# Supplementary Material to A note on oligonucleotide expression values not being normally distributed

| JOHANNA HARDIN* | JASON WILSON |
|---|---|
| Dept. of Mathematics, | Dept. of Mathematics, |
| Pomona College, | Biola University, |
| Claremont, CA 91711 | La Mirada, CA 90639 |
| jo.hardin@pomona.edu | jason.wilson@biola.edu |

January 29, 2009

**Abstract**

**Motivation:**
Novel techniques for analyzing microarray data are constantly being developed. Though many of the methods contribute to biological discoveries, inability to properly evaluate the novel techniques limits their ability to advance science. Because the underlying structure, or distribution, of microarray data is unknown, novel methods are typically tested against the assumed structure of normally distributed data. However, microarray data are not, in fact, normally distributed, and testing against such data can have misleading consequences.

**Results:**
Using an Affymetrix technical replicate Spike-In data set, we showed that oligonucleotide expression values are not universally normally distributed under any of the standard methods for extracting expression values. The resulting data tend to have a large proportion of skew and heavy tailed values. Using data simulated under three models (normal, heavy tailed, and skewed), additionally, we showed that standard methodologies (for differential expression and gene similarity) can give unexpected and misleading results when the data are not normally distributed. Robust methods should be used when analyzing microarray data. Additionally, when evaluating new techniques, skewed and/or heavy tailed data distributions should be considered in simulations.

**Key words:** microarray data; Affymetrix; distributions; non-normality.

*To whom correspondence should be addressed.

# 1 Introduction

Microarray data may be viewed as having two sources of variability; one is due to technology, and one is due to biology. The first is error which would preferably be removed ("noise"), while the second is the truth which is sought ("signal"). In order to develop and recommend statistical methods for working with microarray data (and incidentally many other types of data), it is standard statistical protocol to understand the distribution of the data. For example, if it is known, then parametric methods may be applied; if it is unknown, then non-parametric methods should be applied. The purpose of this paper is to distinguish and study the technical variation/error of Affymetrix microarrays. Three conclusions will be drawn: (1) It is possible to separate and study the technical variability. (2) The technical variability is not normal, it exhibits heavy tail and skew right behavior. Therefore, (3) methods appropriate for non-normality should be used when analyzing Affymetrix, and presumably other, arrays.

Throughout this paper, the word *transformation*, and its cognates, will be used to refer to any published statistical method which calculates a single gene expression value from the entire probe set, for each gene, on an oligonucleotide chip. The "transformation" process includes background correction, adjustment for PM/MM values, normalization, summarization across probes, logarithmic scale change, and other possible calculations. The idea is that one probe set from an oligonucleotide microarray is "transformed" into one gene expression value. The unconventional use of this term was chosen due to the awkwardness of such expressions as "gene expression measurement method." Five popular transformations will be employed in this study (RMA, GCRMA, MAS 5.0, PLIER, and dChip, see Section 2).

In view of the numerous microarray studies conducted over the past decade, attention to the technical/biological variation distinction deserves focus (Giles and Kipling, 2003; Chen et al., 2007). Nevertheless, in order to clarify our contribution to the literature, we offer further distinctions from previous work. In microarray experiments, the sources of technical variation are numerous, including variability due to lab technicians, equipment, protocol, and reagents. In addition to technical and biological variation, study results also differ due to particular choices of microarray platform and transformation method. Platform differences are due to chip design (particularly oligonucleotide selection) and manufacturing. Transformation differences are due to statistical transformation methods. The recent massive MicroArray Quality Control consortium (MAQC), consisting of 137 researchers from 51 institutions, conducted a $1,300^{+}$ chip study, addressed all of the above issues and stands as a benchmark for microarray studies in the near future (MAQC Consortium, 2006a). The MAQC consortium amply documented the reality of divergent study results (p. 1151) as justification for their goal to address the consistency, reliability, and reproducibility of current microarray research across technical, biological, platform, and transformation differences.

In our study, we analyze the distribution of technical variation on the Affymetrix platform for five competing transformations. It is well known that different transformations give different results (Millenaar et al., 2006), which delays establishing a consensus transformation (for example, see the competition at `http://affycomp.biostat.jhsph.edu/`). However, it has been argued that technical variation is negligible relative to biological and is therefore not worth study (Klebanov and Yakovlev, 2007). The relevant MAQC report disagrees with the idea of technical variation being negligible. Table 1 of the MAQC report (p. 1126) summarizes the numbers of genes selected under different conditions and using different transformations. Selection sizes vary by hundreds of genes. The report concludes that, "the variations seen in Figure 2 and Table 1 can also result from differences ... [in] levels of noise in each measurement" (MAQC Consortium, 2006b). Additionally, the Klebanov and Yakovlev (2007) paper, which argues technical error is negligible, suffers from several serious criticisms by reviewers, including the citation of instances where they believe technical error does effect results in practice, directly contradicting the conclusion of the paper (Mushegian, 2007; Koonin, 2007). Furthermore, since Klebanov and Yakovlev (2007) make claims about the relationship between the technical noise, which was present in their data, to biological signal, which was not present in their data, they have made an assertion which could not even in principle be substantiated by their data. In conclusion, technical error is real. Although it is usually small enough to obtain valid results in microarray studies, it is not negligible and therefore it is worthy of further investigation.

The distribution of technical variation affects the distribution of the biological variation, and the sum of the two are commonly assumed to be normal. The assumption of whether transformed oligonucleotide data are normal or not is important for several reasons. One is that new statistical techniques are constantly being developed for analyzing microarray data and are often tested using normally distributed data. For example, in a recent issue of *Bioinformatics*, six papers dealt with microarray data. Four simulated Gaussian data for the purposes of validation and comparison (Nicolau et al., 2007; Wang and Zhu, 2007; Goeman and Bühlmann, 2007; Wong et al., 2007), one sampled parts of microarray images from real data (Song et al., 2007), and one did no simulations (Royce et al., 2007). Another reason the normality assumption is important is that conventional methods for analyzing oligonucleotide data produce differing results when data are truly normal or not. Improper use of statistical applications can lead to invalid biological conclusions. Understanding the distribution of microarray data will lead to more appropriate analyses and more accurate results.

In what follows, we attempt to show that the technical variability of transformed Affymetrix data is not normal and that this fact can have a non-trivial effect on subsequent analysis. Data from other microarray platforms were not included in this study, but it seems reasonable to suspect similar results for them in the absence of evidence to the contrary. The next three sections of this paper correspond to the three conclusions of the opening paragraph. Section 2 describes the work of Giles and Kipling (2003) where they adapted a published spike-in data set to study the technical variation using five transformations, and concluded that the technical variation was normal. We replicate their study, show an error in their analysis, and conclude that the technical variation is, in fact, not normal. Section 3 investigates skew coefficients, kurtosis, and Hogg's Q2 (Hogg et. al. 1975) under the five transformations. Further evidence of non-normality is established and the kind of non-normality is shown to be primarily tail heaviness and right-skewness. Section 4 investigates the consistency and the accuracy of conventional Affymetrix data analysis methods. The methods studied are two-sample t-tests and Wilcoxon Rank-Sum tests for differential expression and Pearson's and Spearman's correlation coefficients for clustering. Lastly, Section 5 concludes with a discussion of the results.

# 2 Establishing Non-normality

In a 2003 *Bioinformatics* paper, Giles and Kipling (GK) concluded there is "strong support for the normality of the data produced by four different algorithms commonly used for extracting expression values" (pp. 2258-9). They began by correctly identifying that the biology and technology contribute two different sources of variability, of which only the biological is of interest. They next argued that although non-parametric methods are powerful enough for large scale experiments (exceeding 50 chips), well-designed small-scale exploratory experiments (under 10 chips) tend to be analyzed using parametric techniques. "There is, therefore, a requirement to address the nature of the data distributions obtained from the underlying microarray technology" (p. 2255). The remainder of their paper examined the distribution induced by technology using an Affymetrix 59 chip spike-in (SI) dataset. The SI dataset was chosen because of the large number of technical replicates available for distribution assessment (not even the MAQC datasets provide as many technical replicates of a single sample). We replicate the work of GK and find an error in their interpretation/method. In correcting the error we are led to the opposite conclusion, namely that transformed oligonucleotide data **are not** normally distributed.

The SI data set was designed by Affymetrix to investigate the expression levels of known concentrations of various transcripts. However, after removing the spiked-in genes, there are 59 technical replicates of each gene. It is available at `http://www.affymetrix.com/support/technical/ sample_data/datasets.affx` and through R's Bioconductor software at `http://bioconductor.org/packages/2.0/data/experiment/html/SpikeIn.html`. We used the data found in Bioconductor and removed the genes corresponding to the 16 spiked-in transcripts documented. GK report removing only 14 spiked-in genes (p. 2255), the difference being genes 33818_at and 546_at (see (Cope et al., 2003) for explanation). As GK, the 67 control genes (with AFFX prefix) were removed. Our analysis and the GK analysis used $12,543$ and $12,545$ genes, respectively. Five transformation algorithms were applied to the SI data: dChip PM-only (Li and Wong, 2001), Affymetrix MicroArray Suite 5.0 (MAS5) (Affymetrix, 2002), Robust Multi-Array Analysis (RMA) (Irizarry et al., 2003), GCRMA (Wu et al., 2004), and Probe Logarithmic Intensity ERror (PLIER) (Affymetrix, 2005). All

| Transformation | Shap-Wilk | JB | Both |
|---|---|---|---|
| RMA | 24.5% | 26.0% | 21.7% |
| GCRMA | 46.8% | 47.2% | 42.7% |
| MAS5 | 46.2% | 33.6% | 32.3% |
| dChip | 29.5% | 25.9% | 22.9% |
| PLIER | 20.1% | 15.2% | 14.2% |
| $t_5$ | 39.3% | 43.7% | 36.6% |
| $\chi^2_3$ | 99.6% | 92.5% | 92.5% |
| Normal(0,1) | 5.0% | 3.7% | 2.3% |

Table 1: **Normal tests of hypotheses.** Each entry represents the percentage of genes whose distribution was significantly different from normality ($p < 0.05$).

the algorithms are available in Bioconductor and may have minor differences from the Affymetrix transformation techniques (whose algorithms are proprietary). Our choices of transformations differ slightly from GK, who used MAS4 and dChip PM-MM, but did not use RMA, GCRMA, or PLIER. Our selection was made to reflect current transformation methods. All work done here and in other sections was in R (R Development Core Team, 2007).

In order to test for normality, GK applied the Shapiro-Wilk (Shapiro and Wilk, 1965) test to every gene ($n = 59$). If the data were independently and normally distributed, there would be approximately 5% of the genes failing significance tests at a 0.05 level. In fact, for the normally simulated data, there were approximately 5% of the genes failing the tests of normality (Table 1). However, for each of the transformations of the SI data, there was a much higher percentage (see Table 1) than expected of genes that were not normally distributed. Our numbers are similar to GK.

In addition, we applied the Jarque-Bera (Jarque and Bera, 1980) test of normality, with similar results. For comparison, Table 1 also gives the empirical percentages of samples ($10,000$ simulations, each of size $n = 59$) that are rejected under three known distributions: Normal(0,1), $t_5$, and $\chi^2_3$. The normal distribution has a reasonable empirical significance level while the other two distributions correctly reject many of the simulated samples. Since the percentages are so high, GK reasoned that the Shapiro-Wilk test was very powerful with $n = 59$ and so perhaps the "magnitude of deviation from normality" of the rejected genes was small (p. 2256). We pass over the lack of attention to the inherent multiple testing problem the analysis, which is treated in Chen, et. al. (2006).

To investigate the "magnitude of deviation from normality," GK pursued a global QQ-plot approach. QQ-plots give an indication of the strength of normality of a sample. A plot of each gene's quantiles against the quantiles of the normal distribution will give evidence of normality if the points fall on a line. Since it was infeasible to visually examine all $12,545$ QQ-plots, they

| Transformation | % < 0.971 | % < 0.984 |
|---|---|---|
| RMA | 17.9% | 35.1% |
| GCRMA | 36.8% | 58.5% |
| MAS5 | 35.3% | 52.0% |
| dChip | 19.7% | 38.3% |
| PLIER | 12.3% | 24.3% |
| $t_5$ | 30.5% | 57.3% |
| $\chi^2_3$ | 95.8% | 99.8% |
| Normal(0,1) | 1.0% | 10.0% |

Table 2: **Correlation quantiles**. Each entry represents the percent of genes whose sample correlation coefficient between the ordered gene expression values and normal quantiles is less than the given column. If the data were normally distributed, we would expect 1% in the first column and 10% in the second column, as seen in the last row which gives the empirical percentages from Looney & Gulledge (1985).

calculated correlation coefficients between the 59 expression values and standard normal quantiles. To make their assessment, GK made histograms of the correlation coefficients (Figure 1, p. 2257), visually inspected the histograms, and concluded that "the vast majority" of MAS4, dChip PM, and dChip PM-MM genes are sufficiently close to normal to conclude that technical error was normal. The fundamental error in their analysis was to ignore the fact that QQ-plots have inherently high positive correlations. Using the tables of Looney and Gulledge (1985), for normally distributed data, only 10% of the QQ-plots will give a correlation coefficient less than 0.984, and only 1% of the QQ-plots will give a correlation coefficient less than 0.971 ($n = 59$ samples). From Table 2 there were many genes less than the given cutoff for each of the transformation techniques. That is, the data were substantially less correlated with normal quantiles than they would be if the data were in fact taken from a normal population. For comparison, Table 2 gives the empirical percentages of samples ($10,000$ simulations, each of size $n = 59$) that are rejected under simulated $t_5$ and $\chi^2_3$ distributions. Again we see that the $t_5$ and $\chi^2_3$ distributions correctly show their departure from normality by having many QQ-plot correlations that are much smaller than would be expected for normal data.

The reason for the error in GK's analysis was that they used no objective measure to determine "magnitude of deviation from normality." The table of Looney and Gulledge (1985) provides this measure. By the results shown in Table 2, the technical error values all exhibit severe non-normality, the amount of which depends upon the transformation method used.

We recognize that the above arguments assume the independence of the genes, and that the structure of gene dependencies is unknown. With independent normally distributed genes, we would expect to reject only 5% of the genes using the hypothesis tests, and we would expect only 1% and 10% of the genes to have correlations less than 0.971 and 0.984, respectively. We attribute our much higher than expected percentages to the non-normality of the data. In an extreme case, if all the genes were perfectly correlated, we would see either 0% rejection or 100% rejection. That is, the rejection level would be based on the variability of only one gene (i.e., all the genes). For both the normality and the higher than expected percentages to hold, around 19.5% of the RMA data would have to be quite strongly dependent on the 5% of the data which naturally rejects normality (due simply to variability). Though there is no doubt some dependence among the genes, we do not believe that dependence alone can account for, say, 24.5% of the RMA data rejecting the null hypothesis of normality.

# 3 Characterizing the Non-normality

Having reasoned that oligonucleotide technical error is not normally distributed, it is important to characterizing the non-normality. It is well known that raw microarray data (across all platforms) are highly skewed (usually skewed right) with many extreme values (Li and Wong, 2001). Often the log transformation is used to offset the skewness. However, as previously mentioned, the resulting distribution has been almost universally designated as normal. Hoyle et. al. (2002) have observed that the bulk of microarray data have a log-normal distribution (corresponding to Winsor's principle, "All observed distributions are Gaussian in the middle" (Tukey, 1960)) while the tails are better described using a power law distribution (Hoyle et al., 2002). Because, typically, interest is in the tails of the data, understanding the tails of the microarray distribution is generally more important than understanding the bulk of the data. To this end, we explore the skewness, peakedness, and tail heaviness of the data using the skew coefficient, kurtosis coefficient, and Hogg's Q2 (Hogg et al., 1975), respectively. Additionally, the technical error data will be compared to known Normal(0,1), heavy-tailed ($t_5$), and skew-right ($\chi^2_3$) distributions.

The SI data was standardized so that each gene was centered at zero (subtracted a 10% trimmed mean) and scaled (divided by the Median Absolute Deviation, MAD). However, neither of the tests of normality nor the correlation coefficient is sensitive to changes in shift or scale. The skewness coefficient, kurtosis coefficient, and Hogg's Q2 are all also shift and scale invariant.

## 3.1 Skewness, Peakedness, and Tail-heaviness

Normal data are not skewed; that is, they have a skewness coefficient of zero. Skewness is typically measured as the third central moment divided by the cube of the standard deviation, $\frac{\mu_3}{\sigma^3}$.

In Figure 1, the technical error is more skewed than expected if normal. The numbers above (below) each plot represent the percentage of sample skewness coefficients for the particular transformation which lie above the $99th$ (below the

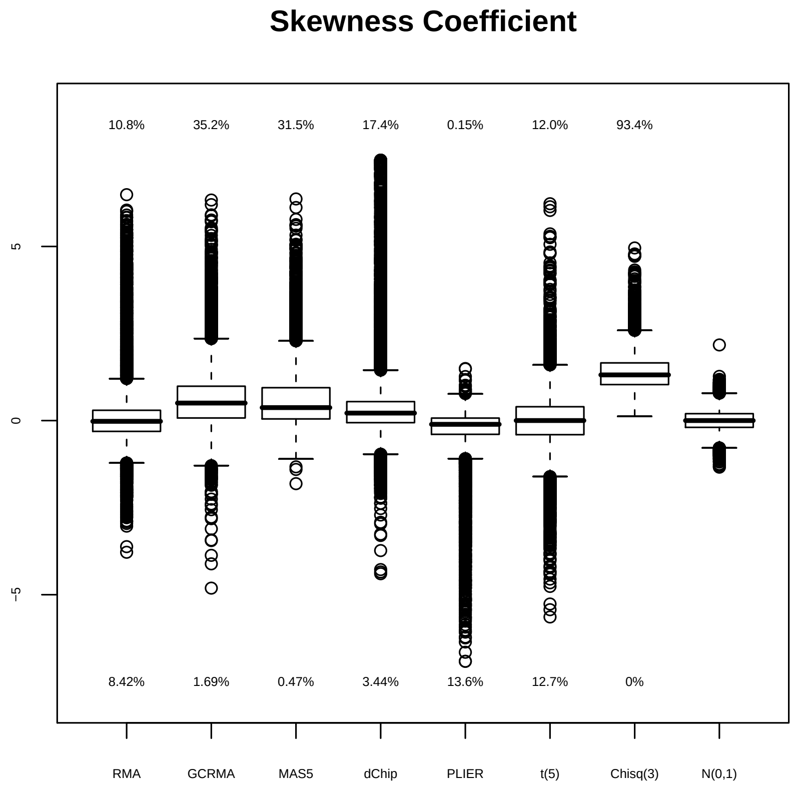

Figure 1: **Skewness boxplots**. The percentage of genes greater than (smaller than) the 99% (1%) of the empirical normal is written above (below) each non-normal boxplot. PLIER is skewed left while the other transformations seem more skewed right.

$1st$) percentile of the normal sample skewness coefficient. The $t_5$ distribution, which is known to be symmetric, and the $\chi^2_3$ distribution, which is known to be skewed, were included as references. The $t_5$, with heavy tails, has about a quarter of its samples with higher skewness than expected under normality. RMA is not substantially different, indicating that the difference from normality might be due to heavy tails or to skewness. GCRMA and MAS5, however, have many genes that are skewed right. The PLIER transformation seems to address the right skewness but has a trade-off of being left skewed.

Another way to observe a departure from normality is through the peakedness of the mound or the heaviness of the tails of the distribution. We measured kurtosis (peakedness) using the fourth central moment divided by the square of the second central moment, $\frac{\mu_4}{\mu_2^2}$. The kurtosis of a normal random variable is 3. We measured tail heaviness using Hogg's Q2, $\frac{\overline{U}_{.05}-\overline{L}_{.05}}{\overline{U}_{.5}-\overline{L}_{.5}}$ (Hogg et al., 1975), where $\overline{U}_p$ is the average of the upper $p * 100\%$ of the data; and $\overline{L}_p$ is the average of the lower $p * 100\%$ of the data. For normal data, Q2 is about 2.6.

In Figures 2 and 3 the kurtosis and tail heaviness are much higher for the technical error than for the normal data. In particular, log scales were applied to display outliers on the boxplots in a visually reasonable manner. The numbers above each plot represent the percentage of sample kurtosis and Q2 coefficients for the particular transformation which lie above the $99th$ percentile of the normal sample kurtosis and Q2 coefficients, respectively.

To characterize the sources of non-normality, the skewness, kurtosis, and tail heaviness for the genes rejected under the Shapiro-Wilk test are tabulated in Table 3. A gene was considered to have high skew, kurtosis, or Q2 if the relevant statistic for that gene fell above the $99^{th}$ percentile for the simulated Gaussian data. As such, inclusion in Table 3 represents extreme deviation from normality. The specific values used were: high skew, 0.74 (low skew, -0.73); high kurtosis, 4.78; and high Q2, 3.38. Under each of the transformations, about two thirds of the rejected genes had skewed data (either positive or negative); large proportions of the rejected genes had high kurtosis; lower, but still substantial, proportions had heavy tails (peakedness and tail heaviness often coincide and we hereafter group them). Any combination of skewness, kurtosis, and tail heaviness is possible for a gene whose technical error distribution deviates from normality. Further study would be required to discriminate all these categories. If such a study were conducted, it should be repeated on a variety of representative technical replicate data sets for validation. From Table 3, we believe it can be reasonably inferred that skewness and tail heaviness are major contributors to the non-normality of the technical error.

## 3.2 Comparison with Known Distributions

In order to obtain some fixed references for the nature of the tail heaviness and skewness detected, we considered the $t_5$ and $\chi^2_3$ distributions in Tables 1, 2, and 3. The $t_5$ distribution represents a heavy tailed distribution whereas the $\chi^2_3$ represents a skew right distribution.

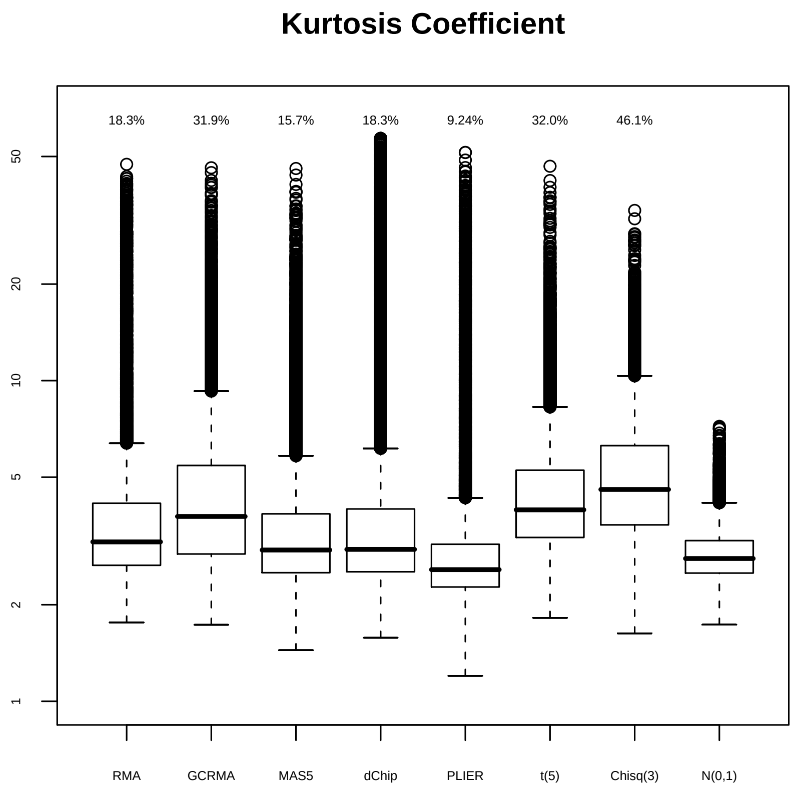

Figure 2: **Kurtosis boxplots**. All transformation methods have a large number of genes with higher kurtosis than expected under normality; the percent of genes with a kurtosis larger than the 99% for the normal is given above each of the boxplots. Note that the y-axis is on a log-scale.

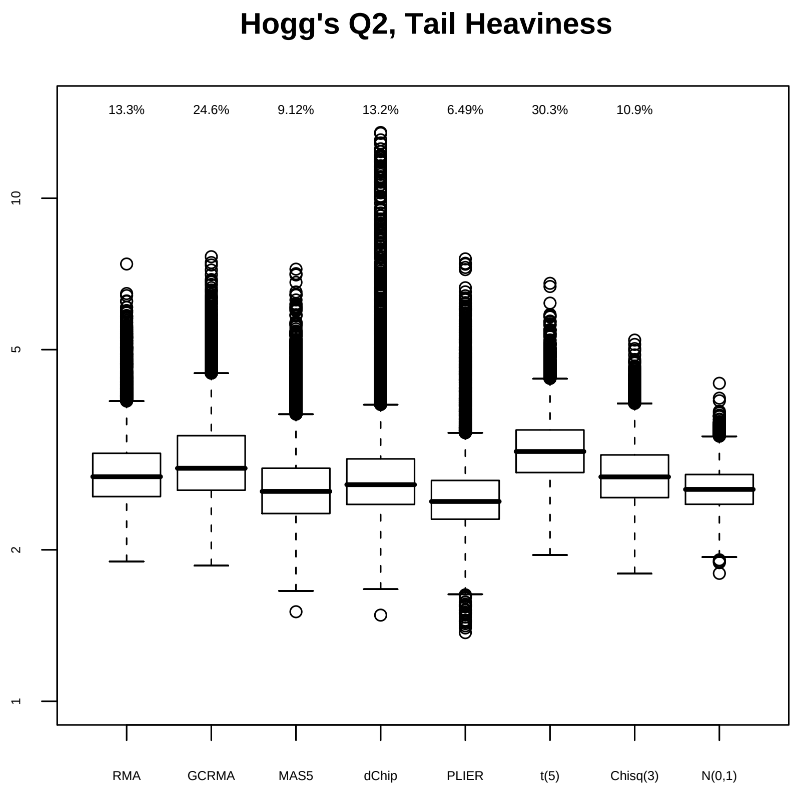

Figure 3: **Tail heaviness boxplots**. All methods have genes with much heavier tails than normal; the percent of genes with a Q2 value larger than the 99% for the normal is given above each of the boxplots. Note that the y-axis is on a log-scale.

| Transf. | Shap-Wilk | Pos Skew | Neg Skew | Any Skew | High Kurt | High Q2 |
|---|---|---|---|---|---|---|
| RMA | 24.5% | 43.9% | 33.1% | 77.0% | 72.5% | 50.2% |
| GCRMA | 46.8% | 74.4% | 3.5% | 77.9% | 67.5% | 51.4% |
| MAS5 | 46.2% | 67.8% | 1.0% | 68.8% | 33.9% | 19.0% |
| dChip | 29.5% | 58.1% | 11.4% | 69.5% | 60.3% | 43.0% |
| PLIER | 20.1% | 0.8% | 66.6% | 67.4% | 45.5% | 31.8% |
| $t_5$ | 39.3% | 12.4% | 12.4% | 24.8% | 32.1% | 31.1% |
| $\chi^2_3$ | 99.6% | 93.5% | 0.0% | 93.5% | 46.2% | 10.9% |

Table 3: **Characterization of non-normal genes**. The first column represents the percentage of genes that rejected normality according to the Shapiro-Wilk test for each of the transformations (and the two simulation distributions). The remaining five columns represent the proportion of the *rejected genes* that were different from what would be expected for Gaussian data.

In Table 1, the simulations of known $t_5$ and $\chi^2_3$ data give a sense of the power of the Shapiro-Wilk and Jarque-Bera tests. It is apparent that it is not feasible to reject all samples of size $n = 59$ taken from non-null distributions. Accordingly, the data in Table 2 indicate that the transformations may have some of the same characteristics as the $t_5$ and $\chi^2_3$ distributions.

Table 3 characterizes the skewness and tail behavior of genes whose data rejected normality. It seems that skewness is a larger component to normality rejection than tail heaviness. Nevertheless, tail heaviness is still a relevant factor. The comparatively large proportion of $\chi^2_3$ samples who rejected normality due to skewness indicates that it is unlikely that skewness is the only component of non-normality contributing to the error deviations we see with the transformed microarray data.

# 4 Effects of Non-normality

From the characterization of the non-normal genes being caused partly by skewness and partly by tail heaviness it becomes important to understand the effects of these attributes on methods typically applied to microarray data. We have referenced papers (in the introduction) which simulated normal data to test novel techniques. The effects non-normality will have on novel techniques is unclear. However, the typical practitioner is probably interested in the effects of non-normality on basic differential expression and gene clustering procedures.

## 4.1 Differential Expression

In order to test the effects of non-normality, we simulated the same three types of data as before: Normal(0,1), $t_5$, and $\chi^2_3$. To be consistent with the Spike-In data, which contained 59 samples, there were 30 samples (arrays) in one group and 29 samples in the other. For each simulation, we repeated the trial $10,000$ times. We simulated genes that were not differentially expressed (no shift) and genes that were differentially expressed (shift of 0.5, 1, or 2). In a typical microarray experiment, we expect to see a combination of genes of both types. The simulated data sets were used to evaluate basic procedures for testing differential expression and computing similarity measures.

### 4.1.1 Consistency

If microarray data are normally distributed, different methods for evaluating differential expression (e.g., t-tests and Wilcoxon rank sum tests) will give similar results (that is, the ordering of the most differentially expressed genes will be conserved across various methods). However, if the data are not normally distributed, methods for evaluating differential expression will give a different ordering of most significantly differentially expressed.

To investigate the consistency of techniques which discover differentially expressed genes, we compared two standard methods: the t-test and the Wilcoxon rank sum (WRS) test. Both methods test for a difference in shift across two samples. If a group of samples is truly differentially expressed, both tests should produce significance. In each of the first set of simulations (Figure 4 and Table 4), we have created two groups which are centered at the same value; that is, there is no differential expression. In this null setting, the p-values should be uniformly distributed from zero to one, and about 5% of the tests should reject the null hypothesis of no differential expression. For each of the distributions and each of the tests, there is a correct amount of error at about 5% (see Tables 4 and 5). There is also a nice spread of p-values across the range of zero to one along each axis. However, the significance of each of the two methods (t-test and WRS test) for a particular simulated dataset (i.e., for a dot on the scatterplot) is not consistent (i.e., do not lie near a 45° line in Figure 4). The normal data were the most consistent of the three (Table 4 and Figure 4(a)). The p-values for the t-test seemed to be correlated (0.90) with the p-values for the WRS test. For the heavy tailed and skew data, there was weaker correlation (0.789 and 0.712, respectively) between the p-values of tests that were assumed to be measuring the same thing (Table 4 and Figures 4(b) & (c)).

### 4.1.2 Power

In the previous subsection, two different methods for testing differential expression were shown to give somewhat inconsistent results in the presence of no differential expression for non-normal data. In order to help determine which method to use, it is important to evaluate the methods under no differential expression (as above) and also under the condition of differential expression. Using the same setup as in Section 4.1.1, we shifted one of the populations (by 0.5, 1, and 2). Given a distribution (e.g., normality), we generated a sample of size 30 from the default distribution and generated a sample of size 29 from the shifted distribution (default + shift). Because there is differential expres-

| | Normal(0,1) | | $t_5$ | | $\chi^2_3$ | |
|---|---|---|---|---|---|---|
| | > .05 t | < .05 t | > .05 t | < .05 t | > .05 t | < .05 t |
| > .05 WRS | 93.6 | 1.1 | 94.1 | 1.4 | 93.6 | 1.4 |
| < .05 WRS | 1.0 | 4.3 | 1.2 | 3.3 | 1.6 | 3.4 |
| correlation | 0.900 | | 0.789 | | 0.712 | |

Table 4: **Quadrant % and correlations.** The top of the table represents the percent of points which fall above and below a significance level of 0.05 for each of the t-test and WRS test. Both tests should reject (p-value < 0.05) the null hypothesis 5% of the time. The last row represents the Pearson correlation for the scatterplot (Figure 4) of p-values for each of the t-test and WRS test.

sion, we will recommend the technique that is able to capture the differential expression most often.

Here, the MAS5 data were not scaled to have a center of zero and standard deviation of one, so the shifts (of 0, 0.5, 1, and 2) were not appropriate for the MAS5 data. Accordingly, for each gene, we multiplied the shift by its median absolute deviation (MAD). The modified shift value was then added to a randomly selected 29 arrays (with a different random selection for each gene). Because the 29 shifted arrays were randomly selected, there should be no underlying differential expression except for the imposed shift value. Other transformation methods (results not shown) produce similar results to MAS5.

For the normal data, the results are consistent for either statistical method (see Table 5). In particular, the t-test is slightly more powerful than WRS, as expected. The t-test should therefore be favored when the data are known to be normal. However, because the gain in power is small, WRS may still be preferable unless the normality assumption can be verified.

When the data are not normally distributed, the two methods for detecting differential expression are much less consistent. For both skewed and heavy tailed data, the WRS test always identified more differentially expressed genes than the t-test. Given that all of the genes were, in fact, differentially expressed, the WRS test was more powerful for the non-normal simulated distributions. Additionally, from tracking the genes which were identified as differentially expressed, the WRS test identified practically all the genes that the t-test identified (see Table 5). Because the t-test genes were almost a subset of the WRS test genes, we conclude the WRS test is a better choice for identifying differentially expressed genes when the data are not normal.

Nonparametric tests tend to be more powerful for data that have heavier tails than the normal distribution. Because the WRS test does not make any assumptions about the underlying error distribution it can be used for any error distribution, and when the data are normal it does almost as well as the t-test. In Section 2 we have argued that, in general, technical error is not normally distributed. It is therefore our recommendation to use robust methods, when possible, to discover differentially expressed genes. When this is not possible, non-robustness of methods should be made explicit.

| | MAS5 Data | | | Normal(0,1) data | | |
|---|---|---|---|---|---|---|
| Shift | t-test % | WRS % | t$\cap$WRS % | t-test % | WRS % | t$\cap$WRS % |
| 0 | 0.046 | 0.049 | 0.035 | 0.054 | 0.053 | 0.043 |
| 0.5 | 0.400 | 0.447 | 0.366 | 0.466 | 0.446 | 0.422 |
| 1 | 0.867 | 0.938 | 0.860 | 0.968 | 0.959 | 0.955 |
| 2 | 0.984 | 0.999 | 0.984 | 1.000 | 1.000 | 1.000 |
| | $t_5$ data | | | $\chi^2_3$ data | | |
| Shift | t-test % | WRS % | t$\cap$WRS % | t-test % | WRS % | t$\cap$WRS % |
| 0 | 0.047 | 0.045 | 0.033 | 0.048 | 0.050 | 0.034 |
| 0.5 | 0.327 | 0.369 | 0.295 | 0.125 | 0.177 | 0.108 |
| 1 | 0.838 | 0.892 | 0.827 | 0.355 | 0.506 | 0.341 |
| 2 | 0.999 | 1.000 | 0.999 | 0.873 | 0.958 | 0.870 |

Table 5: **t-test and WRS test for shifted data**. Two-sample t-tests and Wilcoxon rank sum tests were performed on 10,000 samples where shifts of $\{0, 0.5, 1.0, 2.0\}$ were added to half of the sample. The percentage of rejections for each test at $\alpha = 0.05$, as well as the percentage rejected by both tests are shown.

The conclusions of Section 2 are supported by the behavior of the transformed data to be more consistent with the tail heavy ($t_5$) and skew ($\chi^2_3$) data than with normally distributed data when testing for differential expression (Table 5).

### 4.2 Similarity measures

As with differential expression, different measures of similarity (e.g., Pearson correlation or Spearman correlation) will be consistent for normal data. With non-normal data, however, different measures of similarity will often give inconsistent results (i.e., one high and one low correlation), and will sometimes give opposite results (i.e., one positive and one negative correlation).

To evaluate the degree of inconsistency of different similarity metrics when using relevant non-normal data, we simulated $10,000$ pairs of correlated samples ($n = 59$ arrays in each sample) of Normal(0,1), $t_5$, and $\chi^2_3$ data (Figure 5). For each type of data distribution, we simulated $10,000$ pairs of genes which were correlated at $\rho = 0.0$ and another $10,000$ pairs correlated at $\rho = 0.8$. Pearson correlation is a statistically consistent estimator of the population correlation though Spearman correlation is not. However, Spearman correlation was devised for ranked or noisy data and is often used as a similarity measure in clustering applications.

Figures 5a. and 5d. show the strong consistency (high correlation) of Pearson correlation and Spearman correlation for normally distributed data. The x-axis is the estimated Pearson correlation, the y-axis is the estimated Spearman correlation. If a pair of genes was identified as similar (high correlation) from one measure, the other measure also identified it as similar (also true for gene pairs identified as not similar).

However, if the data were not normally distributed, the methods became much less consistent. For both heavy tailed data and skewed data (Figures 5b. & e. and 5c. & f.) the two measures of similarity were less correlated than the previous normal plots. Particularly with the heavy tailed data, sometimes one measure identified a pair as similar while the other measure did not (see outlying points in Figure 5e). Simulations were repeated for a true population correlation of 0.3 with similar results. Also, Pearson correlation is highly affected by outliers (prevalent in the data, see Figure 1 and (Wang et al., 2007)), so the Pearson correlation estimate of similarity is likely to be inaccurate with heavy tailed data or any data with outliers.

For the Spike-In data, Figure 6 gives both Pearson and Spearman correlations for a random subset of $10,000$ pairs from each of the transformed data sets. That there is a large range of correlations (as to be expected with actual data). For comparison, Normal(0,1) ($\rho = 0$) data were plotted in the lower right corner. The range of the differences (between Spearman and Pearson correlations) is closer to the heavy tailed and skewed data than to that of the normal data. This supports the conclusion of Section 2 that the technical error is not universally normal and it demonstrates that Pearson and Spearman correlation give significantly divergent results on it.

As with tests of differential expression, with microarray data we recommend using robust measures of similarity as input to clustering algorithms. A robust measure of correlation that is consistent for the population correlation, like the translated biweight correlation (Hardin et al., 2007), may give optimal results.

# 5 Discussion

We have demonstrated that transformed microarray data are not, as a rule, normally distributed. Using the 59 technical replicates in the Affymetrix Spike-In data set we showed that none of the standard transformation techniques result in universally normal data. The particular data set is regarded highly enough by the Affymetrix company and academia that it has been used for the development of transformation methods (Cope et al., 2003). For validation, we repeated our methods on another technical replicate data set, the Affymetrix Spike-In 133 data set, with similar results (not shown here). Given the nature of microarray technologies, we presume that our conclusions will hold for other kinds of microarrays, although further study would be required to confirm this. In practice, our work on consulting projects using other microarray data sets is consistent with our conclusion that microarray data are not, in general, normally distributed.

Additionally, our evidence suggests that the main factor leading to non-normality is skewness, and that tail heaviness (presumably in conjunction with peakedness) is also a very substantial factor. Combinations of these factors, as well as other factors require further study.

Finally, not having normal data can yield misleading results for both standard (as shown) and novel methods. We advocate that when testing a novel method, heavier tailed and/or skewed distributions should be used regularly in simulations. It is unclear how newer techniques designed specifically for microarrays (e.g., bagging, boosting, PCA, PLS) will be affected by distributional assumptions, but we hope our results will encourage future researchers to be more realistic in simulating data to test novel methods.

A major goal of this study was to stimulate dialogue about the dangers of assuming the normality of oligonucleotide data. There is ample room for further research, including more precise non-normality characterization, more thorough study of gene dependence, incorporating multiple testing (see Chen, et. al. 2006), and offering a viable alternative distribution to normality for gene expression modeling. The final item is an area of our current investigation.

# Acknowledgments

The authors would like to thank John Kloke for his helpful comments and ideas. The authors would also like to thank Jason's advisor, Xinping Cui, for her support on this project and Barry Arnold for useful discussions on this topic. Conflict of Interest: None declared.

# References

Affymetrix (2002). *Statistical Algorithms Description Document.* Affymetrix, Inc.

Affymetrix (2005). *Guide to Probe Logarithmic Intensity Error (PLIER) Estimation. Technical Report.* Affymetrix, Inc.

Chen, L., Klebanov, L., and Yakovlev, A. (2007). Normality of gene expression revisited. *Journal of Biological Systems*, 15(1):39–48.

Cope, L., Irizarry, R., Jaffee, H., Wu, Z., and Speed, T. (2003). A benchmark for affymetrix genechip expression measures. *Bioinformatics*, 1:1–10.

Giles, P. and Kipling, D. (2003). Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, 19:2254–2262.

Goeman, J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23:980–987.

Hardin, J., Mitani, A., Hicks, L., and VanKoten, B. (2007). A robust measure of correlation between two genes on a microarray. *BMC Bioinformatics*, 8:220.

Hogg, R., Fisher, D., and Randles, R. (1975). A two-sample adaptive distribution-free test. *Journal of the American Statistical Association*, 70:656–661.

Hoyle, D., Rattray, M., Jupp, R., and Brass, A. (2002). Making sense of microarray data distributions. *Bioinformatics*, 18:576–584.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264.

Jarque, C. and Bera, A. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econometric Letters*, 6:255–259.

Klebanov, L. and Yakovlev, A. (2007). How high is the level of technical noise in microarray data? *Biology Direct*, 2(9):1–6.

Koonin, E. V. (2007). Review 3: How high is the level of technical noise in microarray data? *Biology Direct*, 2(9):8–9.

Li, C. and Wong, W. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, 2:Research 0032.

MAQC Consortium (2006a). The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161.

MAQC Consortium (2006b). Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nature Biotechnology*, 24(9):1123–1131.

Millenaar, F. F., Okyere, J., May, S. T., van Zanten, M., Voesenek, L. A., and Peeters, A. J. (2006). How to decide? different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics*, 7(137).

Mushegian, A. (2007). Review 1: How high is the level of technical noise in microarray data? *Biology Direct*, 2(9):6–7.

Nicolau, M., Tibshirani, R., Børresen-Dale, A.-L., and Jeffrey, S. (2007). Disease-specific genomic analysis: identifying the signature of pathologic biology. *Bioinformatics*, 23:957–965.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Royce, T., Rozowsky, J., and Gerstein, M. (2007). Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics*, 23:988–997.

Shapiro, S. and Wilk, M. (1965). An analysis of variance test for normality. *Biometrika*, 52:546–554.

Song, J., Maghsoudi, K., Li, W., Fox, E., Quackenbush, J., and Liu, X. (2007). Microarray blob-defect removal improves array analysis. *Bioinformatics*, 23:966–971.

Tukey, J. (1960). A survey of sampling from contaminated distributions. In Olkin, I., Ghurye, S., Hoeffding, W., Madow, W., and Mann, H., editors, *Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling*, pages 448–485. Stanford, CA: Stanford University Press.

Wang, S. and Zhu, J. (2007). Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, 23:972–979.

Wang, Y., Miao, Z., Pommier, Y., Kawasaki, E., and Player, A. (2007). Characterization of mismatch and high-signal intensity probes associated with affymetrix genechips. *Bioinformatics*, 23:2088–2095.

Wong, D., Wong, F., and Wood, G. (2007). A multi-stage approach to clustering and imputation of gene expression profiles. *Bioinformatics*, 23:998–1005.

Wu, Z., Irizarry, R., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Society*, 99:909–917.

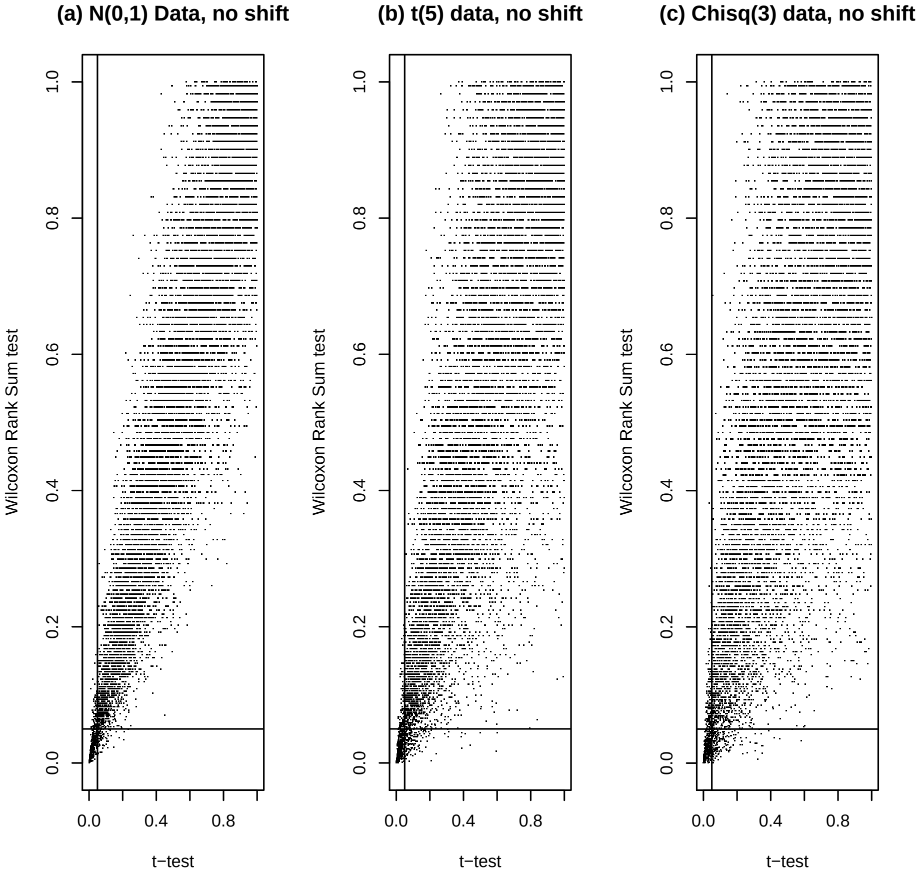

Figure 4: **Scatterplots.** p-values for two tests of differential expression based on data simulated from normal, $t_5$, and $\chi^2_3$ distributions are displayed. Data are simulated so that both groups are centered at zero (no differential expression). Correlations of scatterplots are given in Table 4.

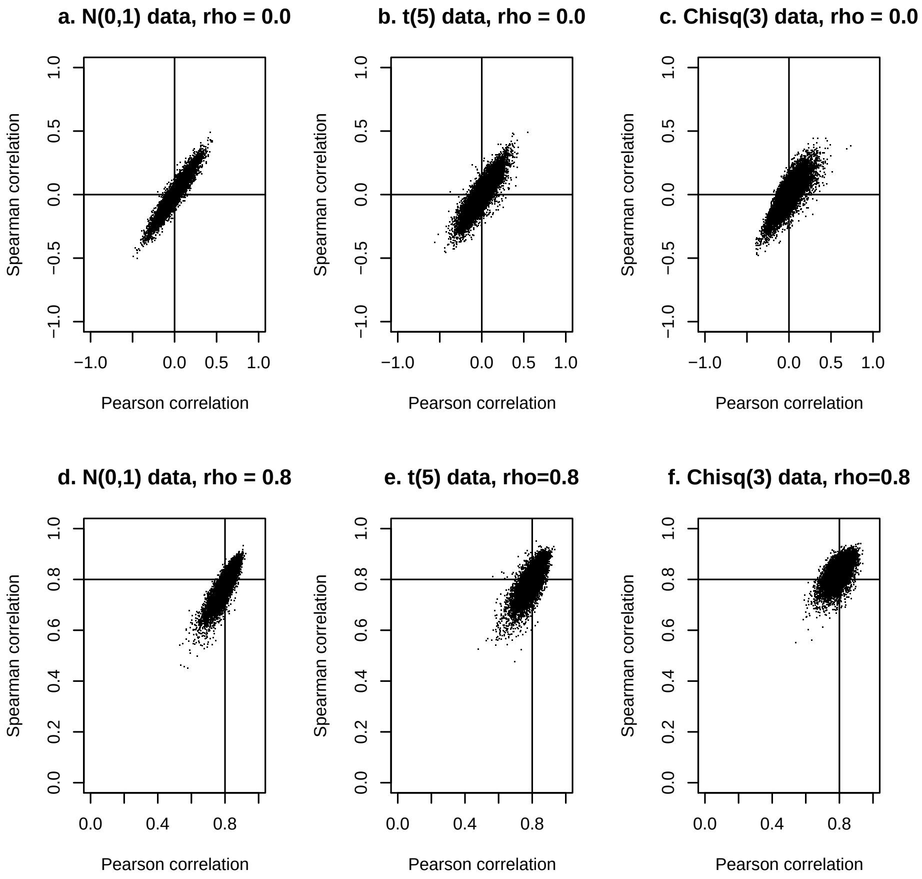

Figure 5: **Correlation Plots**. $\rho = 0.0$ (first row) and $\rho = 0.8$ (second row). Each point represents a pair of genes simulated from a bivariate distribution (Normal(0,1) in first column; $t_5$ in second column; $\chi^2_3$ in third column) with a sample size of $n = 59$. The Pearson correlation and the Spearman correlation give less consistent similarity estimates for the skewed and heavy tailed distributions.

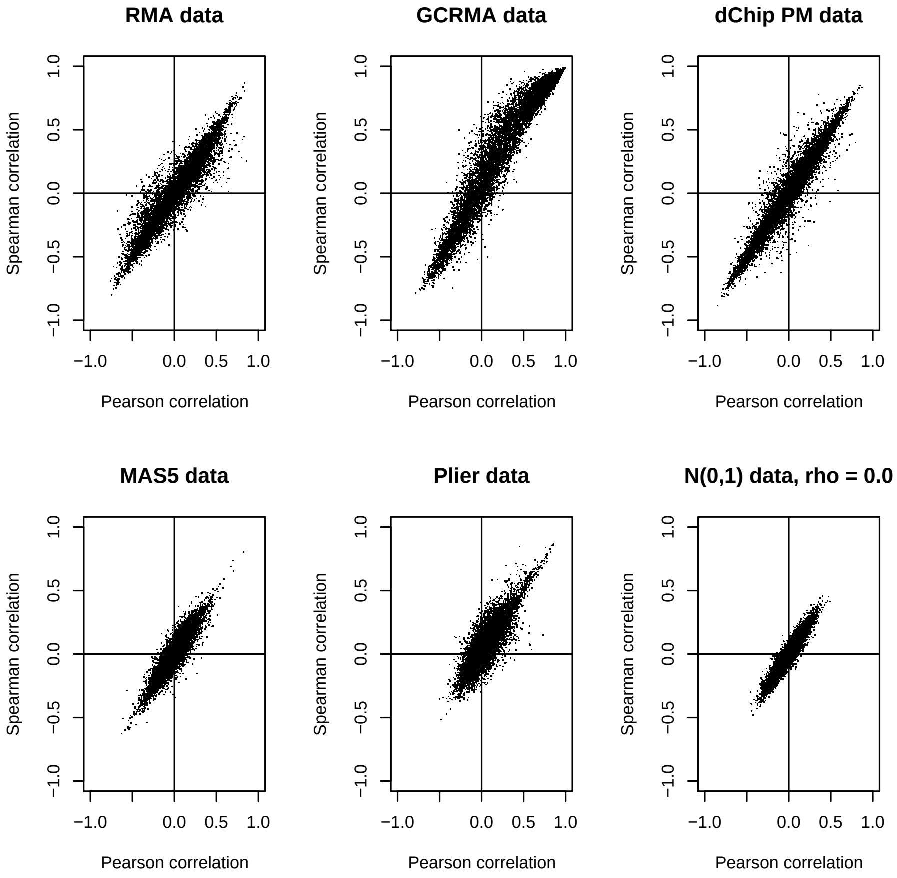

Figure 6: **Correlation Plots for Transformed Data**. Using a random subset of 10,000 pairs of the Spike-In transformed data, all pairwise correlations are shown (both Pearson and Spearman).