8-16-2011

The Importance of Statistical Theory in Outlier Detection

Sarah C. Emerson
Oregon State University, emersosa@stat.oregonstate.edu

Scott S. Emerson
University of Washington, semerson@u.washington.edu

The importance of statistical theory in outlier detection

Sarah C. Emerson

Department of Statistics, Oregon State University, Corvallis, OR 97331, USA

emersosa@stat.oregonstate.edu

Scott S. Emerson

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

Abstract

We explore the performance of the outlier-sum statistic (Tibshirani and Hastie, *Biostatistics* 2007 8:2–8) , a proposed method for identifying genes for which only a subset of a group of samples or patients exhibits differential expression levels. Our discussion focusses on this method as an example of how inattention to standard statistical theory can lead to approaches that exhibit some serious drawbacks. In contrast to the results presented by those authors, when comparing this method to several variations of the t-test, we find that the proposed method offers little benefit even in the most idealized scenarios, and suffers from a number of limitations including difficulty of calibration, high false positive rates owing to its asymmetric treatment of groups, poor power or discriminatory ability under many alternatives, and poorly defined application to one-sample settings. Further issues in the Tibshirani and Hastie paper concern the presentation and accuracy of their simulation results; we were unable to reproduce their findings, and we discuss several undesirable and implausible aspects of their results.

Keywords: Efficiency; Gene expression analysis; Microarray; t-test

1 Introduction

In a recent paper, Tibshirani & Hastie [3] report on the operating characteristics of an "outlier-sum" statistic designed to detect differential gene expression that might occur in only a subset of one group of patients. In their paper, they compare their outlier statistic to the t-test and to the COPA method previously investigated by Tomlins et al. [4]. They present results to suggest that the outlier-sum statistic is superior to those competitors in settings in which a relatively low proportion of patients in the target group exhibit differential gene expression.

There are several aspects of their results that we found surprising, and we therefore undertook more extensive evaluation of their statistic. In doing so, we found that we were unable to reproduce some of the results presented in their paper. This work presents what we believe to be a more accurate comparison of various forms of the t-test to several variations on the outlier-sum statistic. While Tibshirani and Hastie concentrate primarily on the false discovery rate when attention is focused on some number of genes having the largest statistics, we also consider the statistical power of the various statistics to detect whether some specified gene is over-expressed in disease.

We first note that the scientific motivation for this work is well-founded. It is quite plausible that differences in gene expression might be evident in only a subset of diseased subjects owing to heterogeneity in the causes of a phenotypic disease, temporal variation in the pathophysiology of disease due to a single causative agent, variability in patient response to the cause of the disease, variability in patient response to

treatments, variability in environmental exposures related to the expression of disease signs and symptoms, and variability in patients' co-morbid conditions, among others.

In fact, the possibility that only a subset of a group might exhibit change is not a new idea or phenomenon; quite regularly in pharmaceutical trials we expect that only a subset of patients will be responsive to a given treatment. Indeed, this is one reason that t-tests are such valuable tools: the multitude of mechanisms that might lead to effect or lack of effect of a treatment argues that a location shift hypothesis is relatively implausible. The mean, however, is sensitive to a wide variety of effects that a treatment or risk factor might have on the distribution of an outcome variable.

It is true, however, that when distributions are heavily skewed, the effect of outliers on the estimation of the standard errors can be so dramatic as to greatly decrease power to detect differences relative to alternative testing strategies. But even then, some of the conventional wisdom about the effect of outliers on the precision of tests needs to be carefully described. For instance, it is often stated that the Wilcoxon rank sum test will outperform the t-test in the presence of heavy tailed distributions [see 2, Chap. 2.4]. However, that better behavior of the Wilcoxon rank sum test relative to the t-test in the presence of outliers is most pronounced when the alternative conforms to a location shift model under some monotonic transformation. In contrast, when the treatment's effect is to modify the propensity to outliers, the t-test will often outperform the Wilcoxon rank sum test.

Tibshirani and Hastie propose the outlier-sum statistic specifically to detect such phenomena, and consider the performance of this procedure in a limited number of examples. Here we discuss some of the limitations of their proposed method, and investigate its behavior in a wider variety of situations. We find that the performance of the proposed procedure is both highly variable and generally inferior to that of a t-test. The outlier-sum statistic also suffers from a very asymmetrical treatment of the two groups being compared, which results in false positives if both groups have a small subset of "outliers" which are independent of group label. Furthermore, if the statistic is to be actually used in hypothesis testing, calibration of the statistic becomes a problem: the null distribution of the statistic differs greatly depending upon the underlying data distribution. This problem is circumvented in the paper by using the empirical distribution of the statistic in a given data set, but this solution relies on some strong assumptions. Even when the statistic would be used only to rank the genes rather than to perform any inference, the statistic displays very suboptimal behavior in many cases.

We begin by presenting the definition of the outlier-sum statistic, and discussing some possible variations on the statistic in Section 2. The definition of the statistic in the original paper is somewhat ambiguous, so we explore several possible interpretations and some other modifications that might influence the behavior of the procedure. In Section 3, we briefly discuss the use of a t-test as the natural comparator to the outlier-sum statistic. The performance issues will be explored in greater detail in Section 4. We address the presentation and reproducibility of the results in the original paper, and present our quite different simulation results (independently confirmed). We conclude in Section 5 with a discussion of the implications of the results and issues presented here.

2 Outlier-sum Statistic and Variations

Notationally, let $X_{ijk} \overset{\text{iid}}{\sim} G_{ik}$ represent the measurements of the expression of gene i ($i = 1, \ldots, m$) for patient j ($j = 1, \ldots, n_k$) in group k ($k = 1$ for the normal or reference group and $k = 2$ for the disease or treatment group). We assume patients and disease groups are totally independent. We further define for each gene within each disease group moments $E[X_{ijk}] = \mu_{ik}$ and $Var(X_{ijk}) = \sigma^2_{ik}$ and unbiased estimators $\hat{\mu}_{ik} = \bar{X}_{i\cdot k} = \sum_{j=1}^{n_k} X_{ijk}/n_k$ and $s^2_{ik} = \sum_{j=1}^{n_k} \left(X_{ijk} - \bar{X}_{i\cdot k}\right)^2/(n_k - 1)$.

2.1 Standardization of Observations

Owing to the large number of genes investigated and the occasional interest in using the empirical distribution observed across genes to calibrate statistical inference, it is often the case that data analysts standardize the raw data. Tibshirani and Hastie used a standardization based on the median observation and the median

absolute deviation (mad) from the median. For instance, let m_{i1} and $m_{i\cdot}$ be the median gene expression of the ith gene for, respectively, the healthy patients and the combined sample of healthy and diseased patients. Further define d_{i1} to be the median value of the absolute deviations $|X_{ij1} - m_{i1}|$ for $j = 1, \ldots, n_1$, and define $d_{i\cdot}$ to be the median value of the absolute deviations $|X_{ijk} - m_{i\cdot}|$ for $k = 0, 1$ and $j = 1, \ldots, n_k$. Using the mnemonics of 'a' when the median or mad is based on all samples and 'h' when the median or mad is based solely on samples from the healthy population, we can then define standardized observations

$$\widetilde{X}_{ijk}^{(hh)} = \frac{X_{ijk} - m_{i1}}{d_{i1}} \qquad \widetilde{X}_{ijk}^{(ha)} = \frac{X_{ijk} - m_{i1}}{d_{i\cdot}}$$
$$\widetilde{X}_{ijk}^{(ah)} = \frac{X_{ijk} - m_{i\cdot}}{d_{i1}} \qquad \widetilde{X}_{ijk}^{(aa)} = \frac{X_{ijk} - m_{i\cdot}}{d_{i\cdot}}.$$

While the standardizations represented by $\widetilde{X}_{ijk}^{(hh)}$ or $\widetilde{X}_{ijk}^{(aa)}$ might seem the most intuitive, we consider all four in our later investigations, both because Tibshirani and Hastie are somewhat ambiguous in their choice (in the displayed equation defining the standardized values those authors seem to be using $\widetilde{X}_{ijk}^{(ah)}$, but later when describing their simulations, the authors state, "all measurements for a gene are standardized by the overall median and median absolute deviation for that gene") and because we find the relative performance of the alternative definitions depends very much on the prevalence of "outliers".

2.2 Definition of the Outlier-sum Statistic

Tibshirani and Hastie propose inference based on an outlier-sum statistic that involves identifying those standardized observations for gene i that meet a criterion as an outlier, with the value of the statistic given by the sum of any such outliers observed in the diseased group. They define as an outlier any standardized observation that exceeds the 75th percentile of the distribution of standardized observations by more than one interquartile range. Clearly, the definition and behavior of the statistic depends very much on the method by which individual observations are standardized, as well as which sample(s) are used to define the quartiles used in the threshold for outliers. Following Tibshirani and Hastie's general approach, we thus define $\ell_{i1}^{(hh)}$ and $\ell_{i\cdot}^{(hh)}$ to be the 25th percentiles of the standardized gene expression measurements $\widetilde{X}_{ijk}^{(hh)}$ for, respectively, the healthy patients and the combined sample of healthy and diseased patients for gene i. Similarly, we define $u_{i1}^{(hh)}$ and $u_{i\cdot}^{(hh)}$ as the 75th percentiles, and from these quantiles we compute the interquartile ranges $IQR_{i1}^{(hh)} = u_{i1}^{(hh)} - \ell_{i1}^{(hh)}$ and $IQR_{i\cdot}^{(hh)} = u_{i\cdot}^{(hh)} - \ell_{i\cdot}^{(hh)}$. Analogous statistics can be defined in the obvious manner based on each of the remaining three standardizations. Using these definitions, two possible variants of an outlier-sum statistic are given by

$$W_i^{(aaa)} = \sum_{j=1}^{n_2} \widetilde{X}_{ij2}^{(aa)} \mathbb{I}\left\{\widetilde{X}_{ij2}^{(aa)} > u_{i\cdot}^{(aa)} + IQR_{i\cdot}^{(aa)}\right\}$$
$$W_i^{(aha)} = \sum_{j=1}^{n_2} \widetilde{X}_{ij2}^{(ah)} \mathbb{I}\left\{\widetilde{X}_{ij2}^{(ah)} > u_{i\cdot}^{(ah)} + IQR_{i\cdot}^{(ah)}\right\},$$

where $\mathbb{I}$ is an indicator function taking on the value 1 if its argument is true and 0 otherwise. A total of eight possible statistics may be defined: $W^{(aaa)}, W^{(aah)}, W^{(aha)}, W^{(ahh)}, W^{(haa)}, W^{(hah)}, W^{(hha)}$, and $W^{(hhh)}$ where we use superscripts to denote first the median used as the center of the standardization, second the mad used to scale the standardization, and third the upper quartile and interquartile range used to define the threshold for outliers (h when using u_{i1} and IQR_{i1} and a when using $u_{i\cdot}$ and $IQR_{i\cdot}$). It would seem from the descriptions in the paper that Tibshirani and Hastie focused on $W^{(aaa)}$ or possibly $W^{(aha)}$. However, we have explored the performance of the statistic in each of the eight different forms, and have provided results for a representative subset of four. We note that the statistic $W^{(hhh)}$ is similar to, though not exactly the same as, the outlier robust t statistic described by Wu [5], which scales the standardizations by a pooled mad $d_{ip}^* = \text{median}\{|X_{ijk} - m_{ik}| \ : \ k = 1, 2; \ j = 1, \ldots, n_k\}$: the median value across groups of the within group absolute deviations.

3 Inference Based on Means: t-statistics

The standard analysis method that would most typically used to compare distributions would be based on means. Tibshirani and Hastie in their simulations compare the performance of the outlier-sum statistic to that of the t-statistic that presumes equal variances (we note that their displayed equation 2.1 defining the t-statistic is incorrect; the denominator should be $s_i\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$, where s_i is the pooled within-group standard deviation of gene i, and n_1 and n_2 are the sample sizes in groups 1 and 2 respectively). We extend this comparison to two other t-tests, as described below.

A measure of differential expression of gene i based on the mean expression is $\delta_i = \mu_{i2} - \mu_{i1}$, as might be estimated by $\hat{\delta}_i = \hat{\mu}_{i2} - \hat{\mu}_{i1}$ with approximate sampling distribution

$$\hat{\delta}_i \dot{\sim} \mathcal{N}\left(\delta_i, V_i = \frac{\sigma_{i1}^2}{n_1} + \frac{\sigma_{i2}^2}{n_2}\right).$$

This sampling distribution is asymptotically correct and distribution-free within the class of all distributions for gene expression with finite variance. Ordering of the genes with respect to over-expression might therefore be based on a statistic $t_i = \hat{\delta}_i/\sqrt{\hat{V}_i}$. Intuitive choices for $\hat{V}_i$ include:

1. (t-test that presumes equal variances) $T_i^{(e)} = \hat{\delta}_i/\sqrt{\hat{V}_i^{(e)}}$ where $\hat{V}_i^{(e)} = s_{(p)i}^2\,(1/n_1 + 1/n_2)$ and $s_{(p)i}^2 = \left((n_1-1)s_{i1}^2 + (n_2-1)s_{i2}^2\right)/(n_1+n_2-2)$. Ordering of the genes would typically be based on the p-value derived from the t distribution having $d = n_1 + n_2 - 2$ degrees of freedom.

2. (t-test that allows unequal variances) $T_i^{(u)} = \hat{\delta}_i/\sqrt{\hat{V}_i^{(u)}}$ where $\hat{V}_i^{(u)} = \left(s_{i1}^2/n_1 + s_{i2}^2/n_2\right)$. Ordering of the genes would typically be based on the p-value derived from the t distribution having degrees of freedom d given by
$$d = \frac{(s_{i1}^2/n_1 + s_{i2}^2/n_2)^2}{s_{i1}^4/(n_1^2(n_1-1)) + s_{i2}^4/(n_2^2(n_2-1))}$$
(the Satterthwaite approximation).

3. (t-test that uses healthy distribution as reference) $T_i^{(h)} = \hat{\delta}_i/\sqrt{\hat{V}_i^{(h)}}$ where $\hat{V}_i^{(h)} = s_{i1}^2\,(1/n_1 + 1/n_2)$. Ordering of the genes would typically be based on the p-value derived from the t distribution having $d = n_1 - 1$ degrees of freedom.

The p-values based on $T_i^{(e)}$ and $T_i^{(h)}$ are exact when G_{i1} and G_{i2} are normal distributions having $\sigma_{i1}^2 = \sigma_{i2}^2$. In that setting, a hypothesis test based on $T_i^{(e)}$ is the uniformly most powerful test of $H_{0i} : \mu_{i1} = \mu_{i2}$ versus a one-sided alternative, and is the uniformly most powerful unbiased test of H_{0i} versus a two-sided alternative. Under the more general conditions requiring only $\sigma_{ik}^2 < \infty$, as tests of the weak null hypothesis $H_{0i}^{(W)} : \mu_{i1} = \mu_{i2}$, the test based on $T_i^{(u)}$ is asymptotically of the correct size and consistent; the test based on $T_i^{(e)}$ is asymptotically of the correct size and consistent if $n_1 = n_2$; and the test based on $T_i^{(h)}$ is not asymptotically of the correct size. All three tests are asymptotically of the correct size as tests of the strong null hypothesis $H_{0i}^{(S)} : G_{i1} = G_{i2}$, but none are consistent. The test of the strong null based on $T_i^{(h)}$ can be more powerful than the other versions when $\sigma_2^2 > \sigma_1^2$, as long as n_1 is large enough to avoid loss of power from the lower degrees of freedom.

While scientific questions might at times be best addressed by a consistent test of the weak null hypothesis, the question of differential gene expression does not necessarily demand inference about any particular functional of the distribution. Hence, we focus primarily on inference about the strong null. *A priori*, the last version, $T^{(h)}$, would seem best among these t-tests at detecting alternatives where only a subset of the disease group has elevated expression levels. This is because the standard deviation estimate, being based only on the healthy group, will be smaller than that of either of the other two versions when a subset of the disease group displays elevated levels. Therefore, the resulting t statistic will be larger. This improvement

comes at the cost of a very slight loss of power when the entire disease group has elevated expression levels, but as we show in Section 4.6, this difference is negligible in our simulations. We will compare the various versions of the outlier-sum statistic to all three of these t statistics.

4 Theoretical issues with the outlier-sum approach

There are several issues with the outlier-sum statistic that warrant discussion. We first examine the statistical theory relevant to the standardization of observations and definition of outliers and the scientific rationale of one sample versus two sample testing with the outlier-sum statistic. Then, using simulations similar in spirit to, but more extensive than, those used by Tibshirani and Hastie in the two sample setting, we explore the ability to calibrate the statistics for use in inferential hypothesis testing, the distribution of p-values based on the outlier-sum statistic under alternatives, the relative power of the statistic to detect differential gene expression, and the ability of the statistic to correctly identify differentially expressed genes.

4.1 Standardization of data

In devising the outlier-sum statistic, the authors motivate their standardization as a means of putting all genes on the same scale in order to facilitate comparisons across genes. The most common form of standardization for these purposes is the z-score, in which each measurement is standardized using the mean and standard deviation of some reference distribution. For instance, we might define $Z_{ijk}^{(a)} = (X_{ijk} - \hat{\mu}_{i\cdot})/s_{i\cdot}$ to be the z-score for each observation for gene i, standardized by the mean and standard deviation of all measurements when the disease groups are collapsed. More typically, we might scale the standardization using a pooled estimate of the variance $s_{(p)i}^2 = \frac{1}{n_1+n_2+2}\sum_{k=1}^{2}\sum_{j=1}^{n_k}(X_{ijk} - \hat{\mu}_{ik})^2$. For the purposes of identification of outliers, we might instead standardize each measurement to the mean and standard deviation of the healthy group's distribution as $Z_{ijk}^{(h)} = (X_{ijk} - \hat{\mu}_{i1})/s_{i1}$.

When looking for differences in distributions, this latter standardization might have an advantage due to Chebyshev's inequality. That is, if the distributions of gene expression measurements for the diseased and healthy groups are identical, then by Chebyshev's we would expect that

$$P\left(|Z_{ijk}^{(h)}| > c\right) \leq \frac{1}{c^2}.$$

Defining "outliers" based on such a Z-standardization is thus well-founded in statistical theory. Seeing a substantially larger proportion of diseased subjects' standardized measurements greater than some threshold is suggestive of a difference in distributions. A distribution-free hypothesis test could be based on the binomial distribution, though such a test is likely to be conservative due to the fact that Chebyshev's inequality does not produce a very tight bound in most distributions.

We note that there is no clear analogy to Chebyshev's inequality when basing data standardization on the median and median absolute deviation as is used in the outlier-sum statistic. When allowing the standardized values to range over the extended reals, the proportion of a distribution whose standardized values can exceed every value of $c > 0$ is bounded above only by 0.5. That is, for any choice of $\epsilon > 0$, there exists a distribution such that for every choice of $c > 0$ the probability that a standardized measurement exceeds c is $0.5 - \epsilon$. Thus a bound similar in spirit to Chebyshev's inequality would only be that

$$P\left(|\widetilde{X}_{ijk}^{(hh)}| > c\right) < \frac{1}{2},$$

which bound is independent of c.

We do note, however, that as a monotonic transformation of the data, the choice of standardization does not affect the ordering of cases for any given gene, nor will it affect which cases would be categorized as outliers using the criterion defined for the outlier-sum statistic. However, the choice of standardization does have a major effect on the magnitude of the outliers that are eventually summed, and thus the method of standardization will affect how the distribution of the outlier-sum statistic varies with the distribution of observed gene expression measurements. We elaborate further on this below.

4.2 Definition of outliers

The choice of the outlier criterion definition has a much greater effect on the behavior of the outlier-sum statistic. As noted above, when using Chebyshev's inequality as a rationale in the definition of outliers, we can define an outlier for each population based on some absolute threshold of a z-score computed using some reference distribution. Further, the relationships between possible choices for scale in the standardization are well-understood because we can use the relationship $Var(X) = Var(E[X|Y]) + E[Var(X|Y)]$ to relate the total variance in the combined sample to the within group variance for the diseased and healthy samples. For instance, in the setting in which the effect of disease is to generate higher levels of gene expression in some or all individuals (i.e., a location shift effect in a subgroup), we would expect $s_{i\cdot}^2 > s_{i1}^2$ and thus the standardized values $Z_{ij2}^{(a)}$ to tend to be smaller than $Z_{ij2}^{(h)}$.

The properties of the outlier-sum statistic's definition of outliers are more difficult to describe in general. The effect of scaling by the median absolute deviation for the combined sample versus the median absolute deviation for the healthy population alone is not easily quantified. In the setting of an effect of disease that generates an additively higher level of gene expression in some or all individuals, the median absolute deviation for the combined sample can be larger or smaller than the median absolute deviation for the healthy population, depending on the shape of the distribution in the healthy population. For instance, consider a mixture of normals model for the healthy population in which proportion a is distributed $\mathcal{N}(0, \tau^2)$ and proportion $1 - a$ is distributed $\mathcal{N}(1, \tau^2)$. Further suppose that in the diseased population, proportion $1 - p$ has the same distribution as the healthy group, and proportion p has a mean 1 unit higher in a location-shift model. (As discussed below, this normal mixture model might seem an appropriate approximation for the gene expression patterns displayed in Figure 4 of Tibshirani and Hastie.) For the values of $a = 0.4$, $\tau = 0.1$, and $p = 0.5$, the median absolute deviation in the combined population is 7.8% less than the median absolute deviation in the healthy population. In contrast, for the values of $a = 0.2$, $\tau = 0.1$, and $p = 0.75$, the median absolute deviation in the combined population is 74% higher than the median absolute deviation in the healthy population. As p increases toward 1, the median absolute deviation in the combined population can easily be 3 to 6 times higher than that in the healthy population.

Furthermore, as noted above, the probability that an individual observation would be an arbitrarily large number of median absolute deviations from a median is bounded only by 0.5. Hence, even when using the healthy population to standardize the data, defining outliers based on such a standardization could merely be identifying those null distributions having heavier tails, irrespective of whether those larger observations truly represent differential gene expression. For instance, with a standard normal distribution, 2.15% of the distribution meet the outlier-sum statistic's criterion for being high outliers. In the normal mixture model described above, the choices of $a = 0.9$ and $\tau = 0.1$ result in 10.5% of the distribution meeting the criterion as high outliers. A normal mixture model having 24.5%, 51%, and 24.5% of the observations as $\mathcal{N}(-1, \tau^2)$, $\mathcal{N}(0, \tau^2)$, and $\mathcal{N}(1, \tau^2)$, respectively, with $\tau = 0.1$ results in 24.5% of the observations characterized as high outliers by the outlier sum's criterion. This would be true even when the distribution of gene expression is identical in the healthy and diseased groups. Differing shapes of distributions of gene expression among the healthy population can have even more drastic impact on the average value of the standardized gene expression, as the median absolute deviation becomes very small. Code exploring the patterns in median absolute deviation and propensity to outliers for a variety of mixture distributions is available in the supplementary materials.

Hence, this statistic will be unable to distinguish between a gene that has a small percentage of outliers in both disease and normal subjects and a gene that only has outliers in the disease subjects. Thus, if the outliers represent a subpopulation that is independent of disease status, the statistic will still be likely to call a gene with such a subpopulation significant. The phenomenon we describe here can be seen to some extent in the real data example from Section 4 of the original paper, though interpretation is somewhat hampered by discrepancies between the sample size of 14 reported for the "diseased" group in the text and the number of observations displayed in Figure 4 (we believe the error stems from double plotting of vertically jittered "outliers"). In that figure, several of the genes presented seem to have a markedly bimodal distribution in the "healthy" group, with less impressive differences between the "healthy" and "diseased" classes in their ranges of observed values.

4.3 Calibration

One goal of the types of analyses considered here might be to quantify the evidence in support of a hypothesis that some specified gene has a greater tendency toward over-expression in diseased patients. Characterization of the relative advantage of one statistic over another is then typically summarized by the statistical power under various alternatives of a level α hypothesis test.

In order for a statistic to be used for hypothesis testing, it would need to return an appropriately calibrated p-value that is uniformly distributed under the null hypothesis. As noted above, when making inference on means, robust statistical theory identifies the approximate null distribution of t-statistics in moderate sample sizes. On the other hand, there is no statistical theory that defines a null distribution for the outlier-sum statistics that is distribution-free.

At first glance, it can be seen that the outlier-sum statistic is related to a mean. Hence a simple central limit theorem would suggest that the outlier-sum statistic would approximate a normal distribution in large samples, providing the quartiles used to define the threshold for outliers were known and the median absolute deviations used to standardize the measurements were nonzero. Chen et al. [1] rigorously consider the use of estimated quartiles in order to derive a limiting distribution for the outlier-sum statistic for a known distribution of gene expression in a healthy population. However, when the healthy population's distribution is unknown and/or varies across genes their results are not of use.

In fact, calibration of the outlier-sum statistic is extremely difficult. Very different null distributions for the statistic will result from different underlying data distributions. For instance, if all of the data is independent and identically distributed according to a standard normal distribution, the resulting statistics will have a quite different distribution from the case when the data are generated according to, for example, a chi-squared distribution. There is no asymptotic theory to guarantee that with a large enough sample size these differences will cease to matter. Figure 1 illustrates the difficulty in calibrating this statistic: quantile-quantile plots are displayed comparing the null distribution of the outlier-sum statistic ($W^{(aaa)}$) under different data-generating mechanisms to the null distribution of the outlier-sum statistic when the data is standard normal. Similar discrepancies are observed for all of the versions of the outlier-sum statistic that we investigated. The difference in the resulting null distributions is very pronounced, indicating the strong dependence of the distribution of the statistic on the underlying distribution of the data. Furthermore, the differences between the null distributions for the outlier-sum statistic actually get worse with increasing sample sizes. On the other hand, the analogous plots for the t statistic in Figure 1 show very similar null distributions are obtained for all of the data-generating mechanisms, with increasing similarity among the null distributions as the sample sizes are increased.

Furthermore, the use of a permutation approach to calibrate the statistic is complicated when the elevated subset consists of only a few subjects, as the chance that a randomly chosen permutation will place the top few outliers in the disease-labeled group may not be low enough to allow calibration at the desired level. This of course will depend on the relative size of the entire diseased group, and on the number of outliers according to the specified criterion. The authors use an empirical calibration based on other rows of the data-set; this approach will be problematic if the data do not all come from the same distribution, or if there are a reasonable number of non-null cases. Owing to possible variation in gene behavior and correlation among the genes, the empirical distribution of outlier-sum statistics that is computed across genes might very well be inaccurately calibrated.

Instead of attempting to calibrate the outlier-sum statistic, it could alternatively be used only to order the various rows; in this case, the calibration to produce p-values will not matter, but again differing distributions between rows will render this ordering less optimal because of the difference in null distributions discussed above. As we will demonstrate in our simulations below, the outlier-sum statistic frequently produces orderings that are inferior to those of a t-test, in the sense that genes with a differentially expressed subset are less likely to be among the top genes according to the outlier-sum ordering. The operating characteristics of the various statistics can be compared with respect to the statistical power available to detect whether some specified gene is over-expressed in disease or with respect to the false discovery rate when attention is focused on some number of genes having the largest statistics.

4.4 One sample usage

The authors discuss the use of the outlier-sum statistic in a one-sample problem in the simulations in Section 3 of the paper. It is unclear what null hypothesis the statistic is intended to test in this situation. The one-sample t-test is intended to test hypotheses regarding the mean of the underlying distribution, but that would make little sense unless it were known that all genes have the exact same known mean expression levels in the healthy population. Hence, comparisons of the outlier-sum to the one sample t-test seems inappropriate.

One could imagine that the one sample outlier-sum statistic could be used to identify genes where there is a subgroup of outliers as compared to the majority of the disease samples. The authors do not indicate specifically how the statistic would be defined in the one-sample setting (i.e., with no normal subjects). We presume that it would be similarly constructed as the sum of the outliers, with standardization of measurements and definition of outliers calculated from only the diseased subjects. Such an approach would seem to imply that the authors are certain that disease is the major component of variation in expression levels across all genes. Only then should we feel comfortable in regarding that bimodal distributions were attributable to a disease's effect on gene expression. However, this does not seem supportable by our knowledge of many genes that are expressed differentially according to time since meals, time since exercise, diurnal rhythms, monthly cycles, and seasonal exposures, among many other mechanisms.

We further note that the simulation results reported in Table 2 of the original paper appear incorrect. If all of the disease patients have elevated expression levels for a particular gene, then there would be no more outliers than in a null case since the method for generating the alternative distribution was a location shift. Hence, after standardization using only the diseased subjects (and in the one sample setting we would have no other choice), the standardized values should be distributed exactly the same as the non-affected genes. One would then expect a proper p-value to have mean 0.5, median 0.5, and standard deviation 0.289. Were there to be a propensity for many zero-valued outlier-sum statistics to be assigned a p value of 1, that should drive the mean and median higher than 0.5. The lack of variation of median and mean p-values across the cases examined by Tibshirani and Hastie will be revisited in our two sample simulations.

4.5 Sampling distribution of p-values under the alternative

In their paper exploring the outlier sum statistic, Tibshirani and Hastie paper report the mean, median, and standard deviation of p-values in the setting of variably over-expressed genes. This is an unconventional way to compare the efficiency of testing procedures, and does not directly provide information about the quantity that is generally of interest; namely, the power of the procedures. Furthermore, the results presented in that paper did not to us appear correct, even allowing for the unusually small numbers of simulations (50) performed by those authors. The mean and median p-values were reported to vary very little as the proportion of diseased subjects exhibiting over expression of genes varied. Taken at face value, the lower values of standard deviation of p-values in the presence of relatively high mean p-values would suggest that identification of over expressed genes would be extremely rare with the outlier sum statistic. Inspecting the provided quantities, we notice that in Table 1 of the original paper, for the $k = 15$ and $k = 8$ cases, the standard deviations are 0.019 and 0.030, respectively, with means around 0.10. Thus especially for the $k = 15$ simulation it seems very unlikely that many, if any, of the resulting p-values would have been less than the standard 0.05 cut-off (in fact, Chebyshev's inequality shows that at most 11.5% of these p-values could fall below 0.05, if these numbers were correct). Similarly low standard deviations appear in Table 2 for the $k = 4$ and $k = 2$ cases, and so again in these cases it seems likely that the procedure has very low power when a type I error of $\alpha = 0.05$ is desired.

In our own investigations, we have been unable to reproduce the findings of Tibshirani and Hastie. We attempted to perform the exact simulations described in the paper, and we present our results below. Both authors of this manuscript have independently confirmed all of the simulation results presented herein, and commented code used to perform the simulations is available in the supplementary materials.

Briefly, in our attempts to reproduce the results reported in the previous paper, for each simulated experiment we generated a 30×1000 matrix of independent observations drawn from the standard normal

distribution. The first 15 rows of the matrix represent diseased subjects, and the last 15 will be the healthy subjects. Under the location shift model used for the affected subset of diseased subjects, 2 is added to the first r observations in the first column of the matrix. Observations are then standardized by centering and scaling using the eight different versions of the outlier sum statistic outlined in Table 1 of this paper. Here p-values were calculated according to

$$p_1 = \sum_{i=2}^{1000} \mathbb{I}\{W_i > W_1\} + \frac{1}{2} \sum_{i=2}^{1000} \mathbb{I}\{W_i = W_1\},$$

so that the expectation of the p-value in the null case is 0.5 as it should be. We also computed three versions of t-statistics (presuming equal variances, allowing for unequal variances, or using the variability of the healthy cases), and calculated p-values using both the respective t distributions and the empirical distribution across genes.

The results of these simulations are displayed in Table 2, along with the values from Table 2 of the original paper. We show results for 4 versions of the outlier sum statistic, with results for all 8 versions available in supplementary materials. Our simulations produce universally higher summary statistics than those of Tibshirani and Hastie no matter which variant of the outlier sum statistic is used, and in almost all cases our values are larger by a factor of at least two. This discrepancy is much greater than can be explained by random variation in the simulations. We note that additional investigations using alternative definitions of empirical p-values (i.e., methods that handle tied observations differently) also did not agree with the published results.

4.6 Statistical power to detect an over expressed gene

As a more standard comparison of the test procedures, we also estimated the power of a level $\alpha = 0.05$ test for each method, by calculating the proportion of p-values that were below the specified level. The power comparisons for normally distributed gene expression levels are shown in Table 3. The $W^{(aaa)}$ version of the outlier-sum statistic, which we presume to be the version used in the original paper, is the most powerful version of the outlier sum statistic when $r = 2$, attains its best power among the cases examined when $r = 4$, and has no appreciable power when $r = 15$ for the reasons we discussed in Section 4. On the other hand, the variants based on thresholding outliers using the distribution among healthy cases ($W^{(hah)}$ and $W^{(hhh)}$) show increasing power with increasing r and are the more powerful versions of the outlier sum statistic when $r = 4$, 8, or 15.

A comparison of the outlier sum statistics to the t-tests finds that the t-test using variance based on the healthy cases performs comparably to the best of the outlier sum statistics when $r = 4$, has greater power than all other statistics evaluated when $r = 8$, and has comparable power to the uniformly most powerful test (the t-test that presumes equal variances) when $r = 15$.

We extended the power comparisons to data generated according to distributions other than the standard normal distribution, and found that the outlier-sum statistic performed much more poorly when the data comes from a distribution with heavier tails than the normal distribution. Results for data distributed according to the t distribution with 5 degrees of freedom are also shown in Table 3. For this situation, all versions of the outlier-sum statistic have lower power than the $T^{(h)}$ statistic except when $r = 2$, in which case the power of $T^{(h)}$ and $W^{(aaa)}$ are comparably low, and the variants based on thresholding outliers using the distribution among healthy cases ($W^{(hah)}$ and $W^{(hhh)}$ appear most powerful among the outlier sum statistics.

4.7 False discovery rates

In their paper introducing the outlier sum statistic, the authors did not systematically investigate the false discovery rate associated with the use of their statistic compared to that when using a t statistic. Instead they relied on the use of their statistic in a single data set and reported the estimated false discovery rate as the threshold for declaring over expression was varied. In our investigations, we explored the use of these

various statistics to rank the genes resulting from a single experiment, to determine which statistic does the best job of identifying the genes for which a subset of patients have elevated expression levels. This comparison is slightly different from the power comparisons, since for this purpose the exact magnitude of the p-values are irrelevant, and only the relative rankings matter.

Table 3 presents the average false discovery rate among the genes having as low or lower p-values than the case with the 25th lowest p-value rank (that is, the 25 most significant genes). For the results presented in this table, we simulated a setting in which a pathway involving 20 genes might be over-expressed, hence the lowest possible FDR is $\frac{5}{25} = 20\%$. The relative performance of the various statistics is very similar to that seen for the statistical power. With normally distributed data, the $W^{(aaa)}$ version of the outlier sum statistic had the lowest FDR when $r = 2$, but had higher FDR than some other versions of the outlier sum statistic for $r \geq 4$. The FDR of $T^{(h)}$ was comparable to the best of the outlier sum statistics for $r = 4$, and it outperformed all of the outlier sum statistics for $r = 8$ or 15. Furthermore, the FDR for $T^{(h)}$ was nearly as low as that for the uniformly most powerful test. Similarly, when the distribution of gene expression is more heavy tailed, the relative advantages of the t test become more pronounced. These patterns can be seen in Figure 2, which displays the average number of over-expressed genes identified in the 25 most significant for the t and outlier sum statistics as a function of r and the underlying distribution of gene expression in the healthy cases. Immediately apparent from those plots is the changing relative performance among the outlier sum statistics as r is increased: For $r = 2$, there appears a tendency for outlier sum thresholds based on all patients' data to outperform those based on only the healthy cases, while as r increases thresholding based on the healthy data alone appears better. Similarly, the relative advantage of the t-test based on the healthy cases' variance is greater when $r > 2$ for normally distributed gene expression and for all r when the gene expression distribution is more heavily tailed.

5 Discussion

Based on the comparisons presented here and those available as supplemental materials, it seems clear to us that the outlier-sum approach offers very little benefit over more traditional statistical methods under any scenario, and suffers great drawbacks under most. As we mentioned above, the null distribution for the outlier-sum statistic varies greatly depending upon the underlying data distribution. The efficiency (as measured by power to detect an alternative, or false discovery rates) is also very dependent upon the underlying data distribution. Distributions with higher kurtosis than the normal distribution will be more likely to have more extreme outliers even in the null case, and therefore the outlier-sum statistic becomes much less useful. The relative behavior of the various versions of the outlier sum statistic is quite dependent upon the proportion of diseased subjects exhibiting over expression of affected genes. We would not in general recommend that an investigator fish through a battery of statistics trying to optimize the detection of gene expression at different levels of penetrance. Instead, we find that the t-test based on the healthy cases' variance provides much better overall performance. We submit that this finding might well have been expected given the ad hoc nature of the outlier sum statistic and its lack of basis in any well-founded statistical theory.

Lastly, the explorations and remarks presented here, while focused on a particular setting and method, may also be taken more generally. We believe that the ad hoc development of new procedures or test statistics requires careful consideration of the problems that these procedures are appropriate for, and a careful investigation of the performance of the procedures in a variety of settings. Furthermore, that performance should be compared to the most viable of existing methods (an approach analogous to clinical trials of new therapies controlled by the best current standard of care). It is important to think carefully about the goal of the statistic, and how it will perform in many different situations, as the assumptions that dictate the suitability of a given method are often untestable. When a new procedure looks promising after such investigations, we also believe that it is important to identify breaking points for new methods: identifying problems or settings in which a statistic will perform poorly helps to understand when a particular approach is suitable, and whether an existing and better-understood method might instead suffice.

6 Supplementary Materials

Two R code files are provided as supplementary materials. The first file, `MADandOutlierPropensity.R`, explores the relationship between the median absolute deviation of the healthy group and the median absolute deviation of the combined healthy and disease groups for a variety of underlying distributions and differentially expressed subgroup sizes. The results in Section 4.2 were produced using this code, and several other examples are also presented. The second code file, `OutlierSumOperChars.R`, provides code to run the simulations comparing the eight versions of the outlier-sum statistic to the variations on the t-test. Results of many simulations, including those presented in Figure 2 and Tables 2 and 3, are contained in this file. The results were produced using a set random seed, as indicated in the file, for reproducibility.

Acknowledgments

This work was supported by the following grant: National Institutes of Health T32NS048005.

References

[1] Chen, L., Dung-Tsa Chen & Chan, W. (2010). The distribution-based p-value for the outlier sum in differential gene expression analysis. *Biometrika* 97 246–253.

[2] Lehmann, E. L. (2006). *Nonparametrics: statistical methods based on ranks*. New York, NY: Springer.

[3] Tibshirani, R. & Hastie, T. (2007). Outlier sums for differential gene expression analysis. *Biostatistics* 8 2–8.

[4] Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X.-W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. A. & Chinnaiyan, A. M. (2005). Recurrent fusion of $TMPRSS2$ and ETS transcription factor genes in prostate cancer. *Science* 310 644–648.

[5] Wu, B. (2007). Cancer outlier differential gene expression detection. *Biostatistics* 8 566–575.

Outlier-sum Variation	Median	MAD	Quantiles and IQR
$W^{(aaa)}$	All	All	All
$W^{(aah)}$	All	All	Healthy
$W^{(aha)}$	All	Healthy	All
$W^{(ahh)}$	All	Healthy	Healthy
$W^{(haa)}$	Healthy	All	All
$W^{(hah)}$	Healthy	All	Healthy
$W^{(hha)}$	Healthy	Healthy	All
$W^{(hhh)}$	Healthy	Healthy	Healthy

Table 1: Variations on the outlier-sum statistic: which observations from gene i were used to compute each of the standardizing quantities and outlier criterion.

	$r = 15$		$r = 8$		$r = 4$		$r = 2$	
	Mean (SD)	Mdn	Mean (SD)	Mdn	Mean (SD)	Mdn	Mean (SD)	Mdn
$T^{(e)}$	0.000 (0.001)	0.000	0.037 (0.065)	0.013	0.176 (0.180)	0.115	0.316 (0.247)	0.262
$T^{(h)}$	0.000 (0.002)	0.000	0.025 (0.058)	0.005	0.155 (0.188)	0.080	0.307 (0.260)	0.239
$W^{(aaa)}$	0.518 (0.240)	0.673	0.328 (0.284)	0.227	0.236 (0.261)	0.114	0.279 (0.266)	0.161
$W^{(aha)}$	0.495 (0.270)	0.673	0.306 (0.295)	0.152	0.228 (0.263)	0.099	0.280 (0.265)	0.166
$W^{(hah)}$	0.025 (0.085)	0.001	0.074 (0.144)	0.018	0.158 (0.207)	0.070	0.263 (0.248)	0.180
$W^{(hhh)}$	0.031 (0.090)	0.003	0.077 (0.143)	0.023	0.165 (0.205)	0.082	0.272 (0.245)	0.191
T & H	0.110 (0.019)	0.106	0.105 (0.030)	0.094	0.098 (0.130)	0.093	0.100 (0.131)	0.100

Table 2: Comparison of results from 10,000 simulated experiments in which expression is measured on 1,000 genes in 15 diseased and 15 healthy subjects. In each experiment, gene 1 is over-expressed in r of 15 diseased subjects. Gene expression in healthy subjects follows a standard normal distribution. Presented are the mean, median, and standard deviation of asymptotic (t-tests) or empirically computed (outlier-sum statistics) p-values for the single over-expressed gene (gene 1). We also provide the analogous statistics based on the 50 simulations as reported by Tibshirani and Hastie (T&H).

	$r = 15$		$r = 8$		$r = 4$		$r = 2$	
	Pwr	FDR25	Pwr	FDR25	Pwr	FDR25	Pwr	FDR25
Standard Normal Distribution								
$T^{(e)}$	1.000	0.205	0.788	0.573	0.291	0.871	0.129	0.946
$T^{(h)}$	1.000	0.211	0.873	0.475	0.398	0.804	0.173	0.923
$W^{(aaa)}$	0.054	0.974	0.226	0.871	0.327	0.823	0.235	0.892
$W^{(aha)}$	0.100	0.952	0.284	0.843	0.342	0.818	0.221	0.898
$W^{(hah)}$	0.876	0.389	0.647	0.617	0.384	0.799	0.194	0.911
$W^{(hhh)}$	0.834	0.442	0.606	0.650	0.339	0.831	0.168	0.927
t Distribution 5 df								
$T^{(e)}$	0.987	0.258	0.649	0.658	0.239	0.886	0.116	0.948
$T^{(h)}$	0.984	0.295	0.719	0.618	0.304	0.855	0.145	0.937
$W^{(aaa)}$	0.034	0.986	0.126	0.940	0.175	0.918	0.120	0.944
$W^{(aha)}$	0.086	0.963	0.178	0.914	0.195	0.909	0.120	0.946
$W^{(hah)}$	0.729	0.547	0.468	0.751	0.266	0.869	0.140	0.937
$W^{(hhh)}$	0.708	0.550	0.487	0.732	0.253	0.879	0.132	0.944

Table 3: Comparison of selected t statistics and versions of the outlier-sum statistic as a function of r, the number of 15 diseased subjects over-expressing 20 of 1,000 simulated genes. Comparisons are made on the basis of the statistical power (Pwr) of a level 0.05 test to detect over-expression for a single gene, as well as the false discovery rate (FDR25) among genes having p values less than the 25th lowest p-value for each statistic. Results are based on 10,000 simulated experiments, each consisting of measurements of expression of 1,000 genes for 15 healthy and 15 diseased cases. Gene expression in the healthy subjects for each gene is presumed to follow either a standard normal distribution or a t distribution with 5 degrees of freedom. The standard error for all numbers presented in this table is less than 0.0008; these numbers are accurate to at least the third digit.

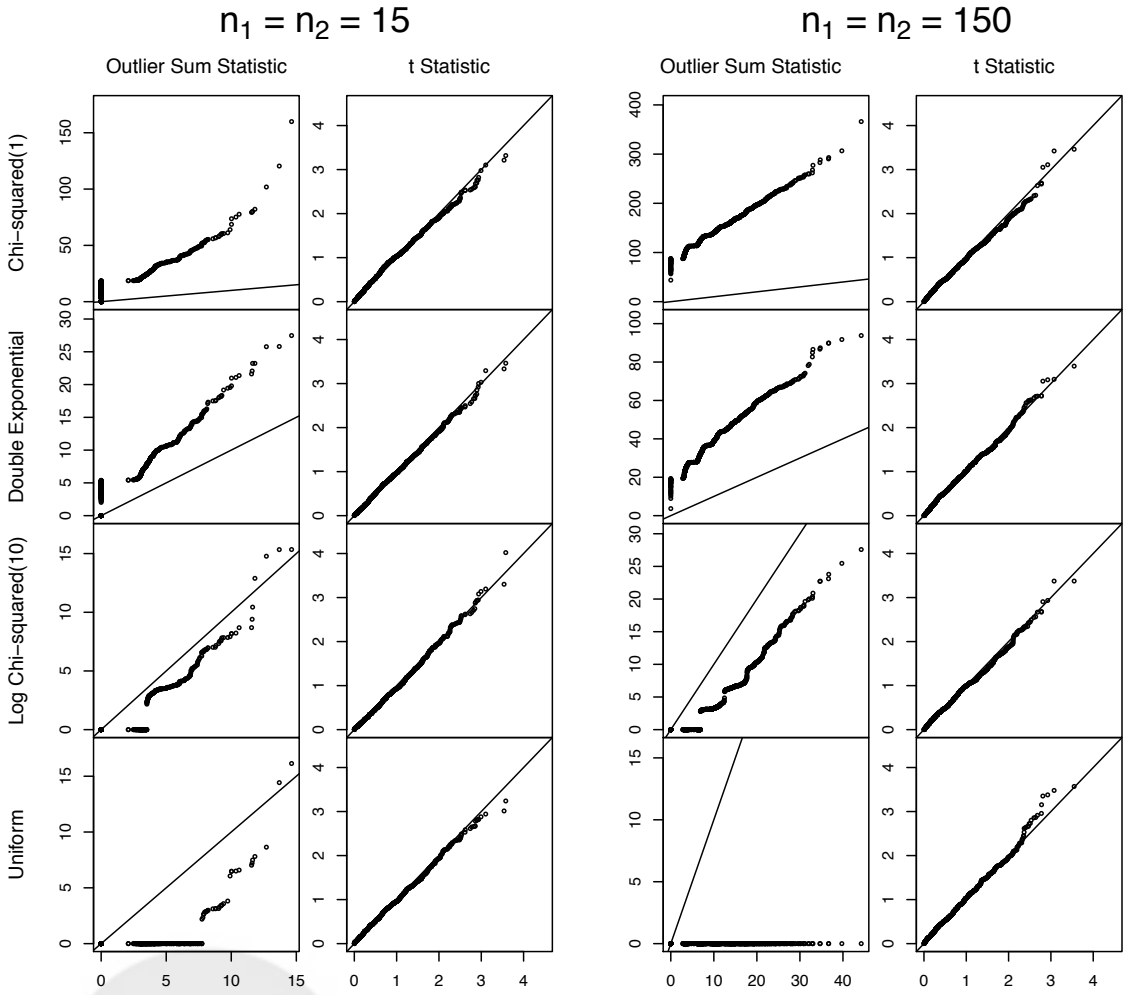

Figure 1: Quantile-quantile plots comparing the null distribution for the outlier-sum statistic when the underlying data distribution changes, and the same plots for the null distribution of the t statistic. For all plots, the reference on the x-axis is the null distribution for the outlier-sum statistic or t statistic when the underlying data is standard normal. The first two columns of plots are for sample sizes of 15 in each group; the last two columns of plots are for sample sizes of 150 in each group.

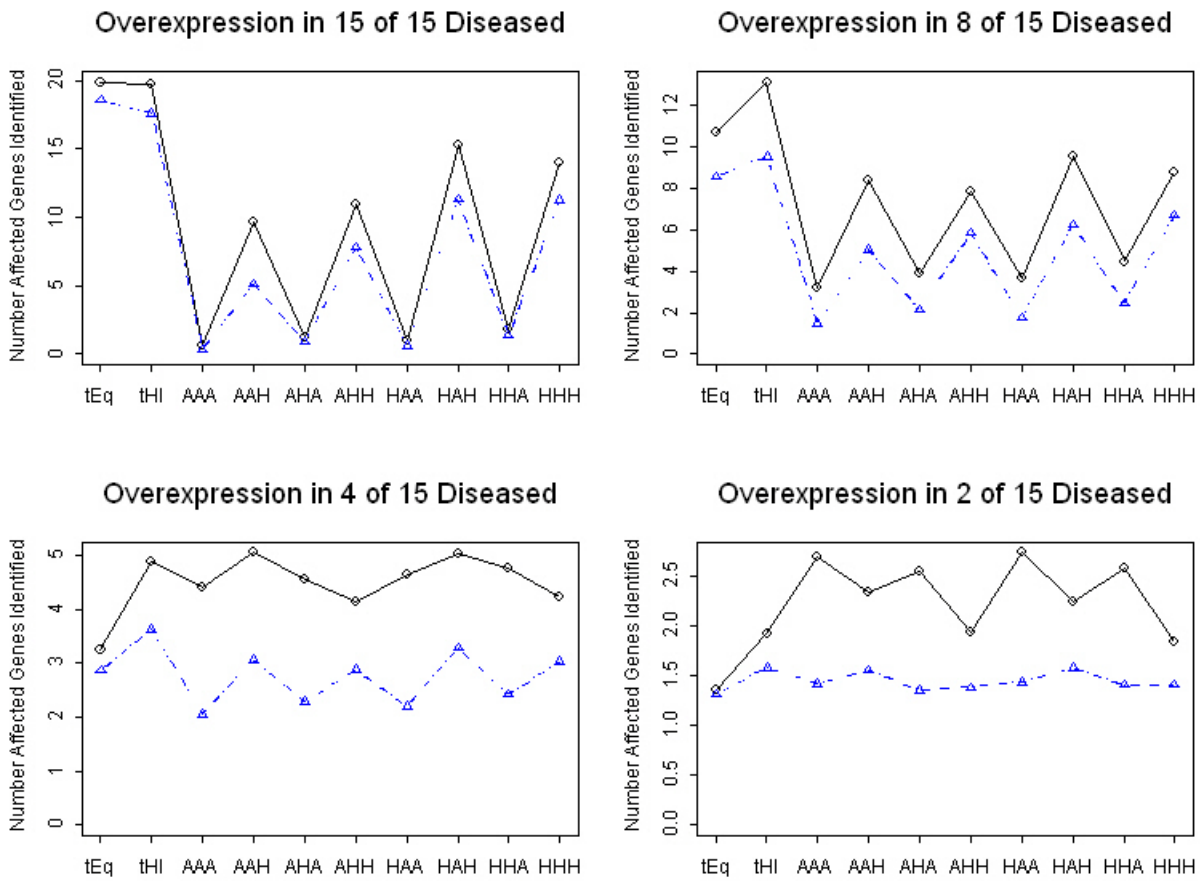

Figure 2: Profile plot of the average number of truly over-expressed genes identified among the genes having as low or lower p-values than the 25th lowest gene for each statistic. Results are based on 10,000 simulated experiments in which expression is measured on 1,000 genes in 15 diseased and 15 healthy subjects. In each experiment, 20 genes are over-expressed in r of 15 diseased subjects. Gene expression in healthy subjects follows a standard normal distribution (black solid line) or a t distribution with 5 degrees of freedom (blue broken line). It should be noted that the scale for the y axis varies in each panel, though the number of over-expressed genes is 20 in each case.