Comparative Evaluation of Tandem MS Search Algorithms Using a Target-Decoy Search Strategy*Ⓢ

Brian M. Balgley‡§, Tom Laudeman¶, Li Yang‖, Tao Song‡, and Cheng S. Lee‖

Peptide identification of tandem mass spectra by a variety of available search algorithms forms the foundation for much of modern day mass spectrometry-based proteomics. Despite the critical importance of proper evaluation and interpretation of the results generated by these algorithms there is still little consistency in their application or understanding of their similarities and differences. A survey was conducted of four tandem mass spectrometry peptide identification search algorithms, including Mascot, Open Mass Spectrometry Search Algorithm, Sequest, and X! Tandem. The same input data, search parameters, and sequence library were used for the searches. Comparisons were based on commonly used scoring methodologies for each algorithm and on the results of a target-decoy approach to sequence library searching. The results indicated that there is little difference in the output of the algorithms so long as consistent scoring procedures are applied. The results showed that some commonly used scoring procedures may lead to excessive false discovery rates. Finally an alternative method for the determination of an optimal cutoff threshold is proposed. ***Molecular & Cellular Proteomics 6:1599–1608, 2007.***

Systems biology has come to occupy a central role in modern biological research. MS-based proteomics is an integral subset of the wide ranging technologies contributing to the systems biology toolbox. MS-based proteomics, in turn, encompasses several commonly used techniques, the most powerful of which is, arguably, high throughput fragmentation of complex protein mixtures that have been enzymatically digested and separated prior to their introduction into a tandem mass spectrometer. These fragmentation events (also referred to as tandem MS or MS^2 events) are submitted to any of a number of available peptide identification search algorithms. Results are returned in the form of peptide sequences with one or more scores with which to evaluate the likelihood that the resulting sequence is correct.

The first use of such an algorithm was reported by Eng *et al.* in 1994 (1). The algorithm, to be named Sequest, enabled the high throughput and automated interpretation of tandem mass spectra against a protein sequence library. The Mascot search algorithm, an outgrowth of the MOWSE (molecular weight search) project (2), was released in 1998 (3). More recently two open source search algorithms have become available including the Open Mass Spectrometry Search Algorithm (OMSSA)[1] (4) and X! Tandem (5) distributed by the National Center for Biotechnology Information (NCBI) and the Global Proteome Machine Organization, respectively.

Although each implementation is different, these search algorithms operate under the same general principles. Namely they first find peptides whose theoretical mass approximates the experimental mass determined by the mass spectrometer for the precursor ion within an *in silico* digest of the sequence library being used. The search space is typically limited by parameters including mass tolerance, enzyme specificity, numbers of missed cleavages, and amino acid modifications. The algorithms next consider the fragment ions generated within the tandem mass spectrometer. The fragment ion masses are similarly compared with an *in silico* fragmentation of the peptides that passed the first criteria. The fragment ions must approximate the masses of the theoretical fragment ion masses within a defined mass tolerance. Typically only a subset of possible fragmentation ions are considered depending on the type of mass spectrometer and the fragmentation mechanism used. For example, only *b* and *y* ions are generally considered in the case of ion trap mass spectrometers using collision-induced dissociation. The details of this implementation differ among the algorithms and are not documented in all cases. In addition, the methods used to assign scores are very different as described in a recent review (6).

As noted, each algorithm outputs one or more scores in response to a query for each tandem mass spectrum. In some cases these scores are dependent on the search parameter inputs and/or sequence library used in addition to the quality

From the ‡Calibrant Biosystems, Gaithersburg, Maryland 20878, ¶Information Technology and Communication, Academic Computing Health Sciences, University of Virginia, Charlottesville, Virginia 22908, and ‖Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742

Received, December 14, 2006, and in revised form, March 28, 2007

Published, MCP Papers in Press, May 28, 2007, DOI 10.1074/mcp.M600469-MCP200

[1] The abbreviations used are: OMSSA, Open Mass Spectrometry Search Algorithm; CIEF, capillary IEF; ΔCn, Sequest delta correlation number; E, expectation; FDR, false discovery rate; HUPO, Human Proteome Organization; nano-RPLC, nanoflow reversed-phase- LC; RSp, Sequest preliminary score ranking; Sp, Sequest preliminary score; Xcorr, Sequest cross-correlation score; TPP, TransProteomic Pipeline; XML, extensible markup language.

of the tandem mass spectrum. This capability is intended to allow the search engine to output a probability consistent with these parameters. However, not all search algorithms consider all parameters in their score assignment. A combination of a lack of documentation together with a general lack of understanding of the underlying algorithms contributes to inconsistent application of any given search algorithm within and across research groups. Ideally a search algorithm would perform an initial search of a relatively small portion of the submitted dataset to determine the appropriate search parameters such as precursor and fragment ion mass accuracy, number of missed cleavages, monoisotopic or average masses, etc. to remove this aspect of user bias. This is now becoming increasingly important as search results make their way into public repositories such as PeptideAtlas (7, 8) and the Global Proteome Machine database (9).

Another source of variability in the search output is in the generation of the list of tandem mass spectra to be searched, typically referred to as a peak list. Peak lists are generated by proprietary programs compatible with the proprietary data files of the manufacturer of the mass spectrometer on which the data were recorded. Open source programs for converting proprietary data files into standardized XML formats such as mzXML and mzData have recently become available as has an open source program to convert mzXML files to peak list files, mzXML2other. There are often as many if not more parameters available to define the creation of a peak list as there are parameters that define the search itself. Again a lack of documentation and understanding of the peak listing algorithms leads to inconsistent use. For example, spectral counting (10) is emerging as a robust and sensitive quantitation method; however, peak listing parameters that have a dramatic impact on the outcome of such a measure are rarely reported. One such parameter available to users of the Thermo peak listing program extract_msn.exe (embedded within Bioworks) determines how many scans of the same precursor mass, within a variable scan window, will be averaged together to create a single query. Certainly choosing a wider window will tend to lessen the number of potential identifications, influencing the quantitative measure relative to a more narrow window or no window at all. This parameter may also affect the outcome of a search as many scans averaged together may vary significantly from any single scan of that grouping. For example, low resolution mass spectrometers routinely used for proteomics studies use a relatively wide isolation window. For any given scan there may be another analyte of similar mass that is isolated for fragmentation together with the analyte ion of interest, resulting in a "distracted" fragmentation of this mixture of analytes, confounding interpretation by the search algorithm.

The final hurdle to realizing a usable dataset is the interpretation of the scores output by the search algorithms. A wide variety of differing scoring procedures have been implemented by differing groups with differing methods for determining false positive rates. This is especially true of the Sequest algorithm as the scores do not reflect the probability of the assignment as is the case with Mascot, OMSSA, and X! Tandem. Many scoring thresholds are used for Sequest, two of which were used by the Human Proteome Organization (HUPO) to interpret results from their Plasma Proteome Project (11) and are valuable as a reference point for that reason. The scoring procedures are referred to as "high confidence" and "low confidence" criteria. Sequest has recently started reporting a probability-based score when used from within the Bioworks Browser (Thermo). Mascot has always provided two criteria for search result evaluation, an identity score and a homology score. The identity score is to be used as a threshold for identification, and the homology score is to be used as a threshold for extensive homology. Mascot has recently started reporting an expectation value in addition to the other scores. OMSSA and X! Tandem both report expectation values as the primary score for result evaluation.

The trend toward probability-based scores is welcome. However, a user must still decide what is an appropriate cutoff criterion. A user might reasonably decide that an expectation or probability value of 0.05 or 0.01, each intended to correspond to a 5% or 1% likelihood of a random match, would be appropriate. However, we have found that these scores can vary substantially with differing search parameters, sequence libraries, and samples as has been reported previously (12, 13). Many groups outside of the search algorithm distributors themselves have made efforts to improve data evaluation methods (14–18), yet the use of these methods has not been widely adopted and certainly not standardized.

One approach that is able to simultaneously deal with variations introduced by samples, analytical techniques, data processing, and search parameters is the target-decoy sequence library search (19–21). In this method the protein sequence library to be searched is copied and reversed, and the reversed copy is appended to the end of the original sequence library. This creates a decoy sequence library within the search set with exactly the same number of proteins, sequence lengths, and amino acid composition as the real, or target, sequence library. Therefore, any variability that would affect a search will be mirrored in the search of the decoy library. The target-decoy search strategy permits an impartial initial assessment of search results and application of cutoffs based on the estimated false discovery rates (FDRs) determined from the search. It is important to note that the FDRs are only estimates in that a hit to a decoy sequence is only a proxy for a "true" false positive and is not itself a false positive, although here we refer to them as such. This initial assessment should be followed by manual inspection of spectra identifying proteins to which biological meaning is attributed. Those proteins should then be further validated by other methods.

Kapp *et al*. (22) have recently evaluated and compared a number of tandem MS search algorithms. However, peak list generation was performed using differing parameters for input into each of the search algorithms, and differing search parameters were applied, such as precursor and fragment ion mass tolerances. Also a validated dataset was established for comparison that may have been biased toward the search algorithms with which it was created. Finally comparisons were made based on a limited set of tandem mass spectra (<4,000).

In this study we evaluated four search algorithms: Mascot, OMSSA, Sequest, and X! Tandem. Sample, data acquisition, data processing (peak listing), search algorithm parameter, and sequence library variables were removed by using a single processed dataset, identical search parameters, and the same target-decoy sequence library in all searches. Results obtained using commonly applied scoring procedures were compared with those obtained using a consistent cutoff as determined by the target-decoy determined FDR. The sensitivity and specificity of each algorithm was evaluated as well as the overlap between algorithms of inferred protein identifications. Finally an alternative, simpler method for evaluating appropriate scoring criteria is proposed.

EXPERIMENTAL PROCEDURES

Microdissected Tissue Sample Preparation and Analysis—Sample preparation and analysis have been described in detail elsewhere (23, 24). Briefly fresh-frozen normal human ovarian epithelium obtained from Dr. Wenxin Zheng's laboratory at the University of Arizona was sectioned and microdissected; ~100,000 cells were procured. Proteins associated with cell pellets were extracted using sodium dodecyl sulfate detergent followed by denaturation, reduction, alkylation, digestion with trypsin, purification on a reversed-phase trap column, and lyophilization to dryness. Approximately 10 μg of the resulting digest was separated into discrete fractions by capillary IEF (CIEF) followed by nanoflow reversed-phase LC (nano-RPLC) coupled with ESI-MS/MS using a linear ion trap (LTQ, ThermoFinnigan, San Jose, CA). Precursor ions were scanned at a mass range of 400–1,400 *m/z*. Data-dependent scanning was enabled with the five most intense ions of the precursor ion scan selected for tandem MS with dynamic exclusion enabled and set to 18 s. MS and tandem MS scan times were set to a maximum of 100 ms. Automatic gain control target settings were 30,000 for full MS scans and 10,000 for MS^2 scans.

Data Analysis—Raw search data files were peak-listed using the version of extract_msn.exe distributed with Bioworks 3.3 (ThermoFinnigan). The command line argument used to run extract_msn was: > -F1 -L0 -B500 -T3500 -M1.00 -C2 -S1 -I10 -C0 -G1.

A detailed description of the command line arguments from the manufacturer may be found in the supplemental methods section. The 154,973 dta files produced by extract_msn were directly used as the input to Sequest. Additionally the dta files were concatenated and converted to mgf file format using an in-house script (convert.pl). This single 1-gigabyte mgf file was used as the input file to the Mascot, OMSSA, and X! Tandem search engines.

The human Swiss-Prot sequence library (November 2004 build) obtained from the European Bioinformatics Institute (ftp.ebi.ac.uk/pub/databases/SPproteomes/fasta/proteomes) was used to create a target-decoy sequence library for searching. An in-house script, rev_swiss_decoyed.pl, was used to copy the sequence library, append the string "_REVERSE" to each protein accession in the copy, reverse every protein sequence string in the copy, and concatenate the now decoyed copy to the original.

Four search engines were evaluated: Mascot 2.0, OMSSA 1.1.0, Sequest/Bioworks 3.3, and X! Tandem 2006.6.1.1. All searches were run locally on a Dell Optiplex GX620 with a 2.8-GHz Pentium D processor and 2 gigabytes of RAM. All searches used the following search parameters: precursor ion mass tolerance, ±1.5 Da; fragment ion mass tolerance, ±0.5 Da; fully tryptic enzyme specificity; one missed cleavage; monoisotopic precursor mass (Mascot, OMSSA, and Sequest); monoisotopic fragment ion mass (OMSSA and Sequest); a fixed modification of cysteine carbamidomethylation; and a variable modification of methionine oxidation. An additional parameter unique to OMSSA was used that requires that one of a variable number of the most intense fragment ions match those of the theoretical peptide. In this case the parameter was set to 5. It should also be noted that X! Tandem applies the 3-Da precursor mass window differently than the other algorithms. The window is set such that the search is defined by a window from −0.5 to +2.5 Da around the precursor ion rather than ±1.5 Da around the precursor in the case of the other algorithms.

Result files were parsed to extract dta name, query number, peptide sequence, protein accession, experimental and theoretical masses, charge, and algorithm-specific scores. All parsers were written in Perl. OMSSA XML result files, omx, were converted to csv format using omx2csv, available from Computational Systems Biology at the University of Virginia. Mascot dat result files were converted to PepXML using the converter available with the TransProteomic Pipeline (TPP) distribution, dat2xml. The TPP is available from the Seattle Proteome Center. PepXML files were converted to csv format using pepxml2csv, also available from Computational Systems Biology at the University of Virginia. Sequest results were exported to Excel format using the built-in feature within Bioworks. X! Tandem XML result files were converted to PepXML using the TPP converter tandem2xml. The parsed data were loaded into a Microsoft Access (2003) database for subsequent analysis and comparison. False discovery rates were calculated according to the method of Elias *et al.* (20): decoy hits were multiplied by 2, and the product was divided by the sum of the target and decoy hits. A table listing all queries that meet the 1% MS^2 FDR threshold of the algorithms discussed is included in the form of an Excel spreadsheet as Supplemental Table 1.

Target and decoy peptide hits were redundantly assigned to all target and decoy Swiss-Prot entries to which they mapped. MS^2 hits indicate a search query (peak list) that results in a peptide assignment at some score. Distinct peptides indicate peptides with differing sequences. Distinct proteins indicate differing Swiss-Prot entries (identifiers). A protein (Swiss-Prot entry) was counted if a single peptide mapped to it whether or not that peptide also mapped to other entries.

RESULTS

Tandem mass spectra were collected over the course of a single proteomics analysis of normal human ovary epithelial cells consisting of 18 CIEF fractions further analyzed by nano-RPLC interfaced with ESI-tandem MS. A common peak list file generated from LTQ raw files was used as input into four tandem mass spectrometry peptide-matching search algorithms: Mascot, OMSSA, Sequest, and X! Tandem. Common search parameters were used for each search including the sequence library, in this case a target-decoy version of the human subset of Swiss-Prot. The score distribution for each of the algorithms is shown in Fig. 1. Hits to both target and decoy peptides are shown.

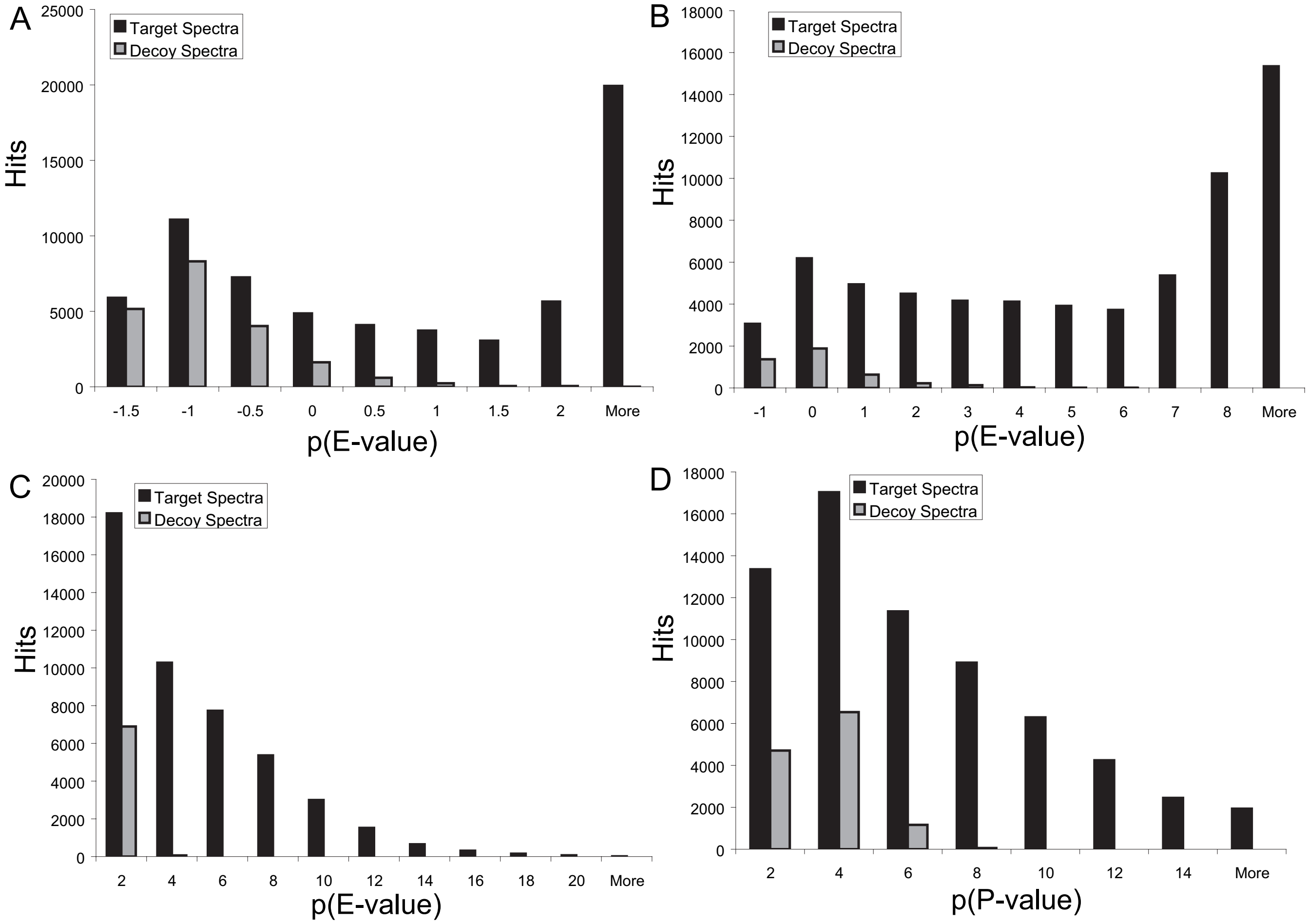

FIG. 1. **Expectation or probability score distributions of the four search algorithms.** Scores are binned based on the −log of the expectation or probability value, that is, higher scores are better. *Panel A*, Mascot E-value scores; *panel B*, OMSSA E-value scores; *panel C*, X! Tandem E-value scores; *panel D*, Sequest *p* value scores.

The results of this search are summarized in Table I. Results are shown only at a 1% MS^2 FDR for each algorithm with the exception of Mascot and Sequest. Mascot, OMSSA, and X! Tandem scores are reported as expectation (E) values, the same scoring method used for Blast local sequence alignment scoring. We observed Mascot scores on a scale from 10^{-2} to 5×10^2, OMSSA scores from 10^{-300} to 2×10^2 (upper limit is user-adjustable), X! Tandem scores from 10^{-20} to 10^{-1} (using a "Find models" setting of 0), and Sequest probability score ranges from 0 to 1. Results obtained across the scoring ranges evaluated may be found in Supplemental Table 2. An expectation value is defined as the probability that the predicted event (in this case a peptide sequence matching an experimental tandem mass spectrum) would occur by chance should the trial be repeated many times. For example, an E-value of 1 indicates that a true outcome is as equally likely as a false outcome, whereas an E-value of 0.01 indicates the predicted outcome will occur by chance one time in 100 given many trials. According to probability theory an E-value is calculated based on the probability of an event occurring at random when a trial of that event is repeated many times. A probability value measures the likelihood that an event will occur given a single trial of that event. For example, given a peptide with a probability assignment of 1×10^{-5} with 10^4 possible matches in the sequence library within the defined search parameters, an E-value of 0.1 would result.

From Fig. 1 it is clear that hits to target and decoy spectra do not conform to the results predicted by the expectation or probability values. At an expectation value of 1 it would be expected that the number of decoy spectra would be half of the number of target spectra (see the 0 bin in Fig. 1: $-\log_{10}(1) = 0$). This is in concordance with an E-value of 1 predicting an equal chance of a random occurrence and the method used to calculate false positives by the target-decoy method: 2 × decoy hits = false positive hits. This occurs for X! Tandem near 0.1. For Mascot this occurs at an E-value near 3, and for OMSSA this occurs at an E-value near 10. This points to the difficulty in predicting the likelihood of the gas-phase kinetic fragmentation mechanisms across the possible peptides (25, 26). Presumably not included in the search

TABLE I

Summary of search results and Sequest scoring matrices

Algorithm	Score	MS^2 FDR	Target			Decoy			Protein FDR	Estimated true proteins
			MS^2 hits	Distinct peptides	Proteins	MS^2 hits	Distinct peptides	Proteins		
		%							%	
Mascot	Identity	0.60	26,596	9,173	1,880	79	49	48	5.00	1,784
Mascot	Identity and homology	0.20	25,802	8,952	1,835	24	18	18	1.90	1,799
Mascot expectation	0.125	1.00	29,463	9,942	2,001	154	98	97	9.20	1,807
OMSSA (expectation)	0.03	1.00	48,328	13,512	2,725	246	127	127	8.90	2,471
X! Tandem (expectation)	0.025	1.00	31,367	10,762	2,089	157	120	119	10.80	1,851
Sequest *p* (probability)	5×10^{-6}	1.09	27,431	9,806	2,082	151	112	110	10.00	1,862
Sequest HUPO high	HUPO high confidence	4.80	36,844	11,892	2,902	908	872	634	35.90	1,634
Sequest HUPO low	HUPO low confidence	13.50	43,199	12,348	4,508	3,128	2,776	2,095	63.50	318
Sequest Xcorr FDR	Xcorr-based FDR	1.20	24,575	10,460	2,007	153	152	105	9.90	1,797

		ΔCn	Xcorr			RSp
			1+	2+	3+	
	HUPO high confidence	0.1	1.9	2.2	3.75	4
	HUPO low confidence	0.1	1.5	2	2.5	NA[a]
	Xcorr-based FDR	0.1	2.9	3.1	3.7	NA

[a] NA, not applicable.

algorithm predictions are the real world "features" existing in a sample. These include any peptides whose characteristics are not defined by the search variables (missed cleavages, post-translational modifications including glycosylations, other charge states, protein contaminants from other species, etc.) as well as common non-peptide contaminants such as plastic polymers and detergents. It should also be noted that the sequence libraries typically used are continually revised as proteins are manually curated and as ORF prediction algorithms change. Also they almost never contain the sequences predicted by the immense variety of single nucleotide polymorphisms. These factors and others contribute to the difficulty in correctly modeling the statistical framework for producing an accurate expectation value. For this reason, a target-decoy search approach has utility in that these features are, theoretically, just as likely to hit a decoy sequence as a target sequence (21).

Mascot provides the user with two scores: an ion score and an E-value. The ion score is compared with two thresholds that are calculated independently for each peptide, the homology score and identity score. Traditionally the identity score is the reported threshold used in most laboratories. The homology score is typically lower than the identity score in the case of longer peptides and higher in the case of shorter peptides. Table I shows the results of a Mascot search using the identity score as the threshold criteria (Mascot identity), using both the identity and homology scores as criteria (Mascot identity and homology) and using the E-value at a threshold found to return a 1% MS^2 FDR. The second scoring criteria using both the identity and homology scores is shown because this is an effective way to remove short peptides, which have a higher tendency to be false positives than longer peptides. This results in a higher confidence dataset with an increase in the estimated number of total proteins discovered.

Sequest presents a special case because many empirically determined threshold scores are used in research laboratories throughout the world. Displayed in Table I are some commonly used thresholds put forth by HUPO, including both a low confidence (Sequest HUPO low) and high confidence (Sequest HUPO high) set of thresholds. Sequest results are also shown using thresholds empirically determined to return a 1% MS^2 FDR for each charge state (Sequest Xcorr FDR). Sequest offers two separate scores. The first score is actually a composite of several scores: The cross-correlation score (Xcorr) is a cross-correlation measure of the theoretical and experimental spectra. Typically a different Xcorr threshold is used for each charge state to reflect the difference in probability of a random match to each. A score measuring the difference between the top two candidate spectra is also given in the form of the delta correlation value (ΔCn) where the scores are normalized such that the top score is set equal to 1 and the difference is taken. Less frequently used is the preliminary score (Sp) and the preliminary score ranking (RSp) that are scores calculated first as a filter to restrict the number of spectra to subject to cross-correlation scoring. The second Sequest score is a probability value score offered in the most recent version of the Bioworks software (Sequest *p*).

Encouragingly the results produced by using a 1% MS^2 FDR are very consistent for each of the algorithms. Mascot, X! Tandem, Sequest *p*, and Sequest Xcorr FDR yielded 9,942–10,762 distinct peptides and ~2,000 proteins, whereas OMSSA identified 13,512 distinct peptides and over 2,700 proteins. The protein FDR for each algorithm is consistent as well, yielding values between 8.9 and 10.8% at a controlled 1% MS^2 FDR. The trend across a range of MS^2 FDR values is shown in Fig. 2. Even across a very wide range there is little

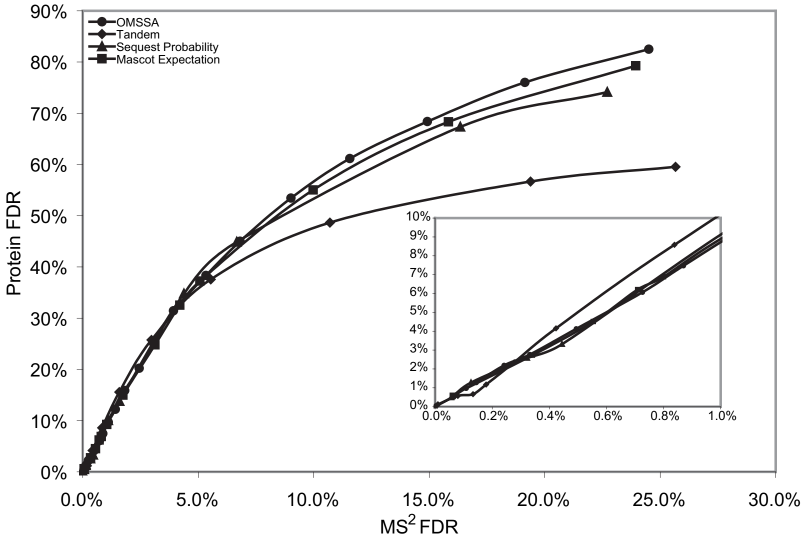

FIG. 2. **Correlations of protein FDR with MS^2 FDR from each search engine.**

difference between search algorithms in the ratio of MS^2 to protein false discovery rates. The figure highlights the importance of strictly controlling the MS^2 FDR when reporting results as small changes in this value have a large impact on protein FDR. For example, a 0.1% MS^2 FDR results in protein FDRs of about 1% for all algorithms. A 1% MS^2 FDR increases the protein FDR to 8–11%. A 5% MS^2 FDR yields a 35–40% protein FDR, and a 10% MS^2 FDR results in protein FDRs in excess of 50%, meaning that half the proteins reported are likely false positives. This has been shown previously in the context of different Sequest Xcorr thresholds (27).

The same effect was observed when comparing Sequest Xcorr-based thresholds at HUPO high and low confidence settings. The HUPO high confidence setting returned an MS^2 FDR of about 5% and a protein FDR of about 36%, whereas the low confidence cutoffs yielded a 13.5% MS^2 FDR and a 63.5% protein FDR. At these values the number of distinct peptide identifications increased ~13 and 18%, respectively, from the values obtained using the cross-correlation-based cutoffs at a 1% MS^2 FDR. However, protein identifications increased substantially to over 2,900 proteins in the case of the high confidence settings and over 4,500 proteins for the low confidence settings, increases of ~50 and 125%, respectively. Correspondingly the number of decoy protein identifications increased ~500 and almost 19,000%, respectively. In this case the low confidence thresholds are clearly set too low, significantly and negatively impacting the quality of search results and consequently the reproducibility of the analysis. The high confidence settings are an improvement; however, a protein FDR of 36% will also make meaningful data interpretation difficult and will confound efforts to obtain reproducible identifications across multiple runs.

In contrast, the widely used Mascot identity score is a relatively conservative threshold, yielding an MS^2 FDR of 0.6% and a protein FDR of 5.0%. Combined with the homology score as an additional threshold Mascot returned even more conservative results with a 0.2% MS^2 FDR and 1.9% protein FDR. However, as can be seen from the estimated number of true proteins identified (equal to target proteins − 2 × decoy proteins), the Mascot identity and homology score yielded more estimated true proteins than the Sequest Xcorr-based scoring methods.

In all cases, combining a 1% MS^2 FDR with the requirement for two distinct peptides per protein lowered the protein FDR to an effective rate of 0%. Mascot E, OMSSA, and X! Tandem each discovered only one decoy protein with two distinct peptides, whereas Sequest *p* discovered two. This method enables production of datasets with very high confidence identifications. However, the method also greatly increases the number of false negatives. For example, OMSSA discovered 127 decoy proteins, indicating that twice as many, 254, were predicted in the set to be false positives. Given that OMSSA discovered 899 proteins with a single distinct peptide, requiring two distinct peptides per protein would eliminate 645 putative true identifications, not an insignificant number.

Other measures of data quality are the average number of MS^2 hits and distinct peptide identifications per protein (Table II). Increases in either of these measures is interpreted as an increase in the confidence of a protein identification (16). In particular, distinct peptide identifications are valued as they lead to increased confidence in and sequence coverage of the predicted protein. By this measure OMSSA and X! Tandem outperform Sequest and Mascot with respect to MS^2 identifications per protein. OMSSA identified 17.8 MS^2 events per protein an average, X! Tandem identified 15.0 MS^2 events, Mascot E identified 14.7 MS^2 events, and Sequest *p* identified 13.2 MS^2 events. The number of MS^2 identifications becomes very important when implementing spectral counting-based quantification because the expression levels of each protein are determined by the number of MS^2 identifications. The algorithms are much more comparable with regard to the number of distinct peptides identified per protein with X! Tandem and Sequest Xcorr FDR yielding 5.2 distinct peptides per protein on average, Mascot E and OMSSA yielding 5.0, and Sequest *p* resulting in 4.7. The lower averages returned by the HUPO scoring thresholds are clearly a result of the high protein false positive rates resulting from use of these scoring procedures. As false positive protein identifications occur randomly throughout the sequence library, false positives do not accumulate at any given protein or subset of proteins, resulting in a large number of proteins identified with a single distinct peptide identification. This leads to lower averages of MS^2 hits per protein and distinct peptides per protein. As shown in Table I, the number of decoy distinct peptide identifications barely exceeds the number of decoy protein identifications.

TABLE II

Comparison of MS^2 and distinct peptide hits per protein by algorithm at various thresholds

Algorithm: threshold	MS^2 hits per protein	Distinct peptide hits per protein
Mascot: identity	14.1	4.9
Mascot: identity and homology	14.1	4.9
Mascot: expectation, 1% MS^2 FDR	14.7	5.0
OMSSA: expectation, 1% MS^2 FDR	17.7	5.0
X! Tandem: expectation, 1% MS^2 FDR	15.0	5.2
Sequest: probability, 1% MS^2 FDR	13.2	4.7
Sequest: HUPO high	12.7	4.1
Sequest: HUPO low	9.6	2.7
Sequest: Xcorr FDR	12.2	5.2

OMSSA delivered the best performance in terms of MS^2 discoveries regardless of MS^2 FDR (Fig. 3). At a 1% MS^2 FDR OMSSA returned 35% more MS^2 hits than X! Tandem, the next best performer. Even at very low false discovery rates ($<$0.1%) OMSSA performed with significantly higher sensitivity than Mascot, Sequest, or X! Tandem. Overall OMSSA returned 48,328 MS^2 hits, 13,512 distinct peptides, and 2,725 proteins with an average of 17.8 MS^2 hits and 5.0 distinct peptides per protein, respectively. Furthermore when evaluating proteins by the number of distinct peptides mapping to each, OMSSA similarly returned the best performance with 1,435 proteins (53%) having three or more distinct peptides, 391 (14%) having two distinct peptides, and 899 (33%) having a single distinct peptide (Table III).

Table IV documents the number of proteins identified by each algorithm both alone and in combination with each of the other algorithms. This is shown graphically by Venn diagram in Fig. 4 where *panel A* shows proteins identified by at least one distinct peptide and *panel B* shows proteins identified by at least two distinct peptides (by that algorithm). Not all over-

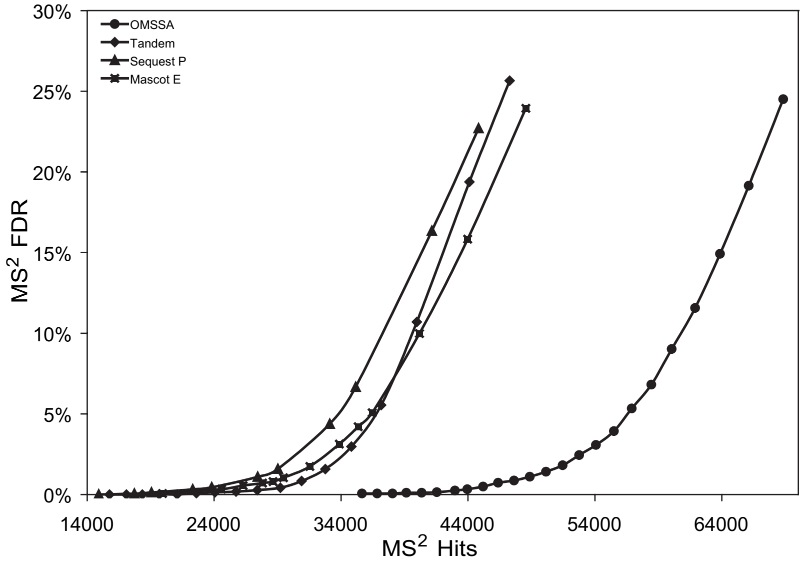

FIG. 3. **Plots of MS^2 hits *versus* MS^2 FDR from each search algorithm.**

TABLE III

Comparison of distinct peptides per protein by algorithm

	Number of proteins with *n* distinct peptides			Percentage of proteins with *n* distinct peptides		
	>2	2	1	>2	2	1
					%	
Mascot: expectation, 1% MS^2 FDR	935	335	731	47	17	37
OMSSA: expectation, 1% MS^2 FDR	1,435	391	899	53	14	33
X! Tandem: expectation, 1% MS^2 FDR	1,008	338	743	48	16	36
Sequest: probability, 1% MS^2 FDR	926	341	814	46	17	41

TABLE IV
Detail of overlapping protein identifications by algorithm

Algorithm(s)	Number of proteins
Number of proteins with one or more distinct peptides	
Found by all algorithms	1,671
X! Tandem	128
OMSSA	573
Sequest	83
Mascot	82
X! Tandem and OMSSA	116
X! Tandem and Sequest	8
X! Tandem and Mascot	16
OMSSA and Sequest	100
OMSSA and Mascot	42
Sequest and Mascot	5
X! Tandem, OMSSA, and Sequest	45
X! Tandem, OMSSA, and Mascot	98
X! Tandem, Sequest, and Mascot	7
OMMSA, Sequest, and Mascot	80
Total proteins discovered	3,054
Number of proteins with two or more distinct peptides	
Found by all algorithms	1,090
X! Tandem	47
OMSSA	405
Sequest	10
Mascot	15
X! Tandem and OMSSA	76
X! Tandem and Sequest	2
X! Tandem and Mascot	5
OMSSA and Sequest	54
OMSSA and Mascot	24
Sequest and Mascot	2
X! Tandem, OMSSA, and Sequest	45
X! Tandem, OMSSA, and Mascot	79
X! Tandem, Sequest, and Mascot	2
OMMSA, Sequest, and Mascot	53
Total proteins discovered	1,909

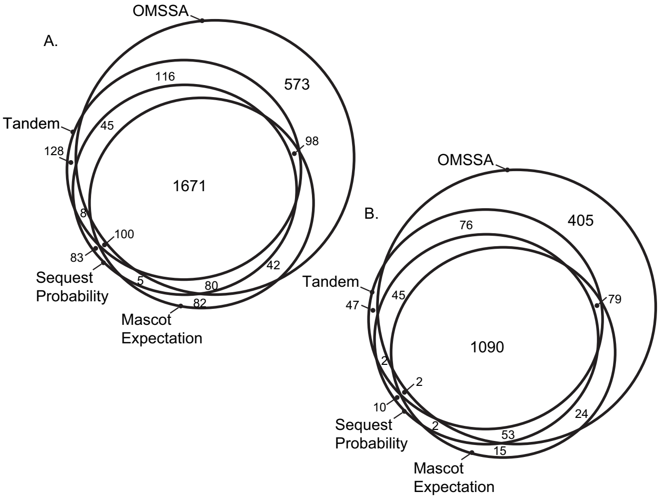

FIG. 4. **Overlap of proteins identified by search algorithms.** *A*, proteins identified by one or more distinct peptides; *B*, proteins identified by two or more distinct peptides.

TABLE V
Comparison of algorithm sensitivity and specificity

TP, true positive; TN, true negative; FP, false positive; FN, false negative.

	TP	TN	FP	FN	Relative sensitivity	Relative specificity
					%	%
Mascot	1,904	9,430	97	751	71.7	99.0
OMSSA	2,598	9,430	127	57	97.9	98.7
Sequest	1,889	9,430	110	766	71.1	98.8
X! Tandem	1,970	9,430	119	685	74.2	98.8

laps are displayed in Fig. 4. There was very good agreement between the algorithms on a core set of identified proteins with all algorithms identifying 1,671 or 1,090 proteins with at least one or two distinct peptides, respectively. Overall 2,761 or 1,837 proteins were identified by at least two of the four algorithms with at least one or two distinct peptides, respectively. OMSSA uniquely identified 573 proteins by one or more distinct peptides or 405 proteins by two or more distinct peptides. In contrast, the numbers of proteins uniquely found by Mascot, Sequest, and X! Tandem were all significantly lower than that of OMSSA alone. It is worthwhile to note that the 573 proteins uniquely identified by OMMSA were identified by a total of 2,468 distinct peptides (see supplemental data). This indicates an average of 4.3 distinct peptides per protein, which is only slightly lower than the 5.0 distinct peptides per protein overall average.

Sensitivity and specificity may be calculated from the numbers of positive and negative measurements, both true and false, in a dataset. In this case the true protein components of the sample are unknown; however, we may use the algorithm-based protein assignments and the predicted number of false positive protein assignments to evaluate the relative sensitivity and relative specificity of each algorithm. The results are shown in Table V. In this case the number of true negatives is set equal to the number of proteins in the Swiss-Prot protein sequence library (12,484) minus the number of proteins discovered by any algorithm (2,655). This latter number is derived from the distinct set of the total number of proteins found by all algorithms (3,054) minus the distinct set of false positive proteins found by all algorithms (399). The number of false negatives is the number of proteins discovered by any algorithm minus the number of proteins discovered by that algorithm. The number of true positives is the number of protein identifications minus the number of predicted false positives from Table I. Through the use of the target-decoy search approach the relative specificity exhibited by all algorithms is excellent at roughly 99%. The relative sensitivities of Mascot and Sequest are similar at about 71%, whereas X! Tandem yields 74% and OMSSA 98%.

An additional observation was made with respect to the ratio of MS^2 discoveries to proteins. For this analysis the ratio reached a maximum corresponding to the 1.0% MS^2 FDR for

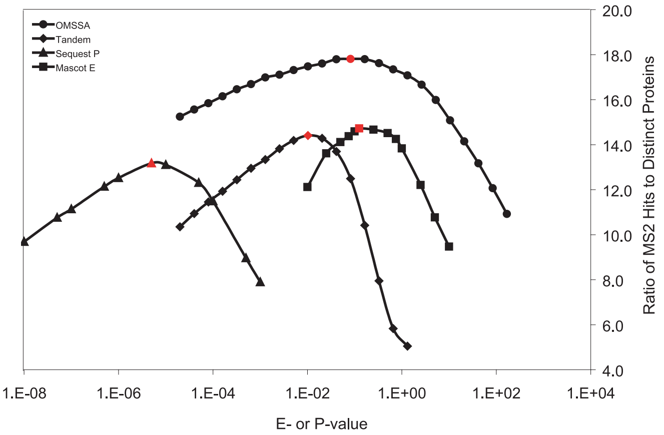

FIG. 5. **Ratio of MS^2 hits to distinct protein identifications as a function of E-value or *p* value for each search algorithm.** Data points highlighted in *red* for all curves correspond to a 1% MS^2 FDR.

each of the algorithms (Fig. 5). The contrast in slopes is the result of using a fixed set of E-values to sample across the results of algorithms that output scores across significantly differing scales. The 1% MS^2 FDR reflects the maximum sensitivity *versus* specificity of an algorithm as observed previously (28), and the MS^2 to protein ratio maximum may offer a method for groups not equipped to perform the decoy sequence library search and follow-on analyses documented herein to evaluate search results and justify reporting and publication decisions. This reinforces the rationale for reporting results at a 1% MS^2 FDR as standard practice. However, this raises the issue of whether the estimated protein FDRs, measured at ~ 10%, are too high. This depends greatly on the intended use of the results. When the results are being used to construct a protein reference library for a sample it is probably best to err on the side of caution and to use a strict filtering criteria such as a 1% MS^2 FDR coupled with a requirement for two distinct peptides per protein. In the cases where the results are being used in conjunction with other, complementary experimental methods, it may be better to use those results as a filter rather than being too strict at the initial, bioinformatics level, which could lead to a large false negative rate.

DISCUSSION

We demonstrated that four commonly used peptide identification search algorithms perform acceptably and more or less comparably when the results are evaluated through the use of a target-decoy protein sequence search library strategy. These results were obtained by analyzing the complex protein mixture extracted from a human tissue by CIEF-nano-RPLC-ESI-MS/MS using a linear ion trap mass spectrometer. A large number of tandem mass spectra (155,973) were collected and output to a peak list file that served as a common input to all search algorithms. Common parameters were used to perform each of the searches, and all searches used an identical target-decoy protein sequence library. These measures served to limit result variability to that derived from the search algorithms themselves. These results also limit how these data may be interpreted. For example, search algorithms may perform differently when using tandem mass spectra resulting from different dissociation mechanisms or different tandem mass spectrometers. It is important to note that only the peptide scoring function of the search algorithms is evaluated in this report. The postsearch algorithms that create the output report(s) for each search platform are not considered. Therefore, results displayed by any of the search platforms may be different from the underlying peptide output evaluated in this report. Needless to say, the output reports displayed derive their input data from the algorithms evaluated here.

The open source OMSSA algorithm from NCBI is exceptional in that its performance significantly exceeded that of each of the other algorithms by almost every measure. Furthermore the large number of MS^2 hits achieved by OMSSA may facilitate the implementation of spectral counting-based quantification approaches. The proteomics community would benefit were OMSSA scoring to be added as a pluggable score (29) in X! Tandem searches or if OMSSA were to be integrated as a search option into the Institute for Systems Biology's TransProteomic Pipeline project.

Regardless of the search algorithm a group may choose to use, it is important that the community arrive at some common standard for data evaluation and reporting. A target-decoy search strategy is one procedure that, as shown, produces relatively reproducible results from different algorithms for the same input data. Evaluating data at the maximum ratio of MS^2 hits per protein is another possibility that would be relatively simple for each of the search algorithm makers to implement in their data viewing and reporting applications. Both of these implementations have the benefit of being sensitive to false positives. That is, the data reported will be impacted by all of the variables leading up to the search from the sample used, to the preparation and analysis methods, to the search parameters entered. This is useful both because false positive rates vary sample to sample and also because some algorithms do not consider all of the search variable parameters in their probability or expectation scoring functions. In this era of high throughput global proteomics studies and cross-platform evaluations it is critical that common data evaluation and reporting procedures are used and that their use is as transparent and standardized as possible.

* Portions of this work were supported by NCI, National Institutes of Health Grants CA107988 and CA103086 and National Center for Research Resources Grants RR021862 and RR021239. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

§ To whom correspondence should be addressed: Calibrant Biosystems, 910 Clopper Rd., Suite 220N, Gaithersburg, MD 20878. Tel.: 301-977-7900 (ext. 14); Fax: 301-977-7981; E-mail: brian.balgley@calibrant.com.

REFERENCES

1. 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
2. Pappin, D. J. C., Rahman, D., Hansen, H. F., Bartlet-Jones, M., Jeffery, W., and Bleasby, A. (1996) Chemistry, mass spectrometry and peptide-mass databases: evolution of methods for the rapid identification and mapping of cellular proteins. *Mass Spectrom. Biol. Sci.* 135–150
3. 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
4. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.* **3,** 958–964
5. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467
6. Sadygov, R. G., Cociorva, D., and Yates, J. R., III (2004) Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **1,** 195–202
7. Desiere, F., Deutsch, E. W., Nesvizhskii, A. I., Mallick, P., King, N. L., Eng, J. K., Aderem, A., Boyle, R., Brunner, E., Donohoe, S., Fausto, N., Hafen, E., Hood, L., Katze, M. G., Kennedy, K. A., Kregenow, F., Lee, H., Lin, B., Martin, D., Ranish, J. A., Rawlings, D. J., Samelson, L. E., Shiio, Y., Watts, J. D., Wollscheid, B., Wright, M. E., Yan, W., Yang, L., Yi, E. C., Zhang, H., and Aebersold, R. (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6,** R9
8. Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., Chen, S., Eddes, J., Loevenich, S. N., and Aebersold, R. (2006) The PeptideAtlas project. *Nucleic Acids Res.* **34,** D655–D658
9. Craig, R., Cortens, J. P., and Beavis, R. C. (2004) Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **3,** 1234–1242
10. Liu, H., Sadygov, R. G., and Yates, J. R., III (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76,** 4193–4201
11. Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W., Menon, R., Hermjakob, H., Apweiler, R., Haab, B. B., Simpson, R. J., Eddes, J. S., Kapp, E. A., Moritz, R. L., Chan, D. W., Rai, A. J., Admon, A., Aebersold, R., Eng, J., Hancock, W. S., Hefta, S. A., Meyer, H., Paik, Y. K., Yoo, J. S., Ping, P., Pounds, J., Adkins, J., Qian, X., Wang, R., Wasinger, V., Wu, C. Y., Zhao, X., Zeng, R., Archakov, A., Tsugita, A., Beer, I., Pandey, A., Pisano, M., Andrews, P., Tammen, H., Speicher, D. W., and Hanash, S. M. (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **5,** 3226–3245
12. Cargile, B. J., Bundy, J. L., and Stephenson, J. L. (2004) Potential for false positive identifications from large databases through tandem mass spectrometry. *J. Proteome Res.* **3,** 1082–1085
13. Qian, W. J., Liu, T., Monroe, M. E., Strittmatter, E. F., Jacobs, J. M., Kangas, L. J., Petritis, K., Camp, D. G., II, and Smith, R. D. (2005) Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J. Proteome Res.* **4,** 53–62
14. Moore, R. E., Young, M. K., and Lee, T. D. (2002) Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **13,** 378–386
15. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392
16. MacCoss, M. J., Wu, C. C., and Yates, J. R., III (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* **74,** 5593–5599
17. Cargile, B. J., Bundy, J. L., Freeman, T. W., and Stephenson, J. L. (2004) Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *J. Proteome Res.* **3,** 112–119
18. Rudnick, P. A., Wang, Y., Evans, E. L., Lee, C. S., and Balgley, B. M. (2005) Large scale analysis of MASCOT results using a mass accuracy-based threshold (MATH) effectively improves data interpretation. *J. Proteome Res.* **4,** 1353–1360
19. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2,** 43–50
20. Elias, J. E., Haas, W., Faherty, B. K., and Gygi, S. P. (2005) Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* **2,** 667–675
21. 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* **3,** 207–214
22. Kapp, E. A., Schutz, F., Connolly, L. M., Chakel, J. A., Meza, J. E., Miller, C. A., Fenyo, D., Eng, J. K., Adkins, J. N., Omenn, G. S., and Simpson, R. J. (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity of analysis. *Proteomics* **5,** 3475–3490
23. Wang, Y., Rudnick, P. A., Evans, E. L., Zhuang, Z., Li, J., DeVoe, D. L., Lee, C. S., and Balgley, B. M. (2005) Proteome analysis of microdissected tumor tissue using a capillary isoelectric focusing-based multidimensional separation platform coupled with ESI-tandem MS. *Anal. Chem.* **77,** 6549–6556
24. Wang, W., Guo, T., Rudnick, P. A., Song, T., Li, J., Zhuang, Z., Zheng, Z., DeVoe, D. V., Lee, C. S., and Balgley, B. M. (2007) Membrane proteome analysis of microdissected ovarian tumor tissues using capillary isoelectric focusing/reversed-phase liquid chromatography-tandem MS. *Anal. Chem.* **79,** 1002–1009
25. Stein, S. E., and Heller, D. N. (2006) On the risk of false positive identification using multiple ion monitoring in qualitative mass spectrometry: large-scale intercomparisons with a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.* **17,** 823–835
26. Ahn, N. G., Shabb, J. B., Old, W. M., and Resing, K. A. (2007) Achieving in-depth proteomics profiling by mass spectrometry. *ACS Chem. Biol.* **2,** 39–52
27. Liu, T., Qian, W. J., Gritsenko, M. A., Xiao, W., Moldawer, L. L., Kaushal, A., Monroe, M. E., Varnum, S. M., Moore, R. J., Purvine, S. O., Maier, R. V., Davis, R. W., Tompkins, R. G., Camp, D. G., and Smith, R. D.; Inflammation and the Host Response to Injury Large Scale Collaborative Research Program (2006) High dynamic range characterization of the trauma patient plasma proteome. *Mol. Cell. Proteomics* **5,** 1899–1913
28. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P., and Gygi, S. P. (2004) Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* **22,** 214–219
29. MacLean, B., Eng, J. K., Beavis, R. C., and McIntosh, M. (2006) General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* **22,** 2830–2832