Journal of proteome research

Article

pubs.acs.org/jpr

Systematic Analysis of Missing Proteins Provides Clues to Help Define All of the Protein-Coding Genes on Human Chromosome 1

Chengpu Zhang,[†,▽] Ning Li,[†,▽] Linhui Zhai,[†,▽] Shaohang Xu,[§,▽] Xiaohui Liu,[∥,▽] Yizhi Cui,[⊥,▽] Jie Ma,[†] Mingfei Han,[†] Jing Jiang,[†] Chunyuan Yang,[†] Fengxu Fan,[†] Liwei Li,[†] Peibin Qin,[†] Qing Yu,[†] Cheng Chang,[†] Na Su,[†] Junjie Zheng,[†] Tao Zhang,[†] Bo Wen,[§] Ruo Zhou,[§] Liang Lin,[§] Zhilong Lin,[§] Baojin Zhou,[§] Yang Zhang,[∥] Guoquan Yan,[∥] Yinkun Liu,[∥] Pengyuan Yang,[∥] Kun Guo,[∥] Wei Gu,[⊥] Yang Chen,[⊥] Gong Zhang,[⊥] Qing-Yu He,[⊥] Songfeng Wu,*[,†] Tong Wang,*[,⊥] Huali Shen,*[,∥] Quanhui Wang,*[,‡,§,○] Yunping 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

[‡]Beijing Institute of Genomics, CAS, No. 1 BeiChen West Road, Beijing 100101, China

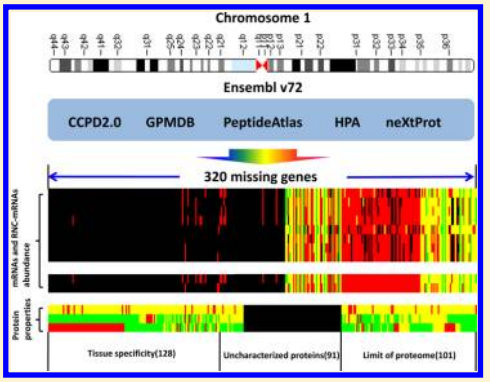
[§]BGI-Shenzhen, Beishan Road, Yantian District, Shenzhen 518083, China

[∥]Institutes of Biomedical Sciences, Department of Chemistry, School of Life Sciences, and Zhongshan Hospital, Fudan University, 130 DongAn Road, Shanghai 200032, 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, Guangzhou 510632, China

Supporting Information

ABSTRACT: Our first proteomic exploration of human chromosome 1 began in 2012 (CCPD 1.0), and the genome-wide characterization of the human proteome through public resources revealed that 32−39% of proteins on chromosome 1 remain unidentified. To characterize all of the missing proteins, we applied an OMICS-integrated analysis of three human liver cell lines (Hep3B, MHCC97H, and HCCLM3) using mRNA and ribosome nascent-chain complex-bound mRNA deep sequencing and proteome profiling, contributing mass spectrometric evidence of 60 additional chromosome 1 gene products. Integration of the annotation information from public databases revealed that 84.6% of genes on chromosome 1 had high-confidence protein evidence. Hierarchical analysis demonstrated that the remaining 320 missing genes were either experimentally or biologically explainable; 128 genes were found to be tissue-specific or rarely expressed in some tissues, whereas 91 proteins were uncharacterized mainly due to database annotation diversity, 89 were genes with low mRNA abundance or unsuitable protein properties, and 12 genes were identifiable theoretically because of a high abundance of mRNAs/RNC-mRNAs and the existence of proteotypic peptides. The relatively large contribution made by the identification of enriched transcription factors suggested specific enrichment of low-abundance protein classes, and SRM/MRM could capture high-priority missing proteins. Detailed analyses of the differentially expressed genes indicated that several gene families located on chromosome 1 may play critical roles in mediating hepatocellular carcinoma invasion and metastasis. All mass spectrometry proteomics data corresponding to our study were deposited in the ProteomeXchange under the identifiers PXD000529, PXD000533, and PXD000535.



KEYWORDS: *proteome, transcriptome, translatome, chromosome, missing protein, HCC*

INTRODUCTION

Following the guidance of the Chromosome-Centric Human Proteome Project (C-HPP),[1,2] the first proteomic exploration of human chromosome 1 was begun in 2012 by the Chinese Human Chromosome Proteome Consortium (CCPC).[3] Proteomic profiling of tissues and representative cells of liver cancer and gastrointestinal cancer (designated CCPD 1.0) generated 12 101 high-confidence identifications of human proteins, covering 59.8% of the protein entries in Swiss-Prot. This number is close to the 12 629 high-confidence proteins reported in PeptideAtlas (ver. Oct. 2012).[4] On the basis of the recent release of the Ensembl database (v72), CCPD 1.0

Special Issue: Chromosome-centric Human Proteome Project

Received: September 2, 2013
Published: November 20, 2013

identified 1248 out of 2071 (60.3%) protein-coding genes on chromosome 1, with 39.7% of the genes still requiring identification.[3] The state of proteome profiling of human protein-coding genes on chromosome 1 as determined by CCPD 1.0 is similar to that of the four public databases widely utilized by the C-HPP consortium, which estimated that 32.1% of the genes still lack protein identification evidence.[5] On the basis of the analysis of CCPD 1.0, as many as 823 genes on chromosome 1 remain unidentified. The first trial of genome-wide characterization of the human proteome also demonstrated that compelling evidence in the public databases was lacking for 656 genes on chromosome 1. Both analyses indicated that chromosome 1, compared with other human chromosomes, has the largest number of missing genes, which is the most intractable issue to address.

Several pilot bioinformatics analyses were conducted to uncover the reasons why those protein-coding genes are hard to identify.[3−5] The results indicated that extreme physicochemical properties, low abundance or transient expression, and sample specificity are the main reasons for the poor detectability of these gene products. For example, the first proteomic exploration of protein-coding genes on chromosome 1 in CCPD 1.0 indicated that the unidentified proteins tend to be low molecular weight (MW), highly hydrophobic, and highly basic. Moreover, several unidentified blocks have been discovered along chromosome 1. These "missing blocks" include several tandem duplicates of preferentially expressed antigen of melanoma (PRAME), olfactory receptor, and putative neuroblastoma breakpoint genes, which are considered tissue- or cell-type-specific.[3] Note that the transcripts of these genes were also found with either low or undetectable signals.[3,5]

Despite the meaningful discoveries already made, the lack of transcriptome data generated from the same samples weakened the confidence of such investigations. A more detailed identification strategy is therefore needed to address the missing proteins encoded by a significant portion of genes on chromosome 1.

In this study, three hepatocellular carcinoma (HCC) cell lines with progressively metastatic potential (Hep3B < MHCC97H < MHCCLM3) were selected for transcriptome, translatome, and proteome profiling. The MHCC97H and HCCLM3 cell lines were established from the same Chinese HCC patient but have different lung metastatic potentials. These cell lines have been widely used as HCC models in vitro.[6,7] Proteome profiling of the three cell lines was reimplemented using high-resolution, high-sensitivity mass spectrometry. Meanwhile, both mRNA and ribosome nascent-chain complex-bound mRNA (RNC-mRNA) sequencing was undertaken for the extracted mRNAs prepared from same samples used for proteomic analyses to accurately quantify the transcripts and translating transcripts.

On the basis of publicly available data regarding protein-coding genes on human chromosome 1 integrated with in-depth analyses of the omics data, the properties of the missing genes on chromosome 1 were comprehensively recharacterized to generate clues for use in future experiments. Furthermore, data saturation resulting from combining various large-scale proteome profiling data sets was investigated, and the potential solutions were discussed. Finally, genes in each of the cell lines that were significantly differentially expressed (SDE) in terms of metastatic potential were also investigated to provide a general description of HCC-related genes on chromosome 1.

MATERIALS AND METHODS

Detailed descriptions regarding cell-culture preparation and methods for proteome, transcriptome, and translatome profiling can be found in the report for a parallel study.[8] In brief, three HCC cell lines (Hep3B, MHCC97H, and HCCLM3) were obtained from the American Type Culture Collection (ATCC, Rockville, MD)[9] and Fudan University.[10] The nonmetastatic Hep3B cell line was established from an 8-year-old black male, whereas MHCC97H and HCCLM3 were clonal cell lines with different metastatic potentials that were derived from the same parent cell. In addition, HCCLM6, a clonal cell line of HCCLM3 with a higher metastatic potential, was also used for mRNA and RNC-mRNA sequencing.

Proteome Profiling of HCC Cell Lines

Transcription factors (TFs) were enriched using an affinity reagent composed of synthetic DNA containing a concatenated tandem array of the consensus transcription factor response elements (catTFRE) for the majority of TF families.[11] The TF-enriched sample was digested with trypsin, cleaned, and separated using an ultra-performance liquid chromatography (UPLC) system (nanoAcquity Ultra Performance LC, Waters, Milford, MA). The eluting peptides were identified using an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific).

For large-scale proteome profiling, proteins extracted from the Hep3B, MHCC97H, and HCCLM3 cell lines were digested with trypsin. The resulting peptide mixture was cleaned and separated by high-pH reverse-phase chromatography and then analyzed using an Orbitrap Q-Exactive mass spectrometer (Thermo Fisher Scientific) and a triple-TOF 5600 mass spectrometer (AB SCIEX). The resulting MS/MS peak lists were searched using the Mascot v2.3.2 local server[12] against the Swiss-Prot database (20 258 protein entries, 2013_06 release) along with 115 common contaminant protein sequences (ftp.thegpm.org/fasta/cRAP). A target-decoy-based strategy was applied to ensure that both the peptide- and protein-level false discovery rates (FDRs) were <1%.[13] Extracting ion-current-based label-free quantification was implemented using SILVER, an unpublished tool developed at the Beijing Proteome Research Center (BPRC). The protein iBAQ index was calculated to compare the abundance of proteins present in different samples.[14]

RNA Sequencing in HCC Cell Lines

Extraction of both mRNA and translating mRNA was performed as previously described.[15] Extracted mRNAs and RNC-mRNAs were sequenced on an Illumina HiSeq-2000 for 50 cycles. High-quality reads that passed the Illumina quality filters were retained for sequence analysis. The sequencing reads were mapped to the Ensembl-v72 RNA reference sequences using the FANSe2 algorithm,[16] with alternative-splicing variants merged.[15] Genes with at least 10 mapped reads were deemed adequate for identification and quantification.[17] Gene expression levels within a single cell line were calculated using the reads per kilobase per million mapped reads (RPKM) method.[18] The integrated mRNA abundance was the sum of the RPKM for the samples divided by the number of samples quantified.

Bioinformatic Analyses of Protein-Coding Genes on Chromosome 1

The ID map from Swiss-Prot accession to Ensembl gene ID was done using the BioMart data mining tool (http://www.

dx.doi.org/10.1021/pr400900j | J. Proteome Res. 2014, 13, 114−125

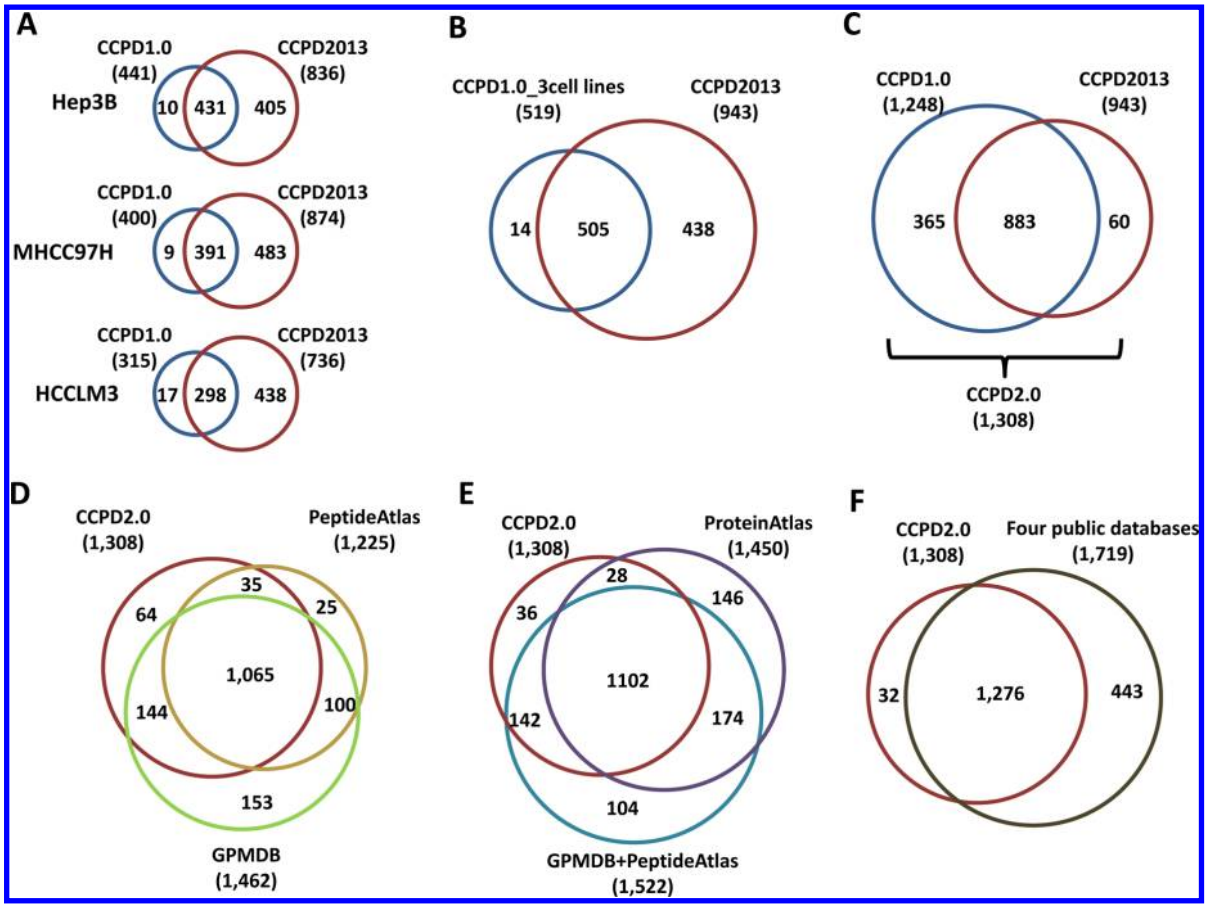


Figure 1. Summary of protein-coding genes on chromosome 1 identified by proteome. (A) Venn diagram of CCPD 2013 and CCPD 1.0 results for each cell line. (B) Venn diagram of CCPD 2013 and CCPD 1.0 results for all three cell lines. (C) Venn diagram of CCPD 2013 and CCPD 1.0. CCPD 2.0 was created as the union of these two data sets. (D) Venn diagram of CCPD 2.0 and two public data sets containing MS evidence: PeptideAtlas and GPMDB. (E) Venn diagram of CCPD 2.0, ProteinAtlas, and the combination of GPMDB and PeptideAtlas. (F) Venn diagram of CCPD 2.0 and four public data sets, including PeptideAtlas, GPMDB, ProteinAtlas, and neXtProt gold.

ensembl.org/info/data/biomart.html). Four public databases, the global proteome machine database (GPMDB,[19] released 2013_7_1), PeptideAtlas[20] (released 2012_12), Human ProteinAtlas (HPA,[21] version 11.0), and neXtProt[22] (released 2013_6_11), were chosen for the analysis of missing proteins on chromosome 1. Detailed information regarding the processing of data from public databases can be found in our companion article.[8]

Analysis of Missing Proteins on Chromosome 1

Tissue specificity was analyzed using the SP TISSUE information in DAVID.[23] Only terms with a p value < 0.05 were retained, and genes associated with only one term were designated as tissue-specific. TMHMM 2.0[24] was used to predict transmembrane helices. Proteins with regions of at least 18 amino acids in transmembrane helices were predicted to be transmembrane proteins. PeptideSieve[25] was employed to predict the proteotypic features of relevant proteins.

Simulation of Peptide Sampling in the Large-Scale Proteome Profiling Experiments

The proteins encoded by genes that could be quantified in the transcriptome or translatome analyses of the three cell lines were theoretically digested into peptides using trypsin. The RPKM value for each gene was inherited by all of its peptides as the weight of random sampling.

$$P_{ij} = \frac{\text{int}(\text{RPKM}_j)}{\sum_{i=1}^{N}\sum_{j=1}^{n}\text{int}(\text{RPKM}_j)}$$

The formula shows the probability of choosing peptide$_j$ in protein$_i$ by random sampling, where N represents the total number of proteins, n represents the number of peptides derived from protein$_i$, and int(RPKM$_j$) represents the nearest integer of the RPKM value of protein$_j$. After sampling, the selected peptide and the corresponding protein were considered to be the "identified peptide" and "identified protein", respectively. By increasing the number of random samplings, a large-scale proteome analysis can be simulated.

Definition of SDE Genes Using Three Omics-Integrated Data

To determine which proteins are differentially expressed based on the proteomics data, we used the significant B algorithm[26] to calculate the significance between two samples. Proteins with a p value < 0.01 were considered to be the products encoded by the SDE genes in subsequent analyses.

For transcriptome and translatome data across the three cell lines, differential expression levels were determined using the edgeR software package, which employs the trimmed means of M-value method based on the negative binomial distribution.[27] Those mRNAs or RNC-mRNAs with a p value < 0.01 were considered to be SDE in subsequent analyses.

dx.doi.org/10.1021/pr400900j | J. Proteome Res. 2014, 13, 114−125

Table 1. Number of Genes/Proteins in the Five Databases and Their Corresponding Identification Results for Chromosome 1[a]

	no. of genes				no. of proteins
	Ensembl	GPMDB[b]	PeptideAtlas	HPA	neXtProt[c]
database	2071	1462	1225	1450	1545
CCPD 1.0	1248	1157	1055	1086	1164
CCPD 2013	943	907	890	869	909
CCPD 2.0	1308	1209	1110	1130	1210

[a]Five databases and their versions are Ensembl (v72, released June 2013), GPMDB (released July 2013), PeptideAtlas (released December 2012), ProteinAtlas (HPA, v. 11.0), and neXtProt (released June 2013). [b]Data set annotated as "Green" was used, with the threshold ">20 Observations and log(e) < −5". [c]Data set annotated as "Protein Evidence" was used.

Table 2. Number of Proteins/Genes in the Five Databases and Their Corresponding Identification Results for Chromosome 1, based on the C-HPP Master Table Guidelines[a]

	all proteins		validated proteins				nonvalidated proteins	
	Ensembl	neXtProt	HPA[b]	PeptideAtlas[c]	GPMDB[d]	neXtProt PE1[e]	neXtProt PE5[e]	"Missings"[f]
database	2059	2061	988	1415	1523	1600	49	412
CCPD 1.0	1263	1252	749	1150	1169	1188	6	58
CCPD 2013	946	930	597	909	914	919	1	10
CCPD 2.0	1322	1310	777	1197	1223	1238	7	65

[a]Five databases and their versions are Ensembl (v72, released June 2013), neXtProt (released September 2013), ProteinAtlas (HPA, v. 11.0), PeptideAtlas (released Augest 2013), and GPMDB (released October 2013). [b]Data set with HPA evidence annotated as "high" or "medium" was used. [c]Data set with protein FDR < 1% was used. [d]Data set annotated as "Green" was used. [e]Protein evidence level used in neXtProt. PE1 is "protein level" and PE5 is "uncertain". [f]Proteins with protein evidence level between 2 and 4.

Confirmation of SDE Genes in HCC Samples Using Public Databases

To confirm the identity of SDE genes discovered in the three-omics data, we utilized LiverAtlas,[28] a comprehensive resource of biomedical knowledge related to the liver and various hepatic diseases that was developed at the BPRC. LiverAtlas incorporates three primary databases (EHCO,[29] OncoDB,[30] and HCC.net[31]) pertaining to hepatic diseases and assigns reliability scores (RSs) to entries for liver-expressed genes and proteins, protein−protein interactions, post-translational modifications, and molecular/genetic events of hepatic diseases. Genes with larger RS values in HCC samples have a higher probability of being SDE.

Genes present in at least two databases of EHCO, OncoDB, and HCC.net and with an RS above 3 in LiverAtlas were analyzed further. On the basis of these filtering criteria, 179 confident SDE genes on chromosome 1 were selected from LiverAtlas.

MetaCore (GeneGo, http://portal.genego.com) was also utilized for the detailed pathway enrichment analysis.

RESULTS AND DISCUSSION

Current Identification Status of Protein-Coding Genes on Chromosome 1

Human chromosome 1 contains 2071 protein-coding genes based on Ensembl annotation (Release-72), and there are 2060 protein entries for chromosome 1 in the Swiss-Prot database. In CCPD 1.0, we identified 1252 Swiss-Prot entries, corresponding to 1227 Ensembl genes.[3] In this study, using high-resolution and high-sensitivity mass spectrometry, we identified 943 genes from the three HCC cell lines (designated CCPD 2013). Over 98% of these genes also had both mRNA and RNC-mRNA evidence, resulting in the identification of 1703 and 1669 genes, respectively, by RNA-Seq (Supplemental Figure S1A in the Supporting Information). Compared with CCPD 1.0, the large-scale proteome profiling of the three cell lines in the present study contributed protein expression evidence of an additional 59 genes on chromosome 1.

For the TF-enrichment experiments, a total of 194 genes were identified, including genes for 6 transcription factors.[32] The *ATF3* gene, which encodes an activating transcription factor, was uniquely contributed by this experiment (Supplemental Figure S1B in the Supporting Information). The abundance of corresponding mRNAs and RNC-mRNAs of the 943 genes for which proteins were identified was distributed in the high RPKM range (Supplemental Figure S1C,D in the Supporting Information). The proportion of genes for which proteins were identified increased as the RPKM values of the mRNAs or RNC-mRNAs increased, which was consistent with several previous studies on human cell lines.[33,34]

In comparison with CCPD 1.0, 405, 483, and 438 more genes were identified in Hep3B, MHCC97H, and HCCLM3, respectively, which resulted in a total contribution of 438 (45.8%) more genes for the three cell lines in CCPD 2013 (Figure 1A,B). The final number of protein-coding genes on chromosome 1 with high-confidence mass spectrometric evidence in CCPD 2.0 (CCPD 1.0 + CCPD 2013) is thus 1308 (Figure 1C).

Combining the data from CCPD 2.0 with the data contained in the public databases GPMDB, PeptideAtlas, ProteinAtlas, and neXtProt indicated that a total of 1586 genes on chromosome 1 (76.6%) have been confidently identified with mass spectrometric evidence. A total of 67.2% of these genes were present in all three databases (CCPD 2.0, GPMDB, and PeptideAtlas) (Figure 1D). Of the 1308 genes identified in CCPD 2.0, 1130 (86.4%) were found with antibody evidence in ProteinAtlas, and CCPD 2.0 provided mass spectrometric evidence for an additional 28 genes that previously had only antibody evidence (Figure 1E). Overall, the five databases contained 1751 (84.6%) of the genes located on chromosome 1, with an additional 32 genes contributed by CCPD 2.0 (Figure 1F). The baselines and identified genes/proteins for chromosome 1, as contained in the Ensembl, GPMDB,

dx.doi.org/10.1021/pr400900j | J. Proteome Res. 2014, 13, 114−125

PeptideAtlas, ProteinAtlas, and neXtProt databases, are shown in Table 1. Following the guideline of the C-HPP Master Table, neXtProt was used as the standard to unify the number of public database and experimental data sets. The aligned results are shown in Table 2.

Analysis of Missing Proteins on Human Chromosome 1

To clarify the status of protein identifications for the coding genes on chromosome 1, we integrated the data from CCPD 2.0 with the data in the four publicly available databases, GPMDB, PeptideAtlas, ProteinAtlas, and neXtProt. For genes located on chromosome 1, we defined missing proteins as gene products with no confident evidence in any of the five databases. A flowchart illustrating the missing protein analysis is shown in Figure 2. All of the missing genes were cataloged

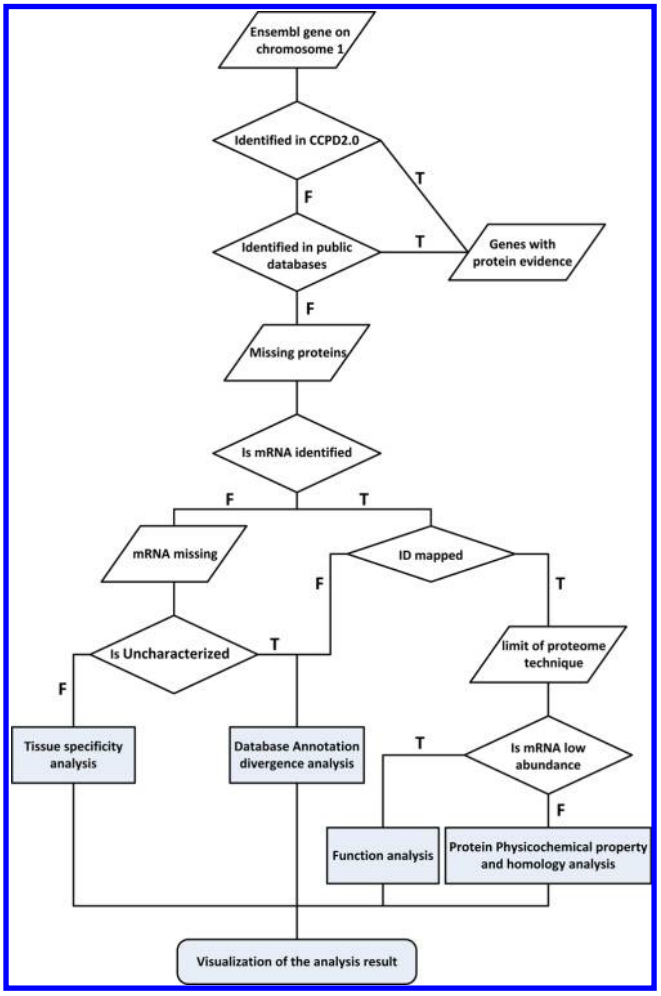


Figure 2. Flowchart illustrating the analysis of missing proteins based on the integration of information contained in five protein databases and the abundance of mRNAs.

comprehensively using mRNA abundance data, Ensembl annotations, and protein physicochemical properties as references, which helped us comprehend the characteristics of the missing genes and design appropriate strategies to acquire protein evidence for these missing genes.

As indicated in Figure 3A, based on the annotations in Ensembl v72, there are a total of 2071 protein-coding genes on chromosome 1, of which 1751 (84.6%) were found to have high-confidence protein evidence in at least one of the five databases. A total of 1607 (77.6%) of the genes on chromosome 1 were found with evidence in at least two of the five databases (Supplemental Figure S2A in the Supporting Information). For CCPD 2.0, 1308 (63.2%) genes were included, of which 32 (1.6%) were uniquely present in CCPD 2.0, with no confident evidence in any of the four public databases. Of the proteins produced from the 32 genes unique to CCPD 2.0, 29 were identified based on at least two peptides, and the average number of identified peptides was 6.7. For the three proteins with only one peptide, manual interpretation of the spectra showed that the data quality could be guaranteed; the spectra for these peptides are presented in Supplemental Figure S3 in the Supporting Information.

In total, 320 (15.5%) genes remained unidentified in the five databases. We therefore checked the corresponding abundance of mRNAs and RNC-mRNAs for these 320 genes and confirmed that 177 of the genes were not actively expressed because they had either no mRNA or no RNC-mRNA evidence (read count 0−10). Of the remaining 143 genes, 42 could not be identified because no corresponding protein ID was present in the Swiss-Prot database, leaving only 101 genes that could potentially be identified because they had both mRNA and RNC-mRNA evidence.

Analysis of Missing Proteins Lacking mRNA Evidence. On the basis of the gene location information and the annotations derived from the Ensembl database, three gene clusters were found among the 177 genes lacking RNA evidence. The first cluster included an olfactory receptor family with 57 genes located adjacent to one another on p36, q23, and q44. These genes have remained unidentified in many recent studies.[5] The other two clusters included PRAME, with 11 genes located on p36, and a cluster of keratin-binding proteins, with 13 genes located on q21. These three blocks of genes (olfactory receptor family, PRAME, and keratin-binding proteins) were also unidentified in CCPD 1.0, further confirming the missing expression of these genes.[3]

In addition, 20 tissue-specific genes annotated by DAVID could be found only in tissues such as the testis, brain, and skeletal muscle (Figure 3B). In total, 128 of 177 genes lacking mRNA evidence were found to be tissue-specific or at least not expressed in the liver. Aside from these genes, 49 uncharacterized genes were found, including 31 genes with no protein ID mapped in the Swiss-Prot database and 18 open reading frame genes, the gene models and functions of which will require further confirmation.

Analysis of the Missing Proteins Due to the Limitations in Proteomic Techniques. Of the 101 genes with both mRNA and RNC-mRNA evidence, 59 had lower mRNA and RNC-mRNA abundances (RPKM < 1) (Figure 3C). On the basis of the annotations of these genes, 21 were found to be potentially present in a transient state in the liver. A total of nine of the genes were tissue-specific, such as the olfactory receptor OR10J1 and the lens epithelial protein LENEP, whereas four genes code for enzymes that may not be constitutively expressed, and eight genes encode scaffolding proteins to provide domains to connect other proteins. Aside from the genes with clearly defined functions, seven newly annotated genes with high sequence similarity to other genes and 11 uncharacterized genes were also found (Supplemental Table S1 in the Supporting Information).

The physical properties of the 42 genes with relatively higher mRNA or RNC-mRNA abundance were further analyzed to

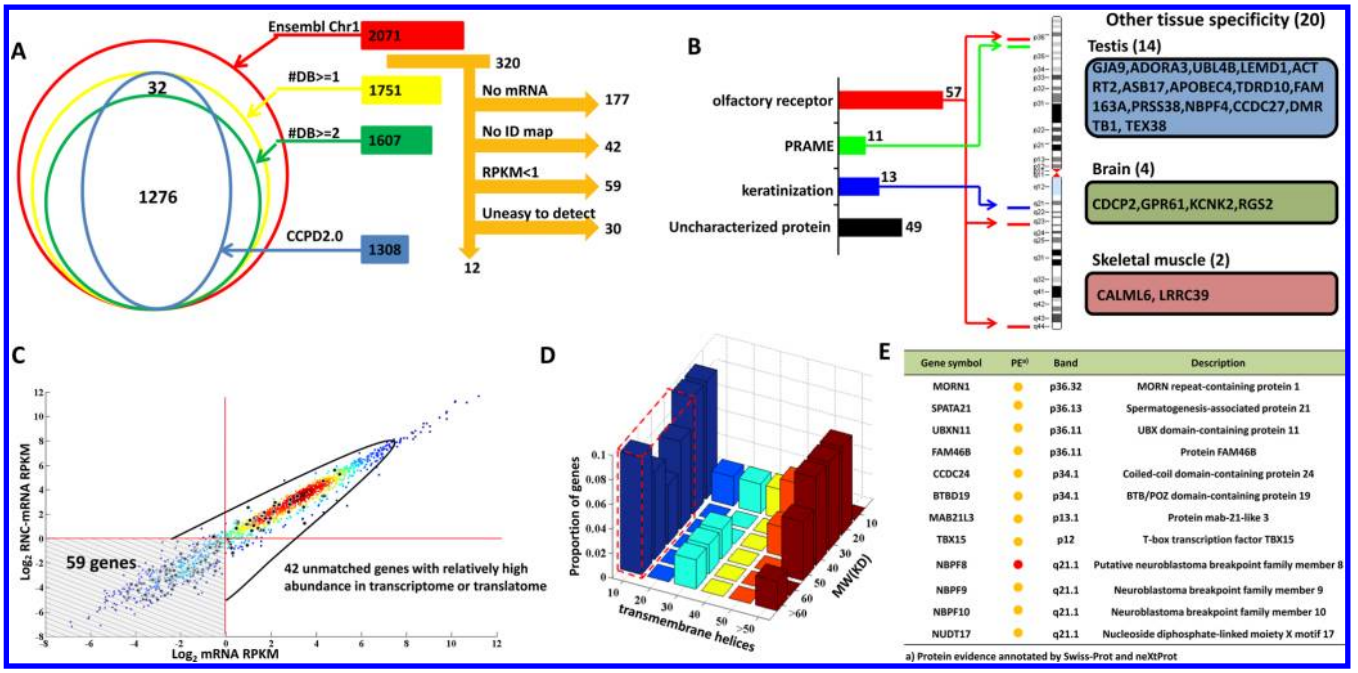


Figure 3. Analysis of the properties of "missing proteins" on human chromosome 1. (A) Overview of the identification evidence for chromosome 1 genes contained in GPMDB, PeptideAtlas, ProteinAtlas, neXtProt (protein evidence), and CCPD 2.0. (B) Analysis of functional clusters and tissue specificity of 177 genes with either no mRNA or RNC-mRNA (ribosome nascent-chain complex-bound) evidence. (C) RPKM of mRNAs and RNC-mRNAs of the remaining 101 genes with no identification evidence available in the four public databases and CCPD 2.0. (D) Transmembrane helices and molecular weight distribution of the products of 42 of the 101 genes with no identification evidence. These genes have mRNA or RNC-mRNA RPKM values larger than 1. (E) "PE" evidence and functional annotation of 12 of the 42 genes described in panel D. These genes have relatively higher mRNA or RNC-mRNA abundance (RPKM > 1), and their protein products were found to have molecular weights above 30 kDa and were predicted to be nonmembrane proteins.

determine the possible reasons why the proteins remain unisolated. As expected (Figure 3D), a majority of these gene products were found to be low-molecular-weight (MW < 30 kDa) transmembrane proteins (transmembrane helices > 18). However, 12 genes that encode nontransmembrane proteins with a molecular weight larger than 30 kDa remained. In contrast with the proteomics data, most of these genes were found to have transcription evidence in both Swiss-Prot and neXtProt (Figure 3E), which is consistent with our observation. Our analysis showed that the products of 10 of the 12 genes had at least three unique proteotypic peptides that might be suited to identification using an SRM/MRM approach (Supplemental Figure S2C,D in the Supporting Information). The information we already obtained suggests that these 12 proteins may have been present in our samples but somehow missed in our data sets. These results provided potential clues for further studies.

Overview of Missing Gene Data Suggesting Strategies for Obtaining Protein Evidence. On the basis of the detailed analyses previously described, overviews of the 320 missing genes on chromosome 1 are provided in Figure 4A and Table S1 in the Supporting Information. Reflecting the capability to detect human proteins, all of the missing genes can be clustered into three categories (missing mRNA, database annotation divergence, and limitations in proteomic techniques) (Figure 4A), according to the indexes of mRNA and RNC-mRNA abundance, database annotations, and protein properties (Figure 4B).

Tissue specificity was the major cause of missing mRNA; even though not all of the 128 genes can be clearly located in specific tissues, they are hardly ever expressed in the liver. In addition, among the tissue-specific proteins expressed by these genes, 104 (84.4%) were found to have a molecular weight of < 30 kDa or to have transmembrane helices with more than 18 amino acids, which may be why these proteins are not present in the public databases (Figure 4A and Supplemental Figure S2B in the Supporting Information). The missing mRNAs prompted us to conduct deep-sequencing transcriptome and translatome analyses in selected tissues before the proteomic analysis.

Of 91 uncharacterized genes, 18 are chromosome 1 open reading frame genes with no clear function, and 73 are genes with no ID mapped in the Swiss-Prot database. Of the 18 genes with associated proteins, 16 are annotated as "transcript level" and the other two as "predicted level" in neXtProt, and all of their proteins are putative and have not been validated. A detailed investigation of the 73 genes with no ID mapped in Swiss-Prot showed that 61 (83.6%) are clone-based genes unique to the Ensembl database. Three of the other 12 are read-through genes that encompass known annotated genes that were identified by CCPD 2.0 and are in public databases, 6 genes are included in UniProt but not reviewed by Swiss-Prot, 2 are pseudogenes, and 1, *TRAX*, is the same as the *TSNAX* gene (Supplemental Table S2 in the Supporting Information). Note that because no confident protein evidence was found by CCPD 2.0 and none is contained in the four public databases, reannotation and verification for the models of these 73 genes are necessary.

The remaining 101 genes were unidentifiable mainly due to limitations in current proteomic techniques. Previous analyses indicated that 59 genes had low mRNA abundance, and a high proportion of the remaining 42 gene products were of low

dx.doi.org/10.1021/pr400900j | J. Proteome Res. 2014, 13, 114−125

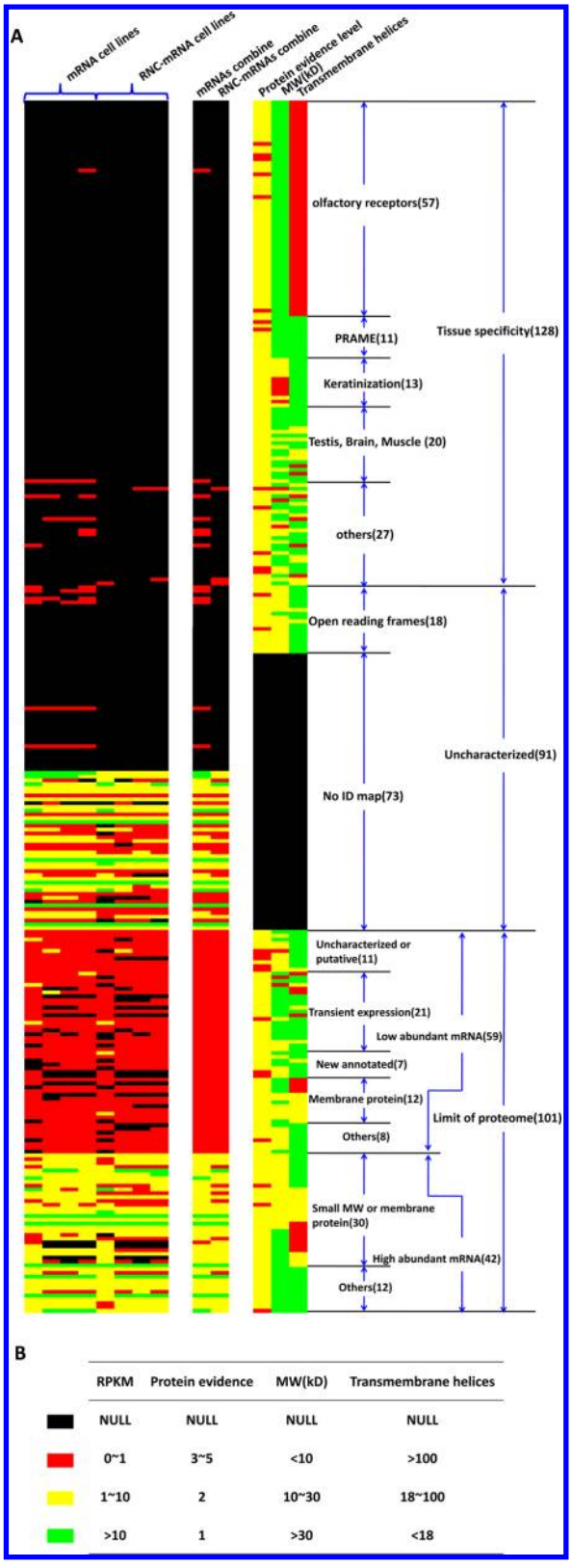


Figure 4. Overview of the analysis results for the 320 missing genes on chromosome 1. (A) Classification and presentation of 320 missing genes based on mRNA abundance, protein evidence level, and protein properties. (B) Legend for gene properties in four different levels.

molecular weight or had a high number of transmembrane helices. To identify the proteins encoded by the genes with low mRNA abundance, it is necessary to employ enrichment strategies, such as the transcription factor profiling that we applied in this study. For low-molecular-weight or highly hydrophobic proteins, special sample separation methods must be developed, and for proteins with highly abundant transcripts and detectable proteotypic peptides, SRM/MRM is an option.

In summary, of the 320 genes on chromosome 1 for which proteins have not been isolated, ~40% (corresponding to ~6.2% of the 2071 total genes) were found to be tissue-specific or only rarely expressed in some tissues, 28.4% encode proteins uncharacterized primarily because of missing ID maps, and the remaining genes are characterized by low mRNA abundance or encoding proteins with unsuited properties. Continued analyses of varied human tissues or different cell types via large-scale proteomic studies may increase the number of identified proteins encoded by genes on chromosome 1; however, the benefits and costs of such studies need to be seriously evaluated.

Approaching the Sample Number-Independent Data Saturation Stage of Protein Identification in Large-Scale Proteome Profiling

With the development of highly sensitive and high-resolution mass spectrometers, such as the Q Exactive and triple-TOF 5600, much deeper coverage of the mammalian proteome has been achieved, as approximately 8000 to 10 000 gene products were identified in a single human cell line.[28,29] With the accumulation of proteome data sets generated from different tissues and cell lines, to date, 50−60% of protein-coding genes have been identified at the proteome level.[3,35] However, because of limitations associated with the detection sensitivity and dynamic range of available mass spectrometers, data saturation could be easily reached. Therefore, the number of identified proteins can only increase slowly, even with the recent significant augmentation of large-scale data sets generated from proteomic studies and rapid advances in instrumentation and experimental technologies. For example, using the genes identified by the combination of transcriptome and translatome analyses as the maximum number expressed in the three liver cell lines we examined, a simple simulation showed that the maximum number of proteins would be 14 326, even if more data sets were included (Figure 5A). Similar data saturation was also observed during CCPD when more samples were included. As indicated in Figure 5B, the curve of the number of accumulated proteins identified by CCPD 2.0 increasing with the number of accumulated peptides was consistent with the simulation results. With the increase in the number of identified proteins, the abundance of mRNAs for the remaining unidentified genes appeared to decline when using the average RPKM of mRNAs from the three human liver cell lines to roughly represent the average abundance of transcripts. A similar phenomenon was also observed for genes located on chromosome 1 (Supplemental Figure S4 in the Supporting Information).

We also determined the average number of newly added proteins per added peptide during the data accumulation. As expected, with the increase in the number of integrated data sets, the average number of newly added proteins per added peptide also reached a saturation level based on results obtained with either simulated or real data sets. There was an exception, however. When the data accumulation nearly reached the saturation phase, the average number of newly added proteins per added peptide in the TF enrichment data set was greater than that in other large-scale proteome data sets (Figure 5C).

dx.doi.org/10.1021/pr400900j | J. Proteome Res. 2014, 13, 114−125

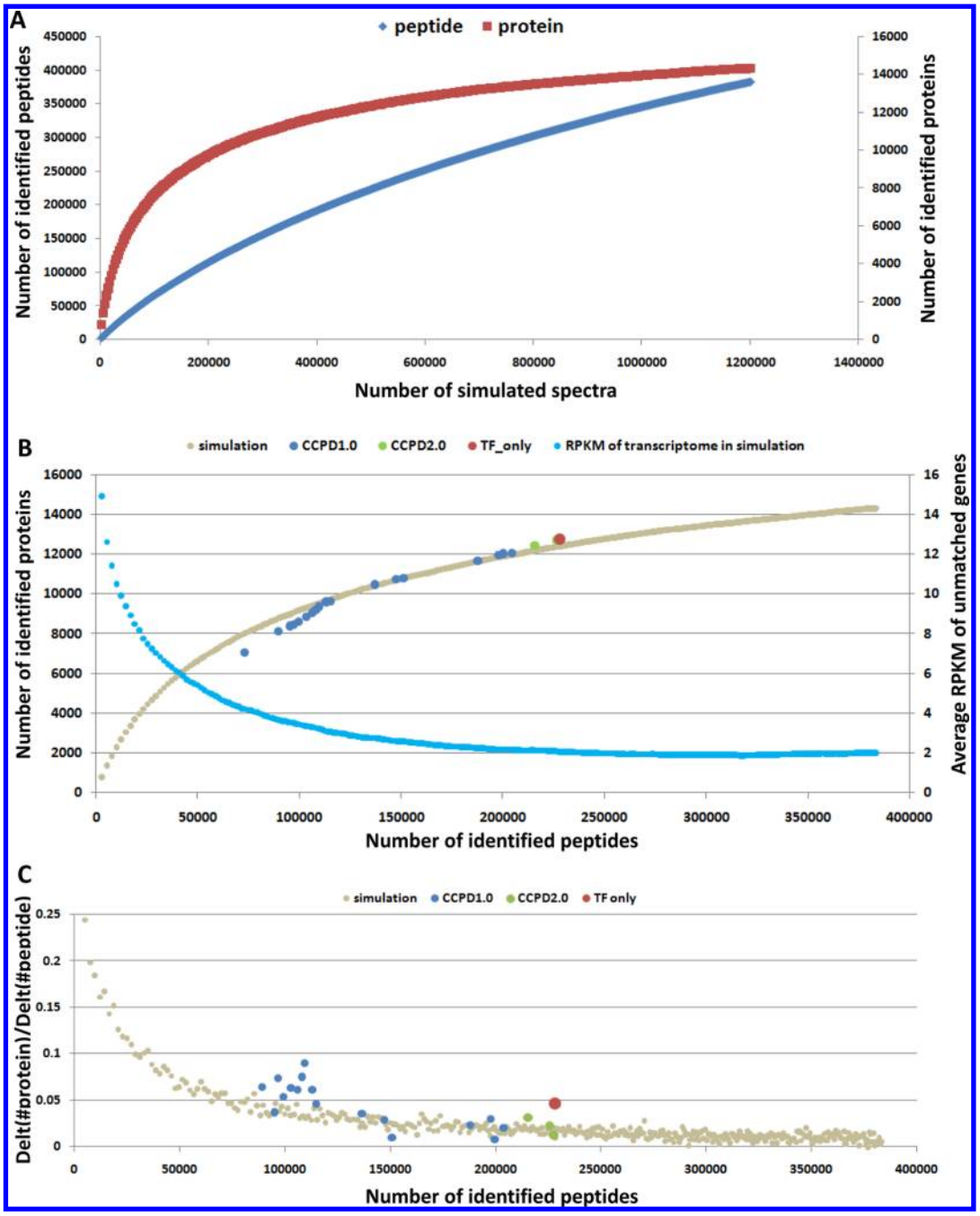


Figure 5. Comparison of the accumulated peptides and proteins in the CCPD data set and simulation data. (A) Number of accumulated identified proteins (red line) and peptides (blue line) with the accumulated sampling times (the number of spectra) derived from simulation results. (B) Number of accumulated proteins and the corresponding RPKM values of unidentified genes with the number of unique identified peptides. (C) Relative number of added proteins per added peptide as the identified peptides accumulate. Brown dots indicate the simulation results; blue dots indicate accumulated proteins identified with each of the 18 samples analyzed by CCPD 1.0 added; green dots indicate accumulated proteins identified with each of three liver cell lines added; red dots indicate data from transcript factor enrichment experiments. During simulations, genes found in the transcriptome or translatome with read counts larger than 10 were used as the sampling pool. The RPKM values were used as the sampling frequency weight. Genes with a high RPKM would have a greater chance of being identified. See the Materials and Methods for more details.

These results indicated that compared with regular large-scale proteomic analyses selection of different approaches based on known information, such as enrichment of low-abundance proteins during sample preparation, might provide a solution to break through the data saturation bottleneck and supplement mass spectrometric evidence for missing proteins. Therefore, targeted proteomics based on specific sample preparation techniques along with MS detection could play a significant role in achieving full coverage of all gene products in the C-HPP study.

Analysis of SDE Genes in Three Human Liver Cell Lines

In this study, we combined the three omics data sets generated from analyses of three human hepatic cell lines to identify the SDE genes on chromosome 1. All genes with a p value < 0.01

dx.doi.org/10.1021/pr400900j | J. Proteome Res. 2014, 13, 114−125

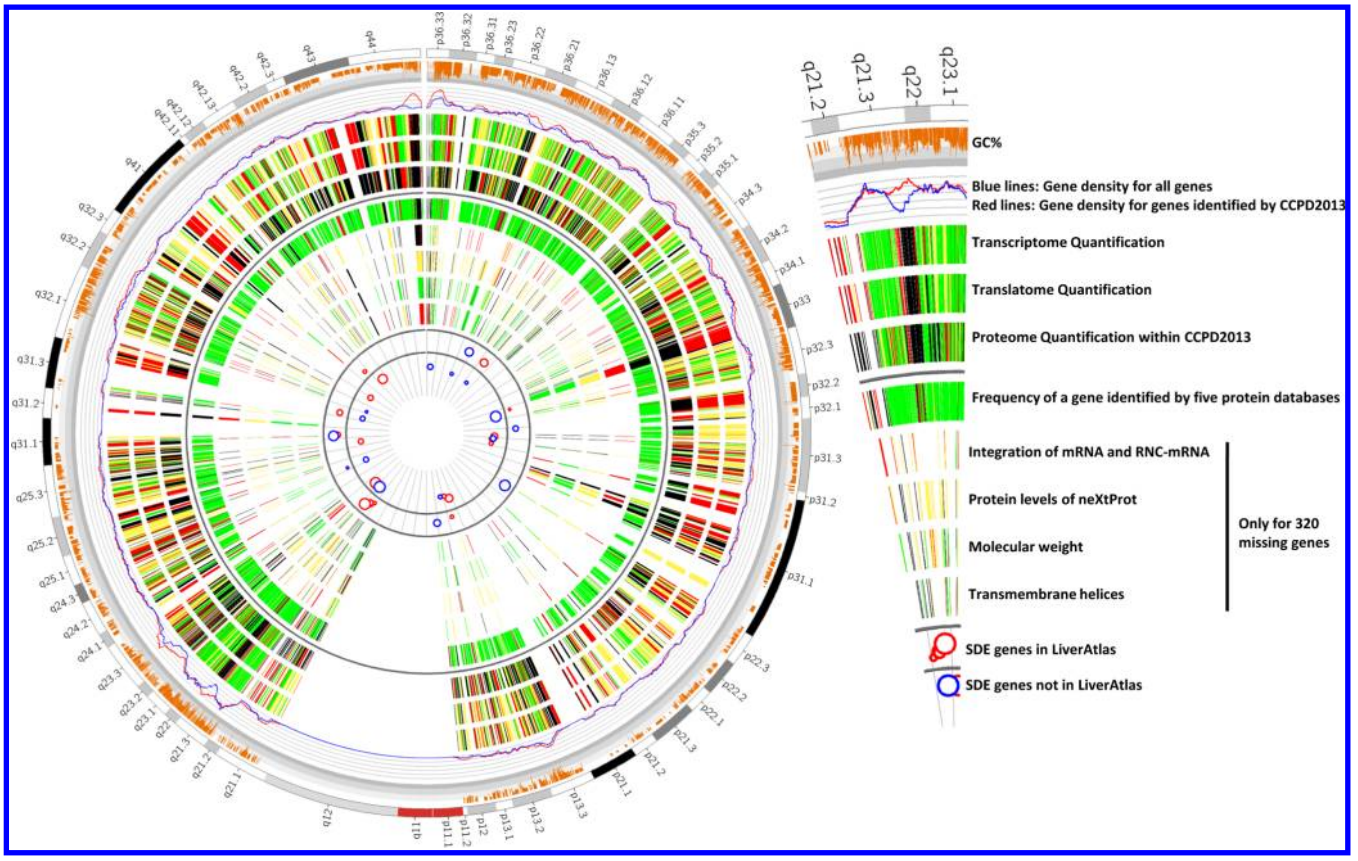


Figure 6. Visualization of the detailed identification and quantification profiles as well as the distribution of SDE genes on chromosome 1. The quantitative results from the three omics data sets were divided into four levels. Green represents the high value (mRNAs and RNC-mRNAs: RPKM > 10; proteome: iBAQ > upper tertile), yellow represents the middle value (mRNAs and RNC-mRNAs: RPKM between 1−10; proteome: iBAQ between upper tertile and lower tertile), red represents the low value (mRNAs and RNC-mRNAs: RPKM between 0−1; proteome: iBAQ < lower tertile), and black represents unidentified genes. The frequency of a gene being identified in the five protein databases was also divided into four levels. Green represents genes present in all five databases, yellow represents genes present in any four databases, red represents genes present in no more than three databases, and black represents genes not present in any of the five databases. For the integration of mRNA and RNC-mRNA quantification data, green represents genes with RPKM values of mRNAs or RNC-mRNAs larger than 10, red represents genes with RPKM values of mRNAs and RNC-mRNAs lower than 1, yellow represents the remaining identified genes, and black represents unidentified genes. The color scheme used to represent neXtProt evidence, molecular weight, and transmembrane helix data is the same used in Figure 4B. For the SDE genes, red represents up-regulation in the cell lines of MHCC97H or HCCLM3 and blue represents down-regulation in the cell lines of MHCC97H or HCCLM3.

were extracted, and Venn diagrams for MHCC97H versus Hep3B and HCCLM3 versus Hep3B are shown in Supplemental Figures S5A and S5B in the Supporting Information. As compared with Hep3B, 140 SDE genes on chromosome 1 were discovered in the proteome data for MHCC97H cells. In total, 43 and 44 SDE genes were confirmed by analyses of the transcriptome and translatome data, respectively. In addition, 291 SDE genes from HCCLM3 were found in the proteome data, 57 and 58 of which were confirmed by analyses of the transcriptome and translatome data, respectively.

To improve the quality of SDE gene identifications based on our data sets, we compared the SDE genes on chromosome 1 that were picked up based on proteome and transcriptome or proteome and translatome data with the 179 confident entries of SDE genes in LiverAtlas. As a result, 50 and 64 SDE genes were found in comparisons of MHCC97H versus Hep3B and HCCLM3 versus Hep3B, respectively, for a total of 72 SDE genes (Supplemental Table S3 in the Supporting Information). Sixteen of these genes were covered by LiverAtlas (Supplemental Figure S5C in the Supporting Information). GO enrichment analysis of these 16 genes showed that many are involved in biology processes such as cell proliferation, regulation of apoptosis, and regulation of cell death, which is consistent with involvement in hepatic disease.

Additionally, of the 34 genes not found in LiverAtlas but identified as SDE genes through comparisons of both HCCLM3 and MHCC97H with Hep3B, 20 genes were found to have no evidence in LiverAtlas. Pathway analysis with MetaCore indicated that many of these genes are involved in the regulation of protein transport, insulin receptor, apoptosis, or cell adhesion network, all of which are important with respect to the occurrence and development of hepatic diseases. Among the 20 genes with no evidence in LiverAtlas, the expression levels of S100A2, S100A3, S100A4, S100A6, and S100A10 of the S100A subfamily were found to be significantly up-regulated in both HCCLM3 and MHCC97H compared with Hep3B. Interestingly, members of the S100A subfamily contain separate domains for EF-hand CaBPs and calcium-binding proteins to balance calcium signaling and sustain intracellular and extracellular homeostasis in vertebrates. The amounts of these proteins were increased in highly metastatic

dx.doi.org/10.1021/pr400900j | J. Proteome Res. 2014, 13, 114−125

cell lines, suggesting that the genes play critical roles in mediating invasion and metastasis in HCC.[36]

The higher expression levels of S100A subfamily proteins in MHCC97H and HCCLM3 could enhance the ability of these cells to bind intracellular free Ca^{2+} $[Ca^{2+}]_i$, which could speed up cell proliferation and cell cycle by activating the PLC-β/IP3/PKC pathway. This also might result in activation of cytoskeleton remodeling processes through the generation of gelsolin and profilin, which are involved in binding actin filaments and promotion of malignant transformation in HCC.[37]

We also found a close relationship between metastasis potential and cholesterol synthesis. The HMGCS2 expression level in HCCLM3 was significantly lower than that in Hep3B and MHCC97H. This cholesterol-synthesis enzyme belongs to the HMG-CoA synthase family. Hypercholesterolemia is an important factor in cancer development and metastasis. For example, stimulation of cholesterol synthesis leads to PI3K/Akt phosphorylation.[38] Cholesterol is involved in the formation of lipid rafts and interacts with the epithelial-extracellular matrix in mediating tumorigenesis, invasion, and metastasis.

Circos Visualization, Database Construction, and Data Download for Transomics Data

The transcriptome, translatome, and proteome identification and quantification profiles of the HCC cell lines examined in this study are vividly shown in Circos view[39] (Figure 6). From the gene density circular distributions and the quantification distributions of the three-omics data, we found several missing regions in bands 1p36.33, 1p36.21, 1q21.3, and 1q23.2, identical to our previous findings.[3] The discrete category of quantitative results could be better to show the consistency of the three-omics data. A summary of information regarding the 320 missing genes is also shown in Circos to present the relationship between protein properties and gene location on chromosome 1. Annotated information for the genes on chromosome 1 derived from public resources and the results of our CCPD 2.0 study can also be found in Supplemental Table S4 in the Supporting Information.

As shown in the inner circles of Figure 6, we added SDE genes on chromosome 1 for the three HCC cell lines we examined, supplementing the list of HCC-related genes in LiverAtlas.

All of the proteome data described in this study were included in the Chinese Chromosome-Centric Human Proteome Database,[3] which was developed for collecting the Chinese C-HPP proteome data. These data can be visualized using ProteomeView (http://proteomeview.hupo.org.cn/chromosome.jsp). The RNA-Seq data are accessible from GEO (Accession number: GSE49994), and the proteomics data are downloadable from ProteomeXchange (http://proteomecentral.proteomexchange.org) using the identifiers PXD000529, PXD000533, and PXD000535[40] and are also available from iProX (http://www.iprox.org), a uniform platform used for collection, storage, and sharing of raw MS files and experimental metadata.

CONCLUSIONS

Using deep-sequencing technology, in this study, we generated transcriptome, translatome, and proteome data sets for Hep3B, MHCC97H, and HCCLM3 cell lines and provided new protein evidence for 60 genes on chromosome 1 in addition to those identified in the CCPD 1.0 study. On the basis of public annotations and CCPD 2.0 as updated in this study, we conclude that 84.6% of the protein-coding genes on chromosome 1 that have been annotated to date have confident protein evidence of their expression. Detailed analyses of the unseen genes revealed that the majority are tissue-specific and might not have been present in the cells used in this study. Using mRNA and RNC-mRNA data for reference, unidentified genes with abundant transcripts and proteotypic peptides could be chosen first for identification using SRM/MRM methods. Furthermore, both simulation analyses and real experimental data sets indicated that combined large-scale data sets would be already saturated; therefore, increases in the number of identified proteins could only be made slowly, even if more large-scale data sets were incorporated. Because of the relatively large contribution made by TF-enrichment experiments, a better approach for the C-HPP would be to focus on low-abundance proteins derived from specific tissues or enriched for using targeted proteomics technologies. We contend that some genes may appear to be unexpressed due to overannotation or incorrect prediction, leading to their omission from the public databases and our CCPD results. Further studies will also focus on proteogenomic analyses of missing proteins that have "unknown" functional annotations but highly abundant transcripts. Our analysis of the metastatic potential of SDE genes in different cell lines demonstrated that several SDE genes located on chromosome 1 might play critical roles in mediating HCC invasion and metastasis. These data might be valuable for mechanistic studies of HCC-related genes.

ASSOCIATED CONTENT

Supporting Information

Comparison of protein coding genes identified on chromosome 1 between transcriptome, translatome, and proteome. Supplementary analysis of "missing proteins" on human chromosome 1. Spectra of the peptides that belong to the proteins with only one peptide identified in CCPD 2.0 uniquely. Relatively added proteins per added peptide when the identified peptides accumulated on chromosome 1. Venn diagram of SDE genes among three OMICS data sets and LiverAtlas on chromosome 1. Details of 320 missing genes on chromosome 1. Supplementary analysis of the genes with no ID mapped to Swiss-Prot protein and not unique in Ensembl database. SDE gene list generated by combining the data of three omics datasets in our experiments and LiverAtlas. Table for visualization contains the public annotation and the rough abundance of proteome, transcriptome, and translatome. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

*Songfeng Wu: Tel: 8610-80705225. E-mail: wusf897@gmail.com.
*Tong Wang: Tel/Fax: 8620-85227039. E-mail: tongwang@jnu.edu.cn.
*Huali Shen: Tel/Fax: 8621-54237961. E-mail: shenhuali@gmail.com.
*Quanhui Wang: Tel/Fax: 8610-80485324. E-mail: wangqh@genomics.cn.
*Yunping Zhu: Tel: 8610-80705225. E-mail: zhuyunping@gmail.com.

*Fuchu He: Tel/Fax: 8610-68171208. E-mail: hefc@nic.bmi.ac.cn.
*Ping Xu: Tel/Fax: 8610-80705155. E-mail: xupingghy@gmail.com.

Present Address

○Quanhui Wang: Beijing Airport Industrial Zone b-6, Shunyi, Beijing 101318, China.

Author Contributions

▽Chengpu Zhang, Ning Li, Linhui Zhai, Shaohang Xu, Xiaohui Liu, and Yizhi Cui 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 the Beijing Proteome Research Center for providing reagents for TF enrichment analysis. This study was supported by the Chinese National Basic Research Program (2011CB910600, 2013CB911200, 2010CB912700, and 2011CB910700), the National High-Tech Research and Development Program (2011AA02A114, 2012AA020201, 2012AA020409, and 2012AA020502), the National Natural Science Foundation of China (31070673, 31170780, 81123001, 21105121, 21275160, 81322028, and 81372135), the National International Cooperation Project (2011-1001), Key Projects in the National Science & Technology Pillar Program (2012BAF14B00), Beijing Municipal Natural Science Foundation 5122013, the State Key Lab Project (SKLP-Y201102), and the Shenzhen Key Laboratory of Transomics Biotechnologies (CXB2011O8250096A).

■ ABBREVIATIONS:

C-HPP, Chromosome-Centric Human Proteome Project; RNC, ribosome-nascent chain; CCPD, Chinese Chromosome Proteome Database; MW, molecular weight; RPKM, reads per kilobase per million mapped reads; SDE genes, significantly differentially expressed genes; HCC, hepatocellular carcinoma; TF, transcription factor

■ REFERENCES

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

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

(3) Wu, S.; Li, N.; Ma, J.; Shen, H.; Jiang, D.; Chang, C.; Zhang, C.; Li, L.; Zhang, H.; Jiang, J.; Xu, Z.; Ping, L.; Chen, T.; Zhang, W.; Zhang, T.; Xing, X.; Yi, T.; Li, Y.; Fan, F.; Li, X.; Zhong, F.; Wang, Q.; Zhang, Y.; Wen, B.; Yan, G.; Lin, L.; Yao, J.; Lin, Z.; Wu, F.; Xie, L.; Yu, H.; Liu, M.; Lu, H.; Mu, H.; Li, D.; Zhu, W.; Zhen, B.; Qian, X.; Qin, J.; Liu, S.; Yang, P.; Zhu, Y.; Xu, P.; He, F. First proteomic exploration of protein-encoding genes on chromosome 1 in human liver, stomach, and colon. *J. Proteome Res.* **2013**, *12* (1), 67–80.

(4) Farrah, T.; Deutsch, E. W.; Hoopmann, M. R.; Hallows, J. L.; Sun, Z.; Huang, C.-Y.; Moritz, R. L. The State of the Human Proteome in 2012 as Viewed through PeptideAtlas. *J. Proteome Res.* **2012**, *12* (1), 162–171.

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

(6) Li, Y.; Tang, Z. Y.; Ye, S. L.; Liu, Y. K.; Chen, J.; Xue, Q.; Gao, D. M.; Bao, W. H. Establishment of cell clones with different metastatic potential from the metastatic hepatocellular carcinoma cell line MHCC97. *World J. Gastroenterol.* **2001**, *7* (5), 630–6.

(7) Tian, J.; Tang, Z. Y.; Ye, S. L.; Liu, Y. K.; Lin, Z. Y.; Chen, J.; Xue, Q. New human hepatocellular carcinoma (HCC) cell line with highly metastatic potential (MHCC97) and its expressions of the factors associated with metastasis. *Br. J. Cancer* **1999**, *81* (5), 814–21.

(8) Chang, C.; Li, L.; Zhang, C.; Wu, S.; Guo, K.; Zi, J.; Chen, Z.; Jiang, J.; Ma, J.; Yu, Q.; Fan, F.; Qin, P.; Han, M.; Su, N.; Chen, T.; Wang, K.; Zhai, L.; Zhang, T.; W., Y.; Xu, Z.; Zhang, Y.; Liu, X.; Zhong, F.; Shen, H.; Wang, Q.; Hou, G.; Zhao, H.; Li, G.; Liu, S.; Gu, W.; Wang, G.; Wang, T.; Zhang, G.; Qian, X.; Liu, Y.; Li, N.; He, Q.; Lin, L.; Yang, P.; Zhu, Y.; He, F.; Xu, P. Systematic analyses of the transcriptome, translatome, and proteome provide a global view and potential strategy for the C-HPP. *J. Proteome Res.* **2013**, DOI: 10.1021/pr4009018.

(9) Khan, M. S.; Knowles, B. B.; Aden, D. P.; Rosner, W. Secretion of testosterone-estradiol-binding globulin by a human hepatoma-derived cell line. *J. Clin. Endocrinol. Metab.* **1981**, *53* (2), 448–9.

(10) Li, Y.; Tian, B.; Yang, J.; Zhao, L.; Wu, X.; Ye, S. L.; Liu, Y. K.; Tang, Z. Y. Stepwise metastatic human hepatocellular carcinoma cell model system with multiple metastatic potentials established through consecutive in vivo selection and studies on metastatic characteristics. *J. Cancer Res. Clin. Oncol.* **2004**, *130* (8), 460–8.

(11) Ding, C.; Chan, D. W.; Liu, W.; Liu, M.; Li, D.; Song, L.; Li, C.; Jin, J.; Malovannaya, A.; Jung, S. Y.; Zhen, B.; Wang, Y.; Qin, J. Proteome-wide profiling of activated transcription factors with a concatenated tandem array of transcription factor response elements. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (17), 6771–6.

(12) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551–3567.

(13) Li, N.; Wu, S.; Zhang, C.; Chang, C.; Zhang, J.; Ma, J.; Li, L.; Qian, X.; Xu, P.; Zhu, Y.; He, F. PepDistiller: A quality control tool to improve the sensitivity and accuracy of peptide identifications in shotgun proteomics. *Proteomics* **2012**, *12* (11), 1720–5.

(14) Schwanhausser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. Global quantification of mammalian gene expression control. *Nature* **2011**, *473* (7347), 337–42.

(15) Wang, T.; Cui, Y.; Jin, J.; Guo, J.; Wang, G.; Yin, X.; He, Q. Y.; Zhang, G. Translating mRNAs strongly correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. *Nucleic Acids Res.* **2013**, *41* (9), 4743–54.

(16) Zhang, G.; Fedyunin, I.; Kirchner, S.; Xiao, C.; Valleriani, A.; Ignatova, Z. FANSe: an accurate algorithm for quantitative mapping of large scale sequencing reads. *Nucleic Acids Res.* **2012**, *40* (11), e83.

(17) Bloom, J. S.; Khan, Z.; Kruglyak, L.; Singh, M.; Caudy, A. A. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics* **2009**, *10*, 221.

(18) Mortazavi, A.; Williams, B. A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **2008**, *5* (7), 621–8.

(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.

(20) Deutsch, E. W. The PeptideAtlas Project. *Methods Mol. Biol.* **2010**, *604*, 285–96.

(21) Uhlen, M.; Oksvold, P.; Fagerberg, L.; Lundberg, E.; Jonasson, K.; Forsberg, M.; Zwahlen, M.; Kampf, C.; Wester, K.; Hober, S.; Wernerus, H.; Bjorling, L.; Ponten, F. Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **2010**, *28* (12), 1248−50.
(22) Lane, L.; Argoud-Puy, G.; Britan, A.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gaudet, P.; Gleizes, A.; Masselot, A.; Zwahlen, C.; Bairoch, A. neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res.* **2012**, *40* (Database issue), D76−83.
(23) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols* **2008**, *4* (1), 44−57.
(24) Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **2001**, *305* (3), 567−80.
(25) Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; Kuster, B.; Aebersold, R. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **2007**, *25* (1), 125−31.
(26) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367−72.
(27) Robinson, M. D.; McCarthy, D. J.; Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26* (1), 139−40.
(28) Zhang, Y.; Yang, C.; Wang, S.; Chen, T.; Li, M.; Wang, X.; Li, D.; Wang, K.; Ma, J.; Wu, S.; Zhang, X.; Zhu, Y.; Wu, J.; He, F. LiverAtlas: a unique integrated knowledge database for systems-level research of liver and hepatic disease. *Liver Int.* **2013**, *33*, 1239−1248.
(29) Hsu, C. N.; Lai, J. M.; Liu, C. H.; Tseng, H. H.; Lin, C. Y.; Lin, K. T.; Yeh, H. H.; Sung, T. Y.; Hsu, W. L.; Su, L. J.; Lee, S. A.; Chen, C. H.; Lee, G. C.; Lee, D. T.; Shiue, Y. L.; Yeh, C. W.; Chang, C. H.; Kao, C. Y.; Huang, C. Y. Detection of the inferred interaction network in hepatocellular carcinoma from EHCO (Encyclopedia of Hepatocellular Carcinoma genes Online). *BMC Bioinformatics* **2007**, *8*, 66.
(30) Su, W. H.; Chao, C. C.; Yeh, S. H.; Chen, D. S.; Chen, P. J.; Jou, Y. S. OncoDB.HCC: an integrated oncogenomic database of hepatocellular carcinoma revealed aberrant cancer target genes and loci. *Nucleic Acids Res.* **2007**, *35* (Database issue), D727−31.
(31) He, B.; Qiu, X.; Li, P.; Wang, L.; Lv, Q.; Shi, T. HCCNet: an integrated network database of hepatocellular carcinoma. *Cell Res.* **2010**, *20* (6), 732−4.
(32) Vaquerizas, J. M.; Kummerfeld, S. K.; Teichmann, S. A.; Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **2009**, *10* (4), 252−63.
(33) Nagaraj, N.; Wisniewski, J. R.; Geiger, T.; Cox, J.; Kircher, M.; Kelso, J.; Paabo, S.; Mann, M. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **2011**, *7*, 548.
(34) Lundberg, E.; Fagerberg, L.; Klevebring, D.; Matic, I.; Geiger, T.; Cox, J.; Algenas, C.; Lundeberg, J.; Mann, M.; Uhlen, M. Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* **2010**, *6*, 450.
(35) Ding, C.; Jiang, J.; Wei, J.; Liu, W.; Zhang, W.; Liu, M.; Fu, T.; Lu, T.; Song, L.; Ying, W.; Chang, C.; Zhang, Y.; Ma, J.; Wei, L.; Malovannaya, A.; Jia, L.; Zhen, B.; Wang, Y.; He, F.; Qian, X.; Qin, J. A fast workflow for identification and quantification of proteomes. *Mol. Cell. Proteomics* **2013**, *12* (8), 2370−80.
(36) Maletzki, C.; Bodammer, P.; Breitruck, A.; Kerkhoff, C. S100 proteins as diagnostic and prognostic markers in colorectal and hepatocellular carcinoma. *Hepatitis Mon.* **2012**, *12* (10 HCC), e7240.
(37) Valenti, G.; Mira, A.; Mastrofrancesco, L.; Lasorsa, D. R.; Ranieri, M.; Svelto, M. Differential modulation of intracellular Ca2+ responses associated with calcium-sensing receptor activation in renal collecting duct cells. *Cell Physiol Biochem* **2010**, *26* (6), 901−12.
(38) Scheinman, E. J.; Rostoker, R.; Leroith, D. Cholesterol affects gene expression of the Jun family in colon carcinoma cells using different signaling pathways. *Mol. Cell. Endocrinol.* **2013**, *374* (1−2), 101−7.
(39) Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S. J.; Marra, M. A. Circos: an information aesthetic for comparative genomics. *Genome Res.* **2009**, *19* (9), 1639−45.
(40) Vizcaino, J. A.; Cote, R. G.; Csordas, A.; Dianes, J. A.; Fabregat, A.; Foster, J. M.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; O'Kelly, G.; Schoenegger, A.; Ovelleiro, D.; Perez-Riverol, Y.; Reisinger, F.; Rios, D.; Wang, R.; Hermjakob, H. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **2013**, *41* (Database issue), D1063−9.