Protannotator: A Semiautomated Pipeline for Chromosome-Wise Functional Annotation of the "Missing" Human Proteome

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · DECEMBER 2013

Impact Factor: 4.25 · DOI: 10.1021/pr400794x · Source: PubMed

6 AUTHORS, INCLUDING:

Gagan Garg
Macquarie University
12 PUBLICATIONS **68** CITATIONS
SEE PROFILE

Brian Risk
Geneffects Software
8 PUBLICATIONS **4,343** CITATIONS
SEE PROFILE

Mark S Baker
Macquarie University
137 PUBLICATIONS **3,266** CITATIONS
SEE PROFILE

Journal of proteome research

Article

pubs.acs.org/jpr

Protannotator: A Semiautomated Pipeline for Chromosome-Wise Functional Annotation of the "Missing" Human Proteome

Mohammad T. Islam,[†,‡,#] Gagan Garg,[†,‡,#] William S. Hancock,[§] Brian A. Risk,[∥] Mark S. Baker,[†] and Shoba Ranganathan*[,†,‡,⊥]

[†]Department of Chemistry and Biomolecular Sciences and [‡]ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, NSW 2109, Australia

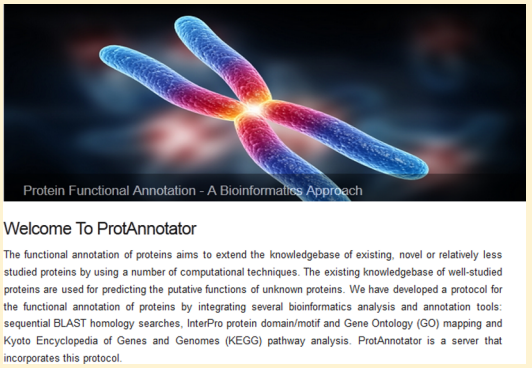
[§]Barnett Institute, Northeastern University, 140 The Fenway, Boston, Massachusetts 02115, United States

[∥]College of Arts and Sciences, Boise State University, 1910 University Drive, Boise, Idaho 83725, United States

[⊥]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 14 Medical Drive, 117599 Singapore

Supporting Information

ABSTRACT: The chromosome-centric human proteome project (C-HPP) aims to define the complete set of proteins encoded in each human chromosome. The neXtProt database (September 2013) lists 20 128 proteins for the human proteome, of which 3831 human proteins (~19%) are considered "missing" according to the standard metrics table (released September 27, 2013). In support of the C-HPP initiative, we have extended the annotation strategy developed for human chromosome 7 "missing" proteins into a semiautomated pipeline to functionally annotate the "missing" human proteome. This pipeline integrates a suite of bioinformatics analysis and annotation software tools to identify homologues and map putative functional signatures, gene ontology, and biochemical pathways. From sequential BLAST searches, we have primarily identified homologues from reviewed nonhuman mammalian proteins with protein evidence for 1271 (33.2%) "missing" proteins, followed by 703 (18.4%) homologues from reviewed nonhuman mammalian proteins and subsequently 564 (14.7%) homologues from reviewed human proteins. Functional annotations for 1945 (50.8%) "missing" proteins were also determined. To accelerate the identification of "missing" proteins from proteomics studies, we generated proteotypic peptides *in silico*. Matching these proteotypic peptides to ENCODE proteogenomic data resulted in proteomic evidence for 107 (2.8%) of the 3831 "missing proteins, while evidence from a recent membrane proteomic study supported the existence for another 15 "missing" proteins. The chromosome-wise functional annotation of all "missing" proteins is freely available to the scientific community through our web server (http://biolinfo.org/protannotator).



KEYWORDS: *Human Proteome Project, human chromosome, missing proteins, sequential BLAST, functional annotation, proteotypic peptides, proteogenomics*

INTRODUCTION

The interpretation of the human genome depends on detailed annotation, usually at the nucleotide level, the protein level, and the process level,[1] for which the functional annotation of proteins is crucial at the process level. Since 2008, the Human Proteome Organization (HUPO) has pursued the comprehensive identification and functional characterization of the human proteome via the Human Proteome Project (HPP),[2] of which the chromosome-centric HPP (C-HPP) approach seeks to catalog the human proteome on the basis of chromosomes.[3−5] The International Chromosome-centric Human Proteome Project (C-HPP), launched in 2012, marks the first step toward the genome-wide chromosome by chromosome characterization of the human proteome.[6] Such an approach would address a key aim of the human genome project, viz. personalized medicine, by providing sensitive and highly specific protein biomarkers for early onset diagnosis, prognosis and treatment of several diseases, providing clinical and translational proteomic solutions.[7]

The three pillars of HPP are mass spectrometric proteomics, antibody/affinity capturing agents, and a knowledgebase,[2] embodied by the neXtProt database,[8] where detailed information on the human proteome is collated, curated, and organized for rapid access of information on a query protein. Our group carried out the functional annotation of missing

Special Issue: Chromosome-centric Human Proteome Project

Received: August 1, 2013
Published: December 6, 2013

proteins of human chromosome 7 (hChr7),[9] developing a sequential BLAST homology search approach, with the neXtProt[8] data available at the time. A list of missing proteins was also compiled and investigated for chromosome 17,[10] where proteins having no proteomic identifications from sources including PeptideAtlas[11] or GPM (http://www.thegpm.org/) were considered "missing." Currently, for the entire human genome, neXtProt (as of September 2013) lists 20 128 proteins, by extending their data sources to incorporate all peptides from PeptideAtlas Human builds (August 2013) as "GOLD" (i.e., <1% error) as well as 20 other studies (unpublished data). Thus, 3831 proteins (~19%; excluding unmapped and redundant sequences) are considered "missing," based on the currently available C-HPP standard metrics table, developed with neXtProt data (http://www.c-hpp.org/gnuboard4/bbs/board.php?bo_table=public).

The UniProtKB/Swiss-Prot database[12] (release 2013_10) with >541 000 entries of reviewed and annotated proteins serves as the highest quality database for bioinformatics studies. Homologous sequences often display identical or similar biological functions. The biological knowledge available in this database can be mined by similarity searches to identify if any of these missing proteins are homologous to similar proteins in higher mammals or other species. BLAST[13,14] programs are widely used for sequence similarity searches, using the default nonredundant (NR) data sets, which include putative, unannotated, or translated coding regions. Thus, a similarity search against NR data sets may result in matches to large numbers of unreviewed or unannotated proteins. Previously, we have annotated less studied organisms such as helminth parasites and fungal pathogens[15−19] by combining similarity searches and functional annotations including gene ontology (GO), biochemical pathways, and functional domains and motifs. Following a targeted BLAST approach, labeled "sequential BLAST search" from our previous hChr7 "missing" protein annotation,[9] where we have run repeated BLAST searches against selected databases providing high-quality reviewed annotations, we now present a semiautomated pipeline for the annotation of "missing" proteins. This is a generic approach and can be adopted for the annotation of any novel proteome, for example, black Périgord truffle (*Tuber melanosporum*).[20] Using this approach, we have annotated the entire set of "missing" proteins in the human proteome. Out of 3831 "missing" proteins, 1271 sequences (33.2%) were homologous to nonhuman reviewed mammalian proteins with proteomic evidence, while 703 proteins (18.4%) had nonhuman reviewed mammalian homologues. 1945 (50.8%) of the "missing" proteins were assigned putative GO and domain/motif annotations, using strict parameters (detailed in Materials and Methods), while 1250 (32.7%) "missing" proteins were mapped to biochemical pathways. We have also generated proteotypic peptides to facilitate proteomic identification of the "missing" proteins. These proteotypic peptides enabled us to garner proteomic evidence for 107 "missing" proteins, using proteogenomic data accurately matching the peptides from the ENCODE project.[21−23] Also, a recent in-depth proteomic study of breast cancer tissues by Muraoka et al.[24] has reported 851 membrane proteins that currently lack evidence by mass spectrometry in the neXtProt database. From this study, we have identified 15 additional "missing" proteins, which together with the ENCODE proteogenomic data have provided proteomic evidence for 122 "missing" proteins. The annotated data for the human proteome have been compiled into a database, which is freely available to the scientific community.

MATERIALS AND METHODS

1. Data Sources

Chromosome reports for each human chromosome were downloaded from the neXtProt database[8] (release September 2013). From these reports, sequences for "missing" proteins were extracted in FASTA format. A number of protein data sets were downloaded from UniProtKB/Swiss-Prot database[12] to our local Linux server for database similarity search. These include nonhuman reviewed mammalian proteins with experimental evidence (14 910 sequences), nonhuman reviewed mammalian proteins (45 926 sequences), human-reviewed proteins (23 515 sequences), and Protein Data Bank (PDB)[25] proteins (260 382 sequences), as in our hChr7 study.[9] We used the PDB to obtain possible matches against proteins with known structures from nonmammalian organisms. Verification data sets used for this study comprise the set of all mammalian proteins (1 155 455 sequences) and the nonmammalian protein set with protein evidence (70 830 sequences).

2. Database Similarity Search

Database similarity search technique is used to identify if a novel sequence is homologous to sequences that are already available in existing databases. BLAST[13,14] is A widely used tool for sequence similarity search. A query sequence is considered to be strongly homologous if it matches against a subject sequence with high significance (E value: 1×10^{-5}, compared with 1×10^{-3})[26] and sequence identity of at least 50%.[27]

We ran BLASTP searches sequentially against the data sets previously described using default parameters with a minimum E value of 1×10^{-5}. Those sequences that yielded no matches against the first data set were matched against the next data set, then the third, and so on. Missing proteins were also functionally annotated based on GO, pathways, and protein domain mapping. This is described in detail in Section 5 (Protannotator bioinformatics pipeline).

3. Functional Annotation of Missing Proteins

"Missing" proteins are provided putative functional annotation by mapping to protein domain, motif, and families. Functional annotations were further strengthened by assigning GO terms to the proteins. InterProScan[28] is widely used for protein functional annotation. It scans the protein sequences using the different protein signature recognition methods (Hidden Markov Model and BLAST) in its 13 protein domain and functional site databases combined in InterPro[29] database.

To obtain the best results, we ran InterProScan with default programs, described in detail in our previous paper,[9] while KOBAS[30] (KEGG Orthology-Based Annotation System, KOBAS-2.0) results provided pathway mapping. The program identifies statistically significantly enriched pathways by first mapping the proteins to genes in KEGG GENES based on BLAST searches followed by mapping against the whole human genome as the background. These programs have been successfully employed for the comprehensive annotation of novel and uncharacterized sequences[15−18] and in recent genome projects of less studied organisms.[31,32]

4. *In Silico* Tryptic Digestion

Protein Digestion Simulator[33] was used with default parameters (fragment mass range of 400−6000 Da; pI range of 0−14; mass

dx.doi.org/10.1021/pr400794x | J. Proteome Res. 2014, 13, 76−83

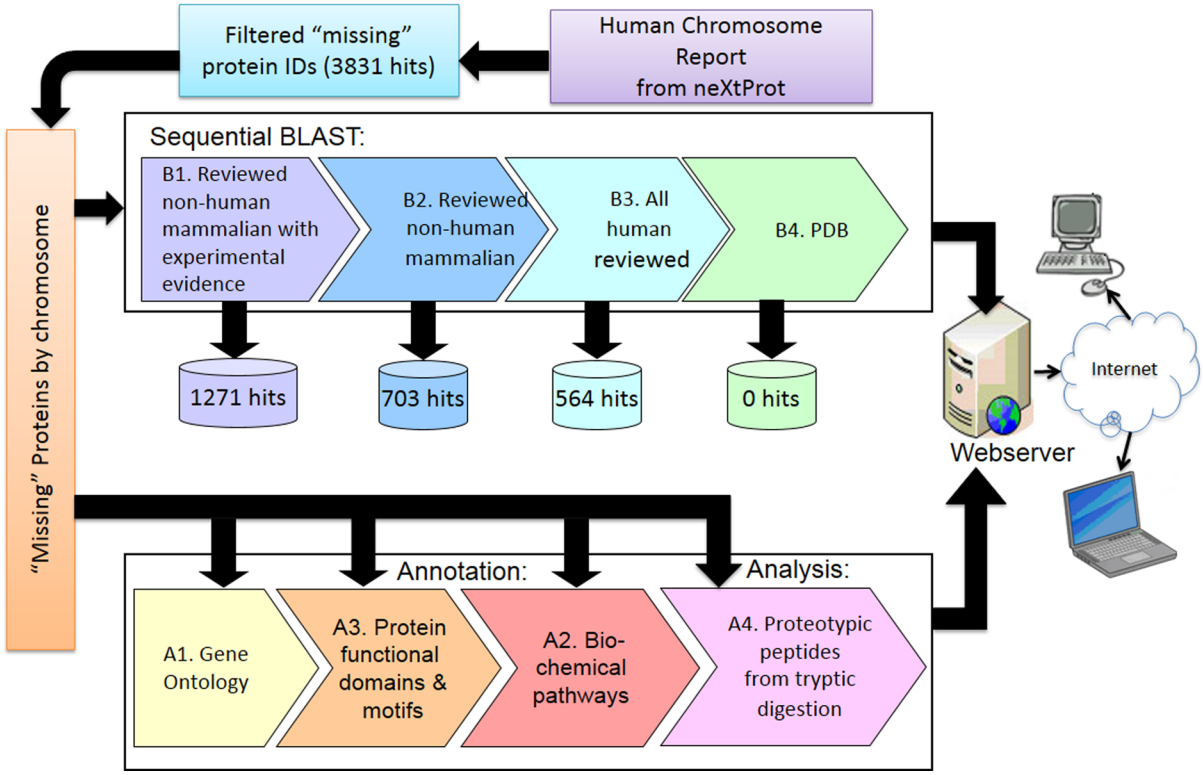


Figure 1. Top-level architecture of the pipeline for annotating human "missing" proteins. Proteins were passed through a series of databases to determine homology (sequential BLAST) as well as annotation databases based on GO, protein functional domains, motifs, and biochemical pathways.

tolerance of 5 ppm; Hopp and Woods[34] hydrophobicity mode) to computationally digest the "missing proteins" with trypsin, to identify proteotypic peptide sequences. Input sequences were validated, and duplicates were removed. These peptide sequences were then matched against ENCODE proteogenomic data,[20] generated on the basis of peptide spectrum scoring system[23] using the Peppy software.[22]

5. Protannotator Bioinformatics Pipeline

We have developed a semiautomated pipeline, called Protannotator, for the "missing" human proteome annotation based on the workflow reported in our previous hChr7 study.[9] All programs and tools used in this study were installed on a Linux cluster running on Ubuntu server operating system. The data are served using the Apache webserver with a PHP front end. The different components of the workflow system are linked using Perl, Python, and bash shell scripts into a workflow. The top-level architecture of the pipeline is illustrated in Figure 1.

In the first step, Protannotator extracted the proteome details of each chromosome from neXtProt, available as chromosome reports. Human proteins were then sorted based on the availability of protein evidence. The system then identified proteins characterized as "missing" proteins based on accession numbers provided for protein evidence level 2−4 (consistent with the recent C-HPP standard metrics table) and extracted their sequences in FASTA format from UniProt using Linux's wget utility.

The protein sequence files thus extracted were processed by the Database Similarity Search (DSS) module of the pipeline. DSS uses a series of sequential BLAST searches to identify high-quality matches. The "missing" proteins are first searched against reviewed nonhuman mammalian proteins with experimental evidence, with the unmatched sequences then searched against nonhuman reviewed mammalian proteins. The unmatched sequences from the second search are then searched against all human reviewed proteins and finally PDB proteins. "Missing" proteins were also searched for sequence similarity against all mammalian proteins and all nonmammalian proteins with proteomic evidence database, for verification.

The Protannotator system then employs InterProScan to characterize "missing" proteins with high-quality annotations based on protein functional domains and motifs along with GO terms. Pathway mapping was then carried out using KOBAS. InterProScan results were processed using IPRStats[35] for compiling statistics from InterProScan results as well as visualization of the output information.

All annotation information was then uploaded to a static webpage for the scientific community to view or download, by chromosome, permitting different C-HPP research groups across the globe to search the information on "missing" proteins for their respective chromosomes.

■ RESULTS AND DISCUSSION

All 20 128 human proteins were sorted based on the availability of protein evidence. 3831 proteins (~19%) were identified as "missing" based on protein evidence level 2−4 (consistent with the recent C-HPP standard metrics table and available as Supporting Information: Table S1). The number of "missing" proteins across the human proteome is steadily decreasing due to the large-scale proteomic effort across the globe. In our previous study of hChr7, 170 proteins were reported as "missing" as compared with 186 "missing" proteins in the current study.

dx.doi.org/10.1021/pr400794x | J. Proteome Res. 2014, 13, 76−83

1. Sequential-BLAST Similarity Search

The first sequential-BLAST run, against reviewed nonhuman mammalian protein sequences with proteomic evidence, resulted in 1271 "missing" proteins (33.2%) with significant matches, with >50% identity and *E* values of 0 to 1×10^{-05} (available from Supporting Information: Table S1). All top hits were selected, and matches with sequence identity ≥50% are considered as significant for this study. The remaining 2560 proteins were then searched against nonhuman reviewed mammalian protein sequences, using BLAST, with significant results for 703 sequences (18.4% of the 3831 "missing" proteins, available from Supporting Information: Table S2). For these matches, *E* values ranged from 0 to 2.00×10^{-6}. The third BLAST search against reviewed human proteins reported matches for 564 sequences (14.7% of the 3831 "missing" proteins, Supporting Information: Table S3), with *E* values ranging from 0 to 2.00×10^{-6}. In the final round of BLAST search, the remaining 1857 sequences were searched against PDB to check for similarity against sequences of known protein structures, with no match identified. Mapping all the "missing" proteins to the validation databases (results not shown) comprising all mammalian proteins and all nonmammalian proteins with experimental protein evidence also yielded a null result. The results from the sequential BLAST searches are shown in Table 1. Compared with the 127 "missing" proteins annotated using the sequential BLAST strategy in our previous hChr7,[9] 134 have been annotated in the current study, possibly as a consequence of the larger BLAST search databases in the current analysis and also due to the increased number of "missing" proteins, according to C-HPP standard metrics table.

Table 1. Sequential BLAST Matches for Human "Missing" Proteins

chromosome	number of missing proteins	reviewed mammalian proteins with experimental evidence	reviewed mammalian proteins	reviewed human proteins
Chr1	412	79	89	55
Chr2	190	55	51	25
Chr3	179	50	43	30
Chr4	130	53	28	10
Chr5	160	69	28	14
Chr6	176	52	31	28
Chr7	186	78	35	21
Chr8	108	36	26	22
Chr9	162	45	35	32
Chr10	152	53	21	41
Chr11	354	89	47	25
Chr12	169	60	35	17
Chr13[a]	53	17	12	13
Chr14	105	22	17	10
Chr15	111	44	22	11
Chr16	139	49	24	31
Chr17	191	50	46	35
Chr18	44	19	6	8
Chr19	369	230	30	25
Chr20	96	24	27	23
Chr21	59	9	8	27
Chr22	87	29	14	20
ChrX	182	52	28	40
ChrY[a]	17	7	0	1

[a]neXtProt Chr13 has one more protein and ChrY one less protein that the C-HPP standard metrics table.

The organism-wise distribution of the first two rounds of BLAST matches is shown in Figure 2, with the largest number of homologues in mouse, followed by rat and cow. Primates are not well-represented in this study, unlike in our previous report on hChr7.[9]

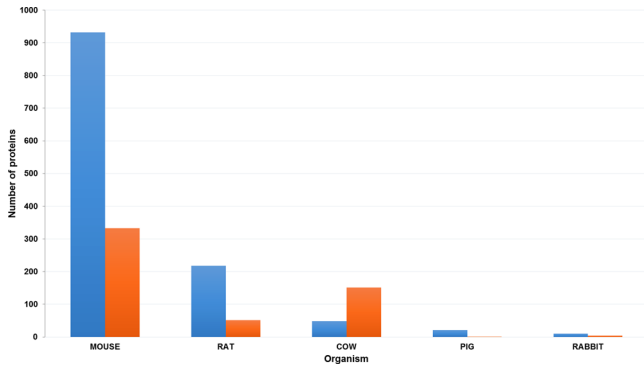


Figure 2. Significant BLAST hits grouped by organism. Blue bars represent the outcome of the first round of sequential BLAST against reviewed nonhuman mammalian proteins with experimental evidence, while the red bars represent the second round of BLAST against reviewed nonhuman mammalian proteins.

2. Functional Annotation

Functional annotation was carried out for all 3831 "missing" proteins, unlike the novel "missing" proteins alone in our previous study of hChr7.[9] InterProScan annotated 1945 "missing" proteins with GO annotation results in mapping of missing proteins to 2269 biological process (BP), 2059 cellular component (CC), and 3731 molecular function (MF) terms. Several GO terms such as *protein binding* and *membrane* reported in recent hChr7-centric proteomic analysis of human colon carcinoma cell lines[36] were also found among the hChr7 "missing" proteins. These 1945 "missing" proteins also mapped to 3019 domains, 4783 families, 162 repeats, 82 conserved sites, 9 binding sites, and 4 active sites. Recently published annotations of male specific chromosome Y proteins[37] were also reflected in chromosome Y "missing" protein mapping. DAZ proteins bind RNA in germ cells and are involved in primordial germ cell population maintenance.[38] The top 15 InterPro codes identified for the human "missing" proteins are shown in Table 2.

"Missing" proteins were also mapped to KEGG biochemical pathways using KOBAS, with 642 proteins annotated with pathway information. *Olfactory receptor* (IPR000725), the second InterPro hit, was listed as the top hit in the KEGG pathway mapping (*Olfactory transduction*). Out of 366 proteins mapped to the *Olfactory receptor family*, 360 were also mapped to the *G-protein-coupled receptor family* (IPR000276). Olfactory receptors were also reported recently in p13.2 and p13.3 regions of chromosome 17.[10] These receptors are associated with the biological process of *G-protein-coupled receptor signaling pathway* and the molecular function of *olfactory receptor activity*. G-protein-coupled receptor signaling pathway and olfactory receptor activity have been reported in genes clustered largely in a localized domain of chromosome 11.[39] *Zinc finger domains* (IPR013087, IPR001841, and IPR007087) comprise another important protein domain, in which zinc plays a structural role for the stability of the small domain. These protein domains are structurally diverse and are present among proteins responsible for a broad range of cellular functions, such as replication and

dx.doi.org/10.1021/pr400794x | J. Proteome Res. 2014, 13, 76−83

Table 2. Top 15 InterProScan Hits for Human "Missing" Proteins

InterPro code	description	number of missing proteins mapped	chromosome(s) mapped
IPR000276	G protein-coupled receptor, rhodopsin-like (family)	445	Chr10
IPR000725	olfactory receptor (family)	366	Chr10
IPR013087	zinc finger C2H2-type/integrase DNA-binding domain	129	Chr6
IPR001909	Krueppel-associated box (domain)	80	Chr6
IPR009057	homeodomain-like (domain)	68	Chr3
IPR001356	homeobox domain	62	Chr3
IPR002110	ankyrin repeat (repeat)	51	Chr8, Chr22
IPR007087	zinc finger, C2H2 (domain)	48	Chr7
IPR002126	cadherin (domain)	45	Chr4
IPR015919	cadherin-like (domain)	45	Chr4
IPR002494	high sulfur keratin associated protein (family)	32	Chr16
IPR001841	zinc finger, RING-type (domain)	31	Chr10
IPR015943	WD40/YVTN repeat-like-containing domain	29	Chr19
IPR011598	Myc-type, basic helix−loop−helix (bHLH) domain	29	Chr19
IPR011992	EF-hand domain pair (domain)	25	Chr19

repair, transcription and translation, metabolism and signaling, cell proliferation, and apoptosis.[40] The *homeobox* domain noted in our results (Table 2) was also recently reported in q21.32 region of chromosome 17.[10]

The sensory system was the main category of KEGG biochemical pathways reports for the "missing" proteins, with 390 proteins mapped to *Olfactory transduction* (372) and *Taste transduction* (18) pathways. Another 13 human "missing" proteins were mapped to pathways involved in Huntington's disease (HD), a neurodegenerative genetic disorder. This result indicates their involvement in human diseases, which needs further proteomic investigation. The top 10 KEGG pathway mappings are shown in Table 3. Recently, chromosome 19

Table 3. Top Ten KEGG Pathways for Human "Missing" Proteins

pathway description	no. of proteins
olfactory transduction	372
neuroactive ligand−receptor interaction	70
metabolic pathways	66
taste transduction	18
GABAergic synapse	17
glutamatergic synapse	16
calcium signaling pathway	16
natural killer cell mediated cytotoxicity	14
retrograde endocannabinoid signaling	14
antigen processing and presentation	13
Huntington's disease	13

genes/proteins have been related to 80 human diseases.[41] We mapped missing proteins from chromosome 19 to Alzheimer's, Parkinson's, and Huntington's diseases. The details of InterProScan and KEGG mapping are documented in the Supporting Information: Tables S4 and S5.

The functional annotation of the 3831 "missing" proteins is summarized in Figure 3, with 608 (15.9%) proteins having GO, InterPro domains as well as biochemical pathway annotations, 1337 (34.9%) proteins have InterPro domains and GO annotations alone, while 642 (16.8%) proteins have only KEGG biochemical pathway annotations. 1244 (32.5%) proteins could not be assigned any functional annotation with the currently available biological knowledge and may be considered novel.

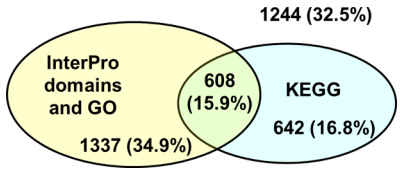


Figure 3. Summary of functional annotations for the "missing" proteins. Gene ontology (GO), functional domains/motif (InterPro) annotations were obtained for 1945 (50.8%) proteins. KEGG pathways (KEGG) annotations were obtained for 1250 (32.6%) proteins. 608 proteins had GO, Interpro domains, and KEGG pathway annotations. 1244 proteins remain unannotated.

3. *In Silico* Tryptic Digestion and ENCODE Proteogenomic Data

The Protein Digestion Simulator[28] was used to generate *in silico* proteotypic peptides, with trypsin selected as the proteolytic enzyme. Monoisotopic masses, pI, and hydrophobicity values for the tryptic peptides were computed (results not shown). These digested peptides were matched against the high-quality proteogenomic peptide data (58 601 records, based on the criteria described elsewhere[23] using the Peppy software[22]) from the ENCODE project[21] for proteomic evidence for the entire set of "missing" proteins. We found 245 peptides that matched the ENCODE data, with 1−44 peptides per protein. We have used the criteria of at least one or more peptide matching accurately (i.e., 100% identity) to the proteogenomic peptides, as we are matching protein sequences, to emulate the false positive discovery rate of Risk et al.,[22] who have set the threshold at >1 peptide matching to six-frame translations of genomic DNA. The peptides provide proteomic evidence of 107 "missing" proteins (with at least one peptide) for review and integration into the neXtProt chromosome summary lists. These peptides were found in 571 locations, with 316 in the positive orientation and 255 in the reverse orientation, that is, coded by the complementary strand. The genomic locations for two proteins (NX_Q9Y5G0 and NX_Q9Y5G1) on hChr5 could not be determined from neXtProt. The mapping results have been summarized in Table 4, and details of the mapping as well as the mapping regions are documented in Supporting Information: Table S6.

We have validated the ENCODE data mapping with neXtProt assigned genomic coordinates for each protein. Of the 571 locations, 202 matched the genomic coordinates

dx.doi.org/10.1021/pr400794x | J. Proteome Res. 2014, 13, 76−83

Table 4. Summary of ENCODE Data Mapping of Peptides from Human "Missing" Proteins

chromosome	number of proteins	number of peptides	positive	negative	number of peptides on both strands	total matched with neXtProt coding region	total unmatched with nextProt coding region
Chr1	7	17	17	18	2	17	18
Chr2	7	27	66	57	21	27	96
Chr3	2	2	1	1	0	2	0
Chr4	3	3	2	1	0	3	0
Chr5[a]	17	29	20	7	0	27	0
Chr6	9	44	72	55	24	3	124
Chr7	9	16	15	17	7	16	16
Chr8	1	1	1	1	1	1	1
Chr9	6	13	15	26	8	13	28
Chr10	5	8	21	3	2	8	16
Chr11	3	4	3	2	1	4	1
Chr12	5	7	4	5	0	7	2
Chr13	3	4	1	3	0	4	0
Chr14	1	1	1	0	0	1	0
Chr15	1	2	2	0	0	2	0
Chr16	3	27	50	27	26	27	50
Chr17	5	8	11	8	4	10	9
Chr18	1	1	0	1	0	1	0
Chr19	6	11	4	10	1	12	2
Chr20	1	1	1	0	0	1	0
Chr21	0	0	0	0	0	0	0
Chr22	3	4	0	1	0	1	0
ChrX	8	14	8	12	1	14	6
ChrY	1	1	1	0	0	1	0
Total	107	245	316	255	98	202	369

[a]For NX_Q9Y5G0 and NX_Q9Y5G1, the genomic location on chromosome 5 is not available in the neXtProt database, although ENCODE proteogenomic data mapped to these proteins.

assigned by neXtProt for 107 proteins, as two proteins (NX_Q9Y5G0, NX_Q9Y5G1) on Chr5 without genomic coordinates in neXtProt were excluded. 98 peptides were found in both orientations, that is, the coding regions covering one ENCODE peptide as well as at least one reverse peptide, requiring further experimental validation. These peptides can be used for the synthesis of antibodies and for future in vitro studies that could lead to proteomic identification of the proteins.

4. Membrane Proteomics "Missing" Protein List Comparison

We have compared the 3831 "missing" proteins with the 851 "missing" proteins identified by Muraoka et al.[24] and found that a total of 17 proteins have been provided proteomic evidence from this study. The chromosome-wise results are presented in Table 5.

Furthermore, we have summarized the proteomic evidence from the ENCODE project[21] and the membrane proteomic study of Muraoka et al.[24] (detailed in Supporting Information: Table S7). Two "missing" proteins (NX_P0CK97 and NX_Q9H0R5) have proteomic evidence from both experimental studies, with 105 "missing proteins uniquely supported by ENCODE data and 15 proteins by membrane proteomic data alone. In all, 122 (3.2%) "missing" proteins now have proteomic evidence

Table 5. Summary of Membrane Proteomic Evidence for Human "Missing" Proteins

chromosome	number of proteins	chromosome	number of proteins
Chr1	3	Chr11	2
Chr6	1	Chr12	3
Chr7	2	Chr16	1
Chr9	1	Chr17	1
Chr10	2	Chr19	1

CONCLUSIONS

We have compiled the chromosome-wise set proteins from the human proteome for 3831 "missing" proteins, as listed in the C-HPP standard metrics table, for *in silico* analysis and annotation. Using selected high-quality protein databases, similarity searches running BLAST sequentially identified homologues with experimental evidence for 33.2% of the "missing" proteins, with another 18.4% mapping to reviewed nonhuman mammalian proteins. As our study has used existing information to identify homologous proteins, further experimental work is required to confirm the existence of the proteins that are not identified by Protannotator, which is outside the scope of the work. However, with homologues identified from higher mammals, these proteins have a high probability of acquiring experimental evidence in the near future. Using a suite of bioinformatics tools, we have assigned putative biological functions in terms of GO and domain/motif signatures for 1945 (50.8%) and biochemical pathways for 1250 (32.6%) of the "missing" sequences. Despite the current level of biological knowledge in the databases, 1244 sequences (32.5%) remain unannotated by our sequential BLAST and computational annotation strategy.

By using a combination of computational tools, close to 50% of "missing" proteins in the human genome have been assigned putative biological functionality, providing valuable clues for

dx.doi.org/10.1021/pr400794x | J. Proteome Res. 2014, 13, 76−83

experimental validation assays. *In silico* tryptic digestion generated proteotypic peptides with which we were able to ascribe proteomic evidence for 107 (2.8%), thereby linking genomics and proteomics via bioinformatics. Additionally, proteomic evidence for another 15 "missing" proteins was provided by the recent membrane protein study of Muraoka et al.,[24] bringing the total of neXtProt "missing" proteins with proteomic evidence to 122. Our results, available freely through Protannotator, will benefit proteomic identification of the human "missing" proteome. The computational approach we have described is generic and can be used to annotate the proteome of any novel organism, such as the black Périgord truffle.[20] We plan to further automate the system (wherever possible) and provide updated information via the Protannotator web portal, to track proteomic or bioinformatics evidence for the unannotated set of "missing" proteins.

■ ASSOCIATED CONTENT

Supporting Information

Table S1: Significant BLAST hits against nonhuman reviewed mammalian proteins with experimental evidence. Table S2: Significant BLAST hits against nonhuman reviewed mammalian proteins. Table S3: Significant BLAST hits against human reviewed proteins. Table S4: Functional annotations of human missing proteins using InterProScan mapping. Table S5: KEGG pathway mapping of "missing" proteins. Table S6: ENCODE proteogenomic peptides mapping to the human "missing" proteins. Table S7: Summary of proteomic evidence for human "missing" proteins from ENCODE proteogenomic and membrane proteomic data. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: shoba.ranganathan@mq.edu.au. Phone: +612-9850-6262. Fax: +612-9850-8313.

Author Contributions

#M.T.I. and G.G. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Financial support was provided by a Macquarie University Research Excellence Scholarship (MQRES) to M.T.I. in the Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia

■ ABBREVIATIONS

C-HPP, Chromosome-centric Human Proteome Project; Chr, chromosome; EC, enzyme code; GO, gene ontology; KOBAS, KEGG Orthology-Based Annotation System; KEGG, Kyoto Encyclopedia of Genes and Genomes; NET, normalized elution time; NR, nonredundant; SCX, strong cation exchange

■ REFERENCES

(1) Stein, L. Genome annotation: from sequence to biology. *Nat. Rev. Genet.* **2001**, *2*, 493−503.

(2) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; Deutsch, E. W.; Domon, B.; Hancock, W.; He, F.; Hochstrasser, D.; Marko-Varga, G.; Salekdeh, G. H.; Sechi, S.; Snyder, M.; Srivastava, S.; Uhlen, M.; Wu, C. H.; Yamamoto, T.; Paik, Y. K.; Omenn, G. S. The human proteome project: current state and future direction. *Mol Cell Proteomics* **2011**, *10* (7), M111 009993.

(3) Hancock, W.; Omenn, G.; Legrain, P.; Paik, Y. K. Proteomics, human proteome project, and chromosomes. *J. Proteome Res.* **2011**, *10*, 210.

(4) Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30* (3), 221−3.

(5) Paik, Y. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Marko-Varga, G.; Aebersold, R.; Bairoch, A.; Yamamoto, T.; Legrain, P.; Lee, H. J.; Na, K.; Jeong, S. K.; He, F.; Binz, P. A.; Nishimura, T.; Keown, P.; Baker, M. S.; Yoo, J. S.; Garin, J.; Archakov, A.; Bergeron, J.; Salekdeh, G. H.; Hancock, W. S. Standard guidelines for the chromosome-centric human proteome project. *J. Proteome Res.* **2012**, *11* (4), 2005−13.

(6) Marko-Varga, G.; Omenn, G. S.; Paik, Y. K.; Hancock, W. S. A first step toward completion of a genome-wide characterization of the human proteome. *J. Proteome Res.* **2013**, *12* (1), 1−5.

(7) Baker, M. S. Building the 'practical' human proteome project - the next big thing in basic and clinical proteomics. *Curr. Opin. Mol. Ther.* **2009**, *11* (6), 600−2.

(8) Gaudet, P.; Argoud-Puy, G.; Cusin, I.; Duek, P.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Zahn-Zabal, M.; Zwahlen, C.; Bairoch, A.; Lane, L. neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* **2013**, *12* (1), 293−8.

(9) Ranganathan, S.; Khan, J. M.; Garg, G.; Baker, M. S. Functional Annotation of the Human Chromosome 7 ″Missing″ Proteins: A Bioinformatics Approach. *J. Proteome Res.* **2013**, 2504−10.

(10) Liu, S.; Im, H.; Bairoch, A.; Cristofanilli, M.; Chen, R.; Deutsch, E. W.; Dalton, S.; Fenyo, D.; Fanayan, S.; Gates, C.; Gaudet, P.; Hincapie, M.; Hanash, S.; Kim, H.; Jeong, S. K.; Lundberg, E.; Mias, G.; Menon, R.; Mu, Z.; Nice, E.; Paik, Y. K.; Uhlen, M.; Wells, L.; Wu, S. L.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Omenn, G. S.; Beavis, R. C.; Hancock, W. S. A chromosome-centric human proteome project (C-HPP) to characterize the sets of proteins encoded in chromosome 17. *J. Proteome Res.* **2013**, *12* (1), 45−57.

(11) Farrah, T.; Deutsch, E. W.; Hoopmann, M. R.; Hallows, J. L.; Sun, Z.; Huang, C. Y.; Moritz, R. L. The state of the human proteome in 2012 as viewed through PeptideAtlas. *J. Proteome Res.* **2013**, *12* (1), 162−71.

(12) UniProt Consortium.. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40* (Database issue), D71−75.

(13) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.

(14) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389−402.

(15) Nagaraj, S. H.; Gasser, R. B.; Ranganathan, S. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinf.* **2007**, *8* (1), 6−21.

(16) Ranganathan, S.; Menon, R.; Gasser, R. B. Advanced *in silico* analysis of expressed sequence tag (EST) data for parasitic nematodes of major socio-economic importance–fundamental insights toward biotechnological outcomes. *Biotechnol. Adv.* **2009**, *27* (4), 439−448.

(17) Garg, G.; Ranganathan, S. *In silico* secretome analysis approach using next generation sequencing transcriptomic data. *BMC Genomics* **2011**, *12* (Suppl 3), S14.

(18) Menon, R.; Garg, G.; Gasser, R. B.; Ranganathan, S. TranSeqAnnotator: large-scale analysis of transcriptomic data. *BMC Bioinf.* **2012**, *13* (Suppl17), S24.

(19) Garg, G.; Ranganathan, S. High-throughput functional annotation and data mining of fungal genomes to identify therapeutic targets. In *Laboratory Protocols in Fungal Biology: Current Methods in Fungal Biology*, Gupta, V. K.; Tuohy, M. G.; Ayyachamy, M.; Turner, K. M.; O'Donovan, A., Eds.; Springer: New York, 2013.
(20) Islam, M. T.; Mohamedali, A.; Garg, G.; Khan, J. M.; Gorse, A. D.; Parsons, J.; Marshall, P.; Ranganathan, S.; Baker, M. S. Unlocking the puzzling biology of the black Périgord truffle *Tuber melanosporum*. *J. Proteome Res.* **2013**, *12* (12), 5349−5356.
(21) Khatun, J.; Yu, Y.; Wrobel, J. A.; Risk, B. A.; Gunawardena, H. P.; Secrest, A.; Spitzer, W. J.; Xie, L.; Wang, L.; Chen, X.; Giddings, M. C. Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. *BMC Genomics* **2013**, *14*, 141.
(22) Risk, B. A.; Spitzer, W. J.; Giddings, M. C. Peppy: proteogenomic search software. *J. Proteome Res.* **2013**, *12* (6), 3019−25.
(23) Risk, B. A.; Edwards, N. J.; Giddings, M. C. A Peptide-Spectrum Scoring System Based on Ion Alignment, Intensity, and Pair Probabilities. *J. Proteome Res.* **2013**, *12* (9), 4240−4247.
(24) Muraoka, S.; Kume, H.; Adachi, J.; Shiromizu, T.; Watanabe, S.; Masuda, T.; Ishihama, Y.; Tomonaga, T. In-depth membrane proteomic study of breast cancer tissues for the generation of a chromosome-based protein list. *J. Proteome Res.* **2013**, *12* (1), 208−13.
(25) Protein Data Bank (PDB): http://www.rcsb.org/pdb/.
(26) Boekhorst, J.; Snel, B. Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. *BMC Bioinf.* **2007**, *8*, 356.
(27) Sander, C.; Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **1991**, *9* (1), 56−68.
(28) Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. InterProScan: protein domains identifier. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W116−20.
(29) Hunter, S.; Jones, P.; Mitchell, A.; Apweiler, R.; Attwood, T. K.; Bateman, A.; Bernard, T.; Binns, D.; Bork, P.; Burge, S.; de Castro, E.; Coggill, P.; Corbett, M.; Das, U.; Daugherty, L.; Duquenne, L.; Finn, R. D.; Fraser, M.; Gough, J.; Haft, D.; Hulo, N.; Kahn, D.; Kelly, E.; Letunic, I.; Lonsdale, D.; Lopez, R.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; McMenamin, C.; Mi, H.; Mutowo-Muellenet, P.; Mulder, N.; Natale, D.; Orengo, C.; Pesseat, S.; Punta, M.; Quinn, A. F.; Rivoire, C.; Sangrador-Vegas, A.; Selengut, J. D.; Sigrist, C. J. A.; Scheremetjew, M.; Tate, J.; Thimmajanarthanan, M.; Thomas, P. D.; Wu, C. H.; Yeats, C.; Yong, S.-Y. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **2012**, *40* (D1), D306−D312.
(30) Xie, C.; Mao, X.; Huang, J.; Ding, Y.; Wu, J.; Dong, S.; Kong, L.; Gao, G.; Li, C. Y.; Wei, L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* **2011**, *39* (Web Server issue), W316−22.
(31) Young, N. D.; Jex, A. R.; Li, B.; Liu, S.; Yang, L.; Xiong, Z.; Li, Y.; Cantacessi, C.; Hall, R. S.; Xu, X.; Chen, F.; Wu, X.; Zerlotini, A.; Oliveira, G.; Hofmann, A.; Zhang, G.; Fang, X.; Kang, Y.; Campbell, B. E.; Loukas, A.; Ranganathan, S.; Rollinson, D.; Rinaldi, G.; Brindley, P. J.; Yang, H.; Wang, J.; Wang, J.; Gasser, R. B. Whole-genome sequence of Schistosoma haematobium. *Nat. Genet.* **2012**, *44* (2), 221−5.
(32) Jex, A. R.; Liu, S.; Li, B.; Young, N. D.; Hall, R. S.; Li, Y.; Yang, L.; Zeng, N.; Xu, X.; Xiong, Z.; Chen, F.; Wu, X.; Zhang, G.; Fang, X.; Kang, Y.; Anderson, G. A.; Harris, T. W.; Campbell, B. E.; Vlaminck, J.; Wang, T.; Cantacessi, C.; Schwarz, E. M.; Ranganathan, S.; Geldhof, P.; Nejsum, P.; Sternberg, P. W.; Yang, H.; Wang, J.; Wang, J.; Gasser, R. B. Ascaris suum draft genome. *Nature* **2011**, *479* (7374), 529−33.
(33) Protein Digestion Simulator: http://omics.pnl.gov/.
(34) Hopp, T. P.; Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 3824−8.
(35) Kelly, R.; Vincent, D.; Friedberg, I. IPRStats: visualization of the functional potential of an InterProScan run. *BMC Bioinf.* **2010**, *11* (Suppl 12), S13.
(36) Fanayan, S.; Smith, J. T.; Sethi, M. K.; Cantor, D.; Goode, R.; Simpson, R. J.; Baker, M. S.; Hancock, W. S.; Nice, E. Chromosome 7-centric analysis of proteomics data from a panel of human colon carcinoma cell lines. *J. Proteome Res.* **2013**, *12* (1), 89−96.
(37) Jangravi, Z.; Alikhani, M.; Arefnezhad, B.; Sharifi Tabar, M.; Taleahmad, S.; Karamzadeh, R.; Jadaliha, M.; Mousavi, S. A.; Ahmadi Rastegar, D.; Parsamatin, P.; Vakilian, H.; Mirshahvaladi, S.; Sabbaghian, M.; Mohseni Meybodi, A.; Mirzaei, M.; Shahhoseini, M.; Ebrahimi, M.; Piryaei, A.; Moosavi-Movahedi, A. A.; Haynes, P. A.; Goodchild, A. K.; Nasr-Esfahani, M. H.; Jabbari, E.; Baharvand, H.; Sedighi Gilani, M. A.; Gourabi, H.; Salekdeh, G. H. A fresh look at the male-specific region of the human Y chromosome. *J. Proteome Res.* **2013**, *12* (1), 6−22.
(38) Reynolds, N.; Cooke, H. Role of the DAZ genes in male fertility. *Reprod. Biomed. Online* **2005**, *10*, 72.
(39) Kwon, K. H.; Kim, J. Y.; Kim, S. Y.; Min, H. K.; Lee, H. J.; Ji, I. J.; Kang, T.; Park, G. W.; An, H. J.; Lee, B.; Ravid, R.; Ferrer, I.; Chung, C. K.; Paik, Y. K.; Hancock, W. S.; Park, Y. M.; Yoo, J. S. Chromosome 11-centric human proteome analysis of human brain hippocampus tissue. *J. Proteome Res.* **2013**, *12* (1), 97−105.
(40) Krishna, S. S.; Majumdar, I.; Grishin, N. V. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Res.* **2003**, Jan 15; *31*(2):532-50.
(41) Nilsson, C. L.; Berven, F.; Selheim, F.; Liu, H.; Moskal, J. R.; Kroes, R. A.; Sulman, E. P.; Conrad, C. A.; Lang, F. F.; Andren, P. E.; Nilsson, A.; Carlsohn, E.; Lilja, H.; Malm, J.; Fenyo, D.; Subramaniyam, D.; Wang, X.; Gonzales-Gonzales, M.; Dasilva, N.; Diez, P.; Fuentes, M.; Vegvari, A.; Sjodin, K.; Welinder, C.; Laurell, T.; Fehniger, T. E.; Lindberg, H.; Rezeli, M.; Edula, G.; Hober, S.; Marko-Varga, G. Chromosome 19 annotations with disease speciation: a first report from the Global Research Consortium. *J. Proteome Res.* **2013**, *12* (1), 135−50.