The Human Plasma Proteome

A NONREDUNDANT LIST DEVELOPED BY COMBINATION OF FOUR SEPARATE SOURCES*[S]

N. Leigh Anderson‡§, Malu Polanski‡, Rembert Pieper¶‖, Tina Gatlin¶‖, Radhakrishna S. Tirumalai, Thomas P. Conrads**, Timothy D. Veenstra**, Joshua N. Adkins‡‡, Joel G. Pounds‡‡, Richard Fagan§§, and Anna Lobley§§**

We have merged four different views of the human plasma proteome, based on different methodologies, into a single nonredundant list of 1175 distinct gene products. The methodologies used were 1) literature search for proteins reported to occur in plasma or serum; 2) multidimensional chromatography of proteins followed by two-dimensional electrophoresis and mass spectroscopy (MS) identification of resolved proteins; 3) tryptic digestion and multidimensional chromatography of peptides followed by MS identification; and 4) tryptic digestion and multidimensional chromatography of peptides from low-molecular-mass plasma components followed by MS identification. Of 1,175 nonredundant gene products, 195 were included in more than one of the four input datasets. Only 46 appeared in all four. Predictions of signal sequence and transmembrane domain occurrence, as well as Genome Ontology annotation assignments, allowed characterization of the nonredundant list and comparison of the data sources. The "nonproteomic" literature (468 input proteins) is strongly biased toward signal sequence-containing extracellular proteins, while the three proteomics methods showed a much higher representation of cellular proteins, including nuclear, cytoplasmic, and kinesin complex proteins. Cytokines and protein hormones were almost completely absent from the proteomics data (presumably due to low abundance), while categories like DNA-binding proteins were almost entirely absent from the literature data (perhaps unexpected and therefore not sought). Most major categories of proteins in the human proteome are represented in plasma, with the distribution at successively deeper layers shifting from mostly extracellular to a distribution more like the whole (primarily cellular) proteome. The resulting nonredundant list confirms the presence of a number of interesting candidate marker proteins in plasma and serum. ***Molecular & Cellular Proteomics 3:311–326, 2004.***

The human plasma proteome is likely to contain most, if not all, human proteins, as well as proteins derived from some viruses, bacteria, and fungi. Many of the human proteins, introduced by low-level tissue leakage, ought to be present at very low concentrations (≪pg/ml), while others, such as albumin, are present in very large amounts (≫mg/ml). Numerous post-translationally modified forms of each protein are likely to be present, along with literally millions of distinct clonal immunoglobulin (Ig)[1] sequences. This complexity and enormous dynamic range make plasma the most difficult specimen to be dealt with by proteomics (1).

At the same time, plasma is the most generally informative proteome from a medical viewpoint. Almost all cells in the body communicate with plasma directly or through extracellular or cerebrospinal fluids, and many release at least part of their contents into plasma upon damage or death. Some medical conditions, such as myocardial infarction, are officially defined based on the increase of a specific protein in the plasma (*e.g.* cardiac troponin-T), and it is difficult to argue convincingly that there is any disease state that does not produce some specific pattern of protein change in the body's working fluid. This immense diagnostic potential has spurred a rapid acceleration in the search for protein disease markers by a wide variety of proteomics strategies.

Current methods of proteomics are only beginning to catalog the contents of plasma. Two-dimensional electrophoresis was able to resolve 40 distinct plasma proteins in 1976 (2), but, because of the dynamic range problem, this number had only grown to 60 in 1992 (3) and is substantially unchanged today, a quarter century later. It is now clear that more than two dimensions of conventional resolution are required to progress beyond this point. Recently, several truly multidimensional survey efforts have been mounted, with the result that the number of distinct proteins detected has increased dramatically. Additional dimensions of separation can be introduced at any of three levels: a) separation of intact proteins, either by specific binding (*e.g.* subtraction of defined high-abundance proteins) or continuous resolution (*e.g.* electrophoresis or chromatography); b) separation of peptides

From ‡The Plasma Proteome Institute, Washington DC 20009-3450; ¶Large Scale Biology Corporation, Proteomics Division, Germantown, MD 20876; **Laboratory of Proteomics and Analytical Technologies, SAIC-Frederick Inc., National Cancer Institute, Frederick, MD 21702-1201; ‡‡Biological Sciences Department, Pacific Northwest National Laboratory, Richland, WA 99352; and §§Inpharmatica Ltd., London, W1T 2NU, United Kingdom

Received, November 29, 2003, and in revised form, January 9, 2004

Published, MCP Papers in Press, January 12, 2004, DOI 10.1074/mcp.M300127-MCP200

[1] The abbreviations used are: Ig, immunoglobulin; MS, mass spectrometry; GO, Genome Ontology; 2DE, two-dimensional electrophoresis; NR, nonredundant; TM, transmembrane; LC, liquid chromatography; MS/MS, tandem MS; IT, ion trap.

derived from plasma proteins, either by specific binding (*e.g.* capture by anti-peptide antibodies) or continuous resolution (*e.g.* chromatography); and c) separation of peptides, and particularly their fragments, by mass spectrometry (MS). Many possible combinations of these dimensions can be implemented, the only limitations being the effort, cost, and time of analyzing many fractions or runs instead of one.

In this article, we have compared and combined data from three different multi-dimensional strategies with data from a fourth, classical source (the protein biochemistry and clinical chemistry literature) to provide a meta-level overview of both the contents and the rate of discovery of new components in plasma. The three experimental datasets are derived from 1) whole protein separation by a three-dimensional process (immunosubtraction/ion exchange/size exclusion) followed by two-dimensional electrophoresis (2DE) followed by MS identification of resolved spots (4); 2) Ig subtraction followed by trypsin digestion followed by two-dimensional liquid chromatography (LC) (ion exchange/reversed phase) followed by tandem MS (MS/MS) (5); and 3) molecular mass fractionation, followed by trypsin digestion followed by two-dimensional LC (cation exchange/reversed phase) followed by MS/MS (6). These three experimental approaches have two features in common (the removal of most Igs, by specific subtraction or size, and the use of MS for molecular identification) but otherwise they span the gamut of proteomics discovery approaches: separation at the protein level, separation at the tryptic peptide level, and a hybrid.

Combining experimental data with literature search results on proteins detected in plasma (representing a large body of accumulated "nonproteomics" data) should provide a broad perspective on plasma contents. Because the same proteins detected by various methods can be referred to by different names or accession numbers, we have used a sequence-based approach to eliminate redundancy and cluster all occurrences of the same protein. The resulting list makes it possible to examine the overlap between the various approaches and to see whether they are biased toward particular classes of proteins. In addition, a pooled nonredundant list should provide a relatively unbiased survey of the kinds of proteins present in plasma, which could have important diagnostic implications. Finally, a large list of proteins actually observed in plasma paves the way for top-down, targeted proteomics approaches to the discovery of disease markers: the development of accurate high-throughput specific assays for selected candidates from this list, as a supplement to the use of single methods for marker discovery in small sample sets. In the longer term, proteins with strong, mechanistic disease relationships may be viable therapeutic candidates as well.

DATA SOURCES AND METHODS

Lit: Literature Search

Manual Medline searches were performed searching for titles or abstracts containing human plasma or serum proteins, excluding articles on membranes, stimulation, drug, and dose. A total of 468 entries were collected, of which 458 had a human sequence accession number in one or more of the major databases.

2DEMS: Separation of Serum Proteins (LC^3/2-DE) + MS/MS Identification

Intact proteins were fractionated by chromatography and 2DE and identified by MS, generating the dataset described by Pieper *et al.* (7). Briefly, human blood sera were obtained in equal volumes from two healthy male donors (ages 40 and 80). Albumin, haptoglobin, transferrin, transthyretin, α-1-anti trypsin, α-1-acid glycoprotein, hemopexin, and α-2-macroglobulin were removed by immunoaffinity chromatography. The immunoaffinity-subtracted serum concentrate was fractionated further by sequential anion exchange and size exclusion chromatography. The resulting 66 samples were individually subjected to 2DE. All visible Coomassie Blue R250 spots were cut out, destained, reduced, alkylated, and digested with trypsin. All extracted peptides were analyzed by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) on a Bruker Biflex or Autoflex mass spectrometer (Bruker, Billerica, MA) and searched against Swiss-Prot. Those samples that did not give positive identification by MALDI-TOF where subjected to LC-MS/MS analysis by ion trap (IT) MS (Thermo Finnegan LCQ, Woburn, MA) and searched against the National Center for Biotechnology Information (NCBI) database using SEQUEST.

LCMS1: Separation of Peptide Digests of Serum Minus Ig (LC^2) + MS/MS Identification

A published dataset prepared by Adkins *et al.*, (5) was used. Briefly, human blood serum was obtained from a healthy anonymous female donor. Igs were depleted by affinity adsorption chromatography using protein A/G. The resulting Ig-depleted plasma was digested with trypsin and separated by strong cation exchange on a polysulfoethyl A column followed by reverse-phase separation on a capillary C18 column. The capillary column was interfaced to an IT-MS (Thermo Finnigan LCQ Deca XP) using electrospray ionization. The IT-MS was configured to perform MS/MS scans on the three most intense precursor masses from a single MS scan. All samples were measured over a mass/charge (*m/z*) range of 400–2,000, with fractions containing high complexity being measured with segmented *m/z* ranges. Tandem mass spectra were analyzed by SEQUEST as described using the NCBI May 2002 database.

LCMS2: Separation of Peptide Digests of Low-Molecular-Mass Serum Proteins (LC^2) + MS/MS Identification

The fourth dataset is that described by Tirumalai *et al.* (6), focused on the lower-molecular-mass plasma proteome. Briefly standard human serum was purchased from the National Institute of Standards and Technology. High-molecular-mass proteins were removed in the presence of acetonitrile using Centriplus centrifugal filters with a molecular mass cutoff of 30 kDa. The low-molecular-mass filtrate was reduced, alkylated, and digested with trypsin. The digested sample was fractionated by strong cation exchange chromatography on a polysulfoethyl A column. Reversed-phase LC was subsequently performed on 300A Jupiter C-18 column coupled on line to an IT-MS (Thermo Finnegan LCQ Deca XP). Each full MS scan was followed by three MS/MS scans where the three most abundant peptide molecular ions were selected. MS/MS spectra were searched against the a human protein database using SEQUEST.

Bioinformatics

Sequence Clustering—The Blastp protein comparison algorithm (8, 9) was used to query the sequence of each protein identified against

a database containing the aggregate sequences of all proteins identified by any method. Sequences sharing greater than 95% identity over an aligned region were grouped into "unique sequence clusters." Sequences were unmasked, and the minimum alignment length considered was 15 aa. This similarity-based approach was sufficient to group identical sequences, sequence fragments, and splice variants. Annotation in the nonredundant table was reported for the "best annotated" protein in the cluster set.

Signal Peptide Prediction—Signal peptides were predicted using the commercially available SignalP version 2.0 neural net and hidden Markov model (HMM) algorithms (10) and sigmask (11) signal masking program developed as part of Inpharmatica's Biopendium (12) protein annotation database. Each sequence received a score of +1 for a statistically significant positive signal peptide prediction from any of the three algorithms. The scores 0, 1, 2, and 3 for a particular sequence were then converted to qualitative terms "no," "possible signal," "signal," or "signal confident," respectively.

Transmembrane Prediction—Transmembrane (TM) regions were predicted using the commercial version of TMHMM version 2.0 algorithm (13). The total number of TM helices predicted per sequence was reported for each protein sequence. When a predicted TM region overlapped a predicted signal sequence (as it did in 40 cases in H_Plasma_NR_v2), this was interpreted as a signal sequence only.

Structural and Sequence-based Domain Annotation—Sequences were scanned against a library of BioPendium and iPSI-BLAST (9, 11)-like protein profiles constructed from SCOP (14), PFAM (15), PRINTS (16), and PROSITE (17) domain families. Hits to these profiles were reported at a statistical e-value cut-off of 1e-5. This cut-off was chosen to maximize profile coverage and minimize the occurrence of false positives. Sequences were not masked for low complexity or coiled coils prior to profile scanning.

Gene Ontology (GO) Term Annotation—NCBI GI number accessions for the sequences were matched to their SPTR (18) equivalents based on sequences sharing >95% sequence identity over 90% of the query sequence length. GO (19) component, process, and function terms were then extracted from text-based annotation files available for download from the GO database ftp site: ftp.geneontology.org/pub/go/gene-associations/gene_association.goa_human. For graphical reporting, a series of GO terms in each category were extracted by text searching of relevant keywords (indicated by the category names on plots) through all the assigned GO definitions. A GO component summary for the whole human proteome was prepared by applying the same approach to the complete GO human database referred to above.

Database Assembly—The nonredundant (NR) plasma database was assembled as a series of tables in a PostgreSQL relational database and queried to derive summary statistics for tables and figures shown here.

RESULTS

Number of Distinct Proteins Detected in Plasma, and the Nature of Nonredundancy

Four sets of accession numbers for proteins occurring in plasma (468 from Lit, 319 from 2DEMS, 607 [reported as 490 nonredundant accessions] from LCMS1, and 341 from LCMS2) were combined to yield 1,735 total initial accessions (Table I). A total of 55 of the input accessions referred to nonhuman sequences, and these were not considered further in the present analysis. A very conservative method of selecting distinct proteins was used in order to avoid counting sequence variants, splice variants, or cleavage products of one gene product as different: any sequences that shared a region larger than 15 aa with greater than 95% sequence identity were assigned to the same cluster and reported as a single entry in the nonredundant set. Fig. 1 shows one result of applying these criteria, in this case resulting in the assignment of 10 initial accessions to a single cluster for haptoglobin, a major plasma protein found in all four initial datasets and whose three separate subunit types are derived from a single translation product. This case also highlights the general observation that not all datasets used the same primary accession database (NCBI GI, Swiss-Prot, or RefSeq as examples). The largest cluster (109 "redundant" entries) is accounted for by Igs, where all the Ig heavy and light chains of all types were clustered together as one entry arbitrarily chosen as S40354 (an Ig κ chain sequence). Thus 6.2% of the input accessions were Igs, despite the fact that each of the experimental methods included steps to remove these molecules.

This approach is more conservative (fewer distinct proteins reported) than the methods used in some of the input data sources, which accounts for the decrease in each set when intra-set redundancy is removed (1,509 human accessions remain). When inter-set redundancies are removed (making the full list nonredundant by the criteria described above), a total of 1,175 distinct proteins remain. The entire nonredundant set, here abbreviated H_Plasma_NR_v2 (H_Plas-

TABLE I

Protein redundancy within and between datasets

The numbers accession numbers contributed by each data source and remaining after specific filter operations are shown.

	Lit	LCMS1	LCMS2	2DEMS	Total
Beginning accessions	468	607	341	319	1735
Minus nonhuman	458	580	330	312	1680
Minus intrasource redundancy and nonhuman accessions	433	475	318	283	1509
Unique to source in NR	284	334	221	141	980
Total combined NR list	–	–	–	–	1175

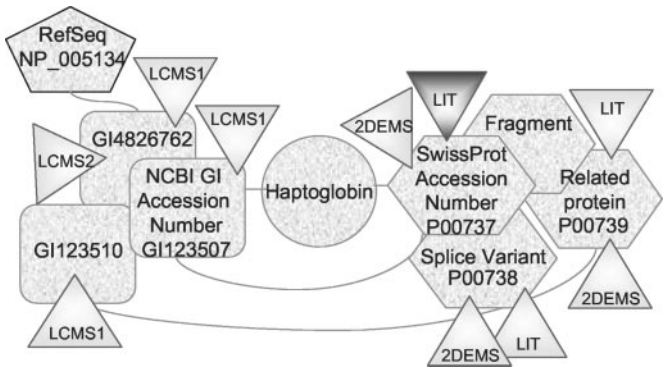

FIG. 1. **Diagram of redundant accessions for haptoglobin.** Various database accessions related to haptoglobin are shown in the rectangles, pentagon, and hexagons. The triangles represent the identifications made in the input datasets. The dark filled triangle represents the accession number upon which all other accession numbers were collapsed to give a single nonredundant accession.

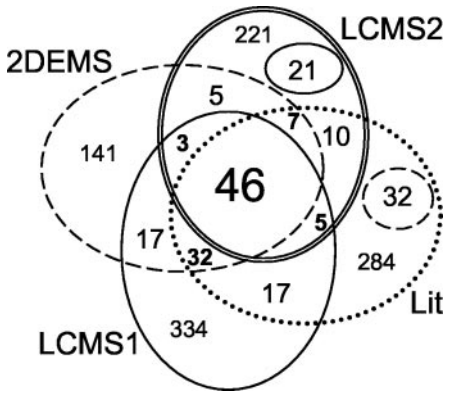

FIG. 2. **Diagram of proteins found in multiple datasets.** All overlaps are shown (2-way, 3-way, and 4-way) for all four input datasets: Lit (*dotted line*), 2DEMS (*dashed line*), LCMS1 (*solid line*), and LCMS2 (*double solid line*). Numbers represent the number of shared accessions in the respective overlapping areas.

ma_NR_v1 being Table I of Ref. 1), is provided as a supplemental data table. Of these, a total of 980 occur in only one source. Because so many entries occur only once, and given the non-zero frequency of false MS identifications, independent confirmation will be required to validate most of this list as true plasma components.

Protein Coverage by Data Source

Of the 1,175 nonredundant human proteins in H_Plasma_NR_v2, 195 entries, or 17%, were present in more than one dataset (set H_Plasma_195: Fig. 2 and Table II). Only 46 (4%) were found in all four sets of accessions (Total_sources = 4, shown in bold type in Table II). Of these only one (inter-α trypsin inhibitor heavy chain H1) is predicted to have even a single transmembrane domain, and only one (the hemoglobin β chain presumably released from red cell lysis) is predicted not to have a signal sequence. These characteristics (presence of signal sequence and absence of transmembrane domains) are those expected for major plasma proteins secreted by organs such as the liver.

An additional 47 proteins (4%) were found in three of the four datasets (Table II). Of these 47 proteins, only three were found in all three experimental datasets but not the literature dataset. These three proteins were pigment epithelium-derived factor; a nonmuscle myosin heavy chain; and secretory, extracellular matrix protein 1. Pigment epithelium-derived factor and secretory, extracellular matrix protein 1 are both secreted proteins (Swiss-Prot annotations), but upon further searching only pigment epithelium-derived factor has been reported as plasma associated (20). Nonmuscle myosin is an intracellular protein (21) possibly derived from platelets. The remaining 43 proteins seen in three datasets were all documented plasma proteins.

A further 102 proteins (9%) were found in two datasets (Table II). Of these, 43 proteins were found in two experimental datasets but not in the literature dataset. These include a number of proteins that would not typically be thought of as likely plasma components, including a chloride channel and a copper-transporting ATPase (with 10 and 7 predicted transmembrane domains, respectively), an oxygen-regulated protein, three hypothetical proteins, and a group of likely nuclear proteins including mismatch repair protein, mitotic kinesin-like protein 1, and centromere protein F.

The remaining 980 proteins (83% of NR) were found in only one of the four input datasets. Of these, 696 proteins (71%) were found only in the experimental sets, with LCMS1, LCMS2, and Lit having similar large percentages of source-unique proteins (70, 69, and 66%, respectively, *versus* 50% for 2DEMS).

Characterization of the Plasma Proteome Via Annotation Statistics

Predicted Signal Sequences—The signal sequence prediction algorithm used yielded four levels of likelihood: "no" (strong probability of no signal sequence), "possible signal," "signal," and "signal confident" (strong probability of a signal sequence) in order of increasing likelihood of a signal sequence (*i.e.* the number of algorithms out of the three used that predict a signal sequence). The procedure does not distinguish between the signal sequences of secreted and membrane-bound proteins (including *e.g.* plasma, Golgi, and mitochondrial membranes), and thus does not directly predict final protein location. Most of the 1,175 H_Plasma_NR_v2 nonredundant sequences (83%) yielded a strong positive or negative prediction (*i.e.* good agreement between the three prediction algorithms used), with these two results occurring in about a 2:3 ratio overall. Approximately 49% of H_Plasma_NR_v2 had no evidence of a signal sequence, while only about 25% of the H_Plasma_195 lacked such evidence; conversely only 34% of H_Plasma_NR_v2 gave a "signal confident" while 54% of H_Plasma_195 gave this signal. Comparing the four data sources over H_Plasma_195 (Fig. 3*A*), entries from all four sources were likely to have signal sequences: the Lit set had the highest bias toward "signal confident" proteins (indicated by a ratio of "signal confident" to "no" signal of 5.05), while 2DEMS showed less bias (ratio of 3.44) and LCMS1 and LCMS2 showed a higher representation of "no" signal predictions (ratios of 2.03 and 1.96, respectively). Comparing the four sources across all the 1,175 H_Plasma_NR_v2 proteins (Fig. 3*B*), these ratios were reduced, reflecting a greater preponderance of "no" signal proteins, but the relative differences between the sources remained, yielding ratios for Lit, 2DEMS, LCMS1, and LCMS2 of 3.85, 0.91, 0.49, and 0.44, respectively.

Predicted Transmembrane Segments—The number of predicted TM segments (0–21) is shown for each of the four datasets (Fig. 4) in comparison with the distribution summarizing The Human Membrane Protein Library (HMPL, cbs.umn.edu/human/info/hs.dat; Fig. 4*B*, *line*). The HMPL con-

TABLE II

Plasma proteins detected in at least two datasets

The table presents a nonredundant list of 195 proteins found in at least two of the four input data sources, alphabetized by protein description (containing name and synonyms). Entries found in all four data sources are shown in bold. Lit, 2DEMS, LCMS1, and LCMS2 columns give the number of accessions in each original data set that were assigned to each NR entry. These are summed across the data sources to yield Total_accessions. Total_sources summarizes the number of sources (here ranging from 2 to 4) in which the entry was found. Signal provides one of four possible values resulting from the signal sequence prediction procedure. TM gives the predicted number of transmembrane segments in the protein.

Accession	Lit	2DEMS	LCMS1	LCMS2	Total_ accessions	Total_ sources	Signal	TM	Description
P10809	1	0	1	1	3	3	No	0	60-kDa heat shock protein, mitochondrial precursor (Hsp60) (60-kDa chaperonin) (CPN60) (Heat shock protein 60) (HSP-60) (mitochondrial matrix protein P1) (P60 lymphocyte protein) (hucha60)
AAB27045	0	0	1	1	2	2	Possible signal	0	70-kDa peroxisomal membrance protein homolog (internal fragment)
P02570	2	4	0	2	8	3	No	0	Actin, cytoplasmic 1 (β-actin)
Q15848	1	1	0	0	2	2	Signal confident	0	Adiponectin precursor (30-kDa adipocyte complement-related protein) (ACRP30) (adipose most abundant gene transcript 1) (apm-1) (gelatin-binding protein)
NP_001124	0	1	1	0	2	2	Signal confident	0	Afamin precursor; α-albumin (*Homo sapiens*)
P02763	1	1	1	0	3	3	Signal confident	0	α-1-acid glycoprotein 1 precursor (AGP 1) (orosomucoid 1) (OMD 1)
P01011	1	1	2	0	4	3	Signal confident	0	α-1-antichymotrypsin precursor (ACT)
P01009	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**α-1-antitrypsin precursor (α-1 protease inhibitor) (α-1-antiproteinase) (PRO0684/ PRO2209)**
P04217	1	1	1	0	3	3	No	0	α-1B-glycoprotein precursor (α-1-B glycoprotein)
P08697	**1**	**1**	**1**	**1**	**4**	**4**	**Signal**	**0**	**α-2-antiplasmin precursor (α-2-plasmin inhibitor) (α-2-PI) (α-2-AP)**
P02765	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**α-2-HS-glycoprotein precursor (Fetuin-A) (α-2-Z-globulin) (Ba-α-2-glycoprotein) (PRO2743)**
P01023	**1**	**1**	**2**	**1**	**5**	**4**	**Signal confident**	**0**	**α-2-macroglobulin precursor (α-2-M)**
P02760	1	1	1	0	3	3	Signal confident	0	AMBP protein precursor [contains α-1-microglobulin (protein HC) (complex-forming glycoprotein heterogeneous in charge) (α-1 microglycoprotein); inter-α-trypsin inhibitor light chain (ITI-LC) (bikunin) (HI-30)]
P01019	**1**	**1**	**1**	**1**	**4**	**4**	**Signal**	**0**	**Angiotensinogen precursor [contains angiotensin I (Ang I); angiotensin II (Ang II); angiotensin III (Ang III) (Des-Asp[1]-angiotensin II)]**
P01008	**1**	**1**	**1**	**1**	**4**	**4**	**Signal**	**0**	**Antithrombin-III precursor (ATIII) (PRO0309)**
P02647	**1**	**1**	**2**	**2**	**6**	**4**	**Signal confident**	**0**	**Apolipoprotein A-I precursor (Apo-AI)**
P02652	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Apolipoprotein A-II precursor (Apo-AII) (apoa-II)**
P06727	**1**	**1**	**1**	**2**	**5**	**4**	**Signal confident**	**0**	**Apolipoprotein A-IV precursor (Apo-AIV)**
P04114	1	1	2	0	4	3	Signal confident	0	Apolipoprotein B-100 precursor (Apo B-100) [contains: apolipoprotein B-48 (Apo B-48)]
P02655	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Apolipoprotein C-II precursor (Apo-CII)**
P02656	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Apolipoprotein C-III precursor (Apo-CIII)**
P05090	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Apolipoprotein D precursor (Apo-D) (apod)**
P02649	**1**	**3**	**2**	**1**	**7**	**4**	**Signal confident**	**0**	**Apolipoprotein E precursor (Apo-E)**
Q13790	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Apolipoprotein F precursor (Apo-F)**
O14791	**1**	**1**	**1**	**1**	**4**	**4**	**Signal**	**0**	**Apolipoprotein L1 precursor (apolipoprotein L-I) (apolipoprotein L) (apol-I) (Apo-L) (apol)**
P08519	1	0	1	0	2	2	Signal	0	Apolipoprotein(a) precursor (EC 3.4.21.−) (Apo(a)) (Lp(a))
P06576	0	1	1	0	2	2	Possible signal	0	ATP synthase β chain, mitochondrial precursor (EC 3.6.3.14)
P01160	1	0	1	0	2	2	Signal confident	0	Atrial natriuretic factor precursor (ANF) (atrial natriuretic peptide) (ANP) (prepronatriodilatin) [contains: cardiodilatin-related peptide (CDP)]
P02749	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**β-2-glycoprotein I precursor (apolipoprotein H) (Apo-H) (B2GPI) (β(2)GPI) (activated protein C-binding protein) (APC inhibitor)**

TABLE II—*continued*

Accession	Lit	2DEMS	LCMS1	LCMS2	Total_ accessions	Total_ sources	Signal	TM	Description
P01884	1	0	0	1	2	2	Signal confident	0	β-2-microglobulin precursor
I39467	0	0	1	1	2	2	No	0	Bullous pemphigoid antigen, human (fragment)
P04003	1	1	1	0	3	3	Signal	0	C4b-binding protein α chain precursor (c4bp) (proline-rich protein) (PRP)
P20851	1	1	0	0	2	2	Signal confident	0	C4b-binding protein β chain precursor
P05109	1	1	0	0	2	2	No	0	Calgranulin A (Migration inhibitory factor-related protein 8) (MRP-8) (cystic fibrosis antigen) (CFAG) (P8) (leukocyte L1 complex light chain) (S100 calcium-binding protein A8) (calprotectin L1L subunit)
NP_001729	0	1	1	0	2	2	No	0	Carbonic anhydrase I; carbonic dehydratase (*Homo sapiens*)
P22792	1	1	1	0	3	3	No	0	Carboxypeptidase N 83-kDa chain (carboxypeptidase N regulatory subunit) (fragment)
P15169	1	1	0	0	2	2	Signal	0	Carboxypeptidase N catalytic chain precursor (EC 3.4.17.3) (arginine carboxypeptidase) (kinase 1) (serum carboxypeptidase N) (SCPN) (anaphylatoxin inactivator) (plasma carboxypeptidase B)
P07339	1	1	0	0	2	2	Signal confident	0	Cathepsin D precursor (EC 3.4.23.5)
P07711	1	1	0	0	2	2	Signal confident	0	Cathepsin L precursor (EC 3.4.22.15) (major excreted protein) (MEP)
P25774	0	1	0	1	2	2	Signal confident	0	Cathepsin S precursor (EC 3.4.22.27)
NP_005185	0	0	1	1	2	2	No	0	CCAAT/enhancer binding protein β, interleukin 6-dependent
O43866	1	1	0	0	2	2	Signal confident	0	CD5 antigen-like precursor (SP-α) (CT-2) (igm-associated peptide)
NP_005187	0	0	2	1	3	2	No	0	Centromere protein F (350/400kd, mitosin); mitosin; centromere
P00450	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Ceruloplasmin precursor (EC 1.16.3.1) (ferroxidase)**
NP_006421	0	0	1	1	2	2	No	0	Chaperonin containing TCP1, subunit 4 (δ); chaperonin
NP_004061	0	0	1	1	2	2	No	10	Chloride channel Ka; chloride channel, kidney, A; hclc-Ka (*Homo sapiens*)
P06276	1	1	0	0	2	2	Signal confident	1	Cholinesterase precursor (EC 3.1.1.8) (acylcholine acylhydrolase) (choline esterase II) (butyrylcholine esterase) (pseudocholinesterase)
P10909	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Clusterin precursor (complement-associated protein SP-40,40) (complement cytolysis inhibitor) (CLI) (NA1 and NA2) (apolipoprotein J) (Apo-J) (TRPM-2)**
P00740	1	1	0	1	3	3	Signal	0	Coagulation factor IX precursor (EC 3.4.21.22) (Christmas factor)
P12259	1	0	1	1	3	3	Signal confident	0	Coagulation factor V precursor (activated protein C cofactor)
P00451	1	0	1	0	2	2	Signal	0	Coagulation factor VIII precursor (procoagulant component) (antihemophilic factor) (AHF)
P00742	1	1	0	0	2	2	Signal confident	0	Coagulation factor X precursor (EC 3.4.21.6) (Stuart factor)
P00748	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Coagulation factor XII precursor (EC 3.4.21.38) (Hageman factor) (HAF)**
P00488	1	1	0	1	3	3	No	0	Coagulation factor XIII A chain precursor (EC 2.3.2.13) (protein-glutamine γ-glutamyltransferase A chain) (transglutaminase A chain)
P05160	1	1	1	0	3	3	Signal confident	0	Coagulation factor XIII B chain precursor (protein-glutamine γ-glutamyltransferase B chain) (transglutaminase B chain) (fibrin stabilizing factor B subunit)
P02462	1	0	1	1	3	3	Signal	0	Collagen α 1(IV) chain precursor
P00736	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Complement C1r component precursor**
P09871	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Complement C1s component precursor (EC 3.4.21.42) (C1 esterase)**
P06681	1	1	1	0	3	3	Signal confident	0	Complement C2 precursor (EC 3.4.21.43) (C3/C5 convertase)
P01024	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Complement C3 precursor (contains C3a anaphylatoxin)**
P01028	**1**	**1**	**2**	**1**	**5**	**4**	**Signal**	**0**	**Complement C4 precursor (contains C4A anaphylatoxin)**
P01031	1	1	1	0	3	3	Signal confident	0	Complement C5 precursor (contains C5a anaphylatoxin)
P13671	**1**	**1**	**1**	**1**	**4**	**4**	**Signal**	**0**	**Complement component C6 precursor**
P10643	1	1	1	0	3	3	Signal confident	0	Complement component C7 precursor
P07357	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Complement component C8 α chain precursor**
P07358	1	1	1	0	3	3	Signal confident	0	Complement component C8 β chain precursor
P07360	1	1	1	0	3	3	Signal confident	0	Complement component C8 γ chain precursor

TABLE II—*continued*

Accession	Lit	2DEMS	LCMS1	LCMS2	Total_ accessions	Total_ sources	Signal	TM	Description
P02748	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Complement component C9 precursor**
P00751	**1**	**1**	**1**	**1**	**4**	**4**	**Signal**	**0**	**Complement factor B precursor (EC 3.4.21.47) (C3/C5 convertase) (properdin factor B) (glycine-rich β glycoprotein) (GBG) (PBF2)**
P08603	1	1	2	0	4	3	Signal confident	0	Complement factor H precursor (H factor 1)
A40455	0	1	1	0	2	2	No	0	Complement factor H-related protein (clone H 36–1) precursor, human (fragment)
P05156	1	3	1	0	5	3	Signal	0	Complement factor I precursor (EC 3.4.21.45) (C3B/C4B inactivator)
P48740	1	1	0	0	2	2	Signal confident	0	Complement-activating component of Ra-reactive factor precursor (EC 3.4.21.−) (Ra-reactive factor serine protease p100) (rarf) (mannan-binding lectin serine protease 1) (mannose-binding protein associated serine protease) (MASP-1)
Q04656	0	1	1	0	2	2	No	7	Copper-transporting ATPase 1 (EC 3.6.3.4) (copper pump 1) (Menkes disease-associated protein)
P08185	1	1	1	0	3	3	Signal	0	Corticosteroid-binding globulin precursor (CBG) (transcortin)
P02741	1	1	0	0	2	2	Signal confident	0	C-reactive protein precursor
P06732	1	1	0	0	2	2	No	0	Creatine kinase, M chain (EC 2.7.3.2) (M-CK)
P49902	1	0	1	0	2	2	No	0	Cytosolic purine 5′-nucleotidase (EC 3.1.3.5)
1DSN	0	1	2	0	3	2	No	0	D60S N-terminal lobe human lactoferrin
P09172	1	1	0	0	2	2	Signal confident	0	Dopamine β-monooxygenase precursor (EC 1.14.17.1) (dopamine β-hydroxylase) (DBH)
JC2521	0	0	1	1	2	2	Possible signal	1	Endothelin converting enzyme (EC 3.4.24.−) 1, umbilical vein endothelial cell form, human
NP_004416	0	1	1	1	3	3	Signal confident	0	Extracellular matrix protein 1 isoform 1 precursor; secretory
P02671	1	0	1	1	3	3	Signal confident	0	Fibrinogen α/α-E chain precursor (contains fibrinopeptide A)
P02675	1	1	0	1	3	3	Signal	0	Fibrinogen β chain precursor (contains fibrinopeptide B)
P02751	**1**	**1**	**3**	**1**	**6**	**4**	**Signal confident**	**0**	**Fibronectin precursor (FN) (cold-insoluble globulin) (CIG)**
P23142	1	1	0	1	3	3	Signal confident	0	Fibulin-1 precursor
O75636	1	1	0	0	2	2	Signal confident	0	Ficolin 3 precursor (collagen/fibrinogen domain-containing protein 3) (collagen/fibrinogen domain-containing lectin 3 P35) (Hakata antigen)
P01225	1	1	0	0	2	2	Signal	0	Follitropin β chain precursor (follicle-stimulating hormone β subunit) (FSH-β) (FSH-B)
P09104	1	1	0	0	2	2	No	0	Gamma enolase (EC 4.2.1.11) (2-phospho-D-glycerate hydrolyase) (neural enolase) (NSE) (enolase 2)
P06396	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Gelsolin precursor, plasma (actin-depolymerizing factor) (ADF) (Brevin) (AGEL)**
P14136	1	1	0	0	2	2	No	0	Glial fibrillary acidic protein, astrocyte (GFAP)
Q04609	1	1	1	0	3	3	Possible signal	1	Glutamate carboxypeptidase II (EC 3.4.17.21)
A47531	0	1	1	0	2	2	Possible signal	1	Glutamyl aminopeptidase (EC 3.4.11.7), human
NP_001494	0	1	1	0	2	2	Signal	0	Glycosylphosphatidylinositol specific phospholipase D1 isoform 1
AAB34872	0	0	5	2	7	2	No	0	GP120, IHRP=ITI heavy chain-related protein (internal fragment)
JW0057	0	1	1	0	2	2	No	0	Gravin, human
P00737	**3**	**3**	**3**	**1**	**10**	**4**	**Signal confident**	**0**	**Haptoglobin-1 precursor**
P01922	1	0	0	1	2	2	No	0	Hemoglobin α chain
P02023	**1**	**1**	**1**	**1**	**4**	**4**	**No**	**0**	**Hemoglobin β chain**
P02790	1	1	1	0	3	3	Signal confident	0	Hemopexin precursor (β-1B-glycoprotein)
P05546	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Heparin cofactor II precursor (HC-II) (protease inhibitor leuserpin 2) (HLS2)**
Q04756	0	1	0	1	2	2	Signal confident	0	Hepatocyte growth factor activator precursor (EC 3.4.21.−) (HGF activator) (HGFA)
Q14520	1	0	1	0	2	2	Signal	0	HGF activator like protein (hyaluronan binding protein 2)
P04196	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Histidine-rich glycoprotein precursor (histidine-proline rich glycoprotein) (HPRG)**
P31151	0	1	1	0	2	2	No	0	Human psoriasin (s100a7)
T14738	0	0	1	1	2	2	Possible signal	0	Hypothetical protein dkfzp564a2416.1, human (fragment)

TABLE II—*continued*

Accession	Lit	2DEMS	LCMS1	LCMS2	Total_ accessions	Total_ sources	Signal	TM	Description
T08772	0	0	1	1	2	2	No	0	Hypothetical protein dkfzp586m121.1, human (fragment)
T00063	0	0	1	1	2	2	No	0	Hypothetical protein KIAA0437, human (fragment)
S40354	**11**	**14**	**78**	**6**	**109**	**4**	**Signal**	**0**	**Immunoglobulin κ chain, human**
P01591	1	1	1	0	3	3	Possible signal	0	Immunoglobulin J chain
P08476	1	0	1	0	2	2	Signal confident	0	Inhibin β A chain precursor (activin β-A chain) (erythroid differentiation protein) (EDF)
P17936	1	0	0	1	2	2	Signal confident	0	Insulin-like growth factor binding protein 3 precursor (IGFBP-3) (IBP-3) (IGF-binding protein 3)
P24593	1	0	1	1	3	3	Signal confident	0	Insulin-like growth factor binding protein 5 precursor (IGFBP-5) (IBP-5) (IGF-binding protein 5)
P35858	1	1	1	0	3	3	Signal confident	0	Insulin-like growth factor binding protein complex acid labile chain precursor (ALS)
P01343	1	0	0	1	2	2	Signal	0	Insulin-like growth factor IA precursor (IGF-IA) (somatomedin C)
P19827	**1**	**1**	**2**	**1**	**5**	**4**	**Signal confident**	**1**	**Inter-α-trypsin inhibitor heavy chain H1 precursor (ITI heavy chain H1) (Inter-α-inhibitor heavy chain 1) (Inter-α-trypsin inhibitor complex component III) (serum-derived hyaluronan-associated protein) (SHAP)**
P19823	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Inter-α-trypsin inhibitor heavy chain H2 precursor (ITI heavy chain H2) (inter-α-inhibitor heavy chain 2) (inter-α-trypsin inhibitor complex component II) (serum-derived hyaluronan-associated protein) (SHAP)**
Q06033	**1**	**1**	**1**	**1**	**4**	**4**	**Signal**	**0**	**Inter-α-trypsin inhibitor heavy chain H3 precursor (ITI heavy chain H3) (Inter-α-inhibitor heavy chain 3) (serum-derived hyaluronan-associated protein) (SHAP)**
Q14624	**1**	**2**	**1**	**1**	**5**	**4**	**Signal**	**0**	**Inter-α-trypsin inhibitor heavy chain H4 precursor**
A33481	1	0	1	0	2	2	No	0	Interferon-induced viral resistance protein mxa, human
P29459	1	0	1	0	2	2	Signal confident	0	Interleukin-12 α chain precursor (IL-12A) (cytotoxic lymphocyte maturation factor 35-kDa subunit) (CLMF p35) (NK cell stimulatory factor chain 1) (NKSF1)
P40933	1	0	0	1	2	2	Signal	0	Interleukin-15 precursor (IL-15)
P05231	1	1	0	0	2	2	Signal confident	0	Interleukin-6 precursor (IL-6) (B-cell stimulatory factor 2) (BSF-2) (interferon β-2) (hybridoma growth factor)
PC1102	0	0	2	1	3	2	No	0	Keratin 10, type I, cytoskeletal (clone HK51), human (fragment)
NP_008985	0	1	1	0	2	2	No	0	Kinesin family member 3A; kinesin family protein 3A (*Homo sapiens*)
P01042	**1**	**1**	**4**	**2**	**8**	**4**	**Signal confident**	**0**	**Kininogen precursor (α-2-thiol proteinase inhibitor) (contains bradykinin)**
P02750	1	1	1	0	3	3	Signal	0	Leucine-rich α-2-glycoprotein precursor (LRG)
P18428	1	1	0	0	2	2	Signal confident	0	Lipopolysaccharide-binding protein precursor (LBP)
P07195	1	1	0	0	2	2	No	0	L-lactate dehydrogenase B chain (EC 1.1.1.27) (LDH-B) (LDH heart subunit) (LDH-H)
P51884	0	1	1	0	2	2	Signal confident	0	Lumican precursor (keratan sulfate proteoglycan lumican) (KSPG lumican)
NP_005920	1	0	1	0	2	2	Signal confident	1	Melanoma-associated antigen p97 isoform 1, precursor
P10636	1	0	1	0	2	2	No	0	Microtubule-associated protein τ (neurofibrillary tangle protein) (Paired helical filament-τ) (PHF-τ)
I37550	0	0	1	1	2	2	No	0	Mismatch repair protein MSH2, human
Q02241	0	0	1	1	2	2	No	0	Mitotic kinesin-like protein-1 (kinesin-like protein 5)
P08571	1	1	0	0	2	2	Signal confident	0	Monocyte differentiation antigen CD14 precursor (myeloid cell-specific leucine-rich glycoprotein)
P35579	0	1	1	1	3	3	No	0	Myosin heavy chain, nonmuscle type A (cellular myosin heavy chain, type A) (nonmuscle myosin heavy chain-A) (NMMHC-A)
P12882	0	1	0	1	2	2	No	0	Myosin heavy chain, skeletal muscle, adult 1 (myosin heavy chain iix/d) (myhc-iix/d)
NP_006380	0	1	1	0	2	2	Signal confident	1	Oxygen regulated protein precursor; oxygen regulated protein
P01270	1	1	0	0	2	2	Signal	0	Parathyroid hormone precursor (Parathyrin) (PTH) (Parathormone)

TABLE II—*continued*

Accession	Lit	2DEMS	LCMS1	LCMS2	Total_ accessions	Total_ sources	Signal	TM	Description
NP_006784	0	0	1	1	2	2	Possible signal	0	Peroxiredoxin 3; antioxidant protein 1; thioredoxin-dependent
Q15648	1	0	0	1	2	2	No	0	Peroxisome proliferator-activated receptor binding protein (PBP) (PPAR binding protein) (thyroid hormone receptor-associated protein complex component TRAP220) (thyroid receptor interacting protein 2) (TRIP2) (p53 regulatory protein RB18A)
P04180	1	1	0	0	2	2	Signal confident	2	Phosphatidylcholine-sterol acyltransferase precursor (EC 2.3.1.43) (lecithin-cholesterol acyltransferase) (phospholipid-cholesterol acyltransferase)
NP_001074	0	1	1	0	2	2	No	0	Phosphodiesterase 5A isoform 1; cgmp-binding cgmp-specific
P15259	1	1	1	0	3	3	No	0	Phosphoglycerate mutase 2 (EC 5.4.2.1) (EC 5.4.2.4) (EC 3.1.3.13) (phosphoglycerate mutase isozyme M) (PGAM-M) (BPG-dependent PGAM 2) (muscle-specific phosphoglycerate mutase)
NP_006209	0	0	1	1	2	2	No	0	Phosphoinositide-3-kinase, catalytic, α polypeptide
P36955	0	1	1	1	3	3	Signal confident	0	Pigment epithelium-derived factor precursor (PEDF) (EPC-1)
P03952	1	1	1	0	3	3	Signal	0	Plasma kallikrein precursor (EC 3.4.21.34) (plasma prekallikrein) (kininogenin) (Fletcher factor)
P05155	1	1	1	0	3	3	Signal confident	0	Plasma protease C1 inhibitor precursor (C1 Inh) (C1Inh)
P02753	1	1	1	0	3	3	Signal confident	0	Plasma retinol-binding protein precursor (PRBP) (RBP) (PRO2222)
P05154	1	1	1	0	3	3	Signal confident	0	Plasma serine protease inhibitor precursor (PCI) (protein C inhibitor) (plasminogen activator inhibitor-3) (PAI3) (acrosomal serine protease inhibitor)
P05121	1	1	0	0	2	2	Signal confident	0	Plasminogen activator inhibitor-1 precursor (PAI-1) (endothelial plasminogen activator inhibitor) (PAI)
P00747	**1**	**1**	**2**	**1**	**5**	**4**	**Signal**	**0**	**Plasminogen precursor (EC 3.4.21.7) (contains angiostatin)**
P02775	1	0	0	1	2	2	Signal	0	Platelet basic protein precursor (PBP) (small inducible cytokine B7) (CXCL7)
P02776	1	0	0	1	2	2	Signal confident	0	Platelet factor 4 precursor (PF-4) (CXCL4) (oncostatin A) (Iroplact)
P43034	1	1	0	0	2	2	No	0	Platelet-activating factor acetylhydrolase IB α subunit (EC 3.1.1.47) (PAF acetylhydrolase 45 kDa subunit) (PAF-AH 45-kDa subunit) (PAF-AH α) (PAFAH α) (Lissencephaly-1 protein) (LIS-1)
NP_006197	0	0	1	1	2	2	Signal confident	1	Platelet-derived growth factor receptor α precursor (*Homo sapiens*)
NP_000436	0	0	1	1	2	2	No	0	Plectin 1, intermediate filament binding protein 500 kDa; plectin 1
P20742	1	1	1	0	3	3	Signal confident	0	Pregnancy zone protein precursor
P07288	2	0	1	0	3	2	Signal confident	0	Prostate-specific antigen precursor (EC 3.4.21.77) (PSA) (γ-seminoprotein) (kallikrein 3) (semenogelase) (seminin) (P-30 antigen)
P07237	1	1	0	0	2	2	Signal confident	0	Protein disulfide isomerase precursor (PDI) (EC 5.3.4.1) (prolyl 4-hydroxylase β subunit) (cellular thyroid hormone binding protein) (P55)
NP_002725	0	1	1	0	2	2	No	0	Protein kinase, camp-dependent, regulatory, type I, α
AAB22439	0	0	1	1	2	2	No	0	Protein tyrosine phosphatase; ptpase (*Homo sapiens*)
P00734	**1**	**1**	**2**	**1**	**5**	**4**	**Signal**	**0**	**Prothrombin precursor (EC 3.4.21.5) (coagulation factor II)**
P22614	2	0	0	1	3	2	Signal confident	0	Putative serum amyloid A-3 protein
P04626	1	0	1	0	2	2	Signal confident	2	Receptor protein-tyrosine kinase erbb-2 precursor (EC 2.7.1.112) (p185erbb2) (NEU proto-oncogene) (C-erbb-2) (tyrosine kinase-type cell surface receptor HER2) (MLN 19)
NP_005397	0	0	1	1	2	2	No	0	Rho-associated, coiled-coil containing protein kinase 1; p160rock
P49908	1	0	1	0	2	2	Signal confident	0	Selenoprotein P precursor (sep)
NP_006206	0	1	1	0	2	2	Signal confident	0	Serine (or cysteine) proteinase inhibitor, clade A (α-1)
P02787	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Serotransferrin precursor (transferrin) (siderophilin) (β-1-metal binding globulin) (PRO1400)**
P02768	1	1	1	0	3	3	Signal confident	0	Serum albumin precursor

TABLE II—*continued*

Accession	Lit	2DEMS	LCMS1	LCMS2	Total_ accessions	Total_ sources	Signal	TM	Description
P02735	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Serum amyloid A protein precursor (SAA) (contains amyloid protein A (amyloid fibril protein AA)**
P35542	1	1	1	0	3	3	Signal	0	Serum amyloid A-4 protein precursor (constitutively expressed serum amyloid A protein) (C-SAA)
P02743	1	1	0	0	2	2	Signal confident	0	Serum amyloid P-component precursor (SAP) (9.5S α-1-glycoprotein)
P27169	1	1	1	0	3	3	Signal	0	Serum paraoxonase/arylesterase 1 (EC 3.1.1.2) (EC 3.1.8.1) (PON 1) (serum aryldiakylphosphatase 1) (A-esterase 1) (aromatic esterase 1) (K-45)
P04278	1	1	0	0	2	2	Signal	0	Sex hormone-binding globulin precursor (SHBG) (sex steroid-binding protein) (SBP) (testis-specific androgen-binding protein) (ABP)
P08240	0	1	0	1	2	2	Possible signal	0	Signal recognition particle receptor α subunit (SR-α) (docking protein α) (DP-α)
AAC83183	0	0	1	1	2	2	No	0	Similar to human hsgcn1 U77700 (PID:g2282576); similar to yeast
P09486	1	1	0	0	2	2	Signal confident	0	SPARC precursor (secreted protein acidic and rich in cysteine) (osteonectin) (ON) (basement membrane protein BM-40)
P29508	2	0	1	0	3	2	No	0	Squamous cell carcinoma antigen 1 (SCCA-1) (protein T4-A)
NP_003060	0	1	1	0	2	2	No	0	SWI/SNF-related matrix-associated actin-dependent regulator of
P05452	1	1	0	0	2	2	Signal confident	0	Tetranectin precursor (TN) (plasminogen-kringle 4 binding protein)
P07996	1	1	0	1	3	3	Signal confident	0	Thrombospondin 1 precursor
P01266	1	0	1	0	2	2	Signal confident	0	Thyroglobulin precursor
P05543	1	1	0	0	2	2	Signal confident	0	Thyroxine-binding globulin precursor (T4-binding globulin)
P02766	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Transthyretin precursor (prealbumin) (TBPA) (TTR) (ATTR)**
P07477	1	0	1	0	2	2	Signal confident	0	Trypsin I precursor (EC 3.4.21.4) (cationic trypsinogen)
P19320	1	0	1	0	2	2	Signal confident	1	Vascular cell adhesion protein 1 precursor (V-CAM 1) (CD106 antigen) (INCAM-100)
P18206	0	1	0	1	2	2	No	0	Vinculin (metavinculin)
P02774	**1**	**1**	**2**	**1**	**5**	**4**	**Signal confident**	**0**	**Vitamin D-binding protein precursor (DBP) (group-specific component) (GC-globulin) (VDB)**
P07225	1	1	1	0	3	3	Signal confident	0	Vitamin K-dependent protein S precursor
P04070	1	1	0	0	2	2	Signal confident	0	Vitamin-K dependent protein C precursor (EC 3.4.21.69) (autoprothrombin IIA) (anticoagulant protein C) (blood coagulation factor XIV)
P04004	**1**	**1**	**1**	**1**	**4**	**4**	**Signal confident**	**0**	**Vitronectin precursor (serum spreading factor) (S-protein) (V75) (contains vitronectin V65 subunit; vitronectin V10 subunit; somatomedin B)**
NP_000213	1	0	0	1	2	2	Signal confident	2	V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
P04275	1	1	0	1	3	3	Signal confident	0	Von Willebrand factor precursor (vwf)
P25311	1	1	1	0	3	3	Signal confident	0	Zinc-α-2-glycoprotein precursor (Zn-α-2-glycoprotein) (Zn-α-2-GP)

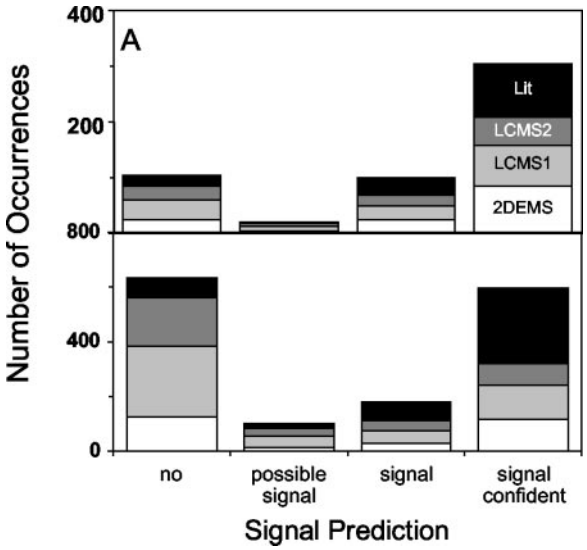

FIG. 3. **Signal sequence predictions for different data sources.** Signal sequences as predicted by three algorithms (SignalP version 2.0 neural net, HMM algorithms, and sigmask signal masking program) in the NR plasma proteome. The four outcomes (no, possible signal, signal, signal confident) correspond to 0, 1, 2, or 3 methods predicting a signal sequence. The Lit dataset is represented in *black*, LCMS2 is *dark gray*, LCMS1 is *light gray*, and 2DEMS is *white*. *A*, signal sequence predictions for proteins repeated in multiple datasets (H_Plasma_195). *B*, signal sequence predictions for nonredundant proteins (H_Plasma_NR_v2). Bar segments should be compared for relative size between sources and outcomes, but total stacked bar height is not relevant because of redundancy between the component sources.

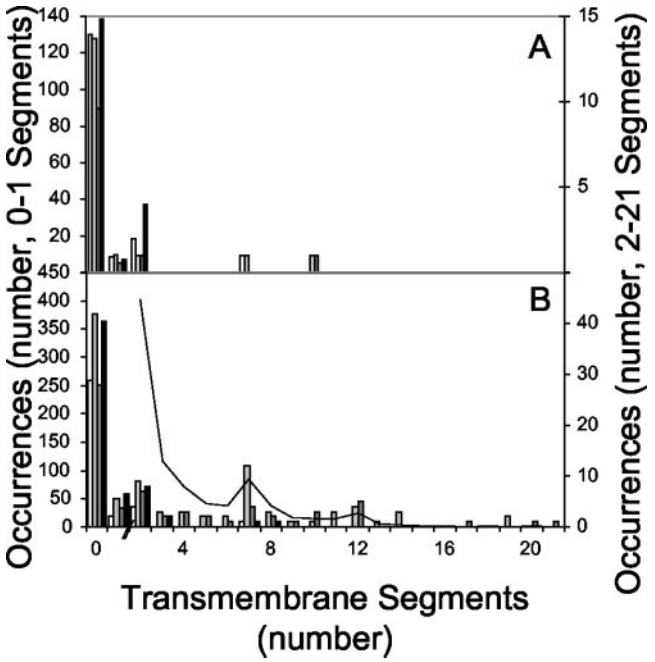

FIG. 4. **Numbers of transmembrane segments as predicted by TMHMM.** The Lit dataset is represented by *black bars*, LCMS2 by *dark gray bars*, LCMS1 by *light gray bars*, and 2DEMS by *white bars*. The *line* represents the TM distribution for all proteins contained in the Human Membrane Protein Library. The *left-hand scale* is for the experimental dataset (0–1 segments) and proteins in the Human Membrane Protein Library. The *right-hand scale* is for the experimental dataset (2–21 segments; ~10× enlarged relative to the left scale in order to show smaller values). *A*, signal sequence predictions for proteins repeated in multiple datasets (H_Plasma_195). *B*, signal sequence predictions for nonredundant proteins (H_Plasma_NR_v2).

tains predictions for a total of 8,817 proteins having 2–38 TM helices, of which 99.7% have 2–21 (covered by the range shown). About 7% of proteins of H_Plasma_195 (19 of 195) contained TM segments (Fig. 4*A*), whereas 18% of 1,175 H_Plasma_NR_v2 proteins had TM segments (Fig. 4*B*). In both cases, most proteins predicted to have a TM segment had only one. The distribution of multiple TM segments generally reflected the shape obtained for HMPL, including peaks at 7 and 12 TM segments. The Lit and 2DEMS datasets contained few transmembrane proteins, whereas the LCMS methods found more, particularly at higher TM segment numbers. LCMS1 detected more proteins with TM segments than LCMS2, which concentrated on smaller proteins.

The relationship between signal sequence and TM segment predictions across H_Plasma_NR_v2 is shown in Table III. Proteins with TM segments make up 11% of "no" signal sequence proteins, and 20% of "signal confident" proteins. The former group (+TM -Sig) shows a much higher representation of multiple TM segments than the latter (+TM +Sig), including a majority of the 7- and 12-TM entries (not all extracellular proteins or domains have a "classical" signal sequence). Sig+ proteins were much more likely (15%) than Sig- proteins (4.5%) to have a single TM segment, typical of many receptors.

GO Component Assignments—Fig. 5 presents a comparison of four different subsets of the human proteome beginning with the whole human proteome and continuing through the H_Plasma_NR and H_plasma_195 sets to those 46 seen in all four of our datasets. In this progression, there is a steady increase in the proportion of extracellular proteins and a steady decrease in the proportions of membrane, mitochondrial, and nuclear proteins. Several categories appear at a higher proportion in H_Plasma_NR_v2 than in either of the smaller versions of the plasma proteome or the whole human proteome: the kinesin complex and lysosomal and cytoskeletal proteins.

Fig. 6 uses the GO component assignments for H_Plasma_NR_v2 to further compare the four input datasets. The Lit proteins are primarily derived from extracellular (50% of the total), membrane, and cytoplasmic categories. 2DEMS accessions showed fewer extracellular and similar membrane and cytoplasmic entries, with substantial increases in a series of other categories including kinesin complex and nuclear. The two LCMS methods showed similar GO component distributions, displaying a smaller proportion of extracellular and cytoplasmic entries and more nuclear and mitochondrial entries than 2DEMS. None of the methods has a distribution close to that for the whole human proteome (Fig. 5*A*), as they would if the MS identifications were random.

TABLE III
Relationship between signal sequence and TM segment predictions over the 1,175 proteins of H_Plasma_NR_v2

Predicted TM segments	Signal prediction				
	No	Possible signal	Signal	Signal confident	Total
0	514	51	85	318	968
1	26	15	18	61	120
2	3	4	5	8	20
3	1	4	–	1	6
4	1	–	2	2	5
5	1	2	1	–	4
6	2	1	–	1	4
7	9	3	2	3	17
8	2	1	–	3	6
9	1	–	–	–	1
10	2	1	–	–	3
11	2	1	–	–	3
12	6	2	1	–	9
13	–	1	–	–	1
14	1	2	–	–	3
15	–	–	–	–	0
16	–	–	–	–	0
17	1	–	–	–	1
18	–	–	–	–	0
19	2	–	–	–	2
20	1	–	–	–	1
21	1	–	–	–	1
Total	576	88	114	397	

GO Function Assignments—A set of eight summary function categories were used to analyze aspects of function over H_Plasma_NR_v2 and to compare the four input datasets (Fig. 7). Proteins with cytokine and hormone activities, which are generally expected to be present at low abundance, were predominantly found in Lit only, while Lit contained almost none of the proteins with DNA-binding activity. Among the other categories reported, the experimental methods performed similarly except for underrepresentation of receptor and DNA-binding proteins in 2DEMS.

GO Process Assignments—Eight selected summary process categories were also used to analyze dataset differences over H_Plasma_NR_v2. As shown in Fig. 8, a number of major GO process categories were more evenly represented across the four datasets, though the representation of Lit proteins seems increased relative to the other sources, possibly due to more extensive process annotation of these molecules.

DISCUSSION

We have assembled an enlarged list of proteins observed in human plasma by combining three published experimental datasets, generated by different proteomics approaches, with a large set of proteins drawn from individual published reports on serum or plasma. Of the combined total of 1,680 human protein accession numbers, 1,175 were judged to be distinct proteins (defining set H_Plasma_NR_v2) using a very conservative set of criteria (95% sequence identity over 15 or more amino acid subsequence). As shown in Fig. 2, the overlap between the four sets is surprisingly small, with only 46 proteins occurring in all four sets, and only 195 proteins (Table II) occurring in more than one set (*i.e.* confirmed by identification using at least two approaches). This result suggests the involvement of one or more of several factors in limiting the overlap between different views of the plasma proteome: 1) the methods used may be different enough to expose quite different subsets of proteins (particularly because only a fraction of peptides observed are typically subjected to MS/MS and identified); 2) the samples used, though all human serum or plasma, may be different in the rank order of medium- and low-abundance protein components; or 3) identifications generated by some proteomics approaches could suffer from more or less random errors associated with mistaken MS/MS hits. We believe the first two factors are likely to account for the relatively small overlap, while the third should be a minimal influence for several reasons. First, the stringency of MS identification criteria used (described in detail in the original publications on each dataset) was reasonably high in each case, and false-positive identifications should represent less than 5–10% of the entries.[2] Second, the distribution of annotation characteristics observed over the 1,175 plasma NR set is quite different from the human proteome as a whole (see dataset results of Fig. 6 *versus* the whole proteome in Fig. 5*A*), which suggests that random human hits are not dominating the accessions. Finally, the sets all show fairly similar levels of difference between every pair (Fig. 2), indicating that none appears to be a major outlier. The observed differences between approaches represented here thus suggest that adding additional methods, or even repetition of these methods, would substantially expand the plasma proteome list from the total we obtained.

The set of 195 accessions that occur in at least two of the four datasets represents a confirmed list of targets that should be accessible for routine measurement by multiple proteomics technologies in human plasma or serum: these proteins have been detected by at least two methods in different laboratories. This set actually comprises more than 195 observable protein subunits because of the collapse of multiple forms into a single entry: for example haptoglobin's α and β chains collapse onto one primary gene product (cleaved after synthesis to yield the individual chains), and all Ig chains (κ and λ light chains, and γ, α, μ, ϵ, and δ heavy chains) are lumped onto one accession because of their sequence similarity. The fact that a total of 96 Ig accessions occurred in the experimental input data (14 in 2DEMS, 76 in LCMS1, and 6 in LCMS2, in addition to the 11 Ig entries in Lit) indicates that a substantial number of heterogeneous Ig sequences remain in plasma/serum samples even after the treatments used in these studies to remove antibodies prior to digestion/fractionation.

H_Plasma_195 contains most of the expected high- and medium-abundance plasma components, but also contains a number of proteins that might not ordinarily be expected to be

[2] J. N. Adkins, manuscript in preparation.

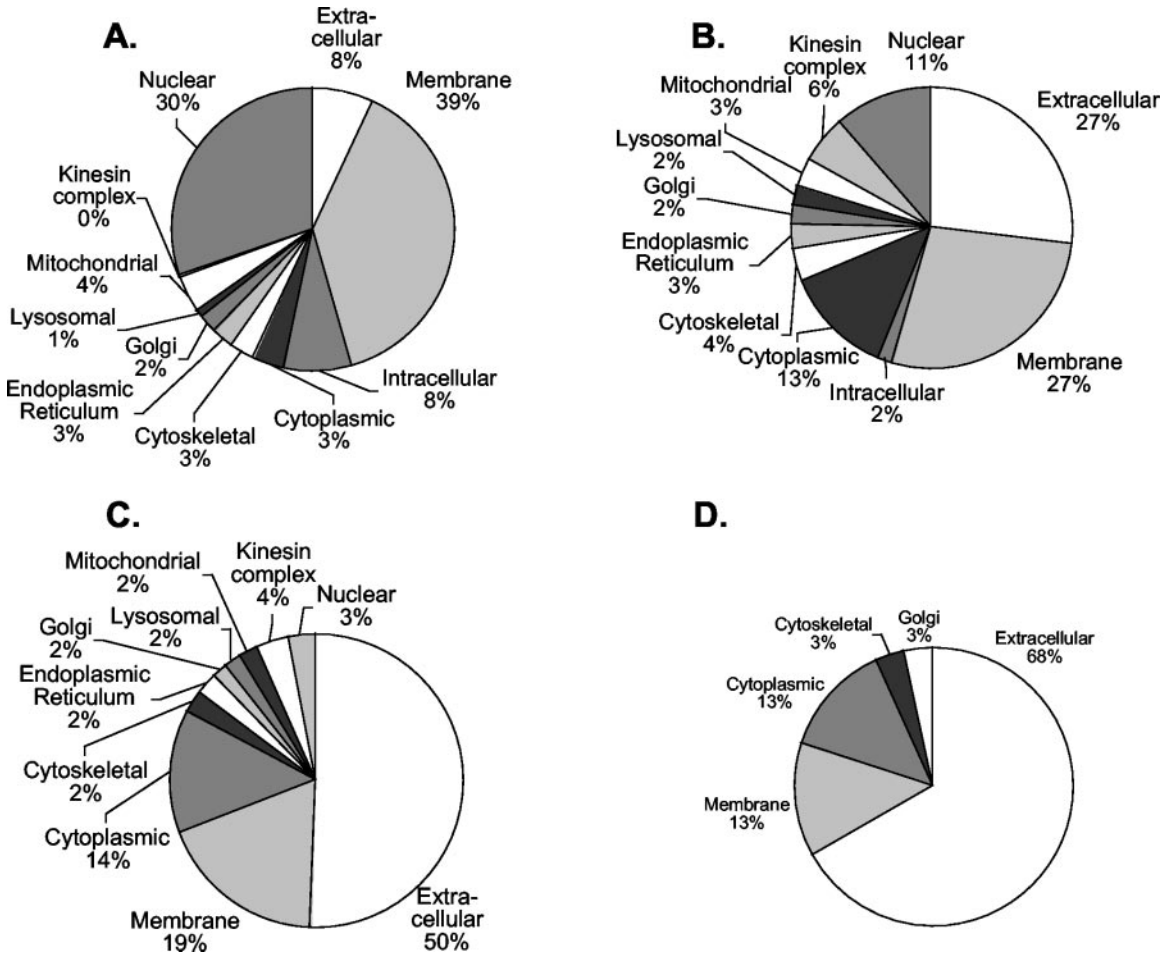

FIG. 5. **Major GO_component categories for proteome subsets.** The proportions of proteins assigned to major component categories are shown for the human proteome (*A*); the nonredundant protein set H_Plasma_NR_v2 (*B*); the H_Plasma_195 proteins seen in more than two datasets (*C*); and the 46 proteins seen in all four datasets (*D*).

abundant enough to appear in a "common component" list. These include adiponectin (involved in the control of fat metabolism and insulin sensitivity), atrial natriuretic factor (a potent vasoactive substance synthesized in mammalian atria and thought to play a key role in cardiovascular homeostasis), various cathepsins (D, L, S), centromere protein F (involved in chromosome segregation during mitosis), creatine kinase M chain (an abundant muscle enzyme), glial fibrilary acid protein (distinguishes astrocytes from other glial cells), psoriasin (S-100 family, highly up-regulated in psoriatic epidermis), interferon-induced viral-resistance protein MxA (confers resistance to influenza virus and vesicular stomatitis virus), melanoma-associated antigen p97 (a proposed cancer marker also expressed in multiple normal tissues), mismatch repair protein MSH2 (involved in postreplication mismatch repair, and whose defective forms are the cause of hereditary nonpolyposis colorectal cancer type 1), oxygen-regulated protein (which plays a pivotal role in cytoprotective cellular mechanisms triggered by oxygen deprivation), peroxisome proliferator-activated receptor binding protein (which plays a role in transcriptional coactivation), prostate-specific antigen (a protease involved in the liquefaction of the seminal coagulum, and one of the few successful cancer diagnostics), selenoprotein P (contains selenocyteines encoded by the opal codon, UGA), signal recognition particle receptor α subunit (an integral membrane protein ensuring, in conjunction with srp, the correct targeting of the nascent secretory proteins to the endoplasmic reticulum membrane system), squamous cell carcinoma antigen 1 (which may act as a protease inhibitor to modulate the host immune response against tumor cells), and V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog (the receptor for stem cell factor). A number of these proteins have obvious relevance to important disease mechanisms, and thus are of potential diagnostic value. Cathepsin S, centromere protein F, psoriasin, mismatch repair protein MSH2, oxygen-regulated protein, and signal recognition particle receptor α subunit did not occur in the Lit accession list, but were rather found via detection in two of the experimental datasets.

Two types of protein features (signal sequences and TM

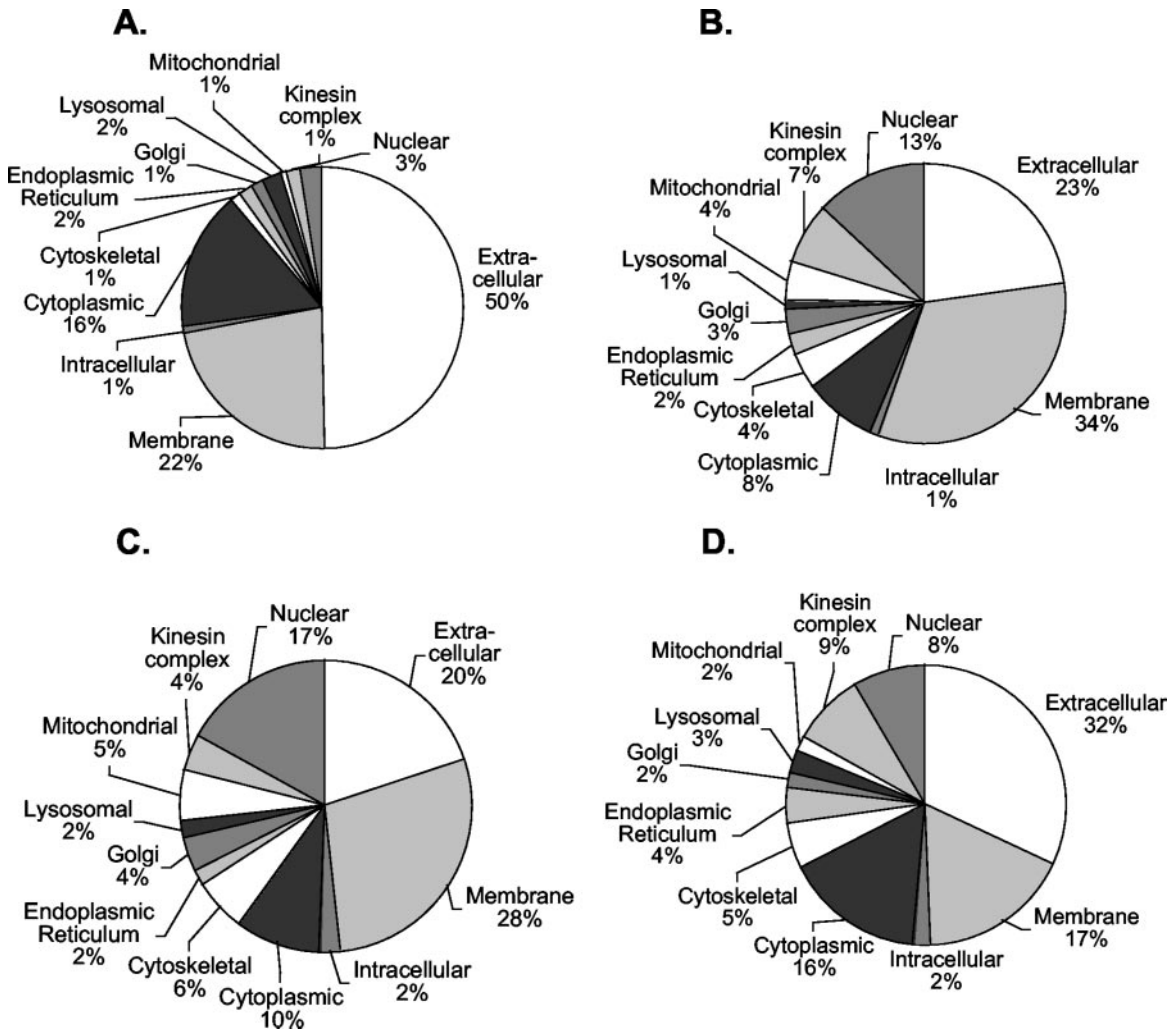

FIG. 6. **Major GO_component categories in H_Plasma_NR_v2 by data source.** Major component categories (same as Fig. 5) are shown for Lit (*A*), LCMS1 (*B*), LCMS2 (*C*), and 2DEMS (*D*) data sources.

domains) were predicted from the H_Plasma_NR_v2 sequences. These two parameters are somewhat related, because transmembrane, as well as secreted, proteins are likely to contain signal sequences. In the 1,175 proteins of H_Plasma_NR_v2, approximately one-third were confidently predicted to contain signal sequences (34% overall, with 32% having 0 or 1 TM segments) as compared with 19% containing signal sequences over the whole human proteome (and only 10% containing signal sequences with 0 or 1 TM segments; R.F., unpublished observation). H_Plasma_NR_v2 is thus substantially (~3:1) enriched in a set of proteins having signal sequences and 0 or 1 TM segments (compared with the genome), which is consistent with the presence of a large number of classical secreted proteins. However, because more than half of H_Plasma_NR_v2 proteins do not contain a signal sequence, the total representation of nonsecreted molecules (presumably cellular constituents) is high.

In the full NR set of 1,175 proteins (H_Plasma_NR_v2), several major groups of proteins occur in patterns that suggest interesting biases between our four data sources. At least 10 transcription factors were observed in the experimental sets (each by only a single method), and none of these were found in the Lit accession set. Similarly, proteins GO-annotated with a DNA-binding function were essentially absent from the Lit set. In contrast, only 4 of 39 cytokines and growth factors included were found in any of the experimental datasets (IL-6, IL-12A, ciliary neurotrophic factor, and FGF-12), while 37 occurred in the Lit set. These results suggest that while the experimental proteomics methods were not sensitive enough to detect most cytokines and hormones, they did detect important classes of proteins not detected in literature reports using targeted assay methods. On a more global level, the distribution of GO_component assignments (Fig. 6) shows substantial differences overall between the set of proteins found in a literature search *versus* the three experimental proteomics technologies. Predicted features of protein sequence also show major source-related differences. Most striking is the fact that the Lit set was strongly biased toward

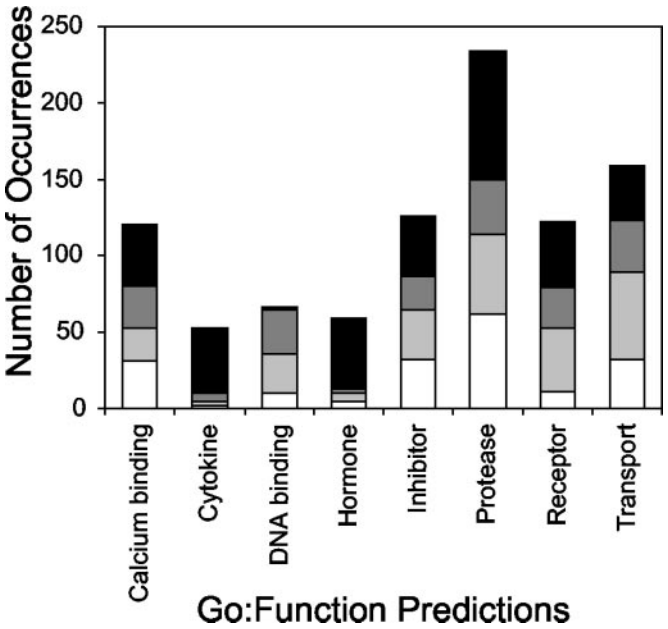

FIG. 7. **Major categories of GO_function assignments over H_Plasma_NR_v2 by data source.** Cellular function categories were extracted from GO_function annotations. The Lit dataset is represented by *black bars*, LCMS2 by *dark gray bars*, LCMS1 by *light gray bars*, and 2DEMS by *white bars*.

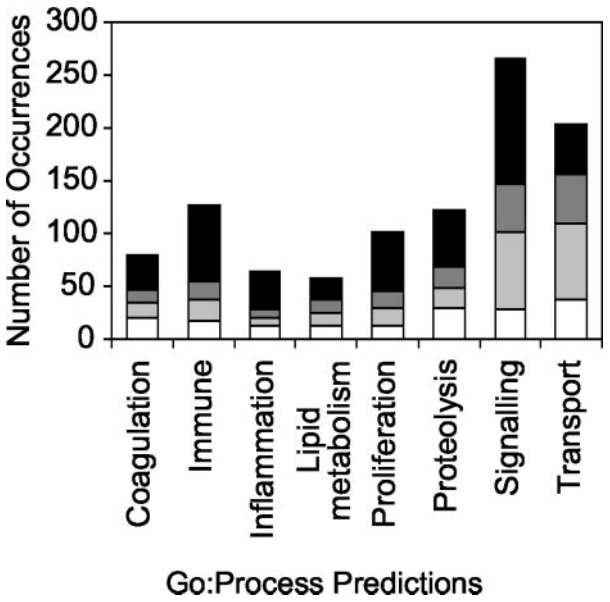

FIG. 8. **Major GO_process categories over H_Plasma_NR_v2 by data source.** The Lit dataset is represented by *black bars*, LCMS2 by *dark gray bars*, LCMS1 by *light gray bars*, and 2DEMS by *white bars*.

proteins that were confidently predicted to have signal sequences (ratio of "signal confident" to "no" of 3.85), while 2DEMS showed little preference (0.91) and the LCMS methods showed a moderate (2 to 1) bias toward proteins without signal sequences (ratios of 0.49 and 0.44). The strong bias of the Lit set toward signal sequences is likely due to the greater ease with which these more soluble proteins can be isolated and studied by biochemical techniques. Similarly, the difference between 2DEMS and LCMS may be due to the failure of many less-soluble proteins to focus in the first (isoelectric focusing) dimension of the 2DE procedure (*e.g.* intact membrane or very large proteins), or else to the presence of numerous isoforms that divide the protein among members of a charge train, and thus decrease the limit of detection (*e.g.* heavily glycosylated extracellular domains cleaved from membrane proteins). By placing fewer requirements on the behavior of sample proteins prior to digestion, and by providing identifications based on a few soluble peptides, the MS-based techniques provided a significantly less biased, though not necessarily more complete, view of the plasma proteome.

Because the four input protein sets were of similar size, it is clear that the literature, viewed as an historical summary of research on proteins in plasma, shows a bias toward secreted proteins and against investigation of cellular proteins in solution in blood. This effect cannot be due to detection sensitivity alone, because the low-abundance cytokines, generally not detected by the experimental proteomics methods, are accessible via immunoassay and widely reported in the literature. The bias may instead be due to a general skepticism that detectable amounts of many cellular proteins are being released into plasma (absent some major cause of tissue damage), or to a view that cellular protein release, if it occurred, would not be especially informative. The present results, built on experimental studies of multiple groups, demonstrate that many (perhaps all) cellular proteins are present in plasma. The demonstrated utility of cardiac muscle protein markers as serum diagnostics for myocardial infraction provides a persuasive argument that many may have diagnostic use.

The next major challenge thus becomes the systematic exploration of protein abundance and structural modification in relation to disease, normal physiological processes, and treatment effects. In a sense this shift can be seen as analogous to the current evolution in pharmaceutical target selection: the genome has provided a wealth (in practical terms an overabundance) of previously unknown therapeutic targets, creating a major challenge in selecting those that are "druggable" and specifically linked to disease mechanisms. A shift, in other words, from protein discovery to target validation. In the context of diagnostics based on proteins in blood, we now have in H_Plasma_NR_v2, and a growing body of other experimental data, a substantial set of candidate disease markers that can be detected in plasma. While these will continue to be supplemented by discovery techniques (22), the stage is set for systematic efforts to validate disease markers for near-term application in clinical trials, medium-term use in disease detection, staging, and therapy selection, and long-term use in population screening. While some of these proteins have been examined individually as potential markers and found to have low sensitivity or specificity as individual tests, growing evidence indicates that these limitations may be overcome using fingerprints of change across panels of proteins that

together better represent patient status. Thus even the present H_Plasma_NR_v2 offers an abundance of candidates deserving of measurement in selected clinical sample sets. Making such measurements, accurately and on a large scale presents a series of technical challenges likely to require substantial efforts for much of the next decade. The results of such a "targeted" proteomics effort can transform diagnostics, improve therapy, and lead to substantial and needed improvements in the economics of healthcare.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

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

§ To whom correspondence should be addressed: The Plasma Proteome Institute, P.O. Box 53450, Washington DC 20009-3450. Tel.: 301-72811451; Fax: 202-234-9175; E-mail: leighanderson@plasmaproteome.org.

‖ Current address: The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850.

REFERENCES

1. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: History, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845–867
2. Anderson, L., and Anderson, N. G. (1977) High resolution two-dimensional electrophoresis of human plasma proteins. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5421–5425
3. Hughes, G. J., Frutiger, S., Paquet, N., Ravier, F, Pasquali, C., Sanchez, J. C., James, R., Tissot, J. D., Bjellqvist, B., and Hochstrasser, D. F. (1992) Plasma protein map: An update by microsequencing. *Electrophoresis* **13,** 707–714
4. Pieper, R., Su, Q., Gatlin, C. L., Huang, S. T., Anderson, N. L., and Steiner, S. (2003) Multi-component immunoaffinity subtraction chromatography: An innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics* **3,** 422–432
5. Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J., Angell, N. H., Smith, R. D., Springer, D. L., and Pounds, J. G. (2002) Toward a human blood serum proteome: Analysis by multidimensional separation coupled with mass spectrometry. *Mol. Cell. Proteomics* **1,** 947–955
6. 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
7. Pieper, R., Gatlin, C. L., Makusky, A. J., Russo, P. S., Schatz, C. R., Miller, S. S., Su, Q., McGrath, A. M., Estock, M. A., Parmar, P. P., Zhao, M., Huang, S. T., Zhou, J., Wang, F., Esquer-Blasco, R., Anderson, N. L., Taylor, J., and Steiner, S. (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
8. 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
9. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402
10. Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10,** 1–6
11. Swindells, M., Rae, M., Pearce, M., Moodie, S., Miller, R., and Leach, P. (2002) Application of high-throughput computing in bioinformatics. *Philos. Transact. Ser. A. Math. Phys. Eng. Sci.* **360,** 1179–1189
12. Michalovich, D., Overington, J., and Fagan, R. (2002) Protein sequence analysis in silico: Application of structure-based bioinformatics to genomic initiatives. *Curr. Opin. Pharmacol.* **2,** 574–580
13. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305,** 567–580
14. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540
15. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002) The pfam protein families database. *Nucleic Acids Res.* **30,** 276–280
16. Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A., and Zygouri, C. (2003) Prints and its automatic supplement, preprints. *Nucleic Acids Res.* **31,** 400–402
17. Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., and Bucher, P. (2002) PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief Bioinform.* **3,** 265–274
18. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31,** 365–370
19. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29
20. Petersen, S. V., Valnickova, Z., and Enghild, J. J. (2003) Pigment-epithelium-derived factor (pedf) occurs at a physiologically relevant concentration in human blood: Purification and characterization. *Biochem. J.* **374,** 199–206
21. Toothaker, L. E., Gonzalez, D. A., Tung, N., Lemons, R. S., Le Beau, M. M., Arnaout, M. A., Clayton, L. K., and Tenen, D. G. (1991) Cellular myosin heavy chain in human leukocytes: Isolation of 5′ cDNA clones, characterization of the protein, chromosomal localization, and up-regulation during myeloid differentiation. *Blood* **78,** 1826–1833
22. Liotta, L. A., Ferrari, M., and Petricoin, E. (2003) Clinical proteomics: Written in blood. *Nature* **425,** 905