BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–6

Oligonucleotide microarray data are not normally distributed

Johanna Hardin [1,*], Jason Wilson [2]

[1]Department of Mathematics, Pomona College, 610 N. College Ave., Claremont, CA 91711,
[2]Department of Statistics, University of California, Riverside, CA 92521

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

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 the science. Because the underlying structure, or distribution, of microarray data is unknown, novel methods are usually tested against the known 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 dataset, oligonucleotide data are shown to not be normally distributed under any of the standard data transformations. The resulting data tend to have heavier tails than normal, and they are often more skewed as well. Using data simulated under three models (normal, heavy tails, and skewed) we show 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, heavy tailed and/or skewed data distributions should be considered in simulations.

Contact: jo.hardin@pomona.edu

1 INTRODUCTION

In the last dozen years, microarray technology has evolved as a powerful tool for gathering genetic information on organisms from yeast to humans. Though the technology is capable of measuring thousands (or tens of thousands) of genes' activities in a single sample simultaneously, the resulting quantitative measurements are often difficult to analyze and interpret. For each gene on a microarray chip the amount of fluorescence (or gene activity) is recorded numerically. The measurements are unitless, but the numbers do give information about relative gene activity when comparing across genes or across samples.

New statistical techniques are constantly being developed for analyzing microarray data. In order to evaluate these techniques, data with known structure is often simulated. Unfortunately, in many cases, the normal distribution is used as test data. For example, in a recent issue of Bioinformatics (Volume 23, Number 8, April 15, 2007), six papers dealt with microarray data. Four of them 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). Raw microarray data are, as a rule, not normally distributed. Even data that have been transformed (note: we prefer the descriptive "transform" to the potentially misleading "normalize") tend to have skew or heavier tails than a normal distribution. Recently, Hoyle et. al. 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.

In a 2003 *Bioinformatics* paper, Giles and Kipling argue that oligonucleotide microarray data are normally distributed Giles and Kipling (2003). Using the same data and similar techniques to those they used, we will argue for the opposite conclusion. Our results are important, because 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.

Even though microarray data are presumed to be normal, there is relatively little literature on evaluating the normality of the data. Part of the problem is that most microarray datasets include large amounts of biological variability and/or small sample sizes. Biological variability makes it difficult to determine the source of the non-normality (non-normal datasets could simply be mixtures of normal datasets). Small samples do not have the power to be able to make claims about the distribution of the data. Fortunately, Affymetrix has provided a dataset of 59 technical replicates. Using this data, we show that such oligonucleotide microarray data are not normally distributed. We go on to show that if the data have heavier tails than normal (e.g., a t distribution) or are skewed (e.g., a Chi-Square distribution), tests for differential expression and measures of similarity can both be affected.

2 METHODS

It is well known that raw microarray data (across all platforms) are highly skewed (usually skewed right) with many extreme

*to whom correspondence should be addressed

values. Often, the log transformation is used to offset the skewness; however, as mentioned in the introduction, the resulting distribution has been almost universally designated as normal. In order to address the irregular structure of microarray data, various transformations have been developed (e.g., dChip C.Li and Wong (2001), Affymetrix MicroArray Suite 5.0 (MAS5) Affymetrix (2002), Robust Multi-Array Analysis (RMA) Irizarry *et al.* (2003), GC-RMA Wu *et al.* (2004), and Probe Logarithmic Intensity ERror (PLIER) Affymetrix (2005) transformations for Affymetrix chips, and loess Yang *et al.* (2002) and variance stabilizing transformations Durbin *et al.* (2002) for cDNA arrays). Though each transformation does improve the data structure, the methods do not necessarily create data which appear to have come from a Gaussian distribution.

2.1 Datasets

We will illustrate the non-normality using the Affymetrix Spike-In (SI) dataset with different transformations. The SI dataset 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. The SI dataset is available on the Affymetrix website (`http:// www.affymetrix.com/ support/ technical/ sample_data/ datasets.affx`) and through the Bioconductor software (`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 and control genes.

Five different algorithms were applied to the SI data. All the algorithms are available in Bioconductor and may have slight differences from the Affymetrix transformation techniques (whose algorithms are proprietary). Expression values were extracted using RMA, GC-RMA, MAS5, PLIER, and PM-only dChip.

2.2 Test of normality

In order to determine whether or not the data are normal we: applied two tests of normality (Shapiro-Wilks Shapiro and Wilks (1965) and Jarque-Bera Jarque and Bera (1980)); investigated qqplots and their associated correlations; and plotted measures of skewness and peaked-ness (skew coefficient, kurtosis, and Hogg's Q2 Hogg *et al.* (1975)). All procedures were done using the statistical software R R Development Core Team (2007).

3 IMPLEMENTATION

3.1 Affymetrix Spike-In Data

The SI data was standardized so that each gene is 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 qq-plots are sensitive to changes in shift or scale. Additionally, skewness coefficient, kurtosis coefficient, and Hogg's Q2 are all also shift and scale invariant.

For each of the methods applied to the data, the 59 arrays are considered to be technical replicates; 12,543 genes can be used to evaluate the structure of the data. So that the variability of the SI data can be compared to the variability of known samples, 12,543 samples of size $n = 59$ were calculated for each of the Uniform(0,1) and Normal(0,1) distributions.

Transformation	Shap-Wilks	JB
RMA	24.5%	26.0%
GCRMA	46.8%	47.2%
MAS5	46.2%	33.6%
dChip	29.5%	25.9%
PLIER	20.1%	15.2%
t_5	39.3%	43.7%
χ^2_3	99.6%	92.5%
Normal(0,1)	5.0%	3.7%

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

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%
χ^2_3	95.8%	99.8%
empirical Normal	1%	10%

Table 2. Correlation quantiles, each entry represents the percent of genes whose normal qq-plot is correlated 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).

3.1.1 Significance Tests of Normality Using both the Shaprio-Wilks (SW) and Jarque-Bera (JB) tests, each of the 12,543 genes was tested against normality. If the data are normally distributed, there should be approximately 5% of the genes failing significance tests at a 0.05 level. In fact, for the normal simulated data, there were approximately 5% of the genes failing the tests of normality. 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. There is no transformation method that truly normalizes the data.

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. Using table of Looney and Gulledge table (1985), for normally distributed data, 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 more genes less than the given cutoff for each of the transformation techniques. That is, the data were less correlated with normal quantiles than they would be if the data were in fact taken from a normal population (conclusions are opposite those of Giles and Kipling 2003, Figure 1).

3.2 Normality statistics

3.2.1 Skewness Normal data are not skewed; that is, they have a skewness coefficient of zero. Skewness is typically measured as the

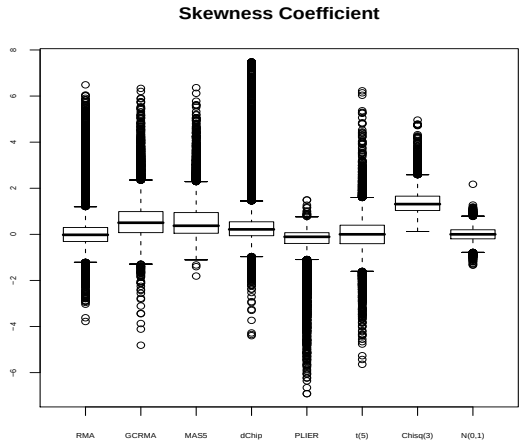

Fig. 1. Skewness boxplots, PLIER is skewed left while the other transformations are skewed right.

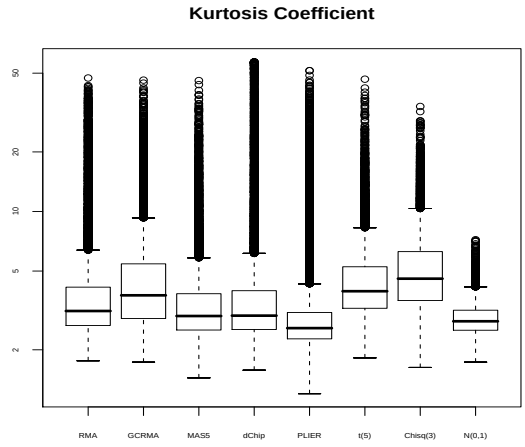

Fig. 2. Kurtosis boxplots, all the transformation methods have a large number of genes with heavier tails than expected under normality.

third central moment divided by the cube of the standard deviation, $\frac{\mu_3}{\sigma^3}$.

In Figure (1), the oligonucleotide data have genes which are much more skewed than expected with normal data. The PLIER transformation seems to address the right skewness but has a tradeoff of being left skewed.

3.2.2 Heavy Tails/Kurtosis Another way to observe a departure from normality is through the peakedness of the bell shaped curve or the heaviness of the tails of the distribution. We measured kurtosis (peakedness) using the 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 Figure (2) we see that the kurtosis is much higher for the oligonucleotide data than for the normal data. That is, the variance of the genes comes from extreme values rather than modest deviations. In Figure (3) we see the same effect which is that many genes have much heavier tails than any of the normal samples.

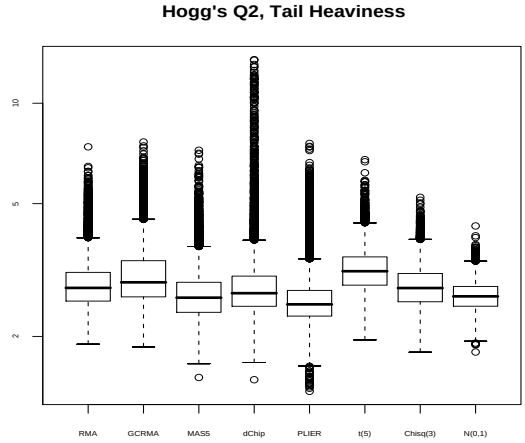

Fig. 3. Tail heaviness boxplots, for each type of data transformation, Hogg's Q2 is calculated. Again, all methods have some genes with much heavier tails than normal.

3.3 Simulated Data

After arguing that microarray data are not, in general, normally distributed, the next logical question is, "Does it matter?" The answer is "yes." We have given references (in the introduction) to papers which simulate normal data to test out novel techniques; we do not know what the effects of non-normality will be on novel techniques. However, the typical practitioner is probably more interested in the effects of non-normality to basic procedures which deal with differential expression and gene clustering.

In order to test the effects of non-normality, we simulated three types of data: Normal(0,1), t_5 (heavy tails), and χ^2_3 (skewed). The simulated datasets were used to evaluate basic procedures for testing differential expression and computing similarity measures.

3.4 Differential Expression

If 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 genes are not normally distributed, methods for evaluating differential expression will give a different ordering of most significantly differentially expressed.

To evaluate the effects of non-normality in testing differential expression, we simulated Normal(0,1), t_5, and χ^2_3 data. We used 30 samples (arrays) in one group and 29 samples in the other (to be consistent with the Spike-In data which contained 59 samples). For each simulation, we repeated the trial 10,000 times. We simulated both groups that were not differentially expressed (no shift) and genes we knew to be 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.

3.4.1 Consistency 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, in fact, differentially expressed, both tests should produce significance. In each of the first set of simulations (Figure 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. We see that for each of the distributions and each of the tests, there is a correct amount of error at about 5% (see Table 3); also, there is a nice spread of p-values across the range of zero to one along each axis.

However, for non-normal data, consistency across the two methods was much less. For normal data (Figure 4), it seems as though the data were somewhat consistent. That is, the p-values for the t-test seemed to be moderately correlated with the p-values for the WRS test. For the skew and heavy tailed data (Figure 4), there was weaker correlation between the p-values of tests that were assumed to be measuring the same thing. When trying to measure differential expression, not all tests are equally valid. The underlying assumptions of the test are important to keep in mind, as different tests can produce extremely different results.

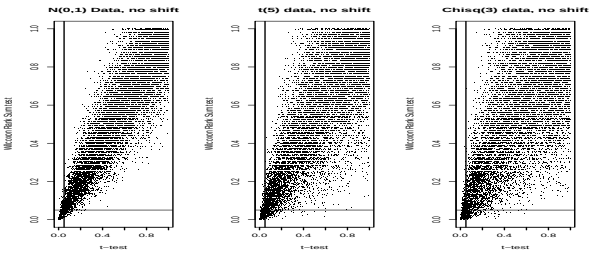

Fig. 4. Scatterplot of p-values for two different tests based on data simulated from normal, t_5, and χ^2_3 distributions. Data are simulated so that both groups are centered at zero (no difference). Both tests should reject (p-value < 0.05) the null hypothesis 5% of the time.

3.4.2 Power In the previous section 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.

	MAS5 Data			Normal(0,1) data		
Shift	t-test %	WRS %	t∩WRS %	t-test %	WRS %	t∩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			Chisquare(3) data		
Shift	t-test %	WRS %	t∩WRS %	t-test %	WRS %	t∩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 3. t-test and WRS test for shifted data, for the parametric data, 10000 rows of 30 deviates were generated for group 1 and 10000 rows of 29 deviates for group 2. Two-sample t-tests and Wilcoxon rank sum tests were performed row-wise. Each test was repeated with shifts of {0, 0.5, 1.0, 2.0} added to the center of group 2. The percentage of rejections for each test at $\alpha = 0.05$, as well as the percentage of rows rejected by both tests are shown. For the MAS5 data all 12543 genes were used and for each gene: (i) the 59 probes were randomly divided into groups of size 30 and 29, and (ii) the shifts were scaled up according to median absolute deviation. Other transformation methods (results not shown) produce similar results to MAS5.

Using the same setup as in Section 3.4.1, 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 expression, we will recommend the technique that is able to capture the differential expression most often.

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 the median absolute deviation (MAD). Then we added the modified shift value 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.

As before, with normal data, the results are fairly consistent for either method (see Table 3). The t-test and the WRS test gave similar p-values for each of the simulated datasets. There does not seem to be a strong argument for one method over another for normally distributed data.

However, when the data are not normally distributed, the two methods for detecting differential expression are not consistent. For both skewed and heavy tailed data, the WRS test consistently identified more differentially expressed genes than the t-test. Because all of the genes were, in fact, differentially expressed, the WRS test was clearly more powerful for the 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 3). Because the t-test genes were 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. We have argued that, in general, microarray data are 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.

3.5 Similarity measures

As with differential expression, with normal data, measures of similarity (e.g., Pearson correlation or Spearman correlation) will be consistent. 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 non-normal data, we simulated 10,000 pairs of correlated samples (59 arrays in each sample) of Normal(0,1), t_5, and χ^2_3 data. For type of data distribution, we simulated 10,000 pairs of genes which were correlated at ρ = 0.0 and another 10,000 pairs correlated at ρ=0.8.

3.5.1 Consistency As with tests for differential expression, two different methods for measuring the similarity between two genes can give inconsistent results. In Figure 5, 10,000 pairs of genes were simulated, each from populations with a correlation of 0.0 (top row) or 0.8 (bottom row). 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. Each point represents a pair of genes simulated from populations with a correlation of 0.0 or 0.8. 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, it was possible to see one measure identifying the pair as similar while the other measure identified them as not similar.

3.5.2 Accuracy For the standard procedure (typically used in gene clustering) of determining similarities, Pearson correlation and Spearman correlation were visually more consistent with normal data and less consistent with non-normal data. Note that with highly correlated skewed data (Figure 5f.) neither method was obviously more accurate. However, with highly correlated heavy tailed data (Figure 5e.), the Spearman correlation measures were closer to the truth (0.8) than the Pearson correlation. Pearson correlation was highly affected by outliers (well known to be prevalent in microarray data), so the Pearson correlation estimate of similarity is likely to be inaccurate with heavy tailed data or any data with outliers.

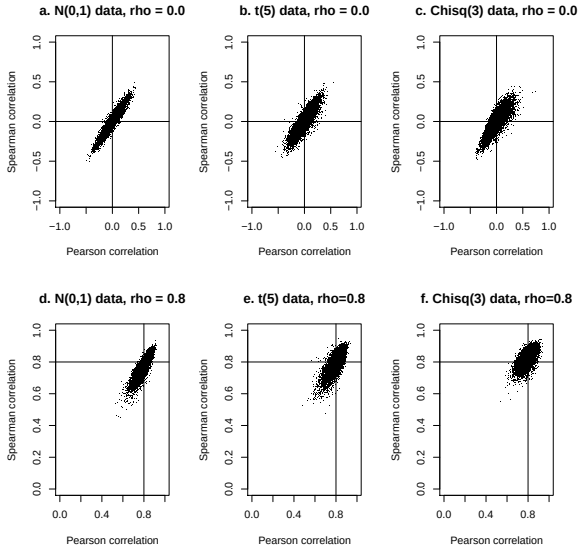

Fig. 5. Correlation Plots, with rho=0 (first row) and rho=0.8 (second row). Each point represents a pair of genes simulated from a bivariate distribution (Normal(0,1) in first column; t_5 in second column; χ^2_3 in third column) with a sample size of $n = 59$. Here, the Pearson correlation and the Spearman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

In Figure 6 we see that there is a large range of correlations (as to be expected with actual data). However, the range of the differences is more similar to the differences seen in heavy tailed and skewed data than to that of the normal data. As a comparison, Normal(0,1) ($\rho = 0$) data were plotted in the lower right corner. The differences in Pearson versus Spearman across the center of the clouds of points are much wider for the transformed data than for the Normal(0,1) data.

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. The simulations were repeated for a true population correlation of 0.3, and results were similar.

4 DISCUSSION

In conclusion, microarray data are not, as a rule, normally distributed. We have used the 59 technical replicates in the Affymetrix Spike-In dataset to show that none of the standard transformation techniques result in normal data. We repeated our methods on the Affymetrix Spike-In 133 dataset with similar results (not shown here). Additionally, our work on consulting projects using other microarray datasets are consistent with our conclusion that microarray data are not, in general, normally distributed.

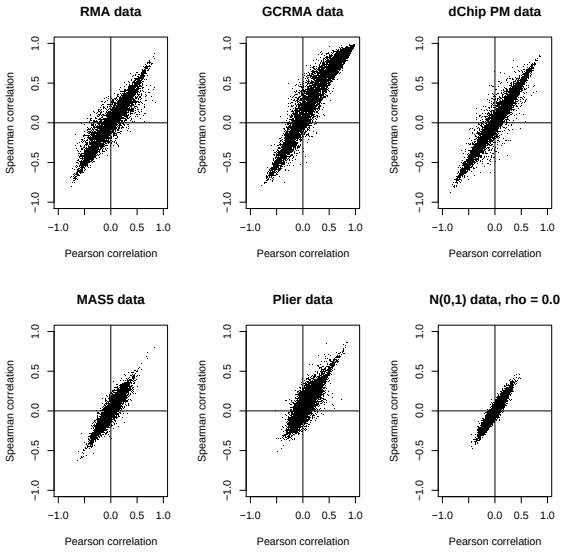

Fig. 6. Correlation Plots for Transformed Data, using a random subset of 10,000 pairs of the Spike-In transformed data, we plotted all pairwise correlations (both Pearson and Spearman).

Not having normal data can yield misleading results for both standard (as shown here) and novel methods. Because our experience is that microarray data tend to have much heavier tails than normal (and that heavy tails and outliers are more disruptive to methods than skewness), we would argue that when testing a novel method, it is important to simulate heavier tailed and/or skewed distributions as "test" microarray data. It is unclear how newer techniques designed specifically for microarrays (e.g., bagging, boosting, PCA, PLS, q-values) will be affected by distributional assumptions, but we hope the results will encourage future researchers to be more realistic in simulating data to test out novel methods.

5 ACKNOWLEDGMENTS

The authors would like to thank John Kloke for his helpful comments and ideas. The authors would also like to thank Xinping Cui and Barry Arnold for their discussions with us on this topic.

REFERENCES

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

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

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

Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**, S105–S110.

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*, pages 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*. in press.

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

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.

Looney, S. and Gulledge, Jr., T. (1985). Use of the correlation coefficient with normal probability plots. *The American Statistician*, **39**, 75–79.

Nicolau, M., Tibshirani, R., Børresen-Dale, A.-L., and Jeffrey, S. (2007). Disease-specific genomic analysis: identifying the signature of pathologic biology. *Bioinformatics*, pages 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*, pages 988–997.

Shapiro, S. and Wilks, 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*, pages 966–971.

Tukey, J. (1960). A survey of sampling from contaminated distributions. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, and H. Mann, 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*, pages 972–979.

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

Yang, Y., Luu, P., David, L., Peng, V., Ngai, J., and Speed, T. (2002). Normalization for cDNA microarray data: a robust composite method for addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**, e15.