BMC Genomics

Research article

Learning from microarray interlaboratory studies: measures of precision for gene expression

David L Duewer*[1], Wendell D Jones[2], Laura H Reid[2] and Marc Salit[3]

Address: [1]Analytical Chemistry Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8390, USA, [2]Expression Analysis, 4324 South Alston Avenue, Suite 101, Durham, North Carolina 27713, USA and [3]Biochemical Science Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8313, USA

Email: David L Duewer* - david.duewer@nist.gov; Wendell D Jones - wjones@ExpressionAnalysis.com; Laura H Reid - Lreid2012@nc.rr.com; Marc Salit - marc.salit@nist.gov

* Corresponding author

Published: 8 April 2009

BMC Genomics 2009, **10**:153 doi:10.1186/1471-2164-10-153

Received: 11 July 2008
Accepted: 8 April 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/153

Abstract

Background: The ability to demonstrate the reproducibility of gene expression microarray results is a critical consideration for the use of microarray technology in clinical applications. While studies have asserted that microarray data can be "highly reproducible" under given conditions, there is little ability to quantitatively compare amongst the various metrics and terminology used to characterize and express measurement performance. Use of standardized conceptual tools can greatly facilitate communication among the user, developer, and regulator stakeholders of the microarray community. While shaped by less highly multiplexed systems, measurement science (metrology) is devoted to establishing a coherent and internationally recognized vocabulary and quantitative practice for the characterization of measurement processes.

Results: The two independent aspects of the metrological concept of "accuracy" are "trueness" (closeness of a measurement to an accepted reference value) and "precision" (the closeness of measurement results to each other). A carefully designed collaborative study enables estimation of a variety of gene expression measurement precision metrics: repeatability, several flavors of intermediate precision, and reproducibility. The three 2004 Expression Analysis Pilot Proficiency Test collaborative studies, each with 13 to 16 participants, provide triplicate microarray measurements on each of two reference RNA pools. Using and modestly extending the consensus ISO 5725 documentary standard, we evaluate the metrological precision figures of merit for individual microarray signal measurement, building from calculations appropriate to single measurement processes, such as technical replicate expression values for individual probes on a microarray, to the estimation and display of precision functions representing all of the probes in a given platform.

Conclusion: With only modest extensions, the established metrological framework can be fruitfully used to characterize the measurement performance of microarray and other highly multiplexed systems. Precision functions, summarizing routine precision metrics estimated from appropriately repeated measurements of one or more reference materials as functions of signal level, are demonstrated and merit further development for characterizing measurement platforms, monitoring changes in measurement system performance, and comparing performance among laboratories or analysts.

Background

The ability to demonstrate the reproducibility of gene expression microarray results is a critical element in their adoption for clinical applications. Several studies have asserted that microarray data can be "highly reproducible" if probe sequences are well-mapped to the genome and standard protocols are followed [1-3]. While largely focused on comparisons among measurement platforms, these and other studies have variously characterized many aspects of microarray performance. However, the microarray community has yet to adopt a standardized terminology and practice for characterizing performance that can facilitate clear communication among the user, developer, and regulator stakeholders.

The measurement science (metrology) community is devoted to establishing a philosophically coherent terminology and practice for characterizing and communicating measurement performance [4]. As the world's largest developer and publisher of international consensus standards, the non-governmental International Organization for Standardization (ISO) plays a critical role in disseminating this guidance [5]. The documentary standard ISO 5725-1 [6] details the basic concepts and estimation techniques for assessing metrological "accuracy" which is defined as a combination of two concepts, "trueness" and "precision." These concepts are formally defined in the Vocabulary of International Metrology (VIM) [7] base document and more cogently described in ISO 3534 [8]: trueness is the closeness of a measurement to an accepted reference value and precision is the closeness of measurement results to each other. While microarrays can generate vastly more data per sample than is typical of the processes that shaped the development of these documents, we believe that this pre-existing metrological framework can be extended to microarrays and other highly multiplexed measurement processes.

Properly designed collaborative studies are one of the very best ways of obtaining the information required to characterize some aspects of measurement performance [9]. The three "rounds" of the Expression Analysis Pilot Proficiency Test evaluated replicate samples of a pair of mixed-tissue RNA pools across multiple participants from June to December of 2004; these studies provide a wealth of information relevant to the estimation of several aspects of within-platform measurement precision and among-participant measurement concordance [10]. While the known relationships between the two RNA pools used in these studies also enable evaluation of several measures of trueness in *differential* expression [11], we here evaluate only the metrological concepts of precision as applied to the underlying *direct* measurements. These concepts provide a foundation for the development of objective expectations for the consistency of microarray results. We anticipate that this foundation will facilitate improving the comparability of microarray measurements over time and place and may lead to the development of new approaches and tools for objectively demonstrating the utility of measures of differential expression and ranked lists of differentially expressed genes.

Results

The precision of a defined measurement process can be characterized using three nested metrics: "repeatability," "intermediate precision," and "reproducibility." These measures of precision are defined in terms of the conditions that apply when the measurements are obtained, including: operator, equipment, calibrations, environmental conditions, and the period of time between measurements. Repeatability is defined as the precision of independent measurements when all conditions are assumed constant and thus do not contribute variability (eg, single participants in a given round). Reproducibility is defined as the precision observed when all conditions are permitted to vary within allowable tolerances and thus to contribute variability (eg, all participants in at least one round). Intermediate precision is a special case, where some specified factors are held constant and others are varied (eg, a single participant in two or more rounds). A variety of analysis of variance approaches are suitable for dissecting such data. ISO 5725-2 details the standard approach used in measurement science to characterize measurement precision for a specific measurement process from the results for a single material in a single study [9,12].

Classical precision metrics

In the context of the microarray platform measurements, the signal from each probeset of the array is a single measurement process. The ISO 5725-2 calculations for a single process in a single study are described in Methods, *Classical Precision Metrics for Single Studies*. Methods, *Classical Precision Metrics for Multiple Studies* extends the calculations to multiple studies that use nominally identical samples. The design elements of the Expression Analysis Pilot Proficiency Test studies that enable the use of these calculations are described in Methods, *Study Design*.

While tedious, none of the classical precision metrics are particularly complex or difficult to calculate. However, keeping track of the nomenclature and symbols used for the various metrics can be challenging. In quick summary: let x_{ijk} represent the k^{th} replicate measurement of a given probeset reported by the j^{th} participant in the i^{th} round. The usual mean, $\bar{x}_{ij}$, and standard deviation, $s(x_{ij})$, of the

replicates estimate the value and technical variation for that probeset for the particular participant at one point in time. The value $\bar{\bar{x}}_i$ and technical variation characteristic of the microarray platform itself are estimated by combining the $\bar{x}_{ij}$ and $s(x_{ij})$ values over all participants; the metrological term for this expected technical variation is *repeatability precision* and is represented as s_{ri}. The variation among the different $\bar{x}_{ij}$ estimates the *between-participant precision* and is represented as s_{Li}. In combination, the two sources of variation estimate the *reproducibility precision* which is represented as s_{Ri}. These single-round estimates can be regarded as characterizing the performance of the given probeset over the (typically short) duration of the study.

When nominally identical samples are analyzed in multiple studies, the resulting multiple $\bar{x}_{ij}$ and $s(x_{ij})$ values can be used to estimate the participant-specific expected value, $\bar{\bar{x}}_j$, repeatability, s_{rj}, and *among-round precision*, s_{Wj}; the combination of these two sources of variation estimate the *intermediate precision over time* for the participant, $s_{I(T)j}$ [13]. Combining the single-round, single-participant estimates over all participants can also estimate the long-term expected value, $\bar{\bar{\bar{x}}}$, repeatability, s_r, between-laboratory precision, s_L, and reproducibility, s_R; these long-term estimates can be regarded as characterizing the intrinsic performance of the platform.

Precision functions

The above classical precision metrics characterize performance for individual processes with multiple nominally identical samples of one material. Characterizing processes as a function of signal level can usefully identify the performance expected for "typical" samples [14]. For measurement methods involving one to a few tens of different measurement processes, this can be accomplished with interlaboratory studies involving a relatively small series of samples of similar matrix composition but with varying levels of the analytes of interest. If the many (thousands to millions) of signals typical of microarrays have very different measurement characteristics, then little simplification to this classical approach is possible. However, estimation of aggregate precision functions from analyses of a single material becomes feasible if the majority of the measurement processes share similar performance characteristics. The expected performance of a "typical" probeset with a "typical" sample can be established by compositing the multitudinous individual estimates for one or a few samples.

The Expression Analysis Pilot studies provide results for 31054 probesets in each of two mixed-tissue pools. In the following, we characterize the various precision metrics as functions of signal level from the 62108 unique combinations of probeset and material.

Discrete Precision Functions

Figure 1 illustrates a simple, empirical approach to evaluating relationships between signal level and precision metrics, where the multitude of individual {signal level, precision} values are reduced to a relative handful of expected values [15,16]. Each datum represents the average signal and standard deviation of the triplicate measurements for one probeset for one of the two mixed-tissue pools reported by participant 12 in Round 1, $\{\bar{x}_{1,12}, s(x_{1,12})\}$. The line displays the expected value of these estimates, $\{<\bar{x}_{1,12}>,<s(x_{1,12})>\}$, calculated as the medians of 100 equally sized groups of the $\{\bar{x}_{1,12}, s(x_{1,12})\}$ after ordering on $\bar{x}_{1,12}$. While many summarization approaches could be used, this binning approach has the benefit of relative familiarity and computational simplicity.

Figure 2 displays all of the precision functions for the three rounds: the $\{<\bar{x}_{ij}>,<s(x_{ij})>\}$ standard deviation functions for the individual participants, the $\{<\bar{\bar{x}}_i>,<s_{ri}>\}$

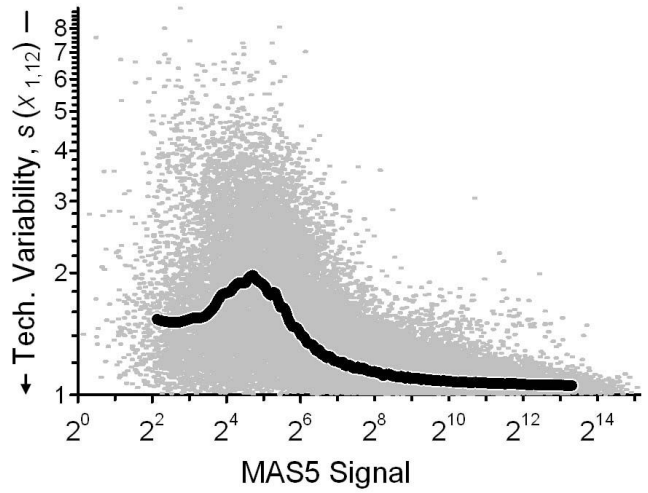

Figure 1
Within-participant repeatability as a function of signal intensity. Each flyspeck denotes the average signal intensity and repeatability, $\{x_{1,12}, s(x_{1,12})\}$, for one of the 62108 sets of measurements made by participant 12 in Round 1. The thick line represents the precision function defined from the median of successive 1/100 binnings of these estimates.

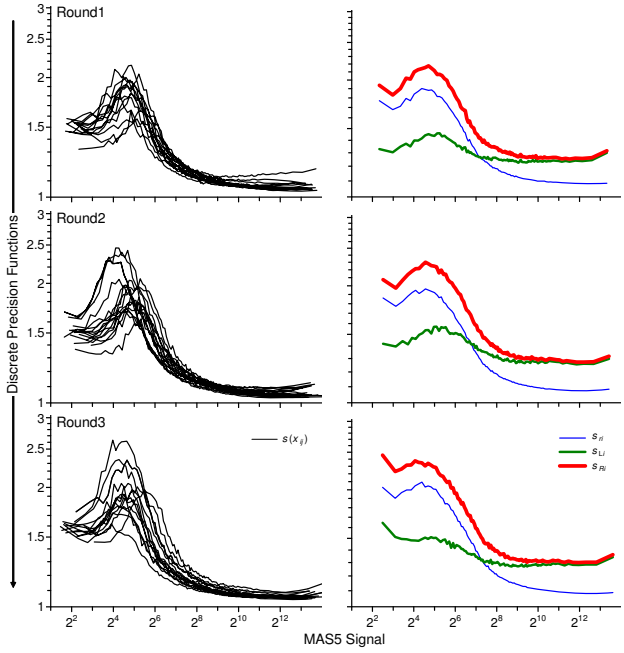

Figure 2

Within-round discrete precision functions. These panels present the single-round discrete precision functions for the three rounds. The panels to the left display participant-average signal intensities and repeatabilities, $\{<\bar{x}_{ij}>,<s(x_{ij})>\}$ for every participant in the round. The panels to the right display round-average intensity and repeatability, $\{<\bar{\bar{x}}_i>,<s_{ri}>\}$, between-participant precision, $\{<\bar{\bar{x}}_i>,<s_{Li}>\}$, and reproducibility, $\{<\bar{\bar{x}}_i>,<s_{ri}>\}$. In all cases the functions are estimated as the medians of successive 1/100 binnings.

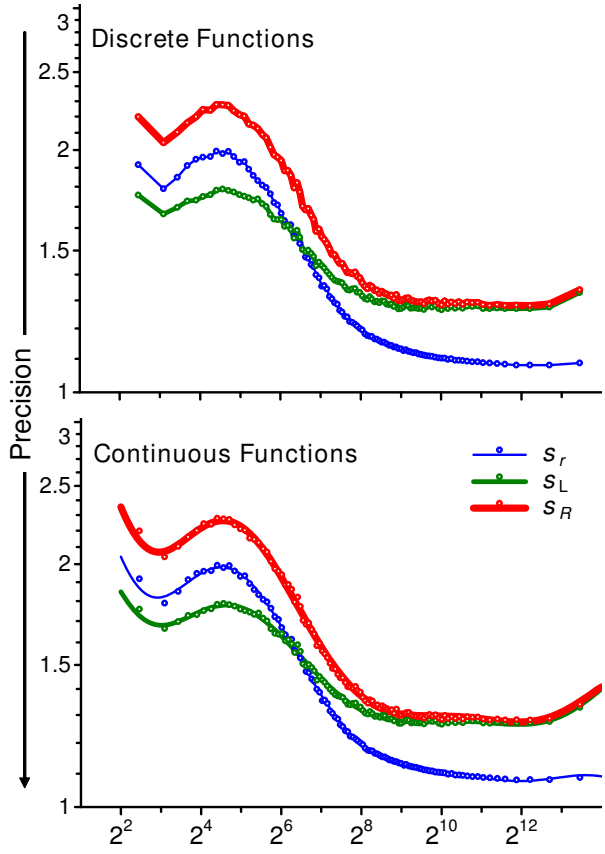

Figure 3

Discrete and continuous precision functions characteristic of the method. The upper panel displays the discrete between-round repeatability $\{<\bar{\bar{x}}>,<s_r>\}$ (thinnest line), between-participant precision $\{<\bar{\bar{x}}>,<s_L>\}$, and reproducibility $\{<\bar{\bar{x}}>,<s_R>\}$ (thickest line) functions. The lower panel displays the same precisions as continuous functions, in the form of 10th-order polynomials. In both panels, the open circles denote the values that define the discrete functions.

repeatabilities, the $\{<\bar{\bar{x}}_i>,<s_{Li}>\}$ between-participant precisions, and the $\{<\bar{\bar{x}}_i>,<s_{Ri}>\}$ reproducibilities. The upper panel of Figure 3 displays the $\{<\bar{\bar{x}}>,<s_r>\}$ between-round repeatability, the $\{<\bar{\bar{x}}>,<s_L>\}$ between-participant precision, and the $\{<\bar{\bar{x}}>,<s_R>\}$ reproducibility functions. The only difference in construction of these discrete functions is use of the appropriate average signal, $\bar{\bar{x}}_i$ or $\bar{\bar{x}}$, rather than the participant average, $\bar{x}_{ij}$, in the initial list creation step.

The pattern and magnitude of the various functions in all three rounds are very similar to each other, indicating that the precision characteristics of the method did not change much over the seven months between Round 1 and Round 3 or with the change in number and identity of the participants. This pattern of technical variability as a function of signal level (in order of lowest to highest signal level: smoothly increasing to a maximum, smoothly decreasing to a relative constant minimum until a small increase at the very highest signal levels) has been often observed with Affymetrix Microarray Suite 5 [17] (MAS5)-processed data [15,18,19]. This complex structure appears to be at least mostly an artifact of MAS5 processing rather than an intrinsic property of the microarray platform, since it is not observed with some other data processing approaches [15,16,20]. Regardless, variability estimated

from short-term studies may underestimate the variability expected over longer periods: while $<s_r>$ for the higher signal levels is a multiplicative factor of $2^{0.14}$ = 1.1 or about 10% relative, $<s_R>$ is a factor of $2^{0.38}$ = 1.3 or about 30% relative. It is perhaps noteworthy that these lessons can only be learned through longitudinal study of measurements from a reasonably well defined population of participants.

Continuous precision functions

While discrete functions are efficient as graphical summaries, they are inconvenient for estimating quantitative values of an expected precision at specific signal levels. For use in further calculations, such as variance-stabilized normalization [21,22], it is desirable to represent the various precision estimates as continuous functions, $\hat{s} = f(x) = f(x)$, where $\hat{s} = f(x)$ is the estimated precision value, x is the signal level, and f is some function parameterized with a relatively small set of coefficients. While a simple four-parameter sigmoidal curve captures much of the structure for most of the discrete functions for the present data, the underlying model assumption of monotonic change as the signal rises from the detection-limit to saturation-limit is not fully adequate to describe MAS5 behavior. Rather than attempting to define and fit more complete theory-based models for particular datasets, platforms, or data processing systems, interpolative empirical functions can readily capture the observed structure. While comparatively crude and unsuitable for extrapolation, even simple polynomials of modestly high-order provide a reasonable qualitative description of the dependence as well as being easily implemented and interpreted. The lower panel of Figure 3 displays 10th-order polynomials parameterized to the discrete between-round precisions shown in the upper panel of Figure 3. At graphical resolution, very similar fits to the discrete functions are provided by polynomials of order 7 and above.

Discussion

Characterizing measurement systems

The between-round repeatability, $\{<\bar{\bar{x}}>,<s_r>\}$, between-participant precision, $\{<\bar{\bar{x}}>,<s_L>\}$, and reproducibility $\{<\bar{\bar{x}}>,<s_R>\}$ functions displayed in Figure 3 provide one definition of the expected measurement precision of signals from one microarray platform, processed with particular software, obtained by a given group of laboratories at a particular point in time. Comparable precision characteristics for other data preprocessing approaches can be estimated by reanalysis of these data. Evaluating the expected characteristics of other platforms will require new studies, with different samples, but comparable functions can be defined given a similar experimental design and sufficient participants of comparable experience. Whether it will be possible to generalize results among related microarray platforms or across data analysis systems is yet to be assessed.

Monitoring performance changes over time

The present data are, however, sufficient to evaluate whether the precision characteristics of the particular platform and signal analysis method are stable over time. The discrete within-round $\{<\bar{\bar{x}}_i>,<s_{ri}>\}$ repeatabilities, $\{<\bar{\bar{x}}_i>,<s_{Li}>\}$ between-participant precisions, and $\{<\bar{\bar{x}}_i>,<s_{Ri}>\}$ reproducibilities that are displayed by each round in Figure 2 are redisplayed in Figure 4 by each precision component. The forms of the functions are very similar over the three studies. Curiously, while the level of within-participant repeatability shows little or no trend, the between-participant precision at signal levels above the median (2^8) appears to have degraded somewhat with time. The invariant repeatability argues against any significant change in the quality of the RNA pools; this plus the distribution of microarrays from multiple lots in Round 1 but single lots in Rounds 2 and 3 argue against significant between-array variability. Thus the small increase in between-participant variability may reflect the changes in the number of participants and the processing protocols that they used; it may also indicate that somewhat less experienced analysts were involved in the later rounds.

Comparing participant precisions

While simple contrasts of repeatability, s_r, and between-participant precision, s_L, (see Methods) enables identification of individual probesets that differ systematically among participants, the variation among the $\{<\bar{x}_{ij}>,<s(x_{ij})>\}$ functions in Figure 2 suggests that the systematic differences among the participants are not confined to just a relatively few measurement processes. Comparison of the within-participant repeatability, $\{<\bar{\bar{x}}_j>,<s_{rj}>\}$, and among-round precision, $\{<\bar{\bar{x}}_j>,<s_{Wj}>\}$, functions for each participant helps identify the nature of long-term changes.

Figure 5 displays these within-participant discrete functions for three exemplar measurement systems. For participant 12 (top panel), s_{r12} is quite good at the highest signal levels and s_{W12} is both quite good and roughly constant

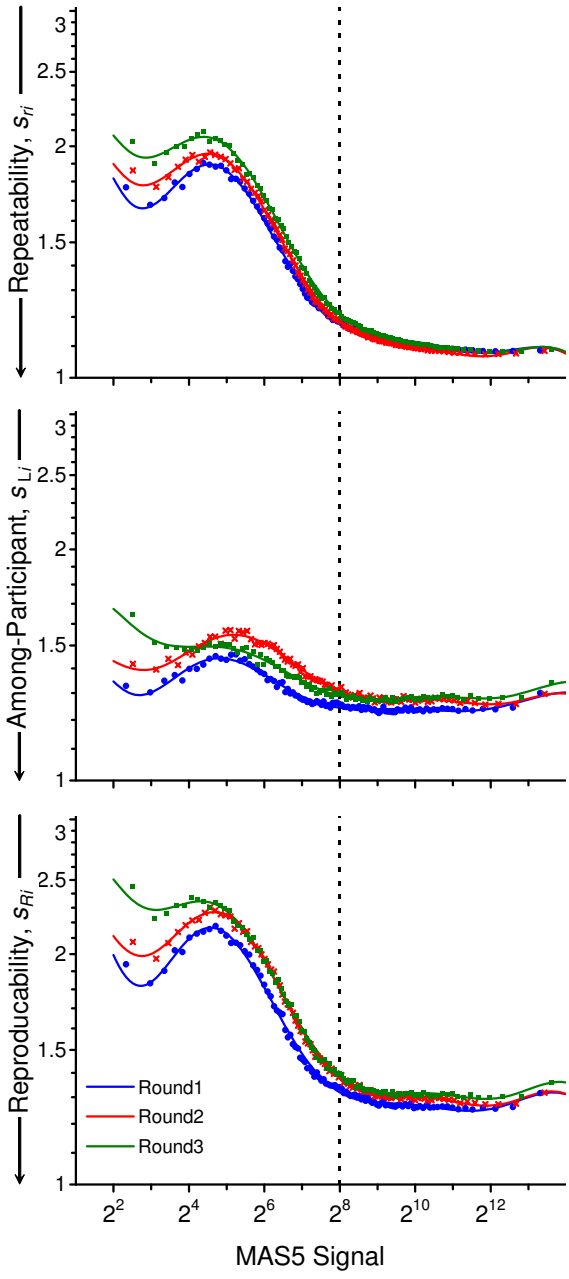

Figure 4
Change in precision over time. These panels redisplay the precision functions shown in the right-side panels of Figure 2 but grouped by the type of estimate rather than by round to facilitate quantitative comparison of change with time. The dotted vertical line denotes the median signal level, 2^8.

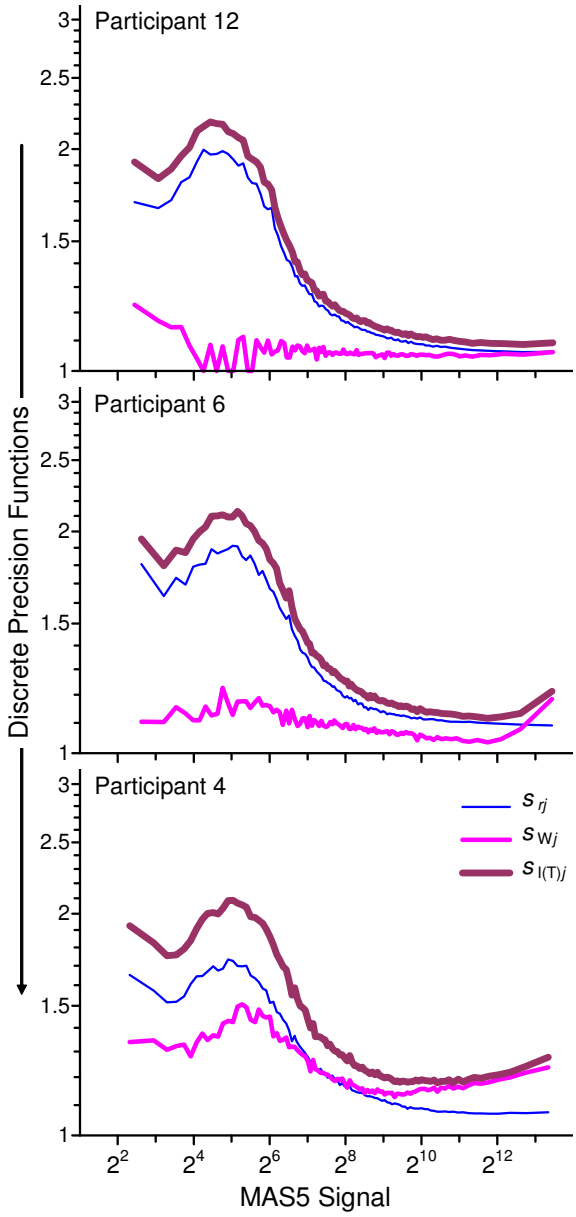

Figure 5
Example discrete precision functions characteristic of the participant. These panels display exemplar within-participant repeatability {$<\bar{\bar{x}}_j>$,$<s_{rj}>$} (thinnest line), among-round precision {$<\bar{\bar{x}}_j>$,$<s_{Wj}>$}, and intermediate precision over time {$<\bar{\bar{x}}_j>$,$<s_{I(T)j}>$}(thickest line) functions for three participants: 12, 6, and 4.

for all levels; this indicates good short-term and excellent long-term control of the measurement process. For participant 6 (middle panel), s_{r6} is somewhat less good then s_{r12} with large signals, and while s_{W6} is excellent for moderately high signals it is systematically less good at both low and very high levels; this suggests somewhat poorer short-term repeatability and long-term changes in the measurement process that impact high and low signals more than those at the mid-levels. We speculate that these changes may be related to scanner performance, influencing both background noise and signal saturation [23]. For participant 4 (bottom panel), s_{r4} is generally as good to somewhat better than for s_{r12} while s_{W4} is considerably poorer for all signal levels; this suggests excellent short-term control but significant differences in the (probably pre-scanner) measurement protocol over the course of the three rounds.

While multivariate approaches could be used to evaluate and display the relative participant performance based on continuous versions of the short-term and long-term precision functions, the basic structure can be readily visualized from the median value of the discrete functions. Figure 6 displays the median within-round repeatability, s_{rj}, and median among-round precision,s_{Wj}. To avoid estimation artifacts, the medians are evaluated over signal levels from 2^8 to 2^{12}. With three exceptions, between-round precision is limited by within-round performance. The excess between-round variability for measurement systems 1B, 4, and 13 likely results from undocumented changes to the systems. The generally poorer precision for systems in protocol group "C" suggests that one or more factor in this group is not well controlled.

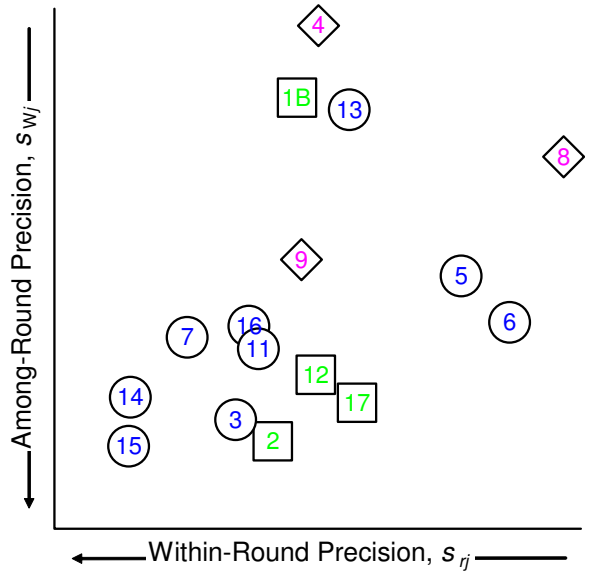

Figure 6
Comparison of median within- and between-round participant precision. This scatterplot displays the median among-round precision, s_{Wj}, as a function of the median within-participant repeatability, s_{rj}, for the 16 measurement systems used in at least two rounds. The open circles denote protocol group "A" measurement systems, open squares denote group "B" systems, and open diamonds denote group "C" systems.

The structure visualized in Figure 6 is congruent with the behavior of the data for the exemplar probeset AFFX_Rat_Hexokinase_5_at discussed in Methods. The association of abstract trends with the behavior of a particular probeset in one sample may facilitate identifying root-causes. The analysis and display tools developed for traditional measurands thus can inform both the development and the interpretation of tools for interlaboratory studies of microarrays.

Conclusion

The established metrological framework for characterizing precision can be applied to results from microarray interlaboratory studies, enabling precision characteristics of microarray results to be expressed in a way that permits comparison to those of other measurement processes. The design of the Expression Analysis Pilot Proficiency Test facilitated assessment of the nested precision metrics of repeatability, intermediate precision over time, and reproducibility – all critical figures of merit of any analytical method. Such studies and figures of merit are essential tools for objective, quantitative performance assessment of individual laboratories, the population of laboratories, and microarray platforms. We believe that continuous precision functions will prove a vital tool for characterizing and comparing measurement platforms and data processing algorithms.

The tools described here for the simplest of microarray signals, are the foundation for further work addressing precision measures for differential expression and differentially expressed gene lists. Figures of merit for these composite signals will support objective performance assessment of the measures behind the biological inference, the reason for performing the measurements in the first place.

Methods

Study design

Expression Analysis, Inc. coordinated three rounds of the Expression Analysis Pilot Proficiency Test in 2004. The first (Round 1) was completed in June, the second (Round 2) in September, and the third (Round 3) in December. All data from these studies are available from ArrayExpress, accession number E-MEXP-1568 http://

Table 1: Study design, participation, and protocol groups

		Protocol Groups[a]			
	Code	Round 1 6/2004	Round 2 8/2004	Round 3 12/2004	#
	1A	A			1
	1B		B	B	2
	2		B	B	2
	3		A	A	2
	4	C	C	C	3
	5	A	A	A	3
	6	A	A	A	3
	7	A	A	A	3
	8	C	C	C	3
	9	C	C	C	3
	10	Y			1
	11	A	A		2
	12	B	B	B	3
	13	A	A		2
	14	A	A	A	3
	15		A	A	2
	16	A	A	A	3
	17		B	B	2
	18			Z	1
Total	19	13	16	15	44

a) Four measurement protocol groups, "A" to "C", were identified on the basis of common suppliers of enzymes, purification kits, instrumentation and similar operating conditions. Two participants used unique protocols that are coded "Y" and "Z".

www.ebi.ac.uk/microarray-as/ae/. The following summarizes the organizational elements pertinent to our analysis.

Participants

Eighteen laboratories participated in at least one of the three rounds, with one laboratory changing its measurement protocol between Round 1 and Round 2. The codes and participation history for the 19 different measurement systems (unique combinations of participant and measurement protocols) are summarized in Table 1. Nine of the measurement systems were used in all three rounds.

Samples

Each participant in each study analyzed three aliquots of two mixed-tissue RNA pools, Mix 1 and Mix 2. The two mixtures were prepared at Expression Analysis by combining different amounts of RNA from rat brain, kidney, liver, and testicle tissues following the design developed by Dr. Karol Thompson and her FDA colleagues [24]. The mixtures were prepared to have total RNA concentrations of 1 μg/μL. Each of the six samples was tagged with unique combinations of five polyadenylated (polyA+) bacterial gene transcripts. The three samples of each mixture were otherwise nominally identical in composition. The sample materials were stored at -80°C by Expression Analysis until shipment and by the participants after receipt. All samples were shipped on dry ice via an overnight delivery service.

Microarrays

Each participant received six Rat230 2.0 GeneChips (Affymetrix, Inc, Santa Clara California, USA) for use in each study, one microarray to be used for each RNA sample. Sixteen arrays were replaced because of various participant-recognized technical problems. To study the influence of lot-to-lot variation, four lots of these arrays were used in Round 1. A single array lot was used in Round 2 and a different single lot in Round 3. The GeneChips are 11-μm format. They were annotated to contain 31099 unique probesets, 27 of which were designed to respond to the polyA+ bacterial control transcripts and 18 to respond to hybridization control transcripts. The remaining 31054 probesets were designed to respond to rat RNA.

Measurement protocols

Each participant prepared biotin-labeled cRNA targets from each sample and hybridized the cRNA to the rat microarrays using his/her own labeling and hybridization reagents. Each participant followed his/her own standard measurement protocol, with the following restrictions: 1) it was strongly recommended that 10 μg of the RNA sample be used to prepare each target, 2) the biotinylated

cRNA was to be prepared using a single round of cDNA synthesis via the Eberwine protocol [25], 3) 20 μg of the fragmented, biotinylated cRNA was to be used with 300 μL of hybridization cocktail including the oligo and eukaryotic controls, and 4) 200 μL of the hybridization cocktail was to be hybridized to each microarray. Three "protocol groups" (coded as "A", "B", and "C") were identified on the basis of the participants' choice of operating conditions and source of enzymes, purification kits, and instrumentation; only two of the 19 measurement systems used during the study were sufficiently different to be considered unique. Table 1 summarizes the use of the different protocols by participant.

Signal estimation and processing

Each participant measured the microarrays following their standard protocols using Affymetrix GeneChip Scanner 3000 instruments with high-resolution software. The resulting "CEL" files for each microarray were sent to Expression Analysis for data processing and evaluation where probeset signal values were evaluated using MAS5 [17]. At NIST, the 55 polyA+ and hybridization controls were deleted from all data sets and the remaining 31054 rat RNA signals were $\log_2$ transformed and centered to have a median $\log_2$(signal) of 8 (ie, a median signal of 2^8 = 256.) This centering value was selected as the integral power of 2 closest to the median of the raw MAS5 signals in all data sets.

Data sets

Forty-four sets of data for the six samples were generated over the course of the three rounds. Each data set consists of 31054 probeset signals for three replicate samples of Mix 1 and Mix 2. Table 1 lists the number of data sets returned in each study.

Classical precision metrics for single studies

Let P represent some one particular measurement process, X the results generated by that process, and x a particular single result. For the data considered here, P is the $\log_2$-transformed MAS5 evaluation of a particular probeset, X the set of $\log_2$-transformed MAS5 results for that probeset in all arrays studied, and x the $\log_2$-transformed MAS5 result for a particular array.

A variety of linear models have been employed to characterize measurement processes from various types of repeated measurements [26]. While complicated models are appropriate when stationary effects (biases) are expected and of interest, given the among-round variation in participants and their measurement protocols, we begin with the general ISO 5725 model. Rather than relying on strong assumptions about the structure of the variance, this simple model is designed to have the general applicability needed for a standard approach. It asserts that each x can be expressed as the sum of three components [12]:

$$x = \bar{x} + B + \varepsilon \tag{1}$$

where $\bar{x}$ is either the true value or more typically a consensus estimate of the quantity being measured, B is the systematic difference (bias) between $\bar{x}$ and the expected result for the given participant's implementation of P (as estimated from replicate measurements), and ε is the random difference between the expectation for the implementation (ie, $\bar{x} + B$) and the given value. For a single material evaluated in a single study (that is, at some given point in time), the ε are assumed to follow a random distribution centered on zero with a standard deviation associated with the (metrological) repeatability precision characteristic of P. Likewise, the variability of the B among all participants is assumed to follow a random distribution centered on zero with a standard deviation associated with the (metrological) reproducibility precision of P. These terms are described more fully in the *Results Section*; we here detail a standard approach to their estimation.

Let x_{ijk} represent the k^{th} of N_m replicate measurements of P reported by the j^{th} participant in the i^{th} study. The standard deviation of the replicates, $s(x_{ij})$, estimates the random variability of P as implemented by the j^{th} participant in the i^{th} study:

$$s\left(x_{ij}\right) = \sqrt{\frac{\sum_{k=1}^{N_m}\left(x_{ijk} - \bar{x}_{ij}\right)^2}{N_m - 1}}; \quad \bar{x}_{ij} = \frac{\sum_{k=1}^{N_m} x_{ijk}}{N_m} \tag{2}$$

where $\bar{x}_{ij}$ is the mean of the replicates. Assuming that the variance magnitude is roughly similar for all participants (as is the case, see Figure 2), the expected random variability of P common to all participants (*i.e.*, its repeatability precision) in the i^{th} study, s_{ri}, can be estimated by combining the individual $s(x_{ij})$ over all N_{pi} participants. While the general formulae detailed in [12] describe variable numbers of replicate measurements for different participants, for these data the same numbers of replicates were reported by every participant in every round. The appropriate formula for pooling variance in this circumstance is:

$$s_{ri} = \sqrt{\frac{\sum_{j=1}^{N_{pi}} s^2\left(x_{ij}\right)}{N_{pi}}}. \quad (3)$$

The s_{ri} can be interpreted as the "average" standard deviation expected for technical replicate measurements made by a typical laboratory, where "typical" is defined by the population of actual participants.

The between-participant precision for the i^{th} study, s_{Li}, is estimated from the standard deviation of the estimated participant-specific biases:

$$s_{Li} = \sqrt{\max\left(0, \frac{\sum_{j=1}^{N_{pi}} \left(\bar{x}_{ij} - \bar{\bar{x}}_i\right)^2}{N_{pi}-1} - \frac{s_{ri}^2}{N_m}\right)}; \quad \bar{\bar{x}}_i = \frac{\sum_{j=1}^{N_{pi}} \bar{x}_{ij}}{N_{pi}} \quad (4)$$

where *max*() is the "take the maximum value of the arguments" function and $\bar{\bar{x}}_i$ is the mean of the participant mean values. The s_{ri}^2/N_m term estimates for the repeatability contribution to the variance of the biases. With atypically noisy replicate measurements, the corrected bias variance is defined as zero – allocating the observed variance to the least-complex source. The s_{Li} estimates the extent of agreement among the various implementations of *P* used by the participants. Ideally, all participants will observe the same mean value for *X* and the value for s_{Li} will be near zero; in practice, studies involving more than one measurement protocol often (by conscious study or from hard experience) discover s_{Li} to be several times larger than s_{ri}.

The reproducibility of *P* during the i^{th} study, s_{Ri}, is estimated by combining the s_{ri} and s_{Li} variance components:

$$s_{Ri} = \sqrt{s_{ri}^2 + s_{Li}^2}. \quad (5)$$

The s_{Ri} combines all of the factors influencing *P* at the time the study was performed; the implementation of *P* in a typical laboratory is expected to yield results that agree with results of other such users within confidence limits appropriate to a normal distribution having mean $\bar{\bar{x}}_i$ and standard deviation s_{Ri}.

The notation used in the above calculations is summarized in Additional file 1. Additional file 2 lists the data and results for the above calculations for one exemplar probeset. Figure 7 displays the relationships among the above estimates for all 31054 probesets in Round 1. While repeatability magnitude is related to signal level, there is considerable variety among the magnitude of the between-participant precision regardless of level. A considerable number of the s_{L1} are plotted along the left-hand margin ($s_{L1} = 2^0 = 1$), a consequence of estimating variance with a relatively small number of replicate measurements.

Additional file 3 summarizes the results of the above calculations for four exemplar probesets in all three rounds. Figure 8 displays these results in a "dot-and-bar" format commonly used with interlaboratory studies. These four probesets were selected as typical of the observed range of the {sL1, sr1} pairs displayed in Figure 7, where the {sL1, sr1} locations are marked with open circles labeled a to d. Exemplar 1, probeset 1379568_at of Mix 2, has very small sL1 and sr1; this represents results that are about the same for all participants, for all replicate samples. Exemplar 2, 1395685_at of Mix 1, has small sL1 but large sr1; this represents results with considerable technical variability but with averages that are about the same for all participants participants. Exemplar 3, 1371165_a_at of Mix 1, has moderate sL1 and sr1; this represents modest variability with some systematic differences among the participants. Exemplar 4, AFFX_Rat_Hexokinase_5_at of Mix 1, has large sL1 but relatively small sr1; this represents results with considerable and quite consistent systematic differences among the participants.

Classical precision metrics for multiple studies

When results for two or more qualitatively similar interlaboratory studies are available, the individually short-term study-specific estimates can be used to define the long-term performance characteristics of the measurement process, *P*, with great confidence. It may also be possible to explore the temporal stability of participant-specific systematic bias, *B*, and random variability, ε. Indeed, an explicit goal of many interlaboratory studies is to help participants identify and minimize sources of systematic difference in their individual implementations of *P* and to establish tighter statistical control over its random influences [27]. Changes in individual performance will manifest may manifest as changes in the study-specific repeatability and reproducibility estimates [28].

Given N_s studies that evaluate identical samples, laboratory-specific repeatabilities can be estimated for all participants that use nominally identical implementations of *P* in at least two of the studies. For the j^{th} such laboratory, s_{rj} is estimated by combining the simple standard devia-

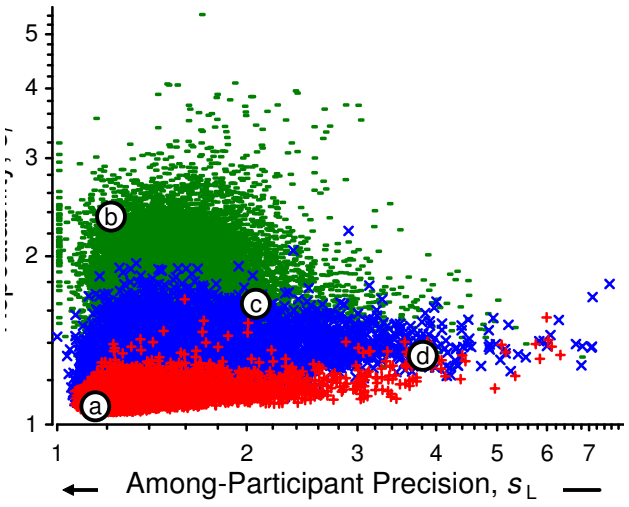

Figure 7

Within-round between-participant and repeatability precisions. This scatterplot displays 62108 repeatabilities for Round 1 data, s_{r1}, as a function of the between-participant precisions, s_{L1}, and the average $\log_2$ signal, $\bar{\bar{x}}_1$, for the Mix 1 and Mix 2 technical replicates. The one-third of probesets with the smallest $\bar{\bar{x}}_1$ are denoted green "-", the middle third are denoted blue "×", and one-third with the largest $\bar{\bar{x}}_1$ are denoted red "+". The labeled open circles denote the {s_{L1}, s_{r1}} location of four exemplar probesets. The precisions are expressed as multiplicative standard deviations: a value of 1 indicates that the values are within a factor of 1 of their mean (i.e., they are identical) whereas a value of 5 indicates a 5-fold spread about their mean.

tions, $s(x_{ij})$, over the N_{sj} studies in which they participated. Since here the same number of replicates were evaluated in each round, the general pooling formula again simplifies to:

$$s_{rj} = \sqrt{\frac{\sum_{i=1}^{N_{sj}} s^2\left(x_{ij}\right)}{N_{sj}}}. \quad (6)$$

In a manner analogous to the between-participant precision described above, laboratory-specific estimates can be obtained for the extent of agreement among results they obtain over relatively long periods of time. The among-round precision for the j^{th} participant, s_{Wj}, is calculated:

$$s_{Wj} = \sqrt{\max\left(0, \frac{\sum_{i=1}^{N_{sj}}\left(\bar{x}_{ij} - \bar{\bar{x}}_j\right)^2}{N_{sj}-1} - \frac{s_{rj}^2}{N_m}\right)}; \quad \bar{\bar{x}}_j = \frac{\sum_{i=1}^{N_{sj}} \bar{x}_{ij}}{N_{sj}} \quad (7)$$

where $\bar{\bar{x}}_j$ is the mean of the mean values for the particular laboratory over all of the studies in which they participated. The intermediate precision over time for the participant, $s_{I(T)j}$ [13], is the combination of s_{rj} and s_{Wj}:

$$s_{I(T)j} = \sqrt{s_{rj}^2 + s_{Wj}^2}. \quad (8)$$

Additional file 4 summarizes these long-term within-participant precision calculations for the four exemplar probesets.

The long-term expected value for X, $\bar{\bar{\bar{x}}}$, and the total number of sets of X values reported by all participants in all of the studies, N_t, can be calculated across the N_p participants or across the N_s studies:

$$\bar{\bar{\bar{x}}} = \frac{\sum_{i=1}^{N_s} N_{pi} \times \bar{\bar{x}}_i}{N_t} = \frac{\sum_{j=1}^{N_p} N_{sj} \times \bar{\bar{x}}_j}{N_t}; \quad N_t = \sum_{i=1}^{N_p} N_{pi} = \sum_{j=1}^{N_s} N_{sj}. \quad (9)$$

The expected long-term repeatability, s_r, is directly calculated from the participation-weighted average of the laboratory-specific repeatability variances; however, the same value is obtained by pooling the study-specific repeatability estimates:

$$s_r = \sqrt{\frac{\sum_{j=1}^{N_p} N_{sj} \times s_{rj}^2}{N_t}} = \sqrt{\frac{\sum_{i=1}^{N_s}\left(N_{pi}-1\right) \times s_{ri}^2}{\sum_{i=1}^{N_s}\left(N_{pi}-1\right)}} \quad (10)$$

Similarly, the expected long-term between-laboratory precision, s_L, is calculated from the study size-weighted average of the between-laboratory precision variances:

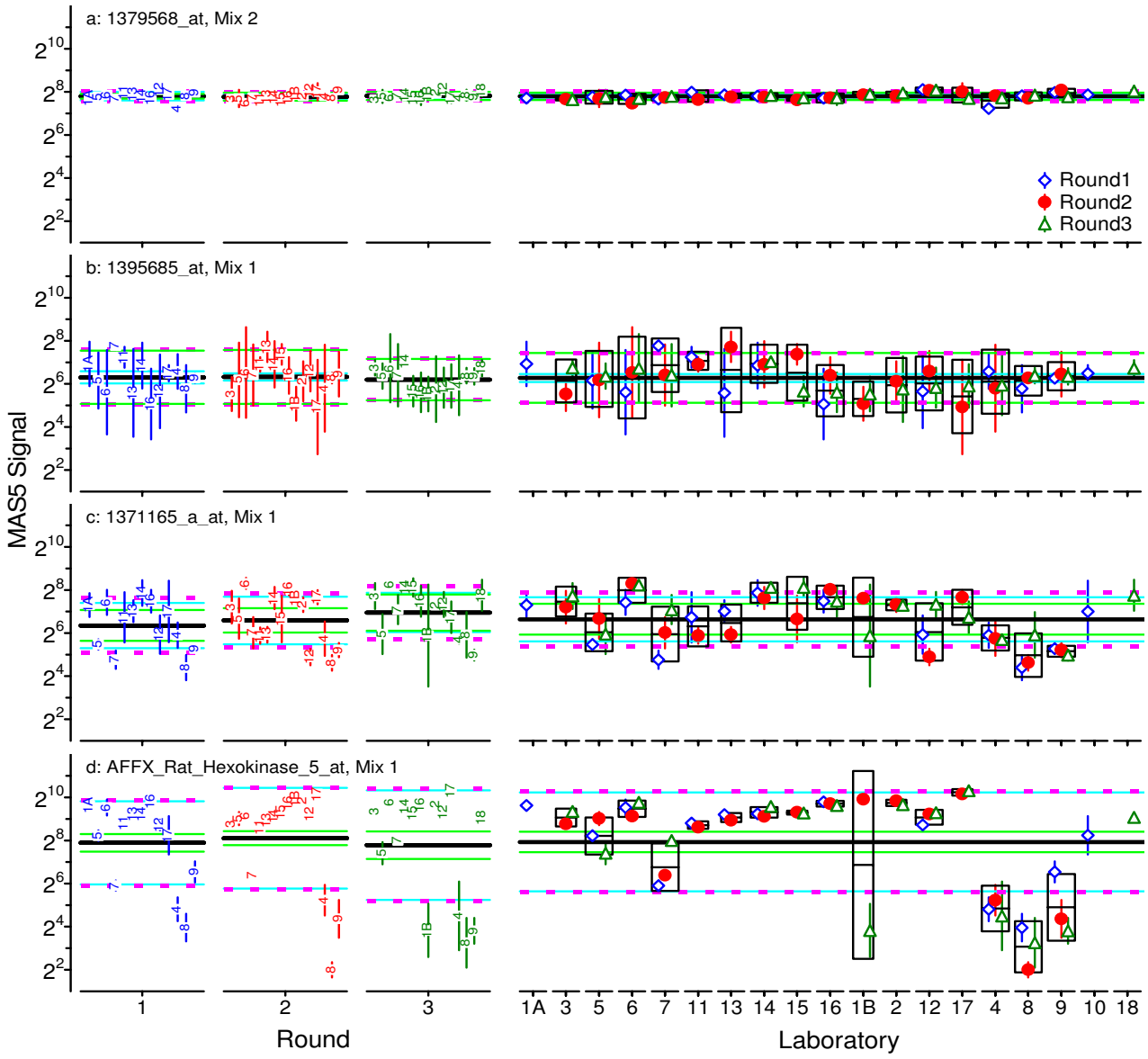

Figure 8

Example graphical analysis of within-round precisions for individual probesets. This "dot-and-bar" chart provides a graphical summary of the participant averages, $\bar{x}_{1i}$, and standard deviations, $s(x_{1i})$, for the four probesets labeled in Figure 7. The within-round averages, $\bar{\bar{x}}_1$, are represented as a thick black line; the repeatabilities, s_{r1}, as solid green lines one precision unit on either side of average; the between-participant precision, s_{L1}, as solid light-blue lines; and the reproducibilities, s_{R1}, as dashed magenta lines. Participant codes are ordered by protocol group: Group A {1A,5,6,7,11,13,14,16}, Group B {12,17}, and Group C {4,8,9}.

$$s_L = \sqrt{\frac{\sum_{i=1}^{N_S} N_{pi} \times s_{Li}^2}{N_t}}. \qquad (11)$$

The expected long-term reproducibility of the measurement process, s_R, can be calculated from the study size-weighted average of the study-specific reproducibility variances or by combining s_r and s_L:

$$s_R = \sqrt{\frac{\sum_{i=1}^{N_s} N_{pi} \times s_{Ri}^2}{N_t}} = \sqrt{s_r^2 + s_L^2}. \quad (12)$$

Figure 9 displays the long-term repeatabilities and between-laboratory precisions characteristic of the microarray platform for all of the 31054 probesets. The relationship of the precision estimates to the signal level is not changed from Figure 7, although the increased number of data used in the estimates can be seen in the much reduced number of probesets with $s_L = 2^0 = 1$. Note that the locations of the exemplar probesets relative to $\{s_L, s_r\}$ are unchanged after three rounds. Figure 10 displays all of the above precision estimates for the four exemplar probesets.

"Outliers"

It is often necessary to identify and remove grossly aberrant values and/or use robust statistical estimation techniques to enable sensible summarization of a given data set. Other than the arrays that were replaced because of participant-recognized technical problems, no "outlier arrays" (in the sense of the majority of MAS5 values for one of the three replicate arrays being quite different from those of its siblings) were identified in the final data. Robust methods, including those advocated in ISO 5725-5 [29], were evaluated and found to yield estimates very similar to those described above.

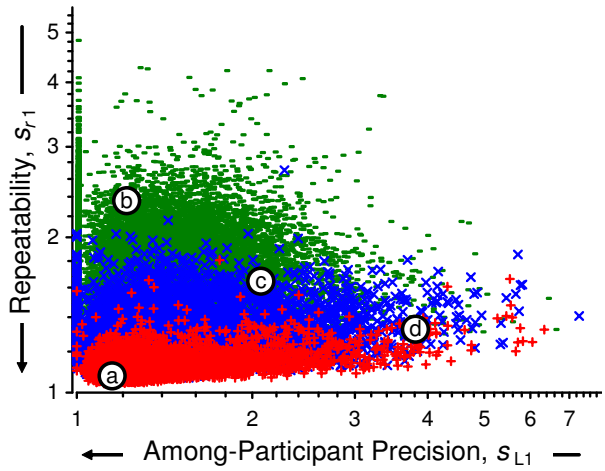

Figure 9
Between-round between-participant and repeatability precisions. This scatterplot is identical in format to Figure 7, only displaying the overall between-participant precision and repeatability estimates pairs, $\{s_L, s_r\}$ rather than just those of Round 1.

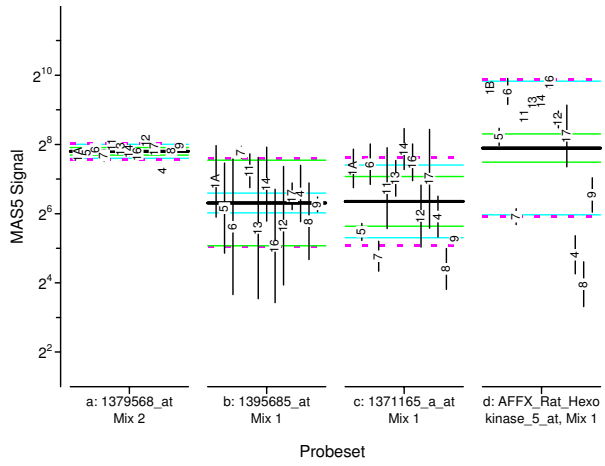

Figure 10
Example graphical analysis of between-round precisions for individual probesets. The left section of each panel displays dot-and-bar charts for the three rounds in the format used in Figure 8. The right section displays the same participant average and standard deviation dot-and-bar data, but grouped by participant rather than round. The mid-line of the box around the data for each participant represents that participant's average, $\bar{\bar{x}}_j$; the lower and upper edges of the box denote the intermediate precision over time, $s_{I(T)j}$. The grand average, $\bar{\bar{\bar{x}}}$, is represented as a think black line; the repeatability, s_r, as solid green lines; the between-participant precision, s_L, as solid light-blue lines; and the reproducibility, s_R, as dashed magenta lines. Participant codes are ordered by protocol group: Group A {1A,3,5,6,7,11,13,14,15,16}, Group B {1B,2,12,17}, Group C {4,8,9}, Y {10}, and Z {18}.

Authors' contributions

DLD participated in the development of the precision estimators, coded the calculations, and drafted the manuscript. WDJ and LHR designed the interlaboratory studies, collected and processed the resulting data, and advised on their interpretation. MS participated in the development of the precision estimators and drafting the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

Summary of notation. *Summarizes the notation used with the various calculations detailed in Methods.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-153-S1.pdf]

Additional file 2

Data and precision calculations for RNA Mix 1, exemplar probeset AFFX_Rat_Hexokinase_5_at. *Provides the data and detailed results of the various calculations for one of the exemplar probesets described in Methods.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-153-S2.pdf]

Additional file 3

Summary of method-related precision estimates for four exemplar probesets. *Provides numeric values for the various method-related precision estimates for the four exemplar probesets described in Methods.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-153-S3.pdf]

Additional file 4

Summary of participant-related precision estimates for four exemplar probesets. *Provides numeric values for the various participant-related precision estimates for four exemplar probesets described in Methods.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-153-S4.pdf]

Acknowledgements

We thank the participants of the interlaboratory studies for their cooperation and support.

Certain commercial equipment, instruments, and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

References

1. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J: **Independence and reproducibility across microarray platforms.** *Nat Methods* 2005, **2:**337-344.
2. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JGN, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang YQ, Ye SQ, Yu W: **Multiple-laboratory comparison of microarray platforms.** *Nat Methods* 2005, **2:**345-349.
3. MAQC Consortium: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.** *Nat Biotech* 2006, **24:**1151-1161.
4. de Bièvre P: **Essential for metrology in chemistry, but not yet achieved: truly internationally understood concepts and associated terms.** *Metrologia* 2008, **45:**335-341.
5. ISO: *ISO in brief. International Standards for a sustainable world. Geneva* 2008 [http://www.iso.org/iso/isoinbrief_2008.pdf].
6. ISO: *ISO 5725-1:1994/Cor 1:1998 Accuracy (trueness and precision) of measurement methods and results. Part 1: General principles and definitions. Geneva* 1994 [http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=11833].
7. JCGM 200-2008: *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM)* 3rd edition. 2008 [http://www.bipm.org/en/publications/guides/vim.html]. Sèvres, France
8. ISO: *ISO 3534-1:2006 Statistics – Vocabulary and symbols – Part 1: General statistical terms and terms used in probability. Geneva* 2006 [http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=40145].
9. ISO: *ISO/TR 22971:2005 Accuracy (trueness and precision) of measurement methods and results – Practical guidance for the use of ISO 5725-2:1994 in designing, implementing and statistically analysing interlaboratory repeatability and reproducibility results. Geneva* 2005 [http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39780].
10. Duewer DL, Kline MC, Sharpless KE, Thomas JB, Gary KT, Sowell AL: **Micronutrients Measurement Quality Assurance Program: Helping participants use interlaboratory comparison exercise results to improve their long-term measurement performance.** *Anal Chem* 1999, **71:**1870-1878.
11. Pine PS, Boedigheimer M, Rosenzweig BA, Turpaz Y, He YD, Delenstarr G, Ganter B, Jarnagin K, Jones WD, Reid LH, Thompson KL: **Use of diagnostic accuracy as a metric for evaluating laboratory proficiency with microarray assays using mixed-tissue RNA reference samples.** *Pharmacogenomics* 2008, **9:**1753-63.
12. ISO: *ISO 5725-2:1994 Accuracy (trueness and precision) of measurement methods and results. Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method. Geneva* 1994 [http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=11835].
13. ISO: *ISO 5725-3:1994 Accuracy (trueness and precision) of measurement methods and results. Part 3: Intermediate measures of the precision of a standard measurement method. Geneva* 1994 [http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=11834].
14. Thompson M, Wood R: **Using uncertainty functions to predict and specify the performance of analytical methods.** *Accredit Qual Assur* 2006, **10:**471-478.
15. Holloway AJ, Oshlack A, Diyagama DS, Bowtell DD, Smyth GK: **Statistical analysis of an RNA titration series evaluates microarray precision and sensitivity on a whole-array basis.** *BMC Bioinformatics* 2006, **7:**511.
16. Held GA, Duggar K, Stolovitzky G: **Comparison of Amersham and Agilent Microarray Technologies Through Quantitative Noise Analysis.** *Omics* 2006, **10:**532-544.
17. Affymetrix, Inc: *Statistical Algorithms Description Document* 2002 [http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf]. Santa Clara, CA USA
18. Mansourian R, Mutch DM, Antille N, Aubert J, Fogel P, Le Goff JM, Moulin J, Petrov A, Rytz A, Voegel JJ, Roberts MA: **The Global Error Assessment (GEA) model for the selection of differentially expressed genes in microarray data.** *Bioinformatics* 2004, **20:**2726-2737.
19. Zhou L, Rocke DM: **An expression index for Affymetrix GeneChips based on the generalized logarithm.** *Bioinformatics* 2005, **21:**3983-3989.
20. Liggett W: **Normalization and technical variation in gene expression measurements.** *J Res Natl Inst Stand Technol* 2006, **111:**361-372 [http://nvl.nist.gov/pub/nistpubs/jres/111/5/V111.N05.A02.pdf].
21. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18:**S96-S104.
22. Weng L, Dai H, Zhan Y, He Y, Stepaniants SB, Bassett DE: **Rosetta error model for gene expression analysis.** *Bioinformatics* 2006, **22(9):**1111-1121.
23. Satterfield MB, Lippa K, Lu JL, Salit M: **Microarray scanner performance over a 5-week period as measured with Cy3 and Cy5 serial dilution dye slides.** *J Res Natl Inst Stand Technol* 2008, **113:**157-174 [http://nvl.nist.gov/pub/nistpubs/jres/113/3/V113.N03.A03.pdf].
24. Thompson KL, Rosenzweig BA, Pine PS, Retief J, Turpaz Y, Afshari CA, Hamadeh HK, Damore MA, Boedigheimer M, Blomme E, Ciurlionis R, Waring JF, Fuscoe JC, Paules R, Tucker CJ, Fare T, Coffey EM, He Y, Collins PJ, Jarnagin K, Fujimoto S, Ganter B, Kiser G, Kaysser-Kranich T, Sina J, Sistare FD: **Use of a mixed tissue RNA design for performance assessments on multiple microarray formats.** *Nucleic Acids Res* 2005, **33:**e187.
25. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P: **Analysis of gene expression in single live neurons.** *Proc Natl Acad Sci U S A.* 1992, **89(7):**3010-3014.

26. Pavese F, Filipe E: **Some metrological considerations about replicated measurements on standards.** *Metrologia* 2006, **43:**419-425.
27. Duewer DL, Kline MC, Sharpless KE, Thomas JB, Stacewicz-Sapuntzakis M, Sowell AL: **NIST Micronutrients Measurement Quality Assurance Program: Characterizing individual participant measurement performance over time.** *Anal Chem* 2000, **72:**3611-3619.
28. Duewer DL, Kline MC, Sharpless KE, Thomas JB: **NIST Micronutrients Measurement Quality Assurance Program: Characterizing the measurement community's performance over time.** *Anal Chem* 2000, **72:**4163-4170.
29. ISO: *ISO 5725-5:1998 Accuracy (trueness and precision) of measurement methods and results. Part 5: Alternative method for the determination of the precision of a standard measurement method. Geneva* 1998 [http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=1384].