

Multiple-laboratory comparison of microarray platforms

Rafael A Irizarry¹, Daniel Warren², Forrest Spencer³, Irene F Kim⁴, Shyam Biswal⁵, Bryan C Frank⁶, Edward Gabrielson⁷, Joe G N Garcia⁸, Joel Geoghegan⁹, Gregory Germino⁴, Constance Griffin¹⁰, Sara C Hilmer¹¹, Eric Hoffman¹¹, Anne E Jedlicka¹², Ernest Kawasaki⁹, Francisco Martínez-Murillo¹³, Laura Morsberger¹⁰, Hannah Lee⁵, David Petersen⁹, John Quackenbush^{6,14}, Alan Scott¹², Michael Wilson^{15,17}, Yanqin Yang², Shui Qing Ye⁸ & Wayne Yu¹⁶

Microarray technology is a powerful tool for measuring RNA expression for thousands of genes at once. Various studies have been published comparing competing platforms with mixed results: some find agreement, others do not. As the number of researchers starting to use microarrays and the number of cross-platform meta-analysis studies rapidly increases, appropriate platform assessments become more important. Here we present results from a comparison study that offers important improvements over those previously described in the literature. In particular, we noticed that none of the previously published papers consider differences between labs. For this study, a consortium of ten laboratories from the Washington, DC–Baltimore, USA, area was formed to compare data obtained from three widely used platforms using identical RNA samples. We used appropriate statistical analysis to demonstrate that there are relatively large differences in data obtained in labs using the same platform, but that the results from the best-performing labs agree rather well.

Microarray technology has become an important tool in medical science and basic biology research. A first time user will find many platform options and little guidance on which is the most appropriate for their application. Various comparison studies have been published presenting contradictory results. Some have observed agreement in results obtained with different platforms^{1–6}, others have not^{7–10}. Here we demonstrate that the disagreement observed in some studies may be due to disputable statistical analyses. In particular, none of the prior studies have considered lab-to-lab variability (lab effect). The lab effect

has been observed in all scientific fields¹¹. Therefore, it is essential to assess this effect before drawing conclusions about platform performances.

A consortium of ten labs from the Washington, DC–Baltimore, USA, area was formed to compare the performance of three leading platforms. Researchers in each lab were given identical RNA samples that were processed according to what was considered best practice in each lab. Affymetrix GeneChips were used in five of the labs (Affymetrix labs 1–5), two-color spotted cDNA arrays were used in three labs (two-color cDNA labs 1–3), and two-color long oligonucleotide arrays were used in two labs (two-color oligo labs 1 and 2). Here we describe the features of our experiment that are necessary for such studies to be informative and a set of simple assessment measures useful for summarizing and interpreting the observed data.

To decide among various strategies for measuring the same quantity, one looks to optimize accuracy and precision. Because in many situations precision can be improved at the cost of accuracy, and vice versa, one tries to find the strategy providing the ‘best’ balance. Because the definition of best depends on the application, it is important to consider precision and accuracy in the context of a realistic problem. We mimicked the most common application of microarray technology: screening for a few candidate genes that appear to be differentially expressed among thousands of genes that are not. In this context, an appropriate comparison experiment requires at least the following three features. (i) To appropriately assess precision we should make a comparison with an *a priori* expectation of no-fold change for most or all genes. (ii) To appropriately assess accuracy, an *a priori* expectation of nonzero

¹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA. ²Department of Surgery, Johns Hopkins University, Baltimore, Maryland 21205, USA. ³McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. ⁴JHU NIDDK Gene Profiling Center, Department of Medicine, Johns Hopkins University, Baltimore, Maryland 21205, USA. ⁵Department of Environmental Health Sciences, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA. ⁶The Institute for Genomic Research, 9712 Medical Center Dr., Rockville, Maryland 20878, USA. ⁷Department of Pathology, Johns Hopkins University, Baltimore, Maryland 21231, USA. ⁸Division of Pulmonary and Critical Care Medicine, Johns Hopkins University School of Medicine, Mason F. Lord Bldg., Center Tower #665, Baltimore, Maryland 21224, USA. ⁹NCI’s Microarray Core Facility, Advanced Technology Center, Gaithersburg, Maryland 20877, USA. ¹⁰Department of Pathology, Johns Hopkins University, School of Medicine, Baltimore, Maryland 21287, USA. ¹¹Research Center for Genetic Medicine, Children’s National Medical Center, George Washington University, Washington, DC 20052, USA. ¹²W. Harry Feinstone Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA. ¹³Department of Molecular Biology and Genetics, Johns Hopkins University, Baltimore, Maryland 21205, USA. ¹⁴Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115-6084, USA. ¹⁵Microarray Research Facility, Research Technologies Branch, DIR, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20892, USA. ¹⁶Oncology Microarray Facility, Johns Hopkins University, Baltimore, Maryland 21231, USA. ¹⁷Present address: Ambion, Inc., Austin, Texas 78744, USA. Correspondence should be addressed to R.A.I. (rafa@jhu.edu).

Table 1 | Assessment measures for all ten labs

Platform	Lab number	Precision		Accuracy signal (s.e.m.)	Proportion of agreement		
		Correlation	s.d.		25	50	100
Affymetrix oligo	1	0.48	0.32	0.62 (0.05)	0.72	0.56	0.54
Affymetrix oligo	2	0.76	0.17	0.64 (0.05)	0.80	0.70	0.70
Affymetrix oligo	3	0.67	0.24	0.66 (0.05)	0.68	0.66	0.60
Affymetrix oligo	4	0.79	0.15	0.59 (0.04)	0.80	0.70	0.65
Affymetrix oligo	5	0.59	0.25	0.58 (0.05)	0.64	0.68	0.55
Two-color cDNA	1	0.65	0.23	0.41 (0.12)	0.68	0.64	0.65
Two-color cDNA	2	0.68	0.21	0.13 (0.04)	0.28	0.30	0.38
Two-color cDNA	3	0.46	0.23	0.54 (0.09)	0.72	0.68	0.50
Two-color oligo	1	0.68	0.51	0.21 (0.09)	0.40	0.36	0.33
Two-color oligo	2	0.90	0.10	0.76 (0.13)	0.44	0.72	0.81

To summarize precision we used the correlation across replicate \log_2 -fold change measurements and standard deviation (s.d.) of the difference between replicate \log_2 -fold change measurements. To quantify accuracy we regressed the observed \log_2 -fold changes of 16 genes against nominal \log_2 -fold changes obtained using RT-PCR. The slope of the regression line defines what we refer to as accuracy signal. The proportion of agreement in interesting genes lists—ranked by fold change—of sizes 25, 50 and 100, created with replicate \log_2 -fold change measurements, are also used to assess precision.

\log -fold change of a few genes is needed. (iii) To be able to distinguish between platform effect and lab effect, at least two labs should provide data from each platform. We have designed the first platform comparison experiment that includes all of these features.

In general, the Affymetrix labs achieved better accuracy and precision. But overall, the best-performing lab was two-color oligo lab 2. Furthermore, two-color cDNA labs 1 and 3 outperformed most Affymetrix oligo labs in some categories. The worst performance was observed from a two-color oligo lab; thus the best and worst overall performance was achieved using the same platform. This underscores the importance of considering the lab effect. In general, we found that the lab had a larger effect on, for example, precision than did the platform, and that the results from the best-performing labs agreed rather well.

RESULTS

Assessment measures and plots

We created two samples in which we expect a few genes to be differentially expressed. To do this we developed a strategy based on mixtures from four knockout human cell lines that resulted in four specific genes with a *a priori* expectation of fold change different

from 1 (**Supplementary Methods** online). We refer to these genes as the altered genes. For each of these two samples we created an exact copy, or technical replicate, for a total of four samples. Exact copies of these four samples were hybridized by the ten labs using their platform of choice, and the resulting data were processed as described. We quantified relative expression between the two duplicate pairs of samples with \log_2 -fold change. This resulted in two replicate \log_2 -fold change measurements for each gene, from each lab.

To summarize precision we used two simple measures: correlation across replicate \log_2 -fold change measurements and standard deviation (s.d.) of the difference between replicate \log_2 -fold change measurements. These assessment measures can also be used to quantify the similarity between measurements made using different platforms. We refer to these two assessment measures as correlation and s.d. (**Table 1**, columns 3 and 4). A box plot of the differences used to compute the s.d. for each lab provides a graphical summary (**Fig. 1a**).

To assess accuracy we validated 16 genes using RT-PCR (**Supplementary Methods**). The 16 genes included the four altered genes, four randomly selected genes from those that were

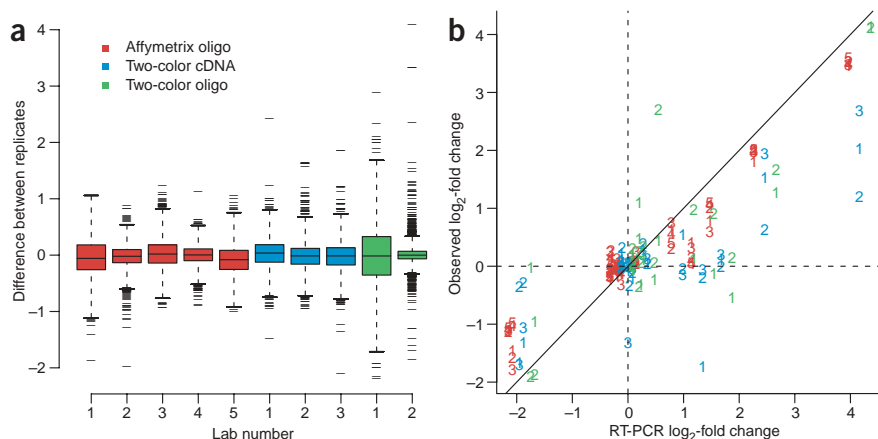


Figure 1 | Precision and accuracy assessment figures. **(a)** Box plot of the difference in \log_2 -fold change between replicate measurements of gene expression from each of the ten labs. The platform used is represented by different colors defined in the figure. **(b)** Observed \log_2 -fold change versus nominal (calculated from RT-PCR experiments) \log_2 -fold change for the four altered genes and 12 other genes. The results for each of the 10 labs are represented by the lab number and color for the different platforms as in **a**. The solid diagonal line is the identity function and represents perfect accuracy.

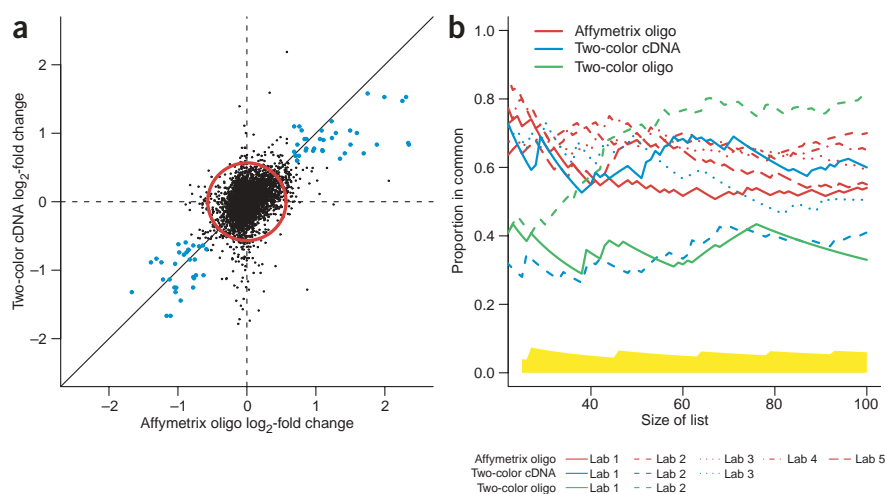


Figure 2 | Plots demonstrating agreement for differentially expressed genes. (a) Scatter plot of observed \log_2 -fold change from two-color cDNA lab 1 and Affymetrix oligo lab 4. Points inside red circle represent genes that do not appear to be differentially expressed. Blue points are genes that appear to be differentially expressed. The solid diagonal line is the identity function and represents perfect accuracy. (b) CAT plot showing agreement between differential expression calls, defined by ranking genes by fold change, using replicate measurements from each lab. We considered list sizes less than 100 because we do not expect more than 100 genes to be differentially expressed, thus correspondence of larger lists is not of interest. The three colors represent the different platforms as in **Figure 1a**. The different line types represent the different labs within each platform so that a color and line-type pair uniquely represents each lab. The yellow strip represents critical values for rejecting the null hypothesis of no agreement at the 0.001 level.

consistently found not to be differentially expressed across all platforms, four genes that were consistently found to be differentially expressed across all platforms, and four genes found to be differentially expressed using one platform and not the others. To quantify accuracy we regressed the observed \log_2 -fold changes of these 16 genes against nominal \log_2 -fold changes obtained by RT-PCR analysis. The slope of the regression line defines our assessment measure, which we refer to as the signal (**Table 1**, column 5). A graphical summary is the scatter plot of the observed versus nominal values obtained by all labs (**Fig. 1b** and **Supplementary Fig. 1** online).

A scatter plot of the \log_2 -fold changes obtained by the best-performing Affymetrix oligo and two-color cDNA labs showed no correlation for about 95% of genes (**Fig. 2a**). These genes had \log_2 -fold changes close to zero and were probably not differentially expressed. Because for these genes it is likely that we measured zero \log_2 -fold change plus random measurement error, we did not expect cross-platform measurements to correlate. But for the few genes that appeared to be differentially expressed there was good agreement. In practice, we typically screen a small subset of genes that appear to be differentially expressed. Therefore, it is more important to assess agreement for genes that are likely to pass this screen. To account for this, we introduced a new descriptive plot: the correspondence at the top (CAT) plot. This plot is useful for comparing two procedures for detecting differentially expressed genes. To create a CAT plot we made a list of n candidate genes for each of the two procedures and plotted the proportion of genes in common against the list size n (**Fig. 2b**). As assessment measures, we reported the value of these curves for list sizes 25, 50 and 100. We refer to these assessment measures as the proportion of agreement (**Table 1**, columns 6, 7 and 8).

Preprocessing

We found that within- and across-platform performance can be greatly improved using alternative preprocessing algorithms to the defaults offered by the array manufacturers. For our analysis, probe-level data from the Affymetrix oligo arrays were preprocessed with the robust multiarray analysis (RMA)¹². Print-tip normalization with no background correction was used to preprocess probe-level data from the two-color platforms¹³. Spot-quality information was ignored because we found it did not have substantial impact on downstream results. Because algorithms implementing these methodologies are available from the Bioconductor project¹⁴, we will refer to them as the Bioconductor procedures. We compared the results obtained with this approach to those obtained with what we consider to be the default approaches: Affymetrix's MAS 5.0 algorithms for Affymetrix oligo arrays and median adjustment normalization with background correction for the two-color technologies. Although in general the default procedures had slightly better accuracy (not statistically significant), the gains in precision given by the

Bioconductor procedures were dramatic. Because of the great improvement provided by the Bioconductor procedures (**Supplementary Fig. 2** online and **Supplementary Table 1** online), we use them for all of the experiments presented in this paper.

Annotation

To match features across platforms, we used mappings that match features to genomic entities that are available from various public databases. Resourcerer¹⁵ provides mappings that link features to UniGene, LocusLink and RefSeq for all the platforms used in our experiment. Resourcerer also provides its own annotation in a

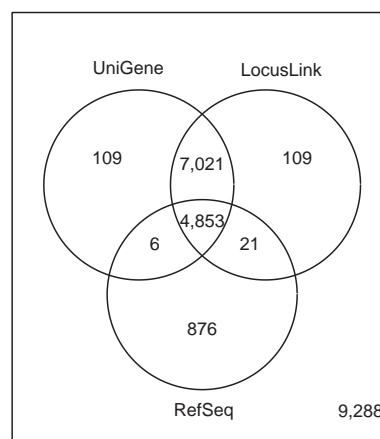


Figure 3 | Venn diagram illustrating agreement between annotation databases. For each mapping (UniGene, LocusLink and RefSeq) we obtained a different set of genes that had identifiers for each platform. This Venn diagram shows the agreement between these three different lists.

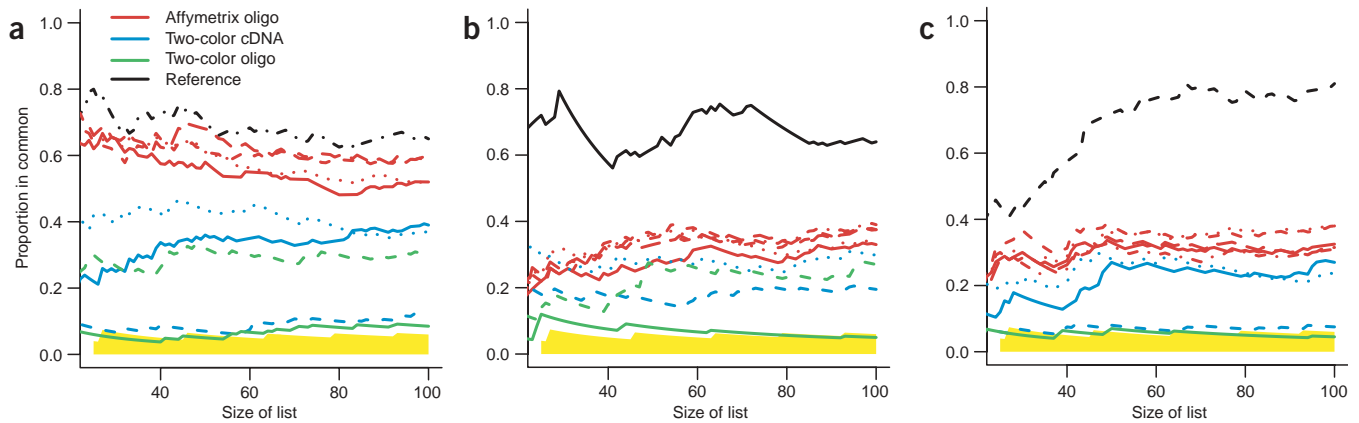


Figure 4 | CAT plots showing agreement in differential expression calls, based on fold change, between each lab and a reference lab. (a–c) The different line types represent the individual labs, and the three colors represent the different platforms as in **Figure 2b**. The black curve is the CAT curve comparing replicates from the reference lab. (a) CAT plot using data from the best-performing Affymetrix oligo lab as the reference. (b) CAT plot using data from the best-performing two-color cDNA lab as the reference. (c) CAT plot using data from the best-performing two-color oligo lab as a reference.

eukaryotic gene orthologs (EGO) database. Unfortunately, none of these mappings are one to one: not all the features in the arrays are annotated and/or some are annotated with more than one genomic identifier. Therefore, for a particular annotation only a subset of the array features will have an entry for each platform. Furthermore, these subsets differ depending on which annotation was used (**Fig. 3**). The annotation used had an effect on the across-platform agreement. For example, the correlation between measurements from Affymetrix oligo lab 4 and two-color cDNA lab 1 was 0.39–0.44 when using UniGene and EGO, respectively. We found that using the genes having entries in all databases for all platforms provided the best agreement. For all the analyses presented here we used the subset of genes obtained from this intersection (**Supplementary Table 2** online).

Platform comparison

Our results demonstrated that precision is comparable across platforms (**Table 1** and **Fig. 1a**). With the exception of two-color oligo lab 1, all the labs performed similarly, and it is clear that the lab effect is stronger than the platform effect. All the labs provided attenuated \log_2 -fold change estimates, and this is consistent with previous observations¹² (**Fig. 1b**). In general, the labs using the Affymetrix platform seem to attain better accuracy than the labs using two-color platforms, although the best signal measure was attained by two-color oligo lab 2. Two-color cDNA

lab 2 and two-color oligo lab 1 were clearly underperforming. The differences in data obtained by the other eight labs were not statistically significant.

We used CAT plots to assess across-platform agreement. It is important to note that these were used to compare results from single array experiments, and thus we did not expect perfect agreement. Note, for example, that the agreement of lists of the top 100 genes created from replicate fold-change measurements ranged from 33–81 percent (**Fig. 1b**). CAT plots comparing across-lab agreement demonstrate that the Affymetrix oligo labs consistently provided results similar to those from the best-performing labs (**Fig. 4**). This suggests that the Affymetrix platform provides by far the most consistent data across labs. Apart from two labs, there appears to be good agreement regardless of the platform used (**Fig. 4** and **Supplementary Table 3** online).

DISCUSSION

We defined a series of assessment measures and plots used to compare three leading microarray platforms. These were justified by questions of scientific interest and have practical interpretations. The signal measure represents the expected \log_2 -fold change in expression of a gene that should be differentially expressed with a nominal fold change of two, and the s.d. measure gives us the expected \log_2 -fold change of a null gene. These two measures gave us a clear idea of the signal-to-noise ratio. Although, overall, the Affymetrix platform performed best, it is important to keep in mind that this platform is typically more expensive than the alternatives.

We also demonstrated that there was relatively good agreement between the Affymetrix labs and the best-performing two-color labs. These results contradict some previously published results that find disagreement across platforms^{7–10}. The conclusions reached by these studies are likely due to three misconceptions. The first misconception is that absolute measurements of gene expression can be used to assess data across platforms. Note that both studies using absolute measurements had found disagreement^{7,10}. Results established based on absolute measurements are misleading because they are adversely affected by platform-dependent probe effects that can be removed by considering relative measurements

Table 2 | Correlation and s.d. measurements computed for absolute and relative measurements of expression

	Correlation		s.d.	
	Absolute	Relative	Absolute	Relative
Affymetrix oligo versus Affymetrix oligo	0.98	0.79	0.16	0.15
Two-color cDNA versus two-color cDNA	0.91	0.65	0.29	0.23
Affymetrix oligo versus two-color cDNA	0.40	0.44	0.91	0.25

Affymetrix oligo lab 4 and two-color cDNA lab 1 were used for this comparison.

of expression. The statistical model used to motivate our assessment measures, described in the Methods section, can be used to demonstrate this point. Note that in all studies interested in differential expression of genes, relative expression is the quantity of interest; thus this type of measurement is always available. The second misconception is that preprocessing has no significant effect on final results. With one exception⁴, all previous studies had used algorithms that have been shown to be inferior to alternatives developed by the academic community^{12,13}. Finally, the third misconception is that platform performance is not affected by lab. The existence of the sizable lab effect was ignored in all previously published comparison studies. This permits the possibility that studies done by, for example, experienced technicians may find agreement and studies done by less-experienced technicians may find disagreement (**Supplementary Fig. 3** online).

Although we found relatively good across-platform agreement, it is quite far from being perfect. In all across-platform comparisons, there was a small group of genes that had relatively large fold changes from data obtained using one platform but not using the others (**Fig. 2b**). We conjecture that some genes were not measured correctly, not because the technologies are not performing adequately, but because transcript information and annotation can still be improved.

Our results provide a useful assessment of three leading technologies and demonstrate the need for continued cross-platform comparisons. In fact, Affymetrix has released a new platform for measuring gene expression in humans, which yields slight improvements in accuracy and precision (**Supplementary Figs. 1** and **4** online and **Supplementary Table 4** online). We expect our study to serve as a starting point for larger, more comprehensive comparisons. Furthermore, our findings show that improved quality assessment standards are needed. Assessments of precision based on comparisons of technical replicates appear to be standard operating procedure among, at least, academic labs. We have demonstrated that precision and accuracy assessments are not informative unless performed simultaneously. We hope that our study serves as motivation to create such standards. This will be essential for the success of microarray technology as a general measurement tool.

METHODS

Data analysis. A commonly used statistical model for microarray data is $Y_{ijk} = \theta_i + \phi_{ij} + \varepsilon_{ijk}$, in which Y_{ijk} represents measurement k of \log_2 -scale expression of gene i measured by platform j . Here θ_i represents absolute gene expression in the \log_2 scale. ϕ_{ij} denotes the platform-specific probe or spot effect. Measurement error is represented by ε_{ijk} . For illustrative purposes we considered each of the effects in this model to be random and statistically independent from each other. We represented their variances with v_θ , v_ϕ and v_ε .

Many researchers have observed a sizeable probe effect in microarray data, which implies that v_ϕ is large¹². This will result in artificially large correlations when comparing absolute measurements obtained using the same platform. To see this, note that within-platform correlation is $\text{corr}(Y_{ij1}, Y_{ij2}) = (v_\theta + v_\phi)/(v_\theta + v_\phi + v_\varepsilon)$. This correlation is typically close to one, but only because v_ϕ can be much larger than v_θ and v_ε . If we compare across platforms, the correlation will not be as large, but only because the probe effect is not common to the two platforms and therefore does not

affect the correlation $\text{corr}(Y_{i1k}, Y_{i2k}) = v_\theta/(v_\theta + v_\phi + v_\varepsilon)$. These theoretical predictions were confirmed empirically (**Table 2**).

A simple solution to the probe effect problem is to consider relative expression instead of absolute expression. Most experiments compare between different samples, thus in general this type of measure is readily available. By considering the difference of the Y_{ijk} from the two samples, the ϕ_{ij} are cancelled out. Because these are \log_2 -scale measurements this difference is simply the \log_2 ratio of the absolute expression levels.

For the relative expression measurements, the within-platform correlations were substantially smaller and the across lab correlation was a bit larger (**Table 2**). We propose that only assessments based on relative expression are useful. All the results presented in this paper deal with relative expression.

Additional methods. Sample preparation, RT-PCR and microarray hybridization and experimental design are described in **Supplementary Methods** online. The code and data necessary to reproduce this work are available online (<http://www.biostat.jhsph.edu/~ririzarr/techcomp>).

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

ACKNOWLEDGMENTS

We thank A. Nones and K. Broman for useful suggestions. The work of R.A.I. is partially funded by the National Institutes of Health Specialized Centers of Clinically Oriented Research (SCCOR) translational research funds (212-2494 and 212-2496). The work of G. Germino and I. Kim was partially funded by NIDDK U24DK58757.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 30 November 2004; accepted 22 March 2005
Published online at <http://www.nature.com/naturemethods/>

- Kane, M. *et al.* Assessment of the sensitivity and specificity of oligonucleotide (50-mer) microarrays. *Nucleic Acids Res.* **28**, 4552–4557 (2000).
- Hughes, T. *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* **19**, 342–347 (2001).
- Yuen, T., Wurmbach, E., Pfeffer, R.L., Ebersole, B.J. & Sealfon, S.C. Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res.* **30**, e48 (2002).
- Barczak, A. *et al.* Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res.* **13**, 1775–1785 (2003).
- Carter, M. *et al.* *In situ*-synthesized novel microarray optimized for mouse stem cell and early developmental expression profiling. *Genome Res.* **13**, 1011–1021 (2003).
- Wang, H. *et al.* Assessing unmodified 70-mer oligonucleotide performance on glass-slide microarrays. *Genome Biol.* **4**, R5 (2003).
- Kuo, W., Jenssen, T., Butte, A., Ohno-Machado, L. & Kohane, I. Analysis of mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405–412 (2002).
- Kothapalli, R., Yoder, S., Mane, S. & Loughran, T.P. Jr. Microarray results: how accurate are they? *BMC Bioinformatics* **3**, 22 (2002).
- Li, J., Pankratz, M. & Johnson, J. Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. *Toxicol. Sci.* **69**, 383–390 (2003).
- Tan, P. *et al.* Evaluation of gene expression measurements from commercial platforms. *Nucleic Acids Res.* **31**, 5676–5684 (2003).
- Youden, W. Enduring values. *Technometrics* **14**, 1–11 (1972).
- Irizarry, R.A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
- Dudoit, S. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).
- Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
- Tsai, J. *et al.* Resourcerer: a database for annotating and linking microarray resources within and across species. *Genome Biol.* **2** software0002.1–0002.4 (2001).