OXFORD

Briefings in Bioinformatics, 2017, 1–11

doi: 10.1093/bib/bbx031
Paper

Identification of differentially expressed peptides in high-throughput proteomics data

Michiel P. van Ooijen, Victor L. Jong, Marinus J.C. Eijkemans, Albert J.R. Heck, Arno C. Andeweg, Nadine A. Binai and Henk-Jan van den Ham

Corresponding author: Henk-Jan van den Ham, Dept Viroscience, Erasmus MC, P.O.Box 2040 3000 CA Rotterdam, The Netherlands. Tel.: +31 10 704 4099; Fax: +31 10 704 4760; E-mail: h.j.vandenham@erasmusmc.nl

Abstract

With the advent of high-throughput proteomics, the type and amount of data pose a significant challenge to statistical approaches used to validate current quantitative analysis. Whereas many studies focus on the analysis at the protein level, the analysis of peptide-level data provides insight into changes at the sub-protein level, including splice variants, isoforms and a range of post-translational modifications. Statistical evaluation of liquid chromatography–mass spectrometry/mass spectrometry peptide-based label-free differential data is most commonly performed using a *t*-test or analysis of variance, often after the application of data imputation to reduce the number of missing values. In high-throughput proteomics, statistical analysis methods and imputation techniques are difficult to evaluate, given the lack of gold standard data sets. Here, we use experimental and resampled data to evaluate the performance of four statistical analysis methods and the added value of imputation, for different numbers of biological replicates. We find that three or four replicates are the minimum requirement for high-throughput data analysis and confident assignment of significant changes. Data imputation does increase sensitivity in some cases, but leads to a much higher actual false discovery rate. Additionally, we find that empirical Bayes method (limma) achieves the highest sensitivity, and we thus recommend its use for performing differential expression analysis at the peptide level.

Key words: statistical analysis; high-throughput proteomics; LC-MS/MS; peptide-level data; imputation

Introduction

Proteomics is the large-scale investigation of proteins that is increasingly being used to investigate a range of biological systems at the protein level [1–9]. Major technological advances in the field of mass spectrometry (MS) have been realized over the past few years, including high-throughput proteomics that is used to obtain a comprehensive view of quantitative protein changes in response to disease and treatment [10, 11]. Proteomics can be used to investigate a range of protein-based changes, including identification and characterization of

Michiel P. van Ooijen is a software engineer at GxP Cloud. Previously, he has worked as a bioinformatics researcher at the Department of Viroscience, Erasmus MC, Rotterdam, The Netherlands.
Victor L. Jong is a researcher at the Department of Biostatistics and Research Support, Julius Center, UMC Utrecht, The Netherlands and the Department of Viroscience, Erasmus MC, Rotterdam. His research focuses on statistical analyses of high-throughput data with special interest in predictive analyses.
Marinus J.C. Eijkemans is a professor of biostatistics at the Julius Center for Health Sciences and Primary Care of the University Medical Center Utrecht, The Netherlands. His main research interest is in prediction and classification in high-dimensional 'omics' data.
Albert J.R. Heck (Utrecht University, The Netherlands) is a professor in Biomolecular Mass Spectrometry and Proteomics at Utrecht University. His group focuses on the development of mass spectrometry-based proteomics technologies to study the structure and function of proteins and proteomes.
Arno C. Andeweg is a molecular virologist at the Department of Viroscience, Erasmus MC, Rotterdam, The Netherlands. His main research interest is the induction and regulation of the host response to viral infections.
Nadine A. Binai has a PhD in Molecular Biology. Previously, she has worked as postdoc in the Biomolecular Mass Spectrometry group at Utrecht University, where she was interested in quantitative phosphoproteomics.
Henk-Jan van den Ham is a researcher at the Department of Viroscience, Erasmus MC, Rotterdam, The Netherlands. His research focuses on the investigation of virus–host interactions using high-throughput data and the development and application of bioinformatics methodology.
Submitted: 6 December 2016; **Received (in revised form):** 22 February 2017

specific isoforms of a protein, detecting changes in whole protein expression induced by a specific treatment, or the phosphorylation status of specific position within a protein (phosphoproteomics), or other post-translational modifications [12–15].

Application and integration of proteomics data

Whereas proteomics has until now been a separate discipline, it is now increasingly being combined with other -omics technologies. Previously, we have seen that proteome and messenger RNA profiling provides valuable but distinct information on disease induced by a vaccination-challenge experiment on respiratory syncytial virus [16] (our unpublished observations). By investigating samples with multiple technologies, more accurate predictions on protein isoforms and location can be made, which greatly enhances the understanding of a biological system more than any single technology could. Database deposition of published proteomics and other -omics data sets allows for maximal dissemination and reutilization of raw data, and facilitates integration of proteomics with other -omics data sets. The optimal storage and dissemination of proteomics data are handled by databases including ProteomeScout [17], PRIDE [18], MassIVE (massive.ucsd.edu), jPOST [19] and PASSEL [20]; many of these databases participate in the ProteomeXchange project [21] (for a review, see [22]).

Until now, proteomics analyses have focused mostly on protein-level data. However, peptide-level analysis is increasingly being applied to study a number of sub-protein problems, including gene isoforms, detecting novel somatic mutations and splice variants in the cancer field and post-translational modifications [23] (Figure 1). Recent developments include the integration of genomics and proteomics data generated from a single sample ('proteogenomics'), which is particularly powerful for the identification of specific disease-related gene isoforms and refining gene models in different circumstances [24–28]. Owing to the increasing sensitivity and accuracy of the instrumentation, personalized proteomes have now become possible, which holds great promise for personalized medicine [29–32].

Proteomics data processing

High-throughput or shotgun profiling of a protein sample involves a number of steps, including digestion, fractionation, MS measurement, followed by processing steps to obtain the final protein measurement. Proteins in the samples are denatured and digested into peptides. This large pool of peptides is typically separated into several peptide fractions by liquid chromatography to improve the resolution of the method. The peptide fractions are then quantified by data-dependent acquisition, which is the most prevalent method to produce peptide-level data. The spectra produced provide information on the quantity and sequence of the peptide liquid chromatography–mass spectrometry/mass spectrometry (LC-MS/MS—for an overview, see [24]).

After acquiring the MS that represent the raw proteome data, several computational processing steps are needed to interpret the data: identification, quantification and summarization of the peptides values into protein expressions. Peptide identification is most commonly done using a database of theoretical spectra that is been generated from a database of relevant protein sequences, such as a database of annotated protein sequences, or a proteome predicted from DNA sequences. Identified proteins can be quantified by counting the number of hits to a particular protein in the database as a measure for abundance, i.e. 'spectral counts'. Alternatively, the signal intensity generated by a peptide in the MS can be used to quantify peptide expression level. To compare these levels across experiments, normalization of the values is typically required, in particular for the signal-based data [33, 34]. Subsequently, the counts or signals can be integrated to obtain a protein expression value; the summarization of multiple peptides into a single protein value has the added advantage of reducing the variance and improving the accuracy of the expression estimate [35]. However, for some applications, such as phosphoproteomics, the proteome data cannot be summarized and are interpreted at the peptide level, typically using signal expression data (Figure 1).

The missing value problem

A specific characteristic of signal-based LC-MS/MS approaches is the stochastic nature of the peptide sequencing, resulting in proteins and peptides not being quantified in all treatments and/or replicates of an experiment. As these datapoints are not quantified (i.e. 'missing'), statistical analysis can often only be performed on a small subset of peptides. This problem is exacerbated as the number of conditions in an experiment increases. To overcome this limitation, these missing values can be substituted to alleviate the missing data problem. Substitution of missing values, i.e. data imputation, is routinely applied in the statistical analysis of LC-MS/MS signal intensity data. Imputation can be done by substituting a run-specific background value into all missing values (i.e. assuming that all missing values are below detection limit, 'halfLocal') [36, 37]. Alternatively, missing values are substituted fitting the apparent truncation of the normal distribution, i.e. supplementing the apparent left-censored data from a restricted normal distribution (random tail imputation, RTI) [38, 39]. In addition, there are imputation techniques, such as multiple imputation, that try to estimate the missing value by substituting values from similar datapoints while taking the uncertainty of the imputation into account [40, 41]. Replacing the missing values in the data set with estimates allows more peptides to be tested and will hence increase the 'yield' of an experiment, which translates to a higher sensitivity. However, if these estimates are inaccurate, this approach may lead to incorrect experimental results.

Challenges in the statistical analysis of high-throughput proteomics data

After processing the spectra into quantitative proteome data, differential expression analysis for signal intensity data is typically performed. The processing and analysis of proteomics data have been debated extensively over the past 15 years, and many processing methods are known and used within the field. Spectral count data are now frequently analysed using count-based RNA sequencing methods [33]. For signal intensity data, classical statistical methodology such as analysis of variance (ANOVA) or a *t*-test is often used. These techniques may not be optimal for proteomics data, in particular when combined with imputing with background values [10, 26]. Alternative statistical analysis strategies have been proposed ([42, 43], for a review, see [33]), including the recently introduced MSstats [44] that uses a linear mixed model approach to model MS data. Methods that have been applied successfully in other fields include limma [45], which was developed for the microarray platform,

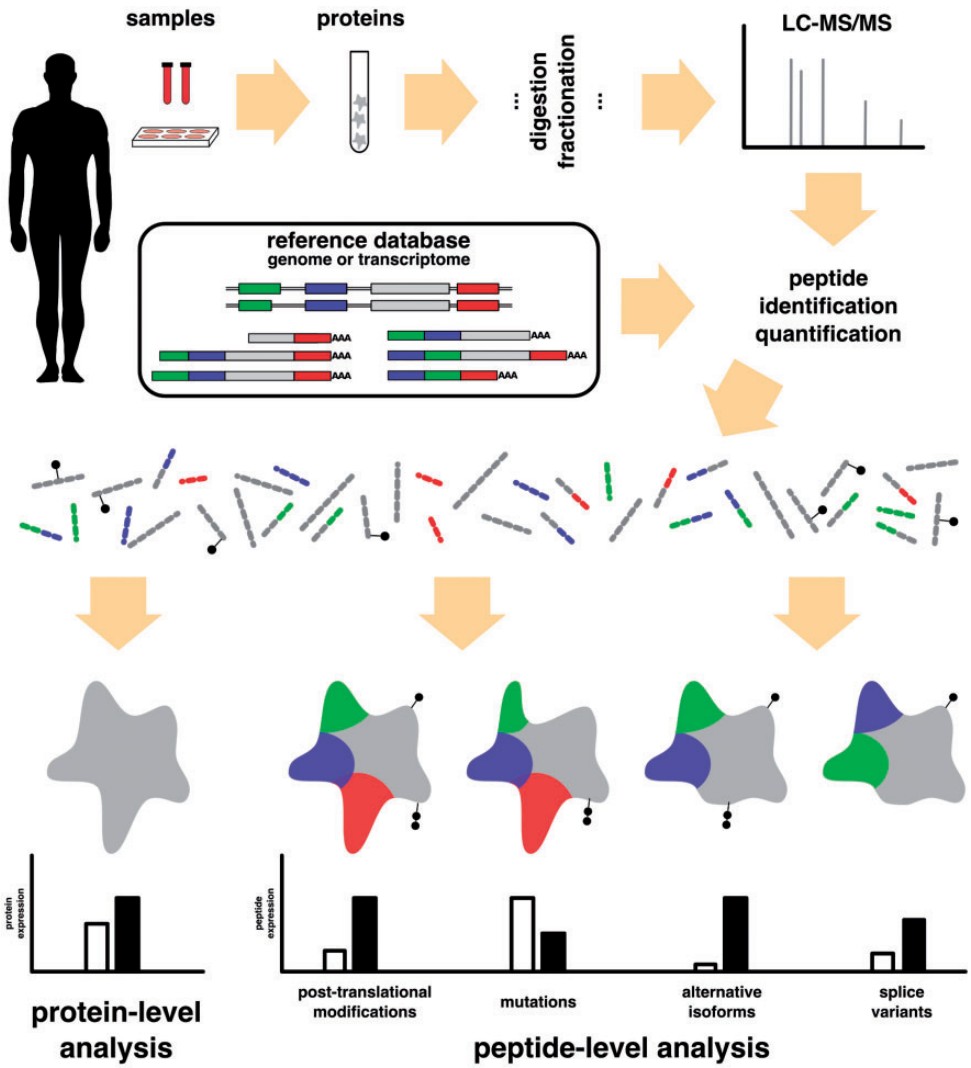

Figure 1. Schematic overview of high-throughput proteomic analysis at the peptide level. Sample protein extracts are digested and (if required) fractionated to obtain peptides. These peptides are subsequently analysed using LC-MS/MS. Peptides are identified by using a reference database of protein sequences. This database contains known or putative sequences, or, in the case of proteogenomics, proteins predicted from the genome or transcriptome of the same individual. Protein-level expression analysis provides information on the protein as a whole; peptide-level analysis provides additional insight into mutations, splice variants, alternative isoforms and a range of post-translational modifications.

but which is now also used for other types of -omics data, including high-throughput proteomics data [46]. Validation of these techniques has been difficult, especially for peptide-level data interpretation. In particular, the application of imputation techniques in combination with statistical analysis has not been studied. Furthermore, the lack of gold standard data sets with a sufficient number of replicates hampers benchmarking efforts. As the volume of data and the reuse of data from protein databases increase, the challenge to establish good practice data processing and analysis methods from both single experiments and meta-analysis studies remains open.

Aim

In this review, we benchmark peptide-level differential statistical analysis methods combined with different imputation techniques to recommend the 'best' method for signal intensity-based LC-MS/MS peptide-level data. Given that there may well be trade-offs between imputation success and the number of replicates [47], or that a particular imputation technique enhances the performance of a specific statistical analysis method, we evaluated all combinations of several imputation techniques and statistical analysis methods. To evaluate these factors, we use published empirical data sets [15, 34] and introduce a new resampling-based technique that allows for gold standard data sets to be generated and evaluated.

Materials and methods

Statistical methods

Different statistical methods can be used to infer which peptides/proteins are differentially expressed in a LC-MS/MS data set. They differ in the underlying assumptions that are being used and the minimal amount of data required to run the test. Next, we briefly describe the statistical methods we used in this work.

Two-sample t-test

Student's *t*-test or the two-sample test is a statistical test in which the equality of means from two sample populations is tested. The main underlying assumption is that both populations have been drawn from a normal distribution, which can be either group-specific or pooled across the experiment. In this study, pooled variance was assumed. Furthermore, the test is limited to cases in which at least two samples are available in both populations. For three-group comparisons, we use ANOVA, which is the multi-group equivalent of the *t*-test.

Empirical Bayes test (limma)

The empirical Bayes method in the Limma R library is similar to a two-sample *t*-test, except that a moderated *t*-statistic is calculated in which posterior residual SDs replace ordinary SDs [45]. By calculating a trend line on peptide means versus variances, a new variance is interpolated for each individual peptide measured. This effectively squeezes the variance of peptides with similar means towards a common value. Compared with the two-sample *t*-test, statistical interference should be more stable when only few measurements are available. And the test is applicable even to peptides for which only a single measurement is available in each class/treatment. Significant differences between peptides, i.e. *P*-values, are calculated from the moderated *t*-statistic. Please see the Supplementary Material for an example of the application of limma analysis to peptide data.

MSstats

Recently, the MSstats package has been developed for statistical inference of differentially expressed proteins and peptides in LC-MS/MS data [44]. MSstats customizes models generated with the R functions lm and lmer. The particular choice of model and its parameters is automatically chosen based on the experimental design (group comparison or time series) and type of LC-MS/MS data supplied (labelled or label-free). MSstats has advanced functionality, such as roll-up from peptide to protein level. Analysis at the protein level would prohibit comparison with the other methods. Therefore, in this project, MSstats is used in its most basic form, using its default settings for label-free LC-MS/MS data.

Generalized linear model with a gamma distribution. When it cannot be assumed that data are normally distributed, generalized linear models (GLMs) provide a framework, which extends linear models to other types of distributions [48]. In contrast to linear models, a GLM is still appropriate when the variance of a variable, e.g. peptide intensity, has a dependence on the mean. Preliminary statistical analysis of the data sets used here showed that a gamma distribution provided a better fit than did a normal distribution (Supplementary Figure S1C). Hence, a gamma distribution was chosen to provide a better fit on a skewed peptide intensity distribution, i.e. in case the peptide intensity distribution would deviate from the normal distribution.

For all statistical methods used, a fixed false discovery rate (FDR) was calculated using the procedure by Benjamini and Hochberg [49].

Construction of data sets

Assessment of differential peptide expression can be performed by a number of statistical methods. To evaluate the ability of these methods to detect differentially expressed peptides, we generated verifiable (i.e. 'gold standard') data sets by simulation and resampling procedures (Figure 2).

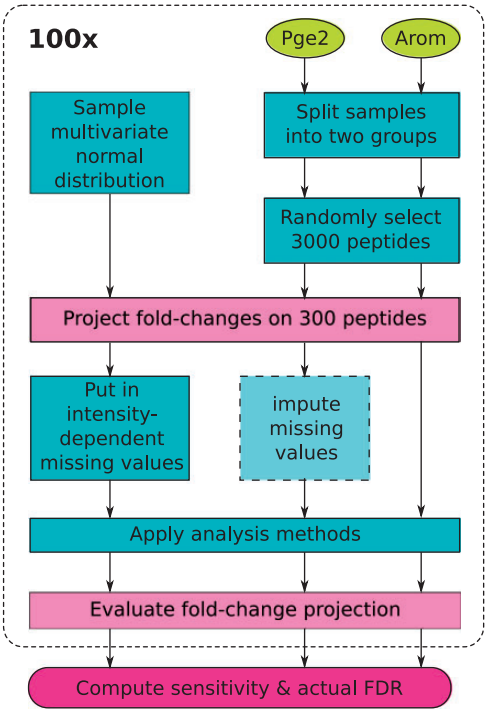

Figure 2. The *ab initio* simulation approach and data set resampling procedures used to generate validation data sets. The PGE2 and AROM data sets are used as input. After simulation/resampling, a fold change is projected onto the data to create differences in peptide expression between groups of samples. Statistical methods of interest can be used to call differentially expressed peptides, which can subsequently be checked against the set of projected changes to compute the sensitivity and actual FDR.

Ab initio proteomics simulation

To have full control on the properties of a simulated proteomics data set and to allow for more biological replicates, an *ab initio* method was developed. Data are simulated by sampling from a multivariate normal distribution. The covariance matrix is constructed from a correlation matrix and a variance vector. The number of runs is two times the number of replicates. Each peptide has a mean sampled from a normal distribution, which then varies across runs according to the covariance matrix. The correlations were sampled from a normal distribution (μ=0; $\sigma = 0.4$), and peptide variances were sampled from an exponential distribution (rate = 8). Three clusters of (up-regulated, down-regulated and noisy) peptides were generated. A range from 2–15 biological replicates was simulated to validate the performance of the four statistical methods. In total, 100 simulations were averaged for each number of biological replicates investigated, and each simulation had 3000 peptides of which 300 were differentially expressed with an absolute $\log_2$ fold change of 1–4, randomly drawn from a uniform distribution [U(1,4)] (for pseudo-code, see Supplementary Material). The resulting data were subjected to a combination of random and intensity-dependent censoring of values to mimic the missing values typically observed in LC-MS/MS data. Although this approach may not perfectly match MS data generation mechanism, its main advantage is that as many samples as required can be generated.

Resampling-based simulation

As our simulation approach does not take into account all sources of variability in real LC-MS/MS data, we also resampled real-life data sets, including the prostaglandin E_2 (PGE2) data set consisting of phosphoproteomic profiling of Jurkat T cells stimulated with PGE_2 [15], and the 'AROM' data set consisting of proteomic profiling of male mice overexpressing human P450 aromatase (AROM) [34]. The PGE2 data set was processed as in the original publication. Briefly, raw data were processed with MaxQuant. MS and MS/MS spectra were searched against concatenated forward-decoy Swiss-Prot *Homo sapiens* database version 2012 09 (40 992 sequences) using the Andromeda search engine. Normalization was performed by subtracting the median of log-transformed intensities for each nano-LC-MS/MS run. The spectra from the AROM data set were searched against the UniProtkB/Swiss-Prot mouse database (16 686 sequences) using the Mascot search engine. Further, filtering and normalization by median scaling were performed in Progenesis 4.0 as described in the original publication [34]. A resampling-based method was developed to retain as many properties of real data as possible. We selected samples from a single treatment group within a data set in such a way that the correlations between peptides and the structure and position of missing values remained unaltered. As we only sample from a single treatment in an experiment, fold changes between treatments cannot play a role. Subsequently, we applied a fold change to part of these samples to create test data sets (Figure 2). Although this approach comes close to real LC-MS/MS data, it is limited by the number of samples in a treatment group (i.e. replicates) that is available in a data set. As we are constructing two artificial test sets from a single treatment group, the maximum test set size is at most half of the number of replicates in the original data set, i.e. an experiment with six replicates can be used to construct two test sets of three samples. All simulation and resampling were carried out 100 times to obtain a robust estimate of methods' performance. Two LC-MS/MS data sets, the PGE2 data set and AROM data set were used as a basis for the resampling procedure (for pseudo-code, see Supplementary Material).

Data imputation

Data imputation is used to postulate the missing values in the data set with estimates that are obtained through a number of different methods. Most of these methods have been applied to MS data and are briefly described below:

halfLocal

The halfLocal (also called localMinimum) method replaces missing values with a run-specific value; here, we use half of the detection limit. All missing values are replaced by this value [37, 50].

Random tail imputation

RTI is a method to provide an estimate for values that are missing because of low expression [38]. These missing values are replaced by random values from a normal distribution around the detection limit. An advantage of this method over halfLocal is that a range of values, rather than a single value, is substituted instead of the missing values.

Multiple imputation

Multiple imputation tries to estimate missing values by substituting values from other samples that are similar [40, 41]. Data were imputed using predictive mean matching method from the mice package [51]. This method uses non-missing peptide values from the same datapoint to identify datapoints that are similar. From these similar datapoints, a value is randomly selected to impute the missing value. This is done multiple times (10 times in our study) to assess the robustness of a single imputation.

Evaluation of statistical and imputation methods

To evaluate the statistical analysis and imputation methods, every comparison was summarized by computing the sensitivity, i.e. the fraction of the 300 differentially expressed peptides that was recovered by an analysis method. For all statistical tests performed, a fixed FDR cut-off of 0.05 was applied. Furthermore, we evaluated the specificity of statistical tests by calculating the actual FDR for each of the 100 simulated or resampled data sets. Because we simulate the protein changes, we can definitively identify a differentially expressed peptide as true positives (TPs) or false positive (FP), and hence determine the actual FDR of a test. Good methods have high sensitivity and an actual FDR that is close to the fixed FDR cut-off that is used (i.e. high specificity). Tukey box plots used to depict the results, where the dark line represents the median, the box represents the interquartile range (IQR) and the whisters represent the last datapoint that is within 1.5*IQR away from the box.

Results and discussion

To assess which statistical analysis methods would be appropriate for LC-MS/MS peptide-level data, we investigated the distribution of our test data sets. Although log-transformed LC-MS/MS data are typically assumed to be normally distributed, this may not always be the case. The simulated *ab initio* data were sampled from a multivariate normal distribution, so is normally distributed as expected (Supplementary Figure S1A). The PGE2 data set is normally distributed, but the AROM data set distribution is better approximated by a gamma function (Supplementary Figure S1B and C). To accommodate these different data distributions, we opted to also evaluate a generalized linear model with a gamma regression approach (GLM-gamma) in addition to the common methodologies applied to peptide-level data, including *t*-tests, limma [45] and MSstats [44].

All methods are validated on simulated LC-MS/MS data sets. These simulated data sets are designed to simultaneously demonstrate the strengths and weaknesses of each method. With a large number of replicates and high fold change between peptides, all methods should be able to perform well. Conversely, when the number of replicates is limited and applying only a low fold change between peptides, all methods are expected to perform poorly.

Comparing two treatments

To evaluate all methods for their performance in a comparison of a treatment and control situation, we first quantify sensitivity of all the methods, given a specific number of replicates. As expected, we see that sensitivity increases as the groups contain more replicates (Figure 3). Although the sensitivity varies per data set, comparing two groups of two replicates does not give appreciable sensitivity in any data set, meaning that two replicates are too little to detect any differentially expressed peptides in these data sets (Figure 3). MSstats shows the lowest sensitivity overall, in particular for the *ab initio* data set. This

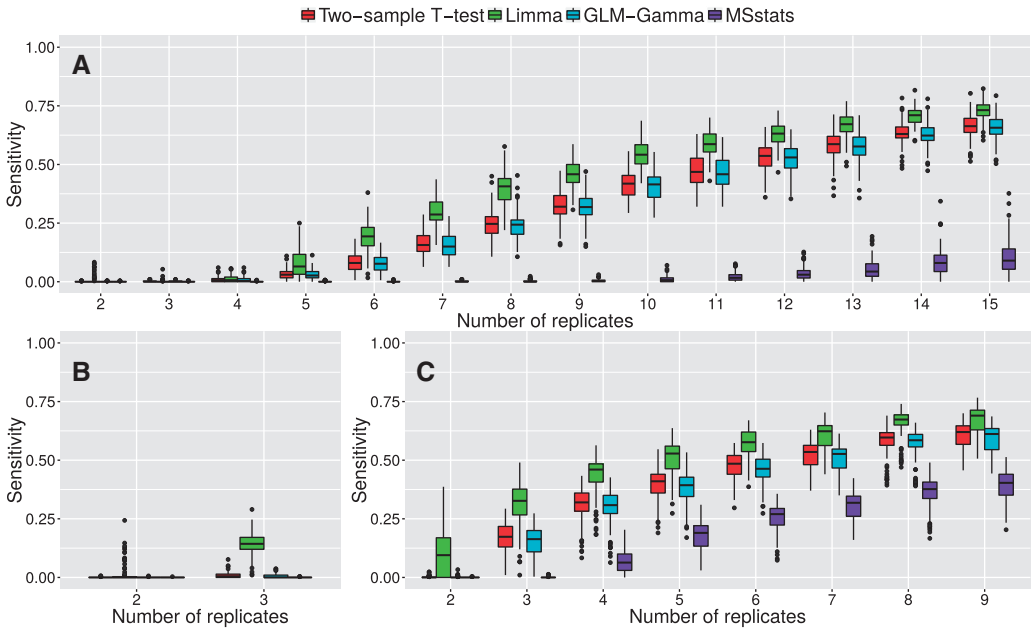

Figure 3. Sensitivity for two-group comparisons, given a particular number of replicates in *ab initio* (**A**), AROM (**B**) or PGE2 (**C**) data. The two-sample *t*-test, limma, GLM-gamma and MSstats were each evaluated 100 times.

suggests that the data structure of the simulated data differs from that of resampled data. Furthermore, we observe that limma achieves the highest sensitivity, in particular when the number of replicates is low. Two-sample *t*-test and gamma regression perform about equally well, but have a lower sensitivity than limma in all comparisons. These findings also hold when the fixed FDR is changed (Supplementary Figure S2), or when the fraction of differentially expressed peptides is varied (Supplementary Figure S3). Sensitivity for limma is particularly high in the case where there are few replicates; as the number of replicates increases, the difference between limma, *t*-tests and gamma regression decreases. This can be explained by the fact that the error sharing among peptides likely improves the accuracy of the peptide variance estimate, which enhances test performance [35]. Furthermore, error sharing allows limma to perform tests with fewer available values than other methods. As expected, these effects diminish as the number of replicates increases.

In addition to evaluating the sensitivity of the different methods, we would also like to check for the correctness of the results (i.e. specificity), given by evaluating the actual FDR. All methods were set to allow for 5% false discoveries (fixed FDR $\leq$ 0.05), and hence all methods should on average have an actual FDR of $\sim$0.05. This is indeed what we observe for most methods and replicates (Figure 4). For two replicates, the actual FDR tends to be much higher, probably because of the low number of differentially expressed peptides identified. For three replicates, all methods tend to control the actual FDR to $\leq$ 0.05, although MSstats does have many instances where the actual FDR is much $>$0.05 in the *ab initio* data set. When these results are combined with the sensitivity, limma is the best method for detecting differentially expressed peptides in a two-group comparison.

The better performance of limma can be explained by the error sharing across peptides, which is a key feature of limma that increases the degrees of freedom as compared with other tests. As it needs fewer degrees of freedom to evaluate a protein, limma should be able to perform tests in peptides that have many missing values, where other methods have too little information. Indeed, we observe that limma is able to test more peptides leading to a higher sensitivity (Supplementary Figure S4).

Multi-group comparisons

In addition to two-group comparisons, many high-throughput proteomics experiments tend to have more treatment groups. Many statistical methods, including those tested here (or their multi-group equivalents), make use of all samples to estimate the peptide variability, and thus a comparison between two groups can be affected by other treatment groups. To evaluate the effect of having more than two treatment groups, we resampled the PGE2 data set as before to have three groups of up to six replicates each. The three groups have differing numbers of differentially expressed peptides between them, with the treatment A versus control and treatment B versus control having 300 and 100 differentially expressed peptides, respectively (Figure 5, inset). Our results show that the statistical analysis methods benefit from having extra treatments, in particular for comparisons with a low number of replicates. For instance, the application of limma to two groups of three samples has a median sensitivity of 0.33, while the three-group comparison has a sensitivity of 0.4. When there are fewer differences between the groups as in treatment B versus control, the sensitivity is lower and the actual FDR is higher, which is expected, given the smaller difference between treatment and control. As before, two replicates are not sufficient for performing statistical analysis in these data sets. In all cases, limma outperforms the other methods in terms of sensitivity, in particular for the case with few replicates (Figure 5). ANOVA analysis does slightly better than GLM-gamma, while MSstats has the lowest sensitivity. In summary, limma is the best statistical analysis method for multi-group experimental designs.

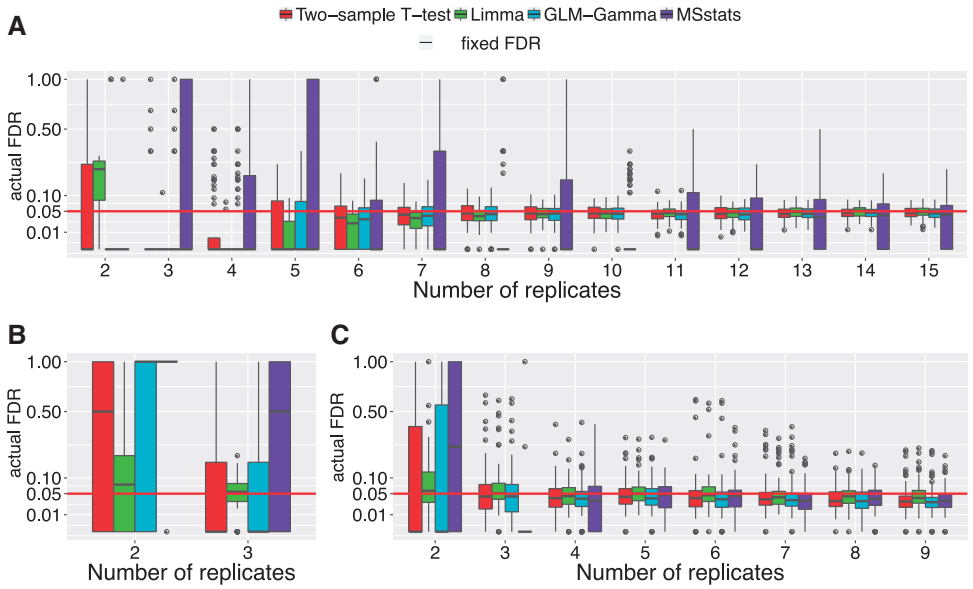

Figure 4. Actual FDR for two-group comparisons, given a particular number of replicates in *ab initio* (**A**), AROM (**B**) or PGE2 (**C**) data. The two-sample *t*-test, limma, GLM-gamma and MSstats were each evaluated 100 times. All methods were evaluated with a fixed FDR set to 0.05 (red line).

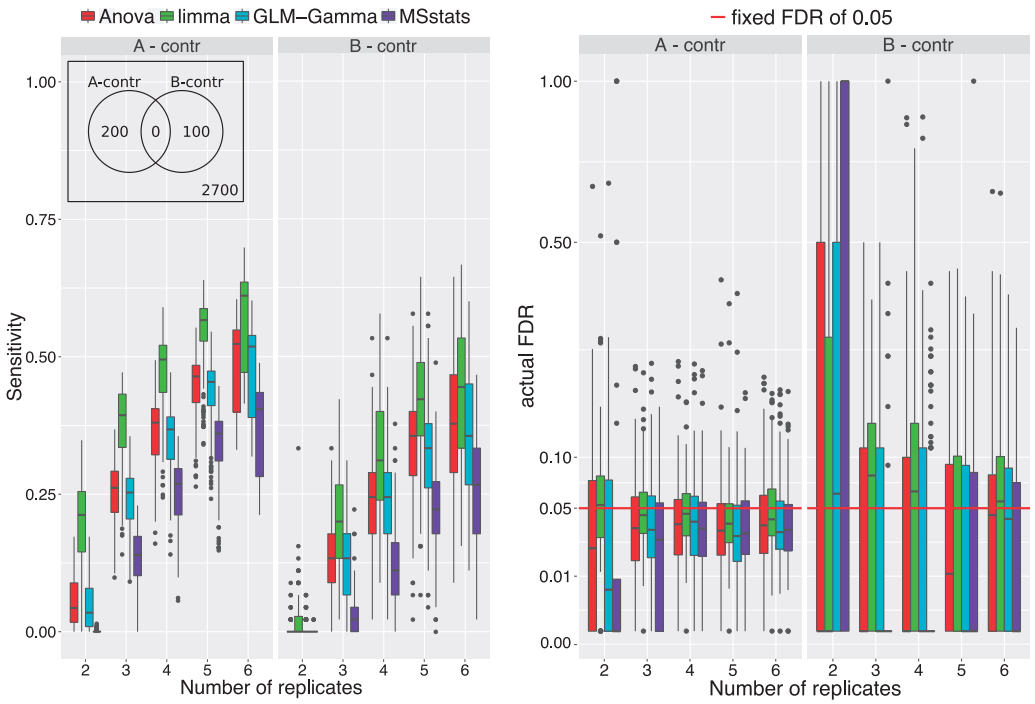

Figure 5. Sensitivity and actual FDR for three-group comparisons, given a particular number of replicates using the resampled PGE2 data set. ANOVA, limma, GLM-gamma and MSstats (red, green, cyan and purple, respectively) were each evaluated 100 times.

Evaluating data imputation

One of the hallmarks of LC-MS/MS data acquired in data-dependent mode is that it has many missing values because peptides are typically not detected in all runs. As a result, none of the evaluated methods recovers all of the differentially expressed peptides projected onto the data sets, even if a large number of replicates are available; for some peptides, there is simply too little data to perform a test on differential expression (Supplementary Figure S4). Although limma manages to perform more tests than the other statistical analysis methods, a large number of peptides cannot be tested for differential expression because of missing values. For instance, a peptide could be highly expressed in only one treatment but not in the other; this would lead to high peptide expression in one treatment and missing expressions in the other, and hence, a test cannot be performed. This has led to the application of imputation techniques to 'fill in' these missing values in the data set. After the application of imputation techniques, all 3000 peptides can be tested by all analysis methods, which is expected to result in an increase in sensitivity.

To test different imputation methods, we used our resampling approach on the PGE2 data set to mimic a normal data set as closely as possible. By applying imputation before performing differentially expression analysis and by evaluating the sensitivity and actual FDR, imputation methods can be compared in terms of increase in sensitivity and control of the actual FDR. Differential expression analysis on data imputed using localMinimum or RTI generally lead to lower sensitivity than when no imputation is applied, regardless of which statistical analysis method is used (Supplementary Figure S5A, B and D). Multiple imputation has a slightly higher sensitivity compared with non-imputed data (Supplementary Figure S5C and D), but the actual FDR is no longer controlled by any of the analysis methods (Supplementary Figure S5G and H).

To evaluate the net effect of imputation on the total number of differential peptides, we looked at the median number of TPs, FPs and false negatives (FNs) after imputation with different methods (Figure 6). We observe that the localMinimum and RTI imputation methods generally do not lead to an increase in the number of differentially expressed peptides found by any statistical analysis method, but lead to a decrease in sensitivity (Supplementary Figure S5A and B versus D). Multiple imputation considerably increases the number of differentially expressed peptides that is identified, regardless of the statistic test that is used. This does not translate to an increase in sensitivity, which only increases marginally (Supplementary Figure S5C versus D). The majority of this gain in differentially expressed peptides is in fact derived from FP hits (red bars). This results in an actual FDR of around 0.2 where 0.05 is used as a cut-off (Supplementary Figure S5G). Overall, we observe that limma finds the highest number of differentially expressed peptides, in particular when the number of replicates is low. In summary, left-tail imputation methods RTI and localMinimum generally do not give better results than when no imputation is applied. Multiple imputation leads to a small increase in the number of TP peptides found, but at the expense of large numbers of FPs (Figure 6).

The effectiveness of imputation techniques has been questioned by several studies that point out the mixed origin of the missing values in LC-MS/MS data [36, 41]. Values can be missing for at least one or two reasons: (1) the peptide may not be selected for quantitation by the MS, which leads to the peptide being absent (missing completely at random, MCAR); or (2) the protein or peptide may not be present or have a low abundance in the sample, and can therefore not be detected (not missing at random, NMAR). Although it is generally thought that Option (1) is less likely compared with Option (2), one cannot determine which of these mechanisms are responsible for a particular missing value. For example, in the case of NMAR, where a peptide is highly expressed in one treatment, but low in another, imputation by filling in a background value will allow a test to be performed and more actual result to be recovered from the experiment. Conversely, if a peptide is expressed at moderate levels across all treatments, but happens by chance to be measured in only one treatment (MCAR), then the application of background imputation techniques will lead to a FP test result. Therefore, the mechanism by which a missing datapoint should be imputed cannot assume either one or the other case, but should account for both MCAR and NMAR. As all variabilities in our test data are derived exclusively from either biological or technical variation to which we have applied fixed fold changes, we are able to definitively identify FPs. With this setup, we see no added value of the application of imputation techniques to LC-MS/MS data.

Conclusions

In this study, we have evaluated several methods for their ability to detect differentially expressed peptides in LC-MS/MS peptide data. Using a simulation and a resampling approach, we are able to definitively measure sensitivity and the actual FDR of both the statistical methods and imputation techniques. We show that limma generally outperforms the other methods while still controlling the actual FDR, both for two- and three-group comparisons. Two replicates are generally insufficient for differential peptide detection, and therefore, we recommend performing experiments with at least three or four replicates. Data imputation leads to a larger number of discoveries in the results, but many are FPs, while the gain in the number of TPs is negligible. Hence, data imputation does not lead to better results in our data resampling evaluation procedure.

While testing the statistical analysis methods and imputation techniques, we tested all combinations of these methods to see if there are particular synergistic imputation technique—analysis method combinations. We have not found such combinations, and we generally see a constant effect of analysis methods, imputation techniques and number of replicates. As expected, we do see that including more replicates leads to a higher sensitivity. We observe that limma has a relatively high sensitivity for a low number of replicates. That is why we recommend limma as first choice for performing peptide-level analysis. For an example of the application of limma on a clinical proteomics data set [52], please see the Supplementary Material.

Evaluation of statistical analysis methods is difficult, as there are a limited number of data sets with known TPs and large biological replicates. We addressed this problem in two ways: (1) by simulating data *ab initio* from a multivariate normal distribution and (2) by using actual proteomics data sets and projecting significant fold changes onto these data sets. Our pure simulation approach illustrates that this approach does not completely capture the structure of real-life data. We think that it is the structure of the missing value distribution in particular that leads to these discordant results between *ab initio* simulations and resampling. In reality, the missing values in LC/MS-MS data are partly induced by intensity (i.e. peptides from non-expressed proteins cannot be measured), and partly by sampling effects (i.e. not all peptides in the sample are measured in every run) and perhaps by other yet unknown mechanisms. Other evaluation methods make use of resampling residuals in linear modelling approaches, but these methods do not account for missing values or any structure therein. Our resampling approach uses only the biological and technical variability from within a single treatment in a data set, which precludes the influence of the original experiment design in our results. By applying synthetic fold changes onto this data set of genuine biological and technical variability, we leave intact the number, position and structure of the missing values, and can hence evaluate the effectiveness of data imputation.

Our results also illustrate that there is variability in performance between resampled data sets derived from the same data set. This illustrates that a method's performance may differ across data sets. However, such performance cannot be verified in a single data set but only in a simulation or

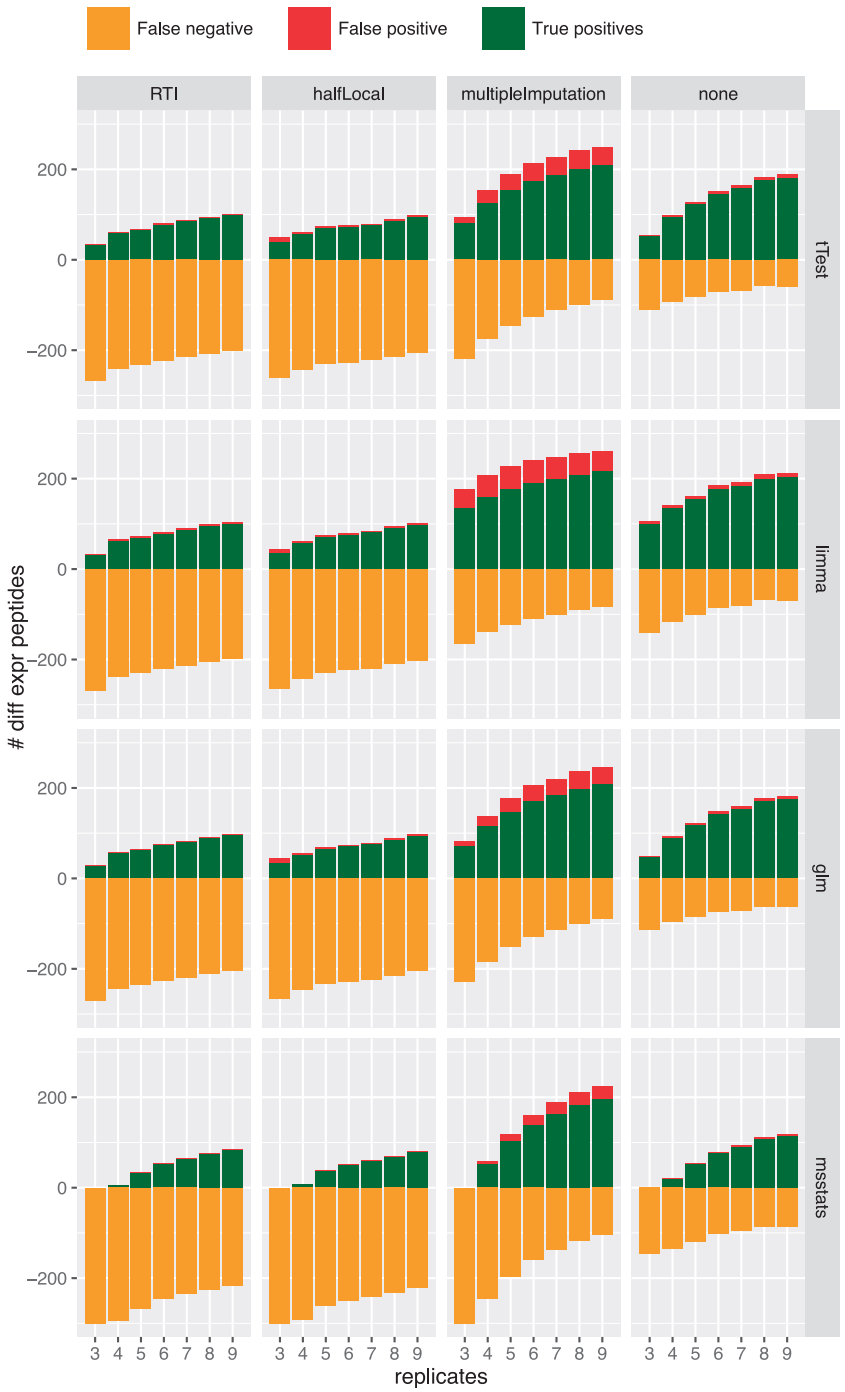

Figure 6. The effect of imputation on the number of TPs (green), FPs (red) and FNs (orange) that the analysis methods detect. As before, there are 300 differentially expressed peptides with a log-fold change of 1–4 of 3000 resampled peptides in two groups of between three and nine replicates. Each bar represents the median of 100 resampled and imputed data sets. The FDR was limited to 0.05, i.e. having 5% FPs (15 peptides) are considered acceptable.

resampling setting such as this one. As such, new methodology being introduced into the field should therefore be validated with an approach such as this to ensure that it generally improves on current practice. Here, we show that the application of imputation to peptide-level data leads to less reliable results. Conversely, the application of limma leads to a major improvement in differential peptide detection, and hence, we recommend its use for the statistical analysis of peptide-level high-throughput proteomics data.

Key Points

- High-throughput proteomics peptide-level data can best be analysed using limma when compared with *t*-tests, ANOVA, gamma regression and MSstats.
- A minimum of three or four replicates is required for achieving acceptable sensitivity and specificity.
- Imputation techniques have no added value, or greatly increase the number of FPs in the results.

Supplementary Data

Supplementary data are available online at BRIBIO online..

Funding

The VIRGO consortium, which is funded by the Netherlands Genomics Initiative and by the Dutch Government (FES0908). The Netherlands Organization for Scientific Research (NWO) supported large-scale proteomics facility Proteins@Work (Project 184.032.201) embedded in The Netherlands Proteomics Centre (to N.A.B. and A.J.R.H).

References

1. Mayya V, Lundgren DH, Hwang S-I, *et al*. Quantitative phosphoproteomic analysis of T cell receptor signaling reveals system-wide modulation of protein-protein interactions. *Sci Sign* 2009;**2**:ra46.
2. Kubach J, Lutter P, Bopp T, *et al*. Human CD4+CD25+ regulatory T cells: proteome analysis identifies galectin-10 as a novel marker essential for their anergy and suppressive function. *Blood* 2007;**110**:1550–8.
3. Brockmeyer C, Paster W, Pepper D, *et al*. T cell receptor (TCR)-induced tyrosine phosphorylation dynamics identifies THEMIS as a new TCR signalosome component. *J Biol Chem* 2011;**286**:7535–47.
4. Filén J-J, Filén S, Moulder R, *et al*. Quantitative proteomics reveals GIMAP family proteins 1 and 4 to be differentially regulated during human T helper cell differentiation. *Mol Cell Proteomics* 2009;**8**:32–44.
5. Satpathy S, Wagner SA, Beli P, *et al*. Systems-wide analysis of BCR signalosomes and downstream phosphorylation and ubiquitylation. *Mol Syst Biol* 2015;**11**:810.
6. Hebert AS, Richards AL, Bailey DJ, *et al*. The one hour yeast proteome. *Mol Cell Proteomics* 2014;**13**:339–347.
7. Kim M-S, Pinto SM, Getnet D, *et al*. A draft map of the human proteome. *Nature* 2014;**509**:575–81.
8. Wilhelm M, Schlegl J, Hahne H, *et al*. Mass-spectrometry-based draft of the human proteome. *Nature* 2014;**509**:582–7.
9. Geyer PE, Kulak NA, Pichler G, *et al*. Plasma proteome profiling to assess human health and disease. *Cell Syst* 2016;**2**:185–95.
10. Zhang Z, Wu S, Stenoien DL, *et al*. High-Throughput Proteomics. *Annu Rev Anal Chem* 2014;**7**:427–54.
11. Geiger T, Velic A, Macek B, *et al*. Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. *Mol Cell Proteomics* 2013;**12**:1709–22.
12. Schmidt A, Kochanowski K, Vedelaar S, *et al*. The quantitative and condition-dependent Escherichia coli proteome. *Nat Biotechnol* 2015;**34**:104–10.
13. Riley NM, Coon JJ. Phosphoproteomics in the age of rapid and deep proteome profiling. *Anal Chem* 2016;**88**:74–94.
14. Olsen JV, Mann M. Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol Cell Proteomics* 2013;**12**:3444–52.
15. de Graaf EL, Giansanti P, Altelaar AFM, *et al*. Single-step enrichment by Ti4+-IMAC and label-free quantitation enables in-depth monitoring of phosphorylation dynamics with high reproducibility and temporal resolution. *Mol Cell Proteomics* 2014;**13**:2426–34.
16. van Diepen A, Brand HK, de Waal L, *et al*. Host proteome correlates of vaccine-mediated enhanced disease in a mouse model of respiratory syncytial virus infection. *J Virol* 2015;**89**:5022–31
17. Matlock MK, Holehouse AS, Naegle KM. ProteomeScout: a repository and analysis resource for post-translational modifications and proteins. *Nucleic Acids Res* 2015;**43**:D521–30.
18. Vizcaino JA, Csordas A, Del-Toro N, *et al*. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* 2016;**44**:D447–56.
19. Okuda S, Watanabe Y, Moriya Y, *et al*. jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Res* 2016;**45**:gkw1080.
20. Farrah T, Deutsch EW, Kreisberg R, *et al*. PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics* 2012;**12**:1170–5.
21. Deutsch EW, Csordas A, Sun Z, *et al*. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* 2016;**45**:gkw936.
22. Perez-Riverol Y, Alpi E, Wang R, *et al*. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* 2015;**15**:930–50.
23. Ruggles KV, Tang Z, Wang X, *et al*. An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Mol Cell Prot* 2016;**15**:1060–71.
24. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Meth* 2014;**11**:1114–1125.
25. Low TY, Heck AJR. Reconciling proteomics with next generation sequencing. *Curr Opin Chem Biol* 2016;**30**:14–20.
26. Kumar D, Bansal G, Narang A, *et al*. Integrating transcriptome and proteome profiling: strategies and applications. *Proteomics* 2016;**16**:2533–44.
27. Locard-Paulet M, Pible O, de Peredo AG, *et al*. Clinical implications of recent advances in proteogenomics. *Exp Rev Prot* 2015;**13**:1–35.
28. Sheynkman GM, Shortreed MR, Cesnik AJ, *et al*. Proteogenomics: integrating next-generation sequencing and mass spectrometry to characterize human proteomic variation. *Ann Rev Anal Chem* 2016;**9**:521–45.
29. Cristobal A, Van Den Toorn HWP, Van De Wetering M, *et al*. Personalized proteome profiles of healthy and tumor human colon organoids reveal both individual diversity and basic features of colorectal cancer. *Cell Rep* 2017;**18**:263–74.
30. Shameer K, Tripathi LP, Kalari KR, *et al*. Interpreting functional effects of coding variants: Challenges in proteome-scale prediction, annotation and assessment. *Brief Bioinform* 2016;**17**:841–62.
31. Zhang B, Wang J, Wang X, *et al*. Proteogenomic characterization of human colon and rectal cancer. *Nature* 2014;**513**:382–7.
32. Malmström L, Bakochi A, Svensson G, *et al*. Quantitative proteogenomics of human pathogens using DIA-MS. *J Proteomics* 2015;**129**:98–107.
33. Blein-Nicolas M, Zivy M. Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics. *Biochem Biophys Acta* 2016;**1864**:883–95.
34. Pursiheimo A, Vehmas AP, Afzal S, *et al*. Optimization of statistical methods impact on quantitative proteomics data. *J Proteome Res* 2015;**14**:4118–26.
35. Ji H, Liu S. Analyzing 'omics data using hierarchical models. *Nat Biotech* 2010;**28**:337–40.
36. Webb-Robertson BJM, Wiberg HK, Matzke MM, *et al*. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res* 2015;**14**:1993–2001.
37. Karpievitch YV, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* 2012;**13(Suppl 1)**:S5.
38. Deeb SJ, D'Souza RCJ, Cox J, *et al*. Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. *Mol Cell Proteomics* 2012;**11**:77–89.

39. Tyanova S, Temu T, Sinitcyn P, *et al*. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 2016;**13**:731–40.
40. Rubin DB. An overview of multiple imputation. *Proc Surv Res Methods Sect Am Stat Assoc* 1988;79–84.
41. Lazar C, Gatto L, Ferro M, *et al*. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *J. Proteome Res* 2016;**15**:1116–25.
42. Nahnsen S, Bielow C, Reinert K, *et al*. Tools for label-free peptide quantification. *Mol Cell Prot* 2013;**12**:549–56.
43. Serang O, Käll L. Solution to statistical challenges in proteomics is more statistics, not less. *J Proteome Res* 2015;**14**:4099–103.
44. Choi M, Chang C-Y, Clough T, *et al*. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* 2014;**30**:2524–6.
45. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;**3**:article3.
46. McClellan EA, Moerland PD, Van Der Spek PJ, *et al*. NetWeAvers: an R package for integrative biological network analysis with mass spectrometry data. *Bioinformatics* 2013;**29**:2946–7.
47. Ryu SY, Qian W-J, Camp DG, *et al*. Detecting differential protein expression in large-scale population proteomics. *Bioinformatics* 2014;**30**:2741–6.
48. Nelder JA, Wedderburn RWM. Generalized linear models. *J Roy Stat Soc* 1972;**135**:370–84.
49. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statis Soc B* 1995;**57**:289–300.
50. Polpitiya AD, Qian WJ, Jaitly N, *et al*. DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics* 2008;**24**:1556–8.
51. Buuren S. v, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw* 2011;**45**, 1–67.
52. Bennike TB, Kastaniegaard K, Padurariu S, *et al*. Proteome stability analysis of snap frozen, RNAlater preserved, and formalin-fixed paraffin-embedded human colon mucosal biopsies. *Data Brief* 2016;**6**:942–7.