JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

ORIGINAL ARTICLE | DOI: 10.5584/jiomics.v2i2.109

A robust permutation test for quantitative SILAC proteomics experiments

Hien D. Nguyen[1,2], Ian A. Wood[2], Michelle M. Hill*[1].

[1]The University of Queensland Diamantina Institute, The University of Queensland, QLD 4102 Australia. [2]School of Mathematics and Physics, University of Queensland, St. Lucia, QLD 4072 Australia.

Received: 06 September 2012 **Accepted:** 02 October 2012 **Available Online**: 14 December 2012

ABSTRACT

Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) along with other relative quantitation methods in proteomics have become important tools in the analysis of cellular and subcelluar functions. Although numerous experimental applications of SILAC have been developed, there is no consensus on the use of statistical procedures to analyze the resulting experimental data. SILAC experiments output relative abundance ratios for proteins to quantify differences in cell populations. These ratios have traditionally been analyzed with fold-change methods and hypothesis testing procedures under Gaussian distribution assumptions.

We find that the normality assumption is invalid and can lead to inaccurate quantitation of the significance of differences between cell populations. As a solution, a permutation based hypothesis test as an alternative for assessing significance is introduced. We develop a distribution-free permutation testing methods for assessing various SILAC experiments. These tests generate p-values which can be easily interpreted and if necessary, the false discovery rate of these p-values can be easily controlled. To compare the permutation test against competing methodology, we used a set of simulations based upon a theoretical model of SILAC ratio data.

Through the simulation studies, we find that the permutation test is generally superior to the competing hypothesis tests across the range of simulation scenarios. We also find that the permutation test is typically more powerful and accurate than the competing methods at the five percent level of significance and averaged over the spectrum of significance levels. Because of the broad superiority of the permutation test and the ease of implementation, we propose the use of the permutation test as a standard measure of protein significance in SILAC experiments.

Keywords: SILAC; quantitative proteomics; robust statistics; permutation test; Gaussian mixture model.

1. Introduction

Mass spectrometry-based proteomics methods have become widely used and highly successful tools for the large-scale study of molecular and cellular biology in recent years. Since mass spectrometry analysis for individual samples are performed sequentially, quantitative comparisons are not inherent in these experiments. Over the past decade, several labeling methodologies have been developed to enable quantitative proteomics analysis using multiplex mass spectrometry. The general approach is to label specific amino acids of different samples with different mass 'tags', allowing the samples to be mixed for the mass spectrometry analysis, providing relative quantitation in terms of a ratio between the samples. One method popular with cell biologists was introduced by the laboratory of Matthias Mann in 2002 [1], termed Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC), and involves the incorporation of isotopic amino acids as the cells produce new proteins during their normal growth. Since proteins are labelled during normal cell metabolism, rather than post-preparation of the proteins of interest, as is used in other quantitative proteomics strategies, SILAC offers the potential to account for variations during protein separation, and is particularly suited for subcellular proteomes [2–4] and time-resolved proteomics [2,5]. Although originally developed for cultured mammalian cells, SILAC has now been applied to other species and organisms, including bacteria [6], fly [7,8], plant [9] and yeast [10]. In

***Corresponding author:** Michelle M Hill. The University of Queensland Diamantina Institute, Level 4 R-Wing, Princess Alexandra Hospital, Ipswich Road, Woolloongabba QLD 4102 Australia. Telephone: +61 7 3176 7456; Email Address: m.hill2@uq.edu.au

addition, the development of SUPER-SILAC methodology has enabled quantitative analysis with non-metabolic samples, such as archival tissue specimen and plasma [11]. A thorough description of the SILAC method can be found in [12].

In a typical two-plex SILAC experiment, samples are labelled in with 'light' and 'heavy' isotopes by feeding the cell or organism with amino acids containing the required isotopes. Lysates are prepared and then combined based on equal protein content. The target cellular or subcellular proteome are isolated, digested with trypsin, and the peptides are analyzed by tandem mass spectrometry (MS/MS) following one or more separation steps, generally by liquid chromatography (LC). Because of the predictability of the mass shift between the peptides coming from the "light" and "heavy" populations, it is possible to distinguish "heavy" and "light" peptide from the survey mass spectra (MS1), and hence calculate a ratio of their relative intensities. A ratio equal to one would indicate that there are no differences in the abundance of the peptide between the two populations whereas a ratio not equal to one would indicate an up or down-regulation of the peptide.

Furthermore, fragmentation of the parent ion (MS/MS spectra) allows the matching of the quantified peptides to a protein sequence, and the inference of a protein's abundance change by averaging over the matched peptides. Protein inference from MS/MS data is generally performed by specialist software, both commercial and open source. Comprehensive comparisons of these algorithms have recently been performed [13,14]. The current work will focus on the statistical evaluation of such quantitative data.

In a typical SILAC experiment, there may be thousands of protein ratios measured [3,8,12]. Because of the large number of protein ratios quantified, and the existence of errors in the computational and experimental processes, it is highly unlikely that all proteins with ratios different to one are actually differently regulated between the two populations. It is therefore necessary to define a method by which proteins can be considered to be differently regulated or not.

A common approach to assessing protein ratios is through the use of a "fold-change" threshold applied over all protein ratios [1,15]. Under a 1.5 threshold scheme as suggested in [15], a protein is considered differently abundant between the whole populations if the protein ratio is either greater than 1.5 or less than the inverse of 1.5 (2/3). Depending on the experiment, there have been reports of uses of thresholds anywhere between 1.3 [12] to 6 [1].

Because of the differences in variability between experiments, an application of a certain threshold to one experiment may result in a high proportion of false positives whereas the same threshold applied to a different experiment might yield a high proportion of false negatives instead. The combination of variability in results, combined with a need for experimenter specification makes the threshold method less useful for high throughput applications where there is a greater priority to mitigate the proportion of false discoveries in an experiment [16,17].

For candidate proteins that will be further validated through orthogonal approaches, such as western blotting, strictly controlling the false discovery rate may not be crucial. However, one of the advantages of omics analyses such as proteomics is the ability to infer pathways and networks from the high throughput data [18,19]. For these workflows, thorough statistical hypothesis testing is crucial in ensuring the true list of altered proteins is carried forward for the network or pathways analyses [20]. There have been a wide variety of hypothesis testing methodologies applied to the assessment of abundance by SILAC ratios such as the one sample t -test [21,22], z-test [23,24], robust z-test (significance A) [17,25,26] and Wilcoxon signed-rank test [22]. The reason for the numerous approaches to hypothesis testing is because of the disagreement about what statistical model and assumption to use for the calculation of p-values. Methods such as the t-test and z-test are based on an assumption that either the protein ratio distribution or peptide ratio is normally distributed.

In this study, we examined the distribution of two-plex SILAC datasets, and demonstrate the assumption violations and deficiencies of the t-test, z-test, robust z-test and Wilcoxon signed-rank test when used with SILAC data. We then propose a new permutation test which does not rely on distributional assumption and demonstrate its utility and superiority over the current methodologies in a simulation study.

2. Material and Methods

Peptide ratios from a SILAC quantitative proteomics experiment take on numbers in the theoretical, noninclusive range of zero to infinity represented by r_i where i is an enumeration of the ratios, taking values one to n. Without loss of generality, we will always consider the peptide ratios as being in the form of heavy over light. The number n denotes the total number of peptide ratios obtained in the experiment. After protein mapping, it is possible to assign each peptide ratio a protein identification number p_i between one and m where m is the total number of proteins identifiable by the peptides quantitated in the experiment. The number p_i is an enumeration of the identifiable proteins in some order such as by the protein's accession number. Together the values r_i and p_i allow us to identify each of the peptide ratios and to which protein they are matched.

Although the distribution of the peptide ratios $\{r\}_{i=1}^{n}$ is bounded to the left by zero, it has often been observed that the distribution of the logarithm of r_i: $l_i=\log(r_i)$, usually takes on a bell-shaped curve [24,27,28]. Under a 1:1 mix between heavy and light cells sampled from the same populations, we would expect the distribution to be centered at zero because the logarithm of one is zero and the abundances of proteins between the two populations should theoretically be the same [27]. Because of these properties, it is often more convenient and meaningful to work with the peptide log-ratios l_i instead of r_i.

2.1. *Experimental Dataset*

For the analysis of the distributions of peptide ratios, we utilise the data set from a published subcellular SILAC study comparing four fractions from prostate cancer PC-3 cells expressing the protein Polymerase I and Transcript Release Factor (PTRF) tagged with Green Fluorescent Protein (GFP), or GFP alone as control [3].

2.2. *One sample t-tests*

The one sample t-test for each protein j, is evaluated by first evaluating the mean over all peptide log-ratios matched to the protein numbered j: $L_j = \sum_{\{l_i \mid p_i = j\}} l_i / n_j$ where $n_j = \#\{l_i \mid p_i = j\}$ is the number of peptides matched to protein j. The protein log-ratios L_j are then divided by the standard error term $s_j / \sqrt{n_j}$ where $s_j = \sqrt{\sum_{\{l_i \mid p_i = j\}} (L_j - 1)^2 / (n_j - 1)}$ is the standard deviation of the peptide log-ratio subsample $\{l_i \mid p_i = j\}$ which yields the test statistic

$$T_j = \frac{L_j}{s_j / \sqrt{n_j}}.$$

Under the null hypothesis that there is no change in abundance of the protein between the two populations and under the assumption that the subsample $\{l_i \mid p_i = j\}$ is normal, the p-value of T_j can be found using the Student-t distribution [29]. It must also be noted that the t-test can only be applied to proteins that are quantitated by two or more peptides because of the necessity to estimate the standard deviation of the log-ratio subsamples.

2.3. *Robust z-tests*

To conduct a robust z-test, each of the proteins j are quantitated by taking the median over all peptide log-ratios matched to the protein numbered j: $M_j = Median(\{l_i \mid p_i = j\})$. An overall median of the protein log-ratios sample : $\{M_j\}_{j=1}^{m} : P_{50}$ alongside with the 15.87 and 84.13 percentiles: $P_{15.87}$ and $P_{84.13}$ can also be calculated. Assuming that the population of protein log-ratios has a Gaussian distribution, the differences P_{50} -$P_{15.87}$ and $P_{84.13}$ - P_{50} are both approximately equal to one standard deviation, and P_{50} is equal to the mean. Using these facts, the z-statistic

$$Z_j = \frac{(M_j - P_{50})}{(P_{84.13} - P_{15.87}) / 2}$$

has a standard Gaussian distribution and can be used to evaluate the p-value of each of the proteins.

As an alternative, [25] specifies the test statistic

$$Z_j = \frac{M_j - P_{50}}{P_{84.13} - P_{50}}$$

for testing the null hypothesis of no difference in abundance versus the alternative hypothesis of greater abundance in the heavy population compared to the light population.

2.4. *Wilcoxon signed-rank tests*

Another robust hypothesis testing method is the Wilcoxon signed-rank test [30]. For proteins, the Wilcoxon signed-rank test is used to test the hypothesis that the abundance is the same between the two cell populations against the hypothesis that the abundance being different. The p-value of the test is evaluated using the signed ranked peptide log-ratio subsample $\{l_i \mid p_i = j\}$ for each protein j through the Wilcoxon rank distribution or a normal approximation.

2.5. *Permutation tests*

We now present a permutation test procedure formulated in the tradition of [31] whereby we resample with replacement from the empirical data in order to evaluate the significance of a test statistic. In the context of bioinformatics, permutation testing has been applied to microarray gene expression studies [32,33], gene ontology and network analyses [34,35], and in proteomics specifically, permutation testing has been applied to false discovery rate control [36,37] and to biomarker discovery [38,39]. In the case of a SILAC protein expression experiment, we want to know the significance of the difference in the abundance of a protein between the heavy and light cell populations. The difference in abundance for a protein j can be measured as some function of the peptide log-ratio subsample $\{l_i \mid p_i = j\}$ such as the mean L_j or the median M_j. As an example, we will only consider L_j in following description of the method.

Under the assumption that the true population distribution of peptide log-ratios for the experiment resembles the sample distribution of the peptide log-ratios [40], we can calculate conservative estimates for the p-values of each protein j using the algorithm in (Figure 1) [41,42] for the null hypothesis that the protein is differently regulated between the two cell populations. Although there are no direct permutation of the protein labels p_i, the random sampling without replacement serves the practical purpose of permuting the class labels by obtaining n_j peptide log-ratios to label as protein j. The algorithm allows for working with data that may or may not need normalization such as 1:1 mixes and other mixing proportions respectively as well as allowing the replacement of the mean protein log-ratio L_j for any other test statistic of interest such as the median M_j.

It is important to note that, unlike the other hypothesis testing procedures, a $\hat{\text{p-value}}_j$ obtained from (Figure 1) is an estimate of the true p-value for protein j with respect to the null hypothesis. In the case of a protein with n_j of one, the estimate $\hat{\text{p-value}}_j$ is the same each time the algorithm is run since the test computes the exact probability of obtaining the test statistic of L_j or more extreme under the null hypothesis in a similar fashion to Fisher's exact test [43]. In the cases where the proteins have subsample sizes n_j greater than one, we cannot calculate the exact p-value since it would require the exhaustive calculation of test statistics for all $n!/((n-n_j)!n_j!)$ possible subsamples from $\{l_i\}_{i=1}^{n}$ of size . This number grows rapidly and can become computationally infeasible for the sizes of most SILAC experiments. Instead of exhaustive computations over all possible subsamples, we

1. Set S as the sample of all peptide log-ratios $\{l_i\}_{i=1}^n$ from the experiment if the data does not require normalization.

a. If the data requires normalization, set S as the normalized sample $\{l_i - M\}_{i=1}^n$ or $\{l_i - \overline{l}\}_{i=1}^n$ where $M = median\left(\{l_i\}_{i=1}^n\right)$ and $\overline{l} = \sum_{i=1}^n \frac{l_i}{n}$ to normalize by the median or the mean respectively.

2. For protein j in the number of proteins one to m:

a. Calculate the number of peptides $n_j = \#\{l_i | p_i = j\}$ in the subsmaple of protein j.

b. Calculate the protein log-ratio $L_j = \sum_{\{l_i | p_i = 1\}} \frac{l_i}{n_j}$ for protein j.

c. If $n_j = 1$:

i. Estimate the p-value of protein j by evaluating $\hat{\text{p-value}}_j = \frac{\sum_{i=1}^n I\{|L_j| > |l_j|\} + 1}{n+1}$ where $I\{|L_j| > |l_j|\} = 1$ if $|L_j| > |l_j|$ is true and $I\{|L_j| > |l_j|\} = 0$ if $|L_j| > |l_j|$ is false.

d. If $n_j > 1$:

i. For b from one to B:

A. Randomly sample without replacement n_j numbers between one and n to obtain the set $\{k_a\}_{a=1}^{n_j}$ where k_a is a number between one and n.

B. For the random peptide log-ratio subsample $S_b = \{k_a\}_{a=1}^{n_j}$ using the random sample $\{k_a\}_{a=1}^{n_j}$.

C. Compute the protein log-ratio of the random subsample by evaluating $L'_b = \sum_{a=1}^{n_j} \frac{l_{k_a}}{n_j}$.

ii. Estimate the p-value by evaluating $\hat{\text{p-value}}_j = \frac{\sum_{b=1}^B I\{|L_j| > |L'_b|\} + 1}{B+1}$ where $I\{|L_j| > |L'_b|\} = 1$ if $|L_j| > |L'_b|$ is true and $I\{|L_j| > |L'_b|\} = 0$ if is $|L_j| > |L'_b|$ false.

Figure 1. *Permutation testing p-value.* Algorithm for assessing the significance of protein log-ratios in SILAC expression experiments.

sample B of these subsamples with replacement and estimate the p-value based on these subsample test statistics instead. It is known that with increasing values of B the estimates $\hat{\text{p-value}}_j$ approach their true values [41] and so B can be set based on obtaining a level of accuracy for $\hat{\text{p-value}}_j$ or based on computational limitations.

2.6. *Simulation setup*

Assuming that the distribution of the peptide log-ratio sample $\{l_i\}_{i=1}^n$ is bell-shaped and has a mean of zero for a 1:1 mix, we can model the distribution of the peptide log-ratios of a quantitative proteomics experiment as a mixture of three bell-shaped distribution functions [44]. The three distributions must consist of a central, zero mean component representing the distribution of peptides which have no meaningful changes in abundance, a distribution with a negative mean which represents negatively changed peptides, and a distribution with positive mean which represents positively changed peptides. In the simplest case, a Gaussian distribution can be used to model each of the bell-shaped components of the log-ratio data [45]. A hypothetical protein log-ratio distribution under this model can be seen in (Figure 2).

The density of log-ratios composes a mixture of three Gaussian distribution functions can be expressed as

$$f(l) = \pi_1 f_G(l; \mu_1, \sigma_1^2) + \pi_2 f_G(l; \mu_2, \sigma_2^2) + \pi_3 f_G(l; \mu_3, \sigma_3^2)$$

whereby the parameters π_1, π_2 and π_3, each taking values between zero and one, denotes the mixing proportion of each of three components respectively, and $\pi_g f_G(l; \mu_g, \sigma_g^2)$ is the Gaussian density of the form

$$f_G(l; \mu_g, \sigma_g^2) = \frac{1}{\sqrt{2\pi\sigma_g^2}} e^{-\frac{1}{2}\left(\frac{l-\mu_g}{\sigma_g}\right)^2}$$

where μ_g and σ_g^2 are the mean and variance of component g respectively, and g is equal to one, two or three.

Of course the underlying distribution which models every experiment is different and will result in a different set of

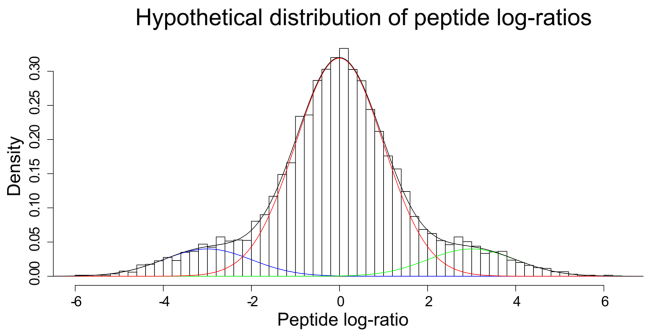

Figure 2. *Hypothetical distribution of peptide log-ratios using a mixture of three Gaussian distributions.* The black curve indicates the distribution of all peptide log-ratios in a protein expression experiment and the histogram is a sample of 10000 log-ratios from this distribution. The red curve indicates the distribution of all peptides that were unchanged in abundance, the blue curve indicates the distribution of negatively changed peptides, and the green curve indicates distribution of positively changed peptides. The black curve can be seen as the sum of the red, blue and green curves.

parameters for each of the mixture distribution components. We concentrate on a family of mixture distributions which we believe are representative of the data from many protein expression experiments. We fix the unchanged peptides distribution as a Gaussian distribution with a mean of zero and standard deviation of one. The positive and negative distributions are set to be symmetrical whereby the positive distribution has a mean of $\mu > 0$ and the negative distribution has a mean of $-\mu$. The variance and mixing proportions of the negative and positive distributions are equal and set to be σ^2 and π respectively. Therefore, we can write this family of peptide log-ratio distributions as

$$f(l) = (1-2\pi) f_G(l;0,1) + \pi f_G(l;-\mu,\sigma_2) + \pi f_G(l;\mu,\sigma_2)$$

whereby we can interpret 2π as the proportion of peptides that have changed in abundance in the experiment, $-\mu$ as the average negative change in the log-ratio and μ as the average positive change in the log-ratio for those peptides that have displayed a change in expression. Similar three component mixture models have been used in previous studies to model the unchanged, up-regulated and down-regulated SILAC log -ratios [44,46,47].

In order to cover as broad a range of distributions from this family as possible, we have chosen to simulate from the cases where μ is equal to 0.5, 1 or 2, σ^2 is equal to 0.25, 1 or 4, and 2π is equal to 0.1, 0.2 or 0.3. We believe these cases to be representative of the variety of relative variability and changes as well as the number of peptides that are differently expressed in an experiment. Simulating data from the experiment requires us to firstly determine the number of identified proteins m which we want to consider. For each of the m proteins j, we then randomly determine whether or not it is unchanged, negatively changed or positively changed with probabilities $1-2\pi$, π and π respectively. Once, we determine behavior of the protein, we generate n_j peptide log-ratios for the protein from a Gaussian distribution with mean and variance corresponding to the behavior. The value of n_j for each protein comes from the geometric density function

$$f(n_j) = (1-\phi)^{n_j - 1}\phi$$

where ϕ is a parameter which can take values between zero and one. We can justify the use of the geometric distribution because the minimum X^2-test for goodness-of-fit [48] indicates that the negative binomial distribution is appropriate for modeling the subsample sizes in the four data sets from [3] (with p-values of 1 in all cases) and the geometric distribution can be seen as a simple case of the negative binomial distribution. The full algorithm can be seen in (Figure 3). We choose to use the values $m = 500$ and $\phi = 0.05$ in all of our simulation scenarios since this results in an average total of 10000 peptide ratios which approximately corresponds to the values observed in the [3] data sets. Each set of parameters were simulated 10 times and some examples of simulated data can be seen in (Figure 4).

2.7. *FDR, TPR and ACC*

For each of the sets simulation parameters, we can evaluate the average estimated false discovery rate (FDR), true positive rate (TPR) and accuracy rate (ACC) [49] across the 10 repetitions. The FDR is a measure of the number of proteins incorrectly determined as changed, as a proportion of the number of proteins considered changed. The TPR is a measure of the number of changed proteins found as a proportion of the number of truly changed proteins, and the ACC is the total number of correctly classified proteins, both changed and unchanged as a proportion of the total number of proteins tested. All three values take on numbers from zero to one whereby a lower number is desirable for the FDR and a higher value indicates better performance for the TPR and ACC.

We can represent the true class labels of each of the m proteins in a simulated data set mathematically as $\{c_j\}_{j=1}^m$ where c_j is zero if the protein is unchanged and one if the protein is changed. Similarly, we can represent the result of the hypothesis tests as $\{t_j\}_{j=1}^m$ where t_j is zero if the p-value is greater or equal to 0.05 and one if the p-value is less than 0.05. Using this notation, the estimated FDR, TPR and ACC can be expressed as

$$\hat{FDR} = 1 - \frac{\sum_{\{j|t_j=1\}} I\{c_j = t_j\}}{\#\{j|t_j = 1\}}, \quad \hat{TPR} = \frac{\sum_{\{j|c_j=1\}} I\{c_j = t_j\}}{\#\{j|c_j = 1\}}, \quad \hat{ACC} = \frac{\sum_{j=1}^m I\{c_j = t_j\}}{m}$$

respectively where $I\{c_j = t_j\} = 1$ if $c_j = t_j$ is true and $I\{c_j = t_j\} = 0$ if $c_j = t_j$ is false. We can denote the average estimated FDR, TPR and ACC as $\bar{FDR}$, $\bar{TPR}$ and $\bar{ACC}$ respectively where they are each evaluated by averaging over the 10 repetitions of each set of simulations. These average measures allow us to comment on the performance of each of the hypothesis tests across the 27 different simulation scenarios when applied using the $\alpha = 0.05$ level of significance.

1. Set the proportion of changed proteins 2π.

2. Set the mean deviation in the expression of peptide log-ratios μ.

3. Set the variance of the changed log-ratio distributions σ^2 .

4. Set the parameter ϕ for generating the size of each protein subsample.

5. Set the number of proteins to simulate m.

6. For protein j in the set of proteins from one to m :

a. Randomly generate a protein subsample of size n_j from the geometric distribution $f(n_j) = (1\text{-}\phi)^{(nj\text{-}1)}\phi$.

b. Randomly Draw R from a uniform distribution in the interval of zero and one.

c. If $0 \leq R \leq 1 - 2\pi$:

i. Randomly generate n_j peptide log-ratios from the density $f_G(l\,; 0,1)$.

d. If $1 - 2\pi < R \leq 1 - \pi$:

i. Randomly generate n_j peptide log-ratios from the density $f_G(l\,; -\mu,\sigma^2)$.

e. If $1 - \pi < R \leq 1$:

i. Randomly generate n_j peptide log-ratios from the density $f_G(l\,; \mu,\sigma^2)$.

Figure 3. Peptide log-ratio simulation algorithm. Algorithm for simulating peptide log-ratio data based upon theoretical Gaussian mixture models.

2.8. *AUC statistic*

The area under the curve (AUC) measurement of the receiver operating characteristic (ROC) curve [50] was also calculated for each simulation scenario. The ROC curve plots the performance of a method in terms of the estimated false positive rate (FPR) and the estimated true positive rate achieved for each level of α used. The FPR can be interpreted as the number of truly unchanged proteins incorrectly classified as being changed as a proportion of the number of truly unchanged proteins. Like the other measurements, the FPR can take values between zero and one whereby lower values are desirable.

While the FPR and TPR are estimations of the probability of a type-I error and the power at every threshold level respectively, the AUC under the ROC curve is an average measure of performance of a method across all possible thresholds for a data set. Hence the AUC measurement has been widely adopted as a summary of performance in machine learning and proteomics applications [49,51–53].

We calculate the AUC for each method on each repetition of each set of simulation parameters. The average AUC measurement $\bar{AUC}$ for each testing method is calculated by averaging over the 10 repetitions of each simulation scenario. An example of ROC curves for an instance of the simulated data can be seen in (Figure 5). The average AUC, combined with the $\bar{FDR}$, $\bar{TPR}$ and $\bar{ACC}$ can give an indication of the relative performance of the permutation test against the competing testing methods.

3. Results

The normality assumption is necessary in guaranteeing accurate inferential results in many hypothesis tests. For this reason, the bell-shape of the distribution of the peptide log-ratios $\{l_i\}_{i=1}^{n}$ makes the assumption of a Gaussian distribution very tempting. It has been noted in the past that a Gaussian distribution is often not a good fit to actual peptide log-ratio data. We investigated this through observing histograms of the peptide-log ratios from our four subcellular fractions

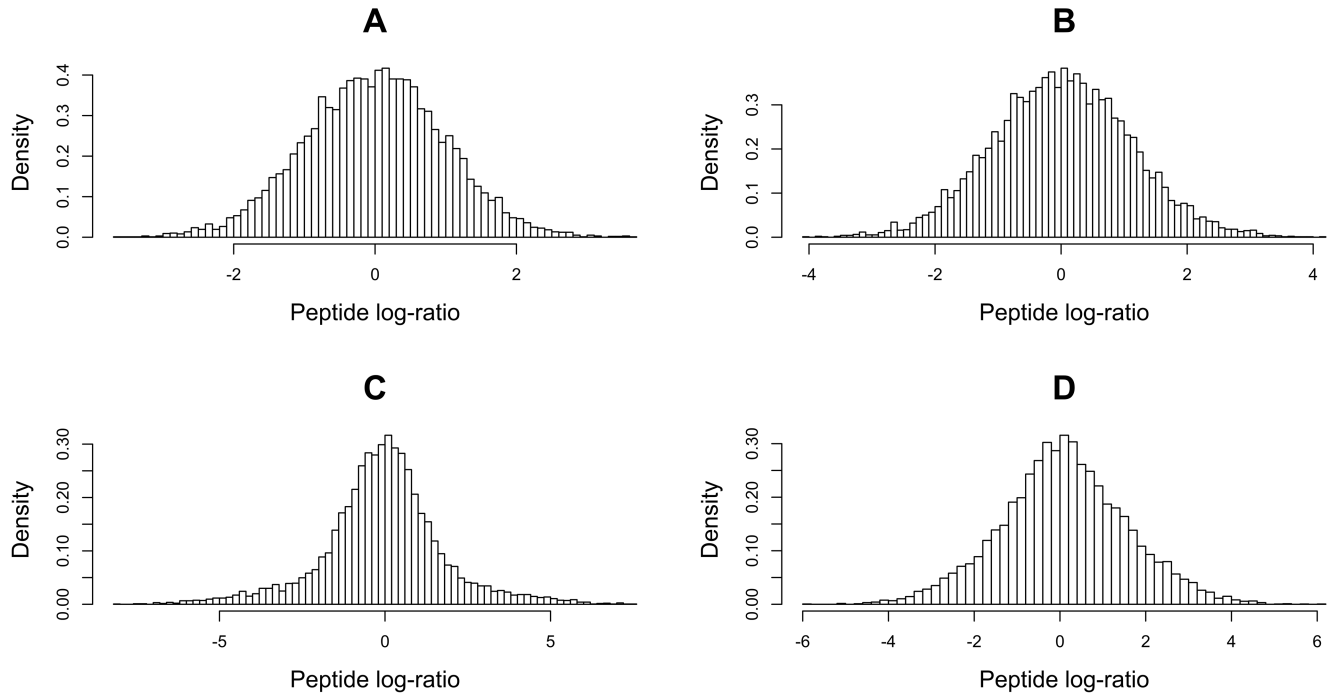

Figure 4. *Histograms of simulated peptide log-ratios under different scenarios.* Each of the histograms are an example of the simulated data of the peptide log-ratios from a protein expression experiment. The set of simulation parameters in each case are A: $2\pi = 0.1$, $\mu = 0.5$ and $\sigma^2 = 0.25$, B: $2\pi = 0.2$, $\mu = 0.1$ and $\sigma^2 = 1$, C: $2\pi = 0.3$, $\mu = 2$ and $\sigma^2 = 4$, and D: $2\pi = 0.3$, $\mu = 2$ and $\sigma^2 = 1$. We can see that changing the parameters can alter the peak, spread and shape of the tails of the distributions in each case.

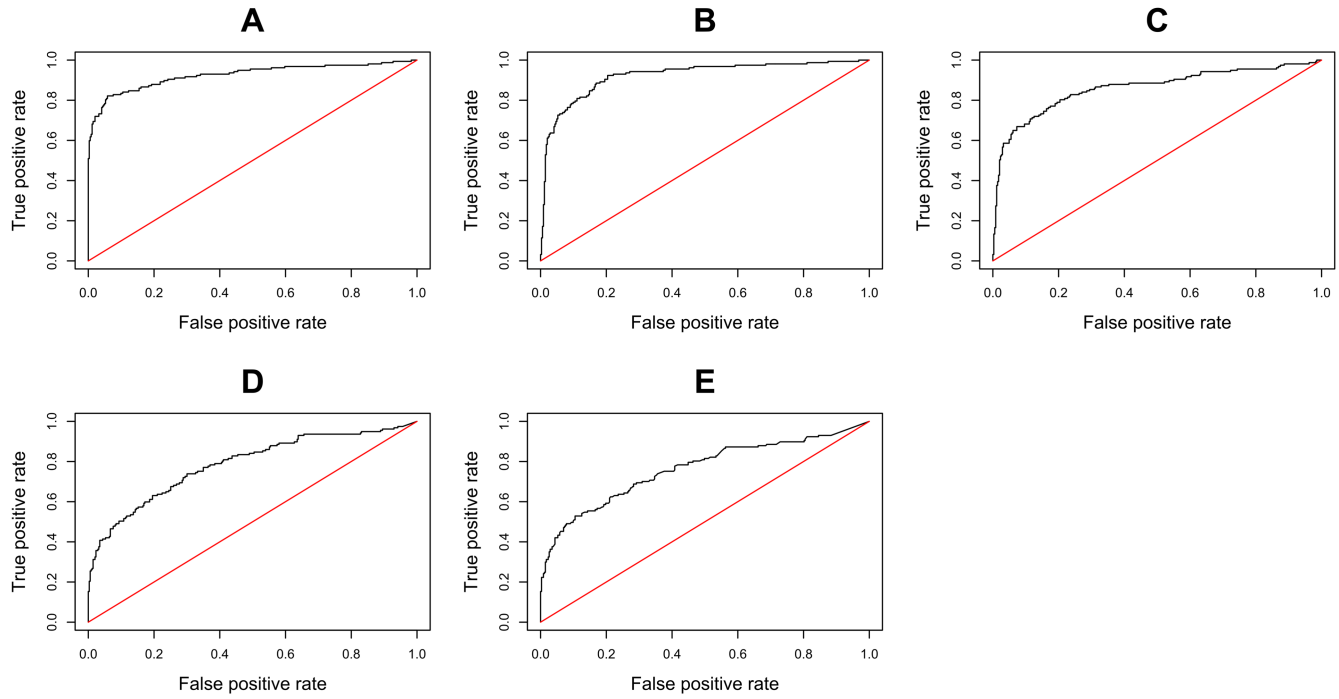

Figure 5. *ROC curves for a simulation using the parameters $2\pi = 0.3$, $\mu = 2$ and $\sigma^2 = 4$.* The ROC curve for each of the method is plotted as the black curve and the red line indicate the ROC curve for random guessing where TPR=FPR. The methods (with the AUC achieved over the data set) are A: permutation test (0.9252), B: z-test (0.9220), C: robust z-test (0.8600), D: t-test (0.7857) and E: Wilcoxon signed-rank test (0.7621).

data sets (Figure 6).

3.1. *One-sample t-tests*

Although the histograms are not Gaussian, traditional insight about the t-test often suggest that if the sample distribution does not deviate too greatly from the Gaussian distribution, a t-test may still be robust enough for use [56,57]. It has been shown that the t-test may still be applied when the data comes from a symmetric distribution which can often be the case when observing peptide log-ratio distributions over all proteins [58].

When we observe the largest peptide log-ratio subsamples from the four subcellular quantitations from [3] (Figure 7), we notice that not only is there strong deviations from symmetry in most cases but also multimodalism which severely violates the normality assumption. These results obtained from the largest subsamples from each of the four quantitations implies to us that it is difficult to assume normality for the peptide log-ratio populations for each of the quantitated proteins and thus invalidates the use of the one-sample t-test.

3.2. *One-sample z-tests and robust z-tests*

There is a similar problem with violations of the normality assumption in the use of the z-test whereby a Gaussian distribution is fitted to the protein log-ratio sample $\{L_j\}_{j=1}^{m}$ and the tails of the fitted distribution are used to calculate the p-value of each protein [60]. We can see the deviations from normality in the protein log-ratio histograms of the four subcellular quantitations from [3] (Figure 8). Because of the lack of fit of the Gaussian distribution to the protein log-ratio samples, the p-values calculated using fitted a curve will not accurately represent the true significance of each of the proteins.

There have been propositions for the use of more robust means of hypothesis testing. One such method is the robust z-test which uses the outlier-insensitive properties of the percentiles of a distribution rather than the mean and standard deviation which are known to be strongly influenced by extreme observations.

Under the assumption of Gaussian protein log-ratios, this z-statistic also has a standard Gaussian distribution and the p-value for each protein can be calculated by evaluating the tail probability P $(z > Z_j)$ where z is a standard Gaussian random variable. However, the method does not address the problem of deviations from normality in protein log-ratio distributions.

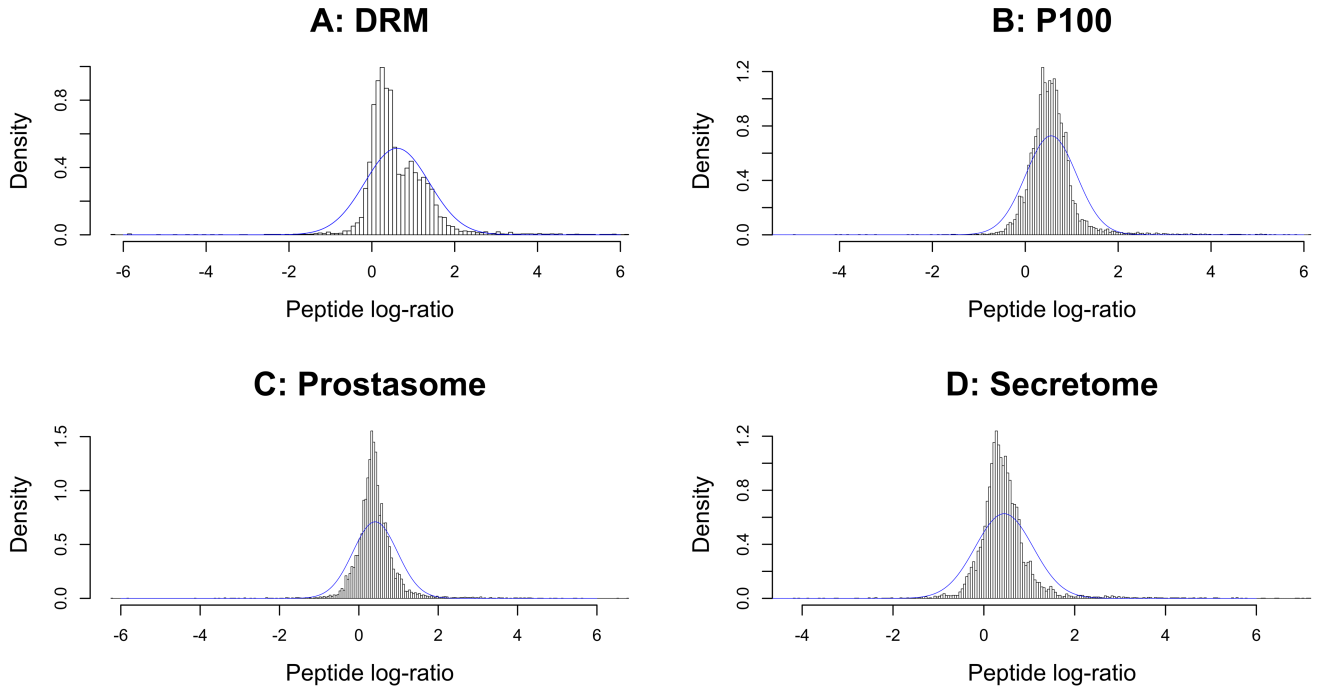

Figure 6. *Histograms and fitted Gaussian distributions to peptide log-ratios of subcellular quantitations.* The experimental data comes from the four subcellular proteome quantitations for the effect of Polymerase I and Transcript Release Factor (PTRF) expression in human prostate carcinoma PC-3 cells. The PTRF expressing cells were labeled with heavy amino acids and the cells without PTRF expression were labeled with light amino acids. The four subcellular fractionals (with number of peptides quantitated in brackets) were A: Detergent-Resistant Membrane (7338), B: Total membrane P100 fraction as described in [54] (10271), C: Prostasome (8682) and D: Secretome (9653). The Lilliefors test for normality p-values [55] for each sample was less than 2.2×10^{-16} indicating that the four sets of peptide log-ratios are severely non-Gaussian.

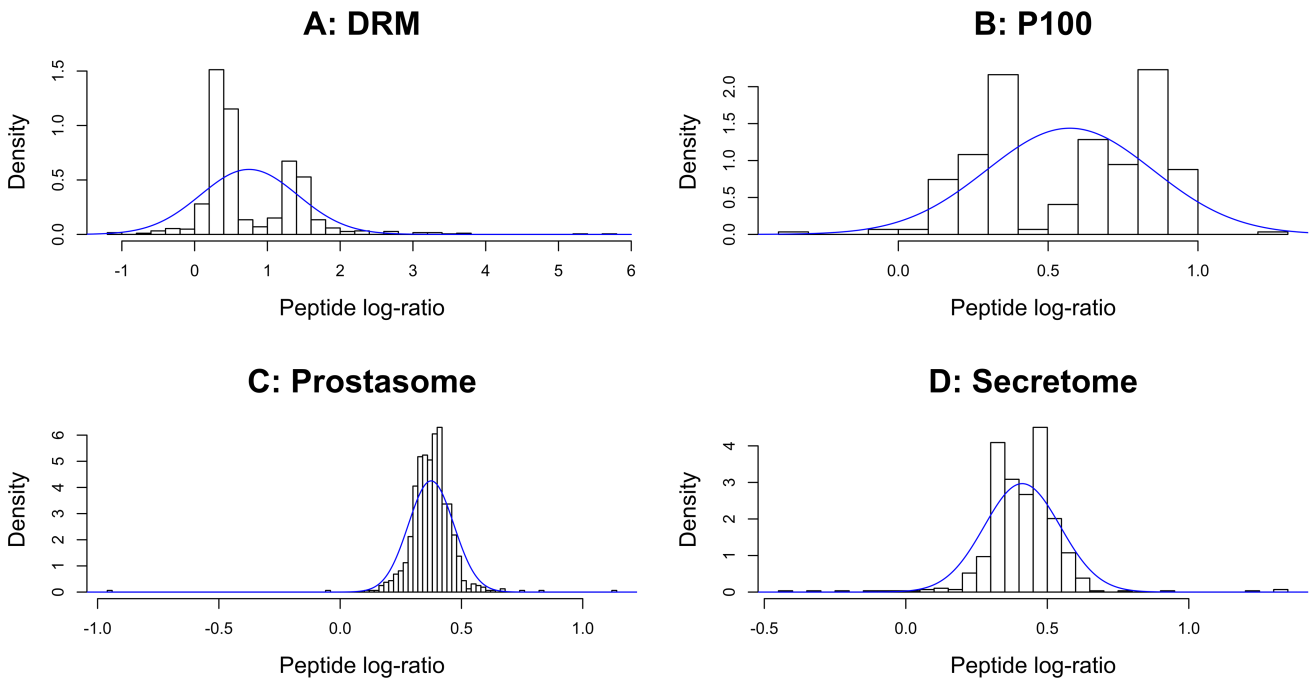

Figure 7. *Histograms and fitted Gaussian distributions to peptide log-ratios of most common proteins.* We plot the histograms of the peptide log-ratio subsample for the most commonly quantitated proteins of each of the quantitated fractions from [3]. The four most common proteins (with the number of peptides in the subsample in brackets) are A: Plectin-1 (929), B: Nucleophosmin (296), C: Thrombospondin-1 (802) and D: Thrombospondin-1 (577) coming from the DRM, P100, Prostasome and Secretome fractions respectively. The p-values for the D'Agostino test for skewness [59] for each subsample is A: less than 2.2×10^{-16}, B: 0.2656, C: less than 2.2×10^{-16} and D: 0.01957.

In order to assess the effects of violations of the normality assumption, we calculate the relative error between the true right-tail probability of a log-ratio of 3 in various Student-t distributions as compared to the right-tail probability calculated using the robust z-test as a ratio of the true probability. It is known that the Student-t distributions have heavier tails than the Gaussian distribution and asymptotically approaches the Gaussian distribution as the number of degrees of freedom approaches infinity [29]. We can see from (Figure 9) that the robust z-test would tend to underestimate the p-value of a protein log-ratio of 3 when the distribution of protein log-ratios has heavier than Gaussian tails. Because of the discrepancies in the p-values, the robust z-test would tend to have a higher proportion of false positives when the protein log-ratios come from distribution with heavier than Gaussian tails.

3.3. Wilcoxon signed-rank tests

The Wilcoxon test is a nonparametric hypothesis testing procedure that does not depend on the distribution of the underlying population. The procedure is therefore robust to deviations from normality and can therefore be implemented to sample data regardless of the underlying population distribution. The cost to this robustness is that the test lacks the power to detect large deviations from the null hypothesis in small samples.

To illustrate the lack of power of the test, suppose that we only observe positive peptide log-ratios in the subsample $\{l_i \mid p_i = j\}$for a protein j. The p-value of the Wilcoxon signed rank test at various number of peptides $\#\{l_i \mid p_i = j\}$ can be seen in (Figure 10) showing that a protein needs to be quantitated by at least six peptides in order to be deemed significantly different within the two cell populations at the five percent level. Proteins which are quantitated only by positive peptide log-ratios are a best case scenario and thus, for proteins that are quantitated by both positive and negative peptide log-ratios, the number of peptides needed for a significant result would be greater than six. The lack of power makes it impossible to find significant differences for rare proteins between the two cell populations and therefore diminishes the usefulness of the procedure.

3.4. *Permutation tests*

The permutation test proposed here is both distribution free and more powerful than the Wilcoxon test in small samples. The lack of distributional assumptions allows the permutation tests to be applicable to SILAC log-ratio data which as we have shown to not meet the required assumptions for conventional hypothesis tests. Additionally, the permutation test appears to be better suited than the other procedures, especially for assessing the significance of proteins that are quantitated by small numbers of peptides.

We now report on the simulation study conducted to assess the performance of the permutation test against the t-

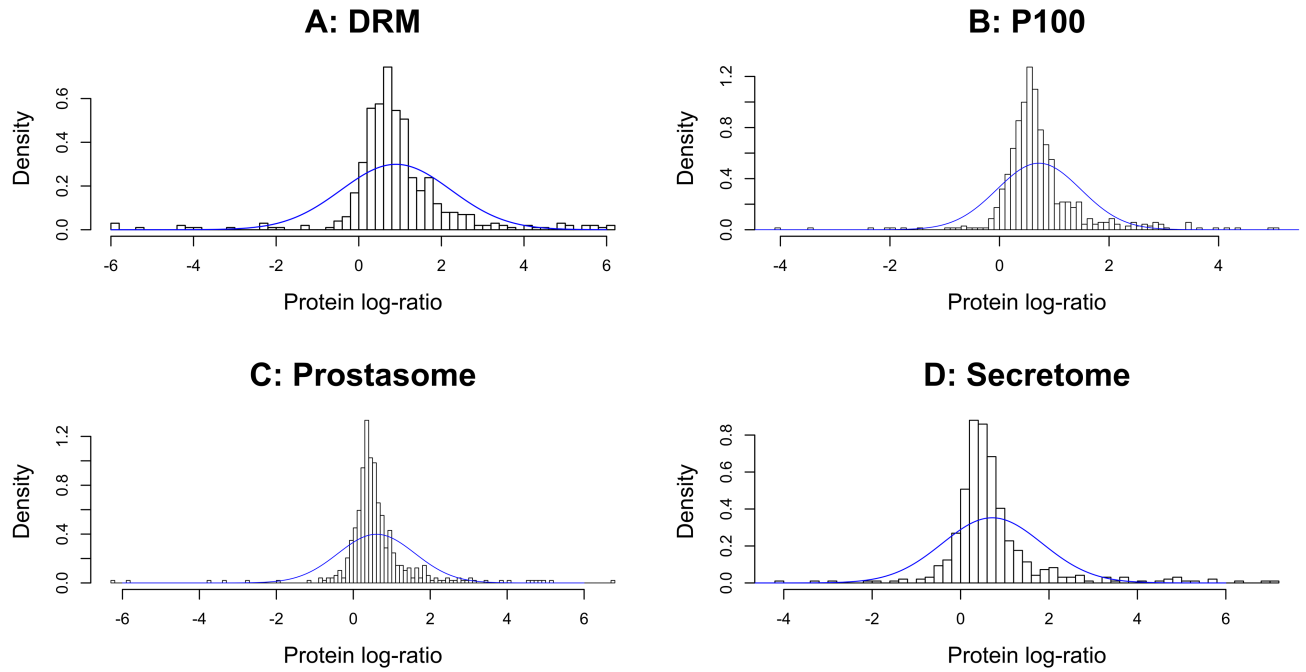

Figure 8. *Histograms and fitted Gaussian distributions to protein log-ratios of subcellular quantitations.* The four subcellular fractionals (with number of protein quantitated in brackets) were A: Detergent-Resistant Membrane (504), B: P100 (691), C: Prostasome (489) and D: Secretome (484). The Lilliefors test for normality p-values for each sample was less than 2.2×10^{-16} shows that the protein log-ratio samples are significantly non-Gaussian in distribution.

test, z-test, robust z-test and Wilcoxon signed-rank test across the 27 simulation scenarios. The performance for each method in each scenario is measured by the FDR, TPR, ACC and AUC averaged over the 10 repetitions.

3.5. *Evaluating competing methodologies*

We firstly assess the performance of each of the methods as applied with the significance level of $\alpha = 0.05$. Since each of the hypothesis testing methods results in a p-value for each protein, we can use the rule that if the p-value is less than 0.05, we consider that the protein is differently abundant between our two samples and if the p-value is greater or equal to 0.05 then we consider that the protein is unchanged. For the t-test, a p-value of 1 is given to any protein that it cannot quantitate since that protein would generally undergo no further assessment and therefore is vacuously deemed unchanged. The number of permutations B is set to 10000 which have been shown to result in sufficiently accurate p-values when tested against a Gaussian null distribution. Results of the FDR, TPR, ACC and AUC across the 27 different simulation scenarios can be found respectively in (Table S1), (Table S2), (Table S3) and (Table S4) in Supplementary Section 1.

To facilitate comparison, we also ranked the five methods for each of the set of parameters, where one represents the method that performed the worst in a scenario and five represents the best method. Ties are each given fractional ranks equal to the average of the tied ranks. The ranked results for

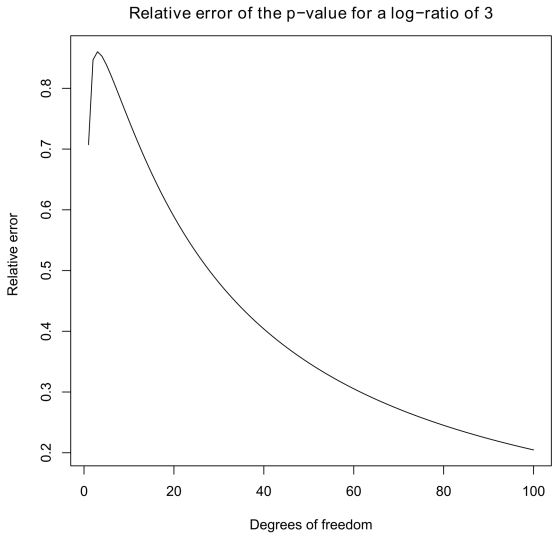

Figure 9. *Relative errors between the true p-value and robust z-test p-value for a log-ratio of 3.* The relative error at is calculated as $(P_t - P_z)/P_t$ where $P_t = P(t < (3\text{-}P_{50})/(P_{84.13} - P_{50}))$, $P_t = P(z < (3\text{-}P_{50})/(P_{84.13} - P_{50}))$, t is a Student-t random variable with degree of freedom ν, z, is a standard Gaussian random variable and the percentiles come from a Student-t distribution with degrees of freedom ν. The relative error for degrees of freedom of one to 100 are graphed showing that the true p-value is always greater than those calculated using the robust z-test. The relative error is decreasing with increasing degrees of freedom because the shapes of the Student-t distribution approaches normality as the degrees of freedom approaches infinity.

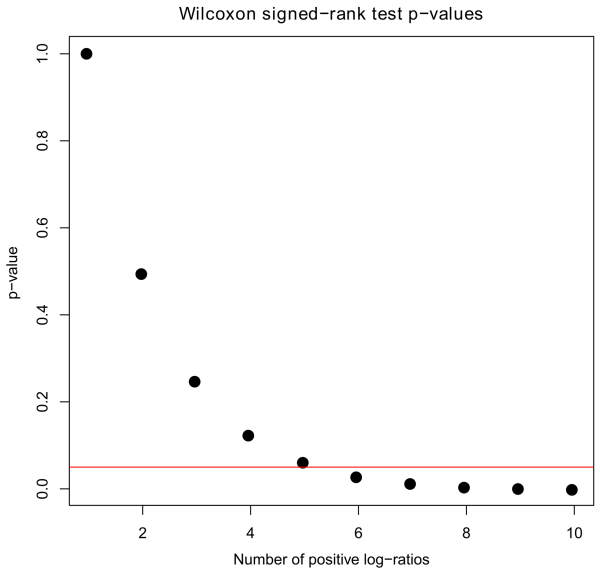

Figure 10. *p-values of a Wilcoxon signed-rank test for various numbers of positive sample observations.* The graph shows the p-values of a Wilcoxon signed-rank test for various sized samples of positive values. The red line indicates the five percent level of significance and indicates that we need at least six observations in order to obtain a p-value less than the significance level. The graph also shows that the p-value is decreasing with respects to larger numbers of observations, indicating that the more positive peptide log-ratios observed, the greater the probability that there is a difference between the protein abundances of the heavy and light cell populations.

the $\overline{\text{FDR}}$, $\overline{\text{TPR}}$ and $\overline{\text{ACC}}$ and $\overline{\text{AUC}}$ across the 27 different simulation scenarios can be found respectively in (Table S5), (Table S6), (Table S7) and (Table S8) in Supplementary Section 1.

Using the rankings, we can calculate the total of the ranks over each of the measurements. These totals give us an indication of the relative performance for each of the testing procedures over all simulation scenarios. The total ranks for each of the six testing methods over the four measurements can be found in (Table 1).

From the total ranks results we can see that the permutation test performed the best in terms of $\overline{\text{FDR}}$ and $\overline{\text{ACC}}$, second best in terms of $\overline{\text{AUC}}$, and third best in terms of. These results are subjected to variability due to the randomized process of the simulation study. It is therefore also interesting to observe the statistical significance of each of the difference between the ranks of the permutation test and those of the competing methods. The paired-sign test [61] is applied to the rank data for each of the four measurements in order to perform this assessment. The p-values for each of the tests can be found in (Table 2).

The results of the hypothesis tests suggest that at the 0.05 level, there is no significant difference between the ranks of the permutation test and the z-test over both the $\overline{\text{TPR}}$ and $\overline{\text{AUC}}$. Therefore, there is not enough evidence to suggest that the z-test outperforms the permutation test in those two measurements.

4. Discussion

The tabulated results show that the permutation test can be seen as being the best performing method or tied for best performing method over the average FDR, ACC and AUC measurements. This indicates that over a variety of possible simulation scenarios, the permutation test can be seen to either equal or outperform the z-test, robust z-test, t-test and Wilcoxon signed-rank test. The performance over the FDR and ACC indicates that the permutation test makes the fewest false identification of changed proteins as well as the fewest mislabeling of proteins across the different simulation parameter sets when applying a $\alpha = 0.05$ threshold. The result from the AUC shows that the permutation test is also the best test to use with any arbitrary threshold as well, indicating that it is not only optimal at the 0.05 level.

The permutation test is only significantly bettered by the robust z-test in terms of ranks of $\overline{\text{TPR}}$. Observing (Table 2), we note that although the permutation test is often beaten by the robust z-test, the absolute value of the $\overline{\text{TPR}}$ in many scenarios are very small. We also note that the robust z-test performs the worst in terms of $\overline{\text{FDR}}$ and thus its strong performance can be attributed to the method's inaccuracy.

We also note that the simulation setup used is favorable to the tests which rely on the assumption of normality. This is because the Gaussian mixture models have tails of similar shapes to single Gaussian distributions which favors the z-test and the robust z-test. The sampling of the peptide log-ratios for each protein subsample from a single Gaussian distribution strongly favors the t-test since the assumption of

Table 1. The total ranks for each of the hypothesis testing procedures over the $\overline{\text{FDR}}$, $\overline{\text{TPR}}$, $\overline{\text{ACC}}$ and $\overline{\text{AUC}}$ measurements.

Measurement — Test	Permutation test	z-test	Robust z-test	t-test	Wilcoxon signed-rank test
$\overline{\text{FDR}}$	124	65	35	107	74
$\overline{\text{TPR}}$	75	90	134	35	71
$\overline{\text{ACC}}$	121	67	47	84.5	85.5
$\overline{\text{AUC}}$	108	115	77	67	38

Table 2. Paired-sign p-values for the differences in ranks over the $\overline{\text{FDR}}$, $\overline{\text{TPR}}$, $\overline{\text{ACC}}$ and $\overline{\text{AUC}}$ measurements.

Measurement — Test	z-test	Robust z-test	t-test	Wilcoxon signed-rank test
$\overline{\text{FDR}}$	4.172×10^{-7}	4.172×10^{-7}	5.925×10^{-3}	4.923×10^{-5}
$\overline{\text{TPR}}$	0.1221	1.49×10^{-8}	1.514×10^{-3}	0.2478
$\overline{\text{ACC}}$	5.648×10^{-6}	4.172×10^{-3}	5.925×10^{-3}	1.514×10^{-3}
$\overline{\text{AUC}}$	1	0.01916	4.923×10^{-5}	4.923×10^{-5}

normality within the subsamples are held. In order to test the performance of the permutation test under deviations from normality, we simulated two additional scenarios where we applied the Laplace distribution [62] and the Student-t distribution with $\nu = 3$ degrees of freedom. The results of these simulations can be found in Supplementary Section 2.

Analyses of these additional simulations show that the permutation test has equivalent performance in both the Gaussian and leptokurtic simulation scenarios. Our choices of possible distributions for modeling the changed peptides distributions are by no means exhaustive. Other examples of possible deviations from normality include modeling the log-ratios with Cauchy distribution mixtures [63], modeling the tails with generalized Pareto distributions [44] or using asymmetric up-regulated and down-regulated distributions.

4.1. *False discovery rate mitigation*

In the simulation scenarios, it is possible to calculate the estimated FDR in each case due to the knowledge of the class labels. However, in an experimental situation, where the true nature of each protein is not known, it is desirable to control the FDR at an acceptable level such that the FDR does not inflate the multiple testing error rate of any downstream analysis. Because the permutation test outputs a p-value, it is easy to perform FDR control either with techniques such as the Benjamini-Hochberg procedure [64] or with empirical Bayes methods [65,66].

In any case, it must be acknowledge that with stricter control of the FDR, there is a trade-off of a reduction in the TPR. Therefore, the cost of controlling the number of false discoveries made is that less of the truly changed proteins will be declared as having a significant change in abundance.

4.2. *Other methodologies*

Apart from the z-test, Robust z-test, t-test and Wilcoxon signed-rank test, there are many alternative methods available for quantitative proteomics experiments. Such methods include spectral counting methods and multiple experimental replicates methods such as those found in [67] and [22]. These methods, unlike the methods for ratios and log-ratios rely on greater volumes of data from multiple repeated experiments or richer experimental data outputs. Such comprehensive data is unavailable in many experimental situations, so we do not discuss these methods further.

Other methods which only rely on single ratio data sets such as likelihood based approaches [44,68]. These methods were not used for comparison because they are often application specific and require statistical expertise which is not available to many researchers. For this reason, we only chose to discuss simple hypothesis testing methods such as permutation testing on log-ratios since this is interpretable and implementable by most researchers using data that is already available to them.

4.3. Broader applications

Although the discussion has focused on permutation testing of SILAC ratio data, as successfully implemented in [3], the permutation test can also be applied to data resulting from other relative quantitation methods such as isobaric tags for relative and absolute quantitation (iTRAQ) [69], isotope-code affinity tags (ICAT) [70], label-free quantitation [67] and tandem mass tags (TMT) [71]. Outside of quantitative proteomics, it is also possible to apply the permutation testing method to fold-change data from microarray gene expression experiments [33].

5. Concluding Remarks

The permutation test which we introduced is distribution free, applicable across a range of relative quantitation methods, and was found to be superior to the other tests across the spectrum of significance level through analyzing the AUC statistics. At the usual 0.05 level it was also found the permutation test generally performed better than the competing methods over the ACC, FDR and TPR criteria.

The advantage of the permutation testing procedure over more technical methods is that it is a simple hypothesis test that outputs a p-value which can be interpreted by scientists. This allows it to be controlled for FDR and inferred in the same way as the p-values of other simple tests.

The permutation method is implementable in any programming environment and an R implementation of the procedure is available upon request. We are currently assembling a server-based web interface to for the test with ease of

deployment in mind. This web interface will provide additional data visualization and summarization facilities along with the computation of the permutation p-values.

6. Supplementary material

Supplementary Material.pdf contains Supplementary Section 1 and Supplementary Section 2.

Acknowledgements

This work was supported by a University of Queensland summer research scholarship to HN, and The University of Queensland Diamantina Institute Fellowship to MH. MH is supported by an NHMRC Career Development Award No. 569512). We gratefully thank Kim-Anh Le Cao for her critique of the manuscript. We would also like to thank the anonymous reviewers for their helpful comments which lead to improvements in the paper.

References

[1] S.-E. Ong, B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, M. Mann, Mol Cell Proteomics 1 (2002) 376–386.

[2] J.S. Andersen, Y.W. Lam, A.K.L. Leung, S.-E. Ong, C.E. Lyon, A.I.L.& M. Mann, Nature 433 (2005) 77–83.

[3] K.L. Inder, Y.Z. Zheng, M.J. Davis, H. Moon, D. Loo, H. Nguyen, J.A. Clements, R.G. Parton, L.J. Foster, M.M. Hill, Mol Cell Proteomics 11 (2012) M111.012245.

[4] S.E. Ong, L.J. Foster, M. Mann, Method Mol Cell Biol 29 (2003) 124–130.

[5] E. Milner, E. Barnea, I. Beer, and A. Admon, Mol Cell Proteomics 5 (2006) 357–365.

[6] B. Soufi, C. Kumar, F. Gnad, M. Mann, I. Mijakovic, and B. Macek, J Proteome Res 9 (2010) 3638–3646.

[7] A. Cuomo, T. Bonaldi, Method Mol Cell Biol 662 (2010) 59–78.

[8] M.D. Sury, J.X. Chen, M. Selbach, Mol Cell Proteomics 9 (2010) 2173–2183.

[9] W. Schütz, N. Hausmann, K. Krug, R. Hampp, B. Macek, Plant Cell 23 (2011) 1701–1705.

[10] A. Gruhler, J.V. Olsen, S. Mohammed, P. Mortensen, N.J. Faergeman, M. Mann, O.N. Jensen, Mol Cell Proteomics 4 (2005) 310–327.

[11] T. Geiger, J. Cox, P. Ostasiewicz, J.R. Wisniewski, M. Mann, Nat Methods 7 (2010) 383–385.

[12] M. Mann, Nat Rev Mol Cell Bio 7 (2006) 952–958.

[13] E.A. Kapp, F. Schütz, L.M. Connolly, J.A. Chake, J.E. Meza, C.A. Miller, D. Fenyo, J.K. Eng, J.N. Adkins, G.S. Omenn, R.J. Simpson, Proteomics 5 (2005) 3475–3490.

[14] K.W. Lau, A.R. Jones, N. Swainston, J.A. Siepen, S.J. Hubbard, Proteomics 7 (2007) 2787–2799.

[15] K.T. Rigbolt, B. Blagoev, Method Mol Cell Biol 658 (2010) 187–204.

[16] H. Choi, A.I. Nesvizhskii, J Proteome Res 7 (2008) 47–50.

[17] J. Graumann, N.C. Hubner, J.B. Kim, K. Ko, M. Moser, C. Kumar, J. Cox, H. Schöler, M. Mann, Mol Cell Proteomics 7 (2008) 672–683.

[18] M. Ackermann, K. Strimmer, BMC Bioinformatics 10 (2009) 47.

[19] D.W. Huang, B.T. Sherman, R.A. Lempicki, Nucleic Acids Res 37 (2009) 1–13.

[20] S.Y. Rhee, V. Wood, K. Dolinski, S. Draghici, Nat Rev Genet 9 (2008) 509–515.

[21] B. van Breukelen, H.W.P. van den Toorn, M.M. Drugan, and A.J.R. Heck, Bioinformatics 25 (2009) 1472–1473.

[22] J.W.H. Wong, M.J. Sullivan, G. Cagne, Brief Bioinform 9 (2007) 156–165.

[23] N.T. Seyfried, Y.M. Gozal, E.B. Dammer, Q. Xia, D.M. Duong, D. Cheng, J.J. Lah, A.I. Levey, and J. Peng, Mol Cell Proteomics 9 (2010) 705–718.

[24] W.-T. Lin, W.-N. Hung, Y.-H. Yian, K.-P. Wu, C.-L. Han, Y.-R. Chen, Y.-J. Chen, T.-Y. Sung, and W.-L. Hsu, J Proteome Res 5 (2006) 2328–2338.

[25] J. Cox, M. Mann, Nat Biotechnol 26 (2008) 1367–1372.

[26] J. Schimmel, K.M. Larsen, I. Matic, M. van Hagen, J. Cox, M. Mann, J.S. Andersen, A.C.O. Vertegaal, Mol Cell Proteomics 7 (2008) 2107–2122.

[27] S. ten Have, S. Boulon, Y. Ahmad, A.I. Lamond, Proteomics 11 (2011) 1153–1159.

[28] P. Mertins, H.C. Eberl, J. Renkawitz, J.V. Olsen, M.L. Tremblay, M. Mann, A. Ullrich, H. Daub, Mol Cell Proteomics 7 (2008) 1763–1777.

[29] Student, Biometrika 6 (1908) 1–25.

[30] F. Wilcoxon, Biometrics Bull 1 (1945) 80–83.

[31] M. Dwass, Ann Math Stat 28 (1957) 181–187.

[32] S. Dudoit, Y.H. Yang, M.J. Callow, T.P. Speed, Stat Sinica 12 (2002) 111–139.

[33] V.G. Tusher, R. Tibshirani, G. Chu, P Natl Acad Sci USA 98 (2001) 5116–5121.

[34] F. Al-Shahrour, R. Díaz-Uriarte, J. Dopazo, Bioinformatics 20 (2004) 578–580.

[35] J.J. Goeman, P. Bühlmann, Bioinformatics 23 (2007) 980–987.

[36] C. Schaab, Method Mol Cell Biol 696 (2011) 41–57.

[37] D.M. Walther, M. Mann, Mol Cell Proteomics 10 (2011) M110.004523.

[38] S. Smit, M.J. van Breemen, H.C.J. Hoefsloot, A.K. Smilde, J.M.F.G. Aerts, C.G. de Koster, Anal Chim Acta 592 (2007) 210–217.

[39] S. Smit, H.C.J. Hoefsloot, A.K. Smilde, J Chromatogr B 866 (2008) 77–88.

[40] A.W. van der Vaart, Asymtotic Statistics, Cambridge University Press, 1998.

[41] M.D. Ernst, Stat Sci 19 (2004) 676–685.

[42] B. Phipson, G.K. Smyth, Stat Appl Genet Mo B 9 (2010) 39.

[43] R.A. Fisher, Statistical Methods for Research Workers, Oliver and Boyd, 1958.

[44] A.A. Margolin, S.-E. Ong, M. Schenone, R. Gould, S.L. Schreiber, S.A. Carr, T.R. Golub, PLoS One 4 (2009) e7454.

[45] G.J. McLachlan, K.E. Basford, Mixture Models: Inference and Applications to Clustering, Marcel Dekker, 1988.

[46] A. von Kriegsheim, D. Baiocchi, M. Birtwistle, D. Sumpton, W. Bienvenut, N. Morrice, K. Yamada, A. Lamond, G. Kalna, R. Orton, D. Gilbert, W. Kolch, Nat Cell Biol 11 (2009) 1458–1464.

[47] S.-E. Ong, M. Schenone, A.A. Margolin, X. Li, K. Do, M.K. Doud, D.R. Mani, L. Kuai, X. Wang, J.L. Wood, N.J. Tolliday, A.N. Koehler, L.A. Marcaurelle, T.R. Golub, R.J. Gould, S.L. Schreiber, S.A. Carr, P Natl Acad Sci USA 106 (2009) 4617–4622.

[48] W.G. Cochran, Ann Math Stat 23 (1952) 315–345.

[49] T. Fawcett, Pattern Recogn Lett 27 (2006) 861–874.

[50] A.P. Bradley, Pattern Recogn 30 (1997) 1145–1159.

[51] N. Chen, W. Sun, X. Deng, Y. Hao, X. Chen, B. Xing, W. Jia, J. Ma, H. Wei, Y. Zhu, X. Qian, Y. Jiang, F. He, Proteomics 8 (2008) 5108–5118.

[52] J. Cui, X. Ma, L. Chen, J. Zhang, BMC Bioinformatics 12 (2011) 439.

[53] S.R. Ramakrishnan, C. Vogel, T. Kwon, L.O. Penalva, E.M. Marcotte, D.P. Miranker, Bioinformatics 25 (2009) 2955–2961.

[54] K.L. Inder, C. Lau, D. Loo, N. Chaudhary, A. Goodall, S. Martin, A. Jones, D. van der Hoeven, R.G. Parton, M.M. Hill, J.F. Hancock, J Biol Chem 284 (2009) 28410–28419.

[55] H.W. Lilliefors, J Am Stat Assoc 62 (1967) 399–402.

[56] J.V. Bradley, Brit J Math Stat Psy 31 (1978) 144–152.

[57] H. Hotelling, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, 1961, .

[58] B. Efron, J Am Stat Assoc 64 (1969) 1278–1302.

[59] R.B. D'Agostino, Biometrika 57 (1970) 679–681.

[60] J.-Y. Zhou, J. Hanfelt, J. Peng, Proteom Clin Appl 1 (2007) 1342–1350.

[61] W.J. Dixon, A.M. Mood, J Am Stat Assoc 41 (1946) 557–566.

[62] C. Forbes, M. Evans, N. Hastings, B. Peacock, Statistical Distributions, Wiley, 2011.

[63] F.P. Breitwieser, A. Mu¨ller, L. Dayon, T. Ko¨cher, A. Hainard, P. Pichler, U. Schmidt-Erfurth, G. Superti-Furga, J.-C. Sanchez, K. Mechtler, K.L. Bennett, and J. Colinge, J Proteome Res 10 (2011) 2758–2766.

[64] Y. Benjamini, Y. Hochberg, J Roy Stat Soc B 57 (1995) 289–300.

[65] B. Efron, J Am Stat Assoc 99 (2004) 96–104.

[66] G.J. McLachlan, R.W. Bean, L.B.-T. Jones, Bioinformatics 22 (2006) 1608–1615.

[67] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, B. Kuster, Anal Bioanal Chem 389 (2007) 1017–1031.

[68] C. Pan, G. Kora, W.H. McDonald, D.L. Tabb, N.C. VerBerkmoes, G.B. Hurst, D.A. Pelletier, N.F. Samatova, R.L. Hettich, Anal Chem 78 (2006) 7121–7131.

[69] P.L. Ross, Y.N. Huang, J.N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, D.J. Pappin, Mol Cell Proteomics 3 (2004) 1154–1169.

[70] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, R. Aebersold, Nat Biotechnol 17 (1999) 994–999.

[71] A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, and C. Hamon, Anal Chem 75 (2003) 1895–1904.