Journal of Proteomics 129 (2015) 83–92

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics☆

Sarah R. Langley [a,b,*], Manuel Mayr [b]

[a] Division of Brain Sciences, Imperial College Faculty of Medicine, London, UK
[b] King's British Heart Foundation Centre, King's College London, London, UK

ARTICLE INFO

Article history:
Received 9 March 2015
Received in revised form 10 July 2015
Accepted 13 July 2015
Available online 18 July 2015

Keywords:
Label-free mass spectrometry
Spectral counts
Differential expression
Statistical methodology

ABSTRACT

Label-free LC-MS/MS proteomics has proven itself to be a powerful method for evaluating protein identification and quantification from complex samples. For comparative proteomics, several methods have been used to detect the differential expression of proteins from such data. We have assessed seven methods used across the literature for detecting differential expression from spectral count quantification: Student's t-test, significance analysis of microarrays (SAM), normalised spectral abundance factor (NSAF), normalised spectral abundance factor-power law global error model (NSAF-PLGEM), spectral index (SpI), DESeq and QSpec. We used 2000 simulated datasets as well as publicly available data from a proteomic standards study to assess the ability of these methods to detect differential expression in varying effect sizes and proportions of differentially expressed proteins. At two false discovery rate (FDR) levels, we find that several of the methods detect differential expression within the data with reasonable precision, others detect differential expression at the expense of low precision, and finally, others which fail to identify any differentially expressed proteins. The inability of these seven methods to fully capture the differential landscape, even at the largest effect size, illustrates some of the limitations of the existing technologies and the statistical methodologies.

Significance: In label-free mass spectrometry experiments, protein identification and quantification have always been important, but there is now a growing focus on comparative proteomics. Detecting differential expression in protein levels can inform on important biological mechanisms and provide direction for further study. Given the high cost and labour intensive nature of validation experiments, statistical methods are important for prioritising proteins of interest. Here, we have performed a comparative analysis to investigate the statistical methodologies for detecting differential expression and provide a reference for future experimental designs.

This article is part of a Special Issue entitled: Computational Proteomics.

1. Introduction

Protein detection and quantification have vastly improved in recent years with the technological advances of mass spectrometry. Liquid chromatography tandem mass spectrometry (LC-MS/MS) has become the method of choice for quantitative proteomics and can now assess protein samples in a bottom-up format with reasonable throughput. There are several methods for tagged or isotope labelled quantification, including isobaric tags for relative and absolute quantitation (iTRAQ) [1], tandem mass tags (TMT) [2] and stable isotope labelling by amino acids in cell culture (SILAC)[3]. These methods offer multiplexing capability at the requirement of more complex protocols and expensive reagents. However, SILAC is unsuitable for clinical samples and the tagged methods have the limitation that co-isolation of multiple precursor ions can interfere with accurate quantitation. Instead, label-free methods aim to provide relative quantification without isotopic labelling and are becoming increasingly popular in proteomics [4–6].

For label-free proteomics, one can quantify proteins by using their spectral counts as an approximation of protein abundance. Spectral counts are simply the total number of spectra per identified protein and can be easily calculated from the detected peptides by LC-MS/MS; within a protein, they can be taken as an semi-quantitative approximation as a protein with higher abundance in one group should have more identified spectra than the protein with lower abundance in another. Several methods have been proposed and applied which take advantage of the relationship between the spectral counts and protein abundance to detect differential expression. The primary goal of a differential expression analysis is to detect as many truly differentially expressed proteins as possible (reducing the number of false negatives or type II errors) while controlling for the number of false positives (type I errors). As label-free methods can quantify hundreds to thousands of proteins, multiple testing corrections must be applied to differential expression

☆ This article is part of a Special Issue entitled: Computational Proteomics.

* Corresponding author at: Duke-NUS Graduate Medical School Singapore, 8 College Road, 169857 Singapore, Republic of Singapore.
E-mail addresses: sarah.r.langley@duke-nus.edu.sg, s.langley@imperial.ac.uk (S.R. Langley).

http://dx.doi.org/10.1016/j.jprot.2015.07.012
1874-3919/

analyses to control the number of false positives or type I errors. One approach is to control the false discovery rate (FDR), which is the expected proportion of false positives within a set of significantly differentially expressed proteins. For example, if one had 100 proteins which are detected as differentially expressed at a 5% FDR, five of them are expected to be false positives. This is a separate FDR measure than the one associated with protein inference and identification.

In this study, we chose seven methods for identifying significant differences in spectral count based protein expression. These methods were chosen from the literature and included methods originally proposed for differential expression analysis in microarrays and RNA-seq as well as those specific to proteomics. We included the significance analysis of microarrays (SAM) [7] and the normalised spectral abundance factor coupled with a power law global error model [8]; both methods were designed for gene expression microarray data and have been used for the analysis of label-free MS proteomics [9]. The spectral index (SpI) [10] and QSpec [11] methods were included as methods which were developed specifically for spectral count quantification and have been used in several studies [12,13]. Others have now taken advantage of the methods developed for RNA-seq experiments and applied them in spectral count proteomic studies [14,15], so we have also included the DESeq method [16]. Finally, we included the *t*-test and normalised spectral abundance factor (NSAF) [17] coupled with the *t*-test. The *t*-test is one of the most commonly used statistical tests and has been used to detect differential protein expression [18].

To evaluate these methods, we used 2000 simulated datasets as well as data with a known spike-in difference from the CPTAC standards assessment [19,20]. We investigated the ability of these seven methods to identify differential expression with respect to several different measures; 1) effect sizes, or the percentage of abundance difference, 2) proportion sizes, or the percentage of proteins within a dataset that are differentially expressed and 3) at two levels of multiple testing corrections. Within this evaluation, we provide insight into the performance of these methods with respect to these measures and suggestions for their use in future proteomic studies.

2. Materials and methods

2.1. Simulated data

Real data from two previously published studies were used as the basis for the simulated data (Fig. 1). The first dataset was from LC-MS/MS study investigating the proteomic changes resulting from the addition of an exogenous matrix metallopeptidase within a population of three cases and three controls [21]. The second was a shotgun proteomic analysis of hibernating arctic squirrels within a population of four cases and four controls [9]. 2000 datasets were simulated — 1000 based on the data from the matrix metallopeptidase study (denoted D1) and 1000 based on the data from the arctic squirrel study (denoted D2). In D1, each of the 1000 datasets consisted of simulated counts from 606 proteins and in D2, each of the 1000 datasets consisted of simulated counts from 3538 proteins. Spectral count data can be modelled as a Poisson distribution where the probability of observing a count, n, with respect to the expected count, γ, is given in Eq. (1).

$$f(n,\gamma) = \frac{\gamma^n e^{-\gamma}}{n!} \tag{1}$$

In our simulations, we set γ to the average spectral count of an individual protein from a real dataset and used it to derive a set of Poisson distributed random deviates to simulate the spectral counts for a given protein. This was to preserve the relationship between the spectral count abundance and protein length, as the number of amino acids is used by several of the statistical methods. To incorporate effect sizes into the simulated data, we randomly sampled 20 simulated proteins and added additional counts to one group as follows

$$\overline{SC}_{i,j} = SC_{i,j} \times (1+p) \tag{2}$$

where $SC_{i,j}$ is the simulated spectral count from the *jth* sample of protein i and p is one of 0.2 (20%), 0.5 (50%), 0.8 (80%), 1 (100%), and 2 (200%). In both D1 and D2, one hundred datasets at each effect level were simulated, resulting in 500 datasets from each. The set of 500 (100

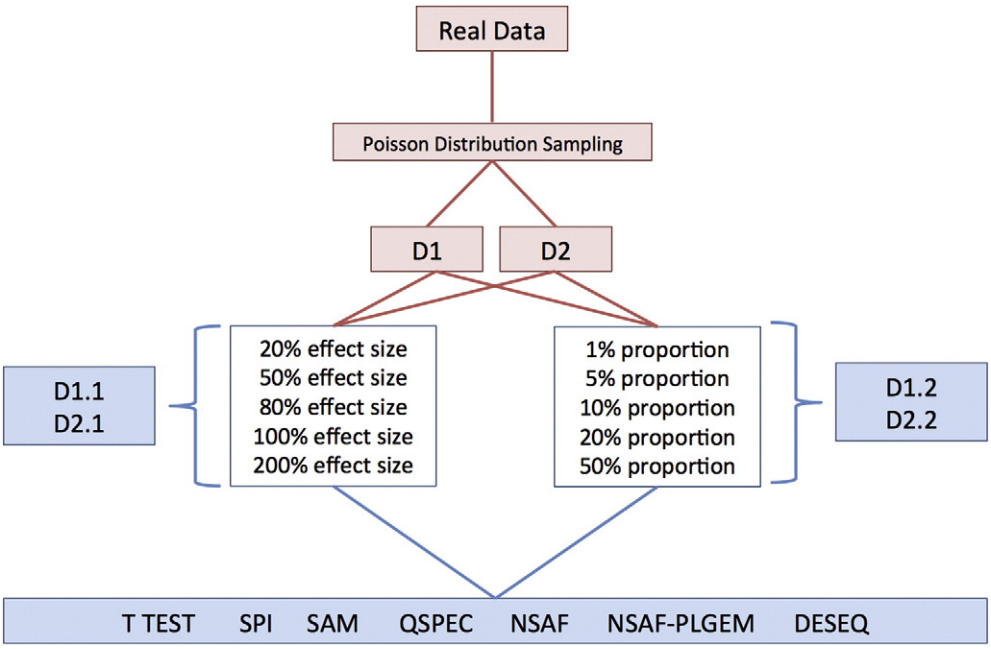

Fig. 1. Simulation data scheme. Overview of the simulated data generation with different effect sizes and different proportions of differentially expressed proteins. *T* TEST — Student's *t*-test; SPI — spectral index; SAM — significance analysis of microarrays; QSPEC — QSpec; NSAF — normalized spectral abundance factor; NSAF-PLGEM — normalized spectral abundance factor-power law global error model; DESEQ — DESeq.

simulated datasets in each of the five effect sizes) from D1 is denoted as D1.1 and similarly, the set of 500 from D2 is denoted as D2.1.

To simulate proportion sizes, five percentages (1%, 5%, 10%, 20%, 50%) of the total number of proteins were randomly sampled from simulated data and for each of those proteins, one of the effect sizes (0.2 (20%), 0.5 (50%), 0.8 (80%), 1 (100%), 2 (200%)) was added following Eq. (2). There were 100 datasets at each of the effect sizes (500 in total) for both D1 and D2. These sets of 500 were denoted D1.2 and D2.2. For D1.2, the number of simulated differentially expressed proteins for each is as follows 1% — 6 proteins, 5% — 30 proteins, 10% — 61 proteins, 20% — 121 proteins and 50% — 303 proteins; for D2.2, the number of simulated differentially expressed proteins for each is as follows 1% — 35 proteins, 5% — 177 proteins, 10% — 355 proteins, 20% — 710 proteins and 50% — 1774 proteins. These 2000 datasets are available at https://zenodo.org/record/19030.

3. CPTAC technology assessment (2006–2011)

Data from the Clinical Proteomic Technologies for Cancer Initiative was used to evaluate the seven differential expression methods with an experimental data set. The CPTAC initiative was established to address the variability and reproducibility issues that are involved in high-throughput proteomics [19,20]. Data from Study 6 was used, where 48 human proteins (Sigma UPS1) at five concentration levels were spiked into yeast samples. mzml files from the LTQ-Orbitrap 56 experiment were downloaded from the CPTAC data portal (https://cptac-data-portal.georgetown.edu/cptac/study/list?scope=Phase+I) and the corresponding fasta file from the Tabb group (http://fenchurch.mc.vanderbilt.edu/misc/20080131-SGD-48-NCI20-BSA-Cntm-reverse.fasta). Spectra to peptide matching was performed using MyriMatch version 2.1 using the default parameters (see Supplemental materials) [22]. Protein inference was performed with IDPicker version 3.1, using a target-decoy strategy with the top-ranked PSM and a parsimonious protein inference algorithm. Again, the default parameters were used (see Supplemental materials) [23]. Briefly, tryptic cleavage defined to be at the carboxyl side of any Lys or Arg, except for those before a Pro and precursor ions were required to be within 10 ppm. A 2% false identification rate and a two spectra minimum were used for protein inference.

4. Statistical methods

The following seven differential expression methods were compared at both a 5% or 1% false discovery rate (FDR) or equivalent for differential expression. The methods were used with their default parameters unless otherwise stated. The ROC curves were visualised using the ROCR package [24].

4.1. Student's t-test

The Student's *t*-test is a hypothesis test that evaluates whether the means from two normally distributed populations are equal. Here, an unpaired, unequal variance, two-tailed Student's *t*-test was used to detect differential expression. The FDR was calculated using the Benjamini–Hochberg method [25]. The implementation was in R.

4.2. Significance analysis of microarrays (SAM)

The significance analysis of microarrays (SAM) was used to detect differential expression. SAM assigns a score to each protein based upon the differences in expression levels relative to the standard deviation and provides a permutation based FDR estimate [7]. Here, SAM was used with the non-parametric Wilcox rank sum statistic as the parametric *t*-test was evaluated independently. The implementation was available in the SigGenes R package in Bioconductor [7].

4.3. Spectral index (SpI)

The spectral index (SpI) method, proposed by Fu et al., calculates a metric (SpI) of protein abundance within each group relative to the number of sample with detectable peptides [10]. The significance of a given SpI is determined empirically via permutation testing. The implementation was in R.

4.4. Normalized spectral abundance factor (NSAF)

The normalized spectral abundance factor (NSAF) values were calculated for each protein identified [17]. Spectral count values of 0 were replaced by an empirically derived fractional value. The value was calculated by determining the smallest value between 0 and 1, which provided the best fit to a normal distribution (Shapiro–Wilks test). The NSAF values were then used with the Student's *t*-test with a Benjamini–Hochberg correction to detect differential expression [25]. The implementation was in R.

4.5. Normalized spectral abundance factor-power law global error model (NSAF-PLGEM)

The NSAF values were calculated for each protein as in the previous section. The NASF values were fit to a power-law global error model and differentially expressed proteins are identified through a permuted signal-to-noise (STN) test statistic, which controls for multiple testing [8]. The implementation was in R and used the PLGEM package in Bioconductor [8].

4.6. QSpec

QSpec calculates differential expression based on a hierarchical Bayes estimation of generalised linear mixed effects model [11]. Per the recommended parameters, 100,000 iterations were used with a 10,000-iteration burn in for the MCMC parameter estimation [11] and the proteins were filtered so that only proteins with counts in two or more samples per group were retained. QSpec version 1.2.2, released as a component of QProt, was used. A z-score thresholding was used to calculate the FDR.

4.7. DESeq

DESeq uses a negative binomial model to test for differential expression in count data using estimates of variance-mean dependence [16]. Due to the low count numbers in proteomics compared to RNA-seq, a Benjamini–Hochberg FDR correction was applied instead of the default [25]. The implementation was in R using the DESeq2 Bioconductor package [16].

5. Results

5.1. Simulated datasets

Here, we used two previously published datasets [9,21] to provide the foundation for the simulation study. Fig. 1 shows the schematic for deriving the simulated data; 1000 datasets were simulated from each with varying levels of effect sizes and proportions of significantly differentially expressed proteins (see Materials and methods). The datasets derived from Stegemann et al. (D1) contained spectral count protein levels across six samples — three in the first group and three in the second group [21]. The first 500 of the 1000 simulated datasets, denoted D1.1, contained 20 differentially expressed proteins at five effect sizes (20%, 50%, 80%, 100%, 200%; 100 datasets at each level, Fig. 1). The second 500 out of the 1000 datasets, denoted D1.2, contained datasets with five different proportions of differentially expressed proteins (1%, 5%, 10%, 20%, 50%; 100 datasets at each level; Fig. 1). The second set of

1000 datasets (D2) was generated from Shao et al. [9]; these simulated datasets contain spectral count levels across eight samples — four in the first group and four in the second. The same simulation schematic was applied, where the first 500 datasets, denoted D2.1, contained 20 simulated or true differentially expressed proteins at five effect sizes and the second 500, denoted D2.2, contained five proportions of differentially expressed proteins (Fig. 1). Each of the seven methods was applied to all 2000 simulated datasets and the significance level was set to the same false discovery rate (FDR) (either 5% or 1%) to ensure comparability across the methods.

Across each set of 100 datasets within the scheme, the number of true positives, false positives, true negative and false negatives returned by each method were averaged to obtain an estimate of the performance of the method. A true positive is where a protein has had an effect size (20%, 50%, 80%, 100%, 200%) added to one group and the given method has detected a significant difference (FDR < 5% or FDR < 1%) between the two groups. In the following text, these averaged values will be used to calculate the sensitivity, specificity and precision as follows,

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

where the *TP* is the average number of true positives, *FP* is the average number of false positives, *TN* is the average number of true negatives and *FN* is the average number of false negatives. Each metric ranges from 0 to 1, with lower values reflecting poor performance.

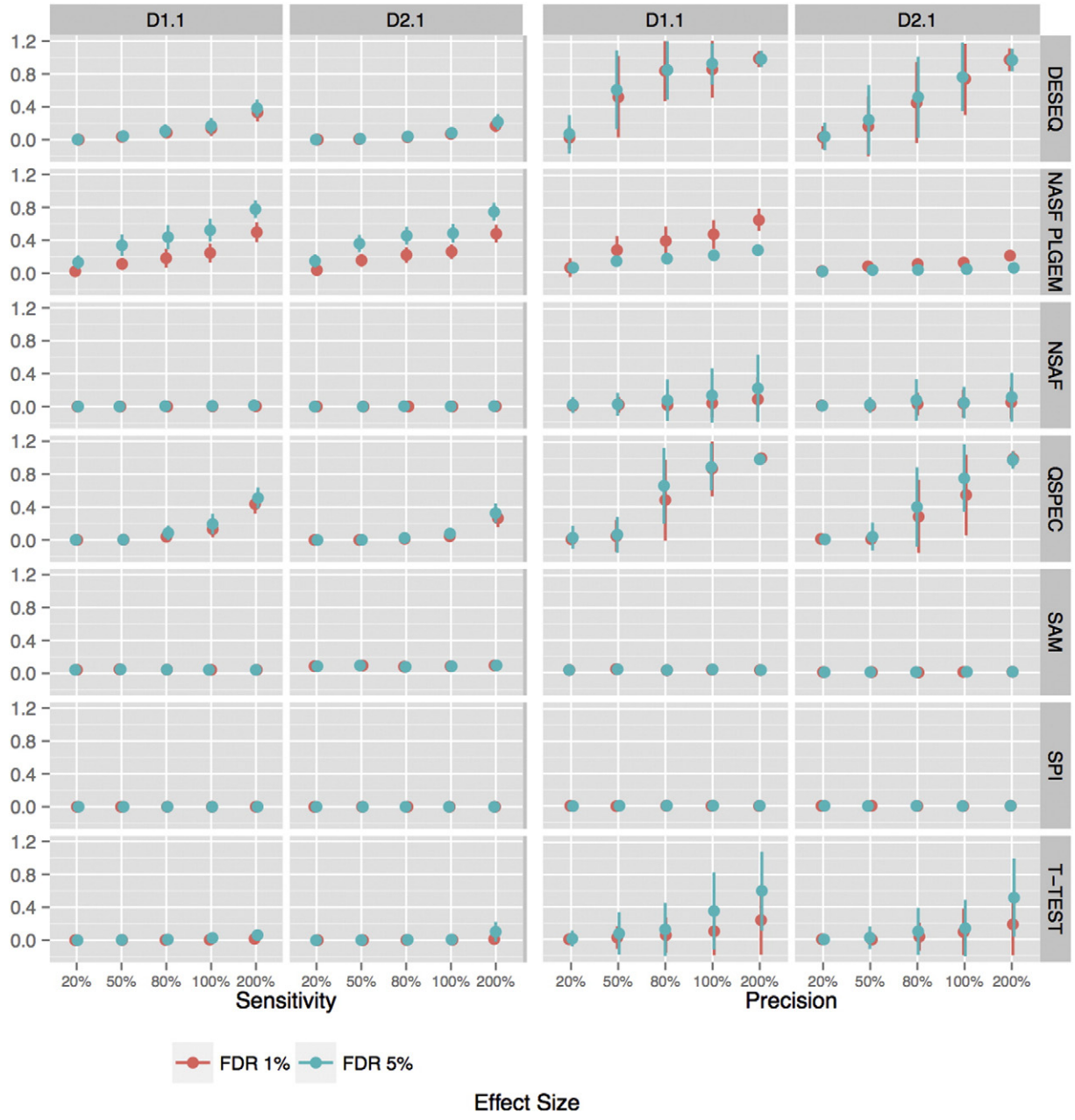

Fig. 2. Sensitivity and precision values for D1.1 and D2.1. The sensitivity and precision values for each the of five effect sizes (20%, 50%, 80%, 100%, 200%) for datasets D1.1 and D2.1. Values are shown for both 1% (red) and 5% (blue) FDR levels; the confidence intervals shown are the standard deviations of the sensitivity and precision values across the one hundred datasets at each effect size.

5.2. Impact of effect size

For each method applied to D1.1, we calculated the sensitivity and precision values for both a 5% FDR (blue) and a 1% FDR (red) level across the seven methods (Fig. 2). For both FDR < 5% and FDR < 1%, the effect size of 20% proved to be at the limit of detection across the methods; the NSAF-PLGEM method was the only one with sensitivity greater than 0.01 (0.13), but at the expense of low precision (0.06) at a 5% FDR. Overall, NSAF-PLGEM identified had the greatest sensitivity (both 5% FDR and 1% FDR) in the five scenarios; however, the corresponding precision values were considerably lower than the other methods (Fig. 2, Supplemental Table S1). The SAM, NSAF and SPI methods were the others to exhibit low precision values across the scenarios, but the sensitivity values did not reach above 0.05 at either a 5% FDR or a 1% FDR level. DESeq and QSpec had the same specificity (1.00) across all five scenarios but DESeq outperformed QSpec with respect to precision and sensitivity at the lower effect sizes (20%, 50% and 80%, Fig. 2 and Supplemental Table S1) while QSpec had a greater sensitivity at the higher effect sizes (100% and 200%, Fig. 2 and Supplemental Table S1). The *t*-test showed high specificity (1.00) and precision (0.73–1.00) across the five effect sizes, but had low sensitivity for each (0.00050–0.061). The receiver-operating characteristic (ROC) curves for the D1.1 with five varying effect sizes (Supplemental Fig. 1a–e; 20%, 50%, 80%, 100%, 200%) illustrate the sensitivity, specificity and area under the curve (AUC) values for each method at each effect size. The SAM method performed poorly with the AUC < 0.50 in each of the effect sizes, and, as the effected sizes increased, the AUC value decreased (Supplemental Table S1).

As in D1.1, the 20% effect size was at the limit of detection in D2.1. For the 20% effect size, both the NSAF-PLGEM and SAM methods had sensitivity greater than 0.01 (Fig. 2), but their respective precision values were both 0.01 at a 5% FDR and 0.01 and 0.02 at a 1% FDR, respectively (Supplemental Table S1). The SpI, SAM and NSAF methods again showed both low sensitivity and low precision values (Fig. 2) for D2.1. ROC curves (Supplemental Fig. 2a–e; 20%, 50%, 80%, 100%, 200%) and corresponding AUC values (Supplemental Table S1) were generated for D2.1. The trends across these methods were in accordance with those found in D1.1 (Supplemental Fig. 1); for six out of seven methods, the AUC values increased with the corresponding increase in effect size, with the SAM method showing the opposite relationship between AUC and effect sizes.

5.3. Impact of proportion size

Here, we varied the proportion of significantly expressed proteins across the five scenarios. Fig. 3 shows the sensitivity and precision of each method across the five proportions for FDR < 5% (blue) and FDR < 1% (red). For each of the methods, the sensitivity values were low and stable across the five proportions and two FDR levels. The precision values, however, varied across the five proportion sizes for each, with the exception of the SpI method. For the NSAF-PLGEM, NSAF, SAM and *t*-test, the increase in precision values followed the increase in the proportion of significantly differentially expressed proteins. For DESeq and QSpec, the highest proportions (20%, 50%) resulted in lower precision values, especially the QSpec method at a 1% FDR (Fig. 3). Supplemental Fig. 3 shows the ROC curves for D1.2 (Supplemental Fig. S3a–e; 1%, 5%, 10%, 20%, 50%). The corresponding sensitivity, specificity and AUC values for each method at each proportion are given in Supplemental Table S2. The AUC values for six out of the seven methods were of a smaller range across proportions than the range across effect sizes (Supplemental Table S2). The SAM method, again, showed a different trend, whereby the AUC values increased as the proportion of differentially expressed proteins increased (AUC = 0.29–AUC = 0.58).

The final set of simulations from Shao et al., D2.2 showed similar trends to those of D1.2 across the sensitivity and precision values (Fig. 3). Supplemental Fig. S4 (a–e; 1%, 5%, 10%, 20%, 50%) and Supplemental Table S2 give the ROC curves and sensitivity, specificity and AUC values. In comparing these AUC values to those from the same proportion levels in D1.2, six of the seven methods showed similar trends. The *t*-test, NSAF, NSAF-PLGEM, SpI and DESeq methods were all of similar ranges across the proportions and the SAM method showed an increase in AUC values as the proportion of differentially expressed proteins increased (AUC = 0.29–AUC = 0.44). The QSpec method, however, deviated from the trend observed in the first dataset. The AUC value corresponding to the final proportion (50%, Supplemental Fig. 4e) was 0.52, which was a decrease from 0.88 in the previous proportion (20%).

5.4. Overlap in methods

We also investigated to what extent the methods overlapped in detecting significant differential expression. The overlap was approximated by consecutively calculating the sensitivity and precision values across the methods within the 100 datasets of a given scenario. The methods were averaged in the order of DESeq, QSpec, *t*-test, NSAF-PLGEM, NSAF, SAM and SpI; the order was based on their individual performances.

The sensitivity and precision values are given in Table 1 for D1.1 and D2.1 and Table 2 for D1.2 and D2.2. For DESeq, QSpec and the *t*-test, the average sensitivity and precision remained stable across the scenarios at both FDR levels. As an example, in the final scenario of D1.1 (FDR < 5%, 200% effect size; Table 1e), the sensitivities for DESeq, DESeq + QSpec and DESeq + QSpec + *t*-test are 0.38, 0.52 and 0.52 with corresponding precision of 0.99 for all three. The increase in sensitivity coupled with high, stable precision indicate that the true differential expression detected by DESeq is also identified by one of the other two, without an increase in the number of false positives. The inclusion of NSAF-PLGEM (DESeq + QSpec + *t*-test + NSAF-PLGEM) increased the average sensitivity from 0.52 to 0.79, but at the decrease in the precision by over two-thirds (from 0.99 to 0.28). The inclusion of the NSAF and SAM methods does not appreciably change the average sensitivity but decreases the average precision (Tables 1 and 2). As the NSAF and SpI methods did not detect many positives, true or false, their inclusion did not alter the average sensitivity or precision. This trend was apparent over the five scenarios in all four datasets, D1.1, D2.1, D1.2 and D2.2 (Tables 1 and 2).

5.5. CPTAC data

Finally, we analysed results from the Clinical Proteomic Technologies for Cancer Initiative Technology Assessment (2006–2011), published by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [19,20], which focused on the pre-analytical and analytical variability intrinsic to proteomic experiments. This CPTAC study was a large-scale investigation involving several institutions with nine individual studies assessing different experimental aspects. We utilised data from the Consortium's Study 6, where the Unbiased Discovery Working Group of the consortium spiked-in human proteins of varying concentrations to evaluate protein identification. 48 human proteins (UPS1) of five varying concentrations were added to a yeast sample consisting of 2522 yeast proteins. For this comparison, we used the results from the LTQ-Orbitrap 56. Using MyriMatch [22] and IDPickers [23] to derive spectral counts from the available spectra data, the LTQ-Orbitrap 56 identified spectra from 43 out of the 48 spiked-in proteins and 1475 out of the 2522 yeast proteins. The yeast samples without the spike-in were used as controls and the yeast plus the 48 protein spike-in samples were considered cases; there were three samples in each group.

We used the yeast proteins combined with the human spiked-in proteins to estimate the type I and type II error rates, where yeast proteins which were identified as significantly differentially expressed can be used as a measure of the false positive rate and the human

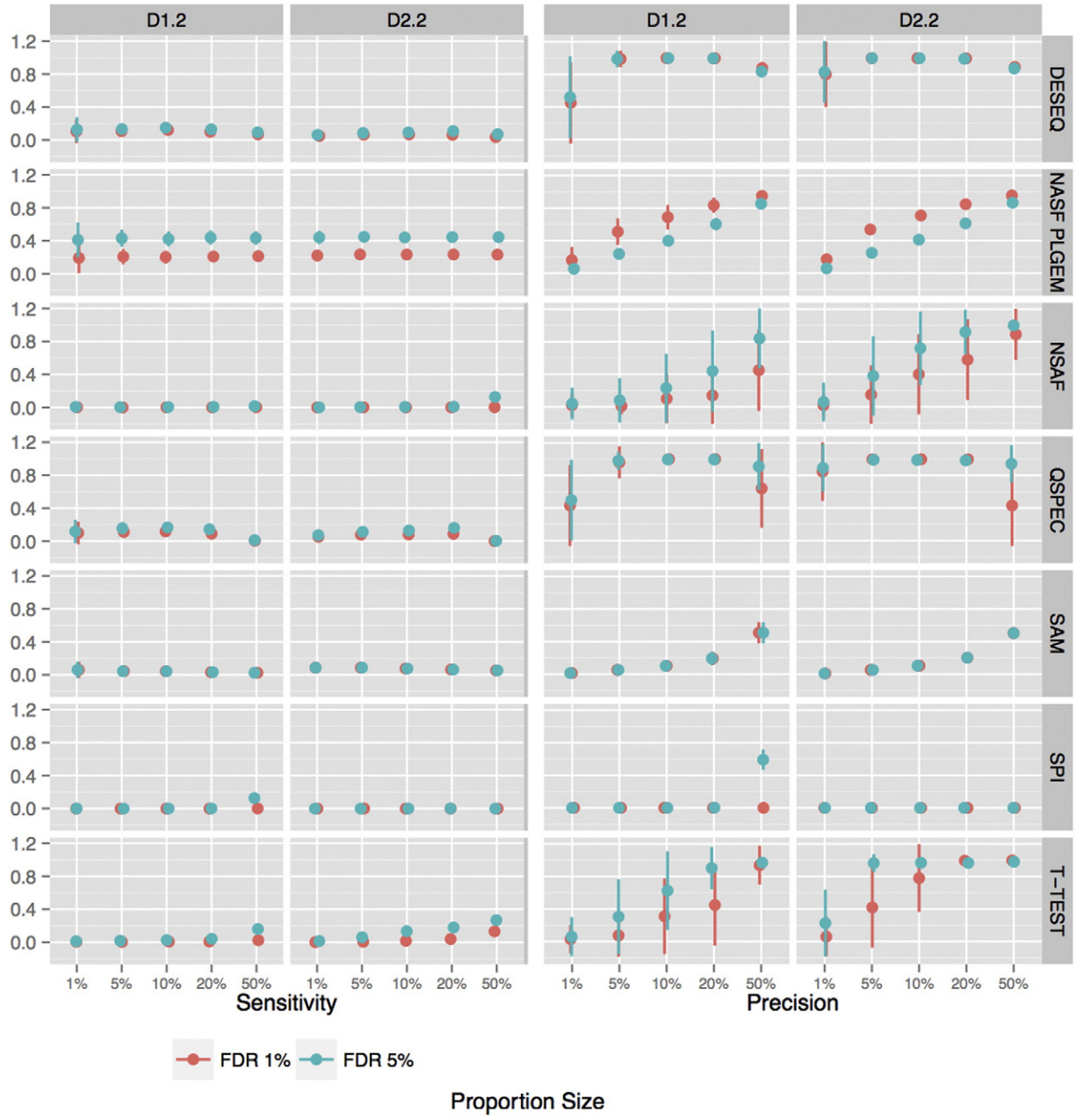

Fig. 3. Sensitivity and precision values for D1.2 and D2.2. The sensitivity and precision values for each the of five proportion sizes (1%, 5%, 10%, 20%, 50%) for datasets D1.2 and D2.2. Values are shown for both 1% (red) and 5% (blue) FDR levels; the confidence intervals shown are the standard deviations of the sensitivity and precision values across the one hundred datasets at proportion size.

proteins which are not identified as significantly expressed can be used as a measure of the false negative rate. As explored in the CPTAC studies [19,20], both the protein identification and differential expression detection were limited at the smallest concentration (0.25 fmol/μL UPS1, Table 3). Only the SAM and NSAF-PLGEM methods detected any true positives (sensitivity of 1 and 0.16, 5% FDR) at this concentration, albeit at the expense of a high false positive rate (precision of 0.25 and 0.07, 5% FDR). At the highest concentration (20 fmol/μL UPS1, Table 3), both the QSpec and the DESeq methods performed reasonably well, with sensitivities of 0.72 and 0.81 and precision of 0.86 and 0.88, respectively. The SAM method identified all 43 proteins at each of the five concentrations and both FDR levels, but did so at the expense of a low precision rate, ranging from 0.25 to 0.16 at a 5% and 1% FDR level. The SpI and NSAF methods failed to detect any differential expression, either true or false positives, at any of the concentration levels investigated.

6. Discussion

Label-free proteomics is becoming a credible alternative for quantifying protein abundance for differential expression analyses, at least in samples of low to medium complexity. Accurate detection of differential expression in these experiments is vital for prioritising downstream validation experiments and for explaining underlying biological phenomena. Given that validation experiments are often costly and labour intensive, detecting proteins that are truly differentially expressed is of utmost importance. Here we have evaluated seven methods for detecting differential expression from spectral count data. These methods were chosen due to their use in the literature and included methods originally proposed for differential expression analysis in microarrays and RNA-seq (SAM, NSAF-PLGEM, DESeq), those specific to proteomics (SpI, NSAF, QSpec) and the commonly used *t*-test. Using the simulated

Table 1
Overlap effect sizes in simulated datasets.

	D1.1				D2.1			
	FDR 1%		FDR 5%		FDR 1%		FDR 5%	
	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
a) 20% effect size[a]								
DESEQ	1×10^{-3}	0.02	3×10^{-3}	0.06	1×10^{-3}	0.02	1×10^{-3}	0.02
DESEQ + QSPEC	1×10^{-3}	0.02	4×10^{-3}	0.08	1×10^{-3}	0.02	1×10^{-3}	0.02
DESEQ + QSPEC + TTEST	1×10^{-3}	0.02	4×10^{-3}	0.08	1×10^{-3}	0.02	1×10^{-3}	0.02
DESEQ + QSPEC + TTEST + NSAF_PLGEM	0.02	0.06	0.13	0.06	0.04	0.02	0.04	0.02
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF	0.02	0.06	0.13	0.06	0.04	0.02	0.08	0.00
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM	0.05	0.03	0.16	0.05	0.08	0.00	0.08	0.00
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM + SPI	0.05	0.03	0.16	0.05	0.08	0.00	0.08	0.00
b) 50% effect size[a]								
DESEQ	0.04	0.52	0.04	0.61	8×10^{-3}	0.16	8×10^{-3}	0.16
DESEQ + QSPEC	0.04	0.52	0.05	0.61	8×10^{-3}	0.16	9×10^{-3}	0.18
DESEQ + QSPEC + TTEST	0.04	0.52	0.05	0.62	8×10^{-3}	0.16	9×10^{-3}	0.18
DESEQ + QSPEC + TTEST + NSAF_PLGEM	0.12	0.29	0.34	0.14	0.16	0.08	0.16	0.08
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF	0.12	0.29	0.34	0.14	0.16	0.08	0.16	0.01
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM	0.12	0.08	0.35	0.11	0.16	0.01	0.16	0.01
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM + SPI	0.12	0.08	0.35	0.11	0.16	0.01	0.16	0.01
c) 80% effect size[a]								
DESEQ	0.09	0.84	0.11	0.85	0.03	0.45	0.04	0.52
DESEQ + QSPEC	0.09	0.84	0.13	0.86	0.03	0.48	0.05	0.58
DESEQ + QSPEC + TTEST	0.09	0.84	0.13	0.87	0.03	0.48	0.05	0.57
DESEQ + QSPEC + TTEST + NSAF_PLGEM	0.21	0.43	0.45	0.18	0.22	0.11	0.46	0.04
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF	0.21	0.43	0.45	0.18	0.22	0.11	0.46	0.04
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM	0.21	0.13	0.46	0.14	0.22	0.01	0.46	0.02
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM + SPI	0.21	0.13	0.46	0.14	0.22	0.01	0.46	0.02
d) 100% effect size[a]								
DESEQ	0.14	0.86	0.17	0.93	0.07	0.74	0.08	0.77
DESEQ + QSPEC	0.16	0.90	0.22	0.93	0.07	0.76	0.10	0.83
DESEQ + QSPEC + TTEST	0.16	0.90	0.22	0.92	0.07	0.76	0.10	0.82
DESEQ + QSPEC + TTEST + NSAF_PLGEM	0.28	0.51	0.54	0.21	0.26	0.12	0.49	0.04
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF	0.28	0.51	0.54	0.21	0.26	0.12	0.49	0.04
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM	0.28	0.17	0.54	0.17	0.27	0.02	0.49	0.02
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM + SPI	0.28	0.17	0.54	0.17	0.27	0.02	0.49	0.02
e) 200% effect size[a]								
DESEQ	0.33	0.99	0.38	0.99	0.17	0.98	0.22	0.98
DESEQ + QSPEC	0.44	1.00	0.52	0.99	0.27	0.99	0.33	0.98
DESEQ + QSPEC + TTEST	0.44	1.00	0.52	0.99	0.27	0.99	0.40	0.98
DESEQ + QSPEC + TTEST + NSAF_PLGEM	0.57	0.68	0.79	0.28	0.50	0.21	0.78	0.06
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF	0.57	0.68	0.79	0.28	0.50	0.21	0.78	0.06
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM	0.57	0.28	0.79	0.22	0.50	0.03	0.78	0.03
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM + SPI	0.57	0.28	0.79	0.22	0.50	0.03	0.79	0.03

[a] The number of differentially expressed proteins in D1.1 and D2.1 is 20.

data, we were able to test how several factors influence the efficacy of statistical methods to detect differential expression in spectral count data. We looked at how each method performs in terms of the effect size and the proportion of total proteins with significant differential expression as well as the total number of identified proteins. We complemented these simulations with spike-in data from the CPTAC standards experiments. While there are considerable differences between the seven methods, there is no singular method which outperformed the rest.

6.1. Effect sizes

We evaluated each of the methods across five effect sizes in two simulated datasets and one real dataset. We started with an effect size of 20% in D1.1 and D2.1, which corresponds to adding 20% of each value in the simulated 'case' group; for the CPATC data, the lowest 'effect size', given in concentrations, was 0.25 fmol/μL UPS1. In both the simulated data and CPATC data at this level, each of the methods had low sensitivity values (Fig. 2, Table 3); the only methods that detected any true differential expression did so at the expense of a high number of false positives. This suggests that if the differences between the groups of samples are expected to be small or marginal, quantification by spectral counts may not be able to provide the resolution to detect them by any statistical method; the technical limitations of the downstream analyses are an important consideration for experimental design. As the experimental technologies develop and the sensitivity and resolution improve, these methods may inherently become more powerful at detecting low levels of differential expression.

As the effect sizes increased, the sensitivity increased for the *t*-test, NSAF-PLGEM, DESeq and QSpec; however, for the simulated data, no method managed to detect all of the differentially expressed proteins at any of the effect sizes at either a FDR < 5% or FDR < 1%. In both the simulated data and the CPTAC data, there is a trade off between the true positives and false positives; this is further illustrated in estimating the overlap between methods. In combining the sensitivity and precision from DESeq, QSpec and the *t*-test, we see that the methods detected true differential expression, as reflected in the sensitivity values, with precision closer to 1 (Table 1). There was an increase in sensitivity when SAM and NSAF-PLGEM were included, but there was a disproportionate decrease in the precision. Estimating overlap is difficult to do with seven methods, and we acknowledge that consecutively averaging the methods is dependent on the order in which the methods are averaged. However, the comparison does provide some insight.

Table 2
Overlap proportion sizes in simulated datasets.

	D1.2				D2.2			
	FDR 1%		FDR 5%		FDR 1%		FDR 5%	
	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
a) 1% proportion[a]								
DESEQ	0.10	0.45	0.13	0.52	0.05	0.80	0.06	0.83
DESEQ + QSPEC	0.12	0.52	0.14	0.59	0.06	0.88	0.08	0.90
DESEQ + QSPEC + TTEST	0.12	0.52	0.14	0.59	0.06	0.88	0.09	0.89
DESEQ + QSPEC + TTEST + NSAF_PLGEM	0.21	0.18	0.42	0.05	0.23	0.17	0.44	0.06
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF	0.21	0.18	0.42	0.05	0.23	0.17	0.44	0.06
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM	0.21	0.04	0.42	0.04	0.23	0.02	0.45	0.03
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM + SPI	0.22	0.04	0.42	0.04	0.23	0.02	0.45	0.03
b) 5% proportion[b]								
DESEQ	0.10	0.99	0.13	0.99	0.06	1.00	0.09	1.00
DESEQ + QSPEC	0.14	1.00	0.18	0.99	0.09	1.00	0.13	0.99
DESEQ + QSPEC + TTEST	0.14	1.00	0.18	0.99	0.09	1.00	0.16	0.98
DESEQ + QSPEC + TTEST + NSAF_PLGEM	0.24	0.55	0.44	0.24	0.24	0.54	0.47	0.26
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF	0.24	0.55	0.44	0.24	0.24	0.54	0.47	0.26
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM	0.24	0.21	0.45	0.19	0.25	0.12	0.48	0.15
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM + SPI	0.24	0.21	0.45	0.19	0.25	0.12	0.48	0.15
c) 10% proportion[c]								
DESEQ	0.12	1.00	0.15	1.00	0.07	1.00	0.09	1.00
DESEQ + QSPEC	0.15	1.00	0.19	1.00	0.09	1.00	0.14	0.99
DESEQ + QSPEC + TTEST	0.15	1.00	0.20	0.99	0.10	1.00	0.24	0.98
DESEQ + QSPEC + TTEST + NSAF_PLGEM	0.24	0.74	0.44	0.41	0.25	0.72	0.51	0.45
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF	0.24	0.74	0.44	0.41	0.25	0.72	0.51	0.45
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM	0.25	0.35	0.45	0.34	0.25	0.24	0.51	0.29
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM + SPI	0.25	0.35	0.45	0.34	0.25	0.24	0.51	0.29
d) 20% proportion[d]								
DESEQ	0.19	1.00	0.13	1.00	0.06	1.00	0.11	0.99
DESEQ + QSPEC	0.12	1.00	0.17	0.99	0.10	1.00	0.19	0.98
DESEQ + QSPEC + TTEST	0.12	1.00	0.18	0.98	0.12	0.99	0.31	0.97
DESEQ + QSPEC + TTEST + NSAF_PLGEM	0.23	0.86	0.45	0.61	0.26	0.86	0.54	0.65
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF	0.23	0.86	0.45	0.61	0.26	0.86	0.54	0.65
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM	0.24	0.55	0.46	0.54	0.27	0.43	0.54	0.49
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM + SPI	0.24	0.55	0.46	0.54	0.27	0.43	0.54	0.49
e) 50% proportion[e]								
DESEQ	0.07	0.88	0.09	0.84	0.04	0.89	0.04	0.99
DESEQ + QSPEC	0.07	0.88	0.09	0.84	0.04	0.89	0.07	0.98
DESEQ + QSPEC + TTEST	0.07	0.88	0.18	0.89	0.14	0.97	0.12	0.97
DESEQ + QSPEC + TTEST + NSAF_PLGEM	0.22	0.92	0.46	0.83	0.32	0.95	0.22	0.65
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF	0.22	0.92	0.46	0.83	0.32	0.95	0.22	0.65
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM	0.23	0.79	0.47	0.80	0.33	0.75	0.22	0.49
DESEQ + QSPEC + TTEST + NSAF_PLGEM + NSAF + SAM + SPI	0.23	0.79	0.47	0.80	0.33	0.75	0.22	0.49

[a] The number of differentially expressed proteins in D1.2 is 6 and in D2.2 is 35.
[b] The number of differentially expressed proteins in D1.2 is 30 and in D2.2 is 177.
[c] The number of differentially expressed proteins in D1.2 is 61 and in D2.2 is 355.
[d] The number of differentially expressed proteins in D1.2 is 121 and in D2.2 is 710.
[e] The number of differentially expressed proteins in D1.2 is 303 and in D2.2 is 1774.

In the CPTAC data, the SAM method detected all 43 differentially expressed proteins at each concentration, but it also detected a large proportion of false positives (Table 3). In accordance with the simulated data, QSpec and DESeq performed reasonably well while SpI and NSAF failed to detect any differential expression. The agreement between the CPTAC data and the simulated data suggests that our simulated datasets via Poisson sampling were representative of experimental label-free proteomic data.

6.2. Proportion sizes

Within D1.2 and D2.2, the increase in the proportion (and total number of differentially expressed proteins) was reflected in the increase of the precision, but not the sensitivity (Fig. 3). Interestingly, the SAM method shows an increase in the precision from the fourth (20%) to the fifth (50%) scenarios (Fig. 3), which is reflected in the increase in the AUC (Supplemental Figs. 4 and 5). Conversely, QSpec and DESeq perform well across the proportion sizes, but there is a decrease in the sensitivity and precision for both at the 50% proportion level. Part of the reasoning behind the QSpec and DESeq methodologies is the ability to address the power lost in small sample sizes by sharing information across all proteins [11,16]. While this approach works well in some scenarios, here, the sharing of information across all proteins may be reducing the power of the model to detect all but the proteins with the greatest differential between the two groups. Pooling information from samples with a high proportion of differential expression to construct a model of no differential expression may influence the model to reject differential expression of smaller effect sizes. Overall, QSpec and DESeq have the highest precision, with stable sensitivity.

6.3. Underlying statistical distributions

The underlying assumptions of the methods seem to play a role in the ability of the methods to detect differential expression and control false positives; both the DESeq and QSpec methods use a distribution appropriate for discrete count data (negative binomial and Poisson) [11,16] rather than normalising the data to approximate a normal distribution as in the NSAF-PLGEM, NSAF and SPI methods. DESeq and QSpec

Table 3
Differentially expressed proteins in CPTAC *^.

		FDR 1%		FDR 5%	
		Sensitivity	Precision	Sensitivity	Precision
TTEST	**A** 0.25 fmol/μL	0.00	0.00	0.00	0.00
	B 0.74 fmol/μL	0.00	0.00	0.00	0.00
	C 2.2 fmol/μL	0.00	0.00	0.00	0.00
	D 6.7 fmol/μL	0.00	0.00	0.00	0.00
	E 20 fmol/μL	0.16	0.88	0.49	0.84
SPI	**A** 0.25 fmol/μL	0.00	0.00	0.00	0.00
	B 0.74 fmol/μL	0.00	0.00	0.00	0.00
	C 2.2 fmol/μL	0.00	0.00	0.00	0.00
	D 6.7 fmol/μL	0.00	0.00	0.00	0.00
	E 20 fmol/μL	0.00	0.00	0.00	0.00
SAM	A 0.25 fmol/μL	1.00	0.25	1.00	0.25
	B 0.74 fmol/μL	1.00	0.13	1.00	0.13
	C 2.2 fmol/μL	1.00	0.13	1.00	0.13
	D 6.7 fmol/μL	1.00	0.19	1.00	0.19
	E 20 fmol/μL	1.00	0.16	1.00	0.16
QSPEC	**A** 0.25 fmol/μL	0.00	0.00	0.00	0.00
	B 0.74 fmol/μL	0.00	0.00	0.00	0.00
	C 2.2 fmol/μL	0.09	0.80	0.14	0.86
	D 6.7 fmol/μL	0.33	0.93	0.42	0.95
	E 20 fmol/μL	0.67	0.88	0.72	0.86
NSAF	**A** 0.25 fmol/μL	0.00	0.00	0.00	0.00
	B 0.74 fmol/μL	0.00	0.00	0.00	0.00
	C 2.2 fmol/μL	0.00	0.00	0.00	0.00
	D 6.7 fmol/μL	0.00	0.00	0.00	0.00
	E 20 fmol/μL	0.00	0.00	0.00	0.00
NSAF PLGEM	**A** 0.25 fmol/μL	0.02	0.12	0.16	0.07
	B 0.74 fmol/μL	0.19	0.35	0.35	0.09
	C 2.2 fmol/μL	0.51	0.55	0.63	0.19
	D 6.7 fmol/μL	0.81	0.67	0.91	0.26
	E 20 fmol/μL	0.93	0.47	0.95	0.19
DESEQ	**A** 0.25 fmol/μL	0.00	0.00	0.00	0.00
	B 0.74 fmol/μL	0.00	0.00	0.00	0.00
	C 2.2 fmol/μL	0.00	0.00	0.02	0.50
	D 6.7 fmol/μL	0.35	0.94	0.44	0.95
	E 20 fmol/μL	0.70	0.85	0.81	0.88

*^ The number of differentially expressed proteins is 43.

perform comparably well with respect to their sensitivity and precision across the varying effects and proportions.

We included the *t*-test as it is one of the most recognised statistical tests and is often included in proteomic software packages, such as Scaffold (Proteome software, Oregon, USA) While not performing as well as QSpec or DESeq, the *t*-test detected true differential expression while controlling the number of false positives, this is despite the spectral count data violating assumption of a normal distribution. Coupling the *t*-test with a normalisation factor (NSAF) actually resulted in lower sensitivity and precision than the *t*-test alone (Figs. 2 and 3).

The SAM method, as implemented here, uses a non-parametric Wilcox rank sum test statistic. The non-parametric method should mitigate the need for normalising to approximate a normal distribution, but it also requires a larger sample size to achieve the same power as a parametric test. Because proteomic experiments are often small in size, we simulated datasets with sample sizes of six and eight, respectively, potentially limiting the power of the SAM method. The SAM method has the option of a (parametric) modified *t*-test rather than the (non-parametric) Wilcox rank sum test. While we did not explicitly test it here, the SAM method coupled with the modified *t*-test may provide more reliable results than SAM with the Wilcox rank sum test.

The SpI method uses an empirical approach to detecting differential expression. At either a 5% or 1% FDR level, the only sensitivity or precision values above zero occurred in D1.2, at a 50% proportion level. The AUC values and gradient of the ROC curves (Supplemental Tables S1 and S2 and Supplemental Figs. S1, S2, S3, S4) suggest that relaxing the FDR threshold may improve the results. We calculated the sensitivity and precision for SpI at an additional FDR level (10%, Supplemental Fig. S5) and found that the relaxed threshold did indeed improve the sensitivity and precision values, but only for D1.1 and D1.2. D2.1 and D2.2 contain approximately five times as many identified proteins as D1.1 and D1.2 which could be driving the difference at the 10% FDR level. These results suggest that the SpI method is influenced to a greater extent by FDR level and the total number of proteins than the other methods.

6.4. Additional statistical and experimental considerations

6.4.1. Sample and group sizes

We did not take into consideration a large variation in sample sizes. Proteomic studies with sample sizes an order of magnitude larger are few, but as technology advances, this may change. Increased sample sizes will increase the power of all methods, but especially non-parametric methods, such as the SAM implementation tested here. We also only considered an experimental design of two groups whereas some may be interested in detecting differential expression between three or more groups. The current implementation of several of the methods selected here only allows for a comparison of two groups; the underlying statistical framework for those may be extended to three or more, but the required implementation is out of the scope of this study.

6.4.2. Experimental protocols

Another important factor, which is addressed to a greater extent in the CPATC Standards Initiative, is the sensitivity and reproducibility of the experimental protocol and of the instrumentation used. For example, out of the 2522 yeast proteins present in the samples, the LTQ-Orbitrap 56 only identified 1475 and, out of the 48 human spiked-in proteins, only 43 were identified. As the sensitivity and range of the mass spectrometers improve and the protocols optimised, the subsequent downstream analyses may also improve and provide opportunities for statistical methods development. The choice of peptide search and protein inference algorithms along with proteomic databases can affect the protein identification and quantification. Mascot [26], SEQUEST [27], X! Tandem [28], and MyriMatch [22] are commonly used search algorithms for spectra matching. Each of the methods varies in the underlying methodology and adjustable parameters (see review by Nobel and MacCoss [29].) Given the recent increase in DNA and RNA sequencing experiments, as well as protein interaction data, several methods and databases have been developed and curated to improve protein inference and the classification of non-unique peptides. Exon–exon junctions, splicing and sequence variants and even novel protein coding genes can be identified using sequencing data (see Wang et al. [30] for a review of data integration for the improvement of peptide and protein identification). Uniprot is one of the most comprehensive and widely used protein databases and is continually being updated and curated [31]. Protein peptide sequences and change and protein IDs can shift or be removed entirely. The protein identification and quantification in an experiment will depend on the database version used; the most up-to-date version will likely produce the most informative results. Finally, large-scale proteome studies, such as the recent tissue-based map of the human proteome [32], may also help to inform new inference algorithms as well as expand proteomic databases.

7. Conclusions

When designing a label-free mass spectrometry experiment, one should also be aware of limitations of the experimental protocol as well as the methods used for downstream analysis. As shown here, the ability of statistical methods to detect differential expression and control false positives is dependent on several factors and varies widely between methods. Several of the methods we evaluated here performed well in terms of detecting true positives while controlling the number of false negatives (DESeq and QSpec) while others detect a higher rate of true positives as the expense of a larger number of false positives (SAM, NSAF-PLGEM). There is no method that outperforms the rest

with respect to 1) effect size, 2) proportion size and 3) influence of the FDR level, but QSpec and DESeq performed comparatively well across the scenarios and we would suggest performing future differential expression analyses with either QSpec or DESeq, or a combination of the two. We would be wary in suggesting that one should use a combination of the results from all seven of the methods, as the high false positive to true positive rate across the seven methods is less than ideal. As illustrated here, spectral count quantification currently works well for larger effect sizes and moderate proportions of differentially expressed proteins. However, advances in the LC-MS/MS instrumentation as well as alternative statistical methodologies may increase the ability to interrogate more complex experimental designs across a wide range of protein abundances.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2015.07.012.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

The authors thank Dr. Nathan Harmston and Dr. Prashant Srivastava for their comments on the manuscript. M. Mayr is a Senior Fellow of the British Heart Foundation (FS/13/2/29892). This work was supported by the Juvenile Diabetes Research Foundation, (17-2011-658) Diabetes UK (12/0004530), the Fondation Leducq (MIRVAD; 13 CVD 02), the National Institute of Health Research Biomedical Research Center based at Guy's and St Thomas' National Health Service Foundation Trust and King's College London in partnership with King's College Hospital and an excellence initiative (Competence Centers for Excellent Technologies — COMET) of the Austrian Research Promotion Agency FFG: "Research Center of Excellence in Vascular Ageing — Tyrol, VASCage" (K-Project Nr. 843536) funded by the BMVIT, BMWFW, the Wirtschaftsagentur Wien and the Standortagentur Tirol.

References

[1] 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, Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents, MCP 3 (2004) 1154–1169.
[2] L. Dayon, J.C. Sanchez, Relative protein quantification by MS/MS using the tandem mass tag technology, Methods Mol. Biol. 893 (2012) 115–127.
[3] M. Mann, Functional and quantitative proteomics using SILAC, Nat. Rev. Mol. Cell Biol. 7 (2006) 952–958.
[4] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, B. Kuster, Quantitative mass spectrometry in proteomics: a critical review, Anal. Bioanal. Chem. 389 (2007) 1017–1031.
[5] M. Wilm, Principles of electrospray ionization, MCP 10 (2011) (M111.009407-M111.009407).
[6] H. Liu, R.G. Sadygov, J.R. Yates, A model for random sampling and estimation of relative protein abundance in shotgun proteomics, Anal. Chem. 76 (2004) 4193–4201.
[7] V.G. Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, PNAS 98 (2001) 5116–5121.
[8] N. Pavelka, M.L. Fournier, S.K. Swanson, M. Pelizzola, P. Ricciardi-Castagnoli, L. Florens, M.P. Washburn, Statistical similarities between transcriptomics and quantitative shotgun proteomics data, MCP 7 (2007) 631–644.
[9] C. Shao, Y. Liu, H. Ruan, Y. Li, H. Wang, F. Kohl, A.V. Goropashnaya, V.B. Fedorov, R. Zeng, B.M. Barnes, J. Yan, Shotgun proteomics analysis of hibernating arctic ground squirrels, MCP 9 (2010) 313–326.
[10] X. Fu, S.A. Gharib, P.S. Green, M.L. Aitken, D.A. Frazer, D.R. Park, T. Vaisar, J.W. Heinecke, Spectral index for assessment of differential protein expression in shotgun proteomics, J. Proteome Res. 7 (2008) 845–854.
[11] H. Choi, D. Fermin, A.I. Nesvizhskii, Significance analysis of spectral count data in label-free shotgun proteomics, MCP 7 (2008) 2373–2385.
[12] S. Cha, M.B. Imielinski, T. Rejtar, E.A. Richardson, D. Thakur, D.C. Sgroi, B.L. Karger, In situ proteomic analysis of human breast cancer epithelial cells using laser capture microdissection: annotation by protein set enrichment analysis and gene ontology, MCP 9 (2010) 2529–2544.
[13] J.M. Elmore, J. Liu, B. Smith, B. Phinney, G. Coaker, Quantitative proteomics reveals dynamic changes in the plasma membrane during *Arabidopsis* immune signaling, MCP 11 (2012) (M111.014555).
[14] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B.M. Beckmann, C. Strein, N.E. Davey, D.T. Humphreys, T. Preiss, L.M. Steinmetz, J. Krijgsveld, M.W. Hentze, Insights into RNA biology from an atlas of mammalian mRNA-binding proteins, Cell 149 (2012) 1393–1406.
[15] A. Castello, R. Horos, C. Strein, B. Fischer, K. Eichelbaum, L.M. Steinmetz, J. Krijgsveld, M.W. Hentze, System-wide identification of RNA-binding proteins by interactome capture, Nat. Protoc. 8 (2013) 491–500.
[16] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (2014) 550.
[17] B. Zybailov, A.L. Mosley, M.E. Sardiu, M.K. Coleman, L. Florens, M.P. Washburn, Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*, J. Proteome Res. 5 (2006) 2339–2347.
[18] J. Barallobre-Barreiro, A. Didangelos, F.A. Schoendube, I. Drozdov, X. Yin, M. Fernández-Caggiano, P. Willeit, V.O. Puntmann, G. Aldama-López, A.M. Shah, N. Doménech, M. Mayr, Proteomics analysis of cardiac extracellular matrix remodeling in a porcine model of ischemia/reperfusion injury, Circulation 125 (6) (2012) 789–802.
[19] D.L. Tabb, L. Vega-Montoto, P.A. Rudnick, A.M. Variyath, A.J.L. Ham, D.M. Bunk, L.E. Kilpatrick, D.D. Billheimer, R.K. Blackman, H.L. Cardasis, S.A. Carr, K.R. Clauser, J.D. Jaffe, K.A. Kowalski, T.A. Neubert, F.E. Regnier, B. Schilling, T.J. Tegeler, M. Wang, P. Wang, J.R. Whiteaker, L.J. Zimmerman, S.J. Fisher, B.W. Gibson, C.R. Kinsinger, M. Mesri, H. Rodriguez, S.E. Stein, P. Tempst, A.G. Paulovich, D.C. Liebler, C. Spiegelman, Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry, J. Proteome Res. 9 (2010) 761–776.
[20] A.G. Paulovich, D. Billheimer, A.J.L. Ham, L. Vega-Montoto, P.A. Rudnick, D.L. Tabb, P. Wang, R.K. Blackman, D.M. Bunk, H.L. Cardasis, K.R. Clauser, C.R. Kinsinger, B. Schilling, T.J. Tegeler, A.M. Variyath, M. Wang, J.R. Whiteaker, L.J. Zimmerman, D. Fenyo, S.A. Carr, S.J. Fisher, B.W. Gibson, M. Mesri, T.A. Neubert, F.E. Regnier, H. Rodriguez, C. Spiegelman, S.E. Stein, P. Tempst, D.C. Liebler, Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance, MCP 9 (2010) 242–254.
[21] C. Stegemann, A. Didangelos, J. Barallobre-Barreiro, S.R. Langley, K. Mandal, M. Jahangiri, M. Mayr, Proteomic identification of matrix metalloproteinase substrates in the human vasculature, Circ. Cardiovasc. Genet. 6 (2013) 106–117.
[22] D.L. Tabb, C.G. Fernando, M.C. Chambers, MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis, J. Proteome Res. 6 (2007) 654–661.
[23] J.D. Holman, Z.-Q. Ma, D.L. Tabb, Identifying proteomic LC-MS/MS data sets with Bumbershoot and.IDPicker, Curr. Protoc. Bioinformatics 37 (13.17) (2012) 13.17.1–13.17.15.
[24] T. Sing, O. Sander, N. Beerenwinkel, T. Lengauer, ROCR: visualizing classifier performance in R, Bioinformatics 21 (20) (2005) 3940–3941.
[25] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. B 57 (1) (1995) 289–300.
[26] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis 20 (1999) 3551–3567.
[27] J.K. Eng, A.L. McCormack, J.R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, J. Am. Soc. Mass Spectrom. 5 (11) (1994) 976–989.
[28] R. Craig, R.C. Beavis, TANDEM: matching proteins with tandem mass spectra, Bioinformatics 20 (2004) 1466–1467.
[29] W.S. Noble, M.J. MacCoss, Computational and statistical analysis of protein mass spectrometry data, PLoS Comput. Biol. 8 (2012) e1002296.
[30] X. Wang, B. Zhang, Integrating genomic, transcriptomic, and interactome data to improve peptide and protein identification in shotgun proteomics, J. Proteome Res. 13 (2014) 2715–2723.
[31] UniProt Consortium, UniProt: a hub for protein information, Nucleic Acids Res. 43 (Database issue) (2015) D204–D212.
[32] M. Uhlén, L. Fagerberg, B.M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C.A.K. Szigyarto, J. Odeberg, D. Djureinovic, J.O. Takanen, S. Hober, T. Alm, P.H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J.M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Pontén, Proteomics. Tissue-based map of the human proteome, Science 347 (2015) 1260419.