Systematic Analyses of the Transcriptome, Translatome, and Proteome Provide a Global View and Potential Strategy for the C-HPP

Cheng Chang,†,# Liwei Li,†,# Chengpu Zhang,†,# Songfeng Wu,*† Kun Guo,*† Jin Zi,*† Zhipeng Chen,†,# Jing Jiang,† Jie Ma,† Qing Yu,† Fengxu Fan,† Peibin Qin,† Mingfei Han,† Na Su,† Tao Chen,† Kang Wang,† Linhui Zhai,† Tao Zhang,† Wantao Ying,† Zhongwei Xu,† Yang Zhang,‡ Yinkun Liu,‡ Xiaohui Liu,‡ Fan Zhong,‡ Huali Shen,‡ Quanhui Wang,§Ⅱ Guixue Hou,§Ⅱ Haiyi Zhao,§ Guilin Li,§ Siqi Liu,§Ⅱ Wei Gu,⊥ Guobin Wang,⊥ Tong Wang,⊥ Gong Zhang,⊥ Xiaohong Qian,⊥ Ning Li,†,⊥ Qing-Yu He,*⊥ Liang Lin,*§ Pengying Zhu,*,† Fuchu He,*,† and Ping Xu,*†

†State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Engineering Research Center for Protein Drugs, National Center for Protein Sciences, Beijing Institute of Radiation Medicine, 27 Taiping Road, Beijing 102206, P. R. China
‡Institutes of Biomedical Sciences, Department of Chemistry and Zhongshan Hospital, Fudan University, 130 DongAn Road, Shanghai 200032, China
§BGI-Shenzhen, Beishan Road, Yantian District, Shenzhen 518083, China.
⊥Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes, Institute of Life and Health Engineering, College of Life Science and Technology, Jinan University, 601 Huangpu Avenue West, Guangzhou 510632, China

Supporting Information

ABSTRACT: To estimate the potential of the state-of-the-art proteomics technologies on full coverage of the encoding gene products, the Chinese Human Chromosome Proteome Consortium (CCPC) applied a multimics strategy to systematically analyze the transcriptome, translatome, and proteome of the same cultured hepatoma cells with varied metastatic potential qualitatively and quantitatively. The results provide a global view of gene expression profiles. The 9064 identified high confident proteins covered 50.2% of all gene products in the translatome. Those proteins with function of adhesion, development, reproduction, and so on are low abundant in transcriptome and translatome but absent in proteome. Taking the translatome as the background of protein expression, we found that the protein abundance plays a decisive role and hydrophobicity has a greater influence than molecular weight and isoelectric point on protein detectability. Thus, the enrichment strategy used for low-abundant transcription factors helped to identify missing proteins. In addition, those peptides with single amino acid polymorphisms played a significant role for the disease research, although they might negligibly contribute to new protein identification. The proteome raw and metadata of proteome were collected using the iProX submission system and submitted to ProteomeXchange (PXD000529, PXD000533, and PXD000535). All detailed information in this study can be accessed from the Chinese Chromosome-Centric Human Proteome Database.

KEYWORDS: proteome, transcriptome, translatome, chromosome, missing protein, SAP
INTRODUCTION

The Chromosome-Centric Human Proteome Project (C-HPP) aims to define the full set of proteins encoded by the human genome.1,2 In contrast with genomic data that are relatively stable in a living organism, expression of encoded genes is time- and space-dependent, and the gene expression profiles at either the transcriptional or translational level vary widely in different cells and tissues. There is strong evidence that gene expression in certain tissues or cells is in homeostasis. For instance, Schaab et al. examined the proteomes from 11 different types of cancer cells and found that an average of 7337 unique proteins were identified in one cell line alone. The analysis of all 11 cell lines resulted in a total of 10 183 nonredundant proteins, in which the identified protein number was saturated easily, even with the addition of more cell lines.3 A logical conclusion is that a cell or tissue somehow possesses a relatively stable amount of the expressed proteins. Estimation of the appropriate size of the proteome is important to understand the status of protein expression and the chromosomal proteome in a target cell or tissue. In other words, as with defining the size of a genome in genomics studies, the investigation of overall gene expression involves determining how many genes are transcribed or translated in a cell or tissue.

The Chinese Human Chromosome Proteome Consortium (CCPC) initiated a collaborative study in mainland China to focus on proteomics in chromosomes 1, 8, and 20. In 2012, the CCPC performed the chromosome proteomic research on three tissues, liver, colon and stomach, as well as on several corresponding cancer cell lines. The CCPC managed to generate a large data set of chromosome proteomics, CCPD1.0, which contains 12 101 unique proteins (peptide and protein FDR < 1%) identified by liquid chromatography-tandem mass spectrometry (LC–MS/MS). When they began to design the next phase of the chromosome proteomics study, a major concern was how to evaluate the number of the translational genes in an individual chromosome. On the basis of such fundamental data, we are able to overview the translation ratio of the encoding genes occurring in a chromosome, but we questioned how many sequence variants could be identified from the translated proteins. Toward this goal, we set three priorities: (1) to focus on similar types of hepatoma cell lines with subtype forms that minimize cell-specific effects on protein expression, (2) to estimate the translational genes within protein signals that are unrestricted, and (3) to enrich low-abundant proteins without biasing the biophysical properties of the modified peptides.

The translation of mRNA into protein is a complicated process. In general, mRNAs carry translational messages and participate in subsequent translational processes. There is an argument regarding translation efficiency from mRNA: one side considers that the total mRNA is able to be completely translated, and the other side debates that only mRNAs bound to ribosomes are translated. Moreover, mRNAs bound to the ribosome-nascent chain complex, RNC-mRNA, may exhibit a much closer correlation with protein synthesis. For example, Wang et al. systematically analyzed the relative abundances of mRNAs, RNC-mRNAs, and proteins on a genome-wide scale, reporting a strong correlation between RNC-mRNAs and proteins in their relative abundances, which could be established using a multivariate linear model that considered mRNA length.7 It is likely that mRNAs in the free or ribosome-bound form carry differences, which reflect the specific characteristics of the translational message transmitted from transcription to translation. An mRNA that could bind to a ribosome is most likely a translational one. An integrated view of mRNA and RNC-mRNA, hence, is expected to better estimate the potential ability of genes to be translated. In this study, we investigate, for the first time, the corresponding expression statuses of mRNA, RNC-mRNA, and protein, encoded by chromosomal genes.

Technological innovation, particularly in MS and LC, is the driving force behind expanding the number of identified proteins. Scientists are keenly aware that protein identification faces several tough challenges, such as proteins with special structures and physicochemical properties, protein half-lives, gene expression that is tissue- or cell-dependent, and technical limitation of separation and detection. These problems have profound effects on the rate of protein identification. We have observed that the biological and biochemical properties of proteins are not easily controlled by experimental design. However, mass spectrometers that rely upon mass signals with high resolution, like Orbitrap Q Exactive MS (Thermo Fisher Scientific, San Jose, CA) and Triple TOF 5600 MS (AB SCIEX, Concord, ON), have achieved stable and reproducible performance in proteomics studies. In a large-scale protein survey, a technique in need of improvement for protein identification is the modification of LC conditions. Affinity chromatography is widely adopted in proteomics to enrich low-abundant and chemically modified proteins.

Affinity chromatography for modified proteins, such as TiO for phosphorylated peptides, lectin for glycosylated peptides, and antibodies for ubiquitinated proteins, was not used, because it may cause enrichment of the modification signals and lead to misinterpretation of data during the global profiling of the proteome. Alternatively, we proposed the transcription factor (TF) enrichment might be used to improve the detection of low-abundant proteins.8 Most TFs are located upstream of the transcriptional cascade and are believed to be of low abundance in tissues and cells; enrichment of TFs is likely to globally increase detection sensitivity.

We have chosen three hepatoma cell lines, Hep3B, MHCC97H, and HCCLM3, which have been generally accepted in liver cancer studies and recognized by their different oncologic behaviors.9,10 After careful cell culture and extraction of mRNAs and proteins, we systematically examined gene expression status at three levels, transcriptome (mRNA), translome (RNC-mRNA), and proteome, in the three cell lines. Bioinformatics analysis including data collection, processing and quality control, and quantitative evaluation was performed at all three levels. On the basis of the integrated -omics data, we obtained an overview of the distribution and function of the proteins, qualitative or quantitative. Meanwhile, the -omics data shed light on proteomics and the C-HPP (aiming to identify and characterize all human protein-coding genes) in terms of low-abundant proteins, missing proteins, and single amino acid polymorphism (SAP). The raw and metadata of the proteome were collected using the iProX submission system (http://www.iprox.org). All of the results from the transcriptome, translome, and proteome as well as their updates can be found in our C-HPP database (http://proteomeview.hupo.org.cn/chromosome/).

MATERIALS AND METHODS

Cells Used in This Study

Human hepatocellular carcinoma (HCC) cell lines were used in the cooperative study, which included two metastatic cell lines (MHCC97H and HCCLM3) and one nonmetastatic cell line (Hep3B). The MHCC97H and HCCLM3 cell lines were...

dx.doi.org/10.1021/pr4009018.J. Proteome Res. 2014, 13, 38−49
generated from the same 39-year-old Chinese male patient with HCC but with different lung metastatic potentials, as determined by the Liver Cancer Institute, Zhongshan Hospital, Fudan University. Both cell lines had an HBV-infected background containing HBsAg and a p53 mutation. However, the Hep3B cells had a p53 deletion, which was established from a black 8-year-old male patient. The cells were all maintained in H-DMEM medium supplemented with 10% fetal bovine serum at 37 °C with 5% CO₂. All of these cells were confirmed to be free of mycoplasma by continuous monitoring of 16S rRNA. The Hep3B, MHCC97H, and HCCLM3 cells (1.5 × 10⁵) were equally divided into three groups for follow-up experiments to examine the deep sequencing of the transcriptome, translatome, and proteome, respectively.

To estimate the number of expressed genes and determine the ability to detect their gene products, we prepared the cultured samples for total mRNA, RNC-mRNA, and protein from the ability to detect their gene products, we prepared the cultured cells for total mRNA, RNC-mRNA, and protein from the same batch of cultured cells in the same laboratory. This process minimized interlaboratory variation, which can confound data integration in chromosome-centric proteome investigations.

RNA Extraction and Whole Transcriptome and Translatome Sequencing

To prepare high-quality samples for the transcriptome, translatome, and proteome studies, we extracted the total mRNA and RNC-mRNA immediately after harvesting the cultured cells with 80% confluence. Total mRNA and RNC-mRNA extraction as well as library construction for RNA-seq were performed as previously described. In brief, equal samples prepared from three independent experiments were pooled. The sequencing libraries for total mRNA and RNC-mRNA were produced using the NEBNext mRNA Sample Prep Master Mix Set (Illumina) and were subsequently sequenced on an Illumina HiSeq-2000 sequencer for 50 cycles. High-quality reads that passed the Illumina quality filters were kept for the sequence analysis.

RNA-seq Data Analysis

To ensure accurate and quantitative mapping, the sequencing reads were mapped to the Ensembl v72 RNA reference sequences using the FANSe2 algorithm (http://bioinformatics.jnu.edu.cn/software//fanse2/) with the following parameters: −L55, −E3, and −S14. Alternative splice variants were merged (the analysis of splice variants is deferred for future studies). Genes with at least 10 mapped reads were deemed confident gene identification and quantification. Gene expression levels within a single cell line were calculated using the reads per kilobase per million reads (RPKM) method. The analysis of expression levels across cell lines was performed using the edgeR software package that implements a trimmed mean of M-value method based on a negative binomial distribution. The nucleotide positions with greater than five times the level of coverage were subjected to single nucleotide variant (SNV) analysis by using the Fisher’s exact test. SNVs were detected when the predominant nucleotide (>50% occurrence) significantly differed from the reference sequence. All sequencing data are available in the Gene Expression Omnibus database (accession number: GSE49994).

Protein and Peptide Fractionation and Mass Spectrometry Analysis

The fresh cells, as previously described, were harvested and lysed in buffer containing 8 M urea, 50 mM NH₄HCO₃, and 5 mM IAA. The total cell lysates (TCLs) were centrifuged at 12 000 g for 10 min at 4 °C to remove cell debris, and 1.5 mg proteins from each cell line were then collected and diluted in 2 M urea, followed by sequential in-solution protein digestion using Lys-C and trypsin at 37 °C. The tryptic peptides were purified on a C₁₈ Sep-Pak column (Waters UK, Manchester, U.K.). The cleaned samples were split into three different aliquots (500 μg each). The paired samples were distributed to two different laboratories at Beijing Proteome Research Center and Fudan University (sites 1 and 2) for protein identification and quantification. The mass spectrometers used in this study were the Orbitrap Q Exactive and the Triple TOF 5600. MHCC97H and HCCLM3 cells were analyzed at site 1 using the Q Exactive, whereas MHCC97H and Hep3B cells were analyzed at site 2 using both instruments. All three MS data sets were produced and analyzed following the same procedures with two technical replicates and 24 fractions per repeat.

To identify more low-abundant proteins, we used the same batch of cultured cells to enrich TFs with the same method as described. In brief, we used 2 mg of TCL and 30 pmol of catTFRE DNA for each reaction to enrich TFs. The trypsin-digested peptide mixture was purified on a homemade C₁₈ StageTip and then analyzed using the LTQ-Orbitrap Velos (Thermo Fisher Scientific, San Jose, CA). For detailed information about peptide mixture prefractonation and MS instrument settings for profiling and the TF-enriched data sets, see the Supporting Information.

Database Searching and Data Processing

The database searching procedure was performed as previously described. The search results were processed using PepDistiller and then only those results were retained with the threshold of protein FDR below 1%. For label-free quantification, the area under the extracted ion chromatograms (XICs) was calculated using the homemade tool SILVER, and then the iBAQ index was determined; the median was normalized to equal (1) to eliminate biases from different experiments for subsequent quantitative proteomics analyses.

To combine the results from different instrument and sites, we determined the combined intensities of MHCC97H cells at site 2 as the average of the results from the Q-Exactive and Triple TOF 5600. Subsequently, a robust regression between the intensities of MHCC97H cells at both sites was implemented (Supplementary Figure S1 in the Supporting Information) with the quantification results at the two sites. According to the linear regression function, the results at site 2 were transformed into those at site 1. This method proved to be reliable, as shown by its good reproducibility in each experiment. (See Supplementary Figure S2 in the Supporting Information; the correlation between replicates in each experiment was higher than 0.9.)

Bioinformatics Analysis of Identified Proteins

Four publicly available and commonly used databases were selected for the comparison of protein identification, including the global proteome machine database (GPMD, release 2013_7_1), PeptideAtlas (Release 2013_08), Human Protein Atlas (HPA, version 11.0), and neXtProt Gold (Release 2013-6-11). For GPMD, the data set annotated as “Green” was used, with a threshold “>20 observations and log(e) < −5”. For HPA, only those proteins with summary evidence of “high” or “medium” were retained. Only those proteins with “Gold” evidence were used for neXtProt.

On the basis of protein sequences, physicochemical properties of molecular weight (MW), hydrophobicity (Hy), and isoelectric point (pI) were calculated using ProPAS software. Methods of hierarchical clustering of the three -omics data quantification...
profiles were previously described. For the gene annotation, gene ontology (GO) analysis was performed using DAVID software v.6.7. The terms with p-value less than 0.01 were considered significantly enriched. Overlap analyses of different data sets were performed using Venn diagram.

**Single Amino Acid Polymorphism Identification**

Using the SNV data of mRNAs that were translated from the Hep3B, MHCC97H, and HCCLM3 cell lines, three sample-specific protein databases were constructed, respectively, according to the Ensembl database. For each hepatoma cell line, proteomic data were inputted to Mascot for database search against the combined original Ensembl and sample-specific databases. The target-decoy strategy was applied to maintain a unique peptide FDR <0.1% after filtering the Mascot ion score, which ensures the accuracy of almost all of the identified peptides. All of the spectra were manually verified to confirm their qualities.

**Proteome Data Collection and Database Construction**

To ensure the standardization and streamline of proteome data collection, the iProX submission system (http://www.iprox.org/), which has been established following the data-sharing policy of the ProteomeXchange consortium, was used to collect and repot proteomics raw data and metadata generated in this second cooperation, and all of the MS proteomics data have also been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) with the following identifiers, PXD000529, PXD000533, and PXD000535, to facilitate the data-sharing within the CCPC and other research groups (Supplementary Table S1 in the Supporting Information). The data were also imported into the C-HPD database that was constructed last year. The detailed identification information of proteins, peptides, and spectra as well as their basic annotation for relevant encoded genes and chromosomes can be found at the C-HPD Web site (http://proteomeview.hupo.org.cn/chromosome).

**RESULTS AND DISCUSSION**

**Strategies of the Trans-Omics Profiling of Human Protein-Coding Genes**

At the initiation of the C-HPP, chromosomes 1, 8, and 20 were studied by different research teams in mainland China. To carry out more efficient research and complement each other’s expertise, the teams decided to collaborate and produce a large-scale data set from 18 different samples for the first cooperation in 2012. In total, we identified 12,101 proteins with high confidence, covering 59.8% of protein entries in Swiss-Prot. Although this number was close to the record of high confident protein number in PeptideAtlas, it was still far away from the C-HPP goal of full coverage for all of the proteins encoded by chromosomal genes.

Gene expression is the prerequisite for identification through proteomics and potential reannotation of chromosomes. To estimate the number of expressed protein-coding genes and the power of current state-of-the-art proteomics technologies for protein identification as well as potential solutions for missing proteins, we applied a three stage multiomics strategy in cultured human cells. In the first stage, mRNA, RNC-mRNA, and protein were extracted from identical samples, and whole transcriptome and translatome libraries were sequenced to saturation using RNA-seq (Figure 1). Thereby, the number of actively transcribed and efficiently translated protein-coding genes in a given state could be estimated as the sum of protein-coding genes expressed in a specific cell. On the basis of such an optimal estimate, a high-coverage proteomics analysis was performed on both the extracted TCL and enriched TFs from the same cultured cells in the second stage. To prevent expression bias of the human genome and to achieve a higher coverage of the proteome, we used three HCC cell lines with no or gradually increasing metastatic potential that resulted from the varied patterns of gene expression. To learn more about the MS preference for protein identification, we tested the same samples on multiple proteomics platforms, including Q Extractive and Triple TOF S600, to partially overcome the premature saturation of distinct protein identifications. In the third stage, systematic data analysis was performed on all identified proteins. This study was expected to provide useful insights into the achievable proteome coverage and potential solutions for missing proteins, differential protein expression, and the relationship between the proteome and metastatic potential and mutations (Figure 1).

In this cooperative research, the data sets for the three -omics were distributed to the teams that were studying chromosomes 1, 8, and 20 for further analysis, focusing on their respective chromosomes and topics of interest, including missing proteins, liver cancer biology, their metastatic potential, and mutation analysis. We focus on general issues, including the qualitative and quantitative exploration of trans-omics data, quantitative
analysis of missing proteins, and sample-specific mutations using trans-omics data, which may provide valuable resources and guides for further C-HPP research.

**High-Quality Transcriptome, Translatome and Proteome Data from Three Hepatoma Cell Lines**

Both transcriptome (mRNA) and translatome (RNC-mRNA) data were produced from the RNA-seq experiments. On average, 170 million reads were obtained for each mRNA and RNC-mRNA sample, and 84% of them were mapped to Ensembl-v72 RNA reference sequences using the FANSe2 algorithm. Compared with previous studies of the transcriptome and translatome,\textsuperscript{18,30} we obtained three to seven times higher throughput. The median of sequencing depth (the average number of times that each nucleotide acid was sequenced) was 22.6, and 62% of the genes reached an average sequencing depth of more than 10, suggesting that the data sets offered a solid basis for downstream analysis (Supplementary Figure S3A,B in the Supporting Information).

On the basis of the high-quality RNA-seq data, the average gene numbers of the transcriptome and translatome for the three cell lines were 17636±114 and 17127±101, respectively (Supplementary Figure S4 in the Supporting Information). Meanwhile, the current throughput of the transcriptome and translatome approached saturation (additional details are provided in the Methods and Supplementary Figure S5 in the Supporting Information).

In the proteome, on the basis of the integration of multisite MS data with technical repeats and low-abundant TF-enriched data, a large-scale data set named CCPD2013 was established, which contained 7,541,744 mass spectra with 1,497,791 peptide-spectrum matches (19.9% against the total mass spectra, Supplementary Table S2 in the Supporting Information) and 110,758 unique peptides, matching 9064 proteins by the parsimony rule with a protein FDR lower than 1% (Supplementary Figure S6 and Supplementary Tables S3 and S4 in the Supporting Information). The CCPD2013 data set was of high quality, as

---

*Figure 2. Trans-omics analysis and comparison of hepatoma cell lines. (A) According to central dogma, genetic information flows from the genome to the proteome through the transcriptome and translatome. Differences between the adjacent -omics are shown in the Figure, and their possible causes are also listed in the Figures. (B) Venn diagram of gene numbers in the three -omes. (C) Identification rate of the three -omes and a public data set (named PublicDB2012) for different chromosomes. The blue bar represents the number of protein-coding genes in the Ensembl database. The PublicDB2012 represents the combination of neXtProt gold (Nov. 2012), Human PeptideAtlas (Dec 2012), GPMDB (Green) (26 Nov. 2012), and Human Protein Atlas Evidence high or medium (Dec. 2012) mentioned in table 1 in a previous paper.\textsuperscript{34}*

---

**Journal of Proteome Research**

Article

dx.doi.org/10.1021/pr4009018 | J. Proteome Res. 2014, 13, 38−49

dx.doi.org/10.1021/pr4009018/1 J. Proteome Res. 2014, 13, 38−49
demonstrated by the large number of unique peptides per protein and high protein sequence coverage (Supplementary Figure S7 in the Supporting Information). On the basis of the high-quality -omics data, further comprehensive analysis focusing on the gene expression status was performed using a trans-omics comparison analysis.

**Trans-Omics Data Revealed Translation Control and Technical Limitation of Proteomics**

In cells, information about expression is transferred from mRNA to RNC-mRNA and then to protein. To obtain a global view of differences in gene expression in the three different -omes, we performed a systematic comparison.

As shown in Supplementary Table S5 in the Supporting Information, translatome data were 4.1% smaller than transcriptome data in Hep3B cells. The percentage was 2.4 and 2.2% for that of MHCC97H and HCCLM3 cell lines, respectively, indicating that there was a small number of genes whose mRNAs could not be transcribed into RNC-mRNAs. In Hep3B, there were 17,023 genes identified in the RNC-mRNA of Hep3B, 50.3% of which were verified by proteome data in our study, as shown in Supplementary Figure S4 in the Supporting Information. For MHCC97H and HCCLM3 cells, the overlap between the proteome and translatome was 52.3 and 44.4%, respectively. By combining the results of the three cell lines, the coverage of the proteome compared with that of the translatome was 54.4%. There were still a large number of genes that could not be identified by proteomics because of the sensitivity issue and bias of current proteomic technologies.

With the three cell lines combined, there were 9918 commonly identified genes in the transcriptome, translatome, and proteome (Figure 2B). We also noticed that some unique genes could be found in each -ome with no clear explanation. For example, there were 134 genes (0.7%) found in the translatome but not in the transcriptome, and 183 genes (1.8%) were found in the proteome only (Figure 2A).

A comparison between the identification rate and the whole genome for each chromosome revealed highly similar distributions of the three -omics data sets (Figure 2C). This high similarity of identification rates between the three -omics indicates that although different chromosomes have different percentages of missing proteins the major reasons were similar including tissue specificity, sensitivity of proteomics, and technical biases. This means that the teams studying different chromosomes could share their methods for finding missing proteins.

Moreover, from a global perspective, the dynamic ranges of the detected intensities (indicating protein abundance) of those overlapping genes from the three -omes were different. It seemed that the intensity ranges expanded gradually from the transcriptome to the translatome and further to the proteome (Methods and Supplementary Figure S8 in the Supporting Information). A similar observation was also previously made. The wider dynamic range of the proteome might hinder the goal of full coverage for the proteome in the C-HPP.

**Trans-Omics Analysis of the Relationship between Protein Abundance and Biological Features**

To assess the reliability of our quantitation information in the three -omes, we constructed hierarchical clustering and the Spearman coefficient matrix for the three hepatoma cell lines, based on abundance. As shown in Figure 3A, the proteome data were disparate from the transcriptome and translatome data due to technological differences (Figure 3A). However, the transcriptome and translatome clustering results were similar, indicating that the expression information for the transcriptome could be conservatively transferred to the translatome.

The clustering of the three cell lines showed that Hep3B cells obviously differed from the other two cell lines. Similar results were found from an analysis of the R^2 coefficient matrix (Supplementary Figure S9 in the Supporting Information). This result is consistent with the origin of these cell lines, as both MHCC97H and HCCLM3 cells were established from a single Chinese male patient who differed from the patient from which Hep3B cells were derived.

An arguable hypothesis suggests that protein abundance might be related to molecular function.32,33 In this study, a stringent experimental design allowed a more accurate determination of the abundance of identified genes. To examine the relationship between protein abundance and biological function, we divided all three -omics data sets into four equal subsets based on abundance: Part 1 for genes with a very low abundances, less than the first abundance quartile (Q1); Part 2 for genes with abundances between Q1 and the median (Q2); Part 3 for genes with a modest abundance between Q2 and the third abundance

*Figure 3. Trans-omics quantitative map and its correlation with gene function. (A) Clustering results between the three data sets and the three cell lines. The color level in the matrix indicates the degree of correlation. Here the Spearman coefficient represents the correlation between two data sets. (B) The -omics data were divided into four parts according to abundances, and the GO analysis of each part was analyzed using David software. Genes with different abundances tended to have different biological functions. T, R, and P represent the transcriptome, translatome, and proteome, respectively.*
quartile (Q3); and Part 4 for genes whose abundance was higher than Q3. Subsequently, enrichment analysis was performed using David software in each subset for three GO categories: biological process (BP), cellular component (CC), and molecular function (MF), and the properties of the main GO terms in each part were calculated as the count of genes in one GO term divided by the number of all genes in each part.

As shown in Figure 3B, the percentage of different GO terms was calculated and is presented by different colors. Compared with the entire distribution of the transcriptome and translatome, the four parts of the proteome data were similar to Parts 2−4 of the transcriptome and translatome because genes with very low abundance in Part 1 of the transcriptome and translatome were hard to find using MS. On the basis of the transcriptome and translatome data, these genes with the lowest abundances in Part 1 were enriched in terms of biological adhesion, response to stimulus, and biological regulation, whereas the high-abundant genes were enriched in metabolic processes, cellular processes, and cellular components (Supplementary Figure S10 in the Supporting Information). These results suggest that new strategies are necessary to identify missing proteins compared with mRNA and RNC-mRNA data in the future.

Quantitative Exploration of Missing Proteins and Their Potential Solutions

The identification of missing proteins is one of the biggest challenges in the C-HPP study. Our previous research revealed that the physicochemical properties of proteins were part of the major reason for protein detectability through MS. Furthermore, the level of gene expression is known to greatly influence the results. However, quantitatively, which factor and how much gene expression affects the ability to detect proteins remain...
unclear. The data sets generated in this study provide a way for us to investigate these issues in a quantitative way.

First, RNC-mRNA abundance was taken as an estimate of protein expression level because these mRNAs were active and ready for translation, which was confirmed with a good correlation between the abundance of RNC-mRNAs and their proteins. To facilitate the analysis, we set three RPKM values as check points: RPKM=10 is the approximate peak of the translatome and proteome; RPKM=0.1 is the lowest abundance for detection of proteins encoded by those genes; and RPKM=0.01 is almost the lowest limit for the translatome through RNA-seq (Figure 4A). In this study, a total of 18 246 genes were identified in the translatome and 9922 genes (54.4%) were identified in the proteome. Among them, 6543 genes had RPKM larger than 10, in which 86.9% were covered by our proteomic analysis; only 1654 genes had RPKM less than 0.1, in which only 3.3% of their gene products could also be detected by MS. However, there were 3494 genes with RPKM of 0.1 to 1 and 6555 genes with RPKM of 1−10, in which 10.1 and 58.4% of their gene products, respectively, were identified in the proteome (Supplementary Table S7 in the Supporting Information). These results suggest that protein abundance plays a major role in protein detectability in an MS-based proteomics study.

We then combined protein physicochemical properties with RNC-mRNA abundance to assess the effects on protein detectability. To prevent a statistical bias, we put at least 250 genes in each individual group (Supplementary Figure S11 in the Supporting Information). As shown in Figure 4B,C, under the same level of RNC-mRNA abundance, proteins with a low MW (<10kD) or high Hy (>0.2) were hard to detect, but little effect could be found among different pIs (Figure 4D). Again, although these three physicochemical properties influenced protein detectability, protein abundance still played the most important role (Figure 4B−D).

To further understand the significance of protein physicochemical properties on the proteins missing in our proteomics study, we calculated the standard deviations (SDs) for the number of genes belonging to different levels of each physicochemical property and drew the changing curves of the SDs with increasing translatome abundance (log10 RKPM) (Figure 4E). The higher the deviation was, the greater the influence of the physicochemical property. Interestingly, the results showed that Hy had the largest impact on the missing proteins, whereas pI had the least impact on detection. This finding may be explainable using a shotgun approach on current proteomics platforms, as some proteins with high Hy or pI tend to have poor solubility or low extraction efficiency, which makes them difficult to detect under current shotgun proteomics approach.

Because protein abundance had a maximal influence on protein identification, strategies that change the relative abundance of proteins in proteomics samples may greatly affect their ability to be identified. Indeed, many enrichment methods have been widely used in the proteomics field to specifically identify proteins, which may also aid the identification of low-abundant proteins in the C-HPP. To test this idea, a new TF-enrichment strategy was also implemented in this cooperative study to identify low-abundant proteins such as TFs. Compared with the CCPD2013 data set, 31 proteins were uniquely identified using this simple enrichment. Among them, 14 (45.2%) were TFs, suggesting that this strategy allows a much greater probability of identifying TFs compared with other regular large-scale proteomics profiling strategies (Figure 5A). Again, the fact that intensity ranks of most TFs in the TF-enriched profile are higher than those in the proteome profile of Hep3B cells further confirms the idea that increasing the relative amounts of low-abundant proteins helps to identify missing proteins (Figure 5B).

**SAP Identification Using Sample-Specific Protein Databases**

In this study, transcriptome, translatome, and proteome data were produced using the same samples, providing an opportunity to explore sample-specific mutations. Here translatome data alone were used to construct a sample-specific database to discover the number and percentage of sample-specific SAPs that could be identified using the proteome data.

There were no more than 116, 129, and 50 unique peptide matches in the sample-specific databases of the Hep3B, MHCC97H, and HCCLM3 cell lines, respectively (Figure 6A). By removing redundant peptides, there were 219 unique peptides identified in the three cell lines, which comprised only 0.4% of all identified peptides. This percentage was even lower than that of theoretical SAPs, which was calculated from the sequenced single nucleotide polymorphisms (SNPs) based on our translatome data. We also found no obvious changes in any of our three cell lines. For example, in the MHCC97H cell line, the percentage of
identified SAPs was 0.3% and the theoretical value was 0.4% (Figure 6B).

Interestingly, there were only 11 SAP peptides identified in the three cell lines (Figure 6A); of these, six mutated sites were not in the termini of the peptides, which could be caused by ambiguous terminal amino acids in the Ensembl database (Supplementary Figure S12 in the Supporting Information). Some of these SAP peptides were identified in several data sets, indicating that rare SAPs might be widely used in these cell lines (Supporting Information).

We also found that 60 peptide pairs (27%) had both wild-type and mutated sequences in our data set (Supplementary Table S8 in the Supporting Information). Of these, 20 peptide pairs (9%) had both wild-type and mutated SAPs identified from the same cell line. In addition, the 60 peptide pairs belonged to 27 different substituted amino acid pairs (Figure 6C). Among them, the D→E mutation was the most popular (eight peptide pairs, 13.3%), followed by the V→A and I→V mutations (seven and five peptide pairs, respectively). Figure 6D–F shows the three most popular cases of the SAP peptide pairs with very similar mass spectra.

We compared our SAP data with the public database COSMIC and found only 13 peptides (6%) that had been previously published and collected36 (Supplementary Table S9 in the Supporting Information); these peptides were detected in breast carcinoma, stomach carcinoma, and so on, arguably suggesting that these mutations might be disease-related. The higher number of SAPs present in the samples indicated the diversity of their sequences. We also noticed that there were some differences between unique SAPs in the varied samples (Figure 6A), but we still do not know whether these SAPs are due to SNPs in the genome or if they were generated during the gene expression process because we do not have genomics sequences for these cell lines.

Although the identification of SAPs had little influence on protein identification, the significance of these mutations on physiological or pathological proteomics study was arguable.
The biological consequences of these mutations are also worth investigating.

**Contribution, Accumulation, and Sharing of the Proteome Data Set (CCPD)**

As previously described, 9064 proteins were identified and analyzed in the CCPD2013 data set. Combining data sets CCPD2013 and CCPD1.0 presented in 2013, the CCPD 2.0 was constructed with 12,740 highly confident identified proteins in total (Figure 7A), covering 62.9% of all human proteins in the Swiss-Prot database (release 2013_06).

Compared with CCPD1.0, CCPD2.0 contains 24,055 additional peptides and 641 new proteins (Figure 7B). Most of these newly identified proteins were annotated due to the presence of these proteins or their transcripts in the neXtProt database (Figure 7C), suggesting the reliability of our proteomics results. After mapping the additional proteins to each chromosome, there were 60, 27, and 17 additional proteins for chromosomes 1, 8, and 20, respectively, thus increasing the coverage to 63.2, 62.1, and 60.5%, respectively (Figure 7D). Compared with the three public databases (GPMDB, PeptideAtlas, and HPA) using the aligned neXtProt IDs, CCPD2.0 contained 324 newly confirmed gene products (Figure 7E). At present, the total number of verified proteins in all of the four databases was 17,026. Although HPA contained fewer proteins than did GPMDB, the percentage of its unique proteins (816, 8.5%) was higher than that of GPMDB (710, 4.8%), indicating that antibody technology could play a complementary role in the search for missing proteins from the MS-based proteomics platform.

Since 2012, the CCPC has produced large-scale data sets of 18 samples from 18 human tissues or cell lines and built the C-HPD database for storing, analyzing, and displaying these data. To further facilitate this collaborative study, the iProX submission system (http://www.iprox.org) was used for the collection, storage, and sharing of MS/MS raw files and experiment metadata. The iProX system will speed data analysis and the sharing of proteomics experiments and will support the long-term cooperation of CCPC in a future C-HPP. The analysis of this collaborative research has been imported into C-HPD for further access and sharing (http://proteomeview.hupo.org.cn/chromosome/).

**CONCLUSIONS**

Using deep sequencing and a systematic analysis of transcriptome, translatome, and proteome data generated from the same samples of human HCC cell lines Hep3B, HCC97H, and HCCLM3 with varied metastatic potentials, we found that an average of 17,636 ± 114 and 17,127 ± 101 genes were actively transcribed and translated, respectively; among them, a total of 9,064 gene products were identified using proteomics, which is ∼53% of the translated genes, and would not be accumulated easily using a regular large-scale proteomics approach due to saturation effects resulting from sensitivity issues associated with current technologies. A feature analysis of missing proteins further revealed that protein abundance plays a major role in MS-based protein identification and missing proteins might be efficiently identified from specifically enriched samples, which was further confirmed after TF analysis, as 31 new low-abundant...
proteins were identified that were missed in previous studies. With the sample-specific databases, we found that SAPs contributed only <0.4% of the total number of identified peptides, suggesting their negligible impact on protein identification. This first systematic proteomics study not only revealed that targeted proteomics approaches combined with high-coverage transcriptome and translatome analyses using deep RNA sequencing greatly aided the identification of missing proteins but also contributed 324 gene products with MS evidence to publicly available proteomics databases. Although some detailed methodology may need to be modified for the analysis of different chromosomes, the conclusions here are highly instructive for conducting efficient proteomics analyses for future C-HPP studies.

■ ASSOCIATED CONTENT

Supporting Information

Quantitative integration methods in multilaboratory. Correlation matrix and the box plot of protein intensities in all experiments. Sequencing depth and saturation analysis of transcriptome and translatome. Venn diagrams of identified gene numbers for the three omics data in three cell lines based on ENSG ID in the Ensembl database. Saturation analysis of RNC-mRNA and mRNA for the three cell lines. Venn diagram of the identified proteins in the three cell lines. Unique peptide number per protein and protein sequence coverage distributions. Abundance dynamic range of the common genes in the three omics data. Scatter distribution between two data sets and the corresponding R2 shown in the symmetrical place. GO enrichment analysis for the genes in Hep3B according to their abundances in the three omics data. Comparative analysis of all genes identified by translatome based on the RNC-mRNAs abundance and the protein physicochemical properties. Identification details of the six mutated peptide pairs in all three cell lines. Cooperative summary of the CCPC in 2013. Identification summary and success rate in the proteomic data set CCPD2013. Detailed information about the identified proteins and peptides in CCPD2013. Gene numbers in transcriptome and translatome for three cell lines based on ENSG ID. Normalized quantitative values of transcriptome and translatome data sets. Number of genes with different RPKM identified in translatome and proteome. Information of the 60 SAP peptide pairs different in the three cell lines. Peptides with SAP that match to cosmic database. Supplementary material and methods. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*Songfeng Wu: Tel/Fax: 8610-80727777-1141. E-mail: wusf897@gmail.com.
*Ning Li: Tel/Fax: 8610-80705155. E-mail: plkuningonly@gmail.com.
*Qing-Yu He: Tel/Fax: 86-20-85227039. E-mail: tqyhe@jnu.edu.cn.
*Liang Lin: Tel/Fax: 86-755-25274284. E-mail: linl@genomics.org.cn.
*Pengyuan Yang: Tel/Fax: 8621-54237961. E-mail: pyyang@fudan.edu.cn.
*Yunping Zhu: Tel: 8610-80705225. E-mail: zhuyunping@gmail.com.
*Fuchu He: Tel/Fax: 8610-68171208. E-mail: hefc@nicbmi.ac.cn.
*Pinguo Xu: Tel: 8610-80705155. E-mail: xupingggyh@gmail.com.

Author Contributions

*Cheng Chang, Liwei Li, Chengpu Zhang, Songfeng Wu, Kun Guo, Jin Zi, and Zhipeng Chen contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge the entire Chinese Human Chromosome Proteome Consortium. We also thank Jun Qin and Chen Ding at Beijing Proteome Research Center for providing reagents for TF enrichment analysis. This work was supported by the Chinese National Basic Research Program (2011CB910600, 2013CB911200, 2013CB910800, and 2010CB912700), the National High-Tech Research and Development Program (2011AA02A114, 2012AA020502, 2012AA020201, and 2012AA020409), the National Natural Science Foundation of China (31070673, 31170780, 81123001, 21105121, 21275160, and 31000379), the China-EU International Collaboration Program (2011-1001), Key Project from the National Science & Technology Pillar Program (2012BAF14B00), the National Science and Technology Major Project of the Ministry of Science and Technology of China (Nos. 2012ZX10004001-003-011 and 2012ZX10002012-006), National Mega Projects for Key Infectious Diseases (2013ZX10003002), the Beijing Municipal Natural Science Foundation (5112012 and 5122013), and the State Key Lab Project (SKLP-Y201102), and the Shenzhen Key Laboratory of Transomics Biotechnologies (No. CXB201108250096A).

■ ABBREVIATIONS

CCPC, Chinese Human Chromosome Proteome Consortium; CCPD, Chinese Chromosome Proteome Data set; C-HPD, Chinese Chromosome-Centric Human Proteome Database; RNC, ribosome-nascent chain; pi, isoelectric point; SAP, single amino acid polymorphism; TF, transcription factor; MW, molecular weight; Hy, hydrophobicity; MS, mass spectrometry

■ REFERENCES


dx.doi.org/10.1021/pr4000918 | Proteome Res. 2014, 13, 38–49


(9) Knowles, B. B.; Aden, D. P. Human hepatoma derived cell line, process for preparation thereof, and uses therefor. 1983.


(19) Craig, R.; Cortens, J. P.; Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* 2004, **3** (6), 1234–42.
