Transferred Subgroup False Discovery Rate for Rare Post-translational Modifications Detected by Mass Spectrometry* Ⓢ

Yan Fu‡§ and Xiaohong Qian¶

In shotgun proteomics, high-throughput mass spectrometry experiments and the subsequent data analysis produce thousands to millions of hypothetical peptide identifications. The common way to estimate the false discovery rate (FDR) of peptide identifications is the target-decoy database search strategy, which is efficient and accurate for large datasets. However, the legitimacy of the target-decoy strategy for protein-modification-centric studies has rarely been rigorously validated. It is often the case that a global FDR is estimated for all peptide identifications including both modified and unmodified peptides, but that only a subgroup of identifications with a certain type of modification is focused on. As revealed recently, the subgroup FDR of modified peptide identifications can differ dramatically from the global FDR at the same score threshold, and thus the former, when it is of interest, should be separately estimated. However, rare modifications often result in a very small number of modified peptide identifications, which makes the direct separate FDR estimation inaccurate because of the inadequate sample size. This paper presents a method called the transferred FDR for accurately estimating the FDR of an arbitrary number of modified peptide identifications. Through flexible use of the empirical data from a target-decoy database search, a theoretical relationship between the subgroup FDR and the global FDR is made computable. Through this relationship, the subgroup FDR can be predicted from the global FDR, allowing one to avoid an inaccurate direct estimation from a limited amount of data. The effectiveness of the method is demonstrated with both simulated and real mass spectra. ***Molecular & Cellular Proteomics 13: 10.1074/mcp.O113.030189, 1359–1368, 2014.***

Post-translational modifications of proteins often play an essential role in the functions of proteins in cells (1). Abnormal modifications can change the properties of proteins, causing serious diseases (2). Because protein modifications are not directly encoded in the nucleotide sequences of organisms, they must be investigated at the protein level. In recent years, mass spectrometry technology has developed rapidly and has become the standard method for identifying proteins and their modifications in biological and clinical samples (3–5).

In shotgun proteomics experiments, proteins are digested into peptide mixtures that are then analyzed via high-throughput liquid chromatography–tandem mass spectrometry, resulting in thousands to millions of tandem mass spectra. To identify the peptide sequences and the modifications on them, the spectra are commonly searched against a protein sequence database (6–8). During the database search, according to the variable modification types specified by the user, all forms of modified candidate peptides are enumerated. For each spectrum, candidate peptides (with possible modifications) from the database are scored according to the quality of their match to the input spectrum. However, for many reasons, the top-scored matches are not always correct peptide identifications, and therefore they must be filtered according to their identification scores (9). Finding an appropriate score threshold that gives the desired false discovery rate (FDR)[1] is a multiple hypothesis testing problem (10–12).

At present, the common way to control the FDR of peptide identifications is an empirical approach called the target-decoy search strategy (13). In this strategy, in addition to the target protein sequences, the mass spectra are also searched against the same number of decoy protein sequences (*e.g.* reverse sequences of the target proteins). Because an incorrect identification has an equal chance of being a match to the target sequences or to the decoy sequences, the number of decoy matches above a score threshold can be used as an estimate of the number of random target matches, and the FDR (of the target matches) can be simply estimated as the number of decoy matches divided by the number of target matches. The target-decoy method, although simple and effective, is applicable to large datasets only. When the number of matches being evaluated is very small, this method becomes inaccurate because of the inadequate sample size (13,

From the ‡National Center for Mathematics and Interdisciplinary Sciences, Key Laboratory of Random Complex Structures and Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; ¶State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China

Received April 22, 2013, and in revised form, October 20, 2013
Published, MCP Papers in Press, November 7, 2013, DOI 10.1074/mcp.O113.030189

[1] The abbreviations used are: FDR, false discovery rate; FDP, false positive proportion.

14). Fortunately, for high-throughput proteomic mass spectrometry experiments, the number of mass spectra is always sufficiently large. Current efforts are mostly devoted to increasing the sensitivity of peptide identification at a given FDR by using various techniques such as machine learning (15).

When the purpose of an experiment is to search for protein modifications, the problem of FDR estimation becomes somewhat complex. In fact, the legality of the target-decoy method for modification-centric studies was not rigorously discussed until very recently (16). At present, for multiple reasons, the identifications of modified and unmodified peptides are usually combined in the search result, and a global FDR is estimated for them in combination, with only a subgroup of identifications with specific modifications being focused on. However, the FDR of modified peptides can be significantly or even extremely different from that of unmodified peptides at the same score threshold. There are three reasons for this fact. First, because the spectra of modified peptides can have their own features (*e.g.* insufficient fragmentation or neutral losses), they can have different score distributions from those of unmodified peptides. Second, because the proportions of modified and unmodified peptides in the protein sample are different, the prior probabilities of obtaining a correct identification are different for modified and unmodified peptides. Third, because the proportions of modified and unmodified candidate peptides in the search space are different, the prior probabilities of obtaining an incorrect identification are also different for modified and unmodified peptides. Therefore, the modified peptide identifications of interest should be extracted from the identification result and subjected to a separate FDR estimation, as pointed out recently (16–18).

The difficulty of separate FDR estimations is highlighted when there are too few modified peptide identifications to allow an accurate estimation. Many protein modifications are present in low abundance in cells but play important biological functions. These rare modifications have very low chances of being detected by mass spectrometry. A crucial question is, if very few modifications are identified from a very large dataset of mass spectra, can they be regarded as correct identifications? There was no answer to this question in the past in terms of FDR control. The target-decoy strategy loses its efficacy in such cases. For example, imagine that we have 10 modified peptide identifications above a score threshold after a search and that all of them are matches to target protein sequences. Can we say that the FDR of these identifications is zero (0/10)? If we decrease the score threshold slightly in such a way that one more modified peptide identification is included but find that that peptide is unfortunately a match to the decoy sequence, then can we say that the FDR of the top 10 target identifications is 10% (1/10)? It is clear here that the inclusion or exclusion of the 11th decoy identification has a great influence on the FDR estimated via the common target-decoy strategy. In fact, according to a binomial model (14), the probability that there are one or more false identifications among the top 10 target matches is as high as 0.5, which means that the real proportion of false discoveries has a half-chance of being no less than 10% (1/10). The appropriate way to estimate the FDR of the 10 target identifications is to give an appropriate estimate of the expected number of false identifications among them, and, most important, this estimate must not be an integer (*e.g.* 0 or 1) but can be a real number between 0 and 1. Note that single-spectrum significance measures (*e.g.* *p* values) are not appropriate for multiple hypothesis testing, not to mention that they can hardly be accurately computed in mass spectrometry.

Separate FDR estimation for grouped multiple hypothesis testing is not new in statistics and bioinformatics. A typical example is the microarray data of mRNAs from different locations in an organism or from genes that are involved in different biological processes (19, 20). Efron (21) recently proposed a method for robust separate FDR estimation for small subgroups in the empirical Bayes framework. The underlying principle of this method is that if we can find the quantitative relationship between the subgroup FDR and the global FDR, the former can be indirectly inferred from the latter instead of being estimated from a limited amount of data. The relationship given by Efron is quite general and makes no use of domain-specific information. Furthermore, it requires known conditional probabilities of null and non-null cases given the score threshold. These probabilities are, however, unavailable in the modified peptide identification problem.

This paper presents a dedicated method for accurate FDR estimation for rare protein modifications detected from large-scale mass spectral data. This method is based on a theoretical relationship between the subgroup FDR of modified peptide identifications and the global FDR of all peptide identifications. To make the relationship computable, the component factors in it are replaced by or fitted from the empirical data of the target-decoy database search results. Most important, the probability that an incorrect identification is an assignment of a modified peptide is approximated by a linear function of the score threshold. By extrapolation, this probability can be reliably obtained for high-tail scores that are suitable as thresholds. The proposed method was validated on both simulated and real mass spectra. To the best of our knowledge, this study is the first effort toward reliable FDR control of rare protein modifications identified from mass spectra. (Note that the error rate control for modification site location is another complex problem (22, 23) and is not the aim of this paper.)

MATERIALS AND METHODS

Suppose that we have searched a set of mass spectra against a target-decoy protein database, with one or more types of variable modifications specified. After the database search, each spectrum

TABLE I
Notations used for modeling

Notation	Meaning
T	An event of true peptide identification
F	An event of false peptide identification
I_k	An identification of a peptide carrying modification k, or a k-modified peptide for short
x	The score threshold used to filter identifications
$FDR(x)$	$= P(F\|X > x)$: the global FDR of all (modified and unmodified) peptide identifications with scores greater than x
$FDR_k(x)$	$= P(F\|I_k, X > x)$: the subgroup FDR of k-modified peptide identifications with scores greater than x
$\lambda_k(x)$	$= P(I_k\|T, X > x)$: the probability that a spectrum is identified as a k-modified peptide given that the identification is true and the score is greater than x
$\gamma_k(x)$	$= P(I_k\|F, X > x)$: the probability that a spectrum is identified as a k-modified peptide given that the identification is false and the score is greater than x

was assigned a (possibly modified) peptide along with an identification score. We are now only interested in those identified peptides that carry a specific type of modification, which is denoted by the symbol k, and we aim to estimate the FDR of these modified peptide identifications according to a score threshold. The notations used for modeling are given in Table I.

Note that the definition of the FDR (as a posterior probability) in Table I is called the Bayesian FDR (21), which is equivalent to the original definition of the FDR (10) when there is a large amount of data that are independent and identically distributed. Using the Bayes rule, we have

$$\begin{aligned}
FDR_k(x) &= P(F|I_k,X > x)\\
&= \frac{P(F,I_k|X > x)}{P(I_k|X > x)}\\
&= \frac{P(I_k|F,X > x)P(F|X > x)}{P(I_k|F,X > x)P(F|X > x) + P(I_k|T,X > x)P(T|X > x)}\\
&= \frac{P(F|X > x)}{P(F|X > x) + \dfrac{P(I_k|T,X > x)}{P(I_k|F,X > x)}(1 - P(F|X > x))}\\
&= \frac{FDR(x)}{FDR(x) + \dfrac{\lambda_k(x)}{\gamma_k(x)}(1 - FDR(x))} \qquad \text{(Eq. 1)}
\end{aligned}$$

The whole dataset of the peptide identifications in practice is usually sufficiently large that $FDR(x)$ can always be accurately estimated by using the target-decoy strategy or more complicated approaches (24, 25). However, the number of identifications of k-modified peptides could be too small to support a separate, accurate estimation of $FDR_k(x)$. Fortunately, Equation 1 implies that if $\lambda_k(x)$ and $\gamma_k(x)$ are known, $FDR_k(x)$ can be indirectly inferred from $FDR(x)$ instead of being estimated from limited data. Equation 1 is a bridge that connects the FDRs of modified and unmodified peptides. The following two subsections present methods for estimating $\lambda_k(x)$ and $\gamma_k(x)$, respectively.

Estimation of $\lambda_k(x)$—Here, $\lambda_k(x)$ is the probability that a true identification with a score greater than x is an assignment of a k-modified peptide. This probability is directly related to the proportion of spectra that are produced from k-modified peptides. This proportion is usually unknown because neither mass spectrometers nor experts can distinguish the spectra from differently modified peptides before they are identified. However, it is possible to estimate $\lambda_k(x)$ from the identification results. In fact, we can use the proportion of k-identified peptides among the target identifications as an estimate of $\lambda_k(x)$.

$$\widehat{\lambda_k(x)} = \frac{N_k(x)(1 - FDR_k(x))}{N(x)(1 - FDR(x))} \qquad \text{(Eq. 2)}$$

where $N(x)$ is the number of target identifications with scores greater than x and $N_k(x)$ is the number of target identifications of modification k with scores greater than x. Equation 2 might appear to be strange because there is a term $FDR_k(x)$ in it that is unknown, and this is why we must estimate $\lambda_k(x)$. Fortunately, by replacing $\lambda_k(x)$ in Equation 1 with $\widehat{\lambda_k(x)}$ in Equation 2 we get

$$\begin{aligned}
FDR_k(x) &= \frac{FDR(x)}{FDR(x) + \dfrac{\dfrac{N_k(x)(1 - FDR_k(x))}{N(x)(1 - FDR(x))}}{\gamma_k(x)}(1 - FDR(x))}\\
&= \frac{\gamma_k(x)FDR(x)}{\gamma_k(x)FDR(x) + \dfrac{N_k(x)}{N(x)}(1 - FDR_k(x))} \qquad \text{(Eq. 3)}
\end{aligned}$$

and therefore

$$FDR_k(x)\left(\gamma_k(x)FDR(x) + \frac{N_k(x)}{N(x)}(1 - FDR_k(x))\right) = \gamma_k(x)FDR(x) \qquad \text{(Eq. 4)}$$

or

$$\frac{N_k(x)}{N(x)}FDR_k(x)(1 - FDR_k(x)) = \gamma_k(x)FDR(x)(1 - FDR_k(x)) \qquad \text{(Eq. 5)}$$

which yields an intermediate estimate of $FDR_k(x)$,

$$\widehat{FDR_k(x)} = \frac{N(x)}{N_k(x)}\gamma_k(x)FDR(x) \qquad \text{(Eq. 6)}$$

This estimate leaves $\gamma_k(x)$ as the only unknown factor on the right-hand side of the equation.

Estimation of $\gamma_k(x)$—Here, $\gamma_k(x)$ is the probability that a spectrum will be identified as a peptide with modification k given that the identification is false and the score is greater than x. This probability is closely related to the proportion of candidate peptides with modification k in the search space, and it is reasonable to use this proportion as an estimate of $\gamma_k(x)$. However, the search space of candidate peptides is hidden from the users, and different search engines could use different implementations when generating candidate peptides. For example, search engines could have their own ways of limiting the maximal number of modifications on a candidate peptide to avoid the combinatorial explosion of the search space. Moreover, the proportion of candidate peptides is a constant that is independent of the identification score x, whereas $\gamma_k(x)$ might not be, as shown by Fig. 1. This paper employs a data-driven approach to the estimation of $\gamma_k(x)$. We know that in the target-decoy database search method, all of the matches to the decoy sequences are definitely false identifications. These false identifications constitute natural training data for estimating $\gamma_k(x)$. According to the observations on real data (see the examples in Fig. 1), $\gamma_k(x)$ can be approximated by a linear function of x,

$$\widehat{\gamma_k(x)} = ax + b \qquad \text{(Eq. 7)}$$

where a and b are coefficients to be determined. Given the results of a database search and an arbitrary value of x, calculating the proportion of k-modified peptides among decoy identifications with scores above x is straightforward. The value of x and the calculated proportion constitute a sample of the linear function. Changing the value of x generates a dataset of training samples from which the estimates of the two coefficients in Equation 7, denoted by $\hat{a}$ and $\hat{b}$, can be readily obtained using least-squares regression. As a result, the final estimate of $FDR_k(x)$ becomes

$$\widehat{FDR_k(x)} = \frac{N(x)}{N_k(x)}(\hat{a}x + \hat{b})FDR(x) \qquad \text{(Eq. 8)}$$

This estimate is called the *transferred FDR* for k-modified peptides, indicating that it is derived from the global FDR rather than estimated completely from data.

When the score threshold x is large, the decoy k-modified peptides will be few, and their proportion will be unstable. As shown in Fig. 1, the major part of the observed data shows good linearity, whereas the high-score tail fluctuates seriously. This fluctuation would make the direct separate FDR estimation very inaccurate. For this reason, we cannot use the decoy identifications above the score threshold x to estimate $FDR_k(x)$ directly, as noted in the Introduction. In fact, it is better not to use large x values when fitting the function in Equation 7. The value of $\gamma_k(x)$ for large x (*e.g.* the value that is suitable to be the score threshold) should be extrapolated from the fitted function.

The values of a and b depend not only on the dataset, but also on the searched database and search parameters. Therefore, they should be estimated on the fly for each search. Moreover, from Fig. 1 we can see that the fitted value of a shows different tendencies for different search engines (*i.e.* negative for Mascot, positive for SEQUEST, and close to zero for pFind). This interesting phenomenon implies that Mascot seems to penalize modified peptides, SEQUEST seems to encourage modified peptides, and pFind seems to treat modified and unmodified peptides fairly.

In order to gain an intuitive idea of the methodology, let us consider an example. In this example, 15,100 spectra of known peptide identities, including 100 spectra of phosphorylated peptides, were searched against a target-decoy database (for more details about the data and the experiment, see the simulated data in "Results"). Above the score threshold $x = 37$, there are 3249 target identifications (3205 unmodified and 44 phosphorylated peptide identifications) and 3 decoy identifications (all unmodified peptide identifications). Then, we have the global FDR estimated as $\widehat{FDR(37)} = 3/3249 \approx 0.00092$ and the separate phosphorylation FDR estimated as $0/44 = 0$. By fitting Equation 7 to all of the decoy identifications, we obtain $\hat{a} = -0.01$ and $\hat{b} = 0.6957$, and thus $\widehat{\gamma_k(37)} = -0.01 \times 37 + 0.6957 = 0.3257$. According to Equation 8, the transferred FDR is $\widehat{FDR_k(37)} = (3249/44) \times 0.3257 \times 0.00092 \approx 0.0222$. When we examined the 3249 target identifications with scores above 37, we found that 5 of them were false, including 1 phosphorylated peptide identification. Therefore, the actual false positive proportion (FDP) of the phosphorylated peptide identifications was $1/44 \approx 0.0227$, which is very close to the estimated transferred FDR (*i.e.* 0.0222) but very different from the estimated global FDR (*i.e.* 0.00092) or separate phosphorylation FDR (*i.e.* 0). Note that the actual FDP of all identifications with scores above 37 was $5/3249 \approx 0.0015$, which is close to the estimated global FDR.

In this study, the accuracy of FDR estimation was assessed by comparing the estimated FDR to the actual FDP. Obviously, a better comparison should be made between the estimated FDR and the actual FDR. However, the actual FDR is in general unknown, because it is by definition the expected FDP, which is not computable in the peptide identification problem. In proteomics, FDR and FDP are often meant to be the same thing, but it is important to keep in mind that they are conceptually different.

RESULTS

In order to validate the above algorithm using mass spectral data, we must be able to accurately judge the correctness of identified modifications so as to compare the estimated FDR to the actual FDP. However, the objective judgment is in general absent for complex protein samples extracted from biological organisms. Therefore, we relied on mass spectra generated from (i) theoretical simulation, (ii) synthesized peptides, and (iii) purified standard proteins.

Three methods for estimating the subgroup FDR of modified peptide identifications are compared:

Global FDR: the FDR that is estimated by using the common target-decoy strategy on all peptide identifications, including modification-containing and modification-free ones.

Separate FDR: the FDR that is estimated by using the common target-decoy strategy solely on identifications with modifications of interest.

Transferred FDR: the FDR that is estimated by using the algorithm described in "Materials and Methods."

Three types of modifications were tested: phosphorylation, carbamylation, and acetylation, with an emphasis on phosphorylation. These types of modifications might not be very rare in nature but are used here to validate the methodology. It is expected that the conclusions obtained with respect to them will apply to other types of modifications. In addition, it is worthwhile to note that phosphorylation was defined as a new type of modification (with a new name) with a 79.966331 Da mass shift on the amino acids S, T, and Y and without specification of neutral losses; thus, the search engines did not know it was phosphorylation. In all of the experiments, the most widely used software for peptide identification, Mascot (version 2.2) (7), was used to identify the spectra. Two other search engines, SEQUEST (version 2.7) (6) and pFind (version 2.6) (8, 26, 27), were additionally used to demonstrate the linearity of the $\gamma_k(x)$ function (Fig. 1). The data used in this paper and the program for FDR estimation are available upon request.

Simulated Data—The simulated data were part of the data used in Ref. 16. Five subsets of simulated spectra were used,

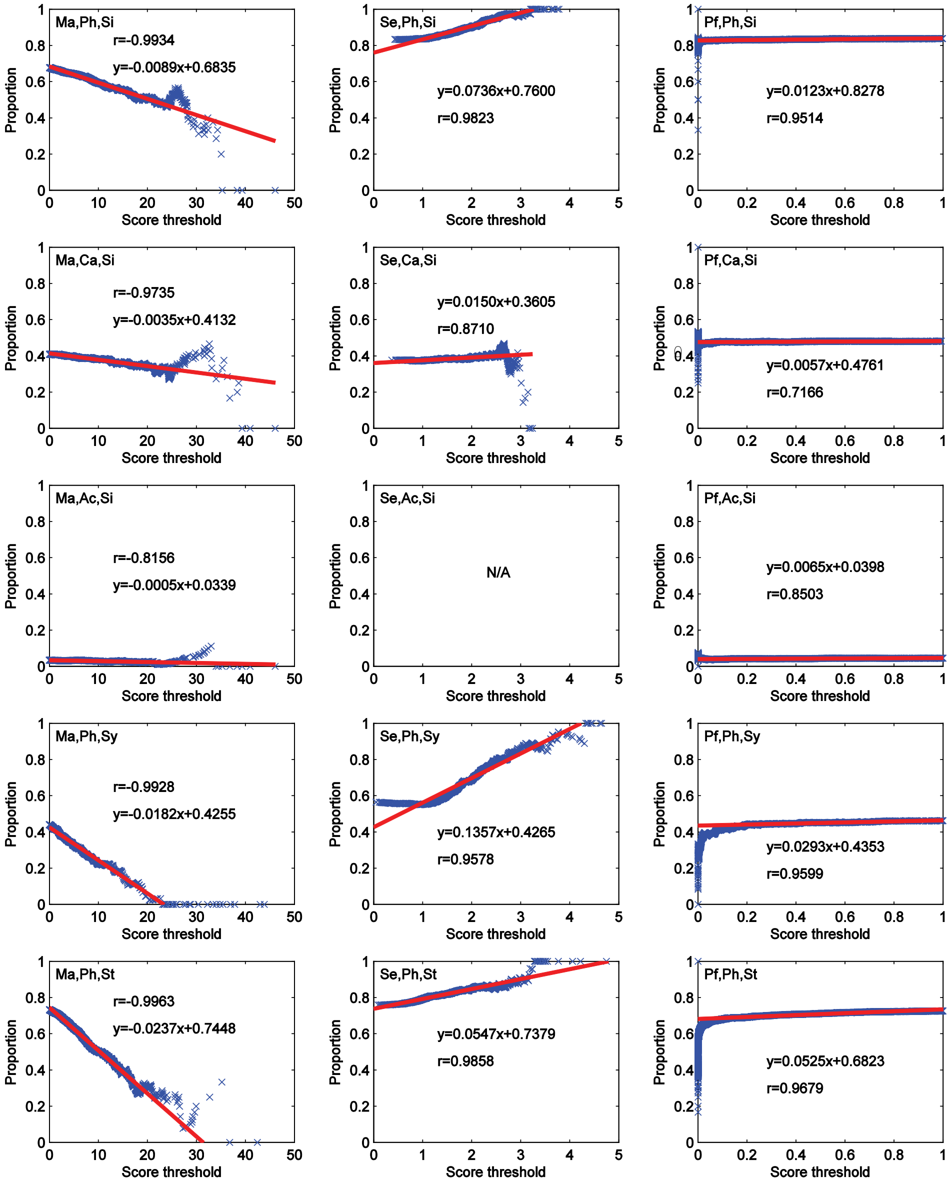


FIG. 1. **Observed proportions of modified peptide identifications among decoy identifications at varying score thresholds and the linear fits.** Correlation coefficients (r) were computed by excluding the high-score tails, which fluctuate stochastically. The abbreviations on the top left corners indicate Mascot (Ma), SEQUEST (Se), pFind (Pf), phosphorylation (Ph), carbamylation (Ca), acetylation (Ac), simulated data (Si), synthesized peptide data (Sy), and standard protein data (St). The scores used were Ion Score (Mascot), Xcorr (SEQUEST), and E-value (pFind). Note that small E-values indicate the significance, and protein N-terminal acetylation is not applicable to SEQUEST.

each of them containing 10,000 mass spectra. The spectra in the first subset ($S_{ph1,10k}$) were predicted from peptides with phosphorylations on S, T, or Y. The spectra in the second subset ($S_{car,10k}$) were from peptides with carbamylations on peptide N termini. The spectra in the third subset ($S_{ace,10k}$) were from peptides with acetylations on protein N termini. The spectra in the fourth subset ($S_{non,10k}$) were from peptides without modifications. The fifth subset ($S_{out,10k}$) included extra spectra from unrelated, unmodified peptides. The protein database contained 100,000 target protein sequences and 100,000 decoy protein sequences. All of the sequences were randomly generated using a Markov chain model trained on the UniProt protein sequence database. The peptides in the first four spectrum subsets, $S_{ph1,1k}$, $S_{car,10k}$, $S_{ace,10k}$, and $S_{non,10k}$, were from the target protein sequences, whereas the peptides in the last spectrum subset, $S_{out,10k}$, were from protein sequences that were not in the database. The peptides that were used for spectrum simulation had unique sequences within each subset and were assumed to carry two charges. Note that if not achieved via simulation, such large-scale nonredundant modification spectra would have hardly been available.

In spectrum simulation, singly charged *b*- and *y*-type fragment ions were predicted, with Gaussian random noises

TABLE II

Results achieved with the three methods for estimating phosphorylation FDRs on simulated data, with the FDR control level set at 1%

n	Global FDR			Separate FDR			Transferred FDR		
	Ave. # of I.D.s (false/all)	Est. error (%)		Ave. # of I.D.s. (false/all)	Est. error (%)		Ave. # of I.D.s. (false/all)	Est. error (%)	
		Mean	S.D.		Mean	S.D.		Mean	S.D.
1	14.14/14.64	−95.69	3.41	1.03/1.43	−67.93	36.74	0.005/0.34	−0.37	5.51
10	14.13/19.15	−73.05	6.72	1.03/5.16	−20.66	13.97	0.008/3.41	−0.18	2.33
100	14.25/64.28	−21.23	2.86	1.04/42.21	−2.44	1.59	0.08/37.85	0.54	0.72
1000	16.20/519.91	−2.12	0.50	5.17/457.44	−0.25	0.27	5.49/460.13	−0.23	0.25
10,000	39.24/5009.82	0.21	0.06	52.27/5136.44	−0.026	0.12	47.73/5100.67	0.06	0.09

Notes: *n*, the number of searched spectra of phosphorylated peptides; Ave. # of I.D.s., average number of false/all identifications of phosphorylated peptides from the target database at 1% estimated FDR; Est. error, FDR estimation error (*i.e.* the estimated FDR minus the actual FDP); Mean and S.D., mean and standard deviation of the estimation errors as a percentage.

added to their theoretical mass-to-charge ratio (*m/z*) values. The intensities of the fragment ions were randomly sampled, with a random proportion of them forced to be zero (to mimic the missing peaks). Additionally, a number of random peaks were added as noise. More details about the data and the database can be found in Ref. 16.

Phosphorylation, peptide N-terminal carbamylation, and protein N-terminal acetylation increase the number of candidate peptides by decreasing degrees. These modifications were first tested individually. In other words, only one of them was set as the variable modification parameter in a database search. For each type of modification, the spectra containing it were added in for searching in increasing numbers ($n = 1$, 10, 100, 1000, 10,000) with the aim of varying the value of $\lambda_k(x)$. For each value of *n*, the error distributions of the three FDR estimation methods were determined by bootstrapping the spectra with 10,000 trials. In each trial, *n* spectra that were randomly sampled from the subset of spectra containing the current type of modification and 15,000 spectra that were randomly selected from the subsets containing no modifications ($S_{non,10k}$ and $S_{out,10k}$) were searched together. Not all of the spectra in $S_{non,10k}$ and $S_{out,10k}$ were used in order to avoid the possibly dominant influence of some special spectra on the result.

Mascot was used to identify the spectra. The precursor and fragment mass matching tolerances were ±3 Da and ±0.5 Da, respectively, and trypsin was specified for *in silico* protein digestion with up to two missed cleavages allowed. After each database search, the modified peptide identifications were filtered according to their scores, and the FDR was estimated using the three methods. Different score thresholds were used for different methods so that the estimated FDR was not more than 1%, and at the same time, the number of identifications was maximized. Finally, the FDR (≤1%) that was estimated via each method was compared with the corresponding actual FDP, and the FDR estimation error was calculated as the former minus the latter. For example, suppose that there are five modified peptide identifications returned above a score threshold in a search, and all of them are target matches, with one of them being a random match (therefore, the FDP is $1/5 = 20\%$). Then, if we use the separate FDR, the estimated FDR for the five modified peptide identifications is $0/5 = 0\%$. Therefore, the FDR estimation error for this case is $0 - 20\% = -20\%$. When the experiment is repeated 10,000 times by bootstrap, we obtain 10,000 FDR estimation errors, which constitute an empirical error distribution.

Tables II through IV compare the three FDR estimation methods for phosphorylation, carbamylation, and acetylation, respectively, in terms of the average numbers of all and false modified peptide identifications as well as the mean and standard deviation of the error distribution. As shown by Table II, when the number (*n*) of phosphorylation spectra was small (mimicking the rareness of protein modifications *e.g.* $n = 1$, 10, or 100), the FDR of phosphorylated peptide identifications was dramatically underestimated by the global FDR. The estimates given by the separate FDRs were better but still deviated greatly from the actual FDP for small *n*. The two methods performed better with increasing *n* and became satisfactorily accurate (a mean error of <1%) when *n* was sufficiently large (10,000 for the global FDR and 1000 for the separate FDR). However, for all of the tested values of *n*, the transferred FDR accurately estimated the FDR of phosphorylated peptide identifications. Moreover, the identification sensitivity was not sacrificed. In other words, significant numbers of phosphorylated peptides were identified with the transferred FDR. Although many more phosphorylated peptides were identified with the global FDR and the separate FDR, these two methods incurred out-of-control error rates.

On carbamylation, the transferred FDR also performed the best among the three methods, as shown by Table III. Again, the estimated global FDR and the separate FDR were on average significantly lower than the actual FDP of carbamylated peptides. With acetylation, the advantages of the transferred FDR were not as clear as with the former two modifications, as shown by Table IV. The transferred FDR performed similarly to the separate FDR. The degraded performance of the transferred FDR for acetylation was possibly due to the small proportion of acetylated peptides in the search space, which limited the accuracy of the estimated $\gamma_k(x)$. In addition, an interesting phenomenon observed is that when the number of acetylation spectra was large, the global FDR seriously overestimated the FDR of acetylated peptide identifications

TABLE III

Results achieved with the three methods for estimating carbamylation FDRs on simulated data, with the FDR control level set at 1% (refer to Table II for definitions of abbreviations)

	Global FDR			Separate FDR			Transferred FDR		
n	Ave. # of I.D.s. (false/all)	Est. error (%)		Ave. # of I.D.s. (false/all)	Est. error (%)		Ave. # of I.D.s. (false/all)	Est. error (%)	
		Mean	S.D.		Mean	S.D.		Mean	S.D.
1	30.15/30.68	−97.27	1.65	0.98/1.36	−71.38	34.94	0.127/0.48	−9.34	27.25
10	30.14/35.48	−84.00	4.11	0.98/4.83	−22.02	13.88	0.12/3.67	−2.82	9.03
100	30.49/83.90	−35.38	3.176	0.99/39.51	−2.51	1.17	1.08/39.46	−2.00	1.369
1000	32.95/571.77	−4.78	0.42	6.83/466.34	−0.59	0.55	8.12/480.59	−0.72	0.27
10,000	52.15/5068.08	−0.03	0.06	78.92/5294.34	−0.50	0.08	76.78/5275.51	−0.47	0.07

TABLE IV

Results of three methods for estimating acetylation FDRs on simulated data, with the FDR control level set at 1% (refer to Table II for definitions of abbreviations)

	Global FDR			Separate FDR			Transferred FDR		
n	Ave. # of I.D.s. (false/all)	Est. error (%)		Ave. # of I.D.s. (false/all)	Est. error (%)		Ave. # of I.D.s. (false/all)	Est. error (%)	
		Mean	S.D.		Mean	S.D.		Mean	S.D.
1	5.41/5.97	−90.16	8.03	1.41/1.90	−70.87	32.04	0.90/1.26	−73.86	33.55
10	5.42/10.99	−49.17	8.37	1.44/6.30	−20.75	13.88	0.90/4.96	−19.99	13.09
100	5.42/61.23	−7.91	1.32	1.43/50.21	−2.66	2.56	2.02/53.50	−2.74	1.30
1000	5.43/566.29	0.03	0.13	14.59/639.71	−1.34	0.26	13.58/635.75	−1.14	0.25
10,000	9.26/5255.26	0.82	0.02	93.91/6749.02	−0.40	0.06	93.64/6750.42	−0.39	0.05

(the estimated FDR was ~1%, whereas the actual FDP was close to zero), and as a result, the number of acetylated peptide identifications was significantly less than with the other two methods. For example, when $n = 10,000$, an average of ~6750 acetylated peptide identifications per trial were obtained with the separate FDR or the transferred FDR, whereas this number was only 5255 with the global FDR, which corresponds to a reduction of 22% in the sensitivity of acetylation detection.

The above experiments were conducted when the FDR control level was set at 1% and only one type of modification was specified in each search. To investigate whether the proposed algorithm worked for more extensive situations, two additional experiments were performed. In the first experiment, the FDR control level was set at 1%, 2%, 3%, 4%, and 5%, and phosphorylation was specified as the variable modification for searches. Fig. 2 depicts box plots of the FDR estimation error distributions; it shows that the transferred FDR was consistently the best among the three methods. Interestingly, the accuracy of the global FDR decreased dramatically with an increasing FDR control level. In the second experiment, phosphorylation, carbamylation, and acetylation were specified as the variable modifications in a single search, and three forms of phosphorylated peptides were considered for FDR control. Fig. 3 shows that the approximate linearity of $\gamma_k(x)$ with regard to the score threshold x held for multiply modified peptides. Table V gives the results of FDR estimation. It is clear that only the transferred FDR accurately estimated the FDRs of all forms of phosphorylated peptide identifications. Note that because every spectrum contained at most one type of modification, any phosphorylated peptide identifications with co-occurring carbamylations and/or acetylations were definitely false identifications.

Synthesized Peptide Data—The second dataset was from the iPRG 2012 study conducted by the Proteome Informatics Research Group of the Association of Biomolecular Resource Facilities. The goal of this study was to evaluate the data analysis capabilities of proteomics researchers in identifying post-translational modifications present at substoichiometric levels within a complex peptide-mixture background. The background of the sample was a proteome lysate of yeast. Synthesized tryptic peptides from non-yeast proteins were spiked into the tryptic yeast lysate. The spiked-in peptides carried a variety of modifications including phosphorylation. The peptide mixture was analyzed by an AB Sciex 5600 TripleTOF mass spectrometer (AB SCIEX, Concord, ON) interfaced with a Waters nanoAcquity UPLC system (Waters, Millford, MA), and a total of 18,009 tandem mass spectra were produced. Twenty-four participants analyzed the data and submitted their results, from which the organizers compiled a list of consensus spectra identified by three or more participants agreeing on peptide sequences. In the current study, 7877 spectra of unmodified yeast peptides and 179 spectra of spiked-in phosphorylated peptides were extracted from the consensus spectra for the experiments. Charges of two and three were assumed for spectra that had undetermined charge states. Other types of modifications were not considered because they had too few consensus spectra for statistical analysis. More details about the dataset and the study can be found at the Association of Biomolecular Resource Facilities website.

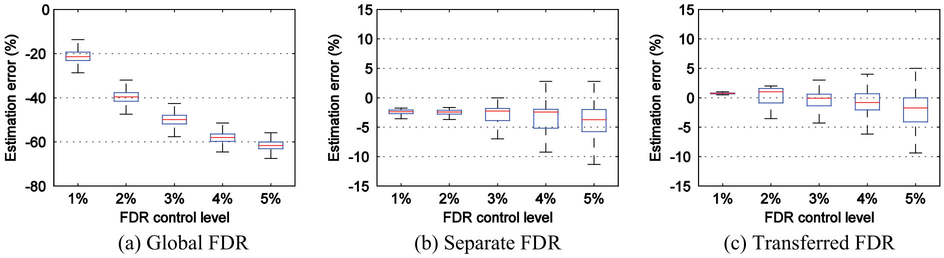


FIG. 2. **Box plots of FDR estimation error distributions for the three methods at different FDR control levels for phosphorylated peptide identification on simulated data.** The number of phosphorylation spectra was 100 in each search.

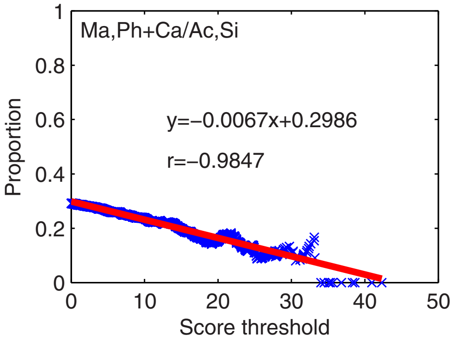


FIG. 3. **Observed proportions of multiply modified peptides (phosphorylation plus carbamylation and/or acetylation) among decoy identifications at varying score thresholds and the linear fit.**

The spectra were searched against the UniProt protein database catenated with the reverse sequences as decoys. Note that the small database provided by the iPRG 2012 study was not appropriate for search here, because it would have resulted in correct identification of almost all of the consensus spectra, which would have made the FDR analysis impossible. The precursor and fragment mass matching tolerances were ±20 ppm and ±0.05 Da, respectively, and trypsin was specified for protein digestion. After the database search, only peptides of more than eight amino acids were reserved in order to avoid short peptides occurring in both forward and reverse sequences in the enormous UniProt database. Similar to the experiments on simulated data, the spectra were sampled for 10,000 trials to generate the error distributions of FDR estimation. In each trial, 30 phosphorylation spectra and 5000 modification-free spectra were randomly selected for search. The FDR control level was set at 1% for all of the FDR estimation methods, and the error was computed as the estimated FDR minus the actual FDP. Fig. 4 shows the error distributions obtained with the three methods. Similar to the results on simulated data, with the global FDR or the separate FDR, the FDR of phosphorylated peptide identifications was greatly underestimated. With the transferred FDR, the FDR was slightly overestimated (within 1%) in most cases and was underestimated by ~5% in a few cases. On average, 14 phosphorylated peptides were identified per trial with the transferred FDR. Obviously, the transferred FDR performed best among the three methods. When the FDR control level was set at other values (*e.g.* 5%), similar results were observed (data not shown).

Standard Protein Data—The third dataset was from the widely used Standard Protein Mix Database, which is a diverse mass spectrum dataset designed for testing peptide and protein identification software tools (28). The proteolytic peptides in a mixture of tryptic digests of 18 purified standard proteins were analyzed using different mass spectrometers and under various conditions. The data used in this paper were from an analysis of the third mixture (mix3) on a Thermo Scientific LTQ-FT mass spectrometer (Thermo Scientific, Bremen, Germany). Ten liquid chromatography–tandem mass spectrometry runs produced 40,376 tandem mass spectra. The raw data can be downloaded from the Seattle Proteome Center Public Data Repository. Although the proteins in this sample were standard proteins, a wide range of modifications were indeed detected, which might have been introduced by sample handling or from contaminant proteins (29, 30). Specifically, there were a few phosphorylation spectra that were very appropriate for the testing purposes of this paper.

The protein sequence database against which the above data were searched included the sequences of the 18 standard proteins, 79 contaminant proteins, and 6758 yeast proteins. The precursor and fragment mass matching tolerances were ±10 ppm and ±0.5 Da, respectively, and trypsin was specified as the digestion enzyme. The yeast sequences were added as background. Any match to the yeast proteins was expected to be a false identification. In contrast, a match to the standard or contaminant proteins had a very small chance of being a false identification. This chance could be approximated based on the ratio of the number of theoretical peptides from the standard or contaminant proteins to the total number of theoretical peptides from all proteins. By calculation (trypsin digestion with up to two missed cleavage sites allowed), this ratio is 10,947/1,089,961 ≈ 0.01004. Therefore, the number of false identifications above a score threshold can be computed as $n_1 + 0.01004n_2$, where n_1 is the number of matches to yeast proteins and n_2 is the number of matches to standard or contaminant proteins.

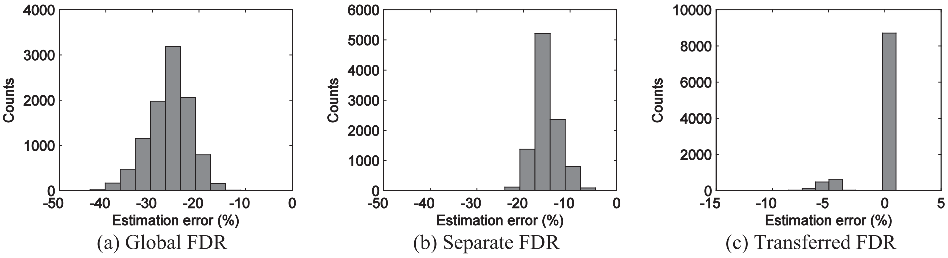
The FDR control level was set at 1% to 5%. At each level, 10,000 trials were run. In each trial, 25,000 spectra were randomly sampled for search. Table VI compares the performances of the three methods for estimating the FDR of phos-

TABLE V

Results achieved with the three methods for estimating the FDRs of different forms of phosphorylated peptide identifications on simulated data when multiple modification types were specified in one search and the FDR control level was set at 1%

Class	Global FDR			Separate FDR			Transferred FDR		
	Ave. # of I.D.s. (false/all)	Est. error (%)		Ave. # of I.D.s. (false/all)	Est. error (%)		Ave. # of I.D.s. (false/all)	Est. error (%)	
		Mean	S.D.		Mean	S.D.		Mean	S.D.
Ph.	19.27/43.34	−43.58	4.85	0.44/19.70	−1.97	4.22	0/16.64	0.11	0.31
Ph. only	12.08/36.19	−32.50	4.81	1.02/21.23	−4.80	3.29	0.0005/17.05	0.29	0.44
Ph. and Ca./Ac.	7.21/7.21	−99.01	0.01	0.38/0.38	−20.91	40.67	0/0	0	0

Notes: Ph., all phosphorylated peptide identifications; Ph. only, phosphorylated peptide identifications without co-occurrences of other types of modifications; Ph. and Ca./Ac, phosphorylated peptide identifications with co-occurring carbamylations and/or acetylations (for definitions of other abbreviations, refer to Table II). In this experiment, for each of 10,000 trials, 15,000 spectra containing no modifications, 50 containing phosphorylations, 50 containing carbamylations, and 50 containing acetylations were randomly selected for search.



(a) Global FDR (b) Separate FDR (c) Transferred FDR

FIG. 4. **Distributions of FDR estimation errors made with the three methods for phosphorylated peptide identifications on synthesized peptide data with the FDR control level set at 1%.**

TABLE VI

Results achieved with three methods for standard protein data, with the FDR control level set at 1% to 5% (refer to Table II for definitions of abbreviations)

FDR level	Global FDR			Separate FDR			Transferred FDR		
	Ave. # of I.D.s. (false/all)	Est. error (%)		Ave. # of I.D.s. (false/all)	Est. error (%)		Ave. # of I.D.s. (false/all)	Est. error (%)	
		Mean	S.D.		Mean	S.D.		Mean	S.D.
1%	20.68/51.78	−38.95	1.90	0.20/19.42	−1.00	0.00	0.24/24.02	−0.05	0.05
2%	85.65/120.22	−69.24	1.27	0.20/19.43	−1.00	0.00	0.25/24.78	0.94	0.06
3%	121.18/156.89	−74.24	0.94	0.20/19.44	−1.00	0.00	0.26/25.42	1.94	0.07
4%	168.60/204.47	−78.45	0.79	0.20/19.54	−0.91	0.62	0.26/25.83	2.94	0.08
5%	210.85/247.47	−80.20	0.70	0.20/19.98	1.81	2.35	0.26/25.89	3.92	0.10

phorylated peptide identifications. We can see that the global FDR was, again, too optimistic an estimate of the FDR of phosphorylated peptide identifications, which resulted in failed FDR control. In contrast, both the separate FDR and the transferred FDR successfully controlled the FDR of phosphorylated peptide identifications (note that overestimation is a successful control). More phosphorylated peptides were identified with the transferred FDR than with the separate FDR. The reason for this is that some decoy phosphorylated peptide identifications had even higher scores than the highest scored, target, but false phosphorylated peptide identification. Consequently, the correct target phosphorylated peptide identifications with lower scores than the decoy phosphorylated peptide identifications could not pass the score threshold set by the separate FDR. According to the transferred FDR, these high-score decoy matches occurred solely by chance and did not justify an equal number of false target matches at the given score threshold. For example, in one trial, the 21st and 22nd highest-scored phosphorylated peptide identifications were both matches to decoy sequences. When the FDR control level was set at 1%, only the first 20 highest-scored phosphorylated peptide identifications were accepted and the estimated separate FDR was 0/20 = 0%. When the FDR control level was set at 5%, the 21st highest-scored phosphorylated peptide identifications passed the score threshold and the estimated separate FDR was 1/20 = 5% (note that the number of reported phosphorylated peptide identifications was still 20, because the 21st highest-scored phosphorylated peptide identification was a decoy match and was discarded).

CONCLUSIONS

This paper presents a solution to the problem of accurate FDR estimation for rare protein modifications detected via high-throughput tandem mass spectrometry. Through flexible use of the empirical data from target-decoy database searches, a computable relationship is derived between the subgroup FDR of modified peptide identifications and the global FDR of all peptide identifications. Through this relationship, the FDR of rare modifications can be transferred from the global FDR, which can be accurately estimated via existing methods, thereby avoiding an inaccurate direct estimation from inadequate data. Experimental results demonstrate that the FDR of modified peptide identifications can be successfully controlled with the transferred FDR without sacrificing sensitivity, whereas the global FDR and the direct separate FDR present large biases when the modifications of interest are rare. Finally, it is worthwhile to note that the proposed method is in principle adaptable to other small-subgroup FDR estimation problems in proteomics, such as the identification of peptides containing specific amino acids or cross-linked peptides, in which cases the mass spectra of such peptides could be rare in a dataset.

* This work was partly supported by the International S&T Cooperation Program of China under Grant No. S2014ZR0097, the National Natural Science Foundation of China under Grant No. 30900262 and the National High Technology Research and Development Program (863) of China under Grant No. 2012AA020201. Yan Fu was supported by the National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences (CAS), and the Key Laboratory of Random Complex Structures and Data Science, CAS.

§ To whom correspondence should be addressed: Yan Fu, No. 55 Zhongguancun East Road, Haidian District, Beijing 100190, China, Tel.: 86–10-62614318, E-mail: yfu@amss.ac.cn.

REFERENCES

1. Walsh, C. T. (2005) *Posttranslational Modification of Proteins: Expanding Nature's Inventory*, Roberts & Company Publishers, Englewood, CO
2. Vidal, C. J. (ed) (2011) *Post-Translational Modifications in Health and Disease, Protein Reviews 13*, Springer, New York
3. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
4. Mann, M., and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **21,** 255–261
5. Witze, E. S., Old, W. M., Resing, K. A., and Ahn, N. G. (2007) Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods* **4,** 798–806
6. Eng, J. K., McCormack, A. L., and Yates, J. R., III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass. Spectrom.* **5,** 976–989
7. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
8. Fu, Y., Yang, Q., Sun, R., Li, D., Zeng, R., Ling, C. X., and Gao, W. (2004) Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* **20,** 1948–1954
9. Nesvizhskii, A. I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4,** 787–797
10. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* **57,** 289–300
11. Choi, H., and Nesvizhskii, A. I. (2008) False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **7,** 47–50
12. Noble, W. S. (2009) How does multiple testing correction work? *Nat. Biotechnol.* **27,** 1135–1137
13. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4,** 207–214
14. Huttlin, E. L., Hegeman, A. D., Harms, A. C., and Sussman, M. R. (2007) Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. *J. Proteome Res.* **6,** 392–398
15. Kall, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4,** 923–925
16. Fu, Y. (2012) Bayesian false discovery rates for post-translational modification proteomics. *Statistics Interface* **5,** 47–59
17. Baker, P. R., Medzihradszky, K. F., and Chalkley, R. J. (2010) Improving software performance for peptide electron transfer dissociation data analysis by implementation of charge state- and sequence-dependent scoring. *Mol. Cell. Proteomics* **9,** 1795–1803
18. Marx, H., Lemeer, S., Schliep, J. E., Matheron, L., Mohammed, S., Cox, J. Mann, M., Heck, A. J., and Kuster, B. (2013) A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat. Biotechnol.* **31,** 557–564
19. Efron, B. (2008) Simultaneous inference: when should hypothesis testing problems be combined? *Ann. Appl. Stat.* **2,** 197–223
20. Hu, J. X., Zhao, H. Y., and Zhou, H. H. (2010) False discovery rate control with groups. *J. Am. Stat. Assoc.* **105,** 1215–1227
21. Efron, B. (2010) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge University Press, New York
22. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24,** 1285–1292
23. Chalkley, R. J., and Clauser, K. R. (2012) Modification site localization scoring: strategies and performance. *Mol. Cell. Proteomics* **11,** 3–14
24. Choi, H., Ghosh, D., and Nesvizhskii, A. I. (2008) Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* **7,** 286–292
25. Kall, L., Storey, J. D., and Noble, W. S. (2008) Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics* **24,** i42-i48
26. Li, D., Fu, Y., Sun, R., Ling, C., Wei, Y., Zhou, H., Zeng, R., Yang, Q., He, S., and Gao, W. (2005) pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* **21,** 3049–3050
27. Wang, L. H., Li, D. Q., Fu, Y., Wang, H. P., Zhang, J. F., Yuan, Z. F., Sun, R. X., Zeng, R., He, S. M., and Gao, W. (2007) pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **21,** 2985–2991
28. Klimek, J., Eddes, J. S., Hohmann, L., Jackson, J., Peterson, A., Letarte, S., Gafken, P. R., Katz, J. E., Mallick, P., Lee, H., Schmidt, A., Ossola, R., Eng, J. K., Aebersold, R., and Martin, D. B. (2008) The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.* **7,** 96–103
29. Fu, Y., Xiu, L.-Y., Jia, W., Ye, D., Sun, R.-X., Qian, X.-H., and He, S.-M. (2011) DeltAMT: a statistical algorithm for fast detection of protein modifications from LC-MS/MS data. *Mol. Cell. Proteomics* **10,** M110.000455
30. Ye, D., Fu, Y., Sun, R. X., Wang, H. P., Yuan, Z. F., Chi, H., and He, S. M. (2010) Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics* **26,** i399–i406