DOI 10.1007/s12177-009-9042-6

Techniques for accurate protein identification in shotgun proteomic studies of human, mouse, bovine, and chicken lenses

Phillip A. Wilmarth · Michael A. Riviere ·
Larry L. David

Received: 5 September 2009 / Accepted: 6 November 2009 / Published online: 12 December 2009

Abstract Analysis of shotgun proteomics datasets requires techniques to distinguish correct peptide identifications from incorrect identifications, such as linear discriminant functions and target/decoy protein databases. We report an efficient, flexible proteomic analysis workflow pipeline that implements these techniques to control both peptide and protein false discovery rates. We demonstrate its performance by analyzing two-dimensional liquid chromatography separations of lens proteins from human, mouse, bovine, and chicken lenses. We compared the use of International Protein Index databases to UniProt databases and no-enzyme SEQUEST searches to tryptic searches. Sequences present in the International Protein Index databases allowed detection of several novel crystallins. An alternate start codon isoform of βA4 was found in human lens. The minor crystallin γN was detected for the first time in bovine and chicken lenses. Chicken γS was identified and is the first member of the γ-crystallin family observed in avian lenses.

Keywords Bioinformatics · False discovery rates · Shotgun proteomics · Lens crystallins

Electronic supplementary material The online version of this article (doi:10.1007/s12177-009-9042-6) contains supplementary material, which is available to authorized users.

P. A. Wilmarth (✉) · M. A. Riviere · L. L. David
Department of Biochemistry and Molecular Biology,
School of Medicine, Oregon Health & Science University,
3181 SW Sam Jackson Park Road,
Portland, OR 97239, USA
e-mail: wilmarth@ohsu.edu

Introduction

Characterizing the identity and relative abundances of proteins in tissues is a logical first step in understanding normal biological and disease processes. Technological advances now allow many proteins to be studied at once rather than isolation, purification, and study of one protein at a time. These proteomic techniques have evolved from electrophoresis-based methods to large-scale mass spectrometry studies capable of cataloging thousands of proteins [1, 2].

The human lens fiber mass is an isolated, relatively simple biological sample that undergoes many age-related changes [3] and is ideally suited for proteomic studies [4]. Human lenses have been extensively studied using electrophoresis and a variety of mass spectrometry approaches [5–8] predominantly to characterize post-translational modifications [9–11] and their likely role in age-related nuclear cataract [12]. Two-dimensional gel electrophoresis (2-DE) maps of young normal human lens have been reported [13] and the major crystallins (highly abundant lens-specific proteins) characterized [14]. Many animal models of lens development and diseases are also used to understand human lens biology, and 2-DE maps have been reported for many species [13, 15–18].

There are several deficiencies in 2-DE studies and alternative proteomic strategies have been developed. Complex peptide mixture analysis is one of the most promising proteomic tools due to advances in instrumentation and bioinformatics. In these "bottom-up" (shotgun) experiments, proteins are enzymatically digested, usually with trypsin, into peptides that are sequenced using mass spectrometry. Multiple dimensions of chromatographic separations (2-DLC) are required to reduce peptide sample complexity and allow sequencing of large numbers of peptides. The separation and mass spectrometry steps can be automated, and the high sensitivity of this technique often results in the identification of several hundred to thousands of proteins per sample.

The advent of extensive genomic and proteomic sequence databases is a key factor in the analyses of the large datasets that shotgun studies produce. Peptide sequences are commonly determined by comparing theoretical fragmentation spectra of peptides present in a protein sequence database to the measured experimental spectra. There are several software packages that perform this database searching such as SEQUEST [19], Mascot [20], and others [21]. Many additional computational steps are necessary to correctly determine peptide sequences from tandem mass spectrometry data and to infer the proteins that gave rise to those peptides [22]. The majority of peptide sequencing events (MS/MS spectra) from most instruments do not produce correct peptide sequences due to many factors such as poor signal to noise, incomplete fragmentations, and non-peptide ions (chemical noise).

Controlling the number of incorrect peptide identifications becomes a major challenge in proteomics, particularly for large datasets. Search program scores can be combined with factors such as peptide mass accuracy and consistency with expected enzymatic cleavage into discriminant functions better able to separate incorrect from correct peptide identifications [23, 24]. Decoy databases, containing protein sequences that are known not to be present in the sample (reversed sequences, randomized sequences, or unrelated species), are now routinely used to estimate global false discovery rates (the fractions of accepted identifications that are incorrect) [25] and adjust thresholds to remove most incorrect identifications. The theoretical peptides generated from the protein database that are compared to measured spectra can be restricted to the subset consistent with enzymatic cleavage (usually trypsin) at both termini, known as a tryptic search. Alternatively, all possible peptide candidates can be considered (a no-enzyme search). Using no-enzyme searches of tryptic digests is another method to improve the accuracy of identifying tryptic peptides and control false discovery rates [26]. This method works because most of the incorrect peptides in no-enzyme searches match to peptides that are not consistent with trypsin cleavage at either terminus, making any peptides that are consistent with tryptic cleavage more likely to be correct. The only drawback of these approaches is that the searches require greater computational time due to the larger number of potential peptide sequences that must be searched.

Proteomic results are also strongly influenced by the completeness and quality of the protein databases used. The International Protein Index (IPI) species-specific databases [27] attempt to be very complete (containing most known and predicted gene products), whereas databases like UniProt Sprot [28] may be less complete but have little redundancy and more extensive protein annotations. More complete databases, like IPI, tend to be larger and require longer search times, which can become an issue with larger datasets. However, important proteins may be missing in UniProt databases and not be detected by search programs.

In this work, we present a straightforward proteomic analysis workflow (PAW) pipeline that used sequence-reversed decoy databases and a discriminant function transformation [24] of SEQUEST scores to maximize peptide identifications while controlling both peptide and protein false discovery rates (FDRs). The PAW pipeline was used to produce accurate whole lens proteomes of young human, mouse, bovine, and chicken lenses. Results obtained by searching larger IPI species-specific databases were compared to searches of databases constructed from the UniProt protein databases to see if more lens proteins could be identified. The more complete IPI databases resulted in the identification of novel crystallins γS in chicken, γN in bovine, and an alternate start codon isoform of βA4 in human.

Materials and methods

Samples and processing

All lens samples used in this study adhered to the ethical standards laid down in the 1964 Declaration of Helsinki for treatment of human subjects and the ARVO Statement for the use of animal subjects. Compliance with NIH guidelines was provided through an institutional review board. The isolation of the water-soluble fraction of whole lenses, protein assay, trypsin digestion, two-dimensional liquid chromatographic separation of peptides, and mass spectrometry were previously described in Wilmarth et al. [11] Single lenses from a 3-day-old human donor and fetal calf (0.65-g wet weight), a dozen 10-week-old chickens, and 234 C57/BL6 mice from 18–28 days of age, collected in the course of other studies were used. In brief, after removal from eye globes, the lens tissues were decapsulated and homogenized in aqueous buffer where soluble proteins were collected following centrifugation. Protein content was then assayed; 2.5 mg protein aliquots were dissolved in 8 M urea, reduced with DTT, alkylated with iodoacetamide, and digested with trypsin overnight. Peptides were separated by strong cation exchange (SCX) chromatography, fractions collected, and 10% of each fraction analyzed by LC/MS. The numbers of SCX fractions for each lens sample are given in Table 1. Reconstituted portions of each SCX fraction were analyzed using shallow 90-min gradient reverse-phase chromatography and peptide sequencing performed by an LCQ Classic ion trap mass spectrometer (ThermoFinnigan, San Jose, CA, USA).

Table 1 Individual lens dataset details listing the number of offline SCX fractions collected and the total number of MS/MS spectra acquired

Lens	SCX fractions	Number of DTAs
Human	32	54,385
Mouse	44	60,387
Bovine, run 1	37	34,953
Bovine, run 2	37	67,905
Chicken	38	58,595

Mass spectrometry and search details

The mass spectrometer was configured to acquire three centroided MS/MS spectra following each survey scan using dynamic exclusion. DTA files were created as previously described [11] and charge state analysis (ZSA, ThermoFinnigan) performed prior to SEQUEST searching (version 28, revision 12, ThermoFinnigan). The numbers of DTA files for each of the lens samples are given in Table 1. SEQUEST parameters were: parent ion tolerance of 2.5 Da, fragment ion tolerance of 1.0 Da, average parent ion masses, monoisotopic fragment ion masses, differential peptide N-terminal modification of +42 Da, static cysteine modification of +57 Da, maximum of two missed cleavages, and either trypsin or no enzyme cleavage specificity as described below.

Protein databases were either IPI species-specific databases [27] or were constructed from UniProt [28] entries. For human and mouse searches, canonical (single, representative sequence) Sprot protein databases were used. Chicken and bovine have less complete Sprot entries, so a combination of (canonical) Sprot and Trembl sequences were used. For all databases used in the searches, 179 common contaminant entries were added and sequence-reversed entries of original database plus contaminants were concatenated to produce databases having equal numbers of target (forward) and decoy (reversed) sequences. Database versions and numbers of protein entries are given in Table 2.

SEQUEST results processing

Python programs were written to convert SEQUEST OUT and DTA file formats to SQT and MS2 formats, respectively [29], to reduce file system overhead. Each SQT file was parsed to compute a discriminant function using the transformations and coefficients from Keller et al. [24]. A slightly modified definition of DeltaCN was used where the top score was compared to the average score of matches 4 to 10 rather than the second best match. The original concept behind DeltaCN is that the top scoring peptide might be correct and all lower scoring peptides are incorrect. However, there are situations where there can be more than one potential correct match (highly homologous peptides, post-translational modifications, permuted amino acids, etc.). This alternative definition was adopted so that the discriminant functions would not overly penalize these situations. The discriminant function coefficient for the DeltaCN term was optimized to maximize correct identifications from standard control mixes [30]. The increase in number of identifications compared to using the original DeltaCN was on the order of 1–2%. Discriminant scores were written to tab-delimited text files for each LC run.

Frequency histograms of the discriminant scores were tabulated and separated by charge states (1, 2, or 3) and by number of tryptic termini [NTT of 0 (non-tryptic), 1 (semi-tryptic), or 2 (fully-tryptic)]. Histograms were also separated into matches to the forward protein sequence entries or to the sequence-reversed entries for a total of nine paired histograms. The current implementation does not support charge states greater than 3+, but extension to higher charge states is possible. Any discriminant scores from peptides present in both forward and reversed sequences were randomly assigned to either database with equal likelihood so discriminant score distributions would be correctly normalized. Histograms were visualized using an Excel workbook template. A minimum peptide length cutoff of six amino acids was used to reduce short, ambiguous peptide identifications.

Peptide FDRs for each of the nine peptide classes were estimated from frequencies in the paired histograms of reversed and forward matches and tabulated as a function of discriminant score. Since we are not estimating peptide identification probabilities, complicated fitting of score distributions was not necessary. Discriminant score thresholds corresponding to desired peptide FDRs were obtained by inspecting tables in the Excel template. Smoothing of score distribution histograms was also used to estimate local error

Table 2 FASTA protein database details

Species	Database	Version	Forward sequences
Human	IPI	3.60	80412
Human	Sprot	57.4	20330
Mouse	IPI	3.60	56701
Mouse	Sprot	57.4	16140
Bovine	IPI	3.46	31501
Bovine	Sprot	57.4	5672
Bovine	Trembl	40.4	9749
Chicken	IPI	3.54	25697
Chicken	Sprot	57.4	2125
Chicken	Trembl	40.4	5849

Species names, database source, database version, and number of proteins sequences in the database are listed rates (to aid threshold selection) based on the relative height of the forward distribution to the sum of forward and reversed distribution heights, similar to how peptide probabilities are estimated in PeptideProphet [24]. After discriminant score thresholds were selected, new SQT, MS2, and discriminant text files were created containing only the top hits that passed the thresholds. To guard against uncertainties in the FDR estimates, any peptide class (typically fully non-tryptic peptides) that contributed less than 1% of total correct identifications for that charge state was excluded from further analysis. The reduction in dataset size after filtering out peptides that did not pass thresholds was typically 85% to 95%. Dataset size reduction is beneficial when processing data from faster scanning instruments or in multi-sample experiments.

Peptide to protein mapping

DTASelect v1.9 [31] was used to compile parsimonious protein lists from the filtered SQT files (cmd>DTASelect -1 0 -2 0 -3 0 –d 0 –o –t 0 –XML) for each 2-DLC sample. A separate Python program was written to parse the DTASelect-filter.xml file (or filtered discriminant text files), apply flexible protein identification criteria, and create tab-delimited text files of protein and peptide results. Minimum distinct (non-identical sequences) peptides per protein counts, minimum unique (peptide present in only one identified protein) peptide per protein counts, minimum number of NTT per distinct peptide, inclusion or exclusion of modified peptides, and inclusion or exclusion of multiple charge states per distinct peptide could be independently applied. Overall dataset normalization of spectral counts [32], and splitting of shared peptide counts on the basis of the unique peptide evidence of the proteins containing the shared peptides, was also incorporated. For example, if two proteins shared five MS/MS spectra with the first protein having nine unique MS/MS spectra and the latter having one unique MS/MS spectrum, the first protein would get 90% of the five shared counts and the second would get 10% of the shared counts. This is an overly simple approximation and there are many situations where it might fail; however, explicit reporting of total counts, unique counts, and these corrected counts along with lists of other proteins having shared peptides helps to identify such situations. A flowchart showing the analysis steps and software components is shown in Fig. 1.

Functional annotations

The Protein Information and Property Explorer (PIPE) [33] was used to add functional annotations to the protein results from the human and mouse lenses. Annotations were not available for bovine or chicken proteins via PIPE. Any missing Entrez gene numbers (necessary for GO term lookup) after a first pass with PIPE were manually entered before functional annotations were added in a second pass. Instead of multiple GO annotations per protein, the PIPE option to use only the most specific GO term was used. This reduced redundancy in the categorizations, but annotations for several crystallins were either missing or inconsistent between species. Therefore, a limited number of manual functional categories were chosen and proteins assigned to categories based on GO annotation information and UniProt database information. Functional summaries were compiled using protein spectral-count-weighted functional category frequencies.

Results

PAW pipeline performance

To benchmark the PAW pipeline performance, an analysis of the yeast dataset used in a recent report by Kall et al. [23] of a semi-supervised classifier program called Percolator was performed. Results from PeptideProphet and DTASelect analyses of the same dataset were also reported in the Percolator paper and could be compared to the PAW analysis. The dataset (35,236 MS/MS scans) and yeast database were downloaded from the article's supporting information and a concatenated forward/reverse database constructed. A no-enzyme SEQUEST search was performed with the parameters used in this work, and the results were processed with the PAW pipeline.

In the Percolator paper, the authors reported identification of 8,197 unique peptides from 12,691 MS/MS identifications at a peptide false discovery rate of 1%. Their classifier function contained 20 terms including three terms related to protein-level information. They reported 10,863 identifications (7,120 unique peptides) from PeptideProphet [24] analysis and 7,583 identifications from DTASelect [31] using default criteria. PAW pipeline numbers at a similar 1% peptide FDR were 7,958 unique peptides from 11,807 identifications. Percolator identified 7% more spectra than the PAW analysis and 3% more unique peptides. Without the three protein-level terms, the number of peptides identified by Percolator was reduced to 11,820, essentially the same as the PAW analysis. The PAW analysis outperformed PeptideProphet and DTASelect by 9% and 56%, respectively.

Inclusion of protein-level terms in the Percolator classifier could allow additional low-quality spectral matches to proteins identified by higher quality spectra and increase the number of spectrum matches (12,691 versus 11,820) with little increase in the number of identified proteins. Therefore, a comparison of PAW pipeline to Percolator results at the protein level was also done. All identifications having a

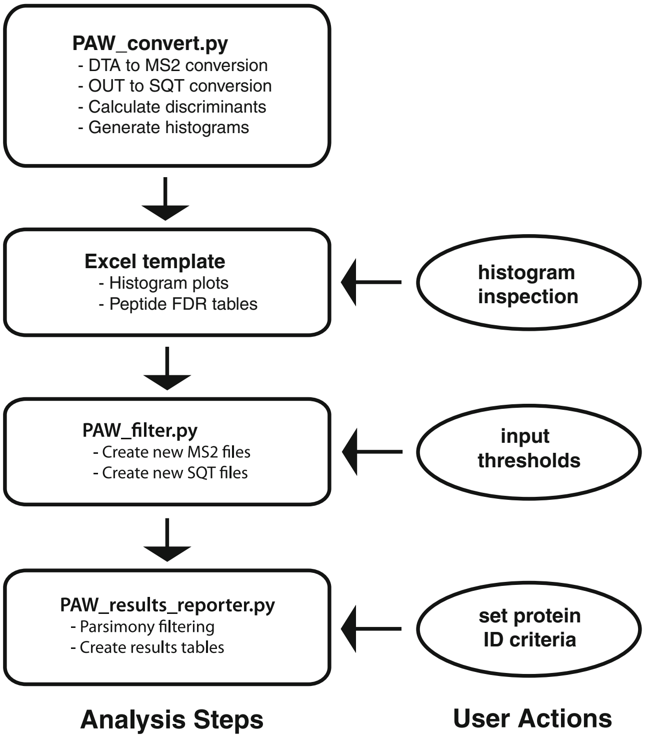


Fig. 1 Schematic of the major analysis steps and required user actions in the PAW pipeline. In each analysis box, python program names are in *bold type* and individual steps are listed. Simple user actions such as running programs, selecting folders, or opening files are not shown

q value less than 0.01 were filtered from the Percolator results, converted to SQT files, and identified proteins compiled using the PAW software. To accurately assess how differences at the peptide level affected the protein identifications, it was important to use the same peptide-to-protein mapping algorithms. Percolator produced 1,059 protein identifications, a modest gain of 2% over the 1,040 identifications from the PAW analysis. This suggests that most of the 884 additional peptide identifications from Percolator (compared to the PAW pipeline) were, in fact, associated with proteins identified by PAW analysis.

These results support the conclusion in a recent paper by Ding et al. [34] that training a classifier on the data being classified, rather than on separate training data, offers little improvement in many cases. The main drawback to supervised classifiers is that appropriate training data must be available for a given mass spectrometer platform. Semi-supervised classifiers are more flexible because they do not require fully classified training data, but they do require a sufficient number of correct identifications for statistical analysis and classification of a portion of the data (usually decoy matches). It is not uncommon in proteomics datasets for numbers of correct identifications to be too low for semi-supervised classifier algorithms to perform correctly. It is also computationally simpler to use a trained discriminant function rather than invoking a separate discriminant function training step. The improvement in results from PAW analysis compared to PeptideProphet were likely due to separation of score histograms by both charge state and NTT, a technique that has been recently incorporated into other analysis software [35].

Comparison between tryptic and no-enzyme searches

Trypsin is the most common enzyme used in proteomic studies because it cuts efficiently with good specificity. Thus, most peptides should be fully tryptic, and nearly all proteins present in samples can be identified on the basis of fully tryptic peptides. Tryptic searches are commonly used in proteomics and are roughly ten times faster than no-enzyme searches, but may miss important peptide identifications such as protein N-terminal peptides when the initial methionine has been removed. There can also be protein processing that produces active proteins from longer sequences, and N- or C-terminal peptides from the processed proteins may no longer match those predicted from the database entries. No-enzyme searches could, in principle, identify such peptides and provide a more complete proteomic picture. In addition, the overwhelming majority of incorrect peptide identifications in no-enzyme searches will be fully non-tryptic, and their exclusion could potentially increase the significance of fully tryptic peptide identifications. Do the longer execution times of no enzyme searches improve the results enough to justify their use?

To address this question, tryptic and no-enzyme searches were performed on the four lens samples using the smaller UniProt databases. Figure 2 shows the differences between the two searches in discriminant score distribution histograms

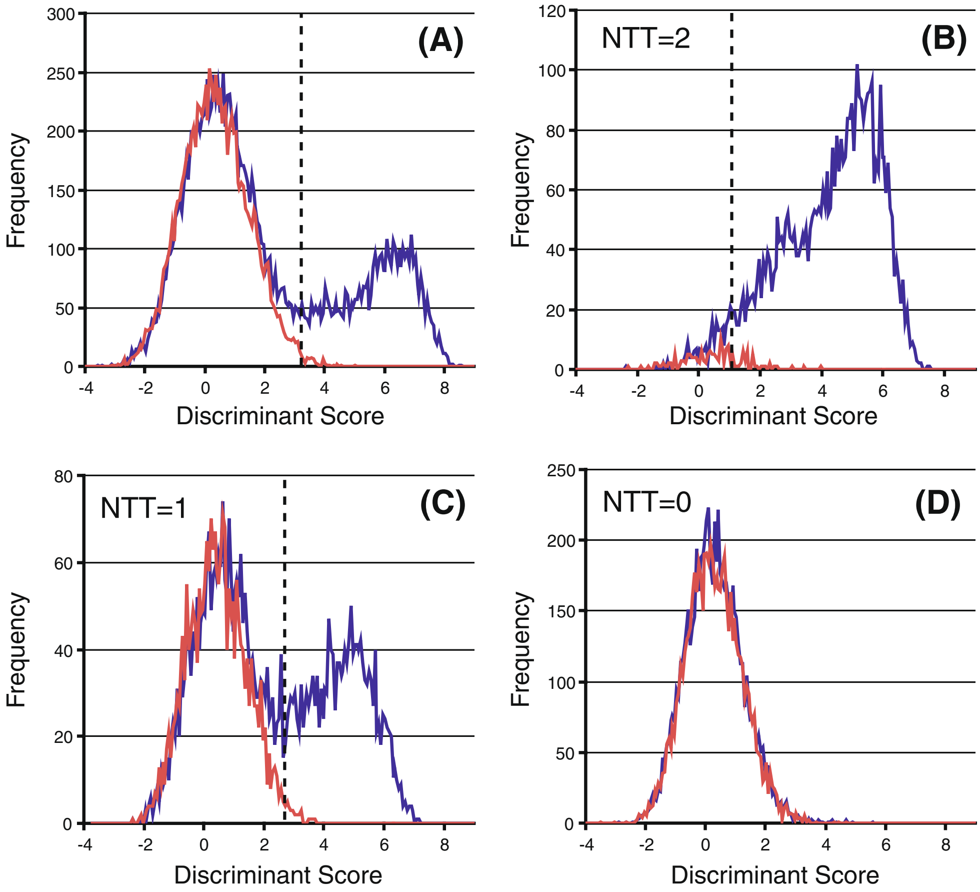


Fig. 2 Score distributions of human lens 2+ peptides from tryptic (**a**) and no-enzyme searches (**b**–**d**) against the human Sprot database with concatenated decoy sequences. Matches to forward database entries are shown in *blue*, and matches to sequence-reversed entries are shown in *red*. *Dotted lines* indicate the locations of the filtering thresholds for each peptide class. Fully non-tryptic peptides (NTT=0) were excluded

for 2+ peptides from human lens data. Essentially all of the fully non-tryptic matches in a no-enzyme search are incorrect, and the matches to the reversed sequences in a concatenated target/decoy database (Fig. 2d, red curve) accurately reproduced the score distribution for the incorrect matches to fully non-tryptic peptides in the forward sequences (Fig. 2d, blue curve). The total numbers of peptides identified in each search are shown in Table 3. In all cases, there were large numbers of semi-tryptic peptides identified with (usually) small decreases in the numbers of fully tryptic peptides. The majority (96–98%) of the semi-tryptic peptides originated from proteins identified on the basis of fully tryptic peptide evidence, and crystallins accounted for 80% to 90% of the semi-tryptic peptides. Thus, the vast majority of semi-tryptic peptides originated from abundant proteins present in the mixture. Due to dynamic exclusion and the relatively low peptide sample complexity in lens, the instrument may have frequently selected very low-intensity ions for fragmentation. Thus, the spectral count numbers of semi-tryptic peptides may have been inflated and not representative of their true abundance levels in the samples. There were no proteins that could be identified at the two-peptide-per-protein level solely on the basis of semi-tryptic peptides.

The numbers of putative correct fully non-tryptic peptides were negligible in these datasets. SEQUEST is capable of performing semi-tryptic searches, which would have been significantly faster than no-enzyme searches. We did not evaluate semi-tryptic searches, but it is very likely that the results would have been similar to the no-enzyme searches. The faster search times of semi-tryptic searches

Table 3 Comparison of identified peptide numbers between tryptic and no-enzyme searches

Lens	Tryptic search	No-enzyme search		Increase in IDs (%)
		Fully tryptic	Semi-tryptic	
Human	7,439 (171)	6,879 (89)	2,538 (85)	27
Mouse	8,770 (177)	8,320 (108)	2,787 (86)	27
Bovine	11,299 (198)	10,832 (135)	4,798 (156)	38
Chicken	4,500 (78)	4,510 (95)	2,903 (90)	65

Reversed peptide matches are shown in parentheses. The increase in total peptide identifications of the no-enzyme search relative to the tryptic search is given in the last column could be an important consideration with large datasets or when using large databases.

It is interesting to note that differences in the fully tryptic peptides identified in the two searches (number of identifications in Table 3 and dramatically different score distributions in Fig. 2a, b) resulted in differences at the protein level. At the individual DTA filename level (MS/MS spectra), the overlap of fully tryptic identifications between the two searches was 90%, with roughly twice as many DTA filenames unique to the tryptic searches as for the no-enzyme searches. The redundant number of MS/MS spectra identified as a distinct peptide sequence in a given charge state can be quite high in shotgun studies, so the overlap at the less redundant peptide level was also checked for fully tryptic identifications. The overlap was a similar 90%, but the numbers of peptides that were unique in each search were more similar to each other. The overlap at the protein level ranged from 84% to 94% when using two fully tryptic peptides per protein as a criterion in both searches. The results of the comparisons for the bovine lens data are shown in Fig. 3, and the other lenses had similar results. Several proteins were uniquely identified in each search and it is likely that all identifications were correct. The tendency of correct peptides to cluster to valid proteins (two peptides per protein) is a powerful classifier of correct protein identifications. This suggested that more complete proteome coverage would have been obtained by combining results from both tryptic and no-enzyme searches.

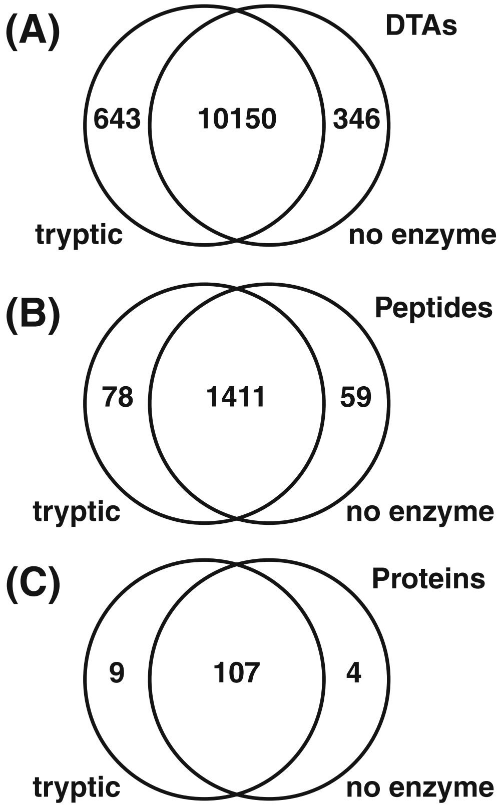


Fig. 3 Overlap between tryptic and no-enzyme SEQUEST search results for bovine lens proteins. **a** Comparison of fully tryptic peptide identifications at the MS/MS level. **b** Comparison of fully tryptic peptide sequences. **c** Comparison of protein identifications

Comparison of IPI and UniProt databases

At first glance, the more complete IPI databases might be expected to result in increased numbers of peptide identifications, but the opposite effect was observed. There were 4.4% more total MS/MS identifications from the UniProt searches than from the IPI searches. Larger databases reduce DeltaCN, on average, resulting in less separation between incorrect and correct distributions and a decrease in sensitivity. The IPI searches did result in increased protein identifications for all four species. The results are summarized in Table 4 where reversed matches are shown in parentheses. Peptide filtering thresholds were based on an estimated local error rate of 0.80 in each peptide class for all searches. Peptide FDRs varied slightly by search but were around 2%. The small number of incorrect peptides in each search resulted in very accurate protein identifications.

Reconciliations of protein identifications between the IPI and the UniProt results were tried. Cross-referencing information in database entries was used to match protein identifications between the two sets of results. This process is challenging because sets of peptides often match to more than one protein, especially in larger databases like IPI. If any member of a redundant protein family in one search matched any member of a redundant protein family member in the other search, the two families were considered as matching. In general, nearly all UniProt proteins matched to entries in IPI databases. This is not surprising since the more complete IPI databases include known sequence information from UniProt. There were identifications that were unique to the larger IPI databases, and some of those are mentioned in the next sections where results from the four species are presented.

Table 4 Comparison between results from no-enzyme SEQUEST searches of IPI or UniProt protein databases at the peptide (MS/MS) level and at the protein level

Species	IPI MS/MS	UniProt MS/MS	IPI Protein	UniProt protein
Human	9,347 (185)	9,414 (174)	115 (1)	113 (1)
Mouse	11,043 (187)	11,107 (194)	164 (0)	156 (0)
Bovine	14,517 (238)	15,630 (291)	123 (0)	116 (0)
Chicken	6,739 (166)	7,413 (185)	69 (0)	62 (0)

Reversed matches are shown in parentheses

Human lens results

The proteins and peptides identified in the IPI searches are listed in Electronic supplementary material (ESM), Supplemental Table 1, and the results from the Sprot search are listed in ESM Supplemental Table 2. The number of proteins identified, about 115, is less than another recent report [36], but depth of proteome coverage is very dependent on the degree of peptide separation and sensitivity of the mass spectrometer. Preliminary analysis of additional young human lenses using a linear ion trap and the PAW pipeline has identified more than 800 proteins (manuscript in preparation). A 2-DE study of the 3-day-old human lens has recently been reported where nearly all visible proteins were identified as crystallins [13], with only one non-crystallin protein detected. This 2-DE picture is in contrast to results from 2-DLC analysis where a much more complex proteome emerges. Functional annotations were added to the identified proteins using PIPE [33], proteins assigned to manual categories based on annotation information, and spectral count weighted functional frequencies computed. The annotated protein list is detailed in ESM, Supplemental Table 3, and a frequency analysis is shown in Fig. 4 where crystallins accounted for 83% of the detected peptides.

The human Sprot database is considered a complete proteome, and we chose only the canonical sequences for this analysis. An additional 14,000 protein isoform sequences are available but were not used. The reconciliation of identified proteins from the two databases indicated that 107 of 109 proteins (after removal of contaminants) identified in the Sprot search were present in the IPI results. Conversely, 107 of 110 IPI proteins could be mapped to Sprot identifications. A new result from one of the unique IPI matches was strong evidence that βA4 has an alternate start codon, resulting in a protein having an additional 12 N-terminal amino acids. Peptide spectral counts suggest that this longer isoform is translated at lower levels than the shorter form of βA4.

Mouse lens results

ESM, Supplemental Tables 4 and 5 list the proteins and peptides identified in the searches against the mouse IPI and Sprot databases, respectively. There were 153 non-redundant proteins (excluding contaminants) identified in the Sprot search, and 150 of those proteins had matching IPI identifications. For the 161 proteins found in the IPI searches, 150 could be mapped to corresponding Sprot entries, with 11 proteins unique to the IPI results. One of those identifications was well-known βA3, the alternate start codon form of βA1, which was not present in the Sprot database. It was not annotated as an isoform of βA1 nor was βA3 present in Trembl. There were no problems with βA1/A3 UniProt sequences for the other three species. Functional annotation analysis was also possible for the mouse lens and is given in ESM, Supplemental Table 6. A frequency analysis similar to that described for human lens above is shown in Fig. 5. As in the human lens, crystallins accounted for 80% of the observed peptides with a relatively even distribution of counts across other categories. Previous 2-DE studies of mouse lens [17] found essentially only crystallins in soluble lens proteins, suggesting that the abundance of most non-crystallins is too low to detect with 2-DE.

Fig. 4 Spectral-count-weighted functional categories for the 3-day-old human lens

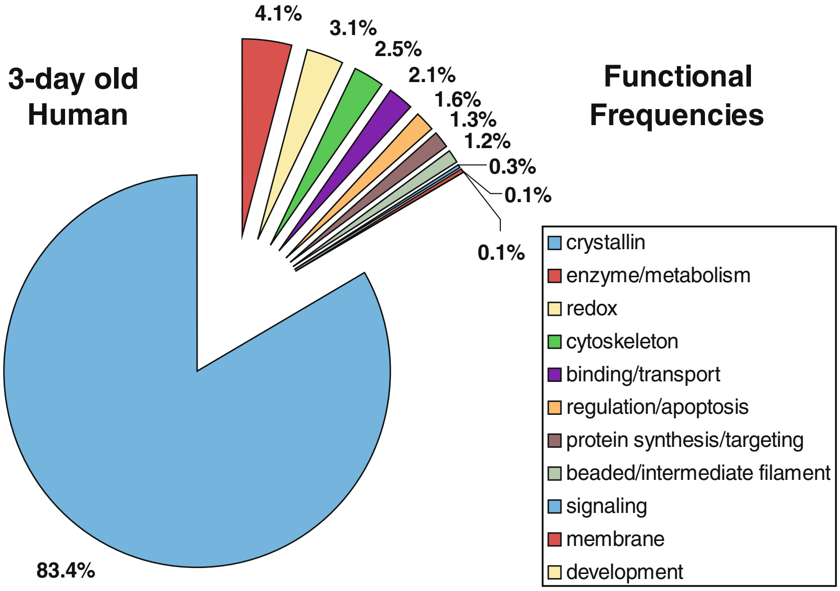
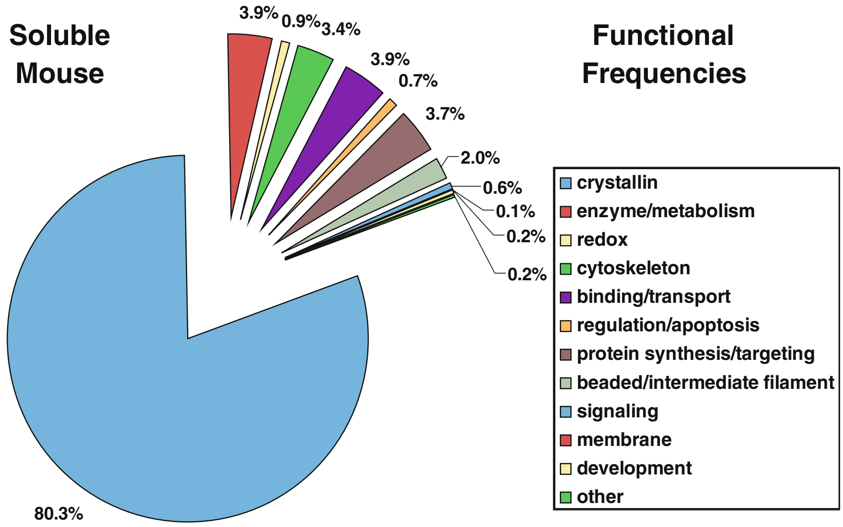


Fig. 5 Spectral-count-weighted functional categories for the soluble mouse lens proteins



Bovine lens results

The results of the IPI and UniProt searches are detailed in ESM, Supplemental Tables 7 and 8, respectively. There were two 2-DLC experiments using bovine lenses. The first experiment had fewer peptide identifications than expected, and chromatogram peak intensities were also lower than anticipated. Therefore, a second experiment was performed. Results from both experiments were easily combined with the flexible workflows possible in the PAW pipeline. Excluding contaminants, 110 of 111 UniProt identifications could be mapped to IPI identifications. There were nine identifications that were unique among the 119 (non-redundant groups without contaminants) IPI protein identifications. Two of the unique IPI identifications were notable. One was a putative identification of γA crystallin in bovine. The γA sequence present in UniProt (P02527) appears to be incorrect and had the highest similarity to γB crystallins in other species. The other important identification was γN crystallin, a minor crystallin recently identified in other species such as mouse and guinea pig [37, 38], but not previously known in the cow.

Chicken lens results

ESM, Supplemental Table 9 contains the chicken lens IPI results and ESM, Supplemental Table 10 has the UniProt results. Young chicken lenses have very high expression of the taxon-specific delta crystallins. Delta crystallin had spectral count numbers roughly twice those of other crystallins. The overabundance of delta crystallin effectively lowers the concentration of all other proteins, and fewer proteins were identified in chicken (60–70) compared to the other species (110–160) under similar experimental conditions. The results from the two databases also showed more differences than the other species, suggesting that protein sequences and annotations may be more variable for chicken. Of the 60 proteins identified in the UniProt analysis, 52 mapped to IPI identifications and eight did not. There were 67 IPI identifications, with 52 mapping to UniProt results and 15 that did not. One of the protein identifications unique to the IPI database had sequence coverage of 77% and total spectral count of 121 and is the first observation of chicken γS crystallin. The total spectral counts of the other major crystallins were about five times larger than γS. There was also evidence for the presence of γN crystallin at low abundance levels. Monomeric γ-crystallins have not been previously reported in avian lenses. A published 2-DE study [18] again detected only major crystallin proteins in chicken lenses. The identification of many proteins besides crystallins and the first observation of chicken γS highlight the increased sensitivity of 2-DLC compared to 2-DE.

Discussion

Controlling peptide and protein error rates

Controlling peptide false discovery rates has received considerable attention in proteomics resulting in many of the techniques used in this work, such as discriminant function transformations [24], target/decoy databases [25], and separation of peptides by NTT [35]. The related problem of protein false discovery rate in large-scale experiments [39] has received less attention. Typically, the numbers of proteins detectible in a sample are small compared to the number of possibilities in the protein database being searched. A reasonable assumption is that random, incorrect peptide matches will be uniformly distributed across the protein database entries, in contrast to the tendency of correct peptide matches to cluster to a small number of proteins actually present in the sample. This tendency had been long recognized and the criterion of two peptides per protein is routinely used to increase accuracy of protein identifications.

A recent report [39] has shown that the number of incorrect protein identifications is a function of the number of incorrect peptide matches and the size of the protein database. Accurate protein identification requires that the absolute number of incorrect peptide matches be small compared to database size and that the number of possible proteins be large compared to the number of proteins in the sample. When these two requirements are met, the two-peptides-per-protein criterion can effectively control protein false discovery rates. Some of the information necessary to meet these requirements is not known until analysis has been completed, creating a Catch-22, but general strategies can be used so that analyses do not have to be routinely repeated. Construction or selection of a fairly complete protein database is often possible, and addition of sequence-reversed entries ensures that the number of protein candidates is usually large compared to the number of proteins detectible in the sample. Discriminant function transformations, target/decoy databases, and separation of peptides by NTT in conjunction with no-enzyme specificity searches are effective at controlling the number of incorrect peptide matches. Heuristics used in this work, such as including only unmodified peptides or fully tryptic peptides for protein identification, effectively reduces the number of incorrect peptides even further.

The peptide false discovery rate that will produce the desired final protein false discovery rate is a function of two main factors: the size of the dataset(s) and the number of datasets included in the analysis. When datasets are very large (from fast-scanning instruments or extensive peptide fractionation), the peptide false discovery rate must be greatly reduced to keep the numbers, not rates, of incorrect peptide matches small. Protein identifications in multi-sample experiments increase asymptotically with increasing sample number [32]. However, incorrect protein matches are unlikely to be the same from sample to sample, so the incorrect matches tend to increase linearly with increasing sample number. If an analysis contains many samples, the protein false discovery rate per sample must be correspondingly reduced so that the final protein false discovery rate is controlled.

Comparison of protein databases

The larger, more complete IPI databases resulted in the discovery of interesting new crystallins in human, bovine, and chicken lenses. They also had many more redundant proteins on average than the UniProt databases, had annotations and web-based database query results that were less informative than UniProt results, and took considerably more time for searches to complete. The more redundant nature of the IPI databases requires very careful control of incorrect peptide identifications or additional assumptions when mapping peptides to proteins [35]. Otherwise, some protein redundancies may be lost due to incorrect “unique” peptides and the protein results artificially inflated.

We chose to present separate lens proteomes for IPI databases and for UniProt databases rather than some form of a combined proteome. Reconciling identifications between the two databases was not trivial and was not deemed accurate enough to produce a combined proteome. The best analysis strategy might be to search IPI databases to gain a more complete picture of the more abundant proteins and then use those results to augment sequences in UniProt databases for final searches. This would ensure that UniProt databases are “complete” for the sample under study, and the higher quality annotations would be easier for follow-up research.

Comparison between tryptic and no-enzyme searches

For these modest-sized datasets, there was less difference between the two search strategies than expected. There was considerable overlap in the fully tryptic peptides that passed thresholds in both searches, but there were also unique fully tryptic peptides that were identified by each search strategy. The small difference at the peptide level also translated into small differences at the protein level, making it hard to conclude which search strategy was better. It is clear from Fig. 2 that score threshold must be set higher for fully tryptic peptides in tryptic searches (Fig. 2a) than for fully tryptic peptides in no-enzyme searches (Fig. 2b). There was a potential gain in no-enzyme searches by reducing the number of incorrect fully tryptic peptides, which allowed lower relative thresholds and retention of a larger fraction of the correct fully tryptic peptides. It has been shown that some fully tryptic peptide identifications get displaced from their top-scoring position in tryptic searches by semi- or non-tryptic results in no-enzyme searches [34], and this will reduce the gain from lowering the thresholds. For these datasets, the two factors seemed to offset each other, and no-enzyme searches did not outperform tryptic searches. If the incorrect distribution is much larger than the correct distribution for fully tryptic peptides from a tryptic search, then there may be more gain from performing a no-enzyme search. In this case, the right-hand tail of the large incorrect distribution may reduce tryptic search sensitivity to a greater extent, and shifting the majority of the incorrect peptides to other peptide classes in no-enzyme searches may be more beneficial.

Lens proteomes

Nearly all 2-DE studies of lens proteins have identified only crystallins, in contrast to these 2-DLC experiments where far more complex proteomes were obtained. Functional annotations could be added for the proteins detected in the human and mouse lenses, and spectral-count-weighted functional frequency analyses were performed. The majority (about 80%) of the observed peptides were associated with crystallins, with other peptides roughly equally scattered across several categories. It is not clear from this study if any of these proteins are actually functional in mature lens fiber cells. It seems more likely that many of these proteins are remnants of the organelle degradation that occurs following fiber cell differentiation. It may be interesting to see if the non-crystallin lens proteome changes with age to assess possible contributions to loss of lens function.

The novel crystallin results added to our understanding of the changes in lens crystallins across species. Alternative start codons are known for many β-crystallins, and the longer isoform of βA4 in human is another example. The relative abundance of the longer form is less than the shorter form, suggesting that the alternate start codon may be very close to the 5′ end similar to the βA1/A3 isoforms. Monomeric γ-crystallins in chicken were predicted from genomic sequences, but this is the first evidence that these genes are expressed. Confirmation of the presence of γS and γN cDNA in chicken lens is planned. Relative abundance levels of γS based on spectral counts indicate lower levels (about 20%) relative to the β-crystallins. The minor crystallin γN has only recently been detected in a few species, and we were able to observe this crystallin in both bovine and chicken lenses for the first time. It had been previously reported in mouse [37] and we were able to confirm that finding. We were not able to detect the presence of γN in human lens.

We have applied powerful new proteomic techniques in combination with rigorous bioinformatics analyses to study the lens from a systems biology point of view. The accurate, more complete lens proteomes presented here resulted in the identification of new crystallins in three species and can serve as starting points for future lens proteomic studies. Additional species can be studied, deeper proteome coverage obtained with newer, more-sensitive instruments, and quantitative proteomic studies performed to address specific lens biological questions.

Supporting information

Full protein and peptide identifications for the human lens are presented in ESM, Supplemental Tables 1, 2, and 3. The mouse results are in ESM, Supplemental Tables 4, 5, and 6. Bovine and chicken proteomes are given in ESM, Supplemental Tables 7, 8, 9 and 10, respectively. The PAW software, for processing SEQUEST search results, is freely available for non-commercial use via written request to the corresponding author.

Acknowledgments This work was supported by the National Institutes of Health grants EY007755 and EY10572. Lenses from mice and chickens were kindly provided by Drs. Steve Bassnett and Thomas Shearer.

References

1. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422:198–207.
2. Han X, Aslanian A, Yates JR 3rd. Mass spectrometry for proteomics. Curr Opin Chem Biol. 2008;12:483–90.
3. Bloemendal H, de Jong W, Jaenicke R, Lubsen NH, Slingsby C, et al. Ageing and vision: structure, stability and function of lens crystallins. Prog Biophys Mol Biol. 2004;86:407–85.
4. Hoehenwarter W, Klose J, Jungblut PR. Eye lens proteomics. Amino Acids. 2006;30:369–89.
5. Lampi KJ, Ma Z, Hanson SR, Azuma M, Shih M, et al. Age-related changes in human lens crystallins identified by two-dimensional electrophoresis and mass spectrometry. Exp Eye Res. 1998;67:31–43.
6. Lapko VN, Purkiss AG, Smith DL, Smith JB. Deamidation in human gamma S-crystallin from cataractous lenses is influenced by surface exposure. Biochemistry. 2002;41:8638–48.
7. Robinson NE, Zabrouskov V, Zhang J, Lampi KJ, Robinson AB. Measurement of deamidation of intact proteins by isotopic envelope and mass defect with ion cyclotron resonance Fourier transform mass spectrometry. Rapid Commun Mass Spectrom. 2006;20:3535–41.
8. Zhang Z, Smith DL, Smith JB. Human beta-crystallins modified by backbone cleavage, deamidation and oxidation are prone to associate. Exp Eye Res. 2003;77:259–72.
9. Dasari S, Wilmarth PA, Rustvold DL, Riviere MA, Nagalla SR, et al. Reliable detection of deamidated peptides from lens crystallin proteins using changes in reversed-phase elution times and parent ion masses. J Proteome Res. 2007;6:3819–26.
10. Hains PG, Truscott RJ. Post-translational modifications in the nuclear region of young, aged, and cataract human lenses. J Proteome Res. 2007;6:3935–43.
11. Wilmarth PA, Tanner S, Dasari S, Nagalla SR, Riviere MA, et al. Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to crystallin insolubility? J Proteome Res. 2006;5:2554–66.
12. Truscott RJ. Age-related nuclear cataract-oxidation is the key. Exp Eye Res. 2005;80:709–25.
13. Nakajima E, David LL, Rivere MA, Azuma M, Shearer TR. Human and non-human primate lenses cultured with calcium ionophore form alphaB-crystallin lacking the C-terminal lysine, a prominent feature of some human cataracts. Invest Ophthalmol Vis Sci 2009 (in press).

14. Lampi KJ, Ma Z, Shih M, Shearer TR, Smith JB, et al. Sequence analysis of betaA3, betaB3, and betaA4 crystallins completes the identification of the major proteins in young human lens. J Biol Chem. 1997;272:2268–75.
15. Lampi KJ, Shih M, Ueda Y, Shearer TR, David LL. Lens proteomics: analysis of rat crystallin sequences and two-dimensional electrophoresis map. Invest Ophthalmol Vis Sci. 2002;43:216–24.
16. Robertson LJ, David LL, Riviere MA, Wilmarth PA, Muir MS, et al. Susceptibility of ovine lens crystallins to proteolytic cleavage during formation of hereditary cataract. Invest Ophthalmol Vis Sci. 2008;49:1016–22.
17. Ueda Y, Duncan MK, David LL. Lens proteomics: the accumulation of crystallin modifications in the mouse lens with age. Invest Ophthalmol Vis Sci. 2002;43:205–15.
18. Wilmarth PA, Taube JR, Riviere MA, Duncan MK, David LL. Proteomic and sequence analysis of chicken lens crystallins reveals alternate splicing and translational forms of beta B2 and beta A2 crystallins. Invest Ophthalmol Vis Sci. 2004;45:2705–15.
19. Eng JK, McCormack AL, Yates JR III. An approach to correlate tandem mass sectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom. 1994;5:976–89.
20. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis. 1999;20:3551–67.
21. Kapp EA, Schutz F, Connolly LM, Chakel JA, Meza JE, et al. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics. 2005;5:3475–90.
22. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics. 2005;4:1419–40.
23. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods. 2007;4:923–5.
24. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002;74:5383–92.
25. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007;4:207–14.
26. Xie H, Griffin TJ. Trade-off between high sensitivity and increased potential for false positive peptide sequence matches using a two-dimensional linear ion trap for tandem mass spectrometry-based proteomics. J Proteome Res. 2006;5:1003–9.
27. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, et al. The International Protein Index: an integrated database for proteomics experiments. Proteomics. 2004;4:1985–8.
28. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, et al. Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics. 2009;10:136.
29. McDonald WH, Tabb DL, Sadygov RG, MacCoss MJ, Venable J, et al. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. Rapid Commun Mass Spectrom. 2004;18:2162–8.
30. Klimek J, Eddes JS, Hohmann L, Jackson J, Peterson A, et al. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. J Proteome Res. 2008;7:96–103.
31. Tabb DL, McDonald WH, Yates JR 3rd. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J Proteome Res. 2002;1:21–6.
32. Liu H, Sadygov RG, Yates JR 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem. 2004;76:4193–201.
33. Ramos H, Shannon P, Aebersold R. The protein information and property explorer: an easy-to-use, rich-client web application for the management and functional analysis of proteomic data. Bioinformatics. 2008;24:2110–11.
34. Ding Y, Choi H, Nesvizhskii AI. Adaptive discriminant function analysis and reranking of MS/MS database search results for improved peptide identification in shotgun proteomics. J Proteome Res. 2008;7:4878–89.
35. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, et al. IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. J Proteome Res. 2009;8:3872–81.
36. Hains PG, Truscott RJ. Proteome analysis of human foetal, aged, and advanced nuclear cataract lenses. Proteomics Clin Appl. 2008;2:1611–9.
37. Wistow G, Wyatt K, David L, Gao C, Bateman O, et al. gammaN-crystallin and the evolution of the betagamma-crystallin superfamily in vertebrates. FEBS J. 2005;272:2276–91.
38. Simpanya MF, Wistow G, Gao J, David LL, Giblin FJ, et al. Expressed sequence tag analysis of guinea pig (*Cavia porcellus*) eye tissues for NEIBank. Mol Vis. 2008;14:2413–27.
39. Reiter L, Claassen M, Schrimpf SP, Jovanovic M, Schmidt A, et al. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Mol Cell Proteomics, 2009;8:2405–2417. doi:10.1074/mcp.M900317-MCP200