Bioinformatics, 31(14), 2015, 2310–2317
doi: 10.1093/bioinformatics/btv154
Advance Access Publication Date: 18 March 2015
Original Paper

OXFORD

Gene expression

Two-group comparisons of zero-inflated intensity values: the choice of test statistic matters

Andreas Gleiss[1], Mohammed Dakna[2], Harald Mischak[2] and Georg Heinze[1,*]

[1]Center for Medical Statistics, Informatics and Intelligent Systems, Medical University Vienna, Austria, Vienna, Austria and [2]Mosaiques Diagnostics and Therapeutics AG, Hannover, Germany

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on April 8, 2014; revised on November 20, 2014; accepted on March 16, 2015

Abstract

Motivation: A special characteristic of data from molecular biology is the frequent occurrence of zero intensity values which can arise either by true absence of a compound or by a signal that is below a technical limit of detection.

Results: While so-called two-part tests compare mixture distributions between groups, one-part tests treat the zero-inflated distributions as left-censored. The left-inflated mixture model combines these two approaches. Both types of distributional assumptions and combinations of both are considered in a simulation study to compare power and estimation of log fold change. We discuss issues of application using an example from peptidomics.

The considered tests generally perform best in scenarios satisfying their respective distributional assumptions. In the absence of distributional assumptions, the two-part Wilcoxon test or the empirical likelihood ratio test is recommended. Assuming a log-normal subdistribution the left-inflated mixture model provides estimates for the proportions of the two considered types of zero intensities.

Availability: R code is available at http://cemsiis.meduniwien.ac.at/en/kb/science-research/software/

Contact: georg.heinze@meduniwien.ac.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The appropriateness of a statistical test for detecting differential expression in omics experiments, e.g. microRNA, metabolomics or peptidomics, is primarily determined by the data distribution. After pre-processing steps intensity values of such experiments typically show a distribution consisting of a certain proportion of values at a point-mass at zero (*point-mass values*, PMVs) and a continuous component. Dakna *et al.* (2010) distinguish between two types of PMVs: *biological* PMVs (BPMVs), meaning that the compound's intensity is absent; and *technical* PMVs (TPMVs) where a compound is present but its signal is below the limit of detection (LOD).

Data taken from http://mosaiques-diagnostics.de/diapatpcms/mosaiquescms/front_content.php?idcat=257 will be used to demonstrate the occurrence of PMVs (see Coon *et al.*, 2008; Siwy *et al.*, 2011). Urine samples from 100 patients with chronic kidney disease (CKD; estimated glomerular filtration rate < 45) and 100 samples from control patients (estimated glomerular filtration rate > 60) are available. The dataset comprises 5,616 peptides of 25 randomly selected CKD and 25 control patients. Across all peptides the median PMV proportion amounts to 96% (quartiles 80%, 100%) in the control and 92% (quartiles 80%, 100%) in the CKD group. For the majority of peptides the PMV proportions in the two groups are

close to each other (quartiles of the intra-compound difference of the number of PMVs are −1 and +2). Figure 1 shows histograms for two selected peptides.

Similarly to the definition of biological and technical PMVs, Taylor and Pollard (2009) distinguish between 'true zeros' and 'truncated values' where the truncation results either from signals below the LOD or from a 'lower bound on meaningful signal set by the researcher'. Examples of such lower bounds are given in Neuhäuser *et al.* (2005) for microarray data. The same authors also discuss DNA methylation data where negative results due to only partially or un-methylated test regions would correspond to BPMVs in our nomenclature. BPMVs are also comparable to the 'rancid part of the sample' in Hallstrom (2010).

In typical omics experiments, both, biological and technical, PMVs may occur. A biological phenomenon may lead to highly expressed compounds in one part of the samples, but to low or absent signals in another part, where different compounds are activated instead. This could be caused by sporadic posttranslational modifications which do not directly connect to the investigated experimental condition. As a result, for many compounds the observed distribution of intensities will be bimodal, with a lower mode exactly at the point mass and an upper mode characterizing the location of the continuous, i.e. highly expressed, part of the distribution. TPMVs may have two reasons: first, in some compounds, the distribution of intensities may be characterized by a high variance and a low mean but without bimodality. In these compounds we will find a high proportion of values below detection limit. Second, it is also plausible that the true intensity of a compound is just above the detection limit, but is not correctly detected because of technical reasons such as interference (e.g. adsorption of the compound or signal suppression) or misinterpretation of the signal (e.g. caused by incorrect charge assignment or error in monoisotopic mass determination). Therefore, statistical test procedures for comparing omics data between groups have to accommodate the information contained not only in the continuous part of the distribution but also the information present in the biological as well as in the TPMVs. Depending on whether the PMVs are considered as technical or biological the distribution of the continuous part is more adequately described by a left-censored or by a fully observed (uncensored) distribution.

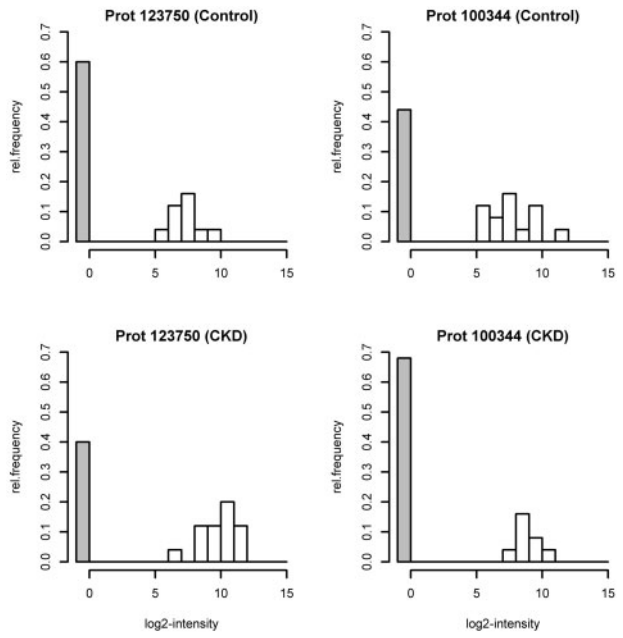

Fig. 1. Histograms of two selected peptides from example dataset (shaded bars represent PMV proportions). Peptide 123750 (left): consonant; Peptide 100344 (right): dissonant

According to the observed distributions of compounds in two groups, Taylor and Pollard (2009) as well as Dakna *et al.* (2010) have distinguished between *consonant* and *dissonant* compounds. The former are characterized by exhibiting the higher PMV proportion in the group with the lower mean in the continuous part, while in the latter case the group with the higher PMV proportion also has the higher mean. This definition does not distinguish between technical and biological PMVs. However, the occurrence of TPMVs naturally corresponds to consonant compounds whereas BPMVs generally allow for both types. In our dataset we found 2942 (52%) consonant and 1153 (21%) dissonant peptides (27% have equal PMV proportions). An example of each type is shown in Figure 1.

There is extensive literature, independent of the context of omics data, on two-group comparisons for data including values below a LOD. A thorough overview is given by Zhang *et al.* (2009), who investigate the following test statistics (among others): various ad-hoc approaches to impute PMVs before applying a *T*-test; Tobit models for an explicit modeling of left-censored distributions; and log-rank type tests applied to the 'flipped' and thus right-censored data (see also Helsel, 2012). In all these tests, in some way or another, the PMVs are seen as being an implicit part of the full distribution instead of as one of two independent components of a mixture distribution. In this sense we will denote these tests as *one-part tests* and view them as primarily targeting distributions that contain TPMVs.

Up to now the test statistics explicitly proposed for omics data comprise only *two-part tests*. These contain two independent components that contribute to a pooled test statistic, one measuring differences in the PMV proportion and the other one in the continuous subdistribution. Two-part test statistics are by construction adequate for cases where PMVs are primarily assumed as biological. Furthermore, the non-parametric empirical likelihood ratio test (ELRT) proposed by Taylor and Pollard (2009) is implicitly of a two-part nature. The only two one-part tests investigated as competitors by these authors are the *t*-test and Wilcoxon's rank–sum test. The same holds for other comparison studies such as that of Dakna *et al.* (2010).

The goal of this paper is to evaluate the various statistical tests proposed for comparing distributions with zero-inflation under scenarios allowing for both types of PMVs. To our best knowledge no simulation study has yet been performed that directly compares two-part and one-part tests for the setting of omics data controlling for biological as well as technical PMVs simultaneously.

The following section gives an overview of the test statistics considered for comparison in this paper. The subsequent two sections present the results of a comprehensive simulation study and of the application to the above mentioned real peptidomics data. The final section will discuss strengths and limitations of our study and draw some conclusions.

2 Methods

We consider experiments where the expressions of K compounds are compared between two groups. The corresponding $\log_2$-transformed expression values are denoted by X_{ki} for $i = 1, \ldots, n_x$ and Y_{ki} for $i = 1, \ldots, n_y$, where n_x and n_y denote the number of samples in the two groups, and $k = 1, \ldots, K$. In the following, we will describe, for each test, the way to obtain a p-value and a corresponding estimator of the log fold change (LFC), i.e. of the location shift between the

groups. This shift concerns the full distributions in the case of one-part tests and the continuous subdistributions for two-part tests. Thus, one-part tests correspond to a null hypothesis on the full-distribution LFC between the groups, including the PMVs at the lower end of the distributions. Conversely, two-part tests test a composite null hypothesis on the subdistribution LFC and the difference in PMV proportions. Thus, estimates of the LFC obtained from the data can either correspond to the full distributions or to the continuous subdistributions.

In the following we omit the index k in this section for simplicity and set $X_i = \pi$, a value selected to represent PMVs (e.g. $\pi = 0$), for $i = 1, \ldots, m_x$ and $Y_i = \pi$ for $i = 1, \ldots, m_y$, assuming $m_x \leq m_y$ without restricting generality.

2.1 One-part tests

Wilcoxon's rank–sum test (W): The normalized and continuity corrected test statistic is given by

$$W = \frac{|U - \mu_W| - 0.5}{\sigma_W}$$

where $U = n_x n_y + n_x(n_x+1)/2 - R_X$ with $R_X = \sum_{i=1}^{n_x} r(X_i)$, the sum of group independent ranks in the first group, $\sigma_W = \sqrt{n_x n_y (n_x + n_y + 1)/12}$ and $\mu_W = n_x\, n_y/2$. The p-value is obtained by comparing W with standard normal quantiles. An estimator of the LFC is calculated as the Hodges–Lehmann estimator of location shift:

$$\hat{\Delta}_W = median\{Y_j - X_i | i = 1, ..., n_x, j = 1, ..., n_y\}$$

Note that this estimate might depend on the choice for π if the PMV proportions are large. In applications appropriate pre-filtering to discard such compounds beforehand will circumvent this problem. For the simulation studies we use appropriately robust measures to summarize LFC estimates.

Adapted T-test (T): To make the standard T-test more competitive for point-mass data we impute PMVs separately in each of the two groups and for each compound in the following way. If the observed proportion m_x/n_x of PMVs in the first group is below 30% then we impute the quantile corresponding to $(m_x/n_x)/2$ from a normal distribution N(median$\{X_i \mid i = 1, \ldots, n_x\}$, 1.4826 MAD$\{X_i \mid i = 1, \ldots, n_x\}$) where MAD denotes the median absolute deviation from the median (1.4826 being the inverse of the MAD of a standard normally distributed random variable and thus a robust estimate of the standard deviation in case of PMVs). Otherwise a value of $\log_2(LOD/2)$ is imputed (cf. the 'Imputed (LOD/2)' method in Zhang *et al.*, 2009). If a LOD is not given a priori then a LOD could be defined as the minimum of all expressed values in the dataset. The same type of imputation is done independently for the second group. By denoting the imputed values as X_i^* for $i = 1, \ldots, m_x$ and Y_i^* for $i = 1, \ldots, m_y$, respectively, an estimator of the LFC is given as the mean difference after imputation:

$$\hat{\Delta}_T = \frac{1}{n_y}\left(\sum_{i=1}^{m_y} Y_i^* + \sum_{i=m_y+1}^{n_y} Y_i\right) - \frac{1}{n_x}\left(\sum_{i=1}^{m_x} X_i^* + \sum_{i=m_x+1}^{n_x} X_i\right)$$

The T-statistic for independent samples with unequal variances

$$T = \hat{\Delta}_T \Big/ \sqrt{\widehat{Var}\left(\hat{\Delta}_T\right)}$$

is compared to a T-distribution with Satterthwaite's approximate degrees of freedom (d.f.) to obtain a p-value.

Truncated Wilcoxon-test (tW): Hallstrom (2010) defined the truncated Wilcoxon rank–sum test as Wilcoxon's rank–sum test applied to the data after removing a maximum common number of PMVs from the two groups. Since we assumed $m_x \leq m_y$ this leaves X_i for $i = m_x + 1, \ldots, n_x$ and Y_i for $i = m_x + 1, \ldots, n_y$. The p-value is computed by the asymptotic Wilcoxon's rank–sum test (based on normal approximation) applied to the truncated data. An estimator of the LFC is calculated as

$$\hat{\Delta}_{tW} = median\{Y_j - X_i | i = m_x + 1, ..., n_x, j = m_x + 1, ..., n_y\}.$$

Moderated T-test (mT): As a standard approach that ignores PMVs we include the moderated T-test using the limma package in R (Smyth, 2005) with all PMVs set to missing. The estimator of the subdistribution LFC is equivalent to that of the two-part T-test (see below).

Tobit-model (Tob): Following Zhang *et al.* (2009) a parametric model based on a left-censored normal distribution is estimated using maximum likelihood (e.g. as implemented in the R-function survreg with option dist = 'gaussian'). The model is given by

$$Z_j = \beta_0 + I(j > n_x)\beta_1 + \varepsilon_j,\ j = 1, ..., n_x + n_y$$

where Z_j denotes the elements of the stacked vector $(X_1, ..., X_{n_x}, Y_1, ..., Y_{n_y})$ and $I(.)$ is the indicator function. ε_j follows the left-censored normal distribution $\max\{N(0, \sigma^2), \lambda - \beta_0 - I(j > n_x)\beta_1\}$ where λ denotes the $\log_2$-transformed LOD. The p-value results from a likelihood ratio test of $\beta_1 = 0$. The model returns a direct estimate $\hat{\Delta}_{Tob} = \hat{\beta}_1$ of the LFC, i.e. of the difference of the means corresponding to the underlying but left-censored normal distributions of the two samples.

2.2 Two-part tests

Following Taylor and Pollard (2009) two-part statistics are defined as the sum of two independent test statistics as first proposed by Lachenbruch (1976). The two parts correspond to the $m_x + m_y$ PMVs on the one hand and to the continuous parts X_i for $i = m_x + 1, \ldots, n_x$ and Y_i for $i = m_y + 1, \ldots, n_y$ on the other hand. The first test statistic is the continuity-corrected version of Pearson's Chi-square test statistic to compare the PMV proportions between the two groups. The second statistic for comparing the continuous parts either results from a T-test or from Wilcoxon's rank–sum test. In the first case the T-statistic for independent samples with unequal variances and in the second case a normalized and continuity-corrected version of Wilcoxon's U-statistic is employed (see Taylor and Pollard, 2009). Both, the test statistic for the binomial part and the square of the test statistic for the continuous part, asymptotically follow a chi-square distribution with 1 d.f. If both components of the two-part test statistic are well defined (i.e. the PMV proportion is not 0 and not 1 in both samples), then an asymptotic p-value can be derived from a chi-square distribution with 2 d.f. Otherwise, the undefined component is set to zero, and the resulting two-part statistic is assumed to follow a chi-square distribution with 1 d.f.

For the *Two-part T-test* (2T) and the *Two-part Wilcoxon-test* (2W) estimators of the subdistribution LFC are calculated as the difference of means and the Hodges–Lehmann estimator, respectively:

$$\hat{\Delta}_{2T} = \sum_{i=m_y+1}^{n_y} Y_i - \sum_{i=m_x+1}^{n_x} X_i$$

$$\hat{\Delta}_{2W} = median\{Y_j - X_i | i = m_x + 1, ..., n_x, j = m_y + 1, ..., n_y\}$$

For the ELRT the test statistic is derived from a likelihood ratio statistic by plugging in the empirical subdistributions of the two groups (see Taylor and Pollard, 2009). Following Lachenbruch

(1976) the likelihoods are constructed such that the binomial parts are explicitly taken into account. In this way the ELRT operates in the manner of a two-part test. No estimate of the subdistribution LFC is directly available for the ELRT due to its non-parametric character with respect to the continuous part. As for the two-part Wilcoxon test the Hodges–Lehmann estimator might be used.

2.3 Left-inflated mixture likelihood ratio test

Finally, we combine the Tobit and the two-part approach by constructing a likelihood that allows for BPMVs and, additionally, models the continuous part with a left-censored distribution. Such a model class is presented by Yang and Simpson (2010) as left-inflated mixture (LIM) models originally proposed in an econometric context. The same model was termed 'Bernoulli/Lognormal mixture model' by Moulton and Halsey (1995).

For the first group with observed values X_i for $i = 1, \ldots, n_x$ the density for the i^{th} observation is given as

$$f_{LIM}(X_i|\mu_x, \sigma_x, p_x; \lambda) = \begin{cases} p_x + (1-p_x)\Phi(\lambda|\mu_x, \sigma_x) & \text{for } X_i \text{ a PMV} \\ (1-p_x)\varphi(X_i|\mu_x, \sigma_x) & \text{otherwise} \end{cases}$$

where p_x denotes the proportion of BPMVs, μ_x and σ_x denote the mean and standard deviation of a (underlying but left-censored) normal subdistribution. $\Phi()$ and φ () denote the cumulative distribution and the density function, respectively, of the normal distribution. An analogous definition is assumed for the second group. Based on these densities we propose a LIM-likelihood ratio test (LIM-LRT) with $\mu_x = \mu_y$ and $p_x = p_y$ (but allowing $\sigma_x \neq \sigma_y$) under the null hypothesis. This test is a special case of the one proposed by Moulton and Halsey (1995). The detection limit λ is the same for both groups under the null as well as under the alternative hypothesis and is set to min $(X_1, ..., X_{n_x}, Y_1, ..., Y_{n_y}) - \varepsilon$ for a small ϵ. The log-transformed likelihood ratio is optimized numerically using the optim() function in R with option method = 'L-BFGS-B' which offers constrained optimization based on a quasi-Newton method. The starting values μ_x^0, μ_y^0, σ_x^0 and σ_y^0 (group-wise under the alternative and ignoring groups under the null hypothesis) can either be set by a Tobit model optimization or in a two-part manner by selecting median and 1.4826 MAD of the non-PMV observations. The starting values for p_x and p_y are then set to their respective maximum likelihood estimates $\hat{p}_x^0 = 1 - (n_x - m_x)/\left[n_x\left(1 - \Phi\left(\lambda|\mu_x^0, \sigma_x^0\right)\right)\right]$ and $\hat{p}_y^0$ analogously. We propose to start with the Tobit based pre-estimates and, in case of non-convergence, have a second try with the two-part pre-estimates. The LIM-LRT also delivers direct estimates for μ_x, μ_y, σ_x, σ_y, p_x and p_y resulting in the subdistribution LFC estimate $\hat{\Delta}_{LIM} = \hat{\mu}_y - \hat{\mu}_x$. The proportion of TPMVs is estimated by $(1 - \hat{p}_x)\Phi\left(\lambda|\hat{\mu}_x, \hat{\sigma}_x\right)$ and analogously for the second group.

Some of the test statistics mentioned in previous investigations are not considered in the present study. While the Kolmogorov-Smirnov test has been reported as being anticonservative (Lachenbruch, 2001), Zhang *et al.* (2009) showed that Gehan's test, the Log-rank and the Peto-Peto test are equivalent to Wilcoxon's rank–sum test in the case of identical LODs in the two groups and no ties in the continuous part. Further proposals of test statistics that employ other than log-normal distributions (log-gamma, log-skew-normal) have been treated in Taylor and Pollard (2009).

In all numerical calculations below we calculate two-sided tests using R 2.13.0 (R Development Core Team, 2011, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/).

3 Results

3.1 Simulations

In the following, we conduct a simulation study to evaluate the statistical methods presented above. Two concepts of generating PMVs are considered: (i) First, a *left-censored distribution* is considered, resulting from a continuous distribution (e.g. log-normal) of which all values below a given LOD are set to zero (or another appropriate PMV outside the range of the continuous part), see, e.g. Zhang *et al.* (2009). (ii) In contrast, one could generate the two parts as two independent components of a *mixture distribution*. A binomial random variable decides whether a signal is not present at all and thus set at the PMV, or this signal's value is independently drawn from some (fully observed) continuous distribution such as the log-normal or gamma (Taylor and Pollard, 2009). An interpretation of the compound null hypothesis for practical use is often difficult. However, simulations can be controlled more directly.

A priori, both concepts do not distinguish between biological and technical PMVs beforehand. However, concept (i) is obviously more naturally linked to TPMVs and concept (ii) to biological ones. Furthermore note that, by construction, the philosophy of using the one-part tests corresponds to concept (i), while the two-part tests naturally correspond to concept (ii). Finally, under concept (ii) consonant and dissonant group differences are equally plausible while under concept (i) dissonant differences are assumed to occur either by chance or as a consequence of a higher variance in the group with the higher location parameter.

In designing a simulation study, the choice of one of the two concepts for data generation will affect its results and conclusions. To our knowledge, the simulation studies presented in papers on newly proposed methods for the two-group problem in omics data are exclusively based on that concept which corresponds to the construction of the respective new test statistic. Neuhäuser *et al.* (2005) compared the two-part Wilcoxon test in two separate types of simulation scenarios, one corresponding to biological, the other to technical zeros but did not offer scenarios including both types at the same time. In practice it is hard to decide which of the two concepts best describes the distributions in a real dataset at hand. In the following we will therefore present a more general simulation setup which allows generating both PMV types simultaneously and covers both concepts as special cases.

We directly simulate on the binary logarithmic ($\log_2$) scale. Thus, results for the log-normal distribution are based on simulated normal distributions. The standard deviations of the underlying normally distributed variables are set to 1 in both groups and for all compounds, making moot the use of moderated test statistics in the simulation (cf. Tusher *et al.*, 2001; Smyth, 2005).

In each simulation run we simulate expression values of K compounds from $n_x + n_y$ samples. For ease of notation K is assumed to be an even number. The simulation parameters comprise λ, the $\log_2$-transformed LOD ≥ 0, the proportions p_x and p_y of BPMVs in the two groups and the subdistribution LFC δ.

For the first group (e.g. the control group) pairs (X_{ki}, C_{ki}^*) are generated where $X_{ki} = \max(U_{ki}, \lambda)$, $C_{ki}^* = I\ (U_{ki} < \lambda \text{ v } C_{ki} = 1)$ with $C_{ki} \sim \text{Bernoulli}(p_x)$ and $U_{ki} \sim N(0, 1)$ if $C_{ki} = 0$ for $k = 1, \ldots, K/2$, $i = 1, \ldots, n_x$ and with $C_{ki} \sim \text{Bernoulli}((p_x + p_y)/2)$ and $U_{ki} \sim N(0, 1)$ if $C_{ki} = 0$ for $k = K/2 + 1, \ldots, K$, $i = 1, \ldots, n_x$.

For the second group (e.g. the diseased group) we generate pairs (Y_{ki}, D_{ki}^*) where $Y_{ki} = \max(V_{ki}, \lambda)$, $D_{ki}^* = I\ (V_{ki} < \lambda \text{ v } D_{ki} = 1)$ with $D_{ki} \sim \text{Bernoulli}(p_y)$, $V_{ki} \sim N(\delta, 1)$ if $D_{ki} = 0$ for $k = 1, \ldots, K/2$, $i = 1, \ldots, n_y$, and with $D_{ki} \sim \text{Bernoulli}((p_x + p_y)/2)$, $V_{ki} \sim N(0, 1)$ if $D_{ki} = 0$ for $k = K/2 + 1, \ldots, K$, $i = 1, \ldots, n_y$.

All pairs with $C^*_{ki} = 1$ or $D^*_{ki} = 1$, respectively, are referred to as PMVs (the corresponding X_{ki} and Y_{ki}, respectively, are not interpreted then). To obtain the most stable estimates of true and false-positive rates half of the compounds exhibit a difference in the proportion of biological zeros and a subdistribution LFC by a value of δ while the remaining half is drawn from equal distributions in both groups.

Note that this general model reduces to concept (i) if $p_x = p_y = 0$ (excluding BPMVs) and to concept (ii) if $\lambda = -\infty$ (excluding TPMVs). All other cases allow for a mixture of concepts (i) and (ii) with biological as well as technical PMVs.

The simulation setup for $n_x = n_y = 25$, is defined by factorial combinations of shift $\delta = 0$, 1, 2, proportions p_x, $p_y = 0.0$, 0.2, 0.5 and $\lambda = -\infty$, -0.5, 0.0. For $n_x = n_y = 10$ only $\delta = 1$ was evaluated. Note that the δ and λ both refer to the $\log_2$-scale. The three chosen values for λ correspond to 0%, 31% and 50%, respectively, of TPMVs if the mean log expression is 0 and to 0%, 7% and 16%, respectively, of TPMVs if the mean log expression is 1. We simulate $K = 500$ independent compounds of which 250 are differentially expressed so that true and false-positive rates (FPR, TPR) can be estimated. There are 500 simulation runs for each scenario. The test statistics are employed as presented in the previous section including the choice of starting values for the LIM-LRT.

For the sake of a compact presentation of results we select eight representative scenarios (Table 1). There are three scenarios without TPMVs (B1 to B3 with decreasing proportion of consonant compounds), one scenario (P) without any mean shift between the groups, one scenario without BPMVs (T) and three mixed scenarios (M1 to M3 again with decreasing proportion of consonant compounds). Note that, compared to the B scenarios, there are fewer BPMVs in the M scenarios. However, the finite λ produces TPMVs in both groups, thus resulting in a lower proportion of dissonant compounds. Scenario T is the prototypic constellation for the Tobit model.

The results for the eight selected scenarios are presented in Figure 2 and Supplementary Figures S1A and S1B depicting TPRs for a FPR of 0.005, as well as the median and the inter-quartile range of the bias in the estimate of the (subdistribution) LFC. The detailed results for all scenarios are summarized in Supplementary Table S1.

Figure 2 demonstrates the strong dependence of TPR on the underlying data-generating model. In general, power is highest with concordant proportions of BPMVs and TPMVs (scenario M1), and lowest when there are differences only in BPMVs (P) or with discordant TPMVs and BPMVs (M3). The investigated tests cluster into two distinct shapes of TPR profiles, one typical for two-part tests, and one for one-part tests. The moderated T-test ignoring PMVs does not fit either type. Naturally, it outperforms all other tests if the proportion of PMVs is equal between groups (B2 and M3), but yields severely lower TPRs in all other scenarios. Generally, the two-part tests, including the ELRT, are more robust to dissonant constellations (B2, B3; cf. also Lachenbruch, 2001), but this advantage is lost with the occurrence of TPMVs (M2, M3). Still, two-part tests have adequate TPRs in most scenarios and generally give the most stable estimates of the subdistribution LFC. By construction, the bias of their estimate depends on the amount of TPMVs and thus on λ and δ. For the two-part T-test the average amount of bias is $\varphi(\alpha_y)/(1-\Phi(\alpha_y)) - \varphi(\alpha_x)/(1-\Phi(\alpha_x))$ where $\alpha_x = (\lambda - \mu_x)/\sigma_x$ with μ_x and σ_x as the parameters of the (left-censored) normal subdistribution and analogously for α_y. In scenarios T and M1 to M3 we have $\lambda = -0.5$, $\mu_x = 0$, $\mu_y = \delta = 1$ and $\sigma_x = \sigma_y = 1$ such that the average bias is -0.37.

In the absence of TPMVs the estimates of the subdistribution LFC associated with the two-part tests are unbiased. If linked with a Hodges–Lehmann estimator the same holds for the ELRT. The truncated version of Wilcoxon's rank–sum test shows little improvement over the original version, and only for dissonant cases. For the one-part tests excluding the moderated T-test the bias in the estimate of the subdistribution LFC is minimal and least variable for equal proportions of BPMVs (senarios B2, T and M2), since the corresponding subdistribution LFC estimates are essentially full-distribution estimates but negative effects level out if BPMVs are equally likely in both groups.

Similar results are observed at other FPRs, as demonstrated for scenario B2 in Supplementary Figure S2 where the TPRs run more or less in parallel when the FPR varies. Supplementary Table S1 shows that the conclusions drawn above also hold for $\delta = 2$ and for $(n_x, n_y) = (10, 10)$.

For the LIM-LRT we finally compare estimated proportions of biological and technical PMVs for the differentially expressed peptides with each scenario's simulation parameters (see Supplementary Fig. S3). For the second (shifted) group the median bias is below 5 % points in all scenarios. For the first (control) group the median bias in the P and B scenarios is only slightly larger. However, in the T and M scenarios the proportions of TPMVs in this group are considerably underestimated and, correspondingly, the proportions of BPMVs are overestimated. Thus in the control group, scenarios involving TPMVs appear 'more biological' than they are due to a lower location parameter which induces more values below the LOD.

Table 1. Overview of eight representative scenarios ($n_x = n_y = 25$)

Scenario	δ	p_x	p_y	λ	% Cons.	% Equal prop.	% Diss.
B1	1	0.5	0.2	$-\infty$	98.0	0.9	1.1
B2	1	0.2	0.2	$-\infty$	43.1	14.1	42.8
B3	1	0.2	0.5	$-\infty$	1.1	0.9	98.0
P	0	0.2	0.5	$-\infty$	49.7	0.9	49.4
T	1	0.0	0.0	-0.5	97.8	1.0	1.1
M1	1	0.2	0.0	-0.5	99.0	0.1	0.9
M2	1	0.2	0.2	-0.5	89.4	4.2	6.4
M3	1	0.0	0.2	-0.5	60.6	11.4	28.1

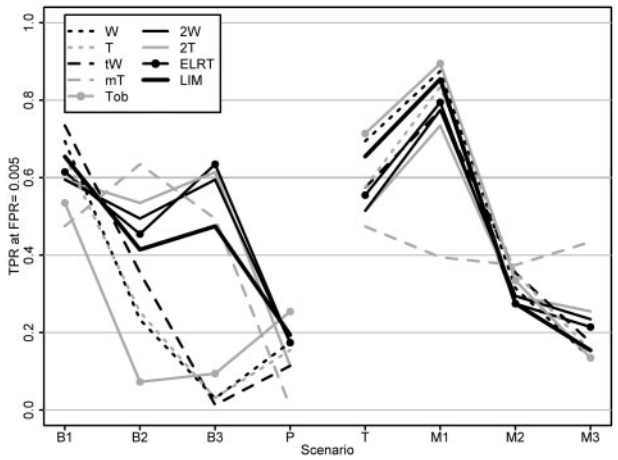

Fig. 2. True-positive rates (TPR) at false positive rate (FPR) = 0.005 for the scenarios described in Table 1. B1 to B3: only BPMVs (consonant to dissonant), P: only BPMVs and $\delta = 0$, T: only TPMVs, M1 to M3: mixed (consonant to dissonant)

To sum up, if there is no prior knowledge about the presence or absence of technical and biological PMVs then the two-part Wilcoxon test and the ELRT seem to be good compromises. The advantage over the two-part *T*-test and the LIM-LRT is their independence from distributional assumptions for the continuous subdistribution. The two-part Wilcoxon test and the ELRT are the most robust tests in dissonant situations and show good performance also in the other scenarios. The bias of the Hodges–Lehmann estimates of the subdistribution LFC is generally low. The LIM-LRT generally has low bias in all scenarios and has higher TPRs in technical and consonant mixed scenarios than the two-part Wilcoxon test. If the assumption of a (left-censored) log-normal subdistribution is reasonable and some technical PMVs cannot be excluded then the LIM-LRT might be preferred.

3.2 Application

In this section the previously introduced and evaluated statistical tests are applied to the motivating dataset presented in the introduction. For this purpose we first restrict the dataset to peptides with a PMV proportion below 70% ignoring group labels. The restricted dataset contains 787 peptides with a median PMV proportion at 52% (quartiles 32% to 62%) including 408 (51.8%) consonant and 322 (40.9%) dissonant peptides and 57 (7.3%) with equal PMV proportions. Table 2 shows the pair-wise correlations between the (uncorrected) *p*-values of the discussed tests. Spearman correlation coefficients of at least 0.80 are observed within the group of one-part tests with a few exceptions as well as within the group of all two-part tests and the LIM-LRT.

The following method for controlling the False Discovery Rate (FDR) is used to adjust for multiple testing among the 1093 peptides: for various thresholds α we estimated false discovery rates (Benjamini and Hochberg, 1995) using

$$\widehat{FDR}(\alpha) = \frac{\hat{F}_{p0}(\alpha)\hat{\pi}_0}{\hat{F}_p(\alpha)}$$

where $\hat{F}_p(\bullet)$ and $\hat{F}_{p0}(\bullet)$ are the empirical cumulative distribution functions of all *p*-values across all peptides in the restricted original dataset and across all peptides and all permutations, respectively, and $\hat{\pi}_0$ is the estimated proportion of genes for which the null hypothesis applies. This procedure resembles the one outlined in Kerr (2009) except of using $\hat{F}_{p0}(\alpha)$ instead of α in the numerator. Finally, $\hat{\pi}_0$ was estimated by $\left[\hat{F}_p\left(\hat{F}_{p0}^{-1}(0.75)\right) - \hat{F}_p\left(\hat{F}_{p0}^{-1}(0.25)\right)\right]/0.5$. *q*-values based on these estimates of the FDR are given.

Table 3 presents the numbers of compounds declared significant based on a FDR of 0.05. As expected from our simulations the two-part tests and the LIM-LRT show a higher proportion of dissonant peptides among those declared significant than the other tests. As seen from the kappa coefficients in Table 2 there is a lower overlap of significant compounds between the LIM-LRT and the group of two-part tests than within the two-part tests.

To assess whether technical or biological PMVs play a larger role in the set of 787 peptides we use the PMV proportion estimates from the LIM models. The median number of technical and biological PMVs is 0.1 (quartiles <0.1, 0.5) and 12.9 (quartiles 8.6, 16.6), respectively, in the CKD group and 0.3 (quartiles <0.1, 1.0) and 9.4 (quartiles 4.1, 13.4), respectively, in the control group. Across all 787 peptides, the median of the proportion of biological among all PMVs within a given peptide is 98.9% (quartiles 95.6%, 99.9%) in the CKD and 95.7% (quartiles 84.9%, 98.9%) in the control group. However, we have learned from Supplementary Figure S3 that the proportion of BPMVs might be overestimated in the respective group with the lower location parameter.

In addition to the prevailing, but not exclusive biological nature of the PMVs in this dataset the continuous parts largely exhibit an approximate (left-censored) log-normal distribution (graphical inspection, results not shown). Thus, the LIM-LRT and the two-part tests (including the ELRT) seem most appropriate for analysis.

The detailed results for the two example peptides presented in Figure 1 are shown in Table 4. There are striking discrepancies between the two-part tests and the LIM-LRT on the one hand and the one-part tests on the other. For the consonant peptide the Tobit

Table 3. Results of all tests for example dataset (significance defined as: *q*-value ≤ 0.05)

	Test	No. significant compounds				Total rel. to 2 × 100	
		Cons.	Equal	Diss.	Total	Sens. (%)	Spec. (%)
One-part tests	W	178	4	38	220	40	87
	T	209	4	88	301	51	83
	tW	185	2	26	213	39	88
	mT	61	6	17	84	25	88
	Tob	148	0	55	203	38	88
Two-part tests	2W	158	10	78	246	35	85
	2T	145	9	67	221	31	84
	ELRT	152	4	83	239	34	84
	LIM	126	3	62	191	29	85
Total		408	57	322			

Sensitivity (Specificity): percent of peptides declared significant (non-significant) in 2 × 100 samples that are also significant (non-significant) in the presented 2 × 25 samples (see text)

Table 2. Upper triangle: pair-wise Spearman correlation coefficients between *p*-values (bold: ≥0.80); lower triangle (italic type): kappa coefficients with respect to significance defined as FDR ≤ 0.05 (bold: ≥0.70)

	W	T	tW	mT	Tob	2W	2T	ELRT	LIM
W		**0.88**	**0.92**	0.22	**0.88**	0.75	0.70	0.73	0.65
T	***0.72***		0.77	0.14	**0.88**	0.68	0.64	0.69	0.63
tW	***0.79***	*0.57*		0.36	0.74	0.78	0.75	0.75	0.65
mT	*0.20*	*0.09*	*0.32*		0.06	0.59	0.67	0.60	0.49
Tob	***0.77***	***0.72***	*0.61*	*0.03*		0.65	0.62	0.70	0.62
2W	*0.67*	*0.52*	*0.69*	*0.38*	*0.58*		**0.96**	**0.94**	**0.81**
2T	*0.59*	*0.44*	*0.65*	*0.48*	*0.50*	***0.83***		**0.97**	**0.82**
ELRT	*0.60*	*0.49*	*0.64*	*0.38*	*0.56*	***0.84***	***0.86***		**0.86**
LIM	*0.57*	*0.44*	*0.57*	*0.37*	*0.53*	***0.71***	***0.74***	***0.76***	

Table 4. *p*-values, *q*-values and (subdistribution) log fold change (LFC) for example peptides (see introduction section)

	Test	Peptide 123 750 (cons.)			Peptide 100 344 (diss.)		
		p	*q*	LFC	*p*	*q*	LFC
One-part tests	W	0.012	0.030	2.21	0.281	0.262	0.00
	T	0.029	0.040	3.06	0.204	0.167	−1.62
	tW	<0.001	0.004	3.75	0.419	0.433	−0.87
	mT	<0.001	0.006	2.39	0.081	0.214	1.15
	Tob	0.049	0.090	4.89	0.139	0.180	−3.83
Two-part tests	2W	0.003	0.006	2.66	0.100	0.096	1.28
	2T	<0.001	0.005	2.39	0.046	0.073	1.15
	ELRT	<0.001	0.012	2.66	0.028	0.063	1.28
	LIM	<0.001	0.009	2.90	0.007	0.029	2.30

Table 5. Differences of (subdistribution) LFC estimates (row minus column). Upper triangle: medians; lower triangle (italic): interquartile ranges

	W	*T*	tW	mT	Tob	2W	2T	ELRT	LIM
W		−1.0	0.0	−0.5	0.9	−0.5	−0.5	−0.5	−0.6
T	*3.8*		1.3	0.7	2.6	0.7	0.7	0.7	0.7
Tw	*2.0*	*7.1*		−0.5	0.7	−0.5	−0.5	−0.5	−0.6
mT	*1.4*	*4.1*	*2.2*		2.0	0.0	0.0	0.0	0.0
Tob	*4.4*	*8.9*	*2.2*	*4.7*		−2.0	−2.0	−2.0	−2.0
2W	*1.3*	*4.1*	*2.1*	*0.2*	*4.7*		0.0	0.0	−0.1
2T	*1.4*	*4.0*	*2.2*	*0.0*	*4.6*	*0.2*		0.0	0.0
ELRT	*1.3*	*4.1*	*2.1*	*0.2*	*4.7*	*0.0*	*0.2*		−0.1
LIM	*1.7*	*4.5*	*2.2*	*0.2*	*4.6*	*0.4*	*0.2*	*0.4*	

model shows the only q-value above 0.05. Except for the one-part Wilcoxon test the one-part tests give higher estimates of the (subdistribution) LFC than the two-part tests. For the dissonant peptide the results are highly inconsistent, most of the estimates associated with one-part tests even give opposite signs due to the high PMV proportion in the CKD group. For the dissonant case we clearly see that the two-part tests are aiming for measuring different features of the data than the remaining tests do.

The LIM-LRT tries to differentiate between biological and technical PMVs. In the CKD group the 10 PMVs for peptide 123 750 and the 17 PMVs for peptide 100 344 are all estimated biological. However, in the control group the 15 PMVs for peptide 123 750 are divided into 13 biological and 2 technical PMVs and the 11 PMVs for peptide 100 344 into 5 biological and 6 technical ones. This results in lower location parameter estimates in the control group and explains the higher estimate of the subdistribution LFC for both peptides compared to the two-part tests. Table 5 shows that the agreement between (subdistribution) LFC estimates is generally high among the two-part tests whereas there are large discrepancies with and among one-part tests (by construction the estimates of mT and 2T and those of 2W and ELRT are identical). The results for the top 25 peptides selected by the LIM-LRT are summarized in Supplementary Figure S4. It can be seen that the proportion of TPMVs among the selected peptides is disproportionately high, 21 of the top 25 peptides exhibit TPMVs in at least one group. Furthermore, 10 of these peptides are dissonant.

Finally, we re-examined the same set of peptides with the same methods in larger groups of 100 CKD cases and 100 controls which contains the 2×25 samples investigated so far as a subsample. As seen in the last two columns of Table 3 the internal consistency in terms of sensitivity relative to the larger groups is generally higher for one-part tests and is smallest for the LIM-LRT. This pattern is quite similar in three other randomly selected 2×25 samples (results not shown). Specificity is generally around 85%.

4 Discussion

This paper tries to fill the gap between proposals for handling PMVs in the context of omics data and those for general or other specific types of data showing a zero-inflation phenomenon. We applied several test statistics in an extensive simulation study and to a real dataset. We observed large differences between the results from one-part and two-part tests. This is a direct consequence of their respective construction.

In the simulation study we generated data simultaneously exhibiting biological and technical PMVs since it is plausible that a (known or unknown) underlying mechanism controls the binomial part independently from the continuous part and from the mechanism producing the TPMVs. In contrast, the 'dependency model' by Hallstrom (2010) tends to produce only consonant distributions despite two independent sources of PMVs.

It is clear from our investigations that the two considered types of PMVs have different implications for the operation characteristics of the considered tests. All tests of the one-part or the two-part types explicitly serve only one type of PMV but may be more or less robust to the other. Our results show that one-part tests give acceptable results for equal proportions of BPMVs. Generally, two-part tests show higher TPRs and the corresponding estimates have low bias depending on the proportion of TPMVs. For a particular data analysis it may often be difficult to judge whether the observed PMVs are technical or biological. The LIM-LRT offers a way to circumvent a decision between one-part and two-part tests by allowing for and explicitly estimating the proportion of both types of PMVs.

The interpretation of the measures of location corresponding to one-part and two-part tests is quite different. For one-part tests location measures, and differences between groups calculated from them, are based on the full distribution of values, including the unobserved signals, which are generally assumed as being left-censored. By contrast, two-part tests supply two separate estimates of group difference, one for the PMV proportion and one corresponding to the difference between the group-specific subdistributions of continuous values. In estimating these group differences, none of the two-part tests distinguishes between technical and biological PMVs unlike the LIM-LRT. By making a distributional assumption on the continuous subdistribution, the LIM-LRT can provide such a distinction, which can be useful in situations where both technical and biological PMVs occur. The group difference in the PMV proportions estimated by the LIM-LRT only incorporates those PMVs that have been classified as biological.

We showed that the distinction between consonant and dissonant compounds has a large impact on test performance. It is, however, unclear if investigators are at all interested in detecting dissonant compounds by rejecting a compound two-part null hypothesis. As described in the introduction dissonant compounds of scientific interest could arise if several compounds are alternatively activated by the same biological phenomenon. This aspect, however, is more appropriately covered by testing whole sets of compounds instead of a single compound at a time. Further investigations are needed to see whether the likelihood ratio nature of the LIM-LRT allows for a useful extension along those lines.

This study did not consider other continuous distributions than the normal distribution (on the $\log_2$ scale). Extensions in this direction have been investigated by Taylor and Pollard (2009) for two-part tests and by Zhang *et al.* (2009) for the Tobit model. These authors observed that parametric test statistics were fairly robust to other distributions. We also did not consider different variability in the two groups and across compounds and set all variances in the simulations equal to one. Neither did we vary the LOD across compounds or samples. In analogy to tW we also investigated a truncated version of the *T*-test (see Supplement) with a performance close to that of tW (results not shown).

In conclusion, the two-part Wilcoxon test and the ELRT show good overall properties without any distributional assumptions for the continuous part. They are recommendable if technical PMVs can be ruled out beforehand. If the assumption of a (left-censored) log-normal subdistribution is reasonable and technical PMVs are possible then the LIM-LRT is an interesting alternative.

Funding

This work was partly supported by the EU's Seventh Framework Programme for Research (FP7), grant HEALTH-F2-2009-241544.

Conflict of Interest: none declared.

References

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Coon,J.J. *et al.* (2008) CE-MS analysis of the human urinary proteome for biomarker discovery and disease diagnostics. *Proteom. Clin. Appl.*, **2**, 964–973.

Dakna,M. *et al.* (2010): Addressing the challenge of defining valid proteomic biomarkers and classifiers. *BMC Bioinformatics*, **11**, 594–609.

Hallstrom,A.P. (2010) A modified Wilcoxon test for non-negative distributions with a clump of zeros. *Stat. Med.*, **29**, 391–400.

Helsel,D.R. (2012) *Statistics for Censored Environmental Data Using Minitab® and R*. 2nd edn. Wiley, Hoboken.

Kerr,K.F. (2009) Comments on the analysis of unbalanced microarray data. *Bioinformatics*, **25**, 2035–2041.

Lachenbruch,P.A. (1976) Analysis of data with clumping at zero. *Biometr. Z.*, **18**, 351–356.

Lachenbruch,P.A. (2001) Comparison of two-part models with competitors. *Stat. Med.*, **20**, 1215–1234.

Moulton,L.U. and Halsey,N.A. (1995) A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*, **51**, 1570–1578.

Neuhäuser,M. *et al.* (2005) Two-part permutation tests for DNA methylation and microarray data. *BMC Bioinformatics*, **6**, 35–41.

Siwy,J. *et al.* (2011) Human urinary peptide database for multiple disease biomarker discovery. *Proteom. Clin. Appl.*, **5**, 367–374.

Smyth,G.K. (2005) Limma: linear models for microarray data. In: Gentleman,R. *et al.* (eds.) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, pp. 397–420.

Taylor,S. and Pollard,K. (2009) Hypothesis tests for point-mass mixture data with application to´omics data with many zero values. *Stat. Appl. Genet. Mo. B.*, **8**, 1–43.

Tusher,V. *et al.* (2001) Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natl. Acad. Sci. USA*, **98**, 5116–121.

Yang,Y. and Simpson,D.G. (2010) Conditional decomposition diagnostics for regression analysis of zero-inflated and left-censored data. *Stat.Methods Med. Res.*, **21**, 393–408.

Zhang, D. *et al.* (2009) Nonparametric methods for measurements below detection limit. *Stat. Med.*, **28**, 700–715.