisons of reproducibility between three sites on the Affymetrix array, two sites on the CodeLink array, and four sites on the Agilent array. To limit the variables within this study, certain array and data processing steps were standardized within platforms (see Materials and Methods).

Coverage of each platform's dynamic range of measurement by the MTRRM analytes was visualized in $\log_2$-scaled scatter plots. Representative data generated across three batch replicates of the MTRRM on Affymetrix RAE230A, Agilent G4130A and CodeLink RU1 arrays are shown in Figure 2. The average signal in Mix1 versus the average signal in Mix2 is graphed for each set of tissue-selective probes and overlaid upon the signal distribution from all probes on each array format. Expression measurements were calculated from the Affymetrix array data using both Microarray Suite 5.0 (MAS5) and PLIER algorithms in order to assess performance of the MTRRM using the manufacturer's standard analysis method and a newer method with improved signal estimates at lower intensities (Figure 2A and B).

The MTRRM analytes spanned the reliable measurement range on all three platforms and allowed for replicate measurements across the dynamic range. Different methods are used to assess the lower limits of signal reliability on each platform. On the Affymetrix platform, reliable measurements are indicated by detection calls, which are assigned by applying a statistical algorithm to PM and MM probe pairs (http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf). On CodeLink and Agilent arrays, signal significance is typically calculated in relation to local background (11,25). Signals consistently detected as absent or not significantly above background on all 6 of the arrays in each of the analysis sets graphed in Figure 2 are shown as empty circles in the scatter plots. The upper end of reliable signal measurement can be indicated by data compression or signal saturation. This effect, most noticeable in the testis-selective probe set on the Affymetrix platform, has been attributed to saturation of binding affinity rather than optical saturation in the high intensity signal range (26). Although some feature extraction software programs remove saturated features from the processed data reports for the purpose of increasing data reproducibility (http://lifesciences.chem.agilent.com/scripts/LiteraturePDF.asp?iWHID=31889), outliers were not removed from any dataset in this study during post-processing to allow comparable assessment of performance of the MTRRM across platforms.

qRT–PCR was used to verify the mixture proportions by measuring the ratios of several tissue-selective transcripts within the MTRRM. Two or more targets per tissue were selected. All targets had median signal intensities and observed signal ratios near the average for their set based on microarray data from each of the three different platforms (see Supplementary Table 3). The nominal input ratios were highly similar to ratios that were measured by qRT–PCR and averaged across each set of tissue-selective targets (Table 1). For the six tissue-selective targets in Table 1 that were

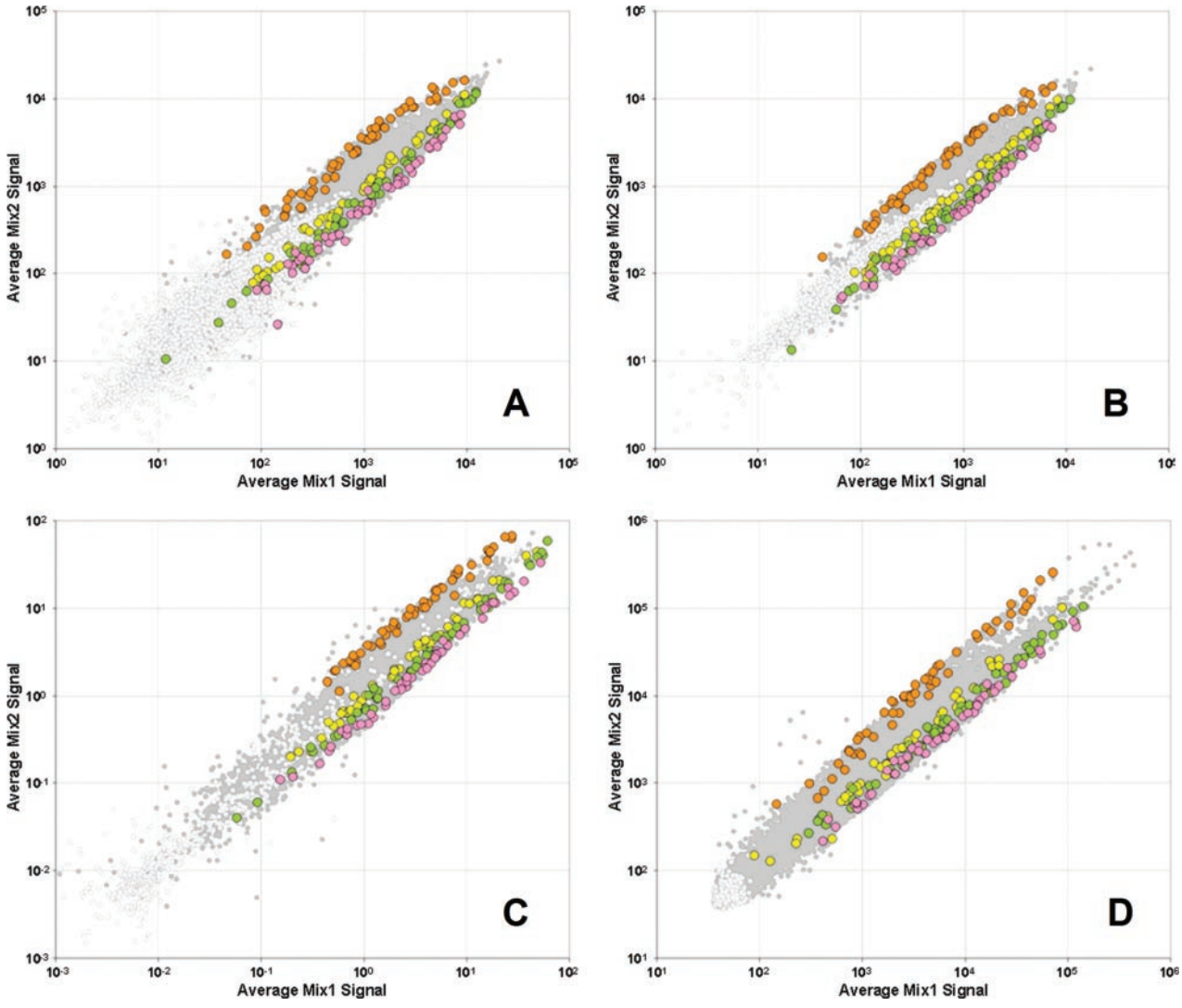

Figure 2. Scatter plots of MTRRM analyte signals on three platform formats. Signal intensities from brain-selective (pink), kidney-selective (yellow), liver-selective (green), testis-selective (orange) and non-selective probes (grey) were averaged across three biological replicate experiments. Non-selective probe signals are designated as 'absent' (empty) if they were consistently assigned a call of 'Absent' on Affymetrix, of 'Empty' on CodeLink, or were not flagged as 'Well Above Background' on Agilent arrays; probe signals not 'absent' by these criteria are designated as 'present' (solid). MTRRM samples were run on Affymetrix RAE230A arrays at site 1 and analyzed using MAS5 (**A**) or PLIER (**B**) algorithms. MTRRM samples were run on CodeLink UniSet Rat I arrays at site 5 (**C**) or on Agilent G4130A arrays at site 9 (**D**).

Table 1. Measurement of accuracy of MTRRM analyte ratios by qRT–PCR

Tissue	Input ratio	qRT–PCR target	qRT–PCR ratio	qRT–PCR tissue average
Brain	2.0	*Chbg*	2.02 ± 0.16 (6)	2.07
		Nef3	2.10 ± 0.21 (2)	
		Nfl	2.09 ± 0.12 (3)	
Kidney	1.0	*Tmem27*	0.97 ± 0.08 (3)	1.05
		Tff3	1.13 ± 0.01 (2)	
Liver	1.5	*Lipc*	1.54 ± 0.01 (3)	1.51
		C9	1.47 ± 0.01 (2)	
Testis	0.25	*Phkg2*	0.27 ± 0.01 (2)	0.27
		Akap4	0.26 ± 0.01 (3)	
		Atp1a4	0.27 ± 0.01 (3)	

qRT–PCR ratios was calculated relative to 18s rRNA levels. The average ratio and SD (if $n > 2$) or range (if $n = 2$) were calculated across the number of replicate experiments indicated in parentheses. The RNA source is MTRRM batch 1.

analyzed in three or more qRT–PCR assays, the mean Mix1: Mix2 ratios were not significantly different from the corresponding input ratios in a one sample *t*-test ($P > 0.01$). These results indicated that a design based on mixing total RNA from different tissues could produce a complex mixture in which tissue-selective components could be individually measured at the same ratios as the overall input proportions.

When measured on microarrays, some degree of attenuation of the ratio of tissue-selective analytes in the MTRRM might be expected from cross-hybridizing transcripts present in the other tissues in the MTRRM. An estimate of the tissue-selective analyte signal in the mixed tissue RM that would be measured on microarrays was modeled for each platform from body map data (Table 2). The model is based on the assumption that the signal of a transcript measured in a complex mixture on a microarray will be additive of each

Table 2. Tissue-selective ratios modeled for non-selective tissue signal contributions using body map data for each platform

Tissue	Input ratio	RAE230A ratio (MAS5.0)	RAE230A ratio (PLIER)	Rat UniSet I ratio	G4130A ratio
Brain	2.0	1.78 ± 0.14	1.73 ± 0.24	1.82 ± 0.13	1.70 ± 0.17
Liver	1.5	1.45 ± 0.07	1.45 ± 0.09	1.45 ± 0.08	1.43 ± 0.08
Kidney	1.0	1.01 ± 0.07	1.02 ± 0.05	1.02 ± 0.09	1.00 ± 0.09
Testis	0.25	0.34 ± 0.09	0.34 ± 0.11	0.33 ± 0.11	0.37 ± 0.10

Ratios shown are the predicted ratios of tissue-selective gene expression in the MTRRM that would be measured on microarrays based on modeling of individual tissue RNA microarray data. Signal estimates were derived for Affymetrix body map data using either MAS5.0 or PLIER algorithms. A SD was calculated across the ~50 $R_{Analyte}$ values for each tissue-selective set.

Table 3. Observed ratio measurements of tissue-selective analytes by microarray platform

Tissue	Modeled ratio	RAE230A ratio	Rat UniSet I ratio	G4130A ratio
Brain	1.75	1.80 ± 0.04	1.81 ± 0.10	1.73 ± 0.02
Liver	1.44	1.40 ± 0.06	1.41 ± 0.06	1.44 ± 0.01
Kidney	1.01	1.01 ± 0.03	1.04 ± 0.01	1.03 ± 0.01
Testis	0.35	0.35 ± 0.01	0.35 ± 0.02	0.34 ± 0.002

The modeled ratio was averaged across three platform formats (from Table 2). The observed ratios are the 10% trimmed mean fold change and SD across three replicate experiments for each set of MTRRM analytes for a representative laboratory per platform (sites 1, 5 and 9 for Affymetrix, CodeLink and Agilent arrays, respectively). Mix1 signals were normalized to the 10% trimmed mean of the kidney-selective analyte signals in Mix2. Affymetrix signal estimates were calculated using MAS5.0.

individual component of the complex mixture, adjusted for its proportion. The modeled ratios for tissue-selective analyte sets were similar across platforms and across two different methods of signal calculation on Affymetrix arrays. The kidney-selective ratio did not differ from the input ratio because, in the model, equivalent amounts of non-kidney-selective signal are added to the kidney-selective signal in both Mix1 and Mix2. The brain-selective signal is compressed from 2-fold to between 1.7–1.8-fold in the model. Most significantly, the modeled testis-selective signal is attenuated from a 1:4 input ratio to a little less than 1:3.

The accuracy of measurement of the ratio of tissue-selective analytes within the MTRRM was next assessed on microarrays. The MTRRM Mix1 and Mix2 samples were designed to be significantly different in three of the four components and are therefore more complex than comparison sets in typical microarray experiments. An additional normalization step was therefore applied to Mix1 that is based on the trimmed mean kidney-selective analyte signal in Mix2, because this tissue RNA was designed to be equivalent between the two mixes. The fold change between Mix1 and Mix2 for each of the four components of the MTRRM were averaged across each set of 46–55 tissue-selective analytes and shown for a representative dataset per platform in Table 3. To reduce the effect of outliers on measurement precision, the average ratios were calculated as the 10% trimmed mean for each tissue-selective analyte set. On each platform, the average ratio measurements were highly similar to the modeled ratio and highly reproducible across batches. Using a one sample *t*-test that compares a population mean to a theoretical mean, the mean ratios of brain- and testes-selective analytes measured on microarrays were significantly different from input ratios ($P < 0.001$), but not from modeled ratios (for the datasets in Table 3). The mean ratios of the liver-selective analytes measured on the three array formats were statistically different from input ratios but not from modeled ratios at a *P*-value threshold of 0.05. Therefore, a more accurate measurement of the designed-in ratios was achieved if the target ratio was adjusted as described above for signal contributions from the four tissue components in the complex mixture. No significant difference was observed in the precision of signal measurements calculated using two different algorithms for Affymetrix signal estimates, MAS5 and PLIER.

The observed ratios for each reference probe, averaged across three replicates and generated at a representative site per platform, are shown in Supplementary Table 3. The probability of detecting a difference in signal between Mix1 and Mix2 was also calculated. In general, most of the reference probes that are selective for the two tissues with designed-in fold change differences $\geqslant 2$ (i.e. brain and testes) were observed to have statistically significant differences in expression at a false discovery rate of 0.05 on all three platforms. Most kidney-selective reference probes were not statistically different between Mix1 and Mix2, which is the anticipated result for a designed target ratio of 1:1. Smaller proportions (50–67%) of liver-selective analytes were observed to be statistically different between Mix1 and Mix2. The 1.5-fold change designed for this set may be near the lower limit that can be distinguished as statistically different under the conditions of this study.

In this study, three independent batches of the MTRRM were run on three different platforms at a total of eight sites. To determine which factors in the study contributed the most variability, two ANOVA-based statistical analyses were applied to Mix1:Mix2 signal ratios for all datasets in this study. The combined results of both ANOVA models are shown in Figure 3. The mean of mean square (graphed on the *y*-axis on a $\log_{10}$ scale) is a measurement of the contribution of each factor on the *x*-axis to the variability in the experiment. A two-way ANOVA showed that the tissue is the major source of variability, which confirmed the expected specificity of results with this mixed tissue RNA design. The next largest source of variation is the interaction between gene and tissue, which is a measure of how well the gene transcripts selected as MTRRM analytes characterize the tissue response. The high mean of mean square value (7.88) for this factor indicates good tissue characterization by the analytes. The actual interaction *P*-value given by the ANOVA model for each gene transcript can be used to rank the performance of each analyte as a tissue representative across all platforms. The gene effect as an independent source of variation is very low (0.269), which indicates that the selected transcripts show a consistent response across all samples and across all platforms. The mixed-model three-way ANOVA showed a very low platform effect (0.143) confirming the results in Table 2 that the ratios of tissue-selective analytes are very similar across platforms. The 3 batches in the experiment are highly reproducible as indicated by a very low batch effect (0.09). The platform-batch interaction effect is even lower (0.082), which indicates that the batch effect of all the combined platform ratios is not platform dependent. Sites, which are nested in platforms, show the

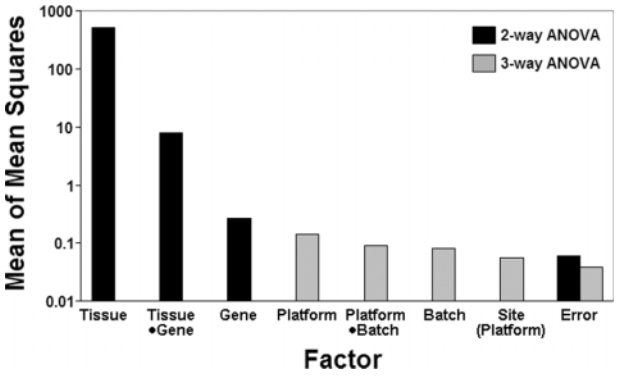

Figure 3. Source of variance in the MTRRM project data. Two ANOVA models were applied to identify the major sources of variability within the MTRRM data collected on three platforms from eight sites using three biological replicate batches of the mixed tissue RNA. A two-way ANOVA model (black bars) was used to study the tissue and gene effects as well as their interactions. A mixed-model three-way ANOVA (grey bars) was applied to determine the contribution to variance by the platform, RNA batch and site effects. The input data for these models are the batch-specific Mix1:Mix2 ratios of the MTRRM tissue-selective analytes. The mean of mean squares plotted on the *y*-axis on a $\log_{10}$ scale is a measure of the contribution of each factor to the variability in the experiment.

lowest effect (0.056), indicating very high reproducibility between laboratories. The very low mean of mean square (0.058 and 0.038) of the residuals (labeled as error in Figure 3) for both models is an indication that the applied models fit well and most of the variability in the experiment can be explained by the tested factors.

For the initial studies on the MTRRM, all RNA was prepared at one site to avoid potential variation in data due to differences in RNA isolation procedure and yield. To determine if the performance of the MTRRM would be affected by RNA source and isolation method, the MTRRM was prepared from two commercial sources of RNA (batches 4 and 5) and at an independent site (batch 6) that used the same RNA isolation and pooling protocols developed for batches 1–3. The MTRRM analyte set should be fairly robust to RNA isolation protocol because the selection process was based in part on body map data generated from different RNA sources. Rat tissue RNA was purchased from two commercial suppliers, mixed-in the specified proportions for the MTRRM [verified by qRT–PCR (data not shown)], and run on Affymetrix RAE230A arrays at site 1. As a benchmark set, an average $\log_2$ ratio was calculated for each of the 203 analytes across batches 1–3 using signal estimates derived by the PLIER algorithm with data generated at site 1. Each of batches 1, 2 and 3 had a tight correlation with the batch average [Pearson correlation coefficient (ρ) $\geqslant$0.99]. Batches 5 and 6 were also highly similar to the batch 1–3 average ($\rho = 0.98$). Of the RNA sources tested, batch 4 had the lowest correlation ($\rho = 0.95$) with batches 1–3, although the microarray QC metrics for batch 4 were comparable with the other batches tested. To assess the stability of the MTRRM after one year of storage at -70°C, batch 1 was retested at site 1 twelve months after its initial assessment. Upon retesting, this batch had a high correlation ($\rho = 0.98$) with batches 1–3 run a year earlier, indicating that the prepared RNA batches of reference material are stable to prolonged storage.

DISCUSSION

This paper describes the formulation of a set of reference materials from readily available sources that can be used for performance measurements on microarrays that differ in design and signal measurement methodology. Cross-platform evaluations have usually been performed using two samples that mimic an experimental design (i.e. comparison of control and perturbed states) for which relatively few changes are subsequently verified as 'real' by qRT–PCR (4,7,8). For evaluating precision within platforms or process drift within a site, a common approach is to measure the correlation of signal values between technical replicates. Unlike these approaches, the MTRRM can provide an assessment of both accuracy and precision for multiple signal ratio measurements in a well-characterized and regenerable sample, allowing comparisons within or between laboratories and platforms. The reference probe set for the MTRRM, defined by cross-platform tissue-selective behavior and mapping to a common curated transcript sequence, provides a basis for measurement of 'true' fold changes in signal level between the two mixes of the MTRRM on multiple platforms. In addition to providing an exemplar set for proficiency assessments, the publicly available dataset generated from this project can serve as a benchmark dataset for evaluating the effect of different data processing choices like signal algorithm and normalization on accuracy and cross-platform agreement. Association of the MTRRM with a performance metric that defines an acceptable range of performance is an area of continuing research that may require accumulation of results from a wider range of conditions and proficiency levels.

The MTRRM provides for measurement of linear range as a function of both signal intensity and fold change, unlike the alternate approaches referred to above. Establishing the reliable measurement range can be critical for applying limitations to biological interpretation of data derived from measurements outside of this range. Deviations from linearity can result from signal noise and assay imprecision as well as from signal plateauing at the upper or lower limits of linearity. Although useful for linear range determinations, for precision or accuracy measurements it may not be desirable to include probes at the lower and higher ends of the signal spectrum that may be more subject to noise or saturation. The effect of relative signal intensity on the measured Mix1:Mix2 ratio can be observed in scatter plots (Figure 2) and in the data presented in Supplementary Table 3.

Although accuracy is an important parameter of microarray performance (20,27), there tends to be an underestimation of fold change with this technology in comparison to measurements made by qRT–PCR (28). Measurement of representative single analytes within the MTRRM using qRT–PCR assays that are each optimized for amplification of individual target transcripts confirmed that the MTRRM contained tissue-selective mRNA at the mixed-in total RNA ratios. In contrast, measurements on microarrays may include a certain level of cross-hybridizing signal, which may be increased in a complex mixture like the MTRRM. The signal attenuation observed with the MTRRM on three different microarray platform formats could be modeled by adding a signal component from non-selective tissues to the tissue-selective signal (Tables 2 and 3). Close agreement with this adjusted target value was achieved on all three platforms.

The formulation and application of the MTRRM produced a set of samples from complex biological sources that have minimal biological variation. The low level of batch-to-batch variance seen for most of the sites in this study and the low site-to-site variance between experienced laboratories indicates that new batches of the MTRRM can be made without adding significantly to performance variance. These batches of the MTRRM can potentially be made at different sites using the same protocol and achieve similar performance. There are steps in the protocol that need to be strictly adhered to in order to reduce variance of results, e.g. quantification of RNA in TNE buffer, and these will continue to be identified as more sites replicate the MTRRM in their own laboratories.

As well as performance level, data comparability is dependent on the use of protocols and reagents optimized to achieve reproducible results and to adherence to protocol standardization across laboratories. In this study, participating laboratories were allowed to use their 'best practice' conditions of labeling and hybridization, although some parameters like image processing method were standardized when found to be a significant source of variation (data not shown). The effect of different array and data processing protocols on the output generated with the MTRRM is under investigation.

A RM for microarrays is of critical importance for evaluating the effect of current and modified protocols, reagent kits and platform design on data comparability and reproducibility. Optimization of microarray technology should be built around achieving detection of the true fold change within a sample, which the design of the MTRRM allows. By its design, the MTRRM can query a broader spectrum of transcript expression than controls made from single tissues. The MTRRM is potentially extensible to other array formats, including cDNA-based arrays, if probe sequences can be matched to the reference probe exemplar sequences in Supplementary Table 1 and the tissue-selective expression of probes can be confirmed through, e.g. body map data that is available for that platform. The design and development of this standard could also serve as a paradigm for similar performance standards for mouse arrays and for human arrays used in clinical applications.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Mark Fielden, Roland Stoughton and David Finkelstein for helpful discussions and Dan Fitzpatrick, Panteha Kiaei, Daniel Baker, Lisa Kivman and Annie Kwok at Amgen Inc. for technical assistance. We thank the US FDA Office of Science and Health Coordination for their support. Funding to pay the Open Access publication charges for this article was provided by the US FDA.

Conflict of interest statement. Gretchen L. Kiser and Tamma Kaysser-Kranich are employees of a microarray manufacturer, GE Healthcare, the makers of CodeLink bioarrays, which were used in this study. For this work, GE Healthcare merely contributed bioarrays and the labor to run them; the data analyses and conclusions were purely based on the resultant raw data and do not reflect any attempt by GE Healthcare to bias the conclusions.

REFERENCES

1. Marshall,E. (2004) Getting the noise out of gene arrays. *Science*, **306**, 630–631.
2. Kuo,W.P., Jenssen,T.K., Butte,A.J., Ohno-Machado,L. and Kohane,I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.
3. Li,J., Pankratz,M. and Johnson,J.A. (2002) Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. *Toxicol. Sci.*, **69**, 383–390.
4. Tan,P.K., Downey,T.J., Spitznagel,E.L.Jr, Xu,P., Fu,D., Dimitrov,D.S., Lempicki,R.A., Raaka,B.M. and Cam,M.C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res*, **31**, 5676–5684.
5. Barczak,A., Rodriguez,M.W., Hanspers,K., Koth,L.L., Tai,Y.C., Bolstad,B.M., Speed,T.P. and Erle,D.J. (2003) Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res.*, **13**, 1775–1785.
6. Bammler,T., Beyer,R.P., Bhattacharya,S., Boorman,G.A., Boyles,A., Bradford,B.U., Bumgarner,R.E., Bushel,P.R., Chaturvedi,K., Choi,D. *et al.* (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods*, **2**, 351–356.
7. Irizarry,R.A., Warren,D., Spencer,F., Kim,I.F., Biswal,S., Frank,B.C., Gabrielson,E., Garcia,J.G., Goeghegan,J., Germino,G. *et al.* (2005) Multiple-laboratory comparison of microarray platforms. *Nature Methods*, **2**, 345–349.
8. Larkin,J.E., Frank,B.C., Gavras,H., Sultana,R. and Quakenbush,J. (2005) Independence and reproducibility across microarray platforms. *Nature Methods*, **2**, 337–344.
9. Shi,L., Tong,W., Goodsaid,F., Frueh,F.W., Fang,H., Han,T., Fuscoe,J.C. and Casciano,D.A. (2004) QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies. *Expert Rev. Mol. Diagn.*, **4**, 761–777.
10. Mecham,B.H., Klus,G.T., Strovel,J., Augustus,M., Byrne,D., Bozso,P., Wetmore,D.Z., Mariani,T.J., Kohane,I.S. and Szallasi,Z. (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res.*, **32**, e74.
11. Shippy,R., Sendera,T.J., Lockner,R., Palaniappan,C., Kaysser-Kranich,T., Watts,G. and Alsobrook,J. (2004) Performance evaluation of commercial short-oligonucleotide microarrays and the impact of noise in making cross-platform correlations. *BMC Genomics*, **5**, 61.
12. Fare,T.L., Coffey,E.M., Dai,H., He,Y.D., Kessler,D.A., Kilian,K.A., Koch,J.E., LeProust,E., Marton,M.J., Meyer,M.R. *et al.* (2003) Effects of atmospheric ozone on microarray data quality. *Anal. Chem.*, **75**, 4672–4675.
13. Dobbin,K.K., Beer,D.G., Meyerson,M., Yeatman,T.J., Gerald,W.L., Jacobson,J.W., Conley,B., Buetow,K.H., Heiskanen,M., Simon,R.M. *et al.* (2005) Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clin. Cancer Res.*, **11**, 565–572.
14. Hoffman,E.P., Awad,T., Palma,J., Webster,T., Hubbell,E., Warrington,J.A., Spira,A., Wright,G., Buckley,J., Triche,T. *et al.* (2004) Expression profiling—best practices for data generation and interpretation in clinical trials. *Nature Rev. Genet.*, **5**, 229–237.
15. Cronin,M., Ghosh,K., Sistare,F., Quackenbush,J., Vilker,V. and O'Connell,C. (2004) Universal RNA reference materials for gene expression. *Clin. Chem.*, **50**, 1464–1471.
16. He,Y.D., Dai,H., Schadt,E.E., Cavet,G., Edwards,S.W., Stepaniants,S.B., Duenwald,S., Kleinhanz,R., Jones,A.R. and Shoemaker,D.D. (2003) Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. *Bioinformatics*, **19**, 956–965.
17. Cope,L.M., Irizarry,R.A., Jaffee,H.A., Wu,Z. and Speed,T.P. (2003) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323–331.
18. Waters,M.D. and Fostel,J.M. (2004) Toxicogenomics and systems toxicology: aims and prospects. *Nature Rev. Genet.*, **5**, 936–948.
19. NCCLS (2003) Evaluation of the Linearity of Quantitative Measurement Procedures: A Statistical Approach; Approved Guideline. NCCLS document EP6-A (ISBN 1-56238-498-8). Wayne, PA, NCCLS.
20. van Bakel,H. and Holstege,F.C. (2004) In control: systematic assessment of microarray performance. *EMBO Rep.*, **5**, 964–969.

21. Ganter,B., Tugendreich,S., Pearson,C.I., Ayanoglu,E., Baumhueter,S., Bostian,K.A., Brady,L., Browne,L.J., Calvin,J.T., Day,G.J. *et al*. (2005) Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol*., **119**, 219–244.
22. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts, and proteins. *Nucleic Acids Res*., **33**, D501–D504.
23. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al*. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol*., **19**, 342–347.
24. Ramakrishnan,R., Dorris,E., Lublinsky,A., Nguyen,A., Domanus,M., Prokhorova,A., Gieser,L., Touma,E., Lockner,R., Tata,M. *et al*. (2002) An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. *Nucleic Acids Res*., **30**, e30.
25. Agilent Technologies (2004) Agilent G2567AA Feature Extraction Software (v 7.5) Manual. Part No. G2566-90011. Santa Clara, CA.
26. Naef,F., Socci,N.D. and Magnasco,M. (2003) A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations. *Bioinformatics*, **19**, 178–184.
27. Moreau,Y., Aerts,S., DeMoor,B., DeStrooper,B. and Dabrowski,M. (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet*., **19**, 570–577.
28. Yuen,T., Wurmbach,E., Pfeffer,R.L., Ebersole,B.J. and Sealfon,S.C. (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res*., **30**, e48.