BuildSummary: Using a Group-Based Approach To Improve the Sensitivity of Peptide/Protein Identification in Shotgun Proteomics

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · JANUARY 2012

Impact Factor: 4.25 · DOI: 10.1021/pr200194p · Source: PubMed

5 AUTHORS, INCLUDING:

Quanhu Sheng
Vanderbilt University
63 PUBLICATIONS **1,407** CITATIONS
SEE PROFILE

Yi-Bo Wu
ETH Zurich
9 PUBLICATIONS **159** CITATIONS
SEE PROFILE

Journal of proteome research

Article

pubs.acs.org/jpr

BuildSummary: Using a Group-Based Approach To Improve the Sensitivity of Peptide/Protein Identification in Shotgun Proteomics

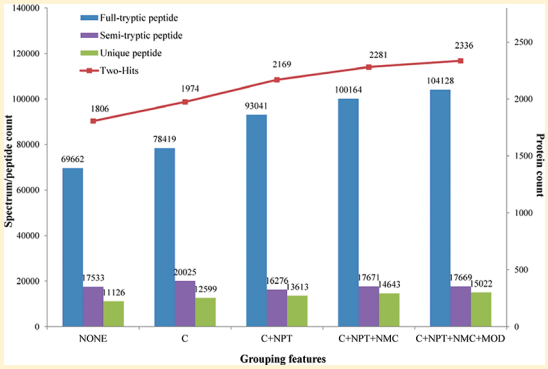
Quanhu Sheng,[†] Jie Dai,[†] Yibo Wu,[†] Haixu Tang,[‡] and Rong Zeng*[,†]

[†]Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

[‡]School of Informatics and Computing, Indiana University, Bloomington, Indiana 47406, United States

Supporting Information

ABSTRACT: The target-decoy database search strategy is widely accepted as a standard method for estimating the false discovery rate (FDR) of peptide identification, based on which peptide-spectrum matches (PSMs) from the target database are filtered. To improve the sensitivity of protein identification given a fixed accuracy (frequently defined by a protein FDR threshold), a postprocessing procedure is often used that integrates results from different peptide search engines that had assayed the same data set. In this work, we show that PSMs that are grouped by the precursor charge, the number of missed internal cleavage sites, the modification state, and the numbers of protease termini and that the proteins grouped by their unique peptide count should be filtered separately according to the given FDR. We also develop an iterative procedure to filter the PSMs and proteins simultaneously, according to the given FDR. Finally, we present a general framework to integrate the results from different peptide search engines using the same FDR threshold. Our method was tested with several shotgun proteomics data sets that were acquired by multiple LC/MS instruments from two different biological samples. The results showed a satisfactory performance. We implemented the method in a user-friendly software package called *BuildSummary*, which can be downloaded for free from http://www.proteomics.ac.cn/software/proteomicstools/index.htm as part of the software suite *ProteomicsTools*.



KEYWORDS: *group-based, BuildSummary, proteomics, protein identification, mass spectrometry*

INTRODUCTION

Some studies have suggested that searching spectra from high-resolution data should be conducted under a stringent precursor mass tolerance (such as 10 ppm for Orbitrap-LTQ data),[1] while other studies have indicated that a large mass tolerance would make it easier to distinguish between positive and negative peptide-spectrum matches (PSMs).[2,3] It has also been suggested that peptide filtering should be carried out on groups of PSMs with different precursor charge states because the (null) score distributions of false PSMs differ substantially among PSMs with distinct charge states.[4] Usually, phosphorylated and unmodified PSMs are analyzed separately, and peptides with one tryptic terminus or multiple missed internal cleavage sites are sometimes discarded by the peptide filtering.[5] However, other studies have shown that a considerable number of nontryptic peptides can be identified in shotgun proteomics experiments.[6,7] Statistical and machine-learning tools have been developed to filter peptide-spectrum matches on the basis of matching scores (e.g., in *PeptideProphet*[4]) and/or other peptide and spectrum features (e.g., in *PeptideProphet*[4] or *Percolator*[8]). However, the optimal criteria for peptide filtering to achieve the maximum number of peptide identifications are still unclear. For example, should we group PSMs by other features prior to peptide filtering, as with precursor charges and modification states?

Even once a reliable set of peptides has been identified, it is not an easy task to compile a reliable list of proteins from those peptides due to two challenges. The first challenge is that some of the identified peptides are shared by two or more proteins in the database, which are called *degenerate* peptides. As a result, the problem of determining which of these proteins are actually present in the sample, which is known as the *protein inference problem*,[9−13] often has multiple solutions. The second challenge is that even after determining the most likely set of proteins by solving the protein inference problem, we need to estimate a false discovery rate (FDR) for the derived proteins, which is referred to as *the protein FDR*. This problem appears straightforward at first glance. One may compute the FDR of an identified protein by multiplying the FDRs of the identified peptides in the protein. However, this simple method may be inaccurate. For example, even under a low peptide FDR (e.g., 0.01), the FDR for *one-hit-wonders*, i.e., proteins with only one identified peptide, can be as high as >0.5.[14] Simply discarding

Received: March 1, 2011

   dx.doi.org/10.1021/pr200194p | *J. Proteome Res.* XXXX, XXX, XXX−XXX

all of these proteins may result in the loss of a large number of correctly identified proteins.[15]

Some algorithms have been developed to integrate search results from multiple search engines.[10,16−20] Resing et al.[16] did not filter peptides/protein pairs on the basis of their estimated FDRs. Edwards et al.[20] applied the random forest technique, which requires careful tuning, to classify PSMs from target and decoy databases. Alves et al.[18] calculated a score from independent *P* values using statistical models, while others have combined an independent *P* value to obtain a final probability based on Bayesian rules[19] or a classification model[17] built from training sets. Either these results cannot incorporate additional search engine results into the methodology or the performance is dependent on the quality of the training set.

Here, we examine the factors that might affect the sensitivity of peptide/protein identification and propose a group-based approach to combine results from different search engines and/or different search parameters using a target-decoy search strategy. Our method compiles a list of identified proteins below a user-defined threshold for a protein FDR while retaining as many identified peptides as possible with a low peptide FDR within those proteins. We compare our method with existing methods, including *PeptideProphet*,[4] *iProphet*, and *ProteinProphet*,[21] using several shotgun proteomics data sets that were acquired from two different biological samples, namely, human plasma and mouse liver. The results show a satisfactory performance. We implemented our method in a user-friendly software package called *BuildSummary*, which can be downloaded for free from our Web site.

MATERIALS AND METHODS

MS/MS Data Sets

The MS/MS data sets were acquired from human plasma samples and the murine 3T3-L1 cell line. The human plasma data set is described in theSupporting Information. The mouse data set was generated by Thermo Orbitrap-LTQ mass spectrometry, and the human plasma samples were acquired using Thermo LCQ, LTQ, and Orbitrap-LTQ mass spectrometry. The mouse data set was generated for the purpose of stable isotope labeling by amino acids in cell culture (SILAC) quantification, as described previously.[22] Only data from the SCX-SAX method (including 20 raw files) was used in this study. A human protein database (size ≈13 MB, containing 20,277 sequences) and a mouse protein database (size ≈10 MB, containing 16,245 sequences) were extracted from the UniProt protein database (http://www.uniprot.org/, version 20100420) and used as the target database for peptide/protein identification. Two concatenated target-decoy databases were generated by reversing the protein sequences in the target database, as described previously.[5]

Database Searching

All acquired MS/MS spectra were preprocessed by TurboR-AW2MGF[23] to generate MASCOT Generic Format (MGF) files.

The MGF files for the human data set were searched against the concatenated target-decoy human protein database using BioWorks 3.2 (Sequest, Thermo Electron, San Jose, CA), MASCOT (version 2.2), and X!Tandem (version 2010.01.01.4). Trypsin was designated as the protease with full-tryptic specified, allowing up to one missed internal cleavage site. Carbamidomethylation was considered as a fixed modification, while phosphorylation of the serine/threonine/tyrosine residues and oxidized methionine were allowed as variable modifications. The precursor mass type was set to average for the LCQ/LTQ data and monoisotopic for the Orbitrap-LTQ data. The precursor mass tolerance was set to 3.0 Da during Sequest/X! Tandem searching and 3.0 Da and 10 ppm during MASCOT searching. N-terminal protein acetylation was allowed as a variable modification during MASCOT searching.

The MGF files for the mouse data set were searched against the concatenated target-decoy mouse protein database using MASCOT (version 2.2). Trypsin was designated as the protease with the semitryptic option specified, allowing up to two missed internal cleavage sites. Carbamidomethylation was considered a fixed modification, while isotope-labeled lysine, N-terminal protein acetylation, phosphorylation of serine/threonine/tyrosine residues, and oxidized methionine were allowed as variable modifications. The precursor mass tolerance was set to 10 ppm, and the fragmentation ion tolerance was set to 0.5 Da.

For the spectra with precursor ions of unassigned charges, we assumed two different precursor charges (+2 and +3) and separately searched with them.

Software Development

We implemented our group-based filtering method in a user-friendly software package called *BuildSummary*. *BuildSummary* was developed using the C# programming language and was compiled in the Microsoft Visual Studio 2010 Professional Edition. The software is fully compatible with Windows-based operating systems with dotNET framework v3.5. It also features an easy installation procedure and provides a user-friendly graphic interface.

RESULTS

Group-Based Filtering Method

Our group-based filtering method consists of six steps to simultaneously achieve the assigned protein FDR cutoff and peptide FDR cutoff for a data set (Figure 1).

(1) Parse search results from the different search engines. Three specific parsers were developed to parse the search results from Sequest/MASCOT and X!Tandem. A combined criteria could be used to filter the PSMs from each search engine, e.g., DeltaC*n* ≥ 0.1 for Sequest and *E*-value ≤0.05 for MASCOT/X!Tandem in addition to a precursor tolerance that can be applied to all of the search engines. For the high-resolution data that is searched with a large mass tolerance, an additional option is provided to filter the PSMs by considering the precursor ion as the second or third isotopic ion of that peptide with a narrow precursor tolerance.

(2) Group and filter the PSMs by the peptide FDR threshold. The PSMs were grouped into different categories prior to FDR filtering according to search engine, precursor charge, the number of missed internal cleavage sites (NMC), the modification state (i.e., whether or not they are post-translationally modified), the number of protease termini, and other, user-specified categories (e.g., different samples or different instruments). The threshold for the peptide FDR is initialized by a user-defined peptide FDR and may be deduced at step 4. The *q*-values of the PSMs in each category are calculated separately;[8,24] the *q*-value of a given PSM is calculated as the peptide FDR of the PSMs with higher scores than the PSM in the category of interest.

dx.doi.org/10.1021/pr200194p | J. Proteome Res. XXXX, XXX, XXX−XXX

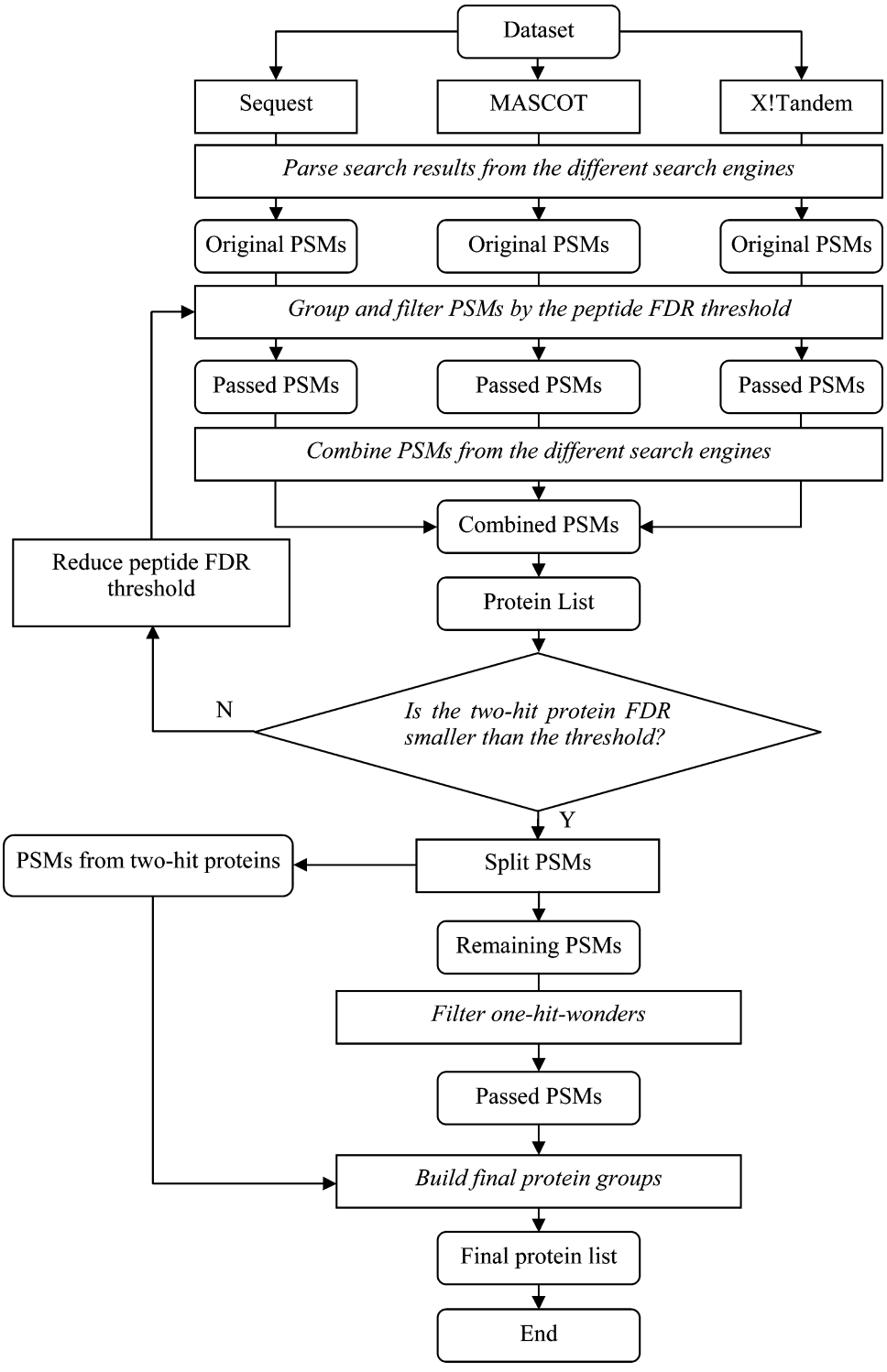


Figure 1. Flow chart of the group-based filtering method.

The peptide FDR can be computed by either eq 1[5] or eq 2[24] and is chosen by the user, where

$$\mathrm{FDR} = 2N(\mathrm{decoy})/(N(\mathrm{decoy}) + N(\mathrm{target})) \quad (1)$$

or

$$\mathrm{FDR} = N(\mathrm{decoy})/N(\mathrm{target}) \quad (2)$$

Here, N(decoy) represents the number of PSMs from the decoy database and N(target) represents the number of PSMs from the target database. If the peptide of a PSM is present in both the target database and the decoy database, then the PSM is considered a decoy entry. For each category, the PSMs with a q-value below the peptide FDR threshold are accepted as *valid* PSMs. Equation 2 was used in the following analysis.

(3) Combine the PSMs from the different search engines. In some cases, both PSMs from both charges of the same

dx.doi.org/10.1021/pr200194p | J. Proteome Res. XXXX, XXX, XXX−XXX

spectrum pass the FDR filter because they are grouped into different categories and filtered separately. Our method keeps the PSM with the better matching score and discards the other one. Furthermore, when multiple search engines are used, the same spectrum may correspond to different PSMs. In this case, two options are provided: the user can either discard all of them or retain the PSM with the lowest *q*-value.

(4) Filter proteins with two or more unique peptide hits. Protein candidates are assembled from valid PSMs. The proteins containing an identical set of valid PSMs are merged into a single protein group. A protein group is considered *redundant*, and therefore removed, if the identified peptides in it are a subset of the identified peptides in another protein group. The protein FDR for nonredundant protein groups with at least two unique peptides is calculated using the same method as used in the peptide FDR calculation (i.e., eq 1 or 2), except that *N*(decoy) now represents the number of protein groups containing at least one protein from the decoy database and that *N*(target) now represents the number of protein groups containing proteins only from the target database. If the resulting FDR is greater than a given threshold for the protein FDR, the threshold for the peptide FDR is reduced, and steps 2−4 are repeated; otherwise, the PSMs of the two-hit protein groups are retained and removed from the list of valid PSMs.

(5) Filter one-hit-wonders. The *q*-values of the remaining, valid PSMs are recalculated on the basis of the unique peptides using the same method as used in the peptide FDR calculation (i.e., eq 1 or 2), except that *N*(decoy) now represents the number of valid PSMs with different peptide sequences from the decoy database and that *N*(target) now represents the number of valid PSMs with different peptide sequences from the target database. The PSMs are retained if their *q*-values are below the given threshold for the protein FDR. Since the PSMs corresponding to protein groups with two or more unique peptides have been removed, the protein groups assembled from the retained PSMs are all one-hit-wonders, and their protein FDR is below the protein FDR threshold.

(6) Build the final protein groups. The PSMs retained in steps 4 and 5 are combined and assembled into the final protein groups.

Search Parameter "Mass Tolerance" and Filter Parameter "Precursor Tolerance". We examined how the search parameter "mass tolerance" and the filter parameter "precursor tolerance" affect protein identification in Orbitrap-LTQ data using the human plasma sample. Both sets of MASCOT results that were searched with a 3 Da and a 10 ppm mass tolerance were used. Table 1 shows the mass accuracy statistics of the PSMs that were searched by a 3 Da mass tolerance and passed with a 1% protein FDR. Iso1/Iso2/Iso3 represents the PSMs with an observed precursor ion around the 20 ppm range of the theoretical mono/second/third isotopic ions of that peptide. Only 1154 PSMs (termed "Other" in Table 1) were not included in those three isotopic groups. Table 1 indicates that there were a substantial number of spectra with precursor mass-to-charges assigned by the second/third isotopic ions but still identified by the search engine when a large mass tolerance was used.

We compared the identification results generated from different combinations of the search and filter precursor parameters that passed a 1% protein FDR threshold (Table 2). Searching with a 3 Da mass tolerance and filtering with a 10 ppm precursor tolerance achieved almost identical peptide/protein identification results as searching directly with a 10 ppm mass tolerance. However, searching with a 3 Da mass tolerance and filtering the mono/second/third isotopic ions with a 10 ppm precursor tolerance achieved the maximum peptide/protein identification sensitivity, except for the one-hit-wonders. Unless specified otherwise, the Orbitrap-LTQ data set from the human plasma sample used in the following analysis was parsed with a 3 Da mass tolerance search result and filtered with 10 ppm precursor tolerance for the isotopic ions.

Table 1. Mass Accuracy Statistics of the PSMs That Were Searched by a 3 Da Mass Tolerance and Passed with a 1% Protein FDR

type[a]	range (ppm)	count	mean (ppm)	SD (ppm)
Iso0	20	27529	0.7	2.6
Iso1	20	6672	4.0	5.3
Iso2	20	949	4.2	6.5
other	NO	1154	−10.8	1004.1

[a]Iso1/Iso2/Iso3 represent the PSMs with an observed precursor ion concentration around the 20 ppm range of the theoretical mono/second/third isotopic ions of that peptide. The mass accuracy of the PSM in the "other" category was calculated from the error between the observed precursor ion and theoretical mono isotopic ion.

Features of PSMs That Affect the Sensitivity of Peptide Identification. We used the mouse data set to confirm that grouping the PSMs by the modification state (MOD) and the precursor charge (Charge) prior to FDR filtering improved the peptide identification sensitivity. We also evaluated two other features, namely, the number of internal missed cleavage sites (NMC) and the number of protease termini (NPT),[4] which are usually directly integrated into the confidence score (Table 3).

Modification State. We examined the PSMs in the group with charge = 3, NMC = 0, and NPC = 2. There were 14,277 PSMs that passed a 1% peptide FDR filtering threshold without grouping PSMs into different categories on the basis of the modification state; these included 1,930 phosphorylated PSMs with an actual peptide FDR of 6.1%. When only the phosphorylated PSMs were considered to be modified (termed MOD[STY] in Table 3), the number of identified peptides increased to 16,151 (an increase of approximately 13.13%) after grouping the PSMs by modification state. More importantly, the number of unique, identified peptides increased from 2,972 to 3,281 (an increase of roughly 10.4%), which might contribute more high-confidence, two-hit proteins. We also examined if grouping the PSMs by the state of other variable modifications would improve the sensitivity of peptide identification because oxidized methionine (termed MOD[M] in Table 3) and isotope-labeled lysine (termed MOD[K] in Table 3) were included in the database search. The results show that grouping by the modification state of the isotope-labeled lysine resulted in no improvement, but grouping by the modification state of oxidized methionine improved the sensitivity. Treating phosphorylated or oxidized PSMs as modified PSMs resulted in the greatest improvement in the sensitivity of peptide identification (termed MOD[MSTY] in Table 3, with a 12.36% increase in identified peptides and a 7.71% increase in unique, identified peptides). All criteria

dx.doi.org/10.1021/pr200194p | J. Proteome Res. XXXX, XXX, XXX−XXX

Table 2. Identification Results from Different Combinations of Search and Filtering Parameters That Passed the 1% Protein FDR Threshold

search	filtering	peptide	unique peptides	protein	two-hit proteins	two-hit protein FDR (%)	one-hit-wonders	one-hit-wonder FDR (%)
10 ppm		31091	1437	223	175	0.57	48	0
3 Da		36304	1466	262	179	0.56	83	0
3 Da	10 ppm[a]	30580	1414	224	172	0	52	0
3 Da	10 ppm + iso[b]	37579	1488	231	181	0.56	50	0

[a]Filtered PSMs with a 10 ppm precursor tolerance while only considering the monoisotopic ion of the precursors. [b]Filtered PSMs with a 10 ppm precursor tolerance while considering the mono/second/third isotopic ions of the precursors.

Table 3. Grouping PSMs by Feature Can Improve Sensitivity

test features		fixed features				ungrouped	grouped			
name	value[e]	charge	NMC	MOD	NPT	peptide FDR (%)	peptide	unique peptides	peptide increase (%)	unique peptide increase (%)
MOD[STY][a]	T	3	0		2	6.10	1490	259	−22.8	−41.67
	F					0.24	14661	3066	18.74	19.21
	all					1	16151	3281	13.13	10.4
MOD[K][b]	T	3	0		2	1.02	3837	1062	−0.49	−0.84
	F					0.99	10421	2572	0	0
	all					1	14258	2967	−0.13	−0.17
MOD[M][c]	T	3	0		2	55.29	18	6	−86.36	−93.88
	F					0.67	14872	3089	5.14	6.96
	all					1	14890	3090	4.29	3.97
MOD[MSTY][d]	T	3	0		2	6.6	1489	262	−24.42	−44.14
	F					0.15	14552	2988	18.24	16.99
	All					1	16041	3201	12.36	7.71
charge	1		0	F	2	8.55	2893	730	−65.04	−65.89
	2					0.25	83379	11257	24.98	33.93
	3+					0.03	13796	2908	77.67	68.48
	all					1	100068	12464	20.56	24.45
NMC	0	2		F	2	0.61	83379	11256	7.57	11.27
	1					10.46	1587	366	−33.51	−49.79
	2					92.45	27	12	−86.76	−91.84
	All					1	84993	11634	6.11	5.84
NPT	1	2	0	F		3.31	16388	3368	−20.06	−28.16
	2					0.34	83379	11256	18.87	25.75
	all					1	99767	14616	10.07	7.23

[a]MOD[STY] indicates that only phosphorylated PSMs were treated as modified. [b]MOD[K] indicates that only the PSMs that have isotope-labeled lysines were treated as modified. [c]MOD[M] indicates that only the oxidized PSMs were treated as modified. [d]MOD[MSTY] indicates that only the oxidized or phosphorylated PSMs were treated as modified. [e]"T" indicates that the modification state is "True". "F" indicates that the modification state is "False". A value of "All" indicates that all PSMs passed the filtering. Charge 1, 2 indicate that the precursor charge is 1, 2, respectively. Charge 3+ indicates that the precursor charge is equal to or larger than 3. NMC 0, 1, 2 indicate that the number of missed internal cleavages is 0, 1, 2, respectively. NPT 1, 2 indicate that the number of protease termini is 1, 2, respectively.

identified as MOD = false in the following sections indicate that the PSMs were not phosphorylated or oxidized.

Precursor Charge. We examined the PSMs in the group with NMC = 0, MOD = false, and NPT = 2. There were 83,006 PSMs and 10,015 unique peptides that passed 1% peptide FDR filtering without grouping PSMs by the precursor charge, which included 8,274 one-charge PSMs with an actual peptide FDR of 8.55%. The number of identified peptides and unique peptides increased to 100,068 (an increase of about 20.56%) and 12,464 (an increase of about 24.45%) after grouping PSMs by the precursor charge.

Number of Internal Missed Cleavage Sites. We examined the PSMs in the group with charge = 2, MOD = false, and NPC = 2. There were 80,099 PSMs and 10,992 unique peptides that passed 1% peptide FDR filtering without grouping the PSMs by the number of internal missed cleavage sites, which included 2,387 PSMs with NMC = 1 and an actual peptide FDR of 10.46% and 204 PSMs with NMC = 2 and an actual peptide FDR of 92.45%. The number of identified peptides and unique, identified peptides increased to 84,993 (an increase of about 6.11%) and 11,634 (an increase of about 5.84%), respectively, after grouping the PSMs by the number of internal missed cleavage sites. Considering the rapid increase in the time it takes to search the database with two internal missed internal cleavage sites and the small improvement in peptide identifications that results, limiting the search to only one internal missed cleavage site in the database search is highly recommended.

Number of Protease Termini (NPT). We examined the PSMs in the group with charge = 2, NMC = 0, and MOD = false. There were 90,643 PSMs and 13,630 unique peptides that passed 1% peptide FDR filtering without grouping the PSMs by the number of protease termini, which included 20,500 semitryptic PSMs with an actual peptide FDR of 3.31%. The number of identified peptides and unique, identified peptides increased to 99,767 (an increase of about 10.07%) and 14,616

dx.doi.org/10.1021/pr200194p | J. Proteome Res. XXXX, XXX, XXX−XXX

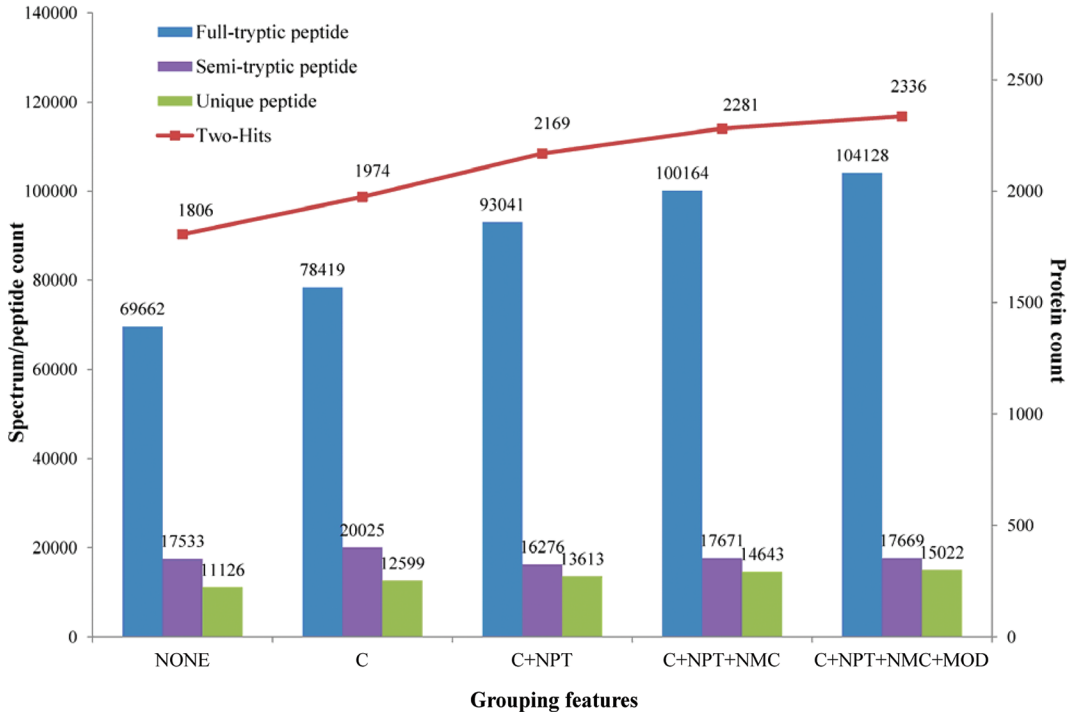


Figure 2. Power of grouping PSMs into different categories: None, all PSMs are in one category; C, PSMs are grouped by the precursor charge; NPT, PSMs are grouped by the number of protease termini; NMC, PSMs are grouped by the number of missed internal cleavage sites; MOD, PSMs are grouped by the modification state (i.e., phosphorylated/oxidated or not). The identified peptides, unique peptides, and two-hit proteins increased significantly after grouping PSMs into different categories and filtering with a 1% protein FDR.

(an increase of about 7.23%), respectively, after grouping the PSMs by the number of protease termini. Because there were still 16,388 semitryptic PSMs that passed 1% FDR filtering using the group-based method, we concluded that semitryptic peptides were indeed present in the sample.

Power of Grouping PSMs into Different Categories

The mouse data set result was filtered with a 1% protein FDR threshold using different grouping criteria (Figure 2). The more features were used to group the PSMs into different categories, the more peptides, unique peptides, and two-hit proteins passed the 1% protein FDR threshold. The improvement in two-hit protein identification demonstrates the advantage of PSM grouping. The results show that PSMs that are grouped by the precursor charge, the number of missed internal cleavage sites, the modification state, and the numbers of protease termini should be filtered separately to achieve the maximum peptide/protein identification sensitivity.

Different Instruments. Sometimes, search results from different instruments using the same sample need to be combined. We thus examined the PSMs in the group with charge = 2, NMC = 0, MOD = false, and NPT = 2 from the MASCOT results with the human data set, which included LCQ, LTQ, and Orbitrap-LTQ data sets (Table 4). Here, the Orbitrap-LTQ data set which searched with 10 ppm mass tolerance was used. Grouping the PSMs by instrument type increased the number of identified peptides from 70,493 to 71,754 (an increase of about 1.79%), but the number of unique peptides only increased slightly from 2,534 to 2,539 (an increase of about 0.20%). Interestingly, the group-based method improved the peptide identification sensitivities of the LCQ and Orbitrap-LTQ data sets significantly (17.32% and 10.91% increases in identified peptides and 32.79% and 15.60% increases in unique peptides, respectively). However, the

Table 4. Grouping PSMs by Instrument Type Can Improve Sensitivity

	ungrouped	grouped			
instrument type	peptide FDR (%)	peptides	unique peptides	peptide increase (%)	unique peptide increase (%)
LCQ	0.31	14890	1053	17.32	32.79
LTQ	1.62	34587	1847	−8.3	−15.51
Orbitrap-LTQ	0.27	22277	1319	10.91	15.6
all	1	71754	2539	1.79	0.2

number of peptides and unique peptides identified from the LTQ data set decreased by 8.30% and 15.51%, respectively. That was possibly caused by the different score distributions between the LCQ, LTQ, and Orbitrap-LTQ data sets (Supporting Information Figure 1), where there was higher decoy peak in the LTQ data set than in the LCQ and Orbitrap-LTQ data sets. We also examined the identified results filtered by 1% protein FDR. Grouping the PSMs by instrument type increased the peptides/unique peptides/target two-hit proteins/target one-hit-wonders from 82,227/1,967/201/54 to 86,201/2,043/206/69. Overall, we still recommend grouping the PSMs from different instruments prior to FDR filtering.

Improvement in Peptide and Protein Identification by the Group-Based Method. We compared the identification results at three filtering levels (i.e., peptide, unique peptide, and protein level) after group-based filtering of two high-resolution data sets (Table 5). For the human data set, only the Orbitrap-LTQ data was used. Two modes were used for filtering at the protein level. "Simple protein" indicated that two-hit proteins and one-hit-wonders were not grouped into different categories prior to FDR filtering, while "Protein" corresponded to the

dx.doi.org/10.1021/pr200194p | J. Proteome Res. XXXX, XXX, XXX−XXX

Table 5. Comparison of Identifications Filtered by Peptide/Unique Peptide/Protein FDR Thresholds

sample	filter level	peptides	unique peptides	proteins	two-hit proteins	two-hit protein FDR (%)	one-hit-wonders	one-hit-wonder FDR (%)
mouse	peptide	129547	17118	4380	2467	1.9	1913	27.96
	unique peptide	106182	13375	3420	2030	0.15	1390	6.84
	simple protein	93556	11538	3118	1819	0	1299	2.28
	protein	120620	14924	3064	2312	0.65	752	0.67
human	peptide	31880	1693	471	175	0.57	296	60.87
	unique peptide	27653	1326	252	163	0	89	18.67
	simple protein	23801	1160	223	148	0	75	1.35
	protein	31327	1445	223	175	0.57	48	0

group-based method. The FDR of one-hit-wonders that passed the 1% peptide FDR threshold was as high as 27.96% in the mouse data set but unacceptably higher (60.87%) in the human data set. Even when filtered with a 1% unique peptide FDR threshold, the FDR of one-hit-wonders was still as high as 6.84% in the mouse data set and 18.67% in the human data set. However, the "Simple protein" mode resulted in a significantly decreased number of PSMs and two-hit proteins as compared to the "Protein" mode. Overall, grouping proteins by the unique peptide counts prior to FDR filtering reduced the FDRs of both the two-hit proteins and the one-hit-wonders below the user-assigned FDR while maximizing the peptide/protein identification sensitivity.

Filtering identification results at the protein level discarded more one-hit-wonders compared to peptide-level filtering. To verify the confidence of the retained one-hit-wonders, a reference protein list was constructed using the 208 two-hit proteins from the human data set, which passed a 1% protein FDR threshold using the group-based method. The one-hit-wonders from any single analysis (i.e., LCQ/LTQ or Orbitrap) that were on this protein list were considered high-confidence proteins.

Table 6 validates the one-hit-wonders that were retained or discarded after group-based filtering at a 1% FDR threshold at the protein level. The percentage of high-confidence one-hit-wonders that were retained and discarded was 50% and 8.7% in the LCQ data set, 23.8% and 2.6% in the LTQ data set, and

Table 6. Confident One-Hit-Wonders Were Enriched by the Group-Based Method

	one-hit-wonders kept (target only)		one-hit-wonders discarded (target only)	
instrument	two-hit[a]	total	two-hit	total
LCQ	13	26	12	138
LTQ	10	42	9	346
Orbitrap-LTQ	14	50	3	55

[a]Two-hit: one-hit-wonders that were two-hit proteins in a combined set of identified proteins (see text for details).

28% and 5.5% in the Orbitrap-LTQ data set, respectively. This implies that the number of high-confidence one-hit-wonders was significantly enriched by the group-based method.

Combining the Results from Multiple Search Engines. Sequest, MASCOT, and X!Tandem were used to search the human data set. The identification results filtered at a 1% protein FDR threshold based on different combinations of search engines are summarized in Table 7.

Table 7 shows that MASCOT achieved the best performance in this data set because it identified the same additional 7.7% target two-hit proteins as X!Tandem than Sequest, which identified 169 target two-hit proteins only, but MASCOT identified more peptides and unique peptides than X!Tandem. Combining Sequest and MASCOT results identified an additional 13.6% target two-hit proteins than by Sequest only. Adding X!Tandem results achieved a total 14.2% increase of target two-hit proteins and 19.1% increase of one-hit-wonders than by Sequest only. Finally, no significant difference was observed when conflicting PSMs (i.e., PSMs from the same spectrum were assigned to different peptides by different search engines) were discarded or when only the PSMs with the lowest q-value were retained.

We also performed overlap analysis of integrated results from two search engines (Supporting Information Figure 2). Since the integration procedure discarded some conflicted spectra in each engine result, the two-hit proteins identified by each engine in combined results decreased as expected. But the complementary engines also contributed more two-hit proteins. Overall, integration of search results from multiple engines identified more proteins.

Comparison with PeptideProphet, ProteinProphet, and iProphet Results. *PeptideProphet* has been widely used to calculate the confidence of PSMs from SEQUEST, MASCOT, and X!Tandem search results. Typically, results are also analyzed by *ProteinProphet*, which assembles the peptides into proteins and computes the confidence of the proteins present in the sample. *iProphet* was designed to improve peptide identification on the basis of the results from *PeptideProphet*. Typically, a probability of 0.9, which corresponds to a FDR threshold of almost 1%, is used in *PeptideProphet/ProteinProphet* analysis. We compared the results from our method with *PeptideProphet/ProteinProphet/iProphet* analysis (TPP v4.4 VUVUZELA rev 1, Build 20101-0121551) using the Orbitrap-LTQ human data set which was searched with 10 ppm mass tolerance (Table 8).

Probability values of 0.842 and 0.823 were used to filter the PSMs refined by *PeptideProphet* and *iProphet*. Both values correspond to a 1% peptide FDR, as described in the distribution tables of their result files. After filtering the results at the peptide level, *PeptideProphet/iProphet* analysis and our method both achieved acceptable actual FDRs for the two-hit proteins (0% for *PeptideProphet*, 0% for *iProphet*, and 0.57% for our method). The results indicate that our method identified more PSMs, unique peptides, and target two-hit proteins than *PeptideProphet/iProphet* while *iProphet* identified the least false positive one-hit-wonders. We also compared our method with *ProteinProphet*. The number of accepted proteins with *ProteinProphet* probabilities greater than 0.9 was less than the proteins identified by our method, and they also had a higher FDR (1.92% for two-hit proteins and 3.77% for one-hit-wonders) than expected. Even after filtering with a *ProteinProphet* probability of 0.99, which corresponds to a 0% FDR threshold,

dx.doi.org/10.1021/pr200194p | J. Proteome Res. XXXX, XXX, XXX–XXX

Table 7. Results from Multiple Search Engines Passed 1% Protein FDR[a]

engines	peptides	unique peptides	proteins	target two-hit proteins		target one-hit-wonders	
				count	increase[b] (%)	count	increase[b] (%)
S/Sequest	37704	1580	237	169	0	68	0
M/MASCOT	**38868**	**1507**	**225**	**182**	**7.7**	**43**	**−36.8**
X/X!Tandem	35987	1443	246	182	7.7	63	−7.4
SM_QVALUE	**44116**	**1686**	**254**	**192**	**13.6**	**62**	**-8.8**
SX_QVALUE	45689	1727	268	190	12.4	77	13.2
MX_QVALUE	42577	1669	256	189	11.8	66	−2.9
SMX_QVALUE	**47208**	**1759**	**275**	**193**	**14.2**	**81**	**19.1**
SMX_DISCARDALL	47119	1755	272	193	14.2	78	14.7

[a]S, Sequest; M, MASCOT; X, X!Tandem. DISCARDALL: If the same spectrum is identified as different peptides by different engines, the corresponding PSMs are discarded. QVALUE: If the same spectrum is identified as different peptides by different engines, the PSM with the lowest *q*-value is kept. [b]Identified protein increase normalized by the number of corresponding proteins identified by Sequest only.

Table 8. Comparison of Our Method and *PeptideProphet/iProphet/ProteinProphet*

filter criteria[a]	method	peptides	unique peptides	target two-hit proteins	two-hit protein FDR (%)	target one-hit-wonders	one-hit-wonder FDR (%)
pepValue ≥0.842[b]	P	27913	1447	157	0	122	36.07
ipepValue ≥0.823[c]	iProphet	31101	1219	142	0	75	8
pepFdr ≤0.01	B	31880	1693	174	0.57	184	60.87
proValue ≥0.9[d]	P	0	0	156	1.92	53	3.77
proFdr ≤0.01	B	31327	1445	174	0.57	48	0
proValue ≥0.96[d]	P	0	0	152	1.32	20	10
proFdr ≤0.00	B	30908	1427	172	0	49	0

[a]pepValue, *PeptideProphet* probability; ipepValue, *iProphet* probability; pepFdr, peptide FDR; proValue, *ProteinPropet* probability; proFdr, protein FDR; B, our group-based method; P, *PeptideProphet/ProteinProphet*. [b]pepValue 0.842 corresponds to a peptide FDR threshold of 0.01 in the *PeptideProphet* results. [c]ipepValue 0.823 corresponds to a peptide FDR threshold of 0.01 in the *iProphet* results. [d]proValue 0.9/0.96 corresponds to a protein FDR threshold of 0.01/0.00 in the *ProteinProphet* results.

there was still a FDR of 1.32% for two-hit proteins and a FDR of 10% for one-hit-wonders.

DISCUSSION

We examined the factors that might affect the sensitivity of peptide identification. Our results showed that the PSM should be grouped not only by precursor charge and modification state but also by the number of missed internal cleavage sites, the number of protease termini, and the instrument type prior to FDR filtering. To improve the identification sensitivity of high-resolution proteomics data, a large mass tolerance in the search procedure and a narrow precursor tolerance that considers second/third isotopic ions during the filtering procedure are recommended.

One-hit-wonders are commonly observed in large-scale proteomics experiments but often have an unacceptably high FDR. Grouping candidate proteins by their unique peptide counts and filtering them separately by the assigned protein FDRs allowed our method to achieve high confidence in the identified one-hit-wonders.

PeptideProphet/iProphet/ProteinProphet analysis is widely used in proteomics research. As compared with the results from these programs using a high-resolution data set, our method showed a satisfactory performance. This indicates that simply grouping the PSMs into different categories and filtering them separately by user-assigned FDR thresholds can achieve a performance that is at least equal to machine-learning or statistic-based methods while allowing more flexibility to extend the method to other search engines.

We have implemented our method in *BuildSummary*, a software tool for assembling protein identifications in shotgun proteomics that is based on the target-decoy search strategy. This strategy allows for the fast and automated extraction of PSMs from Sequest, MASCOT, and X!Tandem search results, PSM filtering on the peptide or protein level at a given FDR threshold, and the integration of search results from different analytical protocols (e.g., different search engines, different search parameters, or different instruments). Since *PeptideProphet* generates a confidence probability for each PSM, it is also supported by *BuildSummary*. *BuildSummary*'s compatibility with other engines is currently being extended. *BuildSummary* can be downloaded for free from http://www.proteomics.ac.cn/software/proteomicstools/index.htm as part of the software suite *ProteomicsTools*.

ASSOCIATED CONTENT

Supporting Information

Description of human plasma dataset preparation and figures showing score distributions of PSMs before and after PSMs grouped by instrument type and two-hit proteins identified by integration of two search engine results. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: zr@sibs.ac.cn. Telephone: +86 215 492 0170. Fax: +86 215 492 0171.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation (30821065 and 91029301), the Basic Research Foundation (2010CB912102, 2011CB910200, and

2011CB910601), the CAS Project (KSCX2-YW-R-106 and KSCX2-YW-R-182), the SIBS Project (2009KIP212), and the Shanghai Government Foundation (10DZ1951202). The authors also gratefully acknowledge the support of the SA-SIBS scholarship program.

REFERENCES

(1) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367−72.

(2) Ding, Y.; Choi, H.; Nesvizhskii, A. I. Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics. *J. Proteome Res.* **2008**, *7* (11), 4878−89.

(3) Hsieh, E. J.; Hoopmann, M. R.; MacLean, B.; MacCoss, M. J. Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J. Proteome Res.* **2010**, *9* (2), 1138−43.

(4) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74* (20), 5383−92.

(5) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207−14.

(6) Tsur, D.; Tanner, S.; Zandi, E.; Bafna, V.; Pevzner, P. A. Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **2005**, *23* (12), 1562−7.

(7) Strader, M. B.; Tabb, D. L.; Hervey, W. J.; Pan, C.; Hurst, G. B. Efficient and specific trypsin digestion of microgram to nanogram quantities of proteins in organic-aqueous solvent systems. *Anal. Chem.* **2006**, *78* (1), 125−34.

(8) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4* (11), 923−5.

(9) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell Proteomics* **2005**, *4* (10), 1419−40.

(10) Alves, P.; Arnold, R. J.; Novotny, M. V.; Radivojac, P.; Reilly, J. P.; Tang, H. Advancement in Protein Inference from Shotgun Proteomics Using Peptide Detectability. *Pacific Symp. Biocomput.* **2007**, *12*, 409−420.

(11) Li, Y. F.; Arnold, R. J.; Li, Y.; Radivojac, P.; Sheng, Q.; Tang, H. A bayesian approach to protein inference problem in shotgun proteomics. *J. Comput. Biol.* **2009**, *16* (8), 1183−93.

(12) Gerster, S.; Qeli, E.; Ahrens, C. H.; Buhlmann, P. Protein and gene model inference based on statistical modeling in k-partite graphs. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (27), 12101−6.

(13) Serang, O.; MacCoss, M. J.; Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.* **2010**, *9* (10), 5346−57.

(14) Balgley, B. M.; Laudeman, T.; Yang, L.; Song, T.; Lee, C. S. Comparative Evaluation of Tandem MS Search Algorithms Using a Target-Decoy Search Strategy. *Mol. Cell Proteomics* **2007**, *6* (9), 1599−608.

(15) Gupta, N.; Pevzner, P. A. False discovery rates of protein identifications: a strike against the two-peptide rule. *J. Proteome Res.* **2009**, *8* (9), 4173−81.

(16) Resing, K. A.; Meyer-Arendt, K.; Mendoza, A. M.; Aveline-Wolf, L. D.; Jonscher, K. R.; Pierce, K. G.; Old, W. M.; Cheung, H. T.; Russell, S.; Wattawa, J. L.; Goehle, G. R.; Knight, R. D.; Ahn, N. G. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **2004**, *76* (13), 3556−68.

(17) Higgs, R. E.; Knierman, M. D.; Freeman, A. B.; Gelbert, L. M.; Patil, S. T.; Hale, J. E. Estimating the statistical significance of peptide identifications from shotgun proteomics experiments. *J. Proteome Res.* **2007**, *6* (5), 1758−67.

(18) Alves, G.; Wu, W. W.; Wang, G.; Shen, R. F.; Yu, Y. K. Enhancing peptide identification confidence by combining search methods. *J. Proteome Res.* **2008**, *7* (8), 3102−13.

(19) Searle, B. C.; Turner, M.; Nesvizhskii, A. I. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.* **2008**, *7* (1), 245−53.

(20) Edwards, N.; Wu, X.; Tseng, C.-W. An unsupervised, modelfree, machine-learning combiner for peptide identifications from tandem mass spectra. *Clin. Proteomics* **2009**, *5*, 23−36.

(21) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75* (17), 4646−58.

(22) Wu, Y. B.; Dai, J.; Yang, X. L.; Li, S. J.; Zhao, S. L.; Sheng, Q. H.; Tang, J. S.; Zheng, G. Y.; Li, Y. X.; Wu, J. R.; Zeng, R. Concurrent quantification of proteome and phosphoproteome to reveal system-wide association of protein phosphorylation and gene expression. *Mol. Cell Proteomics* **2009**, *8* (12), 2809−26.

(23) Mann, B.; Madera, M.; Sheng, Q.; Tang, H.; Mechref, Y.; Novotny, M. V. ProteinQuant Suite: a bundle of automated software tools for label-free quantitative proteomics. *Rapid Commun. Mass Spectrom.* **2008**, *22* (23), 3823−34.

(24) Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7* (1), 29−34.