A comparison of fold-change and the t-statistic for microarray data analysis

Daniela M. Witten [*]

Department of Statistics
Stanford University, Stanford, CA 94305, USA
dwitten@stanford.edu

Robert Tibshirani

Department of Health Research & Policy
and Department of Statistics,
Stanford University, Stanford, CA 94305, USA

February 1, 2007

Abstract

It has recently been suggested that differentially-expressed genes in a microarray experiment are best identified using fold-change, rather than a t-statistic, because the former results in lists of differentially-expressed genes that are more reproducible (Shi et al. 2005, Guo et al. 2006, MAQC Consortium 2006). We argue that reproducibility does not imply accuracy, and we show that the question of whether to use fold-change or a modified t-statistic to identify differentially-expressed genes is essentially biological, rather than statistical. Finally, we demonstrate that the ordinary t-statistic is inferior to the modified t-statistic in terms of both accuracy and reproducibility, and therefore should be avoided in the analysis of microarray data.

[*]To whom correspondence should be addressed. Phone: (650) 248-6323. Fax: (650) 725-8977.

1   Introduction

In the analysis of microarray experiments, the method used to identify genes
that are differentially expressed between conditions can greatly affect the set of
genes that are identified; see, for instance, Jeffery et al. (2006). We used real and
simulated microarray data in order to assess the reproducibility and accuracy of
differentially-expressed gene lists based on the t-statistic, a modified t-statistic,
and two different versions of fold-change. As previous papers have indicated, fold-
change results in more reproducible gene lists than do the ordinary and modified
t-statistics (Shi et al. 2005, Guo et al. 2006, MAQC Consortium 2006). Through
the use of an artificial statistic, we demonstrate that a gene list's reproducibility
does not imply its accuracy. Simulations suggest that whether fold-change or a
modified t-statistic results in more accurate gene lists depends on whether one
is interested in an absolute change in gene expression or in the change in gene
expression relative to the underlying noise in the gene. Therefore, a researcher's
decision to use fold-change or a modified t-statistic should be based on biological,
rather than statistical, considerations. Finally, we determine that the ordinary t-
statistic does not exceed the modified t-statistic in terms of either reproducibility
or accuracy, and thus should not be used in microarray analysis.

2   Statistical measures of differential expression

Let x_{ij} and y_{ij} denote the log_2 expression levels of gene i in replicate j in the
control and treatment, respectively. We define the ordinary two-sample t-statistic
as

$$T_i = \frac{\overline{x_{i\cdot}} - \overline{y_{i\cdot}}}{s_i} \tag{2.1}$$

where s_i is the standard deviation of the replicates for gene i. The modified t-statistic is defined as

$$T_i' = \frac{\overline{x_{i\cdot}} - \overline{y_{i\cdot}}}{s_i + s_o} \qquad (2.2)$$

where s_o is a constant chosen to minimize the coefficient of variation of T_i'. In our analysis, we chose s_o and computed the modified t-statistic using Significance Analysis of Microarrays (Tusher et al. 2001).

There are two definitions of fold-change in the literature. The standard definition of the fold-change for gene i is (e.g. Tusher et al. 2001)

$$FC_i = \frac{\overline{x_{i\cdot}'}}{\overline{y_{i\cdot}'}} \qquad (2.3)$$

where x_{ij}' and y_{ij}' are the raw expression levels of gene i in replicate j in the control and treatment, respectively. On the other hand, in Guo et al. (2006) and Choe et al. (2005), the fold-change for gene i is defined as

$$FC_i = \overline{x_{i\cdot}} - \overline{y_{i\cdot}} \qquad (2.4)$$

We will refer to these versions of fold-change as FC_{ratio} and $\text{FC}_{difference}$, respectively. It is worth noting that as s_o in the denominator of the modified t-statistic is increased, the resulting gene ordering approaches that obtained using $\text{FC}_{difference}$. Of course, the ordinary t-statistic results from setting s_o equal to zero. Therefore, these three statistics can be thought of as the results of plugging three values of s_o into a single function of s_o and the data.

In order to demonstrate that a statistic with high reproducibility is not nec-

essarily accurate, we construct the artificial statistic

$$P_i = (\overline{x_{i\cdot}})^3 - (\overline{y_{i\cdot}})^3 \tag{2.5}$$

where P stands for "power". We do not suggest this statistic for practical use.

3 Overview of the genes selected using the different measures

In order to gain an understanding of the circumstances that result in extreme values of the various measures of differential gene expression, we simulated data so that the log expression values of the replicates for each gene were normally distributed, and computed $FC_{difference}$, FC_{ratio}, the ordinary t-statistic, and the modified t-statistic for each gene. The genes with the most extreme 10% of values for these statistics are plotted in Figure 1. The ordinary t-statistic selects genes with low standard deviations, the fold-changes select genes with large shifts between control and treatment, and the modified t-statistic selects genes with low standard deviations and large shifts. Since the fold-changes and the ordinary t-statistic select completely different sets of genes, a researcher must decide whether a gene's importance is best quantified by the shift in expression or by the shift relative to the standard deviation. The modified t-statistic provides a middle ground between these two possibilities.
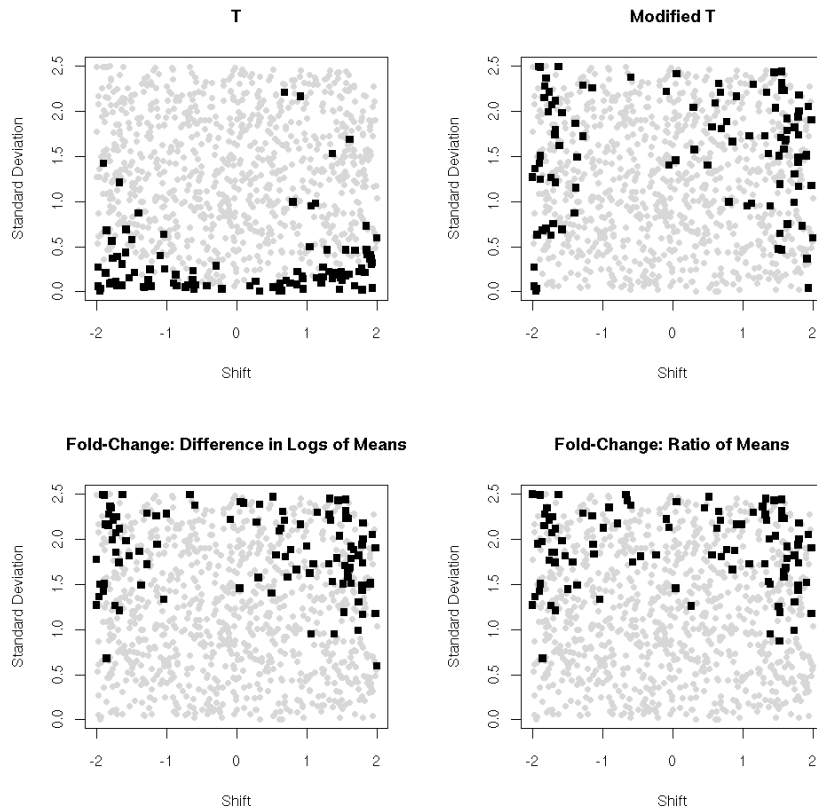
Figure 1: *Data were generated so that the logs of the replicates for each gene were normally distributed. The log standard deviations and log shifts of the simulated genes are shown on the x and y axes. The ten percent of genes with the most extreme values of each statistic are plotted in black; the remaining genes are plotted in gray.*

4 Simulated and real data

Concordance is defined as the percentage of genes in one gene list that also are present in another gene list. For instance, we can compute the concordance between the list of 50 most differentially-expressed genes in one data set and the list of 50 most differentially-expressed genes in another data set. We can measure reproducibility of gene lists obtained using a given statistic by computing the concordance between the gene lists constructed using that statistic for two data sets obtained by performing the same experiment twice, or for two halves of a single data set; this was done in Guo et al. (2006). On the other hand, in order to assess the accuracy of a gene list obtained using a given statistic, we examine a single data set, and compute the concordance between a gene list obtained using that statistic and the true gene list that reflects the actual extent of differential expression of the genes in the data set. Of course, this true gene ordering is unknown for real data. We simulated data in order to compare the accuracies of the statistics in question.

We used two different simulations to generate our data, both of which make reference to a real data set. Our first simulation was based on a normal shift model.

Simulation 1.

1. Fix n, the number of differentially-expressed genes, and k, the number of standard deviations that comprise the shift between control and treatment expression levels for the differentially-expressed genes.

2. For each gene i, draw the mean μ_i from the uniform distribution on the range of observed means of the log data, and draw the standard deviation σ_i from the uniform distribution on the range of observed standard deviations of the log data.

3. For each gene i, simulate log control data $N(\mu_i, \sigma_i{}^2)$ and log treatment data $N(\mu_i + k\sigma_i 1_{i\leq n}, \sigma_i{}^2)$.

Our second simulation maintains the observed control and treatment means and standard deviations for each gene. This preserves an important feature of real data: the shift in expression and the standard deviation are highly correlated.

Simulation 2.

1. For each gene, compute the mean μ_c of the log control data, the standard deviation σ_c of the log control data, the mean μ_t of the log treatment data, and the standard deviation σ_t of the log treatment data.

2. Simulate log control and treatment data as $N(\mu_c, \sigma_c^2)$ and $N(\mu_t, \sigma_t^2)$.

In effect, Simulation 2 uses the observed standard deviation and mean of the actual data as the true standard deviation and mean of the simulated data.

In addition to the data that we simulated, we examined the rat toxicogenomic data set of Guo et al. (2006); in particular, we compared the comfrey-treated liver samples to the control liver samples for the Affymetrix sites. We also used the prostate data set of Singh et al. (2002).

5 Analysis of the accuracy of the different measures

Under the normal shift model (Simulation 1), accuracy (as a function of the number of genes selected) is defined as the percentage of the selected genes with the most extreme values of a given statistic that truly are differentially expressed. (Recall that in this simulation, only a subset of the genes are differentially expressed between control and treatment). Figure 2 indicates that the t-statistics and $FC_{difference}$ have roughly the same accuracy under this model. FC_{ratio} has

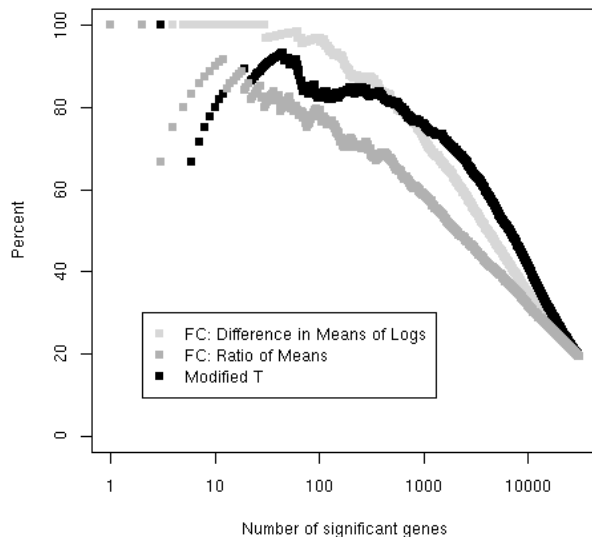lower accuracy. More details can be seen in Supplementary Figure 1.

Figure 2: *The data were generated using Simulation 1 based on the rat toxicoge-nomic data of Guo et al. (2006), with 6,000 out of 31,000 genes differentially expressed, a shift of one standard deviation between treatment and control, and six control and six treatment replicates. The modified t-statistic and $FC_{difference}$ have roughly the same accuracy. FC_{ratio} has slightly lower accuracy. The accuracy of the ordinary t-statistic is equivalent to that of the modified t-statistic.*

In order to compute the accuracy for Simulation 2, we must first develop a notion of what it means for a gene to be differentially expressed between control and treatment. In this case, in contrast with Simulation 1, no gene has identical true control and treatment means. However, for some of the genes, the control and treatment expression are similar enough that the gene is not of interest to a biologist. One way to quantify the extent to which a gene is differentially expressed involves the absolute difference between the true log control and log treatment means, $|\mu_c - \mu_t|$. Another method divides this difference by the pooled standard deviation, σ, resulting in the quantity $|\frac{\mu_c - \mu_t}{\sigma}|$. The former method quantifies the difference in expression, which is purportedly the quantity of interest, whereas the

9

latter attempts to standardize for noisy genes in which large absolute differences in expression may have less meaning due to high levels of variance in expression. The two methods can result in vastly different gene orderings.

Under Simulation 2, when $|\mu_c - \mu_t|$ defines the ordering of differential gene expression, both versions of fold-change are more accurate than the t-statistics. When $|\frac{\mu_c - \mu_t}{\sigma}|$ defines the gene ordering, the t-statistics are more accurate than the fold-changes (Figure 3 and Supplementary Figures 2 and 3). This reflects the fact that the fold-changes and the former ordering concern themselves only with the numerical change in average gene expression, whereas both the t-statistics and the latter ordering essentially compute the number of standard deviations of shift between treatment and control.
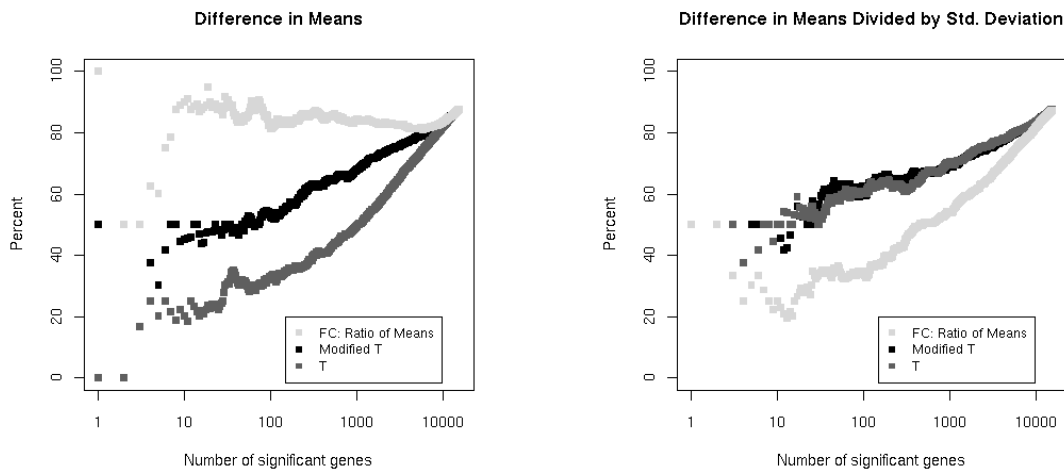
Figure 3: *The data were generated using Simulation 2 based on the rat toxicogenomic data with six control and six treatment replicates (Guo et al. 2006). The figures show the accuracy of the statistics, or the extent to which the gene lists based on a given statistic capture the true ordering of differential gene expression. On the left, this "true" ordering is defined by $|\mu_c - \mu_t|$; on the right, it is defined by $|\frac{\mu_c - \mu_t}{\sigma}|$. In the former case, the fold-changes are more accurate, and in the latter case, the t-statistics are more accurate. Though only one version of fold-change is plotted, the accuracies of the two versions are almost identical.*

	Control	Treatment	$FC_{difference}$	FC_{ratio}	T
Gene 1	150, 200, 250	1, 50, 100	3.51	3.97	1.69
Gene 2	101.1, 101.2, 101.3	100.1, 100.2, 100.3	0.014	1.01	12.25

Table 1: *Three replicates of two genes under control and treatment conditions. Gene 1 results in high values for both versions of fold-change and a lower value of the t-statistic, whereas Gene 2 results in low values of the fold-changes and a high-value of the t-statistic.*

The inherent difference in orderings obtained via fold-changes and the t-statistic can be seen in the following example. Table 1 presents two genes, each of which has three control and three treatment replicates. Gene 1 results in a more extreme value for both versions of fold-change, whereas Gene 2 results in a more extreme value for the t-statistic. Though the fold-changes and the t-statistic result in quite different orderings here, it is not clear which is correct. Gene 2's data are less likely than Gene 1's under the null hypothesis that the control and treatment observations all are independent and identically distributed from the same normal distribution. From this perspective, Gene 2 appears to be more differentially expressed. On the other hand, a biologist might not be particularly interested in an expression shift that is only 1% of the total expression; therefore, Gene 1 might seem to be more differentially expressed.

From this perspective, the question of whether the fold-changes or a modified t-statistic results in more accurate gene orderings is really a biological one, rather than a statistical one, as it depends on what types of expression differences between control and treatment have biological relevance.

In the data that we examined, the accuracy of the ordinary t-statistic never exceeded those of the fold-changes or the modified t-statistic. Therefore, we cannot recommend the use of the ordinary t-statistic on the basis of its accuracy.

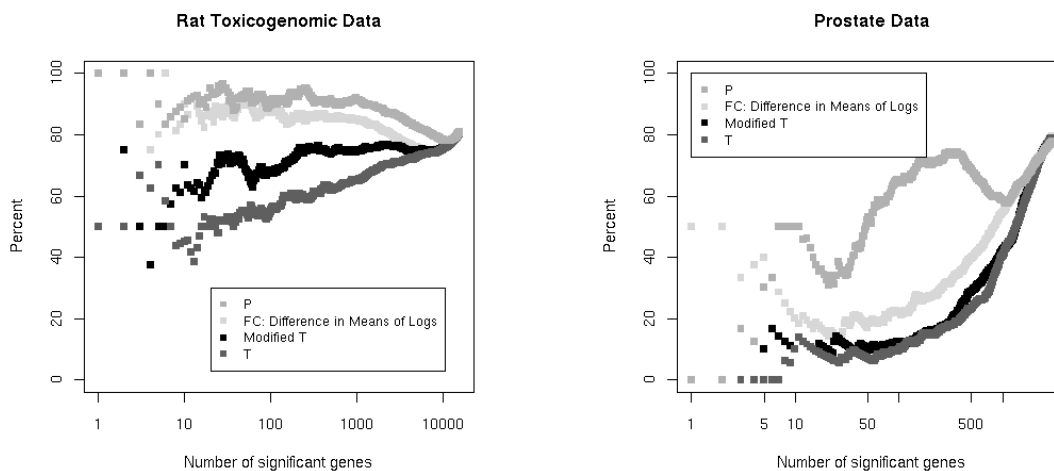# 6 Analysis of the reproducibility of the different measures

Figure 4: *Both figures show the concordance between two gene lists obtained using two separate sets of six control and six treatment replicates. The figure on the left is based on the rat toxicogenomic data of Guo et al. (2006), whereas the figure on the right is based on the prostate data of Singh et al. (2002). In both cases, P is more reproducible than the other statistics. The two versions of fold-change have almost identical reproducibility; only $FC_{difference}$ is plotted.*

Previous studies have shown that $FC_{difference}$ results in more reproducible gene lists than the ordinary and modified t-statistics (Shi et al. 2005, Guo et al. 2006, MAQC Consortium 2006). This held true in the simulated data as well as in the real data sets that we considered. In fact, both versions of fold-change had very high reproducibility. However, we were able to construct a statistic, P, that is even more reproducible than fold-change under all of the circumstances that we investigated (Figure 4, Supplementary Figure 4). Similarly, statistics obtained by replacing the third power in the definition of P with numbers between 1 and 4 also result in quite reproducible gene lists.

The sample mean itself is not a very robust statistic, in that a single observation that differs from the true mean can greatly affect the sample mean. The statistic P cubes the sample mean, thereby substantially increasing the effect of an unusual observation. P can be considered an "extreme" version of fold-change. On the other hand, the statistic obtained by replacing the mean with the median in the equation for fold-change is a "muted" version of fold-change, and is less reproducible than fold-change. These observations suggest that it is fold-change and P's non-robustness that render them reproducible, since noisy genes will consistently result in large values for fold-change and P. Therefore, fold-change and P can be seen as reproducible measures of the noise of a gene. Of course, biologists are interested in measuring differential gene expression, and not noise.

In fact, P has low accuracy in both simulations (Figure 5, Supplementary Figures 1, 2, 3). This emphasizes the distinction between reproducibility and accuracy, and the fact that the former does not imply the latter. Therefore, in making a choice between fold-change and a modified t-statistic, we should not be swayed by the fact that fold-change is more reproducible.

The ordinary t-statistic's reproducibility never exceeded that of the modified t-statistic under the circumstances that we considered.

7   Conclusions

We summarize our findings as follows:

1. Fold-change is defined in two ways in the literature: as the ratio of the mean control and mean treatment observations, and as the difference of the mean log control and mean log treatment data. For the most part, the behaviors of the two versions of fold-change are quite similar. Since $FC_{difference}$ leads
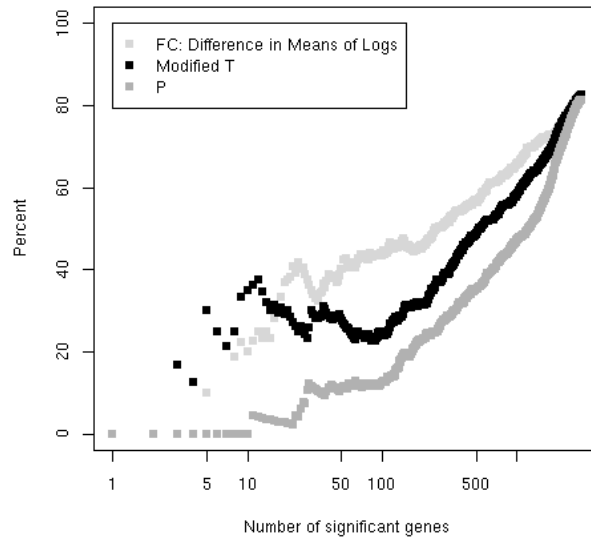
Figure 5: *The data were generated using Simulation 2 based on the prostate data of Singh et al. (2002) with six control and six treatment replicates. The figure shows the accuracy of three statistics when the true gene ordering is defined by $|\mu_c - \mu_t|$. P is not very accurate.*

to the same gene ordering that would result from letting the term s_o in the denominator of the modified t-statistic become arbitrarily large, it follows that $\text{FC}_{difference}$ and the modified t-statistic both are generalized modified t-statistics with two different choices of s_o. Indeed, it is possible that some intermediate value of s_o might result in higher accuracy than either of these statistics.

2. Under the circumstances that we considered, the ordinary t-statistic was never superior to the modified t-statistic in terms of accuracy or reproducibility. Therefore, we suggest that the modified t-statistic or fold-change be used instead of the ordinary t-statistic in microarray data analysis.

3. A statistic's reproducibility does not imply its accuracy; in fact, it is easy to construct a statistic with very high reproducibility and low accuracy. The issues of reproducibility and accuracy should be kept separate when evaluating the performance of a statistic.

4. In the analysis of real microarray data, there is no correct answer as to whether fold-change or the modified t-statistic should be used. However, the question of which statistic to use is very important, as the choice of statistic can dramatically affect the set of genes that is selected. A researcher should choose the measure of differential expression based on the biological system of interest. If large absolute changes in expression are relevant to the system, then fold-change should be used; on the other hand, if changes in expression relative to the underlying noise are important, then a modified t-statistic is preferable.

8 Acknowledgments

References

Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M. & Halfon, M. S. (2005), 'Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset', *Genome Biology* **6**, R16.

Guo, L., Lobenhofer, E. K., Wang, C., Shippy, R., Harris, S. C., Zhang, L., Mei, N., Chen, T., Herman, D., Goodsaid, F. M., Hurban, P., Phillips, K. L., Xu, J., Deng, X. T., Sun, Y. M. A., Tong, W. D., Dragan, Y. P. & Shi, L. M. (2006), 'Rat toxicogenomic study reveals analytical consistency across microarray platforms', *Nature Biotechnology* **24**, 1162–1169.

Jeffery, I. B., Higgins, D. G. & Culhane, A. C. (2006), 'Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data', *BMC Bioinformatics* **7**, 359.

MAQC Consortium (2006), 'The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements', *Nature Biotechnology* **24**, 1151–1161.

Shi, L. M., Tong, W. D., Fang, H., Scherf, U., Han, J., Puri, R. K., Frueh, F. W., Goodsaid, F. M., Guo, L., Su, Z. Q., Han, T., Fuscoe, J. C., Xu, Z. A., Patterson, T. A., Hong, H. X., Xie, Q., Perkins, R. G., Chen, J. J. & Casciano, D. A. (2005), 'Cross-platform comparability of microarray technology: Intra-platform consistency and appropriate data analysis procedures are essential', *BMC Bioinformatics* **6**, S12.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R. & Sellers, W. R. (2002), 'Gene expression correlates of clinical prostate cancer behavior', *Cancer Cell* **1**, 203–9.

Tusher, V. G., Tibshirani, R. & Chu, G. (2001), 'Significance analysis of microarrays applied to the ionizing radiation response', *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116–5121.

List of Figures

List of Tables

Supplementary materials to "A comparison of fold-change and the t-statistic for microarray data analysis"

Daniela M. Witten [*]
Department of Statistics
Stanford University, Stanford, CA 94305, USA
dwitten@stanford.edu

Robert Tibshirani
Department of Health Research & Policy
and Department of Statistics,
Stanford University, Stanford, CA 94305, USA

January 31, 2007

---
[*]To whom correspondence should be addressed. Phone: (650) 248-6323. Fax: (650) 725-8977.
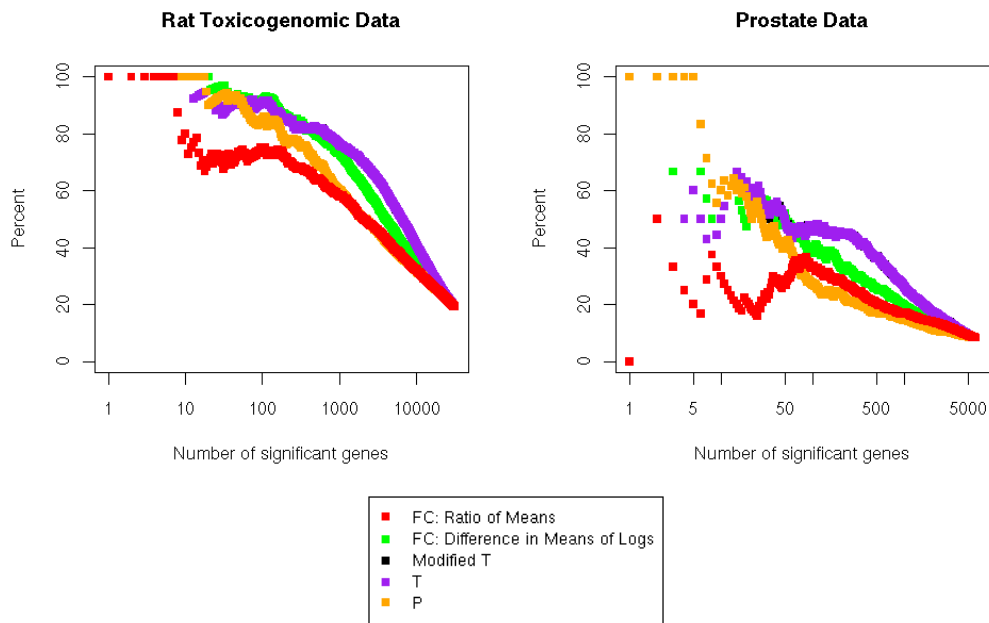
Figure 1: *Both figures show the accuracy of the various statistics on data generated using Simulation 1, with a shift of one standard deviation between treatment and control, and six control and six treatment replicates. The figure on the left is based on the rat toxicogenomic data of Guo et al. (2006) with 6,000 of 31,000 genes differentially-expressed, and the figure on the right is based on the prostate data of Singh et al. (2002) with 500 of 6,000 genes differentially-expressed. The modified t-statistic is hidden by the t-statistic.*
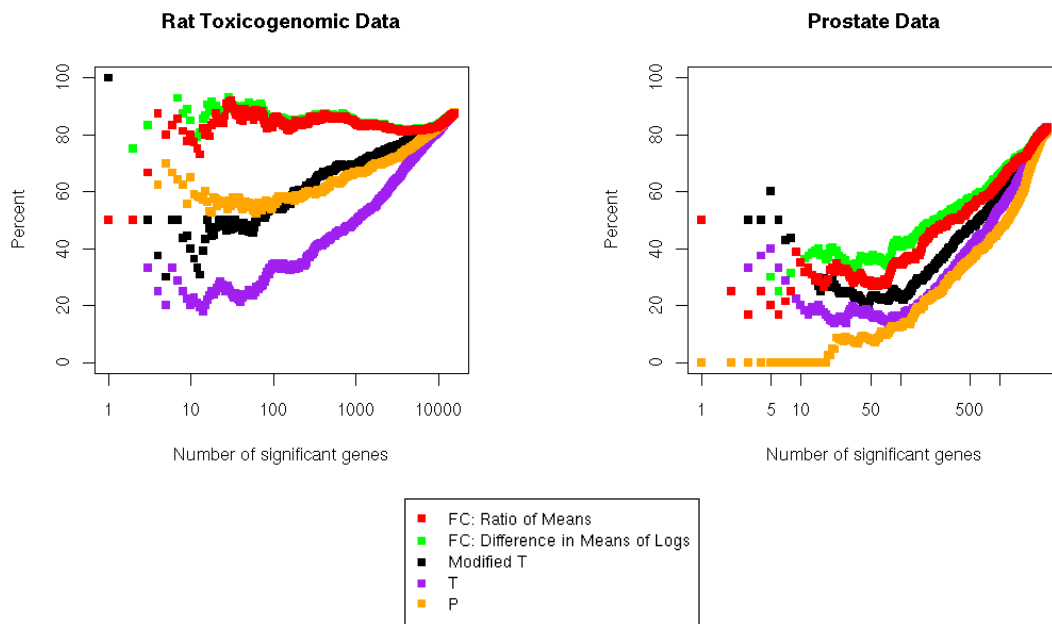
Figure 2: *Both figures show the accuracy of the various statistics on data generated using Simulation 2 with six control and six treatment replicates, when the true gene ordering is defined by* $|\mu_c - \mu_t|$. *The figure on the left is based on the rat toxicogenomic data of Guo et al. (2006), and the figure on the right is based on the prostate data of Singh et al. (2002). The two versions of fold-change are the most accurate.*
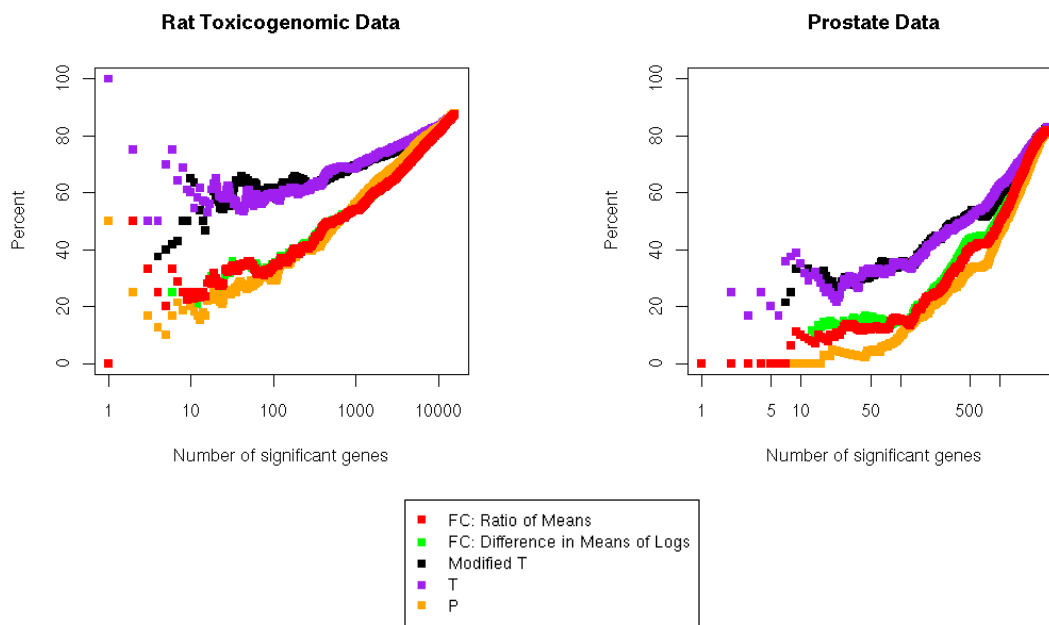
Figure 3: *Both figures show the accuracy of the various statistics on data generated using Simulation 2 with six control and six treatment replicates, when the true gene ordering is defined by $|\frac{\mu_c - \mu_t}{\sigma}|$. The figure on the left is based on the rat toxicogenomic data of Guo et al. (2006), whereas the figure on the right is based on the prostate data of Singh et al. (2002). The t-statistics are the most accurate.*
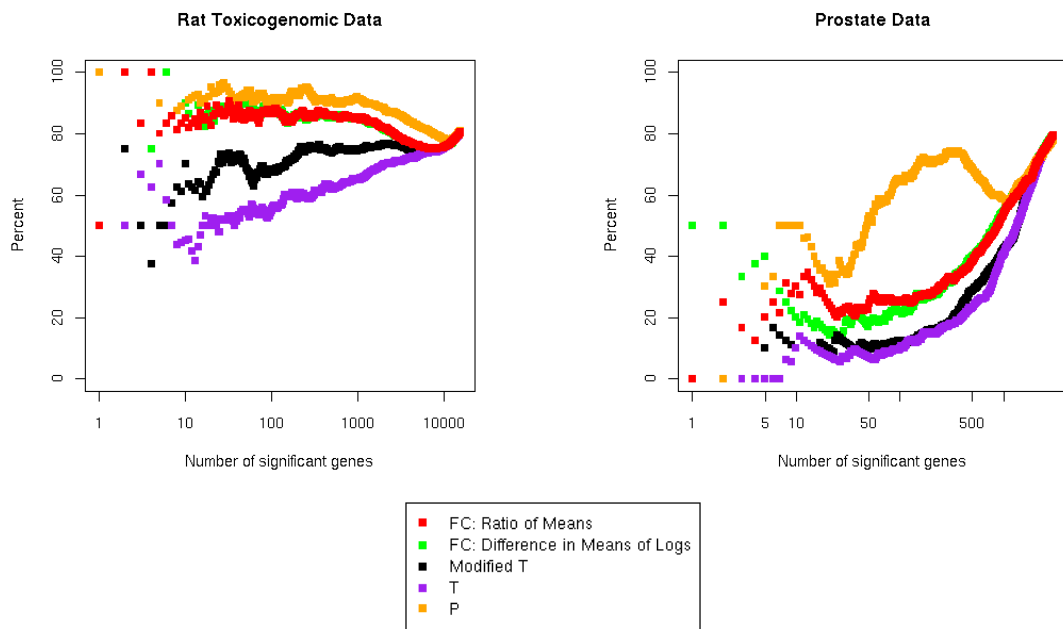
Figure 4: *Both figures show the concordance between two gene lists obtained using two separate sets of six control and six treatment replicates. The figure on the left is based on the rat toxicogenomic data of Guo et al. (2006), whereas the figure on the right is based on the prostate data of Singh et al. (2002). In both cases, P is more reproducible than the other statistics.*