First Proteomic Exploration of Protein-Encoding Genes on Chromosome 1 in Human Liver, Stomach, and Colon

Songfeng Wu,†‡‡ O Ning Li,†‡‡ Jie Ma,†‡‡ Hualin Shen,§‡‡ Dahai Jiang,∥‡‡ Cheng Chang,‡‡ Chengu Zhang,†‡ Liwei Li,†‡ Hongxing Zhang,†‡ JIng Jiang,†‡ Zhongwei Xu,†‡† Lingyan Ping,†‡ Tao Chen,†‡ Wei Zhang,†‡† Tao Zhang,‡‡ Xiaohua Xing,†‡‡ Tailong Yi,‡‡ Yanchang Li,‡‡‡ Fengxu Fan,†‡‡ Xiaqian Li,†‡ Fan Zhong,§ Quanhui Wang,∥‡‡ Yang Zhang,§‡‡ Bo Wen,∥ Guoquan Yan,§ Liang Lin,∥ Jun Yao,§ Zhitong Lin,∥‡‡ Feifei Wu,§ Liqi Xie,§ Hongxiu Yu,§ Mingqi Liu,§ Haojie Lu,§ Hong Mu,∥‡‡ Dong Li,†‡‡ Weimin Zhu,†‡ Bei Zhen,†‡† Xiaohong Qian,†‡‡ Jun Qin,‖‡‡ Siqi Liu,‖‡‡‡ Pengyuan Yang,‖‡‡ Yunping Zhu,‡‡‡ †‡ Ping Xu,‡‡‡ †‡ and Fuchu He,‡‡‡ †‡,§§

†State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China
‡National Engineering Research Center for Protein Drugs, Beijing 102206, China
§Institutes of Biomedical Sciences and Department of Chemistry, 130 DongAn Road, Fudan University, Shanghai 200032, China
§BGI-Shenzhen, ShenZhen 518083, China
∥Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China
‖State Key Laboratory of Molecular Oncology, Cancer Institute & Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100021, China
#Taicang Institute for Life Sciences Information, Taicang 215400, China

ABSTRACT: The launch of the Chromosome-Centric Human Proteome Project provides an opportunity to gain insight into the human proteome. The Chinese Human Chromosome Proteome Consortium has initiated proteomic exploration of protein-encoding genes on human chromosomes 1, 8, and 20. Collaboration within the consortium has generated a comprehensive proteome data set using normal and carcinomatous tissues from human liver, stomach, and colon and 13 cell lines originating in these organs. We identified 12,101 proteins (59.8% coverage against Swiss-Prot human entries) with a protein false discovery rate of less than 1%. On chromosome 1, 1,252 proteins mapping to 1,227 genes, representing 60.9% of human proteome, chromosome, C-HPP, coexpression, data visualization

KEYWORDS: proteome, chromosome, C-HPP, coexpression, data visualization

INTRODUCTION

The Human Genome Project provides fundamental sequence reference resources for large-scale omics research.1,2 The Ensembl human genome database (http://www.ensembl.org/Homo_sapiens/, version 68.37) contains 21,224 known genes, 568 novel genes, 266 putative genes, 14,427 pseudogenes, and 15,952 RNA genes. The field of proteomics benefits not only from improvements in chromatography and mass spectrometry but also from the availability of gene sequences and annotations
derived from genomic research. For example, mass spectra can be matched to peptides more efficiently using protein sequence databases. The Human Plasma Proteome Project, Human Liver Proteome Project (HLPP), Human Brain Proteome Project, and others have used genomic data and generated the corresponding proteome data sets using different samples/tissues. With the advances in rapid, high-resolution protein sequencing technology, protein-coding genes can be identified and quantified more efficiently and precisely. It is time that proteomics "repays" genomics and contributes to the annotation of protein-coding genes.

Over the past few years, the Chromosome-Centric Human Proteome Project (C-HPP), which aims to define the full set of protein-coding genes on each chromosome, has been launched. Twenty-four research groups worldwide have joined the project and contributed to the annotation of different chromosomes, based on research interest in a particular disease or gene cluster.

Digestive cancers, such as hepatocellular carcinoma, gastric cancer, and colorectal cancer, are among the most common and severe diseases in China. To better understand the physiology and pathology of liver diseases, Chinese proteomics groups have focused their efforts, in the past decade, on the HLPP, including generation of proteome expression profiles of the human fetal liver and Chinese adult liver. The Chinese Human Proteome Organization (CNHUPO) engaged in extensive collaboration with the C-HPP and initiated relevant research in China while continuing efforts to advance proteomics projects for chromosomes 1, 8, and 20. The Chinese Human Chromosome Proteome Consortium (CCPC) is coordinated by CNHUPO and includes researchers from Beijing Proteome Research Center, Fudan University, Beijing Genome Institute, and Jinan University. CCPC has established several research groups that are leading proteomics research in China. Like its foreign counterparts, CCPC has come to realize that there is currently no unique approach to a chromosome proteomics project. Under the guidance of C-HPP, we must become creative with experimental design and data analysis.

CCPC has clearly outlined the primary goals of this project. First, C-HPP is not an ordinary project targeting proteomic analysis of a biological sample but is fueled by the conviction that all of the work should be focused on chromosomal features. Second, a major concern of C-HPP is the relationship between the protein-encoding genes on chromosomes and their expression status. A comprehensive analysis of this relationship should take into consideration each aspect of gene expression (e.g., transcription and translation) using qualitative and quantitative information. Third, during C-HPP implementation, we should make full use of the accumulated knowledge in chromosome research, especially knowledge of disease-related chromosomes. In this way, C-HPP’s findings will not be confined to theory but instead will extend to biological or medical applications. For this reason, CCPC used three gastrointestinal tumors and representative cell lines as the starting materials of chromosome proteomic analysis.

Here, we report the results of the first collaboration in CCPC’s effort to construct proteomic expression profiles of samples from human liver, colon, and stomach. Eighteen sets of samples were used, including normal and cancerous tissues and cell lines (eight liver, two colon, and three gastric cell lines). Finally, 12,101 high-confidence proteins were identified with a protein false discovery rate (FDR) less than 1%. Using this high-quality proteomic data set (the Chinese Chromosome Proteome Data set [CCPD]), Beijing Proteome Research Center began to make contributions to proteomic annotation of chromosome 1. The same data set was also used to annotate chromosomes 8 and 20, and the current state of these chromosomes will be described in this volume.

Chromosome 1 is the largest of the human chromosomes, containing seven bands and 249 M base pairs (bp), and is separated into short and long arms. Many disease-related genes are found on human chromosome 1. To date, approximately 460 genetic phenotypes such as Parkinson’s disease, Alzheimer’s disease, autosomal recessive deafness, recurrent pregnancy loss, and Duffy blood group system have been traced to chromosome 1 (see http://www.ncbi.nlm.nih.gov/omim). Chromosome 1 also plays an important role in cancers of the liver, stomach, and colon. For example, because of genomic alterations, at least 785 genes on chromosome 1 are part of liver cancer-related signatures, according to published microarray and proteomic studies. It has been reported that most are located in band 1q21.1-q32.1 and 1q32.1-q44.1. Seven other regions, including 1p36.33-36.31 and 1q31.1-32.1, might harbor putative tumor suppressor genes related to gastric or colon cancers.

In this work, we focused on the generation and systematic analysis of the CCPD. The protein-encoding genes on chromosome 1 were fully annotated using multiple public databases; some properties were calculated on the basis of protein sequences. Unidentified proteins were found to be those with unusual physicochemical properties or low abundance. Combined analysis of the transcriptome and proteome allowed us to deduce the unidentified blocks (a "block" refers to a particular region on chromosome 1 in which proteins of adjacent genes are all qualitatively identified or unidentified) and coexpression clusters (a “cluster” refers to a particular region on chromosome 1 in which adjacent genes are quantitatively coexpressed) on this chromosome. Circular and tabular visualization were used to provide a global view of the data and results. Detailed data can be found in the Chinese Chromosome-Centric Human Proteome Database (C-HPD, http://proteomeview.hupo.org.cn/ chromosome/) and will be available in Chromosome-Assembled human Proteome browsER (CAPER, http://61.50.138.124/index.php) in the future.

As C-HPP investigators learned in Boston, it is essential that we all utilize the agreed-upon standard databases and criteria for the C-HPP chromosome baselines: Ensembl for number of genes (v68, release Jul 2012); neXtProt (release Sep 2012), GPMdb (release Jul 2012), and Peptide Atlas (1% FDR at protein level, release Sep 12) for number and list of proteins confidently identified from MS data sets; and Human Protein Atlas (HPA, v10.0) for antibody-based protein identifications. The baselines of genes/proteins for chromosome 1 from Ensembl, GPMDB, PeptideAtlas, ProteinAtlas, and neXtProt as well as their corresponding identifications in CCPD are shown in Table 1.

**MATERIALS AND METHODS**

**Sample Preparation**

Liver samples were described in HLPP. The eight liver cell lines MHCC97L (97L), MHCC97H (97H), HCCLM3 (LM3), HCCLM6 (LM6), SNU398, SNU449, SNU475, and Hep3B were cultured in Dulbecco’s modified Eagle medium (Gibco, USA) supplemented with 10% fetal bovine serum (Life Technologies) at 37 °C in a humidified atmosphere containing 5% CO₂. Cancerous and adjacent normal tissues of...
Table 1. Number of Genes/Proteins in the Five Databases and Their Corresponding Identified Results for Chromosome 1 “

<table>
<thead>
<tr>
<th>database</th>
<th>no. of genes</th>
<th>no. of proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensembl</td>
<td>2,043</td>
<td>1,455</td>
</tr>
<tr>
<td>GPMDB</td>
<td>1,166</td>
<td>1,441</td>
</tr>
<tr>
<td>PeptideAtlas</td>
<td>1,014</td>
<td>990</td>
</tr>
<tr>
<td>HPA</td>
<td>1,081</td>
<td></td>
</tr>
<tr>
<td>neXtProt</td>
<td>1,081</td>
<td></td>
</tr>
</tbody>
</table>

The five databases and their versions are Ensembl (v68, release Jul 2012), GPMDB (release Jul 2012), PeptideAtlas (release Sep 2012), ProteinAtlas (HPA, v 10.0), and neXtProt (release Sep 2012). The data set annotated as “Green” was used, with the threshold >20 observations and log(e)<-5. “The data set annotated as “protein evidence” was used.

colon and stomach were collected at Beijing Cancer Institute & Hospital and Sun Yat-Sen University Cancer Center under the supervision of the institutions’ ethical review boards. Two colon cell lines, SW480 and HCT116, and three gastric cell lines, AGS, BGC823, and Sgc-7901, were also used in this work. Tissues were lysed as described previously. The quality of the extracted proteins was inspected following SDS-PAGE and Coomassie blue staining. The concentration of extracted protein was determined using the BCA assay. Equal amounts of protein from 3–5 patients were pooled prior to proteomics analysis to reduce individual variation. Detailed information is available in Supporting Information and Methods.

Proteomics Analysis

The protein mixtures were either subjected to gel-assisted digestion or sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) separation followed by in-gel digestion with trypsin as described. The peptide mixtures from gel-assisted digestion were separated by high-pH reverse-phase chromatography. All peptide mixtures were analyzed by liquid chromatography tandem mass spectrometry (LC–MS/MS) in CCPC laboratories. The mass spectrometers used in this project were TripleTOF 5600 (AB SCIEX, Concord, ON), Orbitrap Q-Exactive or LTQ-Orbitrap Velos (Thermo Fisher Scientific). Details of protein digestion, extraction, peptide mixture prefractonation, and MS instrument settings can be found in the Supporting Information and Methods.

All raw files of mass spectra were converted into mzXML and MGF files using the mconvert module in the Trans-Proteomic Pipeline (TPP v4.5.2). The MS/MS peak lists were searched using the Mascot v2.3.2 local server against the database containing sequences of all human proteins from Swiss-Prot (20,231 proteins, 2012_07 release) and common contaminants (115 proteins, ftp://ftp.ebi.ac.uk/pub/databases/asn/). The target-decoy strategy was applied to maintain both peptide and protein FDRs at less than 1%. A normalized label-free quantification method based on the extracted ion chromatograms (XICs) was applied to all confidently identified peptides and proteins. Details of peptide/protein identification quality control methods and XIC-based protein quantification methods are available in Supporting Information and Methods.

Transcriptome Analysis

Transcriptome analysis was performed using the corresponding RNAs isolated from tissues and cell lines. RNA-Seq technology (Illumina HiSeq 2000) was used to generate the transcriptome data for normal and cancerous gastric tissues, gastric cell lines AGS and BGC823, normal and cancerous hepatic tissues, and normal and cancerous colorectal tissues. Affymetrix HG-U133 2.0 Plus oligonucleotide microarrays were used for five hepatic cell lines, MHCC97L, MHCC97H, HCCCLM, HCCLM6, and Hep3B. Microarray data for the other three hepatic cell lines (SNU398, SNU449, and SNU475) were downloaded from http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-37. Transcriptome data for colorectal cell lines were also generated by microarray.

Gene expression levels were determined from RNA-Seq data based on the RPKM (reads per kilobase per million) value. Quantitative microarray data were utilized for further study based on MAS 5.0 (Affymetrix Expression Console Software, http://www.affymetrix.com/). When different samples were compared, RNASeq and microarray raw intensity values were divided by the median value in each sample for normalization. Details of the qualitative and quantitative analysis of transcriptome data can be found in Supporting Information and Methods.

Bioinformatic Annotation of Identified Proteins

To analyze the expression patterns of the proteome and transcriptome, proteins in Swiss-Prot were systematically annotated by combining the information from the Swiss-Prot, Online Mendelian Inheritance in Man (OMIM; NCBI, confirmed Mendelian phenotype), Cancer Gene Census (CGC, Sanger Center; cancer gene information), ProteinAtlas, and Ensembl databases.

Evidence of protein expression, post-translational modifications (PTMs), and protein list were downloaded from Swiss-Prot directly. Antibody evidence was extracted from ProteinAtlas. Disease information was obtained from both OMIM and CGC. Gene position and chromosomal band information was downloaded from the Ensembl database (version 67.37) using BioMart (www.ensembl.org/biomart/martview). The gene symbols were utilized for mapping Ensembl to Swiss-Prot. For curation analysis, the Gene Ontology (GO) terms were downloaded from ftp.ensembl.org/pub/release-67/genbank/homo_sapiens/. The mass spectra evidence was obtained from the global proteome machine database (GPMDB), NIST spectra library (NIST SL) (http://peptide.nist.gov/), and Human PeptideAtlas.

Physicochemical properties and protein detectability were calculated on the basis of the protein sequences. The physicochemical properties of proteins, including molecular weight (MW), hydrophobicity (Hy), and isoelectric point (pl) were computed with ProPAS software. Peptide detectability (pepDetectability) was determined by an support vector machine-based tool, STEPP, and peptide detectability (proDetectability) was deduced using the following formula:

$$proDetectability = \log(\sum_{i=1}^{n} pepDetectability)$$

Here n is the number of all theoretical digested peptides in one protein.

Cluster and Enrichment Analysis

For cluster analysis of protein expression data, two-way hierarchical clustering of proteins was implemented using Cluster 3.0 and the default parameters coupled with centroid linkage.

The GO molecular function and biological process ontologies and pathway analysis was applied through the online tool DAVID coupled with MetaCore from GeneGo, Inc. A p-value of 0.001 was used as a cutoff to achieve the significant categorial annotation.
Identification of Gene Coexpression/Cofunction Clusters

The chromosome 1 gene coexpression clusters were calculated on the basis of the quantitative data from both the proteome and the transcriptome, following Lercher’s method.61 The cofunction clusters were identified according to the shared functions from GO terms.61 Tandem duplicates were defined by protein BLAST62 (word size=2). Protein pairs were considered tandem duplicates if the blast e-value was less than 0.2.63 A detailed description of the method can be found in Supporting Information and Methods.

Data Visualization and Database Construction

As in the gene display method of Paik et al.,13 tabular visualization of genes on chromosome 1 was carried out using conditional formatting of Excel spreadsheets and included the information on protein characterization and the indices of protein and mRNA abundance from this study for each identified gene. There are five levels in Supplementary Table S4. The top level is shown in green, which indicates the highest quality for protein characterization or top 25% protein/mRNA abundance, whereas yellow, red, and gray represent decreased quality or abundance. The dark box in the table indicates that the protein was not characterized or could not be identified in the proteome or transcriptome.

The heatmap of hierarchical clustering of the protein quantitation values, sample correlation coefficient matrix, and gene coexpression clusters of chromosome 1 were displayed with Java TreeView-1.1.6r2 (http://jtreeview.sourceforge.net/).

The global overview of chromosome annotations (from chromosome structure to protein evidence and modifications), detected abundance of protein and mRNA, and expression patterns (unidentified blocks and coexpression clusters) were layered in a circular layout for visualization using Circos.64

The database C-HPD was constructed using the MySQL relational database to store the data and JSP technology to implement the web interface. Currently, the database consists of three parts: basic information for chromosomes, protein and gene annotation for each chromosome, and transcriptive information for each chromosome.

RESULTS AND DISCUSSION

To identify as many proteins as possible for C-HPP projects, we designed the project in such a way as to use diverse samples of diseased tissue. Comprehensive proteome data sets, as well as the corresponding transcriptome data sets, were generated using both tissue and cell lines from three organs of the human digestive system [i.e., liver, colon, and stomach (Figure 1)]. The data were processed for assignment to the database entries with their normalized expression abundance. Strict quality controls were applied to each step of both the experiments and the data analysis process (Figure 1).

Qualitative Analysis of the Identified Proteins

Collaboration within the CCPC produced a large proteome data set, CCPD, containing 6.5 million spectra and 192,546 unique peptides for 12,101 proteins with a protein FDR of less than 1%, corresponding to 11,632 genes (Supplementary Tables S1 and S2). The high quality of protein identification was demonstrated by the large number of unique peptides (Supplementary Figure S1A) and high protein sequence coverage (Supplementary Figure S1B). There were 425 proteins with only one unique peptide in our final protein list after applying a 1% FDR cutoff. The high quality of the matches was also demonstrated by a higher Mascot score (Supplementary Figure S1C).

To compare tissue specificity and increase the coverage of identified proteins to all protein-coding genes in the human genome, three tissues (from the human liver, colon, and stomach) were used (Table 2). The results showed that use of samples from different organs increased the number of identified peptides and proteins, indicating that the combination of experimental results from different samples would be complementary and conducive to chromosome-based proteomics. The number of identified proteins appeared to increase more slowly than the number of peptides (Figure 2A), suggesting that the number of identified proteins tended to reach a plateau with the increase in numbers of tissues or cell lines.

We identified, with high confidence, 9,346 proteins in the liver, 9,506 in the colon, and 7,941 in the stomach. Among these, 6,184 (51.1%) proteins were overlapped by all three
data sets, suggesting a high degree of consistency and confidence of our proteomics platform and protein sequencing power (Figure 2B). Detailed information about the identified proteins and peptides can be found in Supplementary Tables S1 and S2.

Compared with the Swiss-Prot database, the general coverage of identified proteins was approximately 59.8%. The coverage rate for different chromosomes was similar, ranging from 66.6% for chromosome 2 to 41.6% for chromosome 21. For the three chromosomes that CCPC is devoted to, the coverage was 60.9% (chr 1), 58.9% (chr 8), and 58.4% (chr 20, Figure 2C).

**Contributions to the Public Protein/Mass Spectra Databases**

The publicly available protein databases enable the systematic annotation of protein entries from different evidence. For example, Swiss-Prot provides the protein evidence function, which assigns every protein to one of the following levels with decreasing confidence: protein, transcript, homology, prediction, and uncertain. Compared with the protein evidence annotated in Swiss-Prot, 10,264 of 12,101 proteins in our CCPD were annotated as genes with protein-level evidence in the database, accounting for 75.4% of the genes with protein-level evidence in Swiss-Prot. This result again confirmed the high confidence of our identified proteins. Furthermore, 1,837 identified proteins in our data set were annotated in Swiss-Prot as genes with evidence levels of transcript, homology, or even prediction or uncertain (Figure 2D). Therefore, our results provided new evidence for and insight into the functions of these genes.

The public proteomics databases are another valuable resource for protein level validation. As described in Materials and Methods, there are three major mass spectra databases (GPMDB, Human PeptideAtlas, and NIST_SL) that can be used as sources for comparison. Despite the smaller scale and volume of our proteomics data, we consider our contributions to

<table>
<thead>
<tr>
<th>tissue</th>
<th>sample</th>
<th>no. of identified spectra</th>
<th>no. of identified peptides</th>
<th>no. of identified proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>liver</td>
<td>normal tissue</td>
<td>789,201</td>
<td>72,627</td>
<td>7,092</td>
</tr>
<tr>
<td>97H</td>
<td>397,380</td>
<td>27,241</td>
<td>3,738</td>
<td></td>
</tr>
<tr>
<td>97L</td>
<td>289,831</td>
<td>26,809</td>
<td>3,568</td>
<td></td>
</tr>
<tr>
<td>LM3</td>
<td>223,065</td>
<td>15,728</td>
<td>3,014</td>
<td></td>
</tr>
<tr>
<td>LM6</td>
<td>278,896</td>
<td>26,181</td>
<td>3,736</td>
<td></td>
</tr>
<tr>
<td>Hep3B</td>
<td>220,809</td>
<td>25,941</td>
<td>4,011</td>
<td></td>
</tr>
<tr>
<td>SNU339</td>
<td>226,484</td>
<td>25,034</td>
<td>3,910</td>
<td></td>
</tr>
<tr>
<td>SNU449</td>
<td>273,160</td>
<td>24,768</td>
<td>3,994</td>
<td></td>
</tr>
<tr>
<td>SNU475</td>
<td>243,738</td>
<td>23,387</td>
<td>3,804</td>
<td></td>
</tr>
<tr>
<td>colon</td>
<td>normal tissue</td>
<td>562,787</td>
<td>87,955</td>
<td>7,749</td>
</tr>
<tr>
<td>SW480</td>
<td>540,271</td>
<td>36,130</td>
<td>4,429</td>
<td></td>
</tr>
<tr>
<td>HCT116</td>
<td>345,981</td>
<td>46,353</td>
<td>5,398</td>
<td></td>
</tr>
<tr>
<td>stomach</td>
<td>cancerous tissue</td>
<td>220,652</td>
<td>17,529</td>
<td>3,018</td>
</tr>
<tr>
<td>AGS</td>
<td>482,442</td>
<td>50,805</td>
<td>6,236</td>
<td></td>
</tr>
<tr>
<td>BGC823</td>
<td>447,780</td>
<td>51,689</td>
<td>6,130</td>
<td></td>
</tr>
<tr>
<td>Sgc-7901</td>
<td>219,224</td>
<td>31,002</td>
<td>3,734</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>6,554,084</td>
<td>192,546</td>
<td>12,101</td>
</tr>
</tbody>
</table>

**Table 2. Overview of Samples and Data Sets in Proteomic Analysis**
these public resources significant. As shown in Figure 2E, 79,581 unique peptide sequences in our CCPD were not present either in the NIST peptide list or that of Human PeptideAtlas. These peptides were identified with high confidence in our study because 10,741 (1,181 for chromosome 1) were identified by at least one high-accuracy MS/MS spectrum with a Mascot ion score greater than 50. At the protein level, we contributed 564 new identified proteins (58 for chromosome 1) with three or more unique peptides, which are not present in these public databases (Figure 2F). These results suggest the high coverage of our data set at the level of both the individual protein and whole proteome.

**Quantitative Analysis of the Identified Proteins**

The abundance distribution of identified proteins at the proteomic level shows that the proteins on our list span 7 orders of magnitude and that 90% of the proteins fall in the range of 3–4 orders of magnitude (Figure 3A). The Pearson’s correlation coefficient between different samples ranges from 0.21 to 0.96, with the gastric AGS cell line and colon cancer tissue having the lowest correlation coefficient and liver cell lines 97L and LM6 having the highest correlation coefficient (Figure 3B).

To take a global view of the proteomic expression pattern, two-way hierarchical clustering was applied to the samples and identified proteins using the protein concentration. The clustering of 17 samples showed that the organ diversity was more significant than the difference between cancer and normal tissues of the same organ (Figure 3C), which is consistent with the coefficient analysis described above (Figure 3B). As expected, the liver samples were clustered together and appeared as an independent branch. There were some crossovers between the stomach and colon, indicating that to some extent, these two organs might share some physiological functions (Figure 3C). Cluster analysis of different proteins revealed four blocks in which the protein expression levels varied dramatically between different organs (as indicated in Figure 3C). Further research should be carried out for the validation of these function biases.

---

**Figure 3.** Protein abundance distribution and cluster analyses of proteome expression data for liver, stomach, and colon samples. (A) Distribution of protein abundance. (B) Clustering of different samples. (C) Cluster analysis of proteomic data from 17 different samples. The analyses of gene function in different clusters indicate potentially different functions and coordinated coexpression in different tissues.
Protein Identiﬁcations on Chromosome 1

On human chromosome 1, there are 2,057 Swiss-Prot entries (Release 2012_07) and 2,040 protein-coding genes in the Ensembl protein database (Release-67). In this project, we identiﬁed 1,252 Swiss-Prot entries, corresponding to 1,227 Ensembl genes. The baselines and identiﬁed genes/proteins for chromosome 1 from Ensembl, GPMDB, PeptideAtlas, ProteinAtlas, and neXtProt are shown in Table 1.

Enrichment Analysis of Identiﬁed Proteins on Chromosome 1

The enrichment of identiﬁed proteins was analyzed in terms of chromosome structure, protein evidence, and functional pathway.

According to the gene density distribution along chromosome 1 (Figure 4A), the areas with highest gene densities are at the two terminals of the p and q arms (1p3 and 1q4) and at 1q2 close to the centromere. We also noticed that proteins in several regions had been heavily identiﬁed, including those at 1p2 and 1q2-1q3. On the other hand, there were several regions with a very low protein identiﬁcation rate, such as the region of the q terminal, 1q2.

Compared with public data sets (including protein evidence from Swiss-Prot, high quality mass spectra, and antibody and disease data from OMIM and Sanger CGC database, respectively), 70%—80% of annotated proteins were covered by our CCPD (Figure 4B). The results indicate that CCPD is enriched in the known and well-annotated protein list of chromosome 1.

The overlap of the three well-studied protein post-translational modiﬁcation data sets and CCPD show that 857 of the 1,252 identiﬁed proteins on chromosome 1 contain one or more modiﬁcations (Figure 4C). There are only five overlapped proteins among those acetylated or glycosylated. The percentage of all modiﬁed proteins represented by the identiﬁed proteins varies with the type of modiﬁcation. Of the 481 glycoproteins, only 47.8% were identiﬁed. However, 83.6% of the 758 phosphorylated proteins were identiﬁed, as were 96.7% of the acetylated proteins (Figure 4C).
Further pathway enrichment analysis shows that the genes of eleven pathways were significantly enriched on chromosome 1. The ratio of the identified gene number to the total gene number on chromosome 1 varies among the different enriched pathways (Figure 4D). For example, only four proteins (7%) in the olfactory transduction pathway were identified, indicating the obvious absence of the pathway in the three tissues studied. The genes involved in glutathione metabolism and Fc gamma R-mediated phagocytosis were fully identified, indicating that these are highly active pathways in the tissues represented in our data set. These results were consistent with the gene density analysis described above and the unidentified blocks discussed below.

Analysis of Technical and Sampling Bias of Unidentified Proteins on Chromosome 1

The overlap between the CCPD and the public protein/mass spectra data indicates that there is still room for improvement of protein identification. Thus, the analysis of identified and unidentified proteins contributes to a better understanding of technical and sampling bias and might result in greater coverage of protein-coding genes in the future.

Among the 2,057 proteins encoded by genes on chromosome 1 in the Swiss-Prot database, 1,252 proteins (60.87%) were identified and 805 were unidentified proteins in our proteome data set. Compared with the three public mass spectra databases (GPMDB, Human PeptideAtlas, and NIST_SL), 356 of the unidentified proteins were matched to the high-confidence protein list of the public mass spectra databases, and 410 proteins were among the low-confidence results in these databases. However, 39 proteins remained that could not be found in any of the three public databases (Figure 4E). Herein, all proteins on chromosome 1 can be divided into four subgroups, Group A (1,252 identified proteins), Group B (356 unidentified proteins with high-confidence in public databases), Group C (410 unidentified proteins with low-confidence in public databases), and Group D (39 unidentified proteins without mass spectra evidence). The physicochemical properties including MW, Hy, pI, and protein detectability of the four groups were analyzed and compared.

As shown in Figure 4E, the average MW of Group A proteins was higher than that of the three unidentified-protein groups. Further, the 39 proteins not detected in the three public databases had the lowest average MW, indicating that the low-MW proteins tend to be more difficult to identify by mass spectrometry. Bias in Hy, pI, and protein detectability could also be seen with unidentified proteins. In conclusion, the identified proteins tend to be higher in molecular weight, hydrophilicity, acidity, and detectability. The bias in physicochemical properties might be caused by the proteomic technology and might also lead to absence of membrane proteins (Supplementary Figure S2A).

In addition to technical bias, two additional factors may be associated with protein identification. Using mRNA abundance as an index to estimate the abundance of the corresponding proteins, we found that proteins are more difficult to identify...
when the abundance of the corresponding mRNA is low (Figure 4E). Tissue specificity is another factor for unidentified proteins, which tended not to be annotated in the liver, stomach, and colon, according to the ProteinAtlas database, accounting for the absence of these proteins in our data set (Supplementary Figure S2B).

**Block Analysis of the Unidentified Proteins along Chromosome 1**

We found that there are some regions on chromosome 1 marked with ‘dark’ in the tabular visualization, suggesting that few of the genes in those regions are expressed. Here, we call these regions “unidentified protein blocks”. The significant unidentified blocks along chromosomes could be obtained from the number of continuously identified/unidentified genes. On chromosome 1, there are 2,057 proteins, with 1,252 identified and 805 unidentified. The chance that one gene would be missed is $805/2057 = 0.391$. Therefore, the probability of $n$ continuously missed genes is $p = (0.391)^n$. A region is estimated to be significantly unidentified if it contains at least five continuous missed genes ($p < 0.01$). Using this as the cutoff, 18 significant unidentified gene blocks have been identified (Supplementary Table S6).

When these unidentified gene blocks are combined in the tabular visualization (Supplementary Table S4), seven expressing gene blocks are also detected along chromosome 1, with the gene number ranging from 19 to 69, according to the expression levels of the transcriptome and proteome (Figure 5A). Interestingly, of these seven blocks, three are close to the centromere and four are close to the terminals of the long or short arms, which is also consistent with the observation shown in Figure 4A.

In blocks 1, 2, 4, 5, and 7, most genes could not be identified by proteome evidence, and transcriptome evidence was either low or without detectable signal. This observation is consistent with the protein evidence in Swiss-Prot, most of which is from the corresponding transcript. The mass spectra evidence from GPMDB, the NIST spectra library, and the Human PeptideAtlas for these genes also tends to be low-confidence or absent. We further analyzed the sequences of the genes in these blocks and found that most were tandem duplicates. For example, block 7 contains 55 genes, 42 of which belong to the olfactory receptor family and are tandem duplicates. Block 5 is composed of three different gene families, including six CD5 antigen genes, six Fc receptor-like genes, and 16 olfactory receptor genes. The putative neuroblastoma breakpoint family repeat region could be found in Block 3, which was significantly enriched on chromosome 1. A detailed list of the tandem duplicated gene families in each block is shown in Figure 5A.

Two blocks (3 and 6) contain genes with low/absent protein expression but with median or high mRNA expression (Figure 5A). However, analysis of the physicochemical properties and detectability of these proteins showed no obvious bias. More work should be carried out to understand these phenomena.

**Analysis of Gene Coexpression Patterns along Chromosome 1**

In eukaryotes, the order of genes on the chromosomes is not random and genes with similar and/or coordinated expression are often clustered. Although transcriptomic analysis has been
used to establish gene coexpression profiles, research using proteomic data to establish gene coexpression clusters is still in a preliminary stage. Here, we investigated the gene coexpression clusters on human chromosome 1 by analyzing the transcriptome (R) and proteome (P) expression data generated from the collected samples. Further, gene cofunction analysis (F) using GO annotation was used as additional evidence of gene coexpression clusters. Clusters with at least two types of evidence are reported here as gene coexpression clusters with high confidence.

Twenty-three high-quality gene coexpression clusters were found on human chromosome 1, four of which (17%) were confirmed by three types of evidence (R, P, and F). Twelve (52%) found by mRNA expression data were also confirmed by cofunction analysis (R and F), and five (22%) were confirmed by both protein expression data and cofunction analysis (P and F). The two (9%) remaining clusters were confirmed by mRNA and protein expression data (R and P). Detailed information on the coexpression clusters is presented in Supplementary Table S7.

When investigating the details of the 23 coexpression clusters, we found that the four clusters (1, 6, 8, and 19) confirmed by three types of evidence were regions of chromosome 1 with tandem duplicate genes. For example, cluster 1 contains three adjacent genes, C1QA, C1QC, and C1QB, at chromosome 1p36.12 (Figure 6A), all of which are related to the phenotype of C1q deficiency, a rare autosomal recessive disorder associated with several diseases (http://omim.org/entry/613652).

Twelve clusters were found by mRNA expression data and cofunction analysis. Genes in most of these clusters were also found to be tandem duplicates. For example, guanylate-binding proteins located at 1p22.2, amylase protein components at 1p21.1, late cornized envelope proteins at 1q21.3, and olfactory receptor family at 1q23.1 and 1q44 were all found in such types of clusters.

An interesting observation was that although 23 genes of the olfactory receptor family in cluster 23 were found with coexpression patterns using mRNA data, no such proteins were identified in this region, which is also consistent with our earlier analysis.

Proteomic expression data identified five more significant coexpression clusters (2, 3, 10, 11, and 21), which were also verified by cofunction analysis. For example, cluster 3 contains four adjacent genes, NEXN, FUBP1, DNAJB4, and GIPC2 located at 1p31.1 (Figure 6A). The phenotype of NEXN expression is related to cardiomyopathy (http://omim.org/entry/613121). The adjacent FUBP1 is a cancer gene, and mutation of FUBP1 contributes to human oligodendroglioma. FUBP1 and DNAJB4 were also found on the list of genes coexpressed with NEXN in COXPRESdb, a database of coexpressed gene networks in mammals.

Although genes in clusters 18 and 20 are not cofunctionally present, the two gene clusters located at 1q31.2 and 1q32.1 were verified by both protein and mRNA evidence, indicating a strong correlation in expression between the genes and gene product
pairs in these regions. For example, four genes, UCHL5, TROVE2, GLRX2, and CDC73, are in cluster 18, located at 1q31.2 (Figure 6A). CDC73 is a disease-related gene with hyperparathyroidism phenotype. Based on the annotations in COXPRESdb, the other three genes are also on the list of genes coexpressed with CDC73. Cluster 20 contains four genes, LAD1, TNNI1, PHLDA3, and CSRPR1. CSRPR1, located at 1q32.1, may be a cancer suppressor gene involved in the progression of colorectal cancer.50 PHLDA3 was also found coexpressed with CSRPR1 and LAD1 in COXPRESdb. Based on the analysis of mRNA expression data, the four genes in cluster 20 were indeed overexpressed in normal tissues of the colon and underexpressed in colon cancer tissues (Supplementary Table S3). This was also observed in the protein expression data. The abundance of CSRPR1 and LAD1 in normal colon tissue was 1 order of magnitude higher than that in colon cancer tissue (Supplementary Table S1). Figure 6B shows the detailed investigation of XICs with their high-confidence unique peptides for these two gene products identified in colon cancer and normal tissues (see Supplementary Figure S3 for annotated spectra), as we expected. Although 23 clusters were found in this study, these do not represent all of the coexpression clusters on chromosome 1. With the accumulation of transcriptomic data, as well as proteomic data, more tissue- or disease-specific coexpression clusters may be identified. We expect that the proteome expression profile can help to confirm and adjust the gene coexpression clusters found using transcriptomic data, and the integration should shed light on the mechanisms of transcriptional and translational regulation.

Data Visualization and Database Construction

The large number of protein-coding genes on chromosome 1 makes it impossible to display all the information in a single condensed graph. Here, two different visualization methods were applied to present the data and results for different purposes. The circular visualization provides the overview of data and results of analysis in a condensed graph. Here, we integrated the information on protein characterization, protein and mass spectra evidence, expression abundance of the transcriptome and proteome, and the results of analysis of unidentified blocks and coexpression clusters using the circular visualization, which is a good way to provide a panoramic view of chromosome 1 (Figure 7).

The tabular visualization provides an easier way to check the detailed information for each protein (Supplementary Table S4). Most information that was integrated into the circular visualization, with the exception of the results of analysis, is also found in the tabular visualization. However, the table can also be used for data analysis and confirmation. More detailed information and external links can be accessed and queried in the C-HPD. Currently, the content of the database consists of three different types of information, i.e., gene and protein arrangement along chromosomes, powerful annotation of proteins and genes, and expression information for the transcriptome and proteome. C-HPD is now publicly available at http://proteomeview.hupo.org.cn/chromosome. More detailed information and updates of chromosome 1 proteome research can be found in the database.

CONCLUSION

As a cooperative international project, C-HPP will be advanced by the multiple research groups that claim their chromosomes of interest. Accordingly, the CCPC was established and claimed three chromosomes. As the first effort, CCPD was constructed and 12,101 high-confidence proteins were identified. Although a large proportion of identifications could be found in the public mass spectra database, an additional 10,741 peptides and 564 proteins were identified in our data set.

Beijing Proteome Research Center focuses on chromosome 1, for which 1,252 proteins mapping to 1,227 genes were identified in our data set. Using this large data set, multiple analyses were performed. First, the data characterizing protein-coding genes on chromosome 1 were collected, and enrichment analysis was performed, focusing on chromosomal arrangement, protein characterization, and pathway. The bias in protein identification was found to correlate with physicochemical properties and tissue specificity, indicating that additional technologies and tissues are needed in future analyses. The analysis of qualitative identified/unknown blocks and quantitative coexpression clusters shows the nonrandom distribution of genes along the chromosome.

ASSOCIATED CONTENT

Supporting Information

Supplementary materials and methods, figures, and Tables S5—S7. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*(F.H.) Tel and Fax: 8610-68171208. E-mail: hefc@nic.bmi.ac.cn. (P.X.) Tel and Fax: 8610-80705155. E-mail: xupingghy@gmail.com. (Y.Z.) Tel: 8610-80705225. E-mail: zhyunping@gmail.com. (P.Y.) Tel and Fax: 8621-54237961. E-mail: pyyang@fudan.edu.cn. (S.L.) Tel and Fax: 8610-80485324. E-mail: siqiliu@genomics.org.cn.

Author Contributions

These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

Additional Data Available: Supplementary Tables S1—S4 associated with this manuscript may be downloaded from ProteomeCommons.org Tranche using the following hash: mrP+jS5yq1U5+I2f3VdjFReydqIQkIo41tJXWxR74-oyGm34LhmZ2fGJ03hD9S9D+vUV11CXTTuD1Ewv+Ik9NUu05+cQ7MAAAAAAAAewA==

ACKNOWLEDGMENTS

This work was supported by the Chinese National Basic Research Program (2011CB910600, 2013CB911200 and 2011CB910701), the National High-Tech Research and Development Program (2011AA02A114, 2012AA020201, 2012AA020409 and 2012AA020502), the National Natural Science Foundation of China (31070673, 31170831, 31170780, 31203001, 21105121, 21275160 and 9113009), the National International Cooperation Project (2011-1001), Beijing Municipal Natural Science Foundation S122013, Key Projects in the National Science & Technology Pillar Program (2012BAAF14B00), and the State Key Lab Project (SKLP-K200904 and SKLP-Y201102).

ABBREVIATIONS

CCPC, Chinese Human Chromosome Proteome Consortium; CCPD, Chinese Chromosome Proteome Data set; C-HPD, Chinese Chromosome-Centric Human Proteome Database;
REFERENCES


(9) Chinese Human Liver Proteome Profiling Consortium. First Insight into the Human Liver Proteome from PROTEOMESKY-LIVERHu 1.0, a Publicly Available Database. J. Proteome Res. 2010, 9 (1), 79–94.

