April 2007

What Is the Best Reference RNA? And Other Questions Regarding the Design and Analysis of Two-Color Microarray Experiments

Kathleen F. Kerr
University of Washington, katiek@u.washington.edu

Kyle A. Serikawa
University of Washington

Caimiao Wei
M.D. Anderson Cancer Center

Mette A. Peters
University of Washington

Roger E. Bumgarner
University of Washington

Suggested Citation

Kerr, Kathleen F.; Serikawa, Kyle A.; Wei, Caimiao; Peters, Mette A.; and Bumgarner, Roger E., "What Is the Best Reference RNA? And Other Questions Regarding the Design and Analysis of Two-Color Microarray Experiments" (April 2007). *UW Biostatistics Working Paper Series.* Working Paper 305.
https://biostats.bepress.com/uwbiostat/paper305

INTRODUCTION

The reference design, described in (Kerr, M. K. *et al.*, 2001) and (Yang, Y. H. *et al.*, 2002b), is an extremely popular choice for two-color microarray studies. In a reference design, a "reference" RNA is co-hybridized with the RNAs of interest. The design is intuitively appealing because every RNA can be compared with any other RNA since each RNA is directly compared to the reference. While the reference design may be technically less efficient than some other design choices (Kerr, M. K. *et al.*, 2001;Kerr, M. K., 2003a), the efficiency disadvantage can be minor and is often considered negligible in light of the design's advantages: the design is simple and produces data that are easy to analyze compared to more elaborate designs. In addition, a reference design is very flexible, as it is easy to add new or previously unanticipated samples into a study.

In a reference design, the reference RNA is hybridized in one channel of every array in the experiment, so fully half of all hybridizations are to the reference. Since the reference RNA uses such a large proportion of array resources, and since the reference is the linchpin of the design, selecting an RNA to use as the reference is a choice that investigators take very seriously. Some previous work has discussed choices for reference RNAs (Gorreta, F. *et al.*, 2004;He, X. R. *et al.*, 2004;Novoradovskaya, N. *et al.*, 2004;Yang, I. V. *et al.*, 2002). Each of these papers asserts that the most important quality of a reference RNA is that it has good representation of all of the genes on the array, i.e. that most genes give a signal "above background" when hybridized to the reference. This assertion presumably originates with the idea that low-intensity signals are unreliable. Based on this assertion, these studies typically evaluate reference RNAs by examining the percentage of spots on arrays that give signal above some threshold. Note that such an evaluation is several steps removed from evaluating how well a reference facilitates getting accurate answers to a scientific question of interest.

We question whether the most important quality of a reference RNA is broad representation of all genes on the array because this supposition ignores an important fact. Namely, the popular methods for normalizing two-color array data rely on assumptions that most genes are not differentially expressed between the co-hybridized

RNAs, or that the amount of differential expression between them is roughly symmetric (Quackenbush, J., 2002). Our supposition is that the most important quality of a reference RNA is that it satisfies the assumptions of the normalization routine. In other words, a "good" reference will not be too different from the RNAs in the study.

We propose the following reasoning as a basis for evaluating reference RNAs. A reference RNA allows an "indirect" comparison between RNAs of interest. Therefore, it seems reasonable to require that a good reference faithfully reproduce the comparison that would have been acquired had two RNAs of interest been directly compared. One justification for using this criterion is that it favors reference RNAs that give estimates that are not "reference-specific." We comment further on the merits of this method of evaluation in the Discussion. Note a similar principle was applied by König et al (Konig, R. *et al.*, 2004) .

We evaluated the efficacy of three common choices for reference RNAs: (1) a pool of all the non-reference RNAs in our study; (2) placenta RNA; and (3) a commercially-sold RNA that is promoted as a "universal" reference. Our experimental design (see METHODS) employed these three reference RNAs in a design with three different "test-pairs" of other RNAs. The RNAs comprising the test-pairs play the role of the RNAs "of biological interest" in a real study. We chose the "test" RNAs strategically to evaluate our supposition that a reference RNA will perform better when it is more similar to the RNAs of interest. Based on this reasoning, we made specific predictions about the performance of the three reference RNAs prior to collecting the data (see RESULTS).

METHODS

Experimental Design. We used 9 mouse RNAs in this study: 3 "reference" RNAs, and 3 pairs of "test" RNAs. Figure 1 shows the experimental design for one test-pair. The reference RNAs were chosen to represent some common reference choices: (1) a pool of the six "test RNAs"; (2) an aliquot of placenta RNA; and (3) a commercial RNA that is promoted as a "universal" reference. We refer to these as the "pool reference", "placenta reference," and the "commercial reference" in the remainder of this paper. Test RNAs 1a and 1b are from placenta, which was chosen to be most similar to the placenta reference. Test RNAs 2a and 2b are kidney, which was chosen because kidney is a component of the commercial reference. Test RNAs 3a and 3b are lung, which was chosen because the supplier disclosed that lung is a small or nonexistent component of the commercial reference. We refer to these six RNAs as the placenta test-pair, the kidney test-pair, and the lung test-pair.

As shown in Figure 1, there was a dye-swap pair of arrays between each test-pair and a dye-swap pair of arrays between each test RNA and each reference RNA. Our investigation therefore used 14 arrays for each test-pair, for a total of 42 microarrays.

All samples were acquired from a single RNA isolation. Each RNA isolation was divided into two aliquots, one of which was labeled with Cy3 and the other with Cy5. These two dye-labeled aliquots were used in all hybridizations to control for labeling variation when comparing references. For example, test RNA 1 hybridized to the commercial reference is from the same labeling reaction as test RNA 1 hybridized to the placenta reference, so any difference in the performance of the references cannot be explained by labeling inconsistencies of the test RNAs.

All tissues were isolated from normal C57Bl/6J mice. Placenta tissue was isolated from time-mated pregnant mice with E17 embryos attached.

Laboratory Assays. Test RNAs were extracted from intact tissues using the RNeasy Mini Kit (Qiagen). RNAs were quantified via absorption at 260 and 280nm and checked for RNA quality using the Agilent 2100 Bioanalyzer Nano chip (Agilent Technologies). The commercial RNA used was the Universal Mouse Reference RNA from Stratagene (http://www.stratagene.com/homepage/). The Placenta RNA was acquired from Zyagen (http://zyagen.com/). These RNAs were evaluated on the bioanalyzer before use. The pool reference was constructed by combining equal mass aliquots of each of the six test RNAs.

RNA was amplified and labeled using the Agilent low-input fluorescent labeling kit (Agilent Technologies) and an equal amount (0.75ug) of each fluor-labeled target was hybridized according to Agilent's instructions. The arrays in the study were Agilent Whole Genome Mouse Microarrays. Following hybridization and washing, arrays were scanned using the Agilent MicroArray scanner and intensities extracted using Agilent's Feature Extraction software version 7.5. All arrays were checked to ensure that the data spanned the dynamic range appropriately, with a distribution of spot intensities typical of high-quality hybridizations.

Data Pre-Processing. Four versions of "spot intensity" were extracted from the text files generated by Agilent's Feature Extraction software: (1) mean foreground intensity, (2) mean foreground intensity minus median background intensity, (3) median foreground intensity, and (4) mean foreground intensity minus mean background intensity. All control spots were excluded. In subsequent analyses, each of these versions of the data was considered without normalization, or normalized using the the "loess" method (Cui, X. *et al.*, 2003) available in the R add-on package MAANOVA (available at http://www.jax.org/staff/churchill/labsite/software). This normalization is a generalization of the intensity-normalization proposed in (Yang, Y. H. *et al.*, 2002a). We present results for data versions (1) and (2). Results for data version (3) were similar to, but not as good as, the results for version (1), and are not shown. Similarly, results for data version (4) were similar to, but not as good as, the results for version (2), and are not shown.

Note that the Feature Extraction software makes a local measurement of "background" using the pixels around each spot in the microarray image. This background measurement was used for a simple background-subtraction where noted below.

Tissue Specific Genes. To identify tissue specific genes, we utilized Novartis' publicly available SymAtlas (http://symatlas.gnf.org/SymAtlas/). Data for the genes that were mostly highly expressed in lung, kidney or placenta were downloaded. The compilation of tissue specific genes was made by taking the normalized intensities of these genes across all available tissues and visually selecting those genes that appeared to have no or minimal expression in any but the specified tissue.

Low Intensity Genes. A gene is considered "low intensity" according to the following:

1. Average the Cy3 and Cy5 intensities on all arrays and identify the 10th percentile
2. For a particular array, a gene is considered low-intensity if its Cy3 and Cy5 averaged intensity is below the threshold determined in (1.)
3. For a given comparison, a gene is considered low-intensity if it is low-intensity on one or more arrays involved in the comparison.

Data Analysis. For each test-pair of RNAs, we can measure the relative expression between them with a "direct" comparison, using the dye-swap between the test-pair. In addition, we can use the three "indirect" comparisons, via each of the three reference RNAs. Specifically, the indirect logratio for comparing test samples A and B using a particular reference is calculated as:

$$\text{indirect-logratio}_{\text{A vs. B}} = \text{logratio}_{\text{A vs. Ref}} - \text{logratio}_{\text{B vs. Ref.}},$$

where $\text{logratio}_{\text{A vs. Ref}}$ and $\text{logratio}_{\text{B vs. Ref}}$ are the means of the appropriate logratios from the pair of arrays between the test sample and the reference.

For each test-pair of RNAs, we plotted the indirect $\log_2$ratios for each reference RNA against the direct $\log_2$ratios. We summarize these scatterplots with Lin's correlation coefficient (Lin, L. I., 1989). (Whereas Pearson correlation measures how well bivariate data are summarized by a straight line, Lin's correlation measures of how well bivariate data are summarized by a straight line through the origin with slope 1. Lin's correlation coefficient is always less than or equal to Pearson's correlation.) In general, correlation metrics are problematic as a measure of reproducibility because the magnitude of correlation depends on factors such as the spread of the data. However, for a given set of direct logratios, it is reasonable to use correlation to compare different sets of indirect logratios because the direct logratios are held constant.

The experimental design allowed us to investigate other important questions related to microarray data analysis. Because the experimental design contains multiple 3-loops, a test of normalization procedures is how well direct and indirect log-ratios agree. We applied different low-level data processing techniques and compared direct and indirect log-ratios for different methods of processing the data. Specifically, we considered the data with a "loess" method of normalization (Cui, X. *et al.*, 2003) to data with no normalization other than dye-swap averaging. In addition, we considered the data with and without background-adjustment (BA) (local background subtraction). We also investigated genes for which direct and indirect logratios were highly discrepant, and identified characteristics of such genes that may be useful for quality control in future studies.

RESULTS

Predictions and Performance of Reference RNAs. Common methods of normalization for microarray data assume that differential expression between co-hybridized RNAs is roughly symmetric and/or most genes are not differentially expressed (Quackenbush, J., 2002). From this fact we reasoned that an effective reference RNA should not be too dissimilar from the RNAs of interest in a study, since the reference RNA is co-hybridized with every other RNA. Our experimental design and our test RNAs were specifically chosen to evaluate this reasoning. Based on our hypothesis, we predicted that

(a.) The "pool" reference RNA should work well overall.
(b.) The placenta reference RNA should be the best reference for the placenta test RNAs.
(c.) The commercial reference RNA should work well for the kidney test RNAs but not as well for the lung test RNAs, since kidney is a component of the commercial reference but lung is not.

Predictions (a) and (b) were borne out whereas prediction (c) was not. Figures 2 (placenta test-pair), 3 (kidney test-pair), and 4 (lung test-pair) show scatterplots of the direct logratios from the dye-swap between test RNAs (horizontal axis) against the indirect logratios for each of the reference RNAs. These figures support prediction (a):

the pooling strategy is generally effective for constructing an experiment-specific reference RNA. Figure 2 shows that prediction (b) is substantiated, namely that the placenta reference gives the best agreement with the direct logratios for the placenta test pair, although the improvement over the other references is admittedly small.

Figure 3 shows that prediction (c) did not hold. The commercial reference is the worst reference for the kidney test-pair, even though the commercial reference includes kidney as a component. Furthermore, the commercial reference works as well as the other two reference RNAs for the lung-test pairs (Figure 4), even though lung is a small or non-existent component of this reference. In light of our incomplete knowledge of the commercial reference (further details of its composition is proprietary information held by the vendor), we cannot offer further explanation for the performance of the commercial reference. Lung may have an expression profile similar to another tissue that is part of the commercial reference, leading to better-than-expected performance of the commercial reference with the lung test-pair. Alternatively, the poor performance for the kidney test-pair may have been a chance event.

These results are elucidated by considering tissue-specific genes. For the lung test-pair, lung-specific genes are highlighted in Figure 4. Clearly, the pool reference gives the best reproduction of the direct logratios. This is exactly what we would expect, since the pool reference is the only reference containing lung RNA. It also appears that the commercial reference gives the worst agreement for the placenta-specific genes (Figure 2). For the kidney-specific genes, the commercial reference also does poorly, illustrating the overall poor performance of the commercial reference for this test pair.

Data Processing. Figures 2, 3, and 4 are based on intensity-normalized array data that were not background-adjusted. We also constructed versions of these figures using different pre-processings of the array data. Following the methodology of (Qin, L. X. *et al.*, 2004) and (Members of the Toxicogenomics Research Consortium, 2005b), we considered the data with and without background-subtraction and with and without loess normalization. Figure 5 shows the plots for the lung-test pair when background-subtracted spot intensities are used, followed by loess normalization; see Supplementary information for the remaining figures and Table 1 for a summary of the results. Without exception, the best agreement between the direct and indirect logratios was for the version of the data <u>without</u> background-subtraction and <u>with</u> intensity-normalization (in other words, the data as shown in Figures 2, 3, and 4).

	Placenta Test-Pair			Kidney Test-Pair			Lung Test-Pair		
Reference RNA:	Pool	Placenta	Comm	Pool	Placenta	Comm	Pool	Placenta	Comm
Loess normalization, no BA	.77	.80	.77	.65	.64	.51	.79	.79	.80
No normalization, no BA	.34	.68	.76	.09	.50	.18	.63	.63	.57
Loess normalization, with BA	.62	.60	.61	.50	.50	.28	.67	.67	.69
No normalization, with BA	.21	.29	.31	.01	.22	.04	.56	.49	.36

Table 1. Lin's correlation coefficients for four different processings of the data. Uniformly, the best concordance between direct and indirect logratios was for the version of the data that used the global loess normalization and did not use background-adjustment (BA). See Supplementary material for the associated scatterplots. All foreground intensities are the mean spot intensity. The background spot intensity (for the last two rows of the table) is the median intensity of the background pixels for a spot, as determined by the image analysis software.

Correlation coefficients are technically not directly comparable between different pre-processings of the data because the scale of the data changes. However, inspection of the scatterplots (supplementary figures 2, 3, and 4) indicates that the summary of concordance provided by Lin's correlation coefficient is reasonable; agreement between direct and indirect logratios is clearly worsened with alternative pre-processings of the data. An alternative approach is to examine the absolute size of the discrepancy between direct and indirect logratios. Figures 6 and 7 show the normalized data with background-adjustment (Figure 6) and without background-adjustment (figure 7) for the lung test pair. For the background-adjusted data, the discrepancy between indirect and direct logratios is clearly larger on average, dramatically so for low-intensity genes.

Data Quality Control. A noticeable pattern in Figures 2, 3, and 4 is a “+” shape within the scatterplots. For a handful of genes, the direct and indirect logratios are highly discrepant in a particular way: one logratio is very near 0, while the other is not. Examining the logratios for any of these genes on the individual arrays, we observed a clear pattern. Specifically, the logratio was very near 0 for all arrays except one. Obviously these are genes with high variability in measurement, but with a particular kind of variability.

Based on these observations, we propose a filter for identifying suspect measurements for array quality control. For any gene, let $(|LR_1|, |LR_2|, \ldots, |LR_n|)$ be the absolute values of the observed logratio for that gene on the n arrays in an experiment using the normalized data. For each gene, compute the median and skew of these numbers. Figure 8A shows that the skewness tends to increase with the median absolute logratio across arrays. However, for a small number of genes with small median absolute logratio, the skewness is large. If we highlight these genes in the scatterplots of indirect versus direct logratios (Figure 8B), we see that we can identify highly discrepant genes with high sensitivity and specificity. We also examined the array images for these identified genes. In each case we could identify a single spot that was contaminated with dust or otherwise corrupted. These are exactly the kinds of datapoints one wishes to exclude from any analysis.

We contrasted our proposed filter to the five quality control variables provided by the Feature Extraction software. These variables indicate spots for which (1) the within-spot pixel distribution deviates substantially from uniform; (2) the signal has reached saturation; (3) the background pixel distribution deviates substantially from uniform; (4) the background measurement is a population outlier; (5) the spot intensity is a population outlier. The last of these had no flags in our dataset and so was uninformative. Of the remaining four, the only variable that showed any ability to identify highly discrepant genes was (1). Figure 8C shows that this flag reliably identifies genes with highly discrepant results. However, Figure 8C also shows that the specificity of the indicator is quite poor, as a large number of genes with concordant measurements between the direct and indirect logratios are also flagged. It is possible this indicator could be tuned to improve specificity while maintaining its high sensitivity, but there is no tuning option in the software. (See supplementary figures 5-13 for the corresponding results on data filtering for other test pairs and references.)

König et al (Konig, R. *et al.*, 2004) also considered different quality filters. They advocate using the reproducibility of measurements on replicate arrays as a filter. While this is a sensible choice, it requires technical replicates. Most investigators have a limited budget for microarrays and including technical replicates means reducing the number of biological replicates, which is highly undesirable (Kerr, M. K., 2003b). In the absence of technical replicates, König et al (Konig, R. *et al.*, 2004) advocate flagging spots based on the difference between the logratios acquired using mean spot intensities and the median spot intensities. The rationale is that spots with a highly non-symmetric distribution of pixel intensities are likely to be corrupted, and also likely to give a median spot intensity that is far from the mean spot intensity. We flagged spots based on the full range of possible thresholds using König et al's proposed metric, but this method did not approach the accuracy of our proposed filter.

Comments on low-intensity genes. Figure 9 highlights low-intensity genes in the assays for the lung test-pair. The vast majority of these genes appear around the origin in the scatterplot. They are measured as not differentially expressed, and this measurement is reproducible between the direct and indirect logratios. Notice, however, that some of these low-intensity genes appear to be differentially expressed, as measured concordantly by the direct and indirect logratios. Finally, notice that most of the highly discrepant genes are *not* highlighted in Figure 9.

Certainly, it is likely that most of these low-intensity genes are probably not expressed at all, and hence are not differentially expressed. However, the results displayed in Figure 9 have important implications for the practice of discarding low-intensity genes.

Specifically, discarding low-intensity genes may be a rather ineffective filter, since (1) it does not necessarily remove genes with problem measurements, and (2) it can remove potentially interesting differentially-expressed genes.

The scatterplots in Figure 2-4 have a "bulge" around the origin. As shown in Figure 9, low intensity genes tend to appear in this region of the scatterplot. These observations appear to corroborate the conventional wisdom that measurements on low-intensity genes are unreliable, and indirectly corroborate the assumption, which we have challenged, that a good reference RNA will minimize the number of low-intensity spots. However, some care is required in interpreting such plots. First, there are many more points in the middle of the scatterplots, so we expect the total spread there to be bigger. In fact, Figures 6 and 7 show that, on an absolute scale, the genes in the middle of the scatterplots actually give *more* consistent results than the genes at the extremes of the scatterplots.

Figure 10 shows the relationship between the magnitude of the discrepancy between direct and indirect logratios, intensity, and the size of the direct logratio θ. At lower intensities, the magnitude of error rises more quickly with θ than at high intensities. On the other hand, there are more large θ at higher intensities. Said differently, large measurements of differential expression are less reliable for low-intensity genes; however, a lower proportion of low-intensity genes exhibit large changes in expression. The net result is that the average discrepancy is about the same across intensity levels.

Our conclusion is that low-intensity genes are indeed less reliable, but not unreliable. Our data support the statement that "unreliable genes are low-intensity," but not the statement that "low-intensity genes are unreliable."

DISCUSSION

We evaluated several methods for data filtering. In experimental designs that include technical replicates, the distribution of replicates is an effective way to assess data quality and the data can potentially be filtered based on agreement among replicates (Konig, R. *et al.*, 2004). However, not all designs contain such replicates and it is highly desirable to have methods to assess data quality that do not require them. The flags provided by the image analysis software were either ineffective at identifying problem spots, or did not have satisfactory specificity in identifying problem spots. Contrary to conventional wisdom, most low-intensity measurements are reproducible, so simply discarding low intensity genes is an excessively crude filter. We proposed a new indicator to identify genes with suspect measurements. Our filter uses information across arrays rather than relying on spot characteristics or other within-array quantities. Our results suggest this filter is both highly sensitive and highly specific. The proposed filter applies to any experimental design and in particular does not require (technical) replicate arrays. We envision the filter could be used either in an automated fashion to exclude suspect measurements, or to simply identify genes whose measurements need to be manually evaluated for contamination.

A question of great interest to researchers using two-color microarrays is what to use as a reference RNA in a reference design. In this study we evaluated three common choices. On the whole, the data support our premise that a good reference RNA should be similar to the RNAs of interest. However, perhaps the more impressive result is that in two out of three instances, all three reference RNAs performed comparably. These results suggest that practical considerations may be more important than technical considerations in choosing a reference RNA for a microarray study.

We found no advantage to the commercial RNA in our evaluation. The better-than-expected performance of the commercial reference for the lung test-pair is offset by the unexplained poor performance for the kidney test-pair. At the very least, these results support our position that there is no such thing as a “universally best” reference.

One advantage suggested for commercial reference RNAs is that they can facilitate cross-laboratory collaboration if every lab uses the reference design with the same reference RNA. We do not entirely agree with the merit of this argument. First, we think investigators should design microarray experiments to get the best possible data for answering their questions of interest. Designing experiments for hypothetical unplanned collaborations should not be the primary concern. For planned collaborations, investigators may indeed wish to use a common reference RNA, but this need not be a commercially-purchased RNA. A “pooling” strategy could be an economical and effective choice for many studies. Of course, there are always practical considerations that are study-specific. For example, a pooling strategy might make it more difficult to add unanticipated additional samples to a study after the initial hybridizations are completed.

We used the concordance between indirect and direct logratios to evaluate reference RNAs. Figures 2-5 illustrate that this concordance is quite good overall. A disadvantage of this evaluation is that if direct logratios are biased, then we have merely identified which reference RNA reproduces those biases. We offer three counterpoints to this argument. First, as mentioned, references that reproduce direct logratios give study results that are not “reference-specific.” If, instead, results are reference-specific, then they may be inherently non-reproducible because any reference is finite. This violates a fundamental tenet of scientific research. Second, the alternative would be to identify a reference RNA that cancels out any biases in direct logratios. No such claim has ever been made for the existence of such a reference and it seems improbable that one would exist. In other words, it is unlikely that one could ever do better than direct comparisons. Third, it is important to consider the nature of bias in array measurements and whether the bias has any scientific importance.

Elaborating on this third point, under what conditions would one expect indirect and direct logratios to agree? Previous work (Dudley, A. M. *et al.*, 2002;Qin, L. X. *et al.*, 2004;Yuen, T. *et al.*, 2002;Tong, W. D. *et al.*, 2006) has shown that estimates of relative expression from microarrays tend to be attenuated compared to true log-ratios. Some systematic studies (Shi, L. M. *et al.*, 2005;Yuen, T. *et al.*, 2002) further suggest that logratios from arrays are proportional to true logratios:

$\text{observed logratio}_{\text{A vs. B}} = c \times \text{true logratio}_{\text{A vs. B}}$, where c is a positive constant less than 1 (see the proposed model in (Yuen, T. *et al.*, 2002) and figures 5a and 5b in (Shi, L. M. *et* *al.*, 2005)). The constant c does not appear to vary with intensity. Our results are entirely consistent with such a model.

What is the scientific impact of attenuated logratios? Certainly, the answer depends on the goal of a particular study. The goal of many microarray studies is to identify differentially expressed genes. Using spike-in data, Qin et al (Qin, L. X. *et al.*, 2004) showed that much more accurate identification of differentially expressed genes could be made using non-background-adjusted logratios, which have larger attenuation than background-adjusted logratios. Although the background-adjusted logratios have less bias, this is greatly offset by a drastic increase in variability that impedes identification of differentially expressed genes. Other microarray studies are conducted to obtain estimates of relative expression for use in "higher order" analyses such as supervised and unsupervised clustering. In our experience, investigators prefer reliable biased estimates of relative expression to highly variable estimates with less bias. Our data show that the bias in non-background-adjusted logratios is perhaps only slightly larger than the bias in background-adjusted logratios, while the variability in the latter is much higher (see supplemental Figure 1). These results are entirely consistent with previous findings (Qin, L. X. *et al.*, 2004) and (Members of the Toxicogenomics Research Consortium, 2005a). Of course, our evaluation only considered one method of background-adjustment, which was simple background-subtraction. Certainly, more sophisticated methods of background-adjustment may perform better, although none is widely used to our knowledge. Our results also indicate that simple dye-swap averaging is not sufficient to normalize microarray data.

The important problem of assessing and assuring data quality in microarray experiments is being addressed from many angles. One important effort is the External RNA Controls Consortium (Baker, S. C. *et al.*, 2005;The External RNA Controls Consortium, 2005) (ERCC). The ERCC is a community-wide effort to generate a well-characterized set of approximately 100 RNA transcripts for use as external controls in microarray experiments. The primary purpose of these controls is to provide a means to evaluate the performance of gene expression assays, including microarrays. Such controls are extremely important and the products of the ERCC will be valuable to the research community. Note, however, that for methodological validation spike-in studies have disadvantages as well as advantages (Mehta, T. *et al.*, 2004;Qin, L. X. *et al.*, 2004). Thorough validation requires a plurality of approaches.

More interestingly, it is possible that methods for normalizing array data could be developed that are based on controls like the ERCC is developing (The External RNA Controls Consortium, 2005). Such methods could require weaker assumptions than normalization methods currently in use. If so, then it may no longer be a priority to use reference RNAs that are similar to the RNAs in a study. However, given the current state-of-the-art in microarrays, we believe this quality of a suitable reference should be an important consideration in choosing a reference RNA.

Acknowledgment

We thank Dan Bowen-Pope for providing tissue samples. This research was supported by a grant from the National Heart Lung and Blood Institute # HL072370; Public Health Services Grants from the National Institute for Environmental Health Sciences, Grant # NIEHS U19ES011387 to the FHCRC/UW Toxicogenomics Research Consortium; and Grant # NIEHS P30ES07033 to the UW Center for Ecogenetics and Environmental Health.

Reference List

Baker, S. C., Bauer, S. R., Beyer, R. P., Brenton, J. D., Bromley, B., Burrill, J., Causton, H., Conley, M. P., Elespuru, R., Fero, M., Foy, C., Fuscoe, J., Gao, X., Gerhold, D. L., Gilles, P., Goodsaid, F., Guo, X., Hackett, J., Hockett, R. D., Ikonomi, P., Irizarry, R. A., Kawasaki, E. S., Kaysser-Kranich, T., Kerr, K., Kiser, G., Koch, W. H., Lee, K. Y., Liu, C., Liu, Z. L., Lucas, A., Manohar, C. F., Miyada, G., Modrusan, Z., Parkes, H., Puri, R. K., Reid, L., Ryder, T. B., Salit, M., Samaha, R. R., Scherf, U., Sendera, T. J., Setterquist, R. A., Shi, L., Shippy, R., Soriano, J. V., Wagar, E. A., Warrington, J. A., Williams, M., Wilmer, F., Wilson, M., Wolber, P. K., Wu, X., and Zadro, R. (2005). The External RNA Controls Consortium: a progress report. Nat.Methods **2:** 731-734.

Cui, X., Kerr, M. K., and Churchill, G. A. (2003). Transformations for cDNA Microarray Data. Statistical Applications in Genetics and Molecular Biology **2**.

Dudley, A. M., Aach, J., Steffen, M. A., and Church, G. M. (5-28-2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proceedings of the National Academy of Sciences of the United States of America **99:** 7554-7559.

Gorreta, F., Barzaghi, D., VanMeter, A. J., Chandhoke, V., and Del Giacco, L. (2004). Development of a new reference standard for microarray experiments. Biotechniques **36:** 1002-1009.

He, X. R., Zhang, C. L., and Patterson, C. (2004). Universal mouse reference RNA derived from neonatal mice. Biotechniques **37:** 464-468.

Kerr, M. K. (2003a). Design considerations for efficient and effective microarray studies. Biometrics **59:** 822-828.

Kerr, M. K. (2003b). Linear models for microarray data analysis: Hidden similarities and differences. Journal of Computational Biology **10:** 891-901.

Kerr, M. K. and Churchill, G. A. (2001). Experimental design for gene expression microarrays. Biostatistics. **2:** 183-201.

Konig, R., Baldessari, D., Pollet, N., Niehrs, C., and Eils, R. (2004). Reliability of gene expression ratios for cDNA microarrays in multiconditional experiments with a reference design. Nucleic Acids Research **32:** e29.

Lin, L. I. (1989). A Concordance Correlation-Coefficient to Evaluate Reproducibility. Biometrics **45:** 255-268.

Mehta, T., Tanik, M., and Allison, D. B. (2004). Towards sound epistemological foundations of statistical methods for high-dimensional biology. Nat.Genet. **36:** 943-947.

Members of the Toxicogenomics Research Consortium (2005a). Standardizing global gene expression analysis between laboratories and across platforms. Nature Methods **2:** 351-356.

Members of the Toxicogenomics Research Consortium (2005b). Standardizing global gene expression analysis between laboratories and across platforms (vol 2, pg 351, 2005). Nature Methods **2:** 477.

Novoradovskaya, N., Whitfield, M. L., Basehore, L. S., Novoradovsky, A., Pesich, R., Usary, J., Karaca, M., Wong, W. K., Aprelikova, O., Fero, M., Perou, C. M., Botstein, D., and Braman, J. (3-9-2004). Universal Reference RNA as a standard for microarray experiments. Bmc Genomics **5**.

Qin, L. X., Kerr, K. F., and Contributing Members of the Toxicogenomics Research Consortium (2004). Empirical evaluation of data transformations and ranking statistics for microarray analysis. Nucleic Acids Research **32:** 5471-5479.

Quackenbush, J. (2002). Microarray data normalization and transformation. Nat.Genet. **32 Suppl:** 496-501.

Shi, L. M., Tong, W. D., Su, Z. Q., Han, T., Han, J., Puri, R. K., Fang, H., Frueh, F. W., Goodsaid, F. M., Guo, L., Branham, W. S., Chen, J. J., Xu, Z. A., Harris, S. C., Hong, H. X., Xie, Q., Perkins, R. G., and Fuscoe, J. C. (7-15-2005). Microarray scanner calibration curves: characteristics and implications. Bmc Bioinformatics **6**.

The External RNA Controls Consortium (2005). Proposed methods for testing and selecting the ERCC external RNA controls. Bmc Genomics **6:** 150.

Tong, W. D., Lucas, A. B., Shippy, R., Fan, X. H., Fang, H., Hong, H. X., Orr, M. S., Chu, T. M., Guo, X., Collins, P. J., Sun, Y. M. A., Wang, S. J., Bao, W. J., Wolfinger, R. D., Shchegrova, S., Guo, L., Warrington, J. A., and Shi, L. M. (2006). Evaluation of external RNA controls for the assessment of microarray performance. Nature Biotechnology **24:** 1132-1139.

Yang, I. V., Chen, E., Hasseman, J. P., Liang, W., Frank, B. C., Wang, S., Sharov, V., Saeed, A. I., White, J., Li, J., Lee, N. H., Yeatman, T. J., and Quackenbush, J. (10-24-2002). Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol. **3:** research0062.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2-15-2002a). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research **30:** e15.

Yang, Y. H. and Speed, T. (2002b). Design issues for cDNA microarray experiments. Nat.Rev.Genet. **3:** 579-588.

Yuen, T., Wurmbach, E., Pfeffer, R. L., Ebersole, B. J., and Sealfon, S. C. (5-15-2002). Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. Nucleic Acids Research **30**.

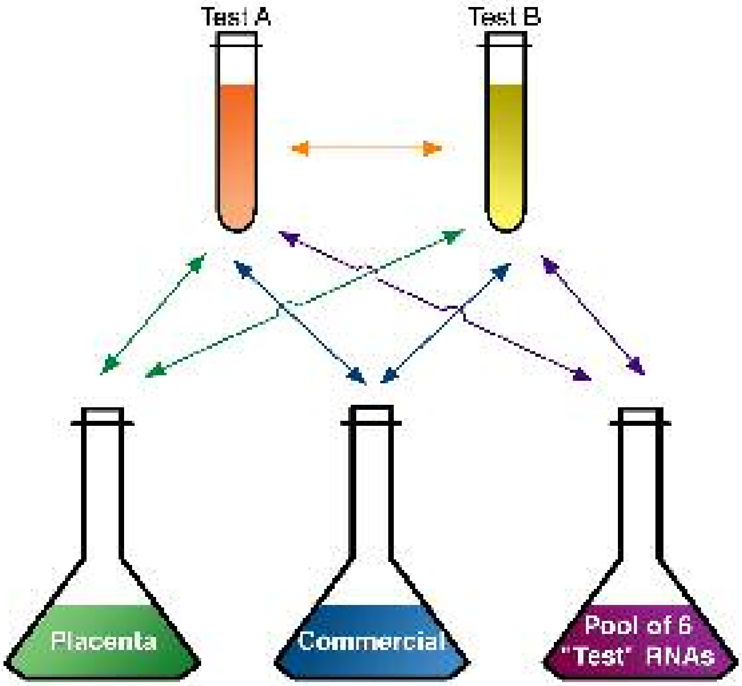

Figure 1. Experimental Design for one test-pair of RNAs. The reference RNAs that were evaluated in this study, Placenta, Commercial, and Pooled Reference, are shown at the bottom of the figure. Each double-headed arrow represents a pair of arrays on which the indicated RNAs are co-hybridized in a dye-swap arrangement. The diagram shows the hybridizations that were performed for a "test" pair of RNAs: each test pair was compared directly on two arrays, and each test RNA was co-hybridized with each reference on two arrays. This experimental design was executed for three test-pairs: a kidney test pair, a placenta test-pair, and a lung test-pair.

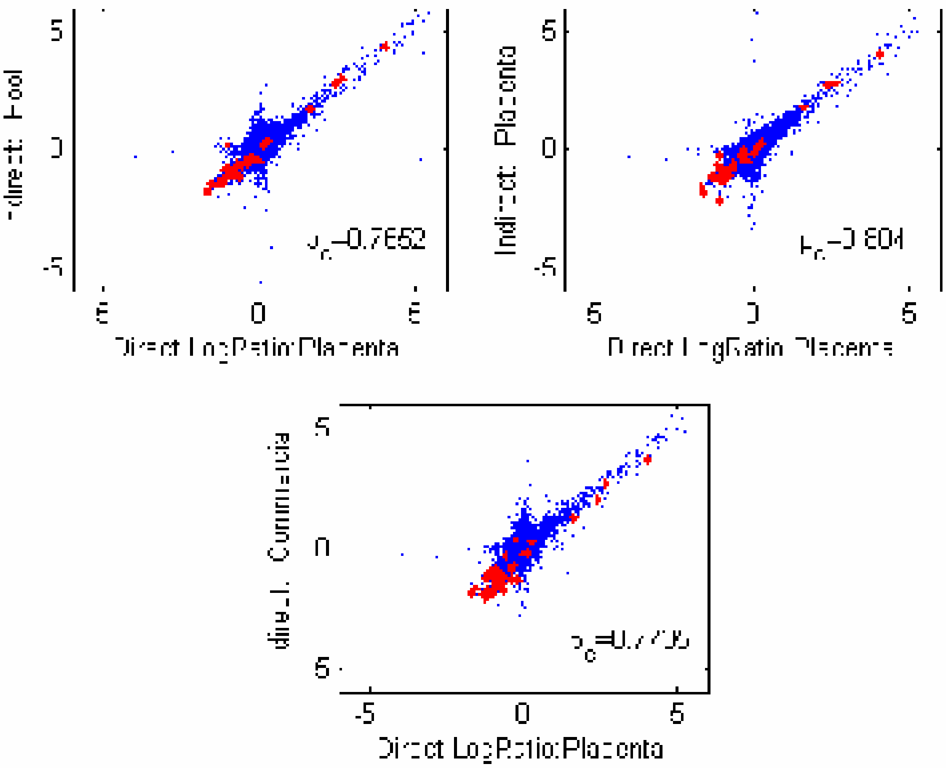

Figure 2. Comparison of the direct $\log_2$ratios to the indirect $\log_2$ratios for the placenta test-pair. Lin's correlation coefficient summarizes the agreement across genes. The level of agreement is similar for all three reference RNAs but highest for the placenta reference. The genes highlighted in red are placenta-specific genes.

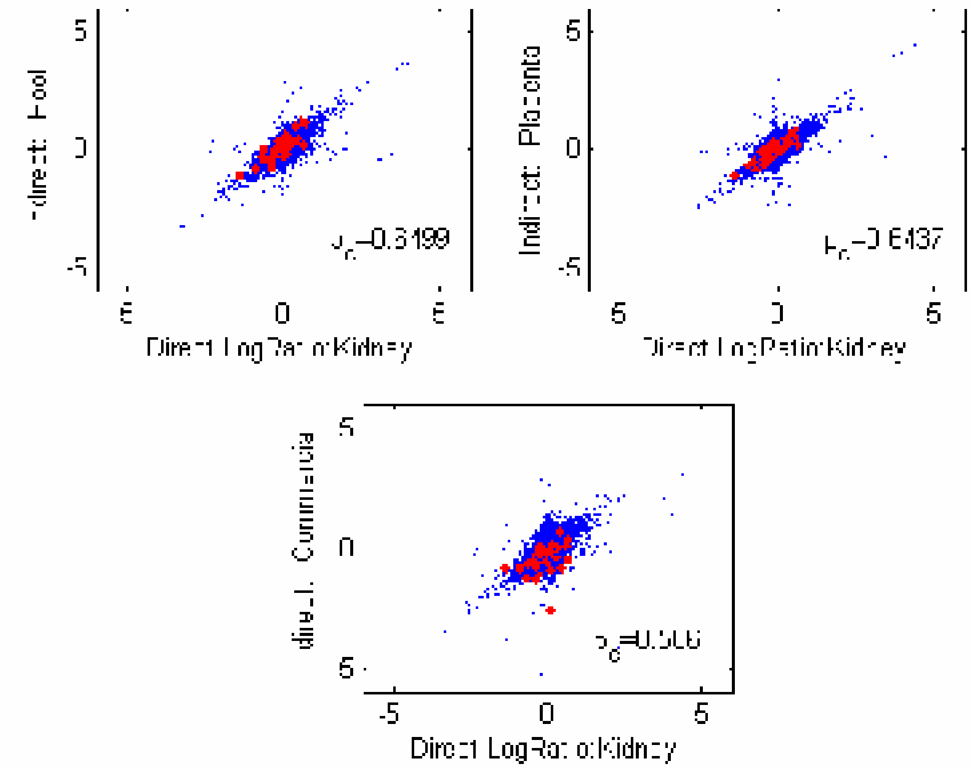

Figure 3. Comparison of the direct $\log_2$ratios and the indirect $\log_2$ratios for the kidney test-pair. The level of agreement is similar for the pool and placenta reference RNAs and noticeably worse for the commercial reference RNA. The genes highlighted in red are kidney-specific genes.

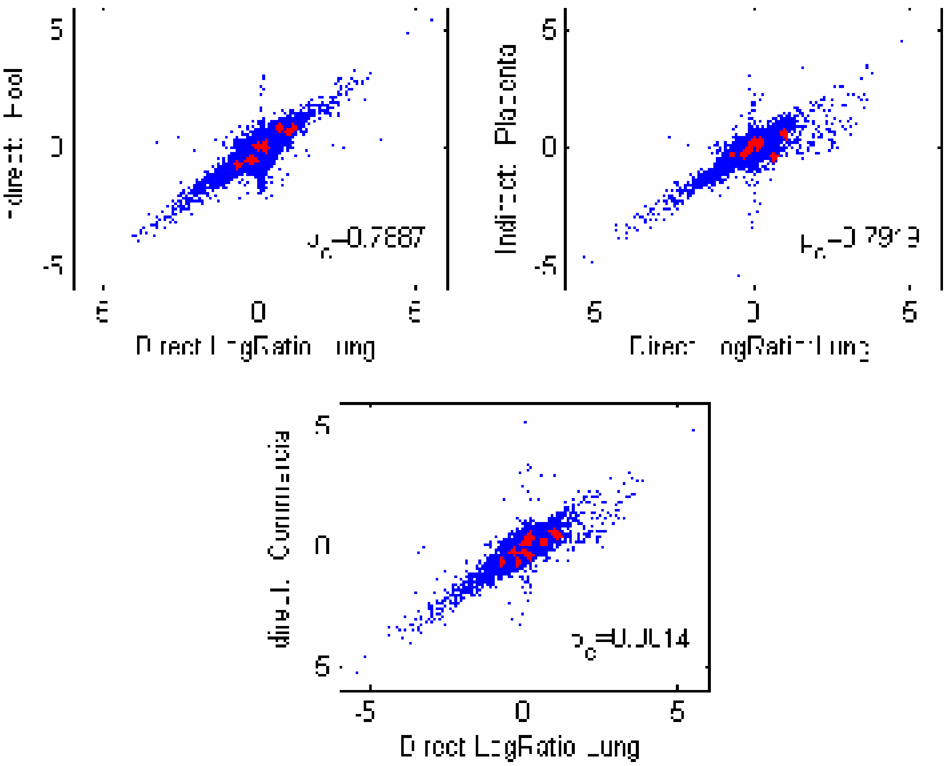

Figure 4. Comparison of the direct $\log_2$ratios to the indirect $\log_2$ratios for the lung test-pair. The level of agreement is similar for all three reference RNAs. The genes highlighted in red are lung-specific genes.

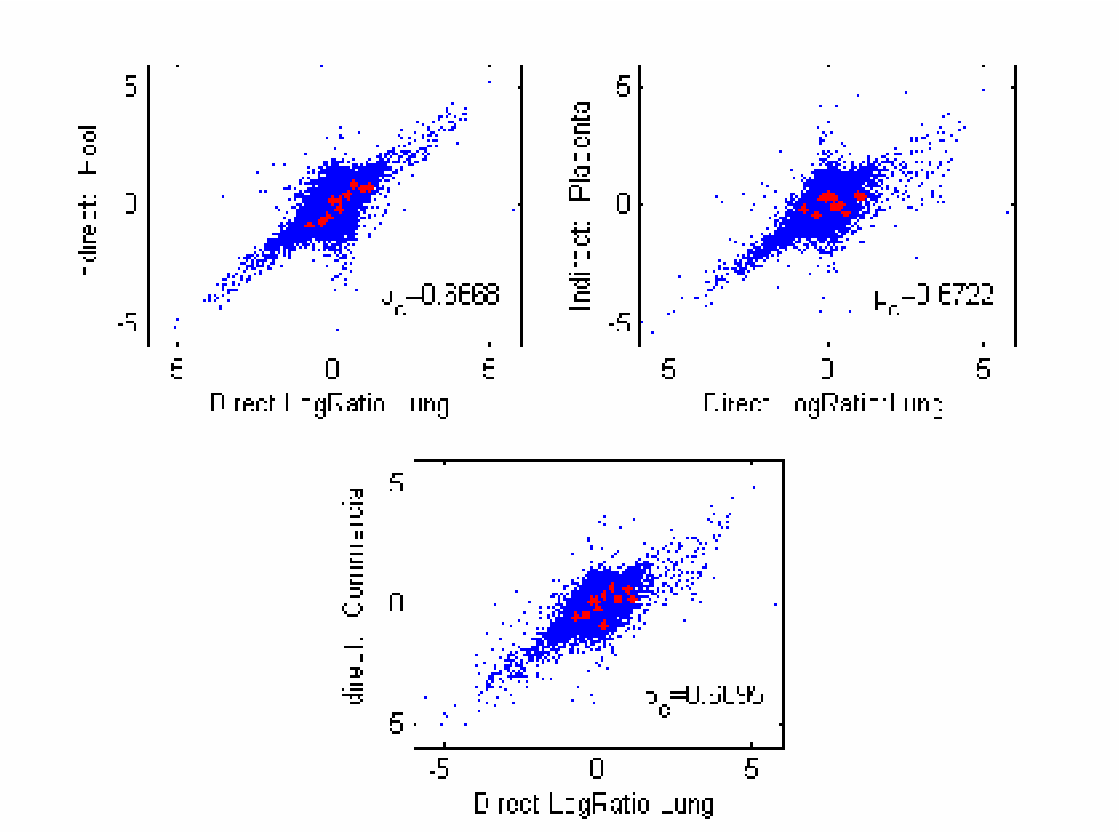

Figure 5. The effect of background subtraction. In this figure we re-plot the data in Figure 4 using background-subtracted intensities. Compared to Figure 4, the variability in the scatterplots is increased and the overall agreement between the direct and indirect logratios is decreased.

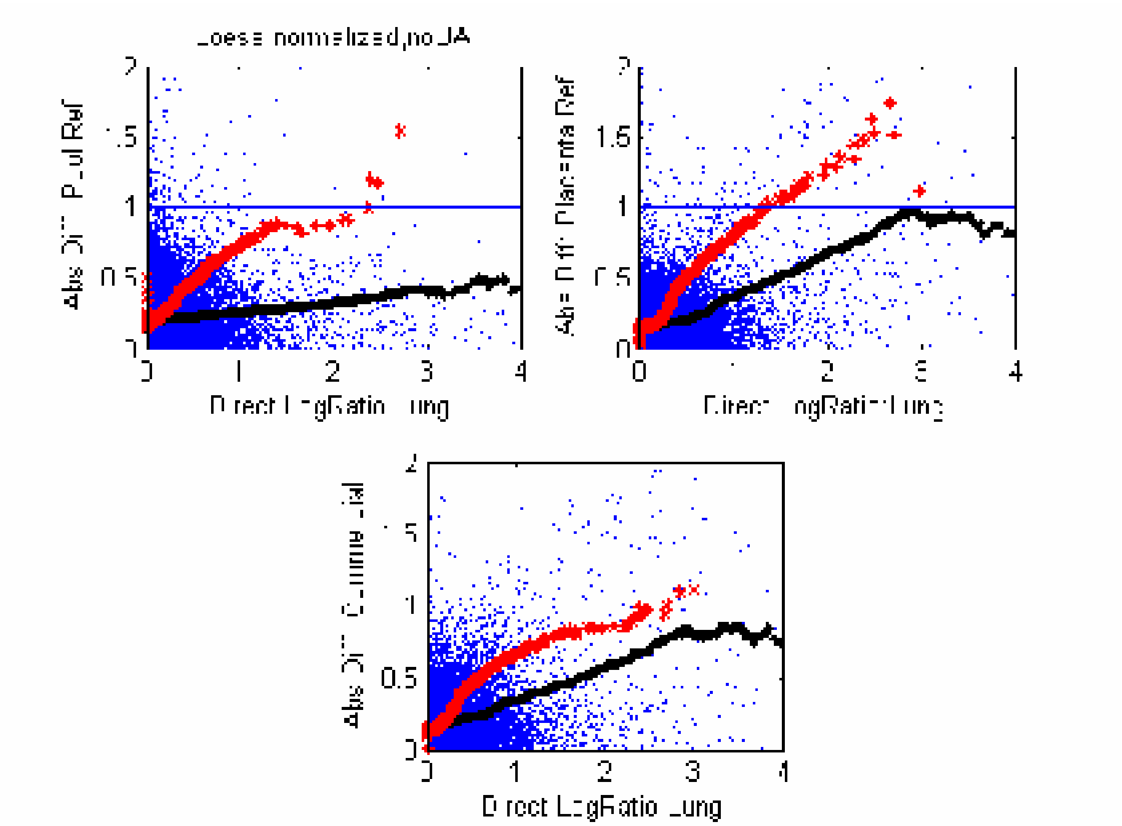

Figure 6. Average discrepancy between direct and indirect logratios for the lung test pair data without background subtraction with intensity-normalization. For the lung test-pair, this plot shows the same data as Figure 4 in a different way. The horizontal axis is the absolute value of the logratio as computed from the dye-swap between the test RNAs. The vertical axis is the absolute value of the difference between this direct logratio and the logratio from an indirect logratio using a reference RNA. The black points are a 10% moving average. The red points are the same moving averaging, but using only the low-intensity genes. On an absolute scale, the average discrepancy increases with the size of the logratio, and the increase is faster for low-intensity genes.

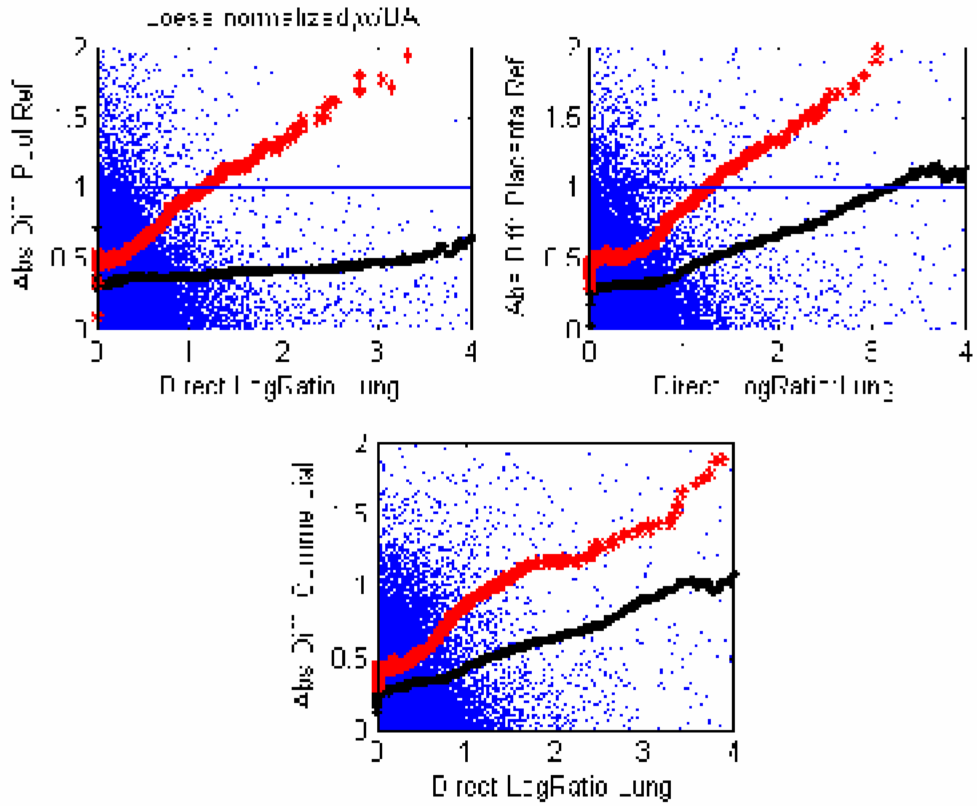

Figure 7. Average discrepancy between direct and indirect logratios for the lung test pair data with background subtraction and intensity-normalization. This is the corresponding plot to Figure 6 for the data with background subtraction. As with the data without background-subtraction, on an absolute scale, the average discrepancy between direct and indirect logratios increases with the size of the logratio, and the increase is faster for low-intensity genes. Comparing this figure to Figure 6, notice that the average discrepancy is larger for this version of the data, dramatically so for low-intensity genes.

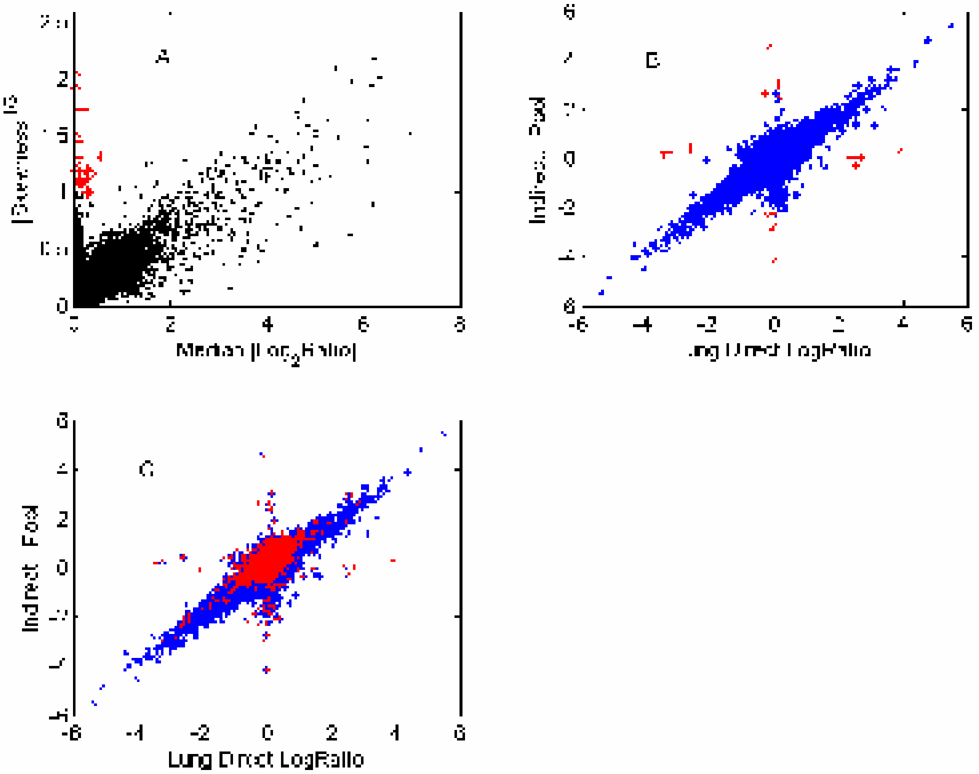

Figure 8. **A**. Absolute skewness of the logratios plotted against the median absolute logratio for the lung test pair and the pool reference. Genes with suspect measurements are those with large skewness but small median absolute logratio. The genes highlighted in red have skewness>1 and median absolute log$_2$ratio<0.9; these are the same genes highlighted in plot B. **B**: Effectiveness of the proposed filter for identifying discrepant data. **C**. Performance of the non-uniformity flag from the Feature Extraction software. Genes plotted in red are those for which Feature Extraction flagged one or more spots on the six arrays contributing to the scatterplot for non-uniformity of pixels. This flag detects most of the genes with discrepant results between the direct and indirect comparisons, but flags many more genes with consistent measurements. The flag has good sensitivity but poor specificity.

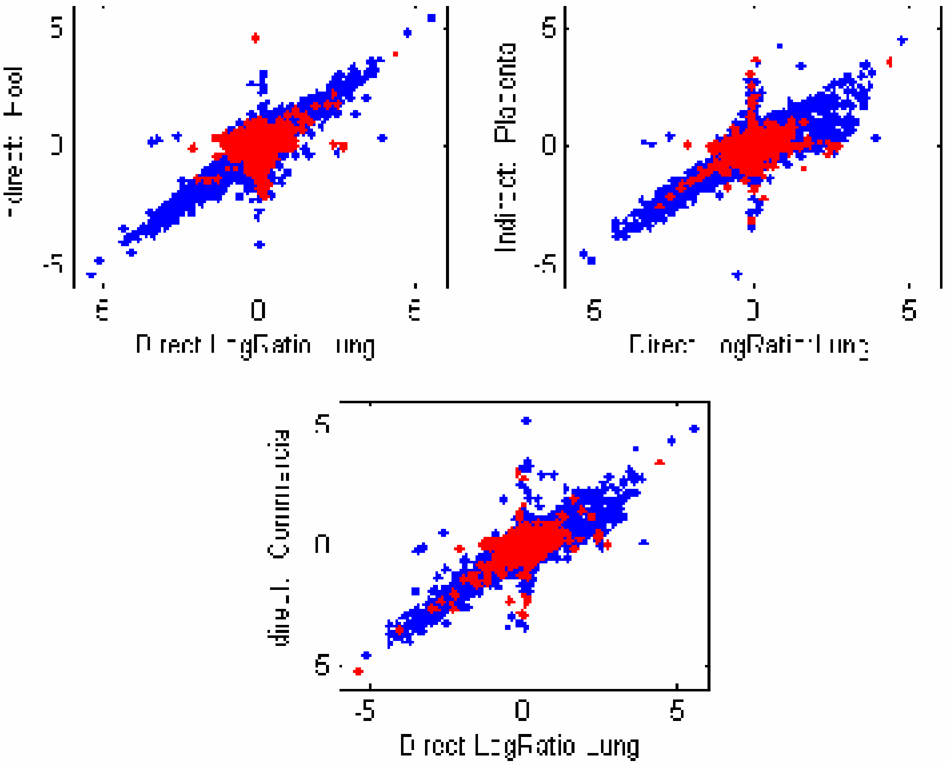

Figure 9. Consistency of results for low-intensity genes. For the lung test pair, a gene is represented in red if its normalized log-intensity ranks in the lowest 10% of all intensities for one or more arrays involved in the specific scatterplot. There are 7795, 7936, and 8013 genes highlighted on the plots, with most highlighted genes clustered around the origin. Note that the direct and indirect logratios are in good agreement for the vast majority of low-intensity genes. Also, some low-intensity genes appear to be differentially expressed as measured concordantly by both the direct and indirect logratios. A final important note is that most of the genes with the largest discrepancy between the direct and indirect logratio are NOT among the low-intensity genes.

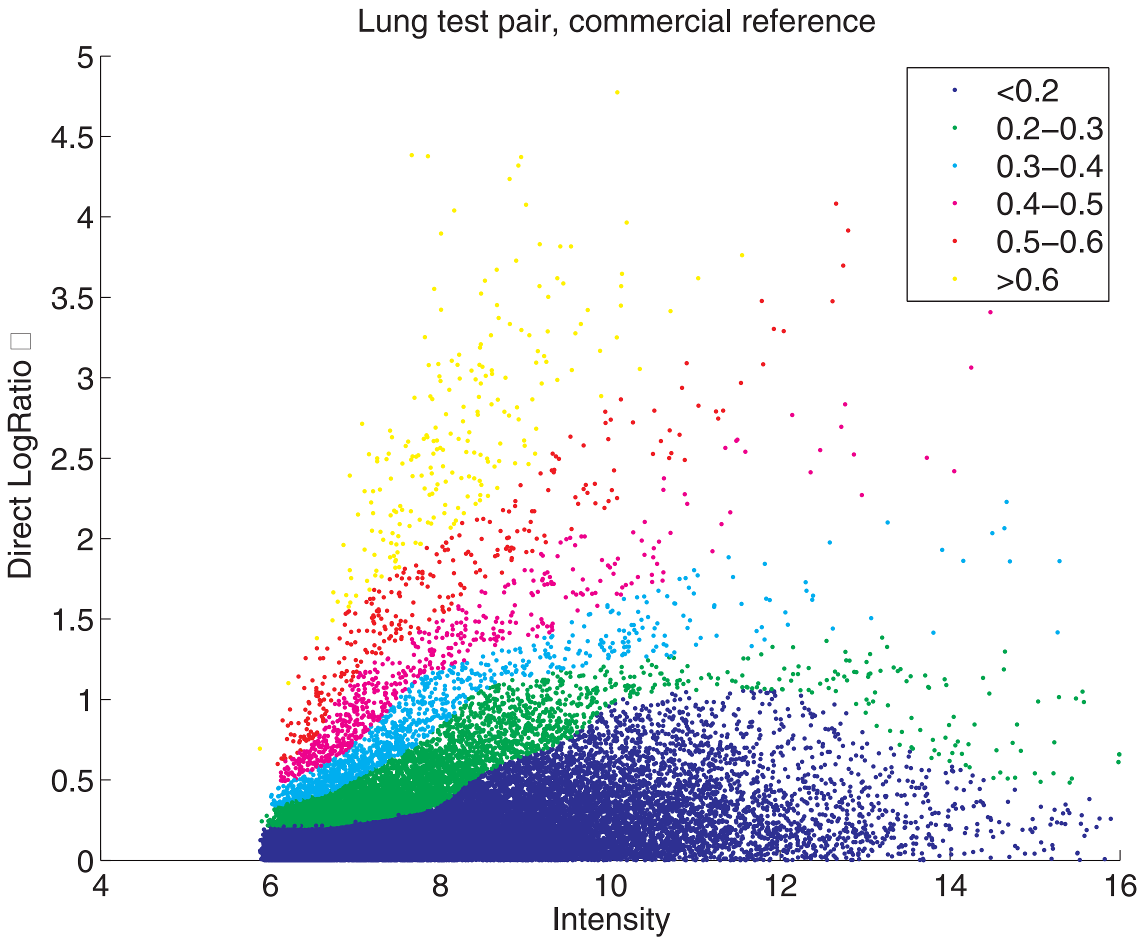

Figure 10. Discrepancy between direct and indirect logratios as related to the direct logratio θ and spot intensity. Each point represents one gene and the color of a point represents the size of the discrepancy.