Detecting Differential Gene Expression in Subgroups of a Disease Population

Citation	Emerson, S. C., & Emerson, S. S. (2013). Detecting differential gene expression in subgroups of a disease population. International Journal of Biostatistics, 9(1), 95-108. doi:10.1515/ijb-2013-0010
DOI	10.1515/ijb-2013-0010
Publisher	Walter de Gruyter GmbH
Version	Version of Record
Citable Link	http://hdl.handle.net/1957/47936
Terms of Use	http://cdss.library.oregonstate.edu/sa-termsofuse

Research Article

Sarah C. Emerson* and Scott S. Emerson

Detecting Differential Gene Expression in Subgroups of a Disease Population

Abstract: In many disease settings, it is likely that only a subset of the disease population will exhibit certain genetic or phenotypic differences from the healthy population. Therefore, when seeking to identify genes or other explanatory factors that might be related to the disease state, we might expect a mixture distribution of the variable of interest in the disease group. A number of methods have been proposed for performing tests to identify situations for which only a subgroup of samples or patients exhibit differential expression levels. Our discussion here focuses on how inattention to standard statistical theory can lead to approaches that exhibit some serious drawbacks. We present and discuss several approaches motivated by theoretical derivations and compare to an ad hoc approach based upon identification of outliers. We find that the outlier-sum statistic proposed by Tibshirani and Hastie offers little benefit over a *t*-test even in the most idealized scenarios and suffers from a number of limitations including difficulty of calibration, lack of robustness to underlying distributions, high false positive rates owing to its asymmetric treatment of groups, and poor power or discriminatory ability under many alternatives.

Keywords: efficiency, gene expression analysis, microarray, *t*-test, hypothesis testing

***Corresponding author: Sarah C. Emerson,** Department of Statistics, Oregon State University, 44 Kidder Hall, Corvallis, OR 97330, USA, E-mail: emersosa@stat.oregonstate.edu
Scott S. Emerson, Department of Biostatistics, University of Washington, Seattle, WA, USA, E-mail: semerson@u.washington.edu

1 Introduction

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, 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. In such a setting, a parametric or semi-parametric location shift hypothesis is relatively implausible. Yet in randomized clinical trials, the most common approach to detecting differential outcomes between treatments remains the *t*-test comparing means. The mean is sensitive to a wide variety of effects that a treatment or risk factor might have on the distribution of an outcome variable. Furthermore, contrary to the impression given by many introductory statistics texts (see Chap. 2.4 [1]), the *t*-test will outperform the Wilcoxon test in the presence of heavy-tailed distributions, when the treatment's effect is to modify the propensity to outliers, rather than to just shift the entire distribution [2].

In contrast, in a recent article, Tibshirani and Hastie [3] propose the outlier-sum statistic specifically to detect such phenomena and consider the performance of this procedure in a limited number of examples. Tibshirani and Hastie [3] compare their outlier statistic to the *t*-test and to the COPA method discussed 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 the problem under consideration. In doing so, we found that we were unable to reproduce some of the results presented in their article. This work presents what we believe to be a more accurate comparison and exploration of the theoretical justification and relative performance in simulations of various forms of the *t*-test and several variations on the outlier-sum statistic. We find that the failure to consider basic statistical theory produces a method that offers little benefit even in the most idealized scenarios and suffers from a number of limitations including difficulty of calibration, lack of robustness to underlying distributions, high false positive rates owing to its asymmetric treatment of groups, and poor power or discriminatory ability under many alternatives. 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.

The organization of this article is as follows. For this setting where it is expected or of interest to detect that an effect will only occur in a subgroup of subjects, we first present and explore slight modifications to the usual t statistic that may be used to increase power for these alternatives. We provide discussion supporting the use of a *t*-test in this setting: it is theoretically justified and well-understood, easily calibrated using methods that are robust to different data distributions, and has competitive and often superior power to other methods for many alternatives. Detailed derivation of the *t*-test as a score test for a particular model of this setting is presented to give further justification for its use in this setting.

We then discuss some of the limitations of the proposed outlier-sum statistic and investigate its behavior in a wider variety of situations. As the definition of the statistic in the original article is somewhat ambiguous, we begin our exploration of this approach by presenting the definition of the outlier-sum statistic and discussing some possible variations on the statistic. We then discuss limitations of the outlier sum that might be anticipated based on general statistical theory. This is followed by investigations of its sampling distribution and the ability to calibrate the operating characteristics to desired levels, as well as comparisons of its statistical power and false discovery rates (FDRs) to that of the *t*-test.

2 Inference based on means: *t* statistics

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). Our interest is determining which genes might have G_{i1} different from G_{i2}. We assume patients and disease groups are totally independent. The standard analysis method that would most typically be used to compare two groups would be based on means. We extend this comparison to two other *t*-tests, as described below.

Define for each gene within each disease group the moments $E(X_{ijk}) = \mu_{ik}$ and $\text{Var}(X_{ijk}) = \sigma^2_{ik}$ which have unbiased estimators

$$\hat{\mu}_{ik} = \bar{X}_{i\cdot k} = \sum_{j=1}^{n_k} X_{ijk}/n_k \quad \text{and} \quad s^2_{ik} = \sum_{j=1}^{n_k} \left(X_{ijk} - \bar{X}_{i\cdot k}\right)^2/(n_k - 1).$$

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 (asymptotic) sampling distribution

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

as long as the distributions of both groups have finite variance. Ordering of the genes with respect to over-expression might, therefore, be based on a statistic $t_i = \hat{\delta}_i/\sqrt{\hat{V}_i}$. Possible choices for $\hat{V}_i$ and the corresponding test statistic include:

1. Equal variance t-test: $T_i^{(e)} = \hat{\delta}_i / \sqrt{\hat{V}_i^{(e)}}$ for $\hat{V}_i^{(e)} = s_{(p)i}^2(1/n_1 + 1/n_2)$ where $s_{(p)i}^2$ is the pooled estimate of variance:

$$s_{(p)i}^2 = \frac{(n_1 - 1)s_{i1}^2 + (n_2 - 1)s_{i2}^2}{n_1 + n_2 - 2}.$$

   Genes would typically be ordered by the p-value derived from the t distribution having $d = n_1 + n_2 - 2$ degrees of freedom.
2. Unequal variance t-test: $T_i^{(u)} = \hat{\delta}_i / \sqrt{\hat{V}_i^{(u)}}$ where $\hat{V}_i^{(u)} = \left(s_{i1}^2/n_1 + s_{i2}^2/n_2\right)$. Genes would typically be ordered by on the p-value derived from the t distribution having degrees of freedom d given by the Welch–Satterthwaite approximation.
3. Healthy reference t-test: $T_i^{(h)} = \hat{\delta}_i / \sqrt{\hat{V}_i^{(h)}}$ where $\hat{V}_i^{(h)} = s_{i1}^2(1/n_1 + 1/n_2)$. Genes would typically be ordered by the p-value derived from the t distribution having $d = n_1 - 1$ degrees of freedom.

The relative performance of these tests may be evaluated in terms of consistency (under a fixed alternative, the probability of rejecting the null tends to one, as the sample size tends to infinity) and exactness (under the null hypothesis, the probability of rejecting the null is the nominal level α) or asymptotic exactness. We will later perform simulations to explore small sample calibration and power. When G_{i1} and G_{i2} are normal distributions having $\sigma_{i1}^2 = \sigma_{i2}^2$, the p-values based on $T_i^{(e)}$ and $T_i^{(h)}$ are exact, and 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 that $\sigma_{ik}^2 < \infty$ for $k = 1, 2$, all three tests are consistent tests of the weak null hypothesis $H_{0i}^{(W)} : \mu_{i1} = \mu_{i2}$; the test based on $T_i^{(u)}$ is asymptotically exact; the test based on $T_i^{(e)}$ is asymptotically exact if $n_1 = n_2$; and the test based on $T_i^{(h)}$ is not asymptotically exact. All three tests are asymptotically exact as tests of the strong null hypothesis $H_{0i}^{(S)} : G_{i1} = G_{i2}$, but none are consistent.

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 5.4, 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.

3 Normal mixture model justification for *t* statistics

Use of the t-test in this problem can also be justified as the score test for a normal mixture model. While the normal mixture model does not satisfy our ideal of distribution-free inference, it does lend some theoretical credence to the utility of the t-test for settings beyond the standard population mean question. As the score test has optimal power against small deviations, at least asymptotically, this derivation suggests that the t-test would be a sound choice for detecting subgroups in the mixture-model setting.

Suppose $X_1, \ldots, X_n$ are independent identically distributed random variables distributed according to $X_i \sim N(\mu, \sigma^2)$. Further, suppose $Y_1, \ldots, Y_m$ are independent identically distributed random variables distributed according to a mixture of normals in which proportion p have $Y_i \sim N(\mu + \delta, \sigma^2)$ and proportion $1 - p$ have $Y_i \sim N(\mu, \sigma^2)$. Assuming σ^2 is known, the likelihood function used to estimate $\vec{\theta} = (\mu, \delta, p)^T$ is thus

$$L(\vec{\theta}|\vec{X},\vec{Y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(X_i-\mu)^2}{2\sigma^2}\right\} \times \prod_{j=1}^{m}\left[p\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(Y_j-\mu-\delta)^2}{2\sigma^2}\right\} + (1-p)\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(Y_j-\mu)^2}{2\sigma^2}\right\}\right].$$

The log likelihood function is, thus, proportional to

$$\mathscr{L}(\vec{\theta}|\vec{X},\vec{Y}) \propto \sum_{i=1}^{n}\left\{-\frac{(X_i-\mu)^2}{2\sigma^2}\right\} + \sum_{j=1}^{m}\log\left[p\exp\left\{-\frac{(Y_j-\mu-\delta)^2}{2\sigma^2}\right\} + (1-p)\exp\left\{-\frac{(Y_j-\mu)^2}{2\sigma^2}\right\}\right].$$

From this, we derive the components of the score vector $\mathscr{U}(\vec{\theta}) \equiv \partial\mathscr{L}/\partial\vec{\theta}$ as

$$\frac{\partial\mathscr{L}}{\partial\mu} = \sum_{i=1}^{n}\left\{\frac{(X_i-\mu)}{\sigma^2}\right\} + \sum_{j=1}^{m}\frac{\left[p\frac{(Y_j-\mu-\delta)}{\sigma^2}\exp\left\{-\frac{(Y_j-\mu-\delta)^2}{2\sigma^2}\right\} + (1-p)\frac{(Y_j-\mu)}{\sigma^2}\exp\left\{-\frac{(Y_j-\mu)^2}{2\sigma^2}\right\}\right]}{\left[p\exp\left\{-\frac{(Y_j-\mu-\delta)^2}{2\sigma^2}\right\} + (1-p)\exp\left\{-\frac{(Y_j-\mu)^2}{2\sigma^2}\right\}\right]},$$

$$\frac{\partial\mathscr{L}}{\partial\delta} = \sum_{j=1}^{m}\frac{\left[p\frac{(Y_j-\mu-\delta)}{\sigma^2}\exp\left\{-\frac{(Y_j-\mu-\delta)^2}{2\sigma^2}\right\}\right]}{\left[p\exp\left\{-\frac{(Y_j-\mu-\delta)^2}{2\sigma^2}\right\} + (1-p)\exp\left\{-\frac{(Y_j-\mu)^2}{2\sigma^2}\right\}\right]},$$

$$\frac{\partial\mathscr{L}}{\partial p} = \sum_{j=1}^{m}\frac{\left[\exp\left\{-\frac{(Y_j-\mu-\delta)^2}{2\sigma^2}\right\} - \exp\left\{-\frac{(Y_j-\mu)^2}{2\sigma^2}\right\}\right]}{\left[p\exp\left\{-\frac{(Y_j-\mu-\delta)^2}{2\sigma^2}\right\} + (1-p)\exp\left\{-\frac{(Y_j-\mu)^2}{2\sigma^2}\right\}\right]}.$$

Using the above, we can easily calculate the score vector under any hypothesized value of $\vec{\theta} = \vec{\theta}_0$. Under the standard regularity conditions, we also know from likelihood theory that when $\vec{\theta} = \vec{\theta}_0$,

$$\mathscr{U}(\vec{\theta}_0) \sim N_3\left(\vec{0}, \mathbf{J}(\vec{\theta}_0)\right),$$

where $\mathbf{J}(\vec{\theta}_0)$ is the information matrix

$$\mathbf{J}(\vec{\theta}_0) \equiv \left[-E\left(\frac{\partial^2}{\partial\vec{\theta}\partial\vec{\theta}^T}\mathscr{L}(\vec{\theta})\right)\right]|_{\vec{\theta}=\vec{\theta}_0}.$$

We are interested in testing the composite null hypothesis $H_0 : \delta = 0$. We would, thus, need to estimate the maximum likelihood estimate $\widehat{\vec{\theta}_0}$. This is easily obtained by noting that under $H_0 : \vec{\theta} = \vec{\theta}_0 = (\mu, \delta = 0, p)^T$, the log likelihood becomes

$$\mathscr{L}(\vec{\theta}_0|\vec{X},\vec{Y}) \propto \sum_{i=1}^{n}\left\{-\frac{(X_i-\mu)^2}{2\sigma^2}\right\} + \sum_{j=1}^{m}\left\{-\frac{(Y_j-\mu)^2}{2\sigma^2}\right\},$$

which does not involve p. It is easily seen, therefore, that $\widehat{\vec{\theta}_0} = (\hat{\mu}_0 = (n\overline{X} + m\overline{Y})/(n+m), \delta_0 = 0, p_0)^T$ maximizes the log likelihood under H_0 for every choice of $p_0 \in [0,1]$. From the above, it is apparent we run into problems due to the nonidentifiability of δ and p under H_0: the model is over-parameterized under the null hypothesis. That is, so long as $p = 0$, δ can be anything, and we have the same likelihood. Alternatively, so long as $\delta = 0$, p can be anything, and we have the same likelihood. Inspecting the formulas, it would seem that we can gain some intuition by deciding to use $H_0 : \delta = 0$ with an arbitrary

fixed nonzero $p = p_0 \in (0,1)$, which will give us a nondegenerate distribution that just differs by a scale factor. That is, for $H_0 : \vec{\theta} = \vec{\theta}_0 = (\mu, \delta = 0, p = p_0)^T$, the log likelihood is still

$$\mathscr{L}(\vec{\theta}_0|\vec{X},\vec{Y}) \propto \sum_{i=1}^{n}\left\{-\frac{(X_i-\mu)^2}{2\sigma^2}\right\} + \sum_{j=1}^{m}\left\{-\frac{(Y_j-\mu)^2}{2\sigma^2}\right\}.$$

This reduced problem satisfies the regularity conditions of likelihood theory. It is easily seen, therefore, that $\widehat{\vec{\theta}_0} = (\hat{\mu}_0 = (n\overline{X} + m\overline{Y})/(n+m), \delta_0 = 0, p_0)^T$. We can, thus, test the null hypothesis using $\mathscr{U}(\widehat{\vec{\theta}_0})$ given by

$$\begin{aligned}
\mathscr{U}_1(\widehat{\vec{\theta}_0}) &= \sum_{i=1}^{n}\left\{\frac{(X_i-\hat{\mu}_0)}{\sigma^2}\right\} \\
&\quad + \sum_{j=1}^{m}\frac{\left[p_0\frac{(Y_j-\hat{\mu}_0-\delta_0)}{\sigma^2}\exp\left\{-\frac{(Y_j-\hat{\mu}_0-\delta_0)^2}{2\sigma^2}\right\} + (1-p_0)\frac{(Y_j-\hat{\mu}_0)}{\sigma^2}\exp\left\{-\frac{(Y_j-\hat{\mu}_0)^2}{2\sigma^2}\right\}\right]}{\left[p_0\exp\left\{-\frac{(Y_j-\hat{\mu}_0-\delta_0)^2}{2\sigma^2}\right\} + (1-p_0)\exp\left\{-\frac{(Y_j-\hat{\mu}_0)^2}{2\sigma^2}\right\}\right]} \\
&= \sum_{i=1}^{n}\left\{\frac{(X_i-\hat{\mu}_0)}{\sigma^2}\right\} + \sum_{j=1}^{m}\left\{\frac{(Y_j-\hat{\mu}_0)}{\sigma^2}\right\} \\
&= \frac{n\overline{X} + m\overline{Y} - (m+n)\hat{\mu}_0}{2\sigma^2} \\
&= 0,
\end{aligned}$$

$$\begin{aligned}
\mathscr{U}_2(\widehat{\vec{\theta}_0}) &= \sum_{j=1}^{m}\frac{\left[p_0\frac{(Y_j-\hat{\mu}_0-\delta_0)}{\sigma^2}\exp\left\{-\frac{(Y_j-\hat{\mu}_0-\delta_0)^2}{2\sigma^2}\right\}\right]}{\left[p_0\exp\left\{-\frac{(Y_j-\hat{\mu}_0-\delta_0)^2}{2\sigma^2}\right\} + (1-p_0)\exp\left\{-\frac{(Y_j-\hat{\mu}_0)^2}{2\sigma^2}\right\}\right]} \\
&= \sum_{j=1}^{m} p_0\frac{(Y_j-\hat{\mu}_0)}{\sigma^2} = p_0\frac{mn}{m+n}\frac{\overline{Y}-\overline{X}}{2\sigma^2}.
\end{aligned}$$

Note that the only non-zero element of the score statistic is proportional to the difference in means between the two groups, and therefore, inference based on the score statistic is equivalent to the two-sample *t*-test in the sense that rejection regions for properly calibrated versions of the score test and the *t*-test will be asymptotically identical.

4 Outlier-sum statistic and variations

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 (IQR). 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.

Tibshirani and Hastie use a standardization based on the median observation and the median absolute deviation (MAD) from the median. Let m_{i1} and $m_{i\cdot}$ be the median gene expression of the *i*th gene for, respectively, the healthy patients and the combined sample of healthy and diseased patients. 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, because both 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".

For each standardization approach, we can also compute the 75th percentile and IQR for gene i in two possible ways: the quartiles may be based on the entire set of standardized values for gene i and may be based on only the standardized values of the healthy group for gene i. Letting $q_{75}^{(aaa)}$ and $\mathrm{IQR}^{(aaa)}$ denote the 75th percentile and IQR of the entire set of the $X^{(aa)}$ standardized values of gene, and letting $q_{75}^{(aah)}$ and $\mathrm{IQR}^{(aah)}$ denote the 75th percentile and IQR of just the healthy set of the $X^{(aa)}$ standardized values of gene i, two possible variants of an outlier-sum statistic are given by

$$W_i^{(aaa)} = \sum_{j=1}^{n_2} \widetilde{X}_{ij2}^{(aa)} \, \mathbb{I}\left\{\widetilde{X}_{ij2}^{(aa)} > q_{75}^{(aaa)} + \mathrm{IQR}^{(aaa)}\right\},$$

$$W_i^{(aaa)} = \sum_{j=1}^{n_2} \widetilde{X}_{ij2}^{(aa)} \, \mathbb{I}\left\{\widetilde{X}_{ij2}^{(aa)} > q_{75}^{(aah)} + \mathrm{IQR}^{(aah)}\right\},$$

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:

$$W^{(aaa)}, \quad W^{(aah)}, \quad W^{(aha)}, \quad W^{(ahh)}, \quad W^{(haa)}, \quad W^{(hah)}, \quad W^{(hha)}, \quad W^{(hhh)},$$

where the superscripts denote first the median used as the center of the standardization, second the MAD used to scale the standardization, and third the quartiles used to define the threshold for outliers. It seems from the descriptions in the article 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, with details and results available in Supplementary Materials. Here, we provide results for $W^{(aaa)}$ and $W^{(hhh)}$. 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

$$d_{ip}^{*} = \mathrm{median}\left\{|X_{ijk} - m_{ik}| \;:\; k = 1, 2;\; j = 1, \ldots, n_k\right\},$$

the median value across groups of the within-group absolute deviations.

5 Theoretical issues with the outlier-sum approach

There are several issues with the outlier-sum statistic that warrant discussion. A more detailed discussion of these issues is available as a technical report [8]. In this article, we will not address the use of this statistic in the one-sample setting, as discussed in Section 3 of the original article, owing to the limitations of space required to discuss the problematic nature of distribution-free outlier detection in that setting: it is unclear what null hypothesis the statistic is intended to test, and the choice of how or whether to standardize the data becomes even more influential.

In the two-sample setting, 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 article, 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 of the original paper by Tibshirani and Hastie (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.

We first examine the statistical theory relevant to the standardization of observations and definition of outliers. 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.

5.1 Standardization of data and detection of outliers

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 standardize 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. 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.

When looking for differences in distributions, this latter standardization might have an advantage due to Chebyshev's inequality: seeing a substantially larger proportion of diseased subjects' standardized measurements greater than some threshold is suggestive of a difference in distributions. Furthermore, by increasing the threshold used to declare outliers, we are guaranteed of decreasing the probability of declaring outliers for every null distribution having a variance. There is no clear analogy to Chebyshev's inequality when basing data standardization on the median and MAD as is used in the outlier-sum statistic. The proportion of a general distribution whose values, standardized by the median and MAD, can exceed every value of $c > 0$ is bounded above only by 0.5. Thus, increasing the threshold is not guaranteed to decrease the proportion of the null distribution declared an outlier.

As a monotonic transformation of the data, the choice of standardization neither does 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.

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. The properties of the outlier-sum statistic's definition of outliers are more difficult to describe in general. The effect of scaling by the MAD for the combined sample versus the MAD for the healthy population alone is not easily quantified. In the setting of an effect of disease that generates higher levels of gene expression in some or all individuals, the MAD for the combined sample can be larger or smaller than the MAD for the healthy population, depending on the shape of the distribution in the healthy population.

Furthermore, as noted above, the probability that an individual observation would be an arbitrarily large number of MADs 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 criterion could merely be identifying those 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 based on the IQR, whereas a normal mixture model having 24.5%, 51%, and 24.5% of the observations as $N(-1, \tau^2)$, $N(0, \tau^2)$, and $N(1, \tau^2)$, respectively, with $\tau = 0.1$ results in 24.5% of the observations characterized as high outliers. 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 MAD becomes very small. Code exploring the patterns in MAD and propensity to outliers for a variety of mixture distributions is available in the Supplementary Materials.

5.2 Calibration

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 statistic 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 MADs used to standardize the measurements were nonzero. Chen et al. [6] 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 or varies across genes their results do not apply.

Calibration of the outlier-sum statistic in the general setting is extremely difficult. Very different null distributions for the statistic will result from different underlying data distributions. For instance, if all the data are 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 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 are standard normal. Similar discrepancies are observed for all 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 that are obtained for all the data-generating mechanisms, with increasing similarity among the null distributions, as the sample sizes are increased.

The use of a permutation approach to calibrate the statistic also performs poorly 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 behavior will depend on the relative size of the entire diseased group and on the number of outliers according to the specified criterion. Tibshirani and Hastie use an empirical calibration based on other rows of the data set; this approach will be very problematic if the data do not all come from the same distribution, or if there are a reasonable number of non-null genes. 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 produce inaccurate calibration.

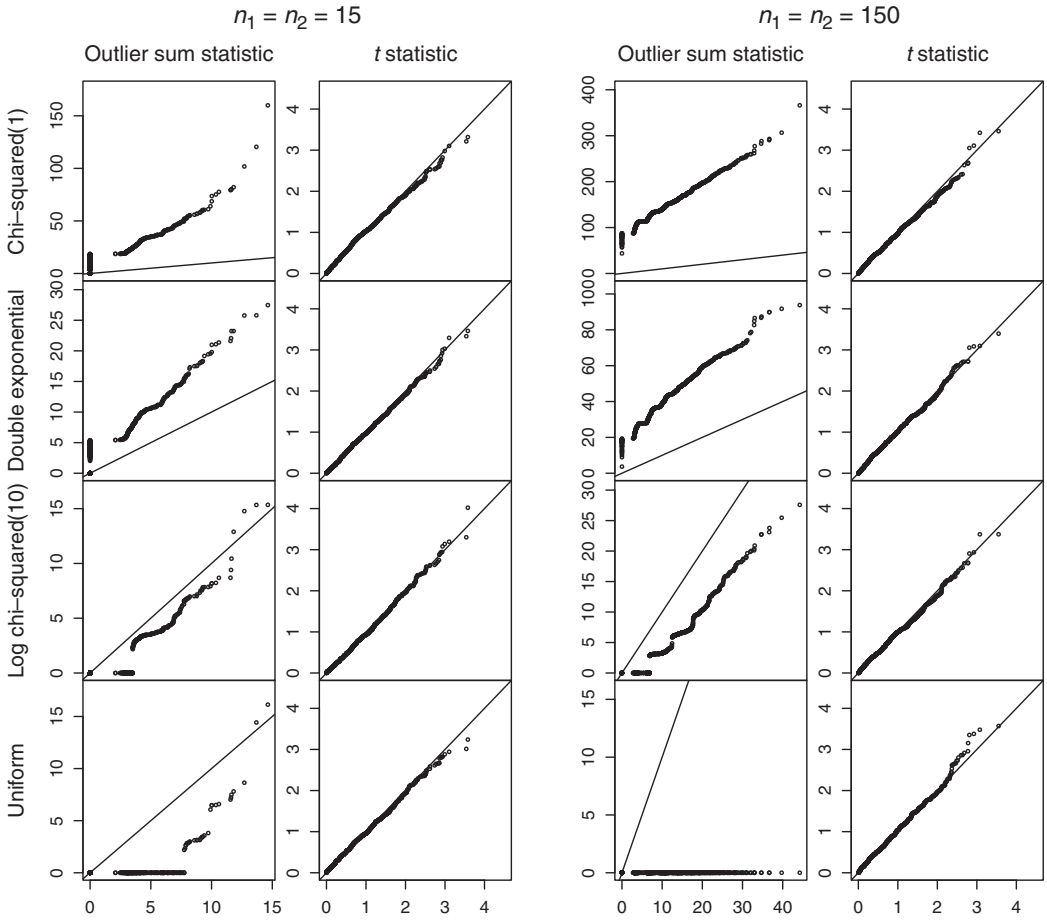

Figure 1 Quantile–quantile plots comparing the null distribution for the outlier-sum statistic when the underlying data distribution changes and the same plots for the null distribution of the *t* statistic. For all plots, the reference on the *x*-axis is the null distribution for the outlier-sum statistic or *t* statistic when the underlying data are 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

The outlier-sum statistic 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 very suboptimal due to 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 FDR when attention is focused on some number of genes having the largest statistics.

5.3 Sampling distribution of *p*-values under the alternative

In their article, Tibshirani and Hastie's article 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 article did not to us appear

Table 1 Comparison of results from 10,000 simulated experiments in which expression is measured on 1,000 genes in 15 diseased and 15 healthy subjects

	$r = 15$		$r = 8$		$r = 4$		$r = 2$	
	Mean (SD)	**Mdn**	**Mean (SD)**	**Mdn**	**Mean (SD)**	**Mdn**	**Mean (SD)**	**Mdn**
$T^{(e)}$	0.000	0.000	0.037	0.013	0.176	0.115	0.316	0.262
	(0.001)		(0.065)		(0.180)		(0.247)	
$T^{(h)}$	0.000	0.000	0.025	0.005	0.155	0.080	0.307	0.239
	(0.002)		(0.058)		(0.188)		(0.260)	
$W^{(aaa)}$	0.518	0.673	0.328	0.227	0.236	0.114	0.279	0.161
	(0.240)		(0.284)		(0.261)		(0.266)	
$W^{(aha)}$	0.495	0.673	0.306	0.152	0.228	0.099	0.280	0.166
	(0.270)		(0.295)		(0.263)		(0.265)	
$W^{(hhh)}$	0.031	0.003	0.077	0.023	0.165	0.082	0.272	0.191
	(0.090)		(0.143)		(0.205)		(0.245)	
T&H	0.106	0.11	0.094	0.105	0.093	0.098	0.100	0.100
	(0.019)		(0.030)		(0.130)		(0.131)	

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. Inspecting the provided quantities, we notice that in Table 1 of the original article, 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 and therefore the power of a level 0.05 test would be tiny. 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.

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 article, and we present our results below. Both authors of this manuscript have independently confirmed all 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 article, for each of 10,000 simulated experiments, we generated a $30 \times 1{,}000$ matrix of independent observations drawn from the standard normal distribution. The first 15 rows of the matrix represent diseased subjects, and the last 15 represent 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, and outliers are identified and summed as described in Section 4. Here p-values were calculated according to

$$p_1 = \frac{1}{999}\left[\sum_{i=2}^{1{,}000} \mathbb{I}\{W_i > W_1\} + \frac{1}{2}\sum_{i=2}^{1{,}000} \mathbb{I}\{W_i = W_1\}\right],$$

so that the expectation of the p-value in the null case is 0.5 as it should be. We also computed the three versions of t statistics ($T_i^{(e)}$, $T_i^{(u)}$, and $T_i^{(h)}$) 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 1, along with the values from Table 2 of the original article. We show results for four versions of the outlier-sum statistic, with results for all eight 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 that can be explained by random variation in the simulations. We note that additional investigations

Table 2 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, where over-expression is a signal of size 2 added to the affected genes and subjects

	Standard normal distribution							
	$r = 15$		$r = 8$		$r = 4$		$r = 2$	
	Pwr	FDR25	Pwr	FDR25	Pwr	FDR25	Pwr	FDR25
$T^{(e)}$	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
$W^{(hhh)}$	0.834	0.442	0.606	0.650	0.339	0.831	0.168	0.927
	t Distribution 5 df							
	$r = 15$		$r = 8$		$r = 4$		$r = 2$	
	Pwr	FDR25	Pwr	FDR25	Pwr	FDR25	Pwr	FDR25
$T^{(e)}$	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
$W^{(hhh)}$	0.708	0.550	0.487	0.732	0.253	0.879	0.132	0.944

using alternative definitions of empirical p-values (i.e. methods that handle tied observations differently) also did not agree with the published results.

In each experiment, gene 1 is over-expressed with a signal of size 2 added to r of 15 diseased subjects. Gene expression in healthy subjects follows a standard normal distribution. 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) are presented. We also provide the analogous statistics based on the 50 simulations as reported by Tibshirani and Hastie (T&H).

5.4 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 2. The $W^{(aaa)}$ version of the outlier-sum statistic, which we presume to be the version used in the original article, 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 5. 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 $T_i^{(h)}$ 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 $r = 15$. Power is low when $r = 2$ or 4 at the effect and sample sizes considered – at these low sample and effect sizes no reasonable test would have good power – but power would obviously improve for both methods if either quantity were increased.

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 come from a distribution with heavier tails than the normal distribution. Results for data distributed according to the t distribution with five degrees of freedom are also shown in Table 2. Similar results were observed as the sample size was increased to 60 in each group, with the proportion of diseased subjects exhibiting over-expressed genes was kept constant.

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 FDR (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 five 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.

5.5 False discovery rates

In their article introducing the outlier-sum statistic, Tibshirani and Hastie did not systematically investigate the FDR 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 FDR 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 2 presents the average FDR 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 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 for $r = 16$. 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.

6 Discussion

Based on the comparisons presented here and those available as Supplementary Materials, it seems clear to us that traditional statistical methods based on sound theoretical justification offer many advantages over

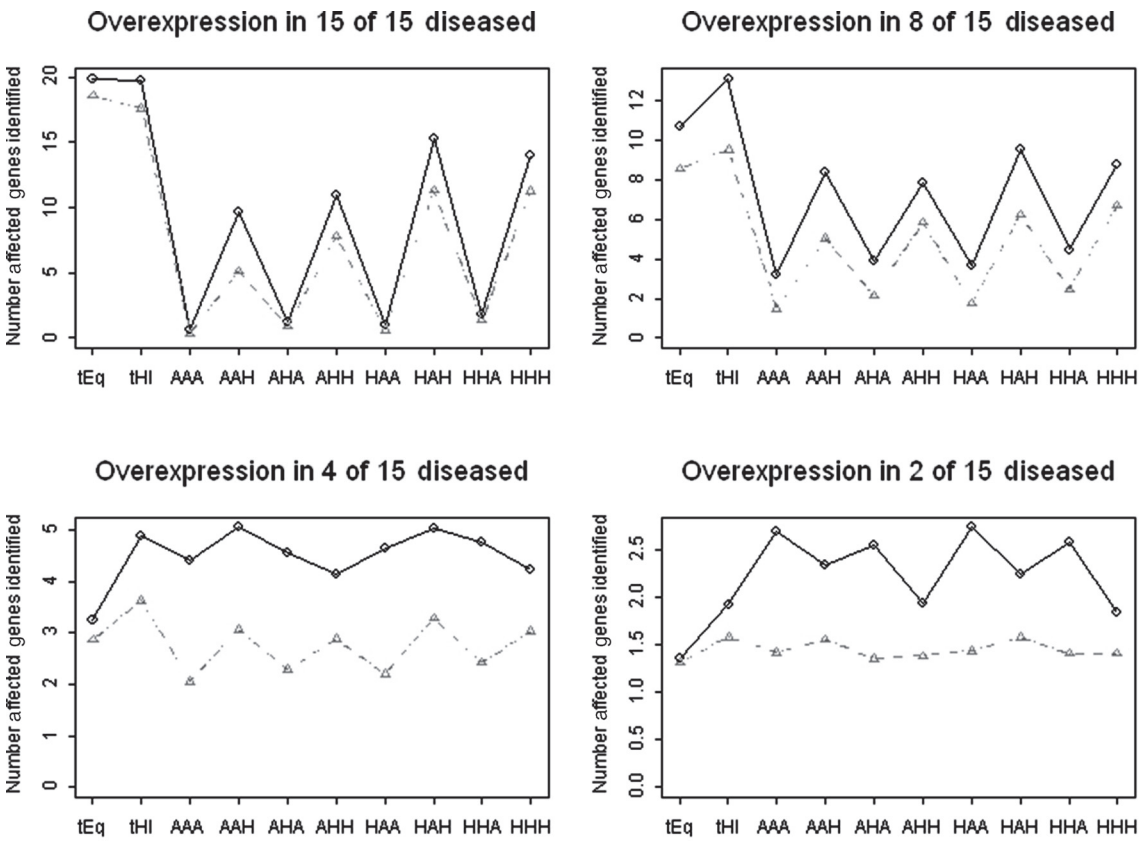

Figure 2 Profile plot of the average number of truly over-expressed genes identified among the genes having as low or lower *p*-values than the 25th lowest gene for each statistic. Results are based on 10,000 simulated experiments in which expression is measured on 1,000 genes in 15 diseased and 15 healthy 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 five 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

ad hoc procedures in terms of both robustness and power. We observed that the outlier-sum approach offers very little benefit over a standard or slightly modified *t*-test 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 either power to detect an alternative or FDRs) 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 fishes 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.

A number of other approaches to problems like this one have been proposed, in Tomlins et al. [4] and Ghosh and Chinnaiyan [7] among others. Ghosh and Chinnaiyan [7] discuss a mixture-model approach where the null distribution for each gene is estimated from the healthy/control population, and then samples are identified as outliers for the gene in question if their values for that gene are at the extremes of the empirical cdf of the control population. In doing so, they do not really use the mixture model in the

formulation of their nonparametric statistic, and, indeed, the mixture parameter and subgroup distribution are nonidentifiable in a nonparametric setting, a point that was apparently not recognized by the authors. In their simulations using mixtures of normals and exponentials, Ghosh and Chinnaiyan [7] consider the outlier-sum statistic and a couple of other statistics that they refer to as "modified *t* statistics", but they do not directly make comparisons to the ordinary *t*-test. However, based on the motivation provided by the derivation for the score test in Section 3, it is to be expected that in general testing based on differences in means would perform better than their statistic under local alternatives. In explorations not presented here, but available from the authors, we have attempted to replicate the simulations in Ghosh and Chinnaiyan [7] and find that this intuition is in fact borne out. However, the additional issues that need to be considered when using the statistics based on the empirical distribution are beyond the scope of this article.

Lastly, the explorations and remarks presented here, while focused on a particular setting and method, are actually intended to 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, as well as 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 help to understand when a particular approach is suitable, and whether an existing and better-understood method might instead suffice.

7 Supplementary materials

Supplementary materials include the R code used to make the comparisons in this paper, as well as additional explorations not reported here.

References

1. Lehmann EL. Nonparametrics: statistical methods based on ranks. New York, NY: Springer, 2006.
2. Emerson SS. Some observations on the Wilcoxon rank sum test. Technical Report 380, University of Washington, 2011. Available at: http://www.bepress.com/uwbiostat/paper380
3. Tibshirani R, Hastie T. Outlier sums for differential gene expression analysis. Biostatistics 2007;8:2–8.
4. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun X-W, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005;310:644–8.
5. Wu B. Cancer outlier differential gene expression detection. Biostatistics 2007;8:566–75.
6. Chen L, Chen D, Chan W. The distribution-based p-value for the outlier sum in differential gene expression analysis. Biometrika 2010;97:246–53.
7. Ghosh D, Chinnaiyan AM. Genomic outlier profile analysis: mixture models, null hypotheses, and nonparametric estimation. Biostatistics 2009;10:60–9.
8. Emerson SC, Emerson SS. The importance of statistical theory in outlier detection. Technical Report 381, University of Washington, 2011. Available at: http://www.bepress.com/uwbiostat/paper381