Clinical Proteomics

ISSN 1542-6416/06/02:185–204/$30.00 (Online)

Original Article

Comparison of Protein Expression Lists From Mass Spectrometry of Human Blood Fluids Using Exact Peptide Sequences Versus BLAST

Peihong Zhu,[1] Peter Bowden,[1] Voitek Pendrak,[1] Herbert Thiele,[2] Du Zhang,[3] Michael Siu,[4] Eleftherios P. Diamandis,[4] and John Marshall[1,4,]*

[1]Department of Chemistry and Biology, Ryerson University, Toronto, Canada; [2]Bruker Daltonics, Bremen, Federal Republic of Germany; [3]Department of Computer Science, California State University, Sacremento; [4]Ontario Cancer Biomarker Network, MaRS, 101 College Street, Toronto, Canada

Abstract

The proteins in blood were all first expressed as mRNAs from genes within cells. There are databases of human proteins that are known to be expressed as mRNA in human cells and tissues. Proteins identified from human blood by the correlation of mass spectra that fail to match human mRNA expression products may not be correct. We compared the proteins identified in human blood by mass spectrometry by 10 different groups by correlation to human and nonhuman nucleic acid sequences. We determined whether the peptides or proteins identified by the different groups mapped to the human known proteins of the Reference Sequence (RefSeq) database. We used Structured Query Language data base searches of the peptide sequences correlated to tandem mass spectrometry spectra and basic local alignment search tool analysis of the identified full length proteins to control for correlation to the wrong peptide sequence or the existence of the same or very similar peptide sequence shared by more than one protein. Mass spectra were correlated against large protein data bases that contain many sequences that may not be expressed in human beings yet the search returned a very high percentage of peptides or proteins that are known to be found in humans. Only about 5% of proteins mapped to hypothetical sequences, which is in agreement with the reported false-positive rate of searching algorithms conditions. The results were highly enriched in secreted and soluble

*Author to whom all correspondence and reprint requests should be addressed:
John Marshall, Department of Chemistry and Biology, Ryerson University, Toronto, Canada.
E-mail: 4marshal@ryerson.ca.

proteins and diminished in insoluble or membrane proteins. Most of the proteins identified were relatively short and showed a similar size distribution compared to the RefSeq database. At least three groups agree on a nonredundant set of 1671 types of proteins and a nonredundant set of 3151 proteins were identified by at least three peptides.

Key Words: Expression; analysis; proteomics; protein; composition; location; function; process; distribution; chi-square statistic; human; serum; LC/LC-MS/MS; computation.

Introduction

High-throughput tandem mass spectrometry (MS/MS) based peptide correlation analysis of complex biological samples yields long lists of proteins ***(1)***. Serum may contain most human proteins ***(2)*** but most are not detectable by chromatography followed by polyacrylamide gel electrophoresis (PAGE) ***(3)***. Liquid chromatography (LC)-based approaches have been shown to detect more proteins than PAGE-based approaches ***(4,5)***. The peptides reported from superabundant proteins have often proved to be purely tryptic peptides with few missed cleavage sites. However, some of the peptides from apparently low abundance proteins have reportedly resulted from nontryptic activities or contained missed cleavage sites. Trypsin cleaves exclusively on the C terminal side of lysine or arginine ***(6)*** and this may lead to the perception that some of the nontryptic peptides reported in sera may be artifacts of searching ***(7)***. The groups have used different sample preparation methods including LC-PAGE ***(3)***, LC/LC-MS/MS ***(5,8–10)***, iso-electric focusing ***(11)***, no sample preparation followed by ultra high-performance liquid chromatography (HPLC) ***(9)***, with fractionation followed by ultra HPLC ***(12)***, and after low molecular mass filtration ***(13)***. We obtained the published protein expression lists generated by 10 research groups from the MS analysis of human blood. The protein expression lists were obtained by searches of MS/MS spectra against both human and nonhuman sequences and against known proteins (NP) vs hypothetical gene products whose expression is unknown (XP). Most proteins have been identified by MS/MS with collision-induced dissociation of tryptic disgests from blood proteins. These fragmentation spectra were correlated to as many proteins as possible from a comprehensive protein databases ***(8)***. A number of correlation algorithms have been developed and tested empirically to determine the scoring parameters that result in acceptable false-positive rates of about 5% ***(14,15–19)*** based on searches of both real and nonphysiological protein databases ***(6,8,17)***, but this alone may not be sufficient to ensure correct identification ***(7)***.

The proteins in blood were all first expressed as mRNAs from genes within cells. There are data banks of known human expression products such as the Reference Sequence (RefSeq) NP database **(Fig. 1)**. Proteins identified from human blood that fail to match expressed human mRNAs may be suspect. We compared the protein identified in human blood by mass spectrometry (MS) from 10 different groups. We determined whether the peptides or proteins identified by the different groups are NP that have been shown to be expressed as mRNA in human beings. There are two ways that correlation of mass spectrometry data to nucleic acid sequences might be in error: the spectra may be correlated to the wrong peptide sequence or the same or very similar peptide sequence might be shared by more than one protein. To control for these two possibilities we used Structured Query Language (SQL) database searches of the peptide sequences and basic local alignment search tool (BLAST) analysis of the identified full

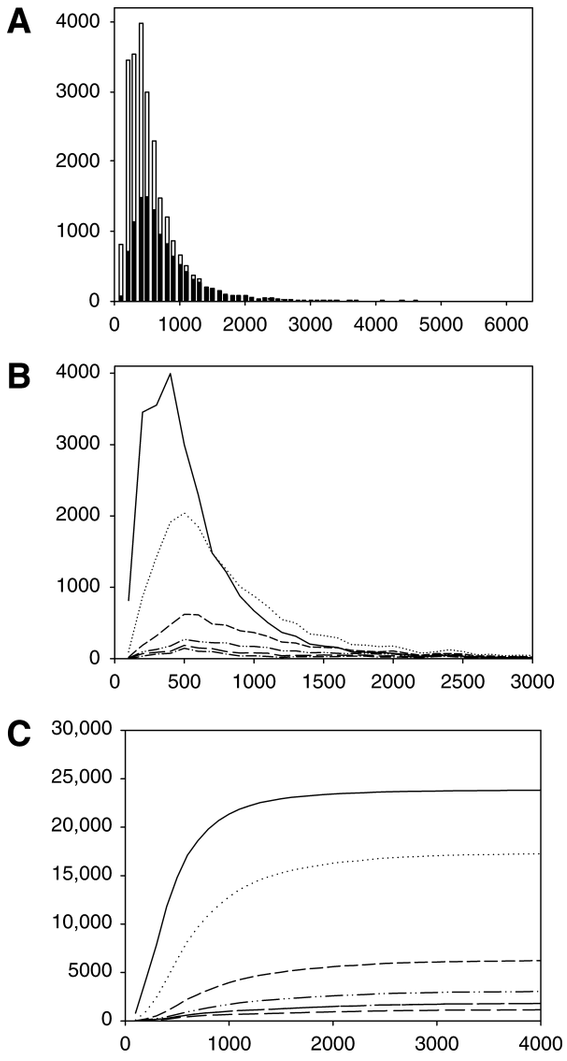

Fig. 1. The size distribution of blood proteins identified by mass spectrometry compared to the RefSeq database. **(A)** Histogram showing the size distribution of the RefSeq database (open bars) vs the nonredundant MS data (closed bars); **(B)** plots of the size distribution of proteins of 3000 amino acids or less by for RefSeq (top solid line) MS vs redundant proteins identified by one to five peptides shown in descending order; **(C)** plots of the running total of size distribution of the MS data from RefSeq (top solid line) vs redundant proteins identified by at least one to five peptides shown in descending order. The axi are number of proteins and abscisa are protein length in amino acids.

length proteins. Different algorithms have been used to combine and compare data sets and to determine high confidence identifications from correlations against known human, nonhuman, or hypothetical protein sequences ***(20)***. We assembled nonredundant lists of normal serum proteins with homology to known human transcripts ***(21)*** or the predicted transcription sites of the human genome ***(22)***.

Materials and Methods

Databases

Publicly available databases of blood proteins from MS analysis that are markedly different have been assembled by the Johns Hopkins/Bio-Informatic Institute and HUPO ***(20,23)***. We obtained the published serum and plasma proteomic data and parsed the accession numbers to obtain FASTA protein sequences ***(5,8,9,11–13,23)***. Where available, the peptide sequences that were correlated to MS/MS fragmentation spectra were parsed. The protein and peptide sequences were used directly for comparisons or transformed to a common RefSeq format by BLAST matching or exact string searches for MS/MS peptide sequences. RefSeq is a comprehensive, nonredundant database of human transcripts together with their genomic loci and gene ontology (GO) annotations ***(24)*** that was used as the basis of comparison between groups and category distributions. Except where indicated, the July 2004 Human RefSeq database was used for these calculations and contains a total of 29,234 entries with 23,888 NP and 5333 XP entries. RefSeq is nonredundant in the sense that few or none of the FASTA sequences are exactly the same but many have very similar sequences, or share subsets of amino acid sequences or are of different lengths or types. We previously obtained a serum proteome via the calculation of many LC-MS/MS runs separately using discrete, individual sequest searches for pure tryptic

peptides with X-Corrs or 1.5, 2.5, or 3.75 for 1+, 2+, and 3+ ion, respectively to yield a list of some 600 proteins after manual editing. We have recalculated these many LC-MS/MS runs together using the SEQUEST software ***(15–17)*** and the conditions previously described ***(5)***. The database was downloaded from NR databse (NCBI) in April 2003 using any FASTA headers that contained human, homo ***(25)***, or sapiens but excluding headers containing virus, viral, or HIV. This resulted in a list of 2681 proteins identified by their NCBI database GI numbers that were presented at the HUPO meeting in 2003. We compared the results of searching a set of 110 LC-MS/MS runs either as individual discrete files or on joint computation of multiple LC-MS/MS runs using the May 2003 database version of RefSeq that contains 18,455 NPs and 18,887 XPs. We found that about 80% of the hits to RefSeq were NP entries. Based on the two-by-two contingency table of NP and XP frequencies, Fisher's exact test indicated that there was a probability value of 3.82E-169 that the multiple run list was randomly sampled from the RefSeq database. The protein list generated from multiple LC-MS/MS calculations was much larger than the list generated from the joint calculation of multiple LC-MS/MS and the rate of XP entries increased from 22 to 25% when runs were calculated together.

Matching to Human RefSeq Proteins

We used two methods, one based on the small peptides identified by MS/MS fragmentation and one based on the full length sequence of the proteins implicated to assemble representative lists of nonredundant proteins for the purpose of comparing across groups.

Exact MS/MS Peptide String Searches

We used the exact peptide sequence correlated in the MS/MS experiment to search for that string of amino acids. The set of peptides discovered by each group were collapsed into the set of longest unique peptide sequences. These sequences were used as search strings against the FASTA files of RefSeq. The amino acid sequences of the resulting RefSeq FASTA files were in turn collapsed into the set of longest representative FASTA sequences that still contained the original FASTA sequence and, therefore, still contained the MS/MS peptides. All these data base steps were performed with a SQL database on a Windows operating system personal computer.

BLASTp Matching

These serum protein lists were compared to each other, as well as to the RefSeq database, using the publicly available BLAST ***(26)*** search engine downloaded from NCBI. We used 75% sequence identity over the full length (FL) of the entire identified protein as a criterion to collapse many proteins into one top scoring entry. The top scoring BLASTp alignment for every query protein was considered a "match" if the FL identity was greater than 75%, and if the alignment contained a perfect match string of at least 20 amino acids ***(26)***.

Comparison Between Experimental Groups

We assembled the complete FASTA sequences from all groups and assembled a nonredundant database using SQL. We then used the exact MS/MS peptides string searches and the BLASTp matching methods against the nonredundant database to determine agreement across groups.

Protein Localization, Gene Ontology Terms, Biological Process, or Molecular Function

The proteins were categorized into intracellular and extracellular compartments by GO terms to classify protein localization in broad categories such as cytoplasmic, membrane,

mitochondrial, nuclear, secreted, and unclassified. To associate the biological processes and molecular functions from the GO ***(27)***, the GO database was downloaded from the Gene Ontology Consortium website and installed as a MySQL database on a Windows 2000 workstation. Following the methods of Camon et al. ***(28)***, the generic GOSlim file (goslim_generic.obo), gene ontology file (gene_ontology.obo), and the Perl utility script (map2slim.pl) were downloaded from the same website. Map2slim.pl script was modified to accommodate our input and output file format. The protein GI numbers identified by correlation analysis were mapped to RefSeq proteins and the corresponding RefSeq ids and their associated GO identifiers were obtained from the gene2refseq, gene2go file downloaded from NCBI. This file was run through the modified map2slim.plscript, which collapsed the original GO ids representing the furthest leaves of the ontology onto the abbreviated GO Slim ontology.

Category Distributions and Chi-Square Statistic

The Chi-square statistic was computed for each GO class, i.e., molecular function, biological process, or cellular location. We used the chi-square statistic to compare distributions of proteins classified into multiple categories compared to the distribution of the entire RefSeq database as the reference. The Chi-square statistic is defined as $\Sigma(O_i - E_i)^2/E_i$, where O_i is the observed frequency of elements in the ith category of the test set, E_i is the expected frequency for the same category, and the sum is overall categories $i = 1, 2,\ldots, N$ in the classification scheme. The significance of Chi-square for each category was evaluated as if it were a two-way classification, that category vs all other, using a cutoff value of chi-square around nine, corresponding to a *p*-value of about 0.0027. The low *p*-value cutoff is used to correct for multiple-hypothesis testing, because there are 27 categories being evaluated. To a first approximation, expected number of false-positives is equal to (*p*-value cutoff)*(Number of categories tested). This approach is conceptually similar to the hypergeometric method that was used previously ***(29,30)***.

Size Distribution of Human Plasma and Serum Database

We used SQL to connect the identified peptides to their full length FASTA sequence and calculated the number of proteins in increments of 100 amino acids in length.

Results

Assembling NR Human RefSeq Databases by Exact MS/MS Peptide String Search

We parsed the files where the MS/MS peptide sequence data were available to collect the set of non-redundant peptides and we refer to the results by the author or institution that supplied the data. We collapsed the peptide data into the longest set of representative peptides, matched these peptides to RefSeq, and then collapsed these results to the longest representative full-length FASTA sequences. The result was that the MS/MS peptide sequences were mapped to the longest sequences in the database that contained the exact MS/MS peptide sequences. Using the data from Table 1 in Shen et al. ***(12)*** as an example, we collapsed 6371 smaller peptides into 4209 representative longer peptides that contained the other sequences. We then found these representative, exact MS/MS peptides in 5959 RefSeq proteins. The protein sequences were likewise collapsed into the 2704 longest representative proteins that still contained the original protein and the MS/MS peptide sequences. The peptide sequences were used to create a nonredundant set of 2704 RefSeq proteins identified by Shen et al. ***(12)*** that still contained an exact match to the peptides from

Table 1
Mapping and Collapse of the MS/MS Peptide Sequences From the Reported Data From Human Blood to the RefSeq Database[a]

									More than 2, 3, 4, or 5 peptides found in a protein			
DBname	Peptides	NRpeptides	Proteins	NRproteins	NP	Percent	XP	Percent	2 pep	3 pep	4 pep	5 pep
Adkins	1397	1208	1738	585	566	96.75%	19	3.25%	144	102	69	60
PPD	1408	251	404	216	207	95.83%	9	4.17%	102	40	7	5
Tirumalai	289	253	348	311	304	97.75%	7	2.25%	5	1	0	0
Marshall	24,972	5180	7489	2571	2375	92.38%	196	7.62%	1798	1154	600	307
PPP	20,732	16,435	24,907	9303	8386	90.14%	916	9.85%	3761	1777	1134	770
Shen c	2453	1746	2439	953	916	96.12%	37	3.88%	167	112	98	83
Shen a	6371	4209	5959	2704	2581	95.45%	123	4.55%	984	305	185	131
Shen b	2444	1802	2516	2258	2184	96.72%	74	3.28%	58	0	0	0
								In total	6669	3844	2592	1771
Total		28,883		12,777	11,388	89.13%	1076	8.42%				
RefSeq				27,705	22,683	81.87%	5,022	18.13%				

[a]The MS/MS spectra obtained were correlated against proteins using MASCOT, Sequest, or other algorithms against databases containing DNA or cDNA sequences from both human and other organisms and containing both known and unknown expression products. Peptides were parsed where reported and installed as SQL databases. The peptides were searched as exact strings against each other to collapse to a nonredundant set of longest representative peptides that contain any shorter exactly matching peptide sequences. The representative peptide sequences were then searched as exact strings against the FASTA files of human RefSeq in another SQL database. The resulting RefSeq FASTA files were searched against each other to collapse into the longest representative sequences that contain any shorter FASTA files with exactly matching sequences. Raw or calculated data sources: Adkins ***(8)***; Chan ***(11)***; Jin ***(10)***; PPD is from the John Hopkins Website; Tirumalai ***(13)***; Marshall ***(5)***; PPP is from the PPP web site; Pieper ***(3)***; Shen a is from Table 1 in **ref. *12***; Shen b is from Table 2 in **ref. *12***; Shen c ***(9)***.

that paper. The groups had an average of 95% NP and 5% XP proteins and thus showed a strong bias toward expressed human RefSeq proteins and away from hypothetical proteins **(Table 1)**. Because the NP entries tended toward agreement between groups but the XP entries tended to diverge between groups the NR set yielded 89% known NP and 8% unknown XP entries.

Comparisons Between Experimental Groups by MS/MS Peptides

Searches against the exact short peptide sequence correlated by MS/MS algorithms **(Table 2)** were used to calculate the agreement on similar proteins across groups. Most groups showed significant overlap by MS/MS peptide searches. The correlated short peptide sequences showed agreement in trends with the full length protein sequences. Most groups showed strong agreement, typically about 60% with the HUPO PPP database, but it is not clear if the data sets were strictly independent in all cases.

Agreement Between BLAST Matching and Exact MS/MS Peptide Matching

The proteins matched to RefSeq by BLAST matching were compared to proteins mapped using the MS/MS peptide sequences where both types of data were available **(Table 3)**.

Peptide sequence string searches yielded considerable agreement with BLAST analysis showing similar trends and related FASTA sequences. All of the FASTA sequences obtained by 75% FL and 20 contiguous amino acids still contained the exact original MS/MS peptide sequences. The exact MS/MS string search and BLASTp matching both obtained a relevant set of nonredundant RefSeq proteins. The data collapsed to the identical RefSeq entry by both methods up to 30–40% of the time and the nonredundant set of proteins defined by BLAST matching still contained the exact MS/MS peptide sequence method in all of the entries. The similar set of related sequences obtained by the BLAST method **(Table 4)** indicates that either method is sufficient to map sets of LC/LC-MS/MS to a representative, nonredundant set of sequences to facilitate comparisons **(Table 5)**. Using the PPP data as an example, the peptide method shows an NR set of some 9303 proteins by MS/MS peptide but an NR set of 6654 proteins by the BLAST method. The number of total nonredundant proteins estimated by the peptide method was 9303 and was in close agreement with the PPP estimate of 9504 proteins. Similarly, the number of proteins with two different peptides in the PPP data set (3020) was similar to the 3761 proteins we observed by the exact MS/MS peptide method and apparently not in sharp disagreement with previous results ***(23)***. In general it appeared that the MS/MS peptide method led to about a 50% larger set of NR protein types compared to the BLAST method.

Assembling NR Human RefSeq Databases by BLAST Matching of Protein FASTA Files

We used BLASTp to match the proteins identified by the correlation analysis of spectra generated by LC-MS/MS against many different protein databases to the human RefSeq database. We used FASTA files from protein gi, IPI, or other database numbers and then mapped to the reference set of 29,234 RefSeq proteins. The mapping to RefSeq was performed by BLASTp matching of 75% full length and by exact peptide sequence string searches of greater than or equal to 20 contiguous amino acids. This mapping was many-to-one, collapsing closely related variants from the original list to a single representative RefSeq entry. We observed that most groups showed a high percentage of identified proteins that have close homologues in the NP RefSeq **(Table 4)**. We observed that all groups obtained a large number of proteins identified in serum that mapped to the RefSeq NP

Table 2
Comparison of Exact MS/MS Peptide Sequence Data From Publicly Available Human Serum or Plasma Data[a]

A DBname	Adkins	PPD	Tirumalai	Marshall	PPP	Shen c	Shen a	Shen b
Adkins	585	63	76	164	424	171	237	88
PPD	63	216	41	67	209	59	83	35
Tirumalai	76	41	311	80	209	76	124	38
Marshall	164	67	80	2571	1663	251	608	375
PPP	424	209	209	1663	9303	636	1594	1156
Shen c	171	59	76	251	636	953	510	185
Shen a	237	83	124	608	1594	510	2704	124
Shen b	88	35	38	375	1156	185	124	2258
Total count	585	216	311	2571	9303	953	2704	2258

B DBname	Adkins	PPD	Tirumalai	Marshall	PPP	Shen c	Shen a	Shen b
Adkins	100	29.2	24.4	6.4	4.6	17.9	8.8	3.9
PPD	10.7	100.0	13.2	2.6	2.2	6.2	3.1	1.6
Tirumalai	13.0	19.0	100.0	3.1	2.2	8.0	4.6	1.7
Marshall	28.0	31.0	25.7	100.0	17.9	26.3	22.5	16.6
PPP	72.5	96.8	67.2	64.7	100.0	66.7	58.9	51.2
Shen c	29.2	27.3	24.4	9.8	6.8	100.0	18.9	8.2
Shen a	40.5	38.4	39.9	23.6	17.1	53.5	100.0	5.5
Shen b	15.0	16.2	12.2	14.6	12.4	19.4	4.6	100.0

[a]Proteins identified by search strings of exact peptide sequences to the full-length human FASTA sequences of RefSeq. The proteins were identified by tandem MS/MS using PAGE or LC with or without sample prefractionation. The MS/MS spectra obtained were correlated against proteins using MASCOT, Sequest, or other algorithms against various databases containing DNA or cDNA sequences from both human and other organisms and containing both hypothetical genes (XP) and confirmed expression products (NP). Nonredundant hits were obtained by collapsing many to the one longest peptide mapping to FASTA files followed by collapsing many to one longest matching FASTA file in RefSeq. The data from each group was directly compared using SQL data base steps and is presented in **(A)** absolute numbers and **(B)** percent overlap. Adkins ***(8)***; Chan ***(11)***; Jin ***(10)***; PPD is from the John Hopkins Website; Tirumalai ***(13)***; Marshall ***(5)***; PPP is from the PPP web site; Pieper ***(3)***; Shen a is from Table 1 in **ref. *12***; Shen b is from Table 2 in **ref. *12***; Shen c ***(9)***.

Table 3
Agreement Between Mapping to RefSeq via Exact MS/MS Peptides Sequences vs BLAST Mapping[a]

DBname	Reported FASTA	NR by BLAST	NR by MS/MS	Common RefSeq Number	Identical percent BLAST	Identical percent MS/MS	MS/MS in BLAST	Percent MS/MS Peptide in BLAST
Adkins	601	490	536	180	29.95	33.5	490	100
Tirumalai	339	323	295	121	35.69	41	323	100
Marshall	2686	2259	2336	334	12.43	14.3	2259	100
Shen c	192	139	888	77	40.1	8.6	139	100
Shen a	2031	1731	2491	842	41.46	33.8	1731	100
Shen b	1636	1515	2058	720	44.01	34.9	1515	100
Total combined	6623	6457	6481	2638	39.83	40.72		

[a]Agreement was calculated by the number of identical RefSeq entries mapped by the BLAST vs exact MS/MS methods and by the existence of the exact MS/MS sequence in the resulting BLAST list. Note that the two methods for generating an NR list absolutely agree at the MS/MS peptide level 100% of the time. Adkins ***(8)***; Chan ***(11)***; Jin ***(10)***; PPD is from the John Hopkins Website; Tirumalai ***(13)***; Marshall ***(5)***; PPP is from the PPP web site; Pieper ***(3)***; Shen a is from Table 1 in **ref. *12***; Shen b is from Table 2 in **ref. *12***; Shen c ***(9)***.

Table 4
The Use of BLAST to Collapse the Full Length FASTA Files of the Publicly Available Human Serum or Plasma Proteins by the Authors Listed[a]

	Published	RefSeq	NR	Redundant	RefSeq hits percent	XP	XP percent	NP	NP percent
Tirumalai	341	313	308	5	91.78	10	3.1	303	96.8
Shen c	1474	126	123	3	8.54	3	2.4	123	97.6
Shen a	3712	2202	1667	535	59.32	110	5	2092	95
Shen b	2438	1816	1434	382	74.5	38	2.1	1778	97.9
Adkins	607	472	461	11	77.8	13	2.8	459	97.2
Pieper	324	303	297	6	93.5	4	1.3	299	98.7
Chan	1444	1246	1238	8	86.3	34	2.7	1212	97.2
Jin	1292	1010	1008	2	78.1	79	7.8	931	92.1
Marshall	2682	1904	1789	115	71	103	5.4	1801	94.6
PPP	15,710	7445	6654	791	47.4	321	4.3	7124	95.7
PPD	7518	6867	4498	2369	91.3	172	2.5	6695	97.5

[a]The MS/MS spectra obtained were correlated against proteins using MASCOT, Sequest, or other algorithms against databases containing DNA or cDNA sequences from both human and other organisms and containing both known and unknown expression products. Nonredundant hits were obtained by many to one mapping of FASTA files to RefSeq using the top BLASTp score that matched => at 75% full length and where there were also at least 20 contiguous amino acids. Adkins ***(8)***; Chan ***(11)***; Jin ***(10)***; PPD is from the John Hopkins Website; Tirumalai ***(13)***; Marshall ***(5)***; PPP is from the PPP web site; Pieper ***(3)***; Shen a is from Table 1 in **ref. *12***; Shen b is from Table 2 in **ref. *12***; Shen c ***(9)***.

Table 5
The Comparison of Overlap Between Data Sets Using the BLAST Method[a]

A

DB Name	Adkins	Chan	Jin	Pieper	PPD	PPP	Shen c	Shen a	Shen b	Tirumalai	Marshall
Adkins	607	190	77	116	634	430	77	454	141	79	241
Chan	228	1444	168	122	2034	925	73	646	256	93	397
Jin	115	188	1292	84	753	801	38	393	177	71	270
Pieper	142	110	70	324	427	311	45	308	66	67	173
PPD	426	1094	385	270	7518	3402	90	1086	651	238	818
PPP	365	691	545	234	4576	15,710	103	1387	896	194	1076
Shen c	96	75	29	48	115	118	1474	268	39	31	97
Shen a	273	380	200	170	1352	1266	118	3172	153	120	471
Shen b	134	208	122	62	1031	1009	36	318	2438	52	256
Tirumalai	108	87	63	72	381	269	30	200	67	341	134
Marshall	440	626	364	274	4172	2372	162	1240	538	200	2682
Total	607	1444	1292	324	7518	15,710	1474	3172	2438	341	2682

B

DB Name	Adkins	Chan	Jin	Pieper	PPD	PPP	Shen c	Shen a	Shen b	Tirumalai	Marshall
Adkins	100	13.2	6	35.8	8.4	2.7	5.2	14.3	5.8	23.2	9
Chan	37.6	100	13	37.7	27.1	5.9	5	20.4	10.5	27.3	14.8
Jin	18.9	13	100	25.9	10	5.1	2.6	12.4	7.3	20.8	10.1
Pieper	23.4	7.6	5.4	100	5.7	2	3.1	9.7	2.7	19.6	6.5
PPD	70.2	75.8	29.8	83.3	100	21.7	6.1	34.2	26.7	69.8	30.5
PPP	60.1	47.9	42.2	72.2	60.9	100	7	43.7	36.8	56.9	40.1
Shen c	15.8	5.2	2.2	14.8	1.5	0.8	100	8.4	1.6	9.1	3.6
Shen a	45	26.3	15.5	52.5	18	8.1	8	100	6.3	35.2	17.6
Shen b	22.1	14.4	9.4	19.1	13.7	6.4	2.4	10	100	15.2	9.5
Tirumalai	17.8	6	4.9	22.2	5.1	1.7	2	6.3	2.7	100	5
Marshall	72.5	43.4	28.2	84.6	55.5	15.1	11	39.1	22.1	58.7	100

[a]Total BLASTp hits from the listed authors that map against other groups. Raw data were obtained from the papers listed by author. Nonredundant hits were obtained by many-to-one mapping against RefSeq taking the top hit. No-redundant percent RefSeq and percent XP were calculated under the conditions: greater that 75 homology over the full length (FL) of the predicted protein product and 20 contiguous amino acids (FL%>=75% AND LPM>=20). Adkins ***(8)***; Chan ***(11)***; Jin ***(10)***; PPD is from the John Hopkins Website; Tirumalai ***(13)***; Marshall ***(5)***; PPP is from the PPP web site; Pieper ***(3)***; Shen a is from Table 1 in **ref. *12***; Shen b is from Table 2 in **ref. *12***; Shen c ***(9)***.

database. Most groups showed a comparably low percentage of XP entries. Two groups showed between 10 and 50% homology of full length FASTA with RefSeq by BLAST mapping. The vast majority of the proteins identified by the various groups, often or mostly against databases that contain a significant proportion of hypothetical and nonhuman entries, had close homologues with exact peptide sequence matches to the known expression products of a human being. In general BLAST mapping resulted in a markedly smaller set of NR proteins compared to peptide string searches.

Comparisons Between Experimental Groups by BLAST of FASTA Sequences

Compared to the smaller residual data sets left by BLAST matching **(Table 5)**, the peptide method **(Table 2)** showed larger absolute numbers of proteins and thus typically greater agreement between groups. Many data sets showed typically from 30 to 40% and up to 80% agreement. Because the exact MS/MS peptide method resulted in a larger set of related sequences that still contained the same MS/MS peptides, there was a concomitantly greater degree of overlap compared to the calculated overlap based on BLAST. There was a general agreement in trends between the two methods.

Category Distributions

String searches of exact peptide sequences that had been correlated with MS/MS fragmentation spectra by different authors led to a related set of protein sequences as compared by BLAST. Peptides are available for only some groups, but all groups have FASTA accession numbers corresponding to full length FASTA files. The BLASTp algorithm ***(26)*** was used to compare all available data sets **(Table 6)**. The proteins listed in RefSeq have the available GO terms and other information linked to the entry. We compared the category distributions of data sets from the groups to that of the RefSeq database. We observed that most groups identified an overabundance of secreted proteins such as extracellular proteins, or proteins that are shed from epithelium such as cytoskeletal proteins and soluble proteins from the cytoplasm. Conversely, ribosomal and mitochrondrial proteins and other insoluble proteins from the membranes appeared either with lower than expected frequencies or were not detected. We observed that Chi-square values for individual groups showed significant deviations from the category distribution of the RefSeq data set and thus did not appear to be randomly distributed over all categories.

NR Human Serum and Plasma Database

Because all groups provided accession numbers, we used the BLAST method to calculate a NR serum and plasma protein database. We collected the full-length protein sequences FASTA files related to the accession numbers reported by each group and combined them into a SQL database. We then used BLAST matching at 75% FL and 20 contiguous AA to determine the level of agreement between all groups. The total reported number of approx 37,000 proteins **(Table 4)** collapsed into some 19,196 proteins after BLASTing. We observed perfect agreement in all 10 groups on a set of four representative proteins, complement C3, gelsolin, transferin, and clusterin isoform 1 **(Table 6)**. Nine groups were in agreement on the top 15 sequences, four groups agreed on the top 455 sequences and at least three groups agreed on a set of 1671 types of proteins using the BLAST method assuming independence between groups and institutions. It was clear that SQL analysis of peptides and BLAST analysis of full-length protein sequences were able to collapse and compare the available data sets against a reference data base in a standardized manner.

Table 6
Category Distributions of GO Slim Terms From 75% FL and 20 AA[a]

GO SLIM	Cellular location C	RefSeq	RefSeq Hz	PPP	PPD	Shen a	Marshall	Shen b	Chan	Jin	Adkins	Tirumalai	Pieper	Shen c X2
GO:0000228	Nuclear chromosome	61	0.00228	0.7	14.9	0.7	2.5	0.1	2.3	.	.	.	.	.
GO:0005575	Cellular_ component	91	0.003402	1.4	0	–0.1	–0.5	0	.	.	.	.	.	.
GO:0005576	Extracellular region	493	0.018431	0	31.3	11	3.7	–5	4.3	3.2	52.9	44.4	123.3	23.2
GO:0005578	Extracellular matrix	609	0.022767	11.4	24.6	30.6	12.6	3	9.3	18.5	33.5	12.5	19	11.9
GO:0005615	Extracellular space	674	0.025197	1.7	85	49.7	25.7	–3.1	53.6	5.9	180.5	101	472.5	188
GO:0005622	Intracellular	1348	0.050394	–11.7	–21.4	–23.5	–8.7	–8.3	–13.4	–2	–14	–11.4	–5.8	.
GO:0005623	Cell	7470	0.279263	–11	–20.3	–0.3	–24.9	–0.2	1.3	–15.5	–4.1	–0.5	–35.3	–0.4
GO:0005634	Nucleus	4312	0.161202	–0.6	–23.7	–14.1	0.5	0	–4.6	–1.1	–21.9	–8.8	–27.9	–0.4
GO:0005635	Nuclear membrane	132	0.004935	3.3	15.2	2.1	0.9	3.3	2.2	.	1.9	.	.	.
GO:0005654	Nucleoplasm	351	0.013122	–2.1	5.8	–3.8	3.4	0	0.1	–0.3	–1.9	.	.	.
GO:0005694	Chromosome	329	0.0123	3.3	–0.6	–0.1	3.4	–0.9	0.3	–0.1	.	.	0	.
GO:0005730	Nucleolus	134	0.00501	–1.1	–1.1	–0.1	–0.4	.	0	.	.	.	.	.
GO:0005737	Cytoplasm	1555	0.058133	24.6	24.4	2.2	1.8	2.6	1.4	13.4	0	0	12.7	0
GO:0005739	Mitochondrion	975	0.03645	–7.3	–25.3	–9.7	–13.5	–4	–13.1	15.2	–0.5	0	–6.4	.
GO:0005764	Lysosome	144	0.005383	–2.1	–0.3	0	–1.4	0		0.2	.	8.2	24.1	.
GO:0005768	Endosome	82	0.003066	0.5	1.1	2.4	0.5	–0.1	.	.	.	.	.	
GO:0005777	Peroxisome	107	0.004	–1.4	–2.1	–1	–1.2			.	5	1.7	.	.
GO:0005783	Endoplasmic reticulum	739	0.027627	–0.6	–9.2	–0.3	–7	1.6	0	–0.5	–2.8	–3.6	0	.
GO:0005794	Golgi apparatus	710	0.026543	3.2	–9.5	–0.1	14.1	–2.3	0	–4	0	1.8	–1.2	.
GO:0005815	Microtubule organizing center	69	0.00258	3.4	5.8	11.5	8.9		.	.	.	.	.	.
GO:0005829	Cytosol	422	0.015776	0	0	0.2	–0.1	0	0.1	7	0	3	5.1	.
GO:0005840	Ribosome	499	0.018655	–44.7	–56.7	–30.9	–12.4	–24.6	–14.5	–0.5	.	.	.	.

(Continued)

Table 6 *(Continued)*

GO SLIM	Cellular location C	RefSeq	RefSeq Hz	PPP	PPD	Shen a	Marshall	Shen b	Chan	Jin	Adkins	Tirumalai	Pieper	Shen c X2
GO:0005856	Cytoskeleton	1377	0.051479	101	137.4	59.1	134.3	16.8	2.4	36.7	14.7	2.8	63.3	4.5
GO:0005886	Plasma membrane	2847	0.106434	0	5.8	0.4	–5.3	22.6	2.1	–4.7	10.2		–2.7	0.4
GO:0005941	Unlocalized protein complex	103	0.003851	0.1	1.6	.	–0.1	0.1	0	0.2		.	.	.
GO:0008372	Cellular component unknown	896	0.033497	–4.2	–0.2	–2.8	–2	–0.2	–1.7	0.2	–5.6	–0.1		.
GO:0016023	Cytoplasmic membrane-bound vesicle	220	0.008225	3.1	2.1	0.3	0.1	0	–0.1	–0.1	1.8	2.7	0.5	

[a]Full length FASTA sequences associated with the accession numbers reported by each author were obtained and mapped to RefSeq as indicate in **Table 3**. The associated GO terms and GO slim terms were then obtained via GOSlim file (generic.0208) and the query utility script (map2slim.pl). A similar process was applied to the entire RefSeq database to generate expected GO frequencies. We then compared the observed to expected frequencies using the chi square test. The larger the absolute chi-square value the greater the difference between the observed and expected population of the category. Extracellular, cellular, cytoplasmic, and cytoskeletal were often over represented whereas mitochondria, ribosome, microtubule, and other insoluble membranes were often under represented or absent. Adkins ***(8)***; Chan ***(11)***; Jin ***(10)***; PPD is from the John Hopkins Website; Tirumalai ***(13)***; Marshall ***(5)***; PPP is from the PPP web site; Pieper ***(3)***; Shen a is from Table 1 in **ref. *12***; Shen b is from Table 2 in **ref. *12***; Shen c ***(9)***. The significance of chi-square for each category may be evaluated as if it were a two-way classification, that category vs all other, using a cutoff value of chi-square around 9, corresponding to a *p*-value of about 0.0027. Positive values indicate more than the expected number of proteins and negative values indicate less than the expected number of proteins compared to the frequency (Hz) in RefSeq.

Table 7
A Nonredundant Database of Human Serum and Plasma Proteins Assembled by BLASTp Matching at the 75% FL and 20 Contiguous AA Level[a]

Groups	Total	Cumulative
10	4	4
9	11	15
8	12	28
7	26	55
6	37	93
5	82	176
4	278	455
3	1215	1671
2	7586	9258
1	9522	18,781

[a]The total and cumulative proteins in agreement between all the experiments summarized to date are calculated alongside the number of groups reporting that type of protein sequence.

Size Distribution of Serum and Plasma Protein Database

We examined the size distribution of the RefSeq NP data base broken into 100 amino acid increments and observed that NP proteins show a maximum of about 400–600 amino acids and that the most populated categories are less than 1000 amino acids in length. The proteins identified by mass spectrometry also show a maximum of 400–600 amino acids and that the population of categories declines to low levels after about 1000 amino acids. There are proteins in the RefSeq data base with sizes up to 33,500 amino acids in length but each size category only contains a small number of proteins and similarly small number of proteins were observed by mass spectrometry. The size distribution of peptides from proteins identified by two or more peptides remains below the size distribution curve of the RefSeq data base.

Discussion

Blood contains a small group of high abundance proteins mixed together with a very diverse group of low abundance proteins that have somehow diffused into the blood from tissues and cells ***(2)***. The abundant proteins that are maintained at high concentrations within the fluid of the blood are constantly secreted into the circulatory system. Endocrine or pericrine proteins in the tissues or circulatory system ***(2)*** may typically have concentrations in the low micromolar to high picomolar range. In contrast, proteins leaking from damaged cells might be in very low concentrations, may not include secretion signals and may include proteins released from the cells owing to pathological processes ***(31)***. If soluble proteins diffuse from cells and travel throughout the body via the blood stream, then all of these proteins in the blood should show a sequence relationship to the cDNAs that are known to be expressed in cells and tissues.

Mapping Exact MS/MS Peptide Sequences to RefSeq

Most groups that provided peptide sequence information showed a majority of peptides in known human proteins in the RefSeq NP. The number of sequences reported by most groups was similar to the number of amino acid sequences found in the set of expressed human proteins. The presence of a minority of hypothetical sequences and sequences from other species that do not match human sequences confirms that the data were searched against broad genetic databases containing many nonhuman sequences as well as predicted gene products. Nevertheless, the MS/MS spectra correlated best to sequences found in the set of known human proteins. We conclude that the majority of groups seemed to show a strong bias toward peptides that exist in the set of proteins that are known to be expressed in humans. For example, the data from **ref. *9*** originated from a search against a library that contained mostly sequences that do not show much homology to the known proteins in

RefSeq **(Table 2)** and yet the spectra from that study were still largely correlated to peptides found in known human expression products. Thus the MS/MS correlation algorithms showed a strong bias against matching fragmentation spectra to physiologically irrelevant XP sequences or sequences not found in human beings. The data from **ref. *9*** in particular seem to provide strong evidence that LC-MS/MS can identify authentic human proteins even when searched against a large and diverse database of sequences that largely show low homology to expressed human proteins.

Agreement Between BLAST and Exact Peptide Sequence Methods

The set of NR proteins obtained by string searches for the peptide sequences was compared to BLAST analysis of the full-length proteins. The absolute number of sequences mapped to RefSeq was much larger using the peptide method because the same peptides sequences may be found in several proteins. The peptide sequences from MS/MS correlation were retained in the NR set of proteins obtained from BLASTp in all cases. The two methods produced similar trends and related sequences but the BLAST method collapsed into smaller number of NR proteins. Both methods are reasonable approaches to mapping existing proteomic data to a nonredundant subset of a reference database for comparison. There are experimental data from more groups available in FASTA format and the BLASTp algorithm is an open-source mainstay of bio-informatic research and thus may be simpler method of comparing data sets for many laboratories.

BLAST Mapping FASTA Files to a Common Set of RefSeq Entries

We observed that the serum proteome expression lists contained relatively few unknown RefSeq XP automated gene predictions as compared to known NP entries. The ratio of XPs to NPs may indicate that many predicted XP proteins are not in fact expressed and therefore are not detected by the LC-MS/MS system. The data referenced ***(2,3,5,8,9,11,31)*** may be interpreted to support the capacity of correlation analysis of LC-MS/MS spectra to identify the expressed proteins of an organism but to avoid implicating the hypothetical proteins that are not actually expressed. A preponderance of known proteins was observed disproportionate to their representation in the databases at the time many of these studies were conducted. The bias toward protein sequences that are known to be expressed in humans instead of hypothetical proteins is consistent with the proposed cut-off scores for search algorithms to accurately identify peptides from complex mixtures most of the time ***(1,14,20)***. The absolute discrepancy between the two related sets of NR proteins defined by BLASTp matching compared to exact peptide searches emphasizes the failure of MS/MS search algorithms to discriminate between similar types of proteins where limited sequence coverage is available ***(5)***.

Agreement Between Groups

We observed that all groups showed some agreement but that different groups observed different subsets of RefSeq. To date there is little replicate data available comparing the effects of sample preparation methods with the list of proteins obtained but, in general, more proteins are observed with replication of the same samples. The simplest explanation of reduced agreement of low abundance proteins might be that each group used different sample preparation techniques, different search engines, and different databases and thus obtained a different subset of the proteome. Additionally, most groups have not exhaustively replicated the entire experiment and so there may be some discrepancy from sampling

error. The effect of sample preparation was perhaps best exemplified by the difference between ultra HPLC without sample preparation ***(9)*** that did not share about 30% of sequences with the same experiment performed after prefractionation ***(12)***.

Category Distributions

The BLAST algorithm produced smaller NR data sets. Full FASTA data were available for all groups. BLAST analysis of the proteins was used to compare all the groups. From the framework set out by **ref. 2** the set of high abundance proteins should be biased toward secreted extracellular proteins and the low abundance proteins should be enhanced in soluble cytoplasmic proteins but diminished in insoluble proteins from membranes. Together, the increased diffusion of soluble proteins and decreased representation of insoluble proteins should result in a nonrandom distribution of serum proteins that should show a significant bias away from a random sampling of the RefSeq database. All groups showed a significant enhancement of extracellular and matrix proteins that represent the high abundance secreted proteins and many groups showed higher frequencies of cytosolic or cytoskeletal proteins. However, most groups showed a strong bias against mitocondrial proteins, ribosomal proteins, membrane proteins, and insoluble proteins. In this sense, the experimental data are in agreement with the concept that serum is comprised of two populations of soluble proteins, one super abundant (secreted proteins) and one low abundant, that have diffused from their cells or organs ***(2)***. However, although many proteins have been identified, there is a need for a more resolved and highly reproduced sampling of serum proteins to describe the composition and structure of serum proteins in detail.

Combined Plasma and Serum Protein Database

Many of the MS/MS identifications reported to date involve proteins where the sequence coverage of MS/MS peptides is too small to determine unequivocally which protein the sequence is derived from and therefore can only be used to report types of proteins ***(5)***. Listing all of the many proteins that may contain the exact MS/MS peptide sequence is not the simplest way to present the data and would imply that more proteins have been identified than the evidence can support. Exact peptide matching also leads to higher calculated overlap values. Thus, there may also be merit in the use of BLAST matching to simplify and compare data sets if precautions are taken to ensure that the resulting set of representative protein sequences still contain the original MS/MS peptides. Several methods have been used to unify the available serum and plasma data with results ranging from as few as 50 to as many as some 3000 proteins in agreement ***(20,31)***. The BLAST and exact peptide sequence methods exhibited the capacity to determine agreement between protein lists and showed that some 1671 types of proteins are shared by at least three groups. So far 3151 proteins have been implicated with three peptides, i.e., about one-tenth of the RefSeq database. We conclude that it will be possible to develop openly available and standardized methods to compare between different populations of results using BLAST matching or exact peptide search strings.

Size Distribution of the Database

Blood is known to contain many very large proteins such as apolipoproteins, complements, and glycoproteins that are known to be processed in the blood. RefSeq NP contains many precursor proteins and other full length protein entries including some proteins that

may contain many thousands of amino acids. Some of these very long proteins may not exist physiologically. Because about 90% of the proteins in RefSeq are less than 1000 amino acids long similar in proportion to the mass spectrometry data, it appears that the data base is not skewed toward longer sequences from random matches *(32)*. The proteins with at least two or three peptides remain well below the size distribution curve of RefSeq, whereas proteins with only one peptides exceeded the size distribution of the RefSeq data base after about 800 amino acids in length. Thus, the proteins identified with two or three peptides may be mostly reliable *(33)*. However, for proteins identified by only one peptides that are greater than 800 amino acids in length it appears that more false data is being collected than the real data obtained. In summary, it would seem that there are reliable peptide ions known from about 3000 proteins that might be detected from and quantified by LC-MS with multiple ion monitoring on a triple stage quadrupole mass spectrometer.

Acknowledgments

We acknowledge Mark Evans, Chris Davies, and especially Ken Kupfer, formerly at Bayer Biotechnology in Berkeley, CA, for contributing bioinformatics methodology and advice, in particular to map proteins to the RefSeq database, manipulate GO annotations, and analyze category distributions. The work was supported by a grant to JGM from NSERC of Canada, the Heart and Stroke Foundation of Canada, and the Ontario Cancer Biomarker Network.

References

1. Washburn, M. P., Wolters, D., and Yates, J. R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247.
2. Anderson, N. L. and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845–867.
3. Pieper, R., Gatlin, C. L., Makusky, A. J., et al. (2003) The human serum proteome: display of nearly 3700 chromatographically separated protein spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins. *Proteomics* **3,** 1345–1364.
4. Koller, A., Washburn, M. P., Lange, B. M., et al. (2002) Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. USA* **99,** 11,969–11,974.
5. Marshall, J., Jankowski, A., Furesz, S., et al. (2004) Human serum proteins preseparated by electrophoresis or chromatography followed by tandem mass spectrometry. *J. Proteome Res.* **3,** 364–382.
6. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3,** 608–614.
7. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol. Cell. Proteomics* **3,** 531–533.
8. Adkins, J. N., Varnum, S. M., Auberry, K. J., et al. (2002) Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol. Cell. Proteomics* **1,** 947–955.
9. Shen, Y., Jacobs, J. M., Camp, D. G., 2nd, et al. (2004) Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. *Anal. Chem.* **76,** 1134–1144.
10. Jin, W. H., Dai, J., Li, S. J., Xia, Q. C., Zou, H. F., and Zeng, R. (2005) Human plasma proteome analysis by multidimensional chromatography prefractionation and linear ion trap mass spectrometry identification. *J. Proteome Res.* **4,** 613–619.
11. Chan, K., Lucas, D. A., Hise D., et al. (2004) Analysis of the human serum proteome. *Clinical Proteomics* **1,** 101–225.
12. Shen, Y., Kim, J., Strittmatter, E. F., et al. (2005) Characterization of the human blood plasma proteome. *Proteomics* **5,** 4034–4045.
13. Tirumalai, R. S., Chan, K. C., Prieto, D. A., Issaq, H. J., Conrads, T. P., and Veenstra, T. D. (2003) Characterization of the low molecular weight human serum proteome. *Mol. Cell. Proteomics* **2,** 1096–1103.

14. 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.
15. Yates, J. R., 3rd (1998) Database searching using mass spectrometry data. *Electrophoresis* **19,** 893–900.
16. Yates, J. R., 3rd, Eng, J. K., McCormack, A. L., and Schieltz, D. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **67,** 1426–1436.
17. Chelius, D., Huhmer, A. F., Shieh, C. H., et al. (2002) Analysis of the adenovirus type 5 proteome by liquid chromatography and tandem mass spectrometry methods. *J. Proteome Res.* **1,** 501–513.
18. 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.
19. Craig, R. and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467.
20. Ping, P., Vondriska, T. M., Creighton, C. J., et al. (2005) A functional annotation of subproteomes in human plasma. *Proteomics* **5,** 3506–3519.
21. Adams, M. D., Soares, M. B., Kerlavage, A. R., Fields, C., and Venter, J. C. (1993) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet.* **4,** 373–380.
22. Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001) The sequence of the human genome. *Science* **291,** 1304–1351.
23. Omenn, G. S., States, D. J., Adamski, M., et al. (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.
24. Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* **28,** 126–128.
25. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
26. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402.
27. Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29.
28. Camon, E., Magrane, M., Barrell, D., et al. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* **13,** 662–672.
29. Boldrick, J. C., Alizadeh, A. A., Diehn, M., et al. (2002) Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proc. Natl. Acad. Sci. USA* **99,** 972–977.
30. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.* **22,** 281–285.
31. Anderson, N. L., Polanski, M., Pieper, R., et al. (2004) The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol. Cell. Proteomics* **3,** 311–326.
32. States, D. J., Omenn, G. S., Blackwell, T. W., et al. (2006) Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat. Biotechnol.* **24,** 333–338.
33. Cargile, B. J., Bundy, J. L., and Stephenson, J. L., Jr. (2004) Potential for false positive identifications from large databases through tandem mass spectrometry. *J. Proteome Res.* **3,** 1082–1085.