Technical vis-à-vis biological variation in gene expression measurements

Walter Liggett
National Institute of Standards and Technology

Ron Peterson
Novartis Institutes for Biomedical Research

Marc Salit
National Institute of Standards and Technology

Abstract

Background:
In characterizing the technical variation in gene expression measurements, one important aspect is comparison with expression differences of potential interest in a biological study. Design of an experiment to investigate this aspect involves choice of samples that are to represent the biological variability.

Results:
In the experiment described here, replicate measurements of gene expression in a group of similarly treated animals provide data for comparison of technical and biological variation. We consider twelve Affymetrix RAE230 2.0 microarray measurements on each of six genetically diverse rats (*Rattus norvegicus*) that were subject to the same control-group treatment. The twelve measurements encompass RNA from the liver, from the kidney, and from two mixtures of both RNAs. Our analysis of these measurements illustrates how one might perform the technical-biological comparison. Under the circumstances of this experiment, the technical variation generally seems small enough to allow the biological variation to be appreciated. However, the sets of twelve measurements are so much alike that the question of whether the biological variation is greater than the technical variation depends on one's perspective. From the perspective of a single-gene, the comparison can go either way. As shown by our data analysis, the technical variation has various aspects including an array that gave results that had to be discarded, some instances of probe saturation, and batch effects related to choice of scanner and fluidics machine. We conclude that, as implemented, the measurement technique would generally be adequate for differentiating animals such as these. However, there is no reason to believe that technical variation can be safely ignored in designing experiments with similar animals.

Conclusions:
Although the question of whether a gene expression measurement system is adequate for a particular purpose is complicated by the high-dimensional nature of microarray measurements, experiments can be performed that provide an answer that is broadly useful if not comprehensive.

Background

Efforts to improve gene expression measurements and the accompanying data analysis methods occupy many researchers [1, 2]. Many efforts involve development of data analysis methods for achieving specific goals in biological research. Others, such as the one reported in this paper, are directed toward methods for overcoming technical variation in gene expression measurements. In general terms, one can speak of biological and technical as a dichotomy in the sources that cause variation in a collection of expression measurements. Biological sources cause variation in mRNA being measured. The effects of technical sources, the technical variation, are observed when the mRNA does not change from measurement to measurement. Thus, technical variation and reproducibility convey the same idea, although reproducibility may also carry an implication of a performance metric.

Among efforts to overcome technical variation, many are studies in which two or more alternative processing methods are compared in terms of their relative performance [3, 4]. A familiar example is comparison of normalization methods [5]. Beyond studies of alternative processing methods are general studies of technical variation [6-8]. An important goal of such studies is identification of sources of technical variation so that decisions can be made about the need for reductions of these sources. General characterization of the technical variation is the purpose of this paper.

In 2004, David Rocke [9] wrote in reference to gene expression measurement, "Most of the variation in any biological experiment or clinical trial is biological. Usually, the largest differences are between organisms, though there are cases in which such factors as diurnal variation can make the within-organism variability larger than the between-organism variability." Statements like this may be responsible for the commonly held attitude that the technical variation in gene expression measurements is a secondary consideration. For example, the MAQC consortium [10] adopted the paper title, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." However, this consortium analyzed measurements from a design with no challenging biological variation and thus with no realistic biological context.

This paper demonstrates how technical variation can be put in the context of experimentally meaningful biological variation. In the experiment discussed here, replicate measurements were made on RNA from six rats that differed genetically but were subjected to similar experimental treatment as part of the control group of a previous study. For each rat, RNA was extracted from the liver and the kidney [11]. The liver RNA sample, the kidney RNA sample as well as two mixtures of the liver and kidney RNA samples were measured in triplicate [12]. This experimental design can be thought of as a six-fold biological replication of the design used in the first phase of the MAQC project [10]. The measurements were made with the Affymetrix RAE230 2.0 microarray. If this measurement system is capable of distinguishing these rats, then this experiment establishes its suitability for many related biological studies.

The iconic issue in metrology is the suitability of a measurement system for an intended purpose. In the case of gene expression measurement, the suitability issue is more complicated than in the case univariate measurements. In the univariate case, the sizes of differences among the objects of interest can be compared to the measurement uncertainty thereby resolving the suitability issue [13]. In the case of univariate measurements, it is moreover true that researchers have in their minds an idea of the uncertainty of familiar measurements. This paradigm applies one gene at a time to gene expression measurements but does not apply generally because researchers in gene expression generally cannot conceptualize with any confidence the collection of gene expression differences that they might find meaningful in an experiment. There are, of course, statistical summaries of collections of gene expression differences, but these may not be useful to someone thinking about the biological goals of an experiment [14]. Thus, the suitability issue is more complicated in the case of high dimensional measurements.

Microarray users are, of course, familiar with the occurrence of defective results from a particular array, but the needed characterization of technical variation extends beyond the identification of defective results [15]. The idea of monitoring a gene expression measurement system is familiar if not often implemented. The value in such monitoring involves characterization of the system variability. What an appropriate characterization might look like can be garnered from the data analysis results presented in this paper. At each monitoring time, one might measure liver mRNA, kidney mRNA, and the two mixtures all in triplicate as was done for each animal in the experiment described here. Of course, there may be alternatives to this set of 12 microarray measurements. We do not present a prescription for gene expression monitoring but ideas on what such a prescription might be.

There is one more aspect to the design of an experiment the purpose of which is comparison of biological variability with technical variation. We have decided to associate biological variability with rat-to-rat differences in the mRNA levels found in their livers and kidneys. To what do we associate technical variation? There is, of course, always the repeatability, but this may be small compared to some batch effects. A familiar type of batch effect is inter-laboratory difference, which was studied in the first phase of the MAQC project [10]. To proceed with the comparison, we must specify batch effects that we want to manifest themselves in the measurement replication [16, 17]. In the experiment discussed here, we do not include inter-laboratory differences but do include the effects of changing scanner and changing fluidics machine. Moreover, we consider the effects of location on a 96-well plate.

Results and discussion

Drawing general conclusions from the data analysis results in this paper requires some familiarity with particulars of the experiment and the data analysis methods employed. We begin this section with two crucial aspects of the experiment, the materials measured and the normalization applied to the microarray intensities.

Materials

The RNA measured was extracted from animals with the following characteristics:

Animal	*Rattus norvegicus*
Strain	Sprague Dawley Crl:CD(SD)
Age	7-8 Weeks
Sex	Male
Body weight range	150 to 300 g
Treatment	20% propylene glycol/80% lactic acid containing 4.3% mannitol, pH 4.0
Administration	Intravenous
Duration	Once per week for 13 weeks

For each animal, the materials measured consisted of RNA extracted from the liver (designated material A), RNA extracted from the kidney (designated material D), and two mixtures of the two RNAs (designated materials B and C). Because only part of the RNA is measured, namely, the mRNA, the mixtures are properly described by the liver mRNA as a fraction of the total mRNA in the mixture [18]. Moreover, because the concentration of mRNA in the total RNA differs between the liver and the kidney, the mRNA fractions differ from the total RNA fractions on which the actual mixing was based. Material B was formed by mixing 0.75 liver RNA with 0.25 kidney RNA. Let the concentration of mRNA in the animal j liver RNA be denoted c_{Aj}. Similarly, let the concentration of mRNA in the animal j kidney RNA be denoted c_{Dj}. The mRNA liver fraction for material B is

$$\frac{0.75}{0.75 + 0.25 c_{Dj} / c_{Aj}}.$$

Material C was formed by mixing 0.25 liver RNA with 0.75 kidney RNA. The mRNA liver fraction for material C is

$$\frac{0.25}{0.25 + 0.75 c_{Dj} / c_{Aj}}.$$

For each animal, we estimated c_{Dj} / c_{Aj} as a step in the normalization.

For each animal, each material was measured three times with Affymetrix Rat230_2 microarrays. Thus, the data set considered here consists of 72 gene expression measurements. A goal of this paper is summarization of these measurements.

The three replicates provide an opportunity to look for discrepant results. For each set of three replicates, we extracted the intensities for the perfect match probes and plotted these one versus the other. The results for animal 3, material B are shown in Figure 1. We see that the results for the third replicate are discrepant. We exclude the third replicate from

the remainder of our data analysis. We offer a two part summary of the 72 measurements: the conclusion that something went wrong with the third replicate on material B of animal 3 and the analysis of the other 71 measurements that follows.

Normalization

The need for normalization of scanner intensities is so firmly established that a comparison of biological variability and technical variation in un-normalized intensities would be of no interest. Normalization, however, reduces both types of variation. For this reason, discussion of the normalization applied must precede discussion of the technical-biological comparison.

We note first that our analysis is of the intensities from the perfect match probes. We ignore the intensities from the mismatch probes. Although the mismatch probes may be useful in some contexts, the perfect match probes alone provide a reasonable basis for comparing the two types of variation [19].

Our normalization method consists of two steps. The first step is a version of the familiar global normalization, which consists of scaling of the intensities separately for each array [20]. For each array, the scaling is chosen so that the logarithms of the intensities center at the same value. The second step is separate adjustment of the normalization of the twelve arrays for each animal so that the relations implied by the RNA mixing can be used as a basis for our analysis.

Several methods have been proposed for normalizing the intensities from Affymetrix arrays. We base our choice of method on the particular structure of the samples, their preparation by mixing. As discussed by Bolstad [20], the normalization methods that are usually considered for Affymetrix arrays distort the relations among the intensities when the measurements are of substantially different mRNA samples. Our normalization method is intended to avoid distortion of the array-to-array intensity relations that the RNA mixing used in sample preparation imply [21, 22]. Thus, our normalization method allows us to make use these intensity relations in our analysis.

For a particular animal, the samples measured are related by the mixing. The second normalization step is based on a model of this relationship. For animal j and probe p of gene g, let the model intensity for the liver sample be θ_{Ajgp} and for the kidney sample θ_{Djgp}. Letting i index the 12 arrays, we have as the mixing model for the scaled intensity

$$x_{Aji}\theta_{Ajgp} + x_{Dji}\theta_{Djgp},$$

where

$$x_{Aji} = \begin{cases} 1 & \text{for material A} \\ \dfrac{0.75}{0.75 + 0.25 c_{Dj} / c_{Aj}} & \text{for material B} \\ \dfrac{0.25}{0.25 + 0.75 c_{Dj} / c_{Aj}} & \text{for material C} \\ 0 & \text{for material D} \end{cases}$$

and

$$x_{Dji} = 1 - x_{Aji}.$$

The unknown parameters in this model are θ_{Ajgp}, θ_{Djgp}, and the ratio c_{Dj} / c_{Aj}.

The second normalization step consists of adjusting the scaled intensities y_{jigp} by forming

$$\frac{y_{jigp} - \eta_{0\,ji}}{\eta_{ji}}.$$

Note that the values of $\eta_{0\,ji}$ and η_{ji} are the same for every probe and gene. The second normalization step includes two parameters $\eta_{0\,ji}$ and η_{ji} instead of just one as would be the case for scale normalization. For each animal, the values of $\eta_{0\,ji}$ and η_{ji} are chosen to make the adjusted scaled intensities fit the mixture model as closely as possible. We denote the chosen values by $\hat{\eta}_{0\,ji}$ and $\hat{\eta}_{ji}$.

Estimation of the unknown parameters in the mixture model is part of the second normalization step. The estimates of θ_{Ajgp} and θ_{Djgp} are denoted by $\hat{\theta}_{Ajgp}$ and $\hat{\theta}_{Djgp}$. From the estimates of the ratios c_{Dj} / c_{Aj}, we obtain estimates of x_{Aji} and x_{Dji}, which we denote by $\hat{x}_{Aji}$ and $\hat{x}_{Dji}$. A fitted value for the intensity of each probe is given by

$$\hat{x}_{Aji}\hat{\theta}_{Ajgp} + \hat{x}_{Dji}\hat{\theta}_{Djgp}.$$

The estimates of the ratios c_{Dj} / c_{Aj} are

Animal	mRNA ratio c_{Dj} / c_{Aj}
1	1.33
2	1.23

3	1.40
4	1.80
5	0.99
6	1.16

The second normalization step leads to normalized intensities that are more in line with the mixture model. How well our normalization method performs is, of course, a legitimate question. In part, the results in this paper provide an answer. We apply our normalization method to each animal individually. As detailed below, each animal gives similar results. If our normalization method were unstable, results for different animals would not be similar. Thus, it seems that our normalization method performs reasonably.

Overview of the gene summaries

Typically, in the analysis of data from Affymetrix arrays, the probe intensities for each gene are summarized to give a single value. The general reason for this, which applies here as well, is that biological variation occurs at the gene level. Probe-to-probe variation for a single gene is purely technical variation. Thus, for the purpose of comparing technical and biological variation, a gene-level analysis seems, at least, to be a good place to start.

In the analysis presented here, we use weighted least squares to summarize the probe intensities for each gene [22]. We obtain the weights from the fitted probe intensities that are part of second step of the normalization.

The basis of our data analysis is more than the gene summaries, which we denote u_{jig}. From the fitted values for the probe intensities, we obtain fitted values for the gene summaries. In addition, we also obtain a regression weight for each gene, which allows us to use weighted least squares for modeling the gene summaries.

A key observation in the experimental results is similarity of the gene expression in the different animals. Figures 2 and 3 present this observation for the liver and the kidney, respectively. For each animal, the base 2 logarithms of the fitted gene expression levels are plotted versus the levels for the other animals. We see that there is a strong relation between the levels for the different animals. The correlation is not perfect, and one can wonder why. Figures 2 and 3 do not offer evidence that the levels for the different animals could be better matched with a different approach to normalization.

Some data analysis purposes require selection of genes with appreciable expression. To select such genes, we average the fitted liver intensities and the fitted kidney intensities over the animals. Figures 2 and 3 suggest that such averages represent in some sense all the animals. We denote the liver average by $\overline{\theta}_{Ag}$ and the kidney average by $\overline{\theta}_{Dg}$. For some purposes, we select genes for which either $\overline{\theta}_{Ag} > 1$ or $\overline{\theta}_{Dg} > 1$. There are 9804 such genes. This number can be compared to the total number, which is 31099. We note that the number of genes for which $\overline{\theta}_{Ag} > 1$ is 7283, and the number of genes for which

$\bar{\theta}_{Dg} > 1$ is 8392. These latter two counts serve to locate the cutoff value in the overall range of expression levels.

Biological-Technical model

Probe summarization provides 18 measurements on liver samples, 3 for each animal. For each gene, we might naively think of these measurements in terms of a one-way analysis of variance with 6 categories each corresponding to an animal and 3 replicates for each category. This thinking is naïve because the replicates cannot be modeled as simply as a one-way analysis of variance assumes. As shown below, there are batch effects in the technical variation that show up as lack of statistical independence among the replicates. Nevertheless, it seems worthwhile to look at the ratio of the estimated animal variance to the estimated replicate variance for each gene.

Figure 4 shows a histogram of the variance ratio for the genes that satisfy $\bar{\theta}_{Ag} > 1$. We note that this histogram shows negative values that result from estimation error in the estimation of the animal variances. The animal variance is greater than the replicate variance if the true ratio is greater than 1. The estimation error clouds the picture somewhat, but it seems clear that the ratio is greater than 1 for a substantial number of genes.

Figure 5 shows a histogram of the animal to replicate variance ratio for the analogous kidney data. The genes chosen are those that satisfy $\bar{\theta}_{Dg} > 1$. Our conclusion from this figure parallels that from Figure 4.

Various factors are associated with the technical variation that perturbs the gene expression measurements considered here. The levels of some of these factors are recorded in the data description. Such factors include the scanner used, the fluidics machine used, and the location of the specimen on the 96-well plate. There may, of course, be other factors that affect the technical variation. Thus, the size of the technical variation observed in this experiment may not be as large as, for example, the size that would be observed in an experiment involving several different laboratories. In fact, experience has shown that inter-laboratory variation is very likely to be larger than within-laboratory variation.

A model that includes both the animal-to-animal variation and the factors that affect the technical variation is what we need to compare the biological variability and the technical variation carefully. The model is of the gene summaries of the normalized probe intensities, which we denote by u_{jig}. For each gene, there are 71 such values, 12 measurements on each of six animals with one value missing, the third replicate on animal 3, material B. The model includes a term for the animal effect, a term for the scanner effect, a term for the fluidics machine effect, and a term for the random error.

Cleary, how the technical variation manifests itself for a particular gene depends on the expression levels in the liver and kidney. We consider the case in which these two expression levels are different. In this case, the most noticeable manifestation of the

technical variation is in the difference between the two levels. In other words, the technical variation manifests itself as variation in the slope of the calibration curve that relates the mRNA species with the observed intensity.

We fit a model with similar terms to each gene in a set chosen to include genes that have appreciable difference in liver and kidney expression levels. From all 31099 genes, we obtain the set of genes considered by first eliminating the genes for which $\overline{\theta}_{Ag} \leq 1$ and $\overline{\theta}_{Dg} \leq 1$. Second, we eliminate the genes for which $\text{abs}\left(\log_2(\overline{\theta}_{Ag}/\overline{\theta}_{Dg})\right) \leq 1$. Note that the genes remaining after the second step show a two-fold difference between the liver intensity and the kidney intensity. Third, we eliminate the genes judged to show saturation for any one of the six animals. The details of judging saturation are discussed in the section on data analysis methods.

In specifying the model, we make use of the liver average by $\overline{\theta}_{Ag}$ and the kidney average by $\overline{\theta}_{Dg}$. For each array, these averages lead to a predicted intensity given by

$$\hat{x}_{Aji}\overline{\theta}_{Ag} + \hat{x}_{Dji}\overline{\theta}_{Dg}.$$

Note the dependence on j due to the variation in the mRNA concentration ratio from animal to animal, which at this point in our analysis we consider known. Think of a deviation that consists of a change in the slope of the calibration curve. Under such a deviation, the model intensity would change to

$$\beta\left(\hat{x}_{Aji}\overline{\theta}_{Ag} + \hat{x}_{Dji}\overline{\theta}_{Dg}\right).$$

We portray the animal-to-animal and technical variation largely in these terms. For the animal-to-animal variation, we also include a deviation in the intercept of the calibration curve

$$\alpha\left(\overline{\theta}_{Ag}/2 + \overline{\theta}_{Dg}/2\right).$$

In the model, the animal effect term is given by

$$\alpha_j\left(\overline{\theta}_{Ag}/2 + \overline{\theta}_{Dg}/2\right) + \beta_j\left(\hat{x}_{Aji}\overline{\theta}_{Ag} + \hat{x}_{Dji}\overline{\theta}_{Dg}\right),$$

where α_j and β_j are the coefficients to be fitted. This parameterization of the animal effect makes the comparison of genes easier. If there were no animal effect, then the coefficient β_j would equal 1 and the coefficient α_j would equal 0.

We parameterize the term for the scanner effect similarly. We let this term be

$$\gamma_i\left(\hat{x}_{Aji}\overline{\theta}_{Ag}+\hat{x}_{Dji}\overline{\theta}_{Dg}\right).$$

The coefficients γ_i depend on the array and not the animal because of the way the experiment was designed. Two different scanners were used in the experiment. The design has an imbalance of little consequence because of the missing value.

We adopt a linear mixed model for the gene summaries [23]. We take the animal effect and the scanner effect to be fixed effects and parameterize them as above. We take the fluidics machine effect to be a random effect. We take its contribution to the model to be given by

$$\varsigma\left(\hat{x}_{Aji}\overline{\theta}_{Ag}\ +\ \hat{x}_{Dji}\overline{\theta}_{Dg}\right),$$

where ς is a random variable. There is a different realization of ς for each of the 12 fluidics machines.

For each gene, we have as our gene summaries model

$$u_{jig}=\alpha_j\left(\overline{\theta}_{Ag}/2+\overline{\theta}_{Dg}/2\right)+(\beta_j+\gamma_i)\left(\hat{x}_{Aji}\overline{\theta}_{Ag}+\hat{x}_{Dji}\overline{\theta}_{Dg}\right)+\varsigma\left(\hat{x}_{Aji}\overline{\theta}_{Ag}+\hat{x}_{Dji}\overline{\theta}_{Dg}\right)+\varepsilon_{ji}$$

In our notation for the unknowns of this model, we have suppressed the dependence on the gene. For a particular i, γ_i takes on one of two values, the one for the particular scanner. The error term ε_{ji} has variance inversely proportional to the regression weight obtained as part of the probe summarization.

The fitting for a particular gene gives a value for each β_j, a total of 6 values. Although there are other possibilities, we portray the animal-to-animal variation with the standard deviation of these 6 values. We summarize over the 2213 genes in our chosen set by means of a histogram.

For each gene, the fitting gives a value of γ_i for one scanner and the negative of that value for the other scanner. We portray the scanner variation in terms of one of these scanner values. Again, we summarize over the 2213 genes by means of a histogram.

Figure 6 shows the scanner histogram and the animal histogram. This figure shows that under the circumstances of the experiment, the scanner variation is small relative to the animal variation. The data do provide evidence that the absolute value of γ_i is greater than 0 for a substantial number of genes. Thus, saying that the scanner variation can be ignored is perhaps going too far.

For each gene, the fitting gives the standard deviation of ς, which we use to portray the fluidics machine variation. Figure 7 shows the histogram of these standard deviations and the animal histogram. This figure shows that under the circumstances of the

experiment, the fluidics machine variation is small relative to the animal variation. The data do provide evidence that the standard deviation of ς is greater than 0 for a substantial number of genes. As with the scanner variation, saying that the fluidics machine variation can be ignored is perhaps going too far.

Finally, we ask about the random variation. This requires some adjustment so that the standard deviation of ε_{ji} can be compared to the animal-to-animal (β_j) standard deviation. The top histogram in Figure 8 is obtained from the standard deviation of ε_{ji} divided by $\hat{x}_{Aji}\overline{\theta}_{Ag} + \hat{x}_{Dji}\overline{\theta}_{Dg}$. Both the numerator and the denominator vary with the array. For this reason, we average over the arrays to obtain the top histogram in Figure 8. The bottom is the animal histogram. Again we see that the animal variation is larger.

It is worth noting that were there a class of genes of particular interest, then figures analogous the Figure 6 to 8 could be generated for this class. These figures would contain histograms involving only the genes in the class of interest.

Whether we have properly chosen the terms for our gene summaries model is an issue. First, for the factors we have included, we have assumed a particular form for the effects as they vary with mixture. Second, there are other factors that we might include such as location on the 96-well plate, instances of measurements repeated on the basis of laboratory inspection, and chip lot. Examination of residuals is the standard approach to this issue.

In the case of high dimensional measurements, examination of residuals differs from examination in the case of univariate measurements. We fit our gene summaries model 2213 times, once for each of 2213 genes. Each gene gives a 71 dimensional vector of residuals. One might guess that any deficiency in our gene summaries model would affect the residuals for many of the genes. Thus, combining the residuals for different genes provides a powerful approach to looking at the fit of our gene summaries model.

One way to obtain this power is to inspect a few components of the singular value decomposition of weighted residuals. Let the residuals given by the difference between the gene summaries and the fitted model be denoted by r_{jig}. We weight these by multiplying by the square root of the regression weight used in fitting the linear mixed model, $\tilde{w}_{jig}$. This gives a 71 by 2213 matrix with elements $r_{jig}\sqrt{\tilde{w}_{jig}}$ to which we apply the singular value decomposition. Thinking of the singular value decomposition in terms of principal components analysis, we observe that the first component explains 27% of the variance. The first three components explain 55%.

The first component of the singular value decomposition is given by $\mathbf{u}_1 d_1 \mathbf{v}_1^{\mathrm{T}}$. If we think of this as the contribution of a factor to the weighted residuals, then the first two terms in this expression give the pattern of variation across the arrays due to the factor and the third term gives the relative amount that the factor contributes to each gene. Figure 9

shows the third term. We plot the elements of $\mathbf{v}_1$ versus $\log_2(\bar{\theta}_{Ag}/2+\bar{\theta}_{Bg}/2)$, which gives an idea of the expression amount for the particular gene. We see that the elements of $\mathbf{v}_1$ are almost all positive and become more positive for more highly expressed genes. This property might help in looking for a factor to explain most of this component. Figure 10 shows the first two terms of the first component with the mixtures labeled. The array-by-array pattern shown in Figure 10 offers few clues as to the factor that drives the first component. A small clue is given by the four measurements that the laboratory substituted for defective measurements. These are shown in red. These values are all high but not so high as to explain most of the variation captured by the first component.

Our consideration of the residuals from the fitted gene summary model shown in Figures 9 and 10 do not offer much in terms of factors missing in the modeling. One could look at more than the first component in hopes that the result would be more informative. One could also include more factors in the gene summary model and look at the corresponding residuals. We have not pursued these ideas.

Conclusions

As illustrated in this paper, a broader investigation of technical variation is possible through test measurement of mRNA from a group of similarly treated animals and of mixtures of mRNA from different animal organs. A group of animals broadens the investigation by providing animal-to-animal differences as a gauge of the technical variation. Mixtures based on different organs broaden the investigation by allowing calibration properties of the measurement system to be checked. In addition, the illustration presented in this paper shows that the normalization method for mixture models used here is effective.

The experiment described here has the virtue of allowing investigation of technical variation in terms of differences among animals in a control group. The use of animals in a control group as a measurement system test seems appropriate because a measurement system capable of detecting differences within such a group can generally be expected to be sufficient for detection of whatever differences there are between such a group and a treated group of animals. By sufficient, we mean that another measurement system would not provide substantially better biological evidence. Thus, the experiment described here can be thought of as a prototype of a pilot study that might be done before a larger study with specific biological objectives involving treated animals.

This paper illustrates approaches to gauging the technical variation in terms of the animal variation, but does not identify a quantitative approach that is to be preferred. In qualitative terms, the gene expression measurement system considered in this paper appears quite consistent in its response to the RNA samples from the six animals chosen. This appearance suggests both that the animal-to-animal differences in the mRNA are small and that the technical variation is, on the same basis, also small. For each gene, we can be more definitive about the comparative size of the animal variability and the technical variation. For many but not all genes, the animal variability is larger than the technical variation. In interpreting an experiment like the one described here, one could

focus on some gene class of particular interest and thus be more definitive in the comparison of animal variability and technical variation.

In judging technical variation in an experiment yet to be performed, the use of standard normalization methods [20] might be anticipated. In this situation, the question arises as to how results such as those presented here can be informative. There does not seem to be an easy answer. An approach might be based on comparison of the intensities obtained under standard normalization with the intensities obtained with the normalization method used here. The comparison should reveal preprocessing problems inherent in the standard normalization methods.

Measurement of mRNA from two different organs as well as mixtures of these two mRNAs adds possibilities to the investigation of the technical variation. Some of the results presented in this paper are essentially single-organ results and others involve measurements on the mRNAs and their mixtures. The animal-to-animal comparisons of expression profiles shown in Figures 2 and 3 as well as the comparison of the among-animal variance with replicate-to-replicate variance shown in Figures 4 and 5 can be considered single-organ results. On the other hand, the comparisons of biological variability with technical variation in Figure 6 to 8 are based on the mRNAs and their mixtures. These comparisons largely involve genes that are more highly expressed in one organ than the other. For such genes, one can check to see if the measurement system has a mean response that is linear in the concentration of the particular mRNA species. When this is true, one can characterize the technical variation in terms of variation of the intercept and slope of the calibration curve.

In this paper, much of the data analysis is applied to the measurements on each animal individually. That the results for the individual animals are as consistent as they are suggests that the data analysis methods we have chosen are reasonably stable. Thus, this paper also illustrates the performance of our methods for normalization and gene summarization.

Data analysis methods

The gene expression measurements on which this paper is based reflect differences between expression in liver and kidney, animal-to-animal variability, and technical variation. In performing the data analysis, the challenge is choice of an approach that separates the animal variability and the technical variation well enough that the two can be compared. The approach we have chosen is based on the general idea that any deviation from the mixture model is due to technical variation.

Our data analysis approach is made up of four parts, normalization of probe intensities, summarization of the probe intensities for each gene, identification of genes for which the mixture model does not fit because of saturation, and characterization of the animal and technical variation in the gene summaries. All of these tasks involve the mixture model. Our normalization approach is based in part on preservation of the fit of the mixture model. The same is true of our summarization approach. Identification of genes that

exhibit saturation allows us to omit such genes from our animal-technical comparison. Our characterization approach, which we have already discussed in detail, is based on a linear mixed model that includes the mixture model.

As discussed previously, our normalization approach consists of two steps. The scaled probe intensities y_{jigp} produced by the first step satisfy

$$\sum_g \left(\sum_p \log_2 (y_{jigp}) \right) = 0 .$$

This equation is equivalent to saying that the scaling consists of centering the logarithms of the probe intensities for each array at their mean. This equation shows that the scaling depends on whether the material measured is liver mRNA, kidney mRNA, or some mixture of the two and that one cannot expect the mixture model to fit the scaled probe intensities. We adjust for this lack of fit in the second step of our normalization.

The second step is an iterative procedure that results in normalized probe intensities given by

$$\frac{y_{jigp} - \hat{\eta}_{0\,ji}}{\hat{\eta}_{ji}}$$

and fitted probe intensities given by

$$\hat{x}_{Aji}\hat{\theta}_{Ajgp} + \hat{x}_{Dji}\hat{\theta}_{Djgp} .$$

Note the dependence or lack of dependence that is indicated by the subscripts attached to the parameters in these expressions. The indices j and i indicate dependence on the array with j indexing animal and i indexing mixtures and their replicates. The indices g and p indicate dependence on the probe with g indexing gene and p indexing the probes for this gene. We see that the arithmetic operation that constitutes our second normalization step is the same for every probe on an array. We see that the fitted intensities for each animal can be obtained from the fitted intensities for the liver $\hat{\theta}_{Ajgp}$ and the fitted intensities for the kidney $\hat{\theta}_{Djgp}$.

An understanding of our second step stems from the relation between the normalized intensities and the fitted intensities that the iteration produces upon convergence. Consider first the normalization parameters $\hat{\eta}_{0\,ji}$ and $\hat{\eta}_{ji}$. We choose these parameters so that the normalized intensities are as close to the fitted intensities as possible. In mathematical terms, we want to choose values of $\eta_{0\,ji}$ and η_{ji} so that $\left(y_{jigp} - \eta_{0\,ji}\right)/\eta_{ji}$ is as close to the fitted intensities as possible. In other words, the normalization parameters $\hat{\eta}_{0\,ji}$ and $\hat{\eta}_{ji}$ are given by fitting $\eta_{0\,ji} + \eta_{ji}\left(\hat{x}_{Aji}\hat{\theta}_{Ajgp} + \hat{x}_{Dji}\hat{\theta}_{Djgp}\right)$ to y_{jigp} for each array. The fitting is by weighted least squares with weights [24, 25]

$$1 \Big/ \left(\left(\max\left(\hat{x}_{Aji} \hat{\theta}_{Ajgp} + \hat{x}_{Dji} \hat{\theta}_{Djgp}, 0 \right) \right)^2 + \delta^2 \right).$$

Consider second the fitted intensities for the liver and kidney, $\hat{\theta}_{Ajgp}$ and $\hat{\theta}_{Djgp}$. Whereas the normalization parameters are obtained array by array, these fitted intensities are obtained probe by probe for each animal. The fitted intensities are given by fitting $\hat{x}_{Aji} \theta_{Ajgp} + \hat{x}_{Dji} \theta_{Djgp}$ to $\left(y_{jigp} - \hat{\eta}_{0\,ji} \right) / \hat{\eta}_{ji}$ for each probe. The fitting is by weighted least squares with the weights given above.

Consider third the estimate of the ratio of mRNA concentrations $\left(c_{Dj} / c_{Aj} \right)$. This ratio is given by fitting $x_{Aji} \hat{\theta}_{Ajgp} + x_{Dji} \hat{\theta}_{Djgp}$ to $\left(y_{jigp} - \hat{\eta}_{0\,ji} \right) / \hat{\eta}_{ji}$ for each animal, where x_{Aji} and x_{Dji} are functions of the concentration ratio as discussed previously. The fitting is by weighted non-linear least squares with the weights given above.

This description of the relation between the normalized intensities and the fitted intensities provides the basis for a computational algorithm [22]. Note that in each type of fitting, some of the initially unknown parameters are fit and others are held constant. Our iterative procedure for normalization involves these types of fitting applied one after the other in a repetitive sequence. Eventually all the fitted parameters converge so that the above relations between the normalized intensities and the fitted intensities hold.

The next part of our data analysis is summarization of the probe intensities for each gene. The Affymetrix array used in the experiment discussed here provides several (typically 11) probes as sensors for the mRNA for each gene. It would, of course, be feasible to investigate animal variability and technical variation on the probe intensity level. However, because biological variation occurs on the gene level, Affymetrix array processing typically includes summarization over the probes corresponding to each gene. For this reason, comparison of animal variability and technical variation is more meaningful in terms of the underlying biology if performed on the gene summary level.

We perform probe summarization with weighted least squares [22]. The use of weighting is motivated by the fact that higher probe intensities are more variable than lower probe intensities. This same fact motivates the use of the logarithm in other approaches to probe summarization. Carroll and Ruppert [26] discuss weighting and transformation as alternatives. For animal j, array i, and gene g, our probe summary is given by

$$u_{jig} = \frac{\sum_p w_{jgp} \left((y_{jigp} - \hat{\eta}_{0\,ji}) / \hat{\eta}_{ji} \right)}{\sum_p w_{jgp}},$$

where w_{jgp} is the weight for probe p.

We calculate the weights from the fitted probe intensities in a way analogous to calculation of weights for the least squares computations in our normalization approach. Think about the fitted probe intensities $\hat{x}_{Aji}\hat{\theta}_{Ajgp} + \hat{x}_{Dji}\hat{\theta}_{Djgp}$ as a two-way table indexed by i and p. We average this table over i so that the weights do not depend on i. With weights independent of i, the summarization does not interfere with the fit of the mixture model. We have

$$1/w_{jgp} = \left(\max\left(\frac{1}{12}\sum\nolimits_i \hat{x}_{Aji}\hat{\theta}_{Ajgp} + \hat{x}_{Dji}\hat{\theta}_{Djgp}, 0\right)\right)^2 + \delta^2.$$

Probe summarization methods for Affymetrix arrays typically give just a gene index value analogous to u_{jig}. With our method, we can obtain something more. From the fitted probe intensities, we can obtain fitted gene intensities

$$\hat{\mu}_{jig} = \frac{\sum_p w_{gp}\left(\hat{x}_{Aji}\hat{\theta}_{Ajgp} + \hat{x}_{Dji}\hat{\theta}_{Djgp}\right)}{\sum_p w_{gp}}.$$

Moreover, we can obtain weights for use in fitting models to the u_{jig}

$$\tilde{w}_{jig} = \frac{\left(\sum_p w_{jgp}\right)^2}{\sum_p w_{jgp}^2\left(\left(\max\left(\hat{x}_{Aji}\hat{\theta}_{Ajgp} + \hat{x}_{Dji}\hat{\theta}_{Djgp}, 0\right)\right)^2 + \delta^2\right)}.$$

The third part of our data analysis involves identification of genes for which the mixture model does not fit because of saturation. It is well known that for some probes, the relation between the response and the concentration of the mRNA species in the sample is not linear because of saturation [27]. The degree of saturation in array-wide terms can be adjusted through adjustment of the sample preparation protocol. However, this protocol is usually not adjusted to the extent necessary to eliminate all saturation. Thus, we can expect to identify some genes that exhibit saturation. Treating such genes as a case separate from the remainder of the genes seems to be a reasonable option in the data analysis.

For a particular animal j and a particular gene g, there are 12 values of u_{jig} (or 11 in the case of animal 3 as discussed in Section 2). One or more aspects of the technical variation cause these 12 values to deviate from the mixture model. One aspect is saturation, and another is random error. We are going to test the null hypothesis that the random error is sufficient to explain the lack of fit of the mixture model. From the outset, we must acknowledge that there is an ambiguity in our investigation of saturation because some aspects of the technical variation besides saturation may cause rejection of

this null hypothesis. For each of the 4 mixtures, there are 3 replicate measurements. Because lack of fit of the mixture model does not influence the differences among replicate measurements on the same mixture, these differences provide an estimate of the variance of the random error. Our test consists of determining whether this estimate is sufficient to account for the lack of fit of the mixture model to the averages of the mixture replicates.

Computation of the lack-of-fit test statistic involves familiar statistical ideas from linear regression. For a particular animal j and a particular gene g, two sums of squares must be computed. One sum of squares can be obtained from the replicates as follows: For each set of three replicates, we compute the standard deviation, which we denote by $s_A,\ s_B,\ s_C,\ s_D$. Let the 4 distinct values of w_{jig} be denoted by $w_A,\ w_B,\ w_C,\ w_D$. Except for animal 3, which has only 11 values, this sum of squares is given by

$$S_1 = 2w_A s_A^2 + 2w_B s_B^2 + 2w_C s_C^2 + 2w_D s_D^2 .$$

The other sum of squares is a result of the fitting of the mixture model. We have

$$S_2 = \min_{\theta_A,\theta_B} \sum_i w_{jig} \left(u_{jig} - x_{Aji}\theta_A - x_{Dji}\theta_D\right)^2 .$$

The test statistic is given by

$$\left((S_2 - S_1)/2\right)/(S_1/8) .$$

Under the null hypothesis, this statistic is F distributed with 2 and 8 degrees of freedom. Thus, a p value can be obtained for each animal and gene. The formulas for animal 3 differ in some details.

This lack-of-fit F test does not tell us all that we would like to know about the character of the lack of fit such as whether the problem is saturation. For a particular gene, consider the case in which the liver expression level is greater than the kidney expression level. In this case, saturation shows up in the liver level with the fitted value based on the linear model greater than the observed value. We can assess this by forming the standardized residual for the liver level [22]. In a similar fashion, we can treat the case in which the kidney level is greater than the liver level by forming the standardized residual for the kidney level.

We combine the lack-of-fit F test and the appropriate standardized residual in Figures 11 to 16. Each of these figures has two plots, one for genes with the liver level greater and one for gene with the kidney level greater. For each gene, the p value of the lack-of-fit F test is plotted versus the appropriate standardized residual. Small p values show lack of fit, and standardized residuals with substantial negative values show saturation. We see that most genes with significant lack of fit have standardized residuals that suggest saturation. There are, however, some genes with significant lack of fit that do not have associated evidence of saturation. These genes are puzzling, but perhaps there are so few

of them that individual attention is not worthwhile. We note that the results of Zheng, et al. [28] seem consonant with Figures 11-16.

Each figure in the sequence reflects results for a different animal. That the figures are so much alike seems remarkable. Interestingly, the genes that show saturation have considerable overlap from animal to animal. Consider genes for which the lack-of-fit p value is less than 0.01 and the standardized residual is less that -2. There are 368 genes for which this is true for at least one animal. Of these, 225 genes exhibit no overlap, that is, they qualify on the basis of only one animal. Another 59 genes qualify on the basis of all six animals. The other results are

Number of animals	6	5	4	3	2	1
Number of genes	59	13	20	21	30	225

The total number of genes considered in forming this table is 9804. Thus, one might conclude that many of the genes that appear with no overlap are included in the table only because of noise.

Characterization of the animal and technical variation in the gene summaries involves fitting a linear mixed model as described previously. The linear mixed model for the normalized intensities u_{jig} is given there in a standard form that permits fitting with available software [23].

Generally, Figures 2, 3, and 11 through 16 show that the animal-to-animal differences in the gene expression measurements are small. Of course, choice of a way to quantify this notion is difficult because the data are high dimensional. As what is perhaps a secondary aspect, this experiment provides a test of the normalization method. Recall that the second step of the normalization method was applied to each animal individually. Apparently, one can say that the normalization method did not make the results for the different animals seem different. Thus, the normalization method appears to be reasonably stable.

Disclaimer:
Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

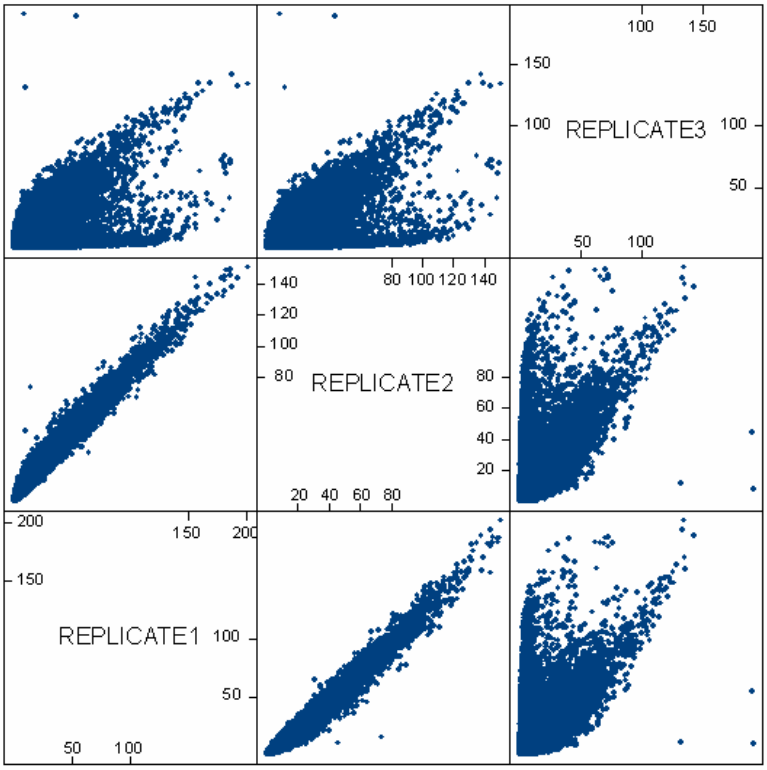

Figure 1. Replicates for animal 3, material B. Intensities of perfect-match probes plotted one replicate versus another.

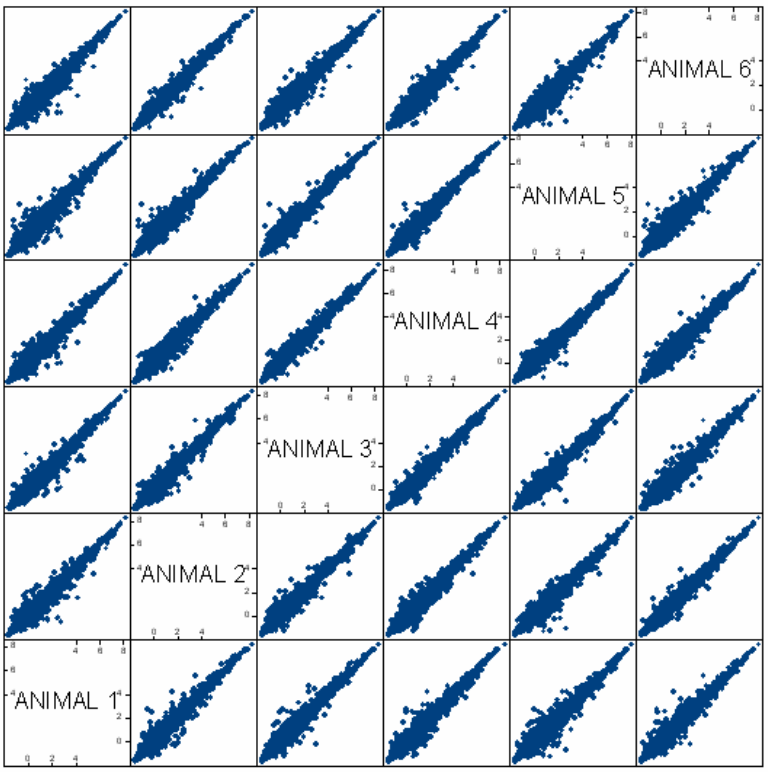

Figure 2. Fitted gene summary values for the liver material from each animal. Log base 2 intensities plotted one animal versus another.

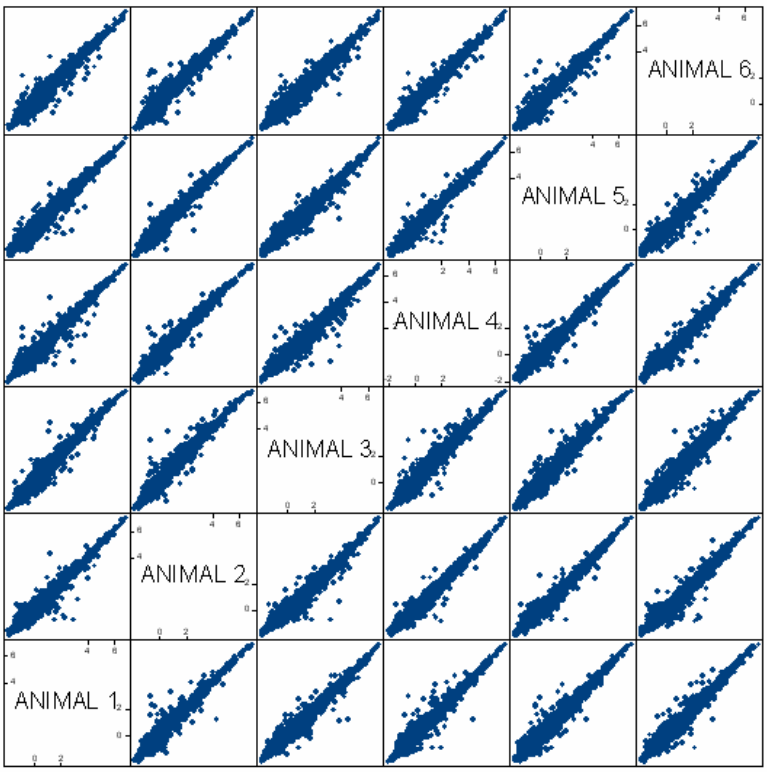

Figure 3. Fitted gene summary values for the kidney material from each animal. Log base 2 intensities plotted one animal versus another.

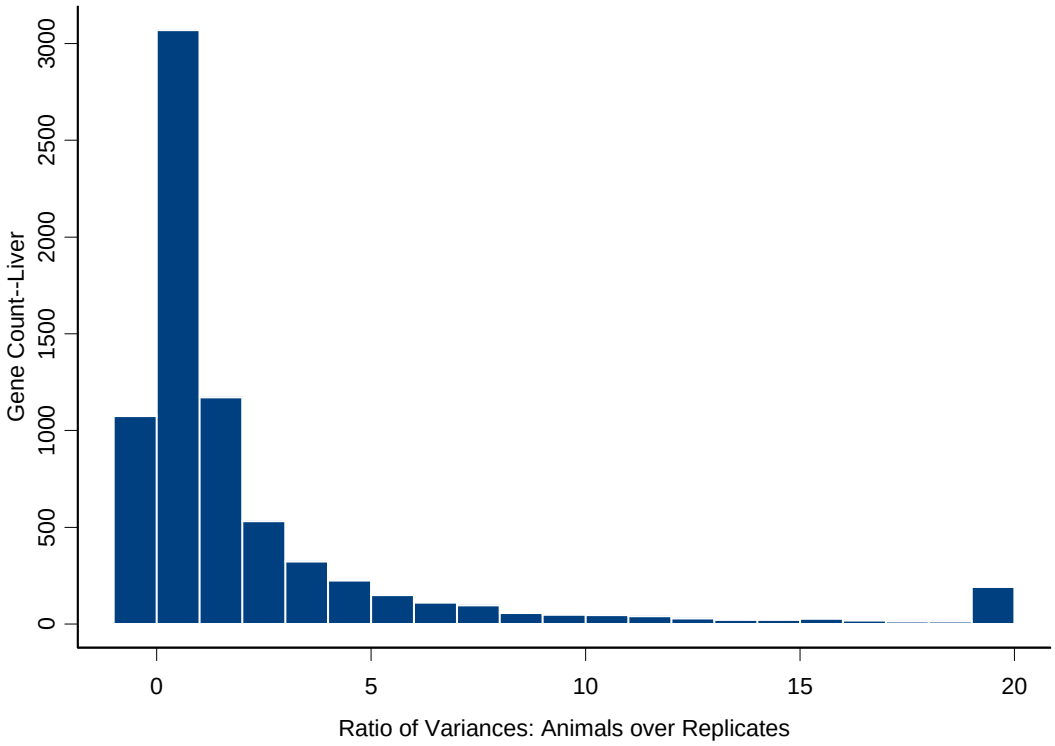

Figure 4. For the liver materials, histogram of the ratio of the animal-to animal variance to the replicate-to-replicate variance. The ratio is truncated at 20.

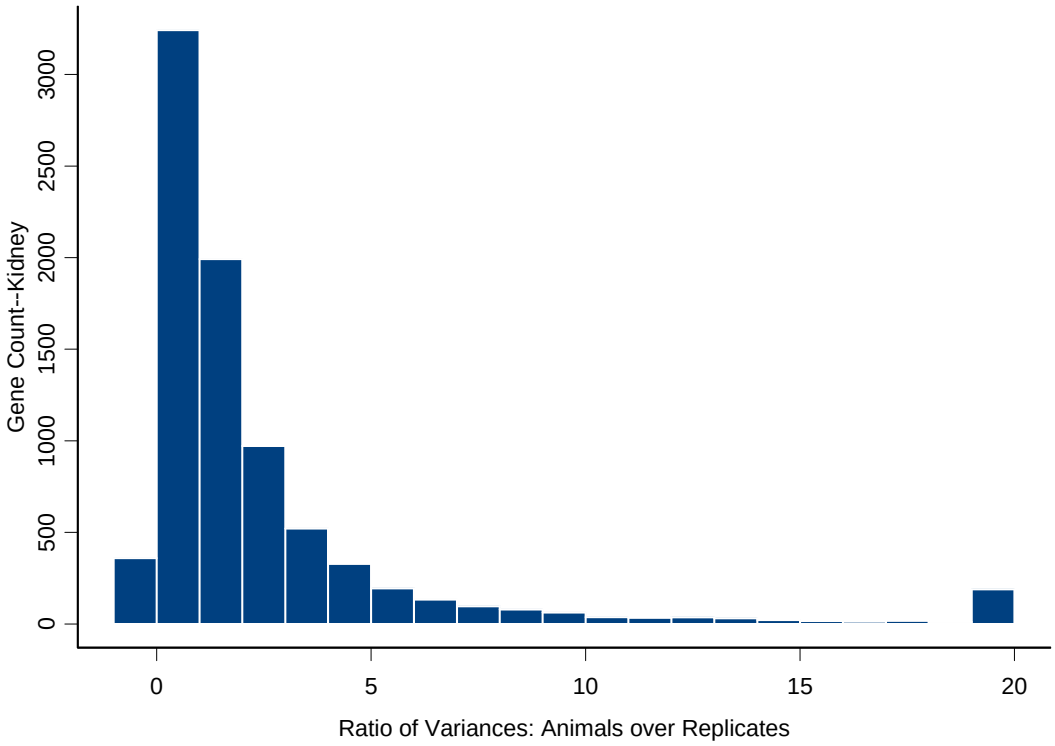

Figure 5. For the kidney materials, the ratio of the animal-to animal variance to the replicate-to-replicate variance. The ratio is truncated at 20.

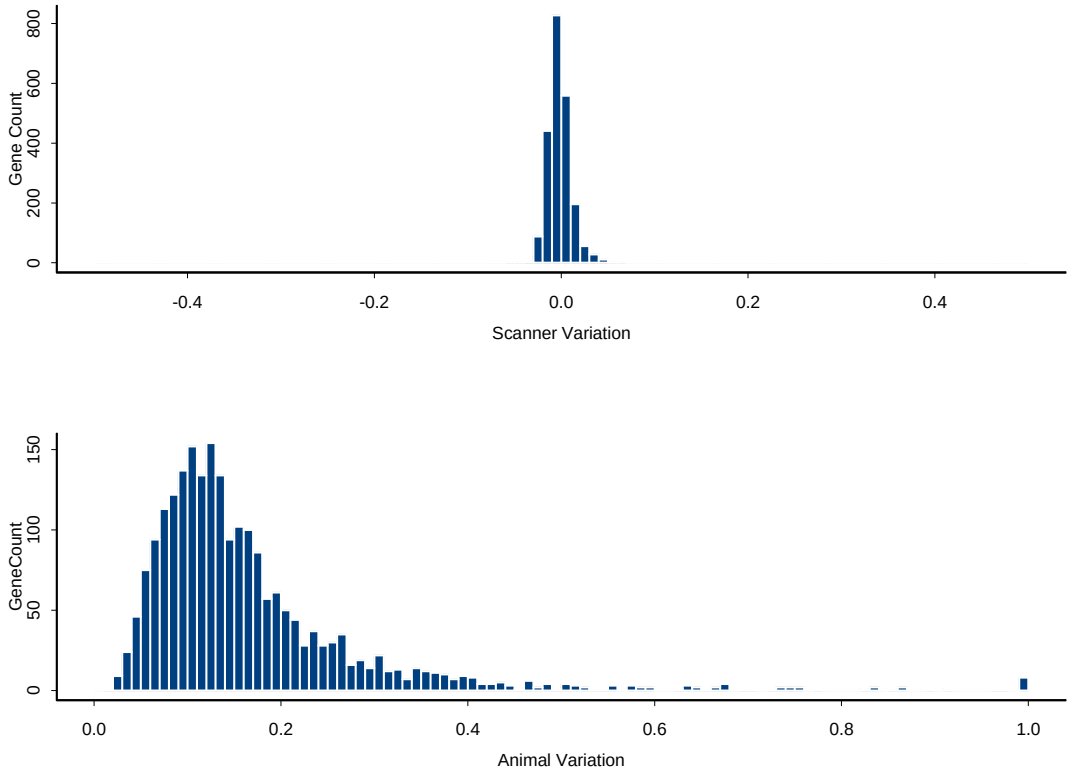

Figure 6. Variation in terms of the fractional change in calibration curve slope. Difference between scanners versus animal-to-animal standard deviation. Animal standard deviation truncated at 1.

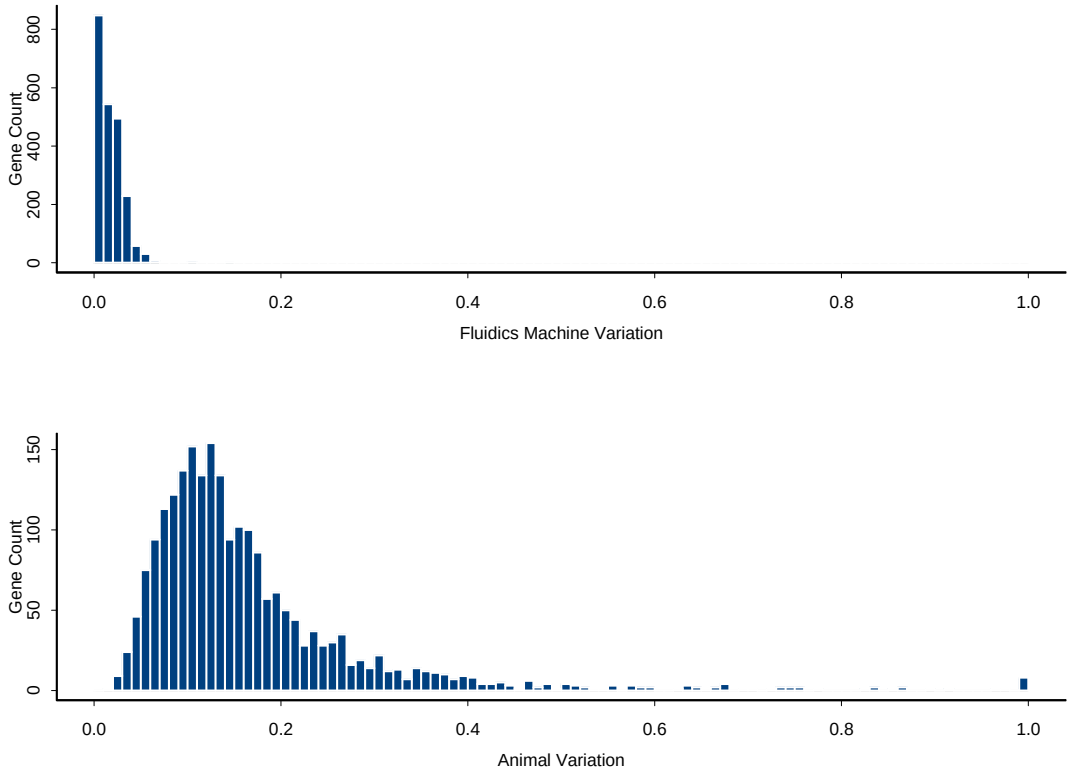

Figure 7. Variation in terms of the fractional change in calibration curve slope. Fluidic machine standard deviation versus animal-to-animal standard deviation. Animal standard deviation truncated at 1.

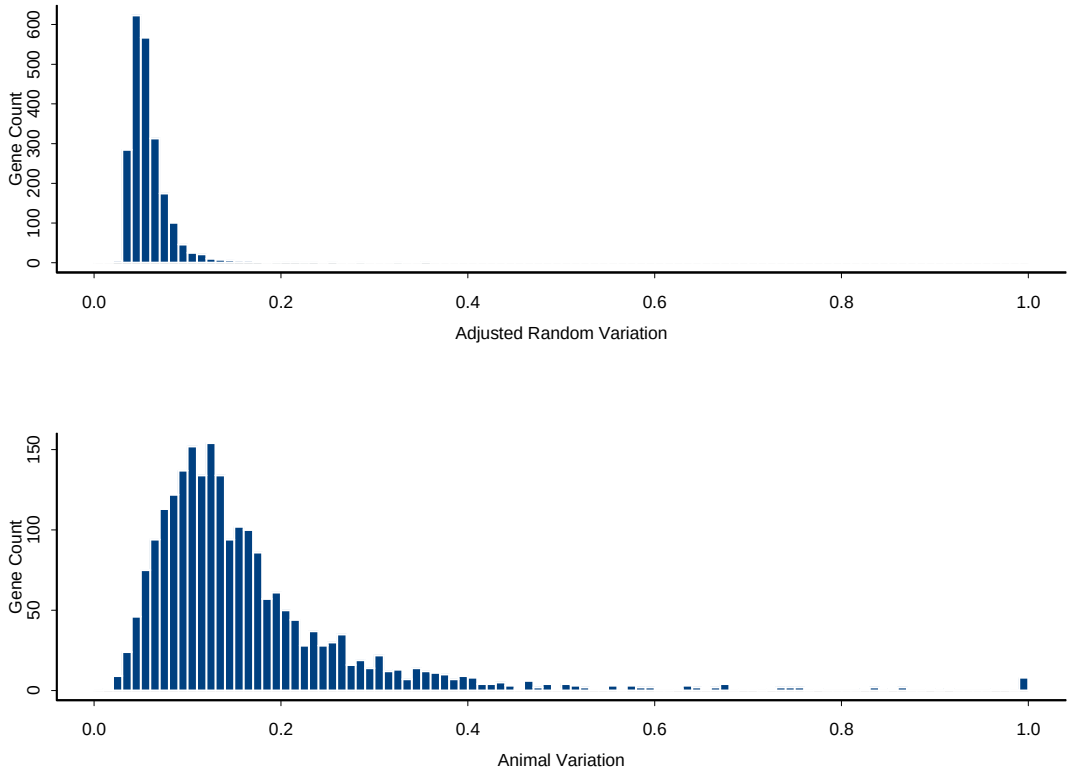

Figure 8. Variation in terms of the fractional change in calibration curve slope. Average standard error for the animal fitted values versus animal-to-animal standard deviation. Animal standard deviation truncated at 1.

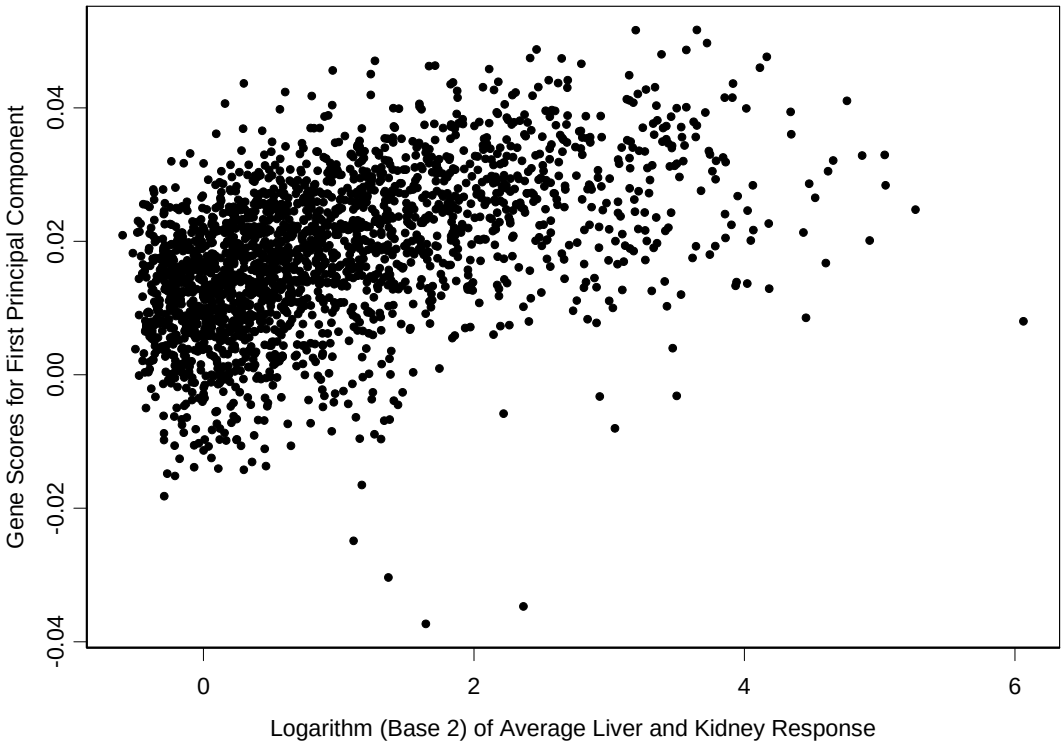

Figure 9. Gene-to-gene variation in the contribution of the first principal component.

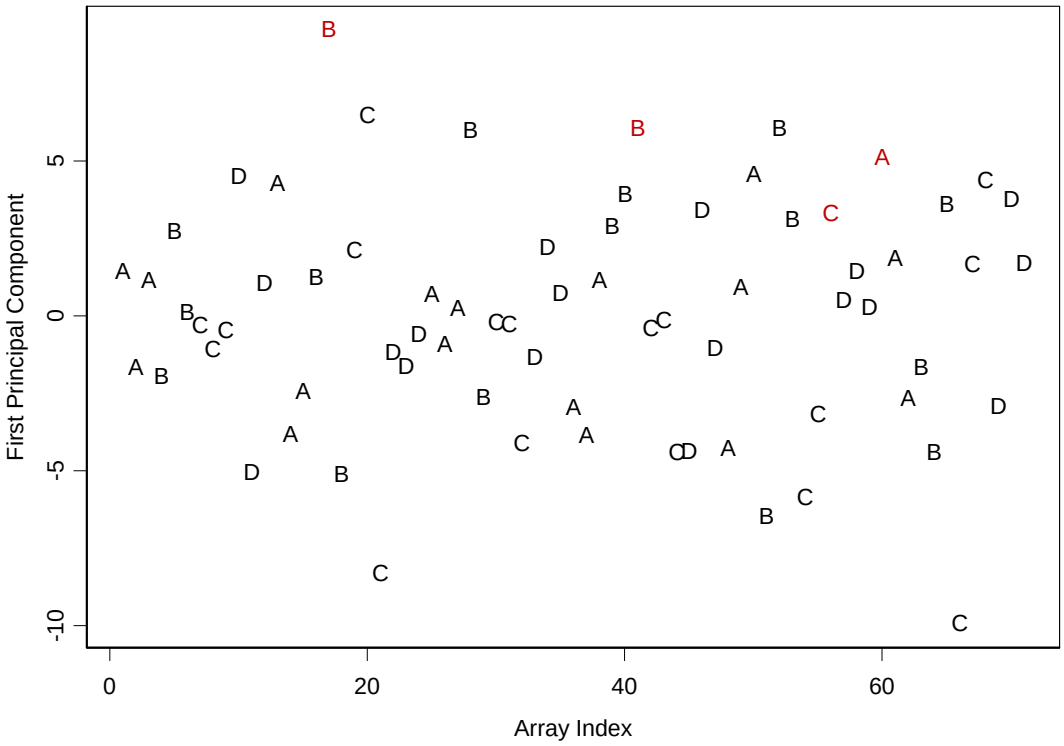

Figure 10. Array-to-array contribution of the first principal component with the mixture labeled. The red points correspond to measurements that were repeated when the original measurements were judged defective.

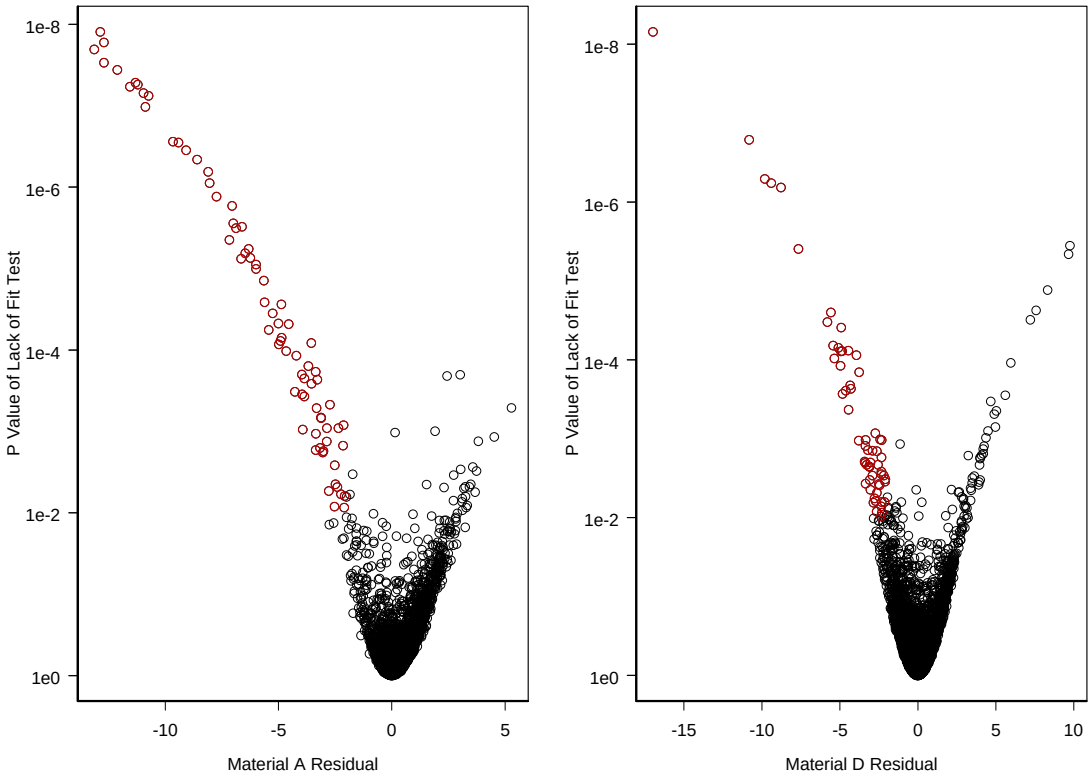

Figure 11. For animal 1, gene-by-gene characterization of the fit to the linear mixture model. Lack-of-fit p value versus the standardized residual. Indication of saturation shown by the genes in red.

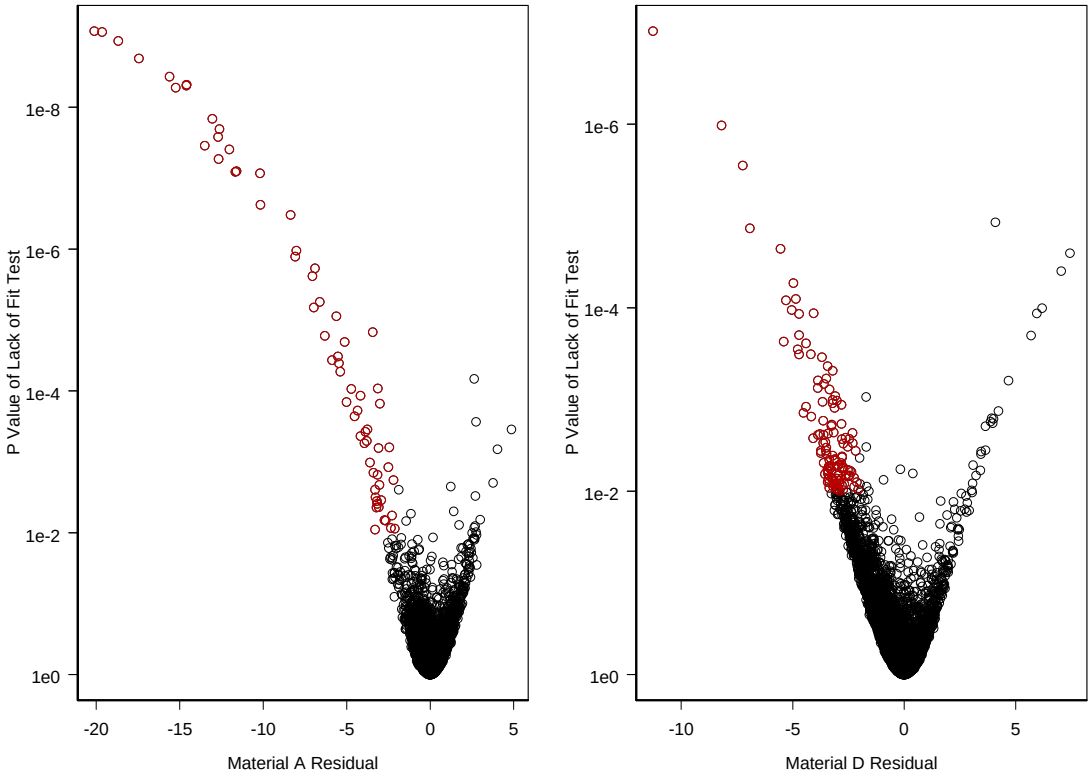

Figure 12. For animal 2, gene-by-gene characterization of the fit to the linear mixture model. Lack-of-fit p value versus the standardized residual. Indication of saturation shown by the genes in red.

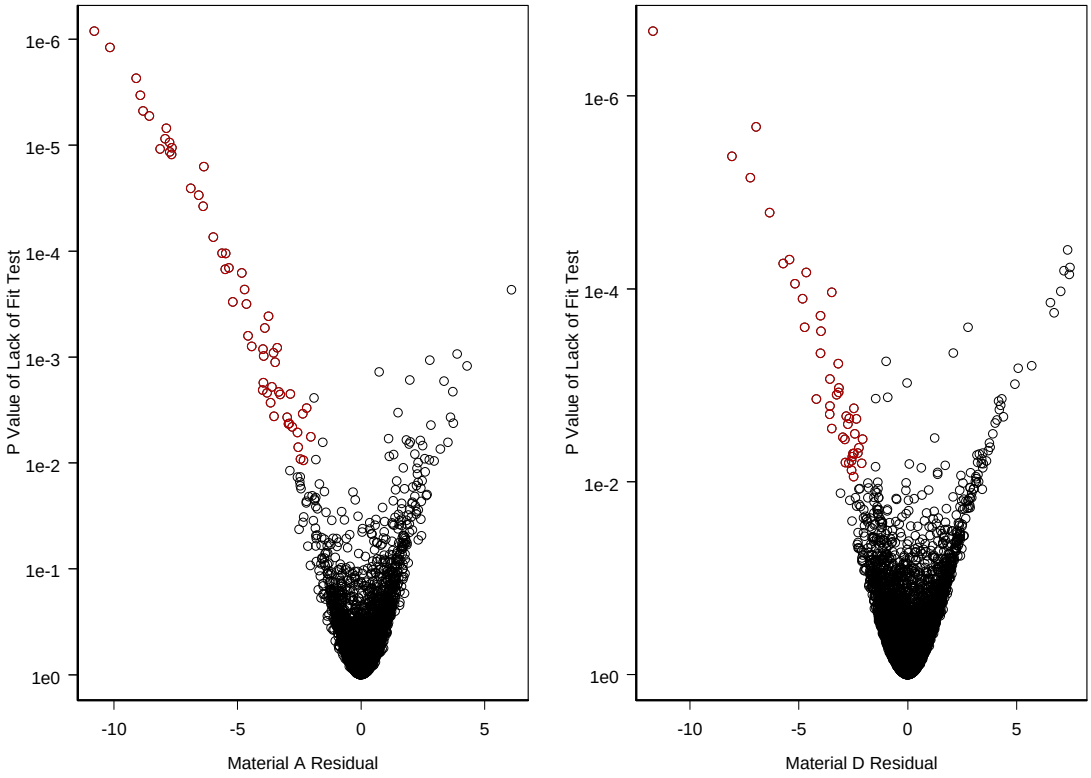

Figure 13. For animal 3, gene-by-gene characterization of the fit to the linear mixture model. Lack-of-fit p value versus the standardized residual. Indication of saturation shown by the genes in red.

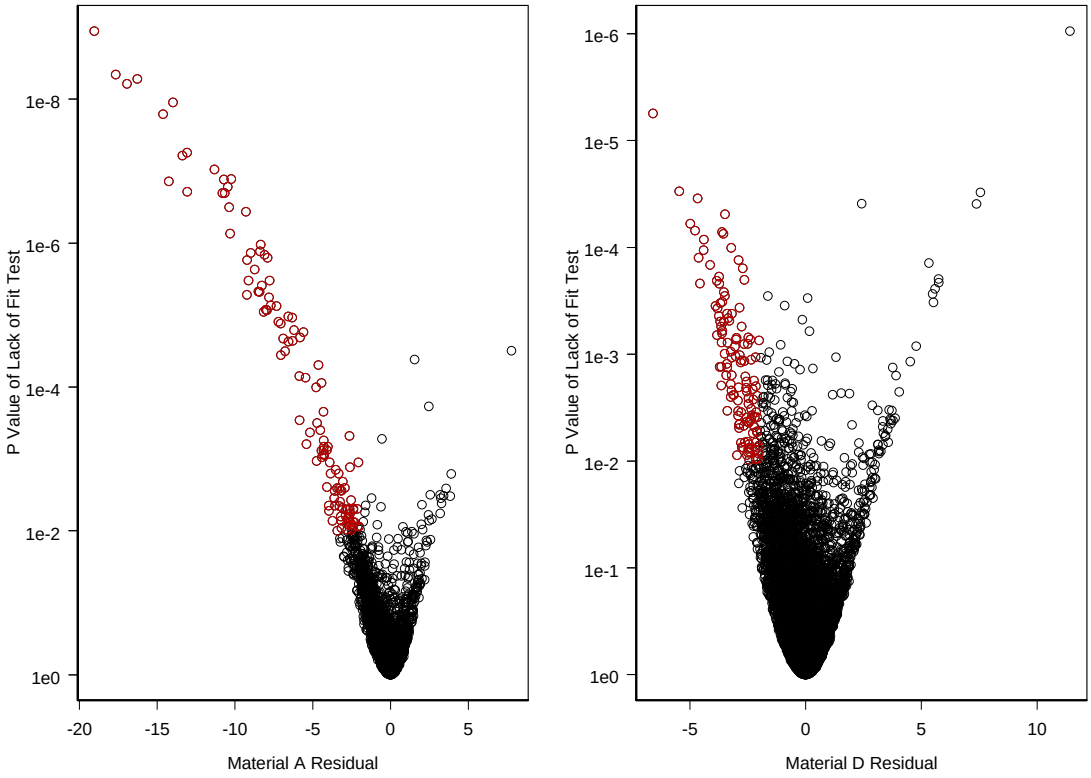

Figure 14. For animal 4, gene-by-gene characterization of the fit to the linear mixture model. Lack-of-fit p value versus the standardized residual. Indication of saturation shown by the genes in red.

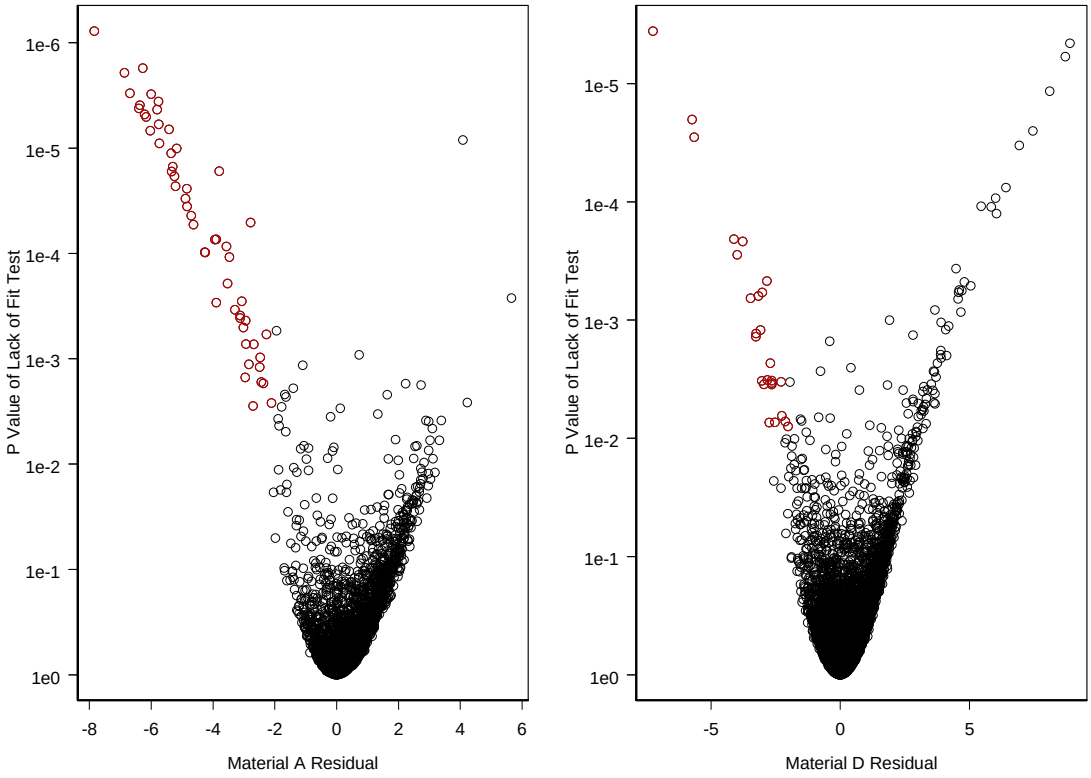

Figure 15. For animal 5, gene-by-gene characterization of the fit to the linear mixture model. Lack-of-fit p value versus the standardized residual. Indication of saturation shown by the genes in red.

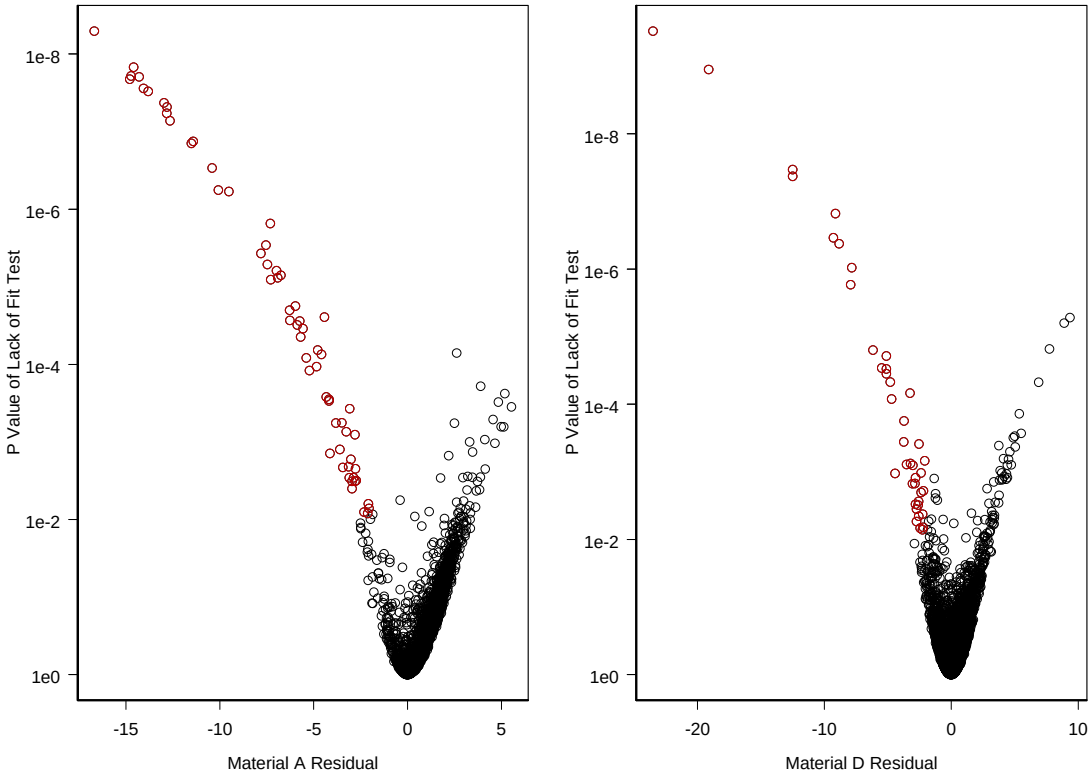

Figure 16. For animal 6, gene-by-gene characterization of the fit to the linear mixture model. Lack-of-fit p value versus the standardized residual. Indication of saturation shown by the genes in red.

References

[1] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics,* vol. 7, pp. 55-65, 2006.

[2] D. B. Allison, G. P. Page, T. M. Beasley, and J. W. Edwards, *DNA microarrays and related genomics techniques: Design, analysis, and interpretation of experiments*. Boca Raton, FL: CRC Press, 2006.

[3] M. Bylesjo, D. Eriksson, A. Sjodin, S. Jansson, T. Moritz, and J. Trygg, "Orthogonal projections to latent structures as a strategy for microarray data normalization," *BMC Bioinformatics,* vol. 8: Art. No. 207, 2007.

[4] M. Bylesjo, M. Rantalainen, J. K. Nicholson, E. Holmes, and J. Trygg, "K-OPLS package: Kernal-based orthogonal projections to latent structures for prediction and interpretation in feature space," *BMC Bioinformatics,* vol. 9: Art. No. 106, 2008.

[5] R. A. Irizarry, Z. Wu, and H. A. Jaffee, "Comparison of Affymetrix GeneChip expression measures," *Bioinformatics,* vol. 22, pp. 789-794, 2006.

[6] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Peterson, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu, "Multiple-laboratory comparison of microarray platforms," *Nature Methods,* vol. 2, pp. 345-349, 2005.

[7] M. B. Satterfield, K. Lippa, Z. Q. Lu, and M. L. Salit, "Microarray scanner performance over a five-week period as measured with Cy5 and Cy3 serial dilution slides," *Journal of Research of the National Institute of Standards and Technology,* vol. 113, pp. 157-174, 2008.

[8] I. Y. Yang, "Use of external controls in microarray experiments," *Methods in Enzymology,* vol. 411, pp. 50-63, 2006.

[9] D. M. Rocke, "Design and analysis of experiments with high throughput biological assay data," *Seminars in Cell & Development Biology,* vol. 15, pp. 703-713, 2004.

[10] MAQC Consortium, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology,* vol. 24, pp. 1151-1161, 2006.

[11] M. J. Boedigheimer, R. D. Wolfinger, M. B. Bass, P. R. Bushel, J. W. Chou, M. Cooper, J. C. Corton, J. Fostel, S. Hester, J. S. Lee, F. Liu, J. Liu, H.-R. Qian, J. Quackenbush, S. Pettit, and K. L. Thompson, "Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories," *BMC Genomics,* vol. 9 Art. No. 285, 2008.

[12] K. L. Thompson, B. A. Rosenzweig, P. S. Pine, J. Retief, Y. Turpaz, C. A. Afshari, H. K. Hamadeh, M. A. Damore, M. Boedigheimer, E. Blomme, R. Cuurlionis, J. F. Waring, J. C. Fuscoe, R. Paules, C. J. Tucker, T. Fare, E. M. Coffey, Y. He, P. J. Collins, K. Jarnagin, S. Fujimoto, B. Ganter, G. Kiser, T. Kaysser-Kranich, J. Sina, and F. D. Sistare, "Use of mixed tissue RNA design for performance assessments on multiple microarray formats," *Nucleic Acids Research*, vol. 33, 2005.

[13] International Organization for Standardization (ISO), *Guide to the expression of uncertainty in measurement*. Geneva, Switzerland, 1995.

[14] D. Cook, H. Hofmann, E.-K. Lee, H. Yang, B. Nikolau, and E. Wurtele, "Exploring gene expression data, using plots," *Journal of Data Science*, vol. 5, pp. 151-182, 2007.

[15] D. Finkelstein, M. Janis, A. Williams, K. Steiger, and J. Retief, "Microarray quality control and assessment," in *DNA microarrays and related genomics techniques*, D. B. Allison, G. P. Page, T. M. Beasley, and J. W. Edwards, Eds. Boca Raton, FL: CRC Press, 2006.

[16] J. J. Chen, R. R. Delongchamp, C.-A. Tsai, H.-m. Hsueh, F. Sistare, K. L. Thompson, V. G. Desai, and J. C. Fuscoe, "Analysis of variance components in gene expression data," *Bioinformatics*, vol. 20, pp. 1436-1446, 2004.

[17] K.-Y. Kim, D. H. Ki, H. J. Jeong, H.-C. Jeung, H. C. Chung, and S. Y. Rha, "Novel and simple transformation algorithm for combining microarray data sets," *BMC Bioinformatics*, vol. 8 Art. No. 218, 2006.

[18] R. Shippy, S. Fulmer-Smentek, R. V. Jensen, W. D. Jones, P. K. Wolber, C. D. Johnson, P. S. Pine, C. Boysen, X. Guo, E. Chudin, Y. A. Sun, J. C. Willey, J. Thierry-Mieg, D. Thierry-Mieg, R. A. Setterquist, M. Wilson, A. B. Lucas, N. Novoradovskaya, A. Papallo, Y. Turpaz, S. C. Baker, J. A. Warrington, L. Shi, and D. Herman, "Using RNA sample titrations to assess microarray platform performance and normalization techniques," *Nature Biotechnology*, vol. 24, pp. 1123-1131, 2006.

[19] Z. Wu and R. A. Irizarry, "Stochastic models inspired by hybridization theory for short oligonucleotide arrays," *Journal of Computational Biology*, vol. 12, pp. 882-893, 2005.

[20] B. Bolstad, "Preprocessing and normalization for Affymetrix GeneChip expression microarrays," in *Methods in microarray normalization*, P. Stafford, Ed. Boca Raton, FL: CRC Press, 2008.

[21] W. Liggett, "Normalization and technical variation in gene expression measurements," *Journal of Research of the National Institute of Standards and Technology*, vol. 111, pp. 361-372, 2006.

[22] W. Liggett, "Technical variation in the modeling of the joint expression of several genes," in *Methods in microarray normalization*, P. Stafford, Ed. Boca Raton, FL: CRC press, 2008, pp. 121-150.

[23] J. C. Pinheiro and D. M. Bates, *Mixed-Effects Models in S and S-PLUS*. New York: Springer Verlag, 2000.

[24] W. Huber, A. v. Heydebreck, H. Sultmann, A. Poustka, and M. Vingron, "Variance stabilization applied to microarray data calibration and to the

quantification of differential expresssion," *Bioinformatics*, vol. 18, Suppl. 1, pp. S96-S104, 2002.

[25] D. M. Rocke and B. Durbin, "A model for measurement error for gene expression arrays," *Journal of Computational Biology*, vol. 8, pp. 557-569, 2001.

[26] R. J. Carroll and D. Ruppert, *Transformation and weighting in regression*. New York: Chapman and Hall, 1988.

[27] E. Chudin, S. Kruglyak, S. C. Baker, S. Oeser, D. Barker, and T. K. McDaniel, "A model of technical variation of microarray signals," *Journal of Computational Biology*, vol. 13, pp. 996-1003, 2006.

[28] X. Zheng, H.-C. Huang, W. Li, P. Liu, Q.-Z. Li, and Y. Liu, "Modeling nonlinearity in dilution design microarray data," *Bioinformatics*, vol. 23, pp. 1339-1347, 2007.