Table 1 Living liver donor quality of life scale factor structure ($n = 432$)

Factor names and questions	Extracted factors							Cronbach's α-coefficient	Descriptive statistics	
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7		Mean value	Standard deviation
Factor 1: damage from the operation								0.868	4.50	0.76
* Chose actions with the least burden	0.875									
* Took longer breaks than usual	0.820									
* Changed behavior to avoid pain	0.790									
* Had difficulty performing casual movements	0.510					0.224				
* Could not eat to their satisfaction	0.415					0.355				
Factor 2: scars								0.868	3.92	1.04
* Felt tightness in the wound		0.819								
* Felt discomfort in the wound area		0.784								
* Unintentionally shielded the wound	0.320	0.691								
* Felt a lack of strength in abdominal muscles	0.234	0.336								
Factor 3: satisfaction								0.722	4.38	0.72
Felt glad to have become a donor			0.826							
Felt happy to have donated liver			0.823							
Felt that they had gained something personally by becoming a donor			0.491							
Felt satisfied with the results of the transplant surgery			0.381							
Factor 4: burden								0.670	4.23	0.85
* Experienced unexpected inconveniences after becoming a donor				0.685						
* Were troubled by intra-abdominal discomfort				0.631		0.210				
* Were financially affected by the surgery				0.549						
* Felt depressed at the sight of their scars				0.527						
Factor 5: after-effects								0.790	4.14	0.89
* Felt a change in their physical constitution					0.726					
* Felt after-effects of the surgery		0.365			0.579					
* Were troubled by epigastric region body shape changes					0.416					
Felt just as healthy as before surgery	0.212				0.352	0.299				
Factor 6: digestive symptoms								0.431	4.29	0.66
* Experienced heartburn						0.590				
* Experienced diarrhea accompanied by abdominal pain				0.222		0.405				
* Felt discomfort caused by constipation						0.354				
Factor 7: lack of understanding of donor health								0.746	3.07	1.25
People around them are attentive to their health							0.774			
People close to them are understanding of the donor's physical status							0.737			

Table 1 continued

Factor names and questions	Extracted factors							Cronbach's α-coefficient	Descriptive statistics	
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7		Mean value	Standard deviation
Eigenvalues	7.65	2.76	1.57	1.40	1.18	1.10	1.01			
Cumulative contribution ratio (%)	29.44	40.06	46.11	51.51	56.04	60.27	64.16			

Factor extraction method: major factor method, rotation method: promax method with Kaiser's normalization

Factor loading <0.02 omitted

* Denotes reversed items (reversed)

significant correlation among all subscales, except for "lack of understanding of donor health" (Table 3). "Damage from the operation" was highly correlated with the "role limitations because of physical problems" ($r = 0.680$) and "body pain" ($r = 0.606$) in SF-36. "Scars" exhibited a strong correlation with "body pain" ($r = 0.632$) in SF-36. We did not expect a correlation for "lack of understanding of donor health" because it was a social element; however, we noted a significant correlation with "general health perceptions" and "vitality", suggestive of a relationship. These results confirmed the validity of the LLD-QOL scale with the SF-36 scale, indicating that it could be considered as a scale for measuring QOL. These results also confirmed that the scale had criterion-related validity.

Verification of reliability

Verification of internal consistency

We calculated the α-coefficients of the subscales and found that "damage from the operation", "scars", "satisfaction", "after-effects", and "lack of understanding of donor health" had higher than optimal values (0.7–0.8). This internal consistency confirmed their reliability. Conversely, the fourth factor, "burden", exhibited a lower value of 0.670, while the sixth factor, "digestive symptoms", had a much lower value of 0.431 (see Table 1).

Verification by the test–retest method

Pearson product–moment correlation coefficients of initial survey and second survey LLD-QOL scale subscale results exhibited a strong correlation (0.749–0.918), confirming its stability as a scale through test–retest reliability.

Discussion

We analyzed the elements of LLD-specific QOL statistically, then developed a scale and investigated its reliability and validity through a large-scale survey.

Investigation of the LLD-QOL scale conceptual framework

Our QOL scale comprises seven subscales: "damage from the operation", "scars", "satisfaction", "burden", "after-effects", "digestive symptoms", and "lack of understanding of donor health". Good QOL denotes minimal postoperative effect by "damage from the operation" or "scars", unnoticeable "after-effects" and "digestive symptoms", high "satisfaction" as a donor, little sense of "burden", and the

Table 2 Correlation coefficients between subscales ($n = 432$)

	Damage from the operation	Scars	Satisfaction	Burden	After-effects	Digestive symptoms	Lack of understanding of donor health
Damage from the operation							
Scars	0.70***						
Satisfaction	0.06	0.11*					
Burden	0.39***	0.37***	0.21***				
After-effects	0.58***	0.66***	0.21***	0.47***			
Digestive symptoms	0.36***	0.40***	0.11*	0.27***	0.45***		
Lack of understanding of donor health	−0.02	0.07	0.31***	0.07	0.14**	0.09	

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 3 Correlation coefficients of living liver donor quality of life scale and SF-36 subscale scores ($n = 432$)

	SF-36 subscale scores							
	Physical functioning	Role limitations because of physical problems	Body pain	General health perceptions	Vitality	Social functioning	Role limitations because of emotional problems	Mental health
LLD-QOL Scale								
Subscale Scores								
Damage from the operation	0.49***	0.68***	0.61***	0.41***	0.45***	0.51***	0.54***	0.38***
Scars	0.39***	0.53***	0.63***	0.37***	0.40***	0.37***	0.45***	0.35***
Satisfaction	0.19***	0.14**	0.12*	0.32***	0.23***	0.16**	0.15**	0.23***
Burden	0.38***	0.38***	0.43***	0.43***	0.39***	0.42***	0.39***	0.38***
After-effects	0.49***	0.54***	0.57***	0.52***	0.53***	0.36***	0.43***	0.47***
Digestive symptoms	0.25***	0.28***	0.34***	0.38***	0.40***	0.28***	0.35***	0.40***
Lack of understanding of donor health	−0.01	0.00	0.02	0.17***	0.16**	0.08	0.08	0.16

LLD living liver donor, *QOL* quality of life, *SF-36* short form 36 questionnaire

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

feeling that one is understood by one's family, friends, and colleagues. Although there was some repetition between concepts and some differences in expressions used in the conceptual framework in our previous study [16], we managed to create a factor structure with almost no inconsistencies. The factors termed "damage from the operation" and "scars" were almost identical to the factors identified in our previous research [16]. Both of these factors exhibited differences that appeared attributable to the number of postoperative years, while changes over time in effects resulting from the invasiveness of surgery were also noted. The confirmed correlation with SF-36 subscales gave credence to the fact that there were no inconsistencies in the concepts.

"Satisfaction" and "burden" included concepts reflecting a sense of satisfaction about decision making, transplantation therapy-related feelings brought about by financial burden, and feelings about organ resection and loss, as exhibited in our previous study [16]. Although these psychosocial elements differed from the frameworks indicated by prior research, we felt that it was acceptable to use these new classifications based on the simplicity of "satisfaction" and "burden" as well as statistical evidence. Interestingly, the "sense of satisfaction" and "sense of burden" were not one-dimensional concepts. In other words, because a person can be burdened heavily, but still feel a strong sense of satisfaction, we believe that it is necessary to provide support that decreases the burden regardless of high or low levels of satisfaction. Although "after-effects" can be considered part of the concept of "damage from the operation", according to prior research [15], the difficulty of identifying long-term physical changes as having been clearly caused by the surgery suggests that they are perceived as part of a framework separate from that of the direct effects. Using the frameworks indicated by our results, changes caused by surgery directly are easier to

comprehend, making the LLD-QOL scale more meaningful. The concept of "digestive symptoms" was also indicated by our previous study [15]. The results led to the inclusion of heartburn, diarrhea, and constipation as one factor, which suggests the existence of an enabling factor common to all three symptoms other than their feature of being digestive. It is therefore highly significant to comprehend these as one effect of hepatectomy.

Finally, we restricted "lack of understanding of donor health" to the perceptions of donors on how they were treated by close family and friends, excluding medical professionals. It appears valid to consider the influence of medical professionals as unimportant because transplants were performed between relatives and the subjects had hardly any contact with medical professionals after their rehabilitation into society. Thus, LLD QOL can be considered statistically supported by the seven elements according to these research results; however, because we investigated only the correlation between factors at the present stage, covariance structure analysis of models between factors and further concept analysis of the QOL structure of LLDs is required in the future.

Usefulness as a scale: reliability and validity

Our LLD-QOL scale appears to have adequate construct and criterion-related validity. Thus, it may be considered as a valid scale for evaluating QOL. Moreover, fair test–retest stability and internal consistency allowed us to confirm that the scale was reliable and appropriate for clinical practice. However, one should exercise caution with the subscale "digestive symptoms". The α-coefficient of the items included in this subscale, "heartburn", "diarrhea", and "constipation", was 0.431, denoting low internal consistency. The reasons cited for the low α-coefficient include the small number of questions (3) and low correlation coefficients between the questions. To increase the α-coefficient, one could add items that ask for similar information, although this could result in the scale becoming lengthy. The low correlation also suggests that factors are conceived by subjects in a multilateral manner, which is an advantage. Subjects may experience severe constipation but no diarrhea, or vice versa, making the low correlation between items appear valid. The results of factor analysis also showed that factor loading of the three items was not particularly low. Considering that there was no problem with validity or test–retest reliability, it does not appear necessary to view this as a problem; however, when using the scale, one should interpret the results carefully, considering the possibility that differences are not being sensitively reflected, and comparatively examining each item for differences. Even if no significant difference is noted, it is logical to interpret results conservatively.

The subscale "burden" had an α-coefficient of 0.67, which was slightly lower than the optimum level of 0.7; however, because there was no problem with validity, it can be considered to be within the permissible range. Meanwhile, "damage from the operation", "scars", "satisfaction", "after-effects", and "lack of understanding of donor health" exhibited adequate reliability, and these subscales could be used independently. It may also be possible to use subscales independently when comparing differences in QOL according to surgical technique. Because we have not yet examined the predictive validity of the scale, scale sensitivity according to postoperative changes over time needs to be investigated with further verification of its usefulness. Reliability also needs to be increased further, considering the future creation and addition of questions with reasonable similarity for the digestive symptoms subscale. Moreover, we note that the QOL conceptual framework and questions have not been subjected to evaluation by a psychiatrist. Thus, it will be necessary to conduct revisions in line with pre- and postoperative psychiatric evaluations and follow up of donors by psychiatrists. Nevertheless, although some points regarding interpretation need to be considered, there are no problems with validity and reliability when using this scale, which can be used to effectively evaluate the QOL of LLDs.

Conclusions

We developed and confirmed the validity and reliability of a LLD-specific QOL scale comprising 26 items and 7 subscales. This scale enables us to assess the QOL specifically of LLDs.

References

1. The Japanese Liver Transplantation Society. Liver transplantation in Japa—Registry by the Japanese Liver Transplantation Society (in Japanese with English abstract). Ishoku (Japanese J Transpl). 2009;45:621–32.
2. DuBay DA, Holtzman S, Adcock L, Abbey SE, Greenwood S, Macleod C, et al. Adult right-lobe living liver donors: quality of life, attitudes and predictors of donor outcomes. Am J Transplant. 2009;9:169–78.
3. Erim Y, Beckmann M, Kroencke S, Valentin-Gamazo C, Malago M, Broering D. Psychological strain in urgent indications for living donor liver transplantation. Liver Transpl. 2007;13:886–95.
4. Kroencke S, Wilms C, Broering D, Rogiers X, Schulz K-H. Psychosocial aspects of pediatric living donor liver transplantation. Liver Transpl. 2006;12:1661–6.

5. Basaran O, Karakayali H, Emiroglu R, Tezel E, Moray G, Haberal M. Donor safety and quality of life after left hepatic lobe donation in living-donor liver transplantation. Transplant Proc. 2003;35:2768–9.

6. Karliova M, Malago M, Valentin-Gamazo C, Reimer J, Treichel U, Franke GH, et al. Living-related liver transplantation from the view of the donor: a 1-year follow-up survey. Transplantation. 2002;73:1799–804.

7. Kim-Schluger L, Florman SS, Schiano T, O'Rourke M, Gagliardi R, Drooker M, et al. Quality of life after lobectomy for adult liver transplantation. Transplantation. 2002;73:1593–7.

8. Trotter JF, Talamantes M, McClure M, Wachs M, Bak T, Trouillot T, et al. Right hepatic lobe donation for living donor liver transplantation: impact on donor quality of life. Liver Transpl. 2001;7:485–93.

9. Chan SC, Liu CL, Lo CM, Lam BK, Lee EW, Fan ST. Donor quality of life before and after adult-to-adult right liver live donor liver transplantation. Liver Transpl. 2006;12:1529–36.

10. Verbesey JE, Simpson MA, Pomposelli JJ, Richman E, Bracken AM, Garrigan K, et al. Living donor adult liver transplantation: a longitudinal study of the donor's quality of life. Am J Transplant. 2005;5:2770–7.

11. Nagai S, Nakamura T, Kamei H, Ito K, Kanematsu N, Kato C, et al. "Transplantation QOL" practical and mental QOL of liver transplant patients and living donors (in Japanese). Ishoku (Japanese J Transpl). 2008;43:264–75.

12. The Japanese Liver Transplantation Society. The Japanese Liver Transplantation Society live liver donor survey (in Japanese). 2005. pp. 2–41.

13. Yuan Y, Gotoh M. Biliary complications in living liver donors. Surg Today. 2010;40(5):411–7.

14. Fukuhara S. Why QOL now? Its position as a patient-based outcome. In: Ikegami, N, Fukuhara S, Shimozuma K, Ikeda S, editors. Handbook of QOL evaluation for clinical use (in Japanese). Tokyo: Igaku-Shoin; 2001. pp. 2–7.

15. Parikh ND, Ladner D, Abecassis M, Butt Z. Quality of life for donors after living donor liver transplantation: a review of the literature. Liver Transpl. 2010;16(12):1352–8.

16. Morooka Y, Umeshita K, Hagiwara K, Ogawa K. The structure of quality of life in living liver donors (in Japanese with English abstract). Ishoku (Japanese J Transpl). 2011;48(2–3):147–53.

17. Fukuhara S, Suzukamo Y. Manual of SF-36v2 Japanese version. Institute for Health Outcomes and Process Evaluation Research: Kyoto; 2004. pp. 7–145.

Identification of Missing Proteins in the neXtProt Database and Unregistered Phosphopeptides in the PhosphoSitePlus Database As Part of the Chromosome-Centric Human Proteome Project

Takashi Shiromizu,[†] Jun Adachi,[†] Shio Watanabe,[†] Tatsuo Murakami,[†] Takahisa Kuga,[†] Satoshi Muraoka,[†] and Takeshi Tomonaga*[†,‡]

[†]Laboratory of Proteome Research, National Institute of Biomedical Innovation, Ibaraki, Osaka, Japan
[‡]Clinical Proteomics Research Center, Chiba University Hospital, Chiba, Japan

Ⓢ Supporting Information

ABSTRACT: The Chromosome-Centric Human Proteome Project (C-HPP) is an international effort for creating an annotated proteomic catalog for each chromosome. The first step of the C-HPP project is to find evidence of expression of all proteins encoded on each chromosome. C-HPP also prioritizes particular protein subsets, such as those with post-translational modifications (PTMs) and those found in low abundance. As participants in C-HPP, we integrated proteomic and phosphoproteomic analysis results from chromosome-independent biomarker discovery research to create a chromosome-based list of proteins and phosphorylation sites. Data were integrated from five independent colorectal cancer (CRC) samples (three types of clinical tissue and two types of cell lines) and lead to the identification of 11,278 proteins, including 8,305 phosphoproteins and 28,205 phosphorylation sites; all of these were categorized on a chromosome-by-chromosome basis. In total, 3,033 "missing proteins", i.e., proteins that currently lack evidence by mass spectrometry, in the neXtProt database and 12,852 unknown phosphorylation sites not registered in the PhosphoSitePlus database were identified. Our in-depth phosphoproteomic study represents a significant contribution to C-HPP. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium with the data set identifier PXD000089

KEYWORDS: Chromosome-Centric Human Proteome Project, missing protein, phosphopeptide, IMAC, colorectal cancer, FASP, neXtProt, PhosphoSitePlus
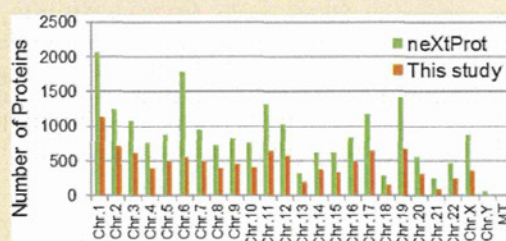
■ INTRODUCTION

The Chromosome-Centric Human Proteome Project (C-HPP) is a worldwide effort by proteomics researchers to create expression profiles of the approximately 20,000 genes encoded on all human chromosomes and build a database.[1] Protein expression patterns are closely associated with the location of the gene on a chromosome and are correlated with diseases associated with chromosomal abnormalities. Therefore, a comprehensive understanding of the protein expression profile of each chromosome is critical for biological studies and clinical research. The initial aim of C-HPP was to identify at least one protein isoform for every gene encoded by the human genome. Proteins not detected by antibody or proteomic analysis using mass spectrometry are called "missing proteins".[2] Currently, there are about 6,000 missing proteins among all of the proteins in the neXtProt database.[3] One reason why missing proteins are undetectable is that protein expression differs significantly between tissue and cell types. Although the number of proteins that can be identified in a single analysis has greatly increased due to recent advances in mass spectrometric techniques, complete expression profiles of all proteins will require the integration and analysis of data from a wide variety of samples.

C-HPP also aims to map specific protein variations such as post-translational modifications (PTMs), alternative splicing, and protease-processed variants.[2] Protein phosphorylation is a key regulator of cellular signal transduction processes, and its deregulation is involved in the onset and progression of various human diseases such as cancer and inflammatory and metabolic disorders.[4−7] Recent advances in proteomics, especially phosphopeptide enrichment strategies such as immobilized metal ion affinity chromatography (IMAC) and TiO$_2$ affinity chromatography,[8] have enabled the identification of up to several thousands of site-specific phosphorylation events within one large-scale analysis.[9−19]

As participants in C-HPP, we have integrated proteomic and phosphoproteomic analysis data from human colorectal cancer tissue and cell lines and created a chromosome-based list of identified proteins. Newly detected proteins and phosphorylated peptides were identified from the neXtProt and PhosphoSitePlus databases.

■ MATERIALS AND METHODS

Tissue and Cell Culture Samples

Colorectal cancer tissue and tumor-adjacent normal tissue samples were obtained from 44 patients at Chiba University School of Medicine. Tissue samples were frozen in liquid nitrogen and stored at −80 °C until analysis. Written informed consent was obtained from each patient before surgery, and the protocol was approved by the ethics committees of the Proteome Research Center, National Institute of Biomedical Innovation, and the Chiba University School of Medicine. Cell cultures used were HCT116, SW480, and SW620. HCT116, a colorectal cancer cell line, was grown in RPMI 1640 medium with 10% fetal bovine serum (Invitrogen, Carlsbad, CA, USA) and penicillin/streptomycin (Invitrogen). Cells were maintained at 37 °C in an incubator supplemented with 5% CO_2 until they grew to 80% confluence. SW480 and SW620, colon cancer cell lines, were grown at 37 °C and 5% CO_2 for at least five passages in SILAC media (R1780-RPMI-1640 without arginine, lysine, leucine (Sigma−Aldrich Corp., St. Louis, MO, USA) with 10% dialyzed fetal bovine serum (Invitrogen) and 100 U/mL penicillin/streptomycin (Invitrogen)) containing 84 mg/L L-arginine (Arg0) and 40 mg/L L-lysine (Lys0) (light), or $^{13}C_6^{15}N_4$-L-arginine (Arg10) and $^{13}C_6$-L-lysine (Lys6) (heavy) and 50 mg/L L-leucine.

Protein Extraction and Digestion

Protein extraction and proteolytic digestion were performed using a filter-assisted sample preparation (FASP) protocol.[20] Tissue samples or pellets of cultured cells were homogenized by sonication in FASP buffer [1% SDS, 0.1 M DTT, in 0.1 M Tris/HCl, pH 7.6 and PhosSTOP phosphatase inhibitor cocktail (Roche, Mannheim, Germany)]. Protein concentration was determined using a DC protein assay kit (Bio-Rad, Richmond, CA, USA). A total of 10 mg (for phosphoproteomic analysis) or 100 μg (for proteomic analysis) of extracted proteins was digested using 1:100 (w/w) trypsin (proteomics grade; Roche) for 12 h at 37 °C. Digested peptides were concentrated and purified using a C18 Sep-PAK cartridge (Waters, Milford, MA, USA).

Phosphopeptide Enrichment

Phosphopeptide enrichment was performed using immobilized Fe(III) affinity chromatography (Fe-IMAC) as described previously.[21] The Fe-IMAC resin was prepared from Probond (Nickel-Chelating Resin; Invitrogen) by substituting Ni^{2+} on the resin with Fe^{3+}. Ni^{2+} was released from Probond upon treatment with 50 mM EDTA-2Na, and then Fe^{3+} was chelated to the ion-free resin upon incubation with 100 mM $FeCl_3$ in 0.1% acetic acid. The Fe-IMAC resin was packed into an open column for large-scale enrichment. Following equilibration of the resin with loading solution (60% acetonitrile/0.1% TFA), the peptide mixture was loaded onto the IMAC column. After washing with loading solution (9 times the volume of the IMAC resin) and 0.1% TFA (3 times the volume of the IMAC resin), phosphopeptides were eluted using 1% phosphoric acid (2 times the volume of the IMAC resin).

iTRAQ Labeling

Enriched peptides were labeled with isobaric tags for relative and absolute quantification reagents (iTRAQ 4 plex; Applied Biosystems, Foster City, CA, USA) according to the manufacturer's instructions. Peptide mixtures desalted with C18 Stage-Tips were incubated in the iTRAQ reagents for 1 h.

iTRAQ 115, 116, and 117 were used for labeling individual samples, and iTRAQ 114 was used as the reference sample, a mixture of aliquots of all samples. The reaction was terminated by the addition of an equal volume of distilled water. The labeled samples were combined, acidified with trifluoroacetic acid, and desalted with C18 Stage-Tips.

### Strong Cation Exchange Chromatography (SCX)

The peptides were fractionated using a HPLC system (Shimadzu Prominence UFLC) fitted with an SCX column (50 mm × 2.1 mm, 5 μm, 300 Å, ZORBAX 300SCX; Agilent Technology). The mobile phases consisted of buffer A [25% acetonitrile and 10 mM KH_2PO_4 (pH 3)] and B [25% acetonitrile, 10 mM KH_2PO_4 (pH 3), and 1 M KCl]. The labeled peptides were dissolved in 200 μL of buffer A and separated at a flow rate of 200 μL/min using a four-step linear gradient: 0% B for 30 min, 0% to 10% B in 15 min, 10% to 25% B in 10 min, 25% to 40% B in 5 min, 40% to 100% B in 5 min, and 100% B for 10 min. Fractions were collected and desalted using C18-Stage Tips (number of fractions: CRC tissues_1 peptides, 30 fractions; CRC tissues_1 phosphopeptides, 25 fractions; HCT116 peptides, 34 fractions; HCT116 phosphopeptides, 32 fractions; SW480 + SW620 peptides, 25 fractions; SW480 + SW620 phosphopeptides, 30 fractions; CRC tissues_2 non-tumor phosphopeptides, 30 fractions; CRC tissues_2 tumor phosphopeptides, 30 fractions).

LC−MS/MS Analysis

Fractionated peptides were analyzed using an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a nanoLC interface (AMR, Tokyo, Japan), a nanoHPLC system (Michrom Paradigm MS2) and an HTC-PAL autosampler (CTC, Analytics, Zwingen, Switzerland). The analytical column was made in-house by packing L-column2 C18 particles (Chemical Evaluation and Research Institute (CERI), Tokyo, Japan), into a self-pulled needle (200 mm length × 100 μm inner diameter). The mobile phases consisted of buffer A (0.1% formic acid and 2% acetonitrile) and B (0.1% formic acid and 90% acetonitrile). Samples dissolved in buffer A were loaded onto a trap column (0.3 × 5 mm, L-column ODS; CERI). The nanoLC gradient was delivered at 500 nL/min and consisted of a linear gradient of buffer B developed from 5% to 30% B in 180 min. A spray voltage of 2000 V was applied.

Full MS scans were performed using an orbitrap mass analyzer (scan range m/z 350−1500, with 30 K fwhm resolution at m/z 400). The 10 most intense precursor ions were selected for the MS/MS scans, which were performed using collision-induced dissociation (CID) and higher energy collision-induced dissociation (HCD, 7500 fwhm resolution at m/z 400) for each precursor ion. The dynamic exclusion option was implemented with a repeat count of 1 and exclusion duration of 60 s. Automated gain control (AGC) was set to $1.00e + 06$ for full MS, $1.00e + 04$ for CID MS/MS, and $5.00e + 04$ for HCD MS/MS. The normalized collision energy values were set to 35% for CID and 50% for HCD.

The CID and HCD raw spectra were extracted and searched separately against UniProtKB/Swiss-Prot (release-2010_05), which contains 20,295 sequences (the forward and reverse-decoy) of *Homo sapiens*, using Proteome Discoverer 1.3 (Thermo Fisher Scientific) and Mascot v2.3. The precursor mass tolerance was set to 7 ppm, and fragment ion mass tolerance was set to 0.5 Da for CID and 0.01 Da for HCD. The search parameters allowed two missed cleavage for trypsin,
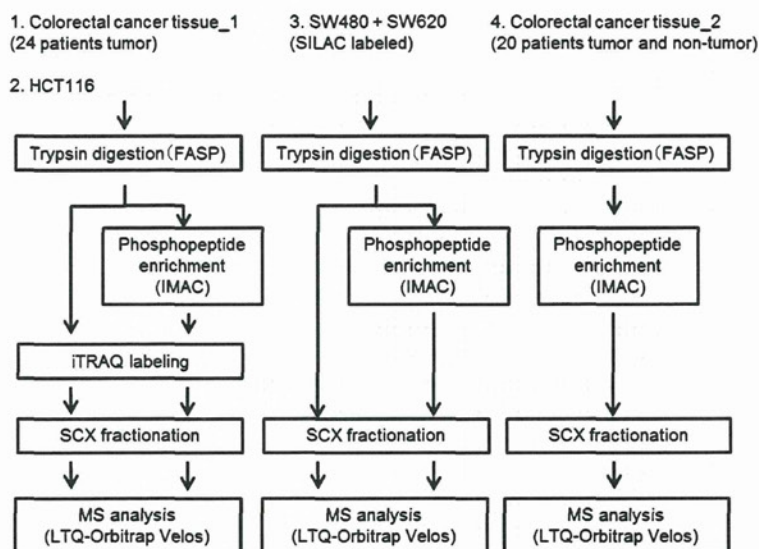
Figure 1. Schematic representation of the experimental work flow for the proteomic and phosphoproteomic analyses of the four experiments. SW480 + SW620: a mixture of protein extracts obtained from SW480 and SW620 cells. After trypsin digestion, each sample was separated for proteomic (100 μg) or phosphoproteomic (10 mg) analysis. Digested samples were separated by using an SCX column. LC–MS/MS, requiring 3-h runs, was performed using an LTQ-Orbitrap Velos.
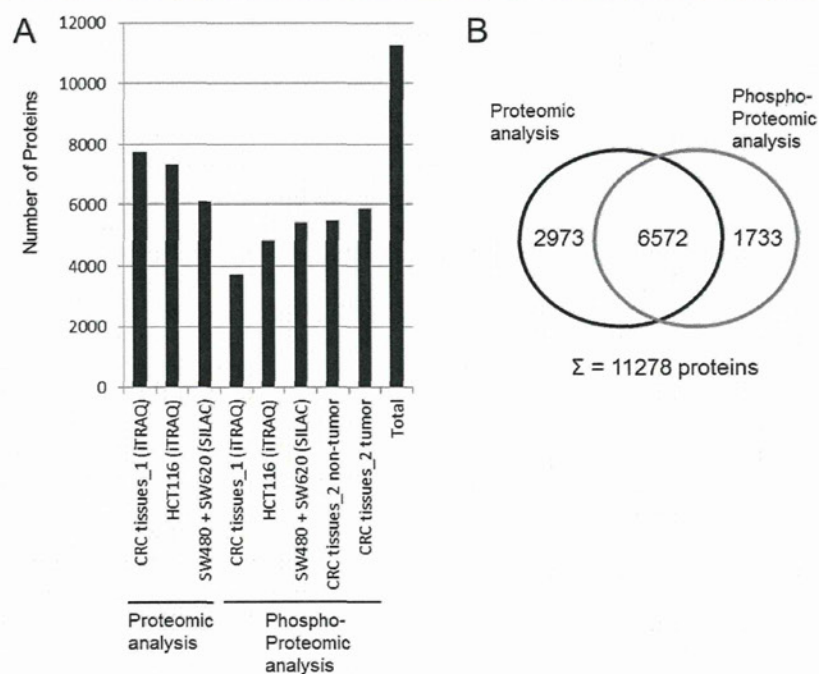
Figure 2. Number and overlap of identified proteins from the proteomic and phosphoproteomic analyses. (A) Number of identified proteins from the proteomic and phosphoproteomic analyses of the 8 data sets. (B) Proportion of proteins identified in each analysis and overlap between proteins identified by the proteomic and phosphoproteomic analyses.

fixed modifications (carbamidomethylation at cysteine), and variable modifications (oxidation at methionine). Fixed modifications were set for CRC tissue and HCT116 (iTRAQ labeling at lysine and the N-terminal residue) and SW480 + SW620 (SILAC labeling 13C(6) 15N(4) Arg, 13C(6) Lys). Variable modifications were added for phosphoproteomic analysis (phosphorylation at serine, threonine, and tyrosine). In the workflow of Proteome Discoverer 1.3, following the Mascot search, the phosphorylated sites on the identified peptides were assigned again using the PhosphoRS algorithm, which calculated the possibility of the phosphorylated site from the spectra matched to the identified peptides.[22] The score threshold for peptide identification was set at 1% false-discovery rate (FDR) and 75% phosphoRS site probability. FDR was calculated using the Percolator algorithm for peptide sequence analysis. Percolator uses >30 features of a peptide spectral match (PSM) to distinguish true positives from random matches.
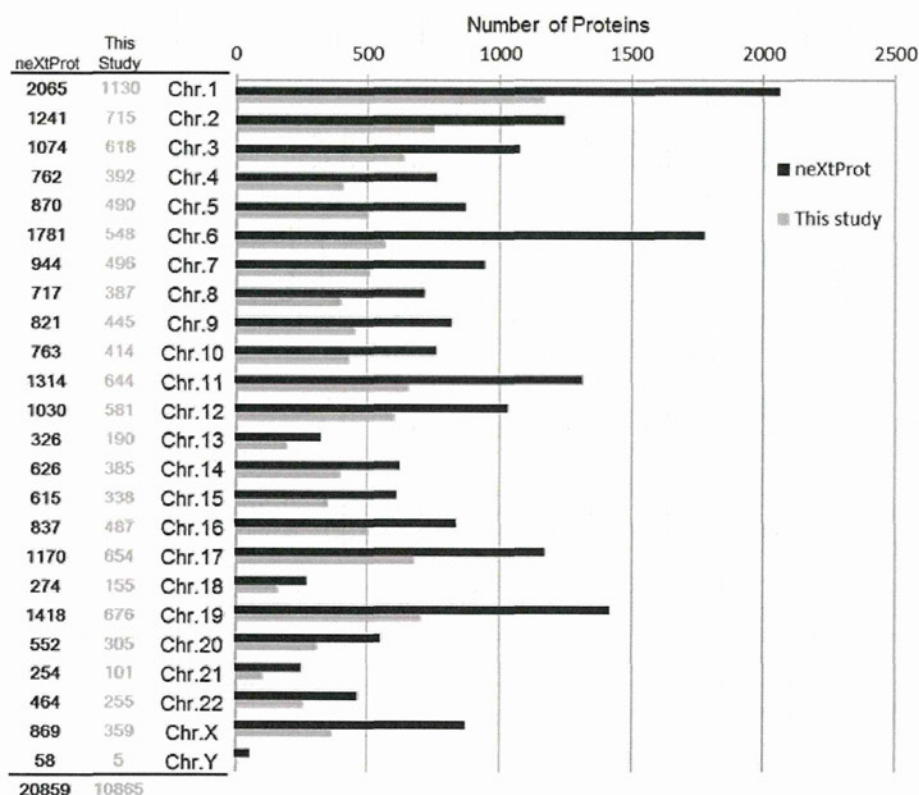
Figure 3. Chromosomal distribution of the identified proteins (gray) in relation to total proteins (black) registered in the neXtProt database.

Bioinformatics Analysis

Chromosomal locations and missing protein analyses of identified proteins were elucidated using the neXtProt database (http://www.nextprot.org/db/), and identified phosphorylation sites were elucidated using the PhosphoSitePlus database (http://www.phosphosite.org/). The function of identified missing proteins was elucidated by ingenuity pathway analysis software (Ingenuity Systems, Redwood City, USA).

Stable Isotope-Labeled Peptides

Stable isotope-labeled standard peptides (SIS peptides, crude grade) were synthesized (Thermo Fisher Scientific, Ulm, Germany). A single lysine was replaced by isotope-labeled lysine ($^{13}C_6$, 98%, $^{15}N_2$, 98%). The SIS peptides were dissolved in distilled water at a concentration of 1 $\mu g/\mu L$ and stored at -80 °C. A mixture of these SIS peptides was added to colorectal carcinoma phosphoproteomic samples.

■ RESULTS

As part of the C-HPP project, we combined the eight data sets from four different experiments obtained from colorectal cancer tissue and colon cancer cells; these experiments included three quantitative analyses and one non-quantitative analysis. Colorectal cancer tissues and colon cancer cells were first solubilized and trypsin-digested using the FASP method.[20] Phosphopeptides were then enriched using the IMAC method. These peptides and phosphopeptides were fractionated on a Strong Cation-Exchange (SCX) column before LC−MS/MS using an LTQ-Orbitrap mass spectrometer (Figure 1). Proteome Discoverer 1.3 software was used to analyze the RAW data files, Mascot was used as the search engine, and UniProtKB/Swiss-Prot (release-2010_05) was the database. Following data

integration, 11,278 proteins were identified with Peptide FDR ≤1.0 containing at least one unique peptide corresponding to one protein in the database (Figure 2A, Supplementary Table 1−4). Of these, 8,305 proteins were identified as phosphorylated. Among the total identified proteins, 673 proteins were identified only with CID, and 386 proteins were identified only with HCD. Also, 4924 phosphopeptides were identified only with CID, and 3538 phosphopeptides were identified only with HCD. A total of 6,572 proteins were commonly identified in the proteomic and phosphoproteomic analyses (Figure 2B). However, a proportion of proteins were found not to overlap in the analyses. This is probably due to the abundance and complexity of the proteins and phosphoproteins in the samples, which prevent proteomics and phosphoproteomics to identify all of the proteins and phosphoproteins present.

Quantitative analyses were performed to investigate the differences between metastatic and non-metastatic cases by using clinical tissue and two types of cultured cells (a mixture of SW620 + SW480 and HCT116 cells). Clinical tissue samples of primary colorectal cancer obtained from 12 patients with or without metastasis were pooled. Cancers without metastasis were labeled with iTRAQ 114 or 116, and those with metastasis were labeled with iTRAQ 115 or 117. We also performed quantitative analyses between metastatic and non-metastatic cell lines. HCT116 metastatic clone was established by orthotopic implantation model mouse, and its protein expression was compared with that of the parent clone. SW620 cell line is a lymph node metastatic variant of SW480. HCT116 parent clone was labeled with iTRAQ 114 or 115, and metastatic clone was labeled with iTRAQ 116 or 117. SW480 and SW620 were reciprocally labeled with light and heavy

Table 1. Number of Identified Proteins in Each Chromosome

			proteomic analysis			phosphoproteomic analysis				
	neXtProt	total	CRC tissue_1	HCT116	SW480 + SW620	CRC tissue_1	HCT116	SW480 + SW620	CRC tissue_2 non-tumor	CRC tissue_2 tumor
Chr.1	2065	1171	808	767	611	356	494	554	568	598
Chr.2	1241	753	516	481	416	259	305	378	371	414
Chr.3	1074	642	445	425	360	209	275	310	321	341
Chr.4	762	412	267	251	185	126	178	176	210	224
Chr.5	870	508	339	319	285	168	210	249	255	281
Chr.6	1781	569	395	347	302	186	233	260	272	291
Chr.7	944	511	364	319	305	160	204	269	244	253
Chr.8	717	402	270	263	196	125	177	188	176	199
Chr.9	821	458	312	294	229	161	197	212	230	225
Chr.10	763	434	298	300	254	131	186	223	232	245
Chr.11	1314	660	477	447	388	227	299	340	352	360
Chr.12	1030	603	399	409	330	207	274	296	294	316
Chr.13	326	194	140	127	114	62	80	97	87	99
Chr.14	626	400	283	252	218	132	166	180	203	210
Chr.15	615	353	237	234	181	121	163	184	174	197
Chr.16	837	508	341	333	260	159	230	245	236	258
Chr.17	1170	678	437	484	399	239	330	339	347	361
Chr.18	274	161	102	112	74	50	67	64	70	77
Chr.19	1418	707	479	471	362	253	336	332	341	369
Chr.20	552	313	223	194	182	119	137	151	138	160
Chr.21	254	103	78	68	64	36	41	51	47	53
Chr.22	464	262	187	182	153	93	109	123	129	130
Chr.X	869	369	261	222	222	120	144	189	175	188
Chr.Y	58	6	4	2	1	1	0	0	1	2
NA[a]		101	73	30	17	11	11	17	18	16
total	20845	11278	7735	7333	6108	3355	4352	5427	5491	5867

[a]Not applicable in neXtProt.

stable isotope amino acids (lysine and arginine). HCT116 has a mutation in codon 13 of the ras protooncogene, while SW480 and SW620 have a mutation in codon 12. Among 8305 proteins and 28,205 phosphopeptides, 472 proteins and 2547 phosphopeptides showed >2-fold differences between meta-static and non-metastatic tissues and cell lines (either upregulated or downregulated).

A total of 20,845 proteins have been registered in the neXtProt database. Proteins identified in this study were referred to the database and accounted for 53.6% (11,177/20,845) of all the proteins registered in the neXtProt database; their chromosomal locations are shown in Figure 3 and Table 1. Of the proteins registered in the neXtProt database, the expression of 14,612 proteins (70.1% of the total of 20,845 proteins) has been confirmed by mass spectrometry or antibody assay (protein level 'yes'), whereas 10,649 proteins (51.1% of the total of 20,845 proteins) have been identified only by MS analysis (proteomic level 'yes') (Table. 2). Cross-checking the 11,278 proteins identified in this study with the neXtProt database revealed 1,145 proteins currently lacking evidence of protein expression by mass spectrometry or

Table 2. Number of Proteins Identified at the Protein or Proteomic Level

evidence		neXtProt	this study
protein level	yes	14612	10032
	no	6233	1145
proteomics level	yes	10649	8144
	no	10196	3033

antibody assays, and 3,033 proteins lacking evidence by mass spectrometry. These "missing proteins (protein level = no and proteomic level = no)" are listed on a chromosome-by-chromosome basis (Figure 4).

In contrast, 28,205 phosphorylation sites were identified (Supplementary Table 5). When these phosphopeptides were cross-checked with the PhosphoSitePlus database, 15,353 registered phosphorylation sites were identified, or 12.2% of all registered sites in PhosphoSitePlus (15,353/125,433). Of these, 12,852 sites were not registered in PhosphoSitePlus (Figure 5A). The chromosomal locations of these phosphor-ylation sites are shown in Figure 5B. In order to verify the accuracy of the identified phosphopeptides, lysine at the C-terminus of two peptides (LYNSEESRPYTNK, SASQS-SLDKLDQELK) was labeled by stable isotope ($^{13}C_6$, $^{15}N_2$). The SIS peptides were added to the extract of colorectal cancer tissue, and annotated mass spectra and extracted ion chromato-gram data of SIS peptides were compared to those of nonlabeled endogenous peptides (Supplementary Figure 1).

Non-quantitative analyses were also performed using pooled colorectal carcinoma tissues and tumor-adjacent normal tissues 5−10 cm remote from the tumor. To investigate the association between phosphoproteins and biological function, gene ontology analysis was performed by using Ingenuity Pathway Analysis (IPA) software. Specifically identified phosphoproteins in normal (636 proteins) and carcinoma tissues (1020 proteins) were also analyzed by IPA. Molecular functions involved in cell cycle (normal = 9 proteins, tumor = 132 proteins; $p < 0.01$ Fisher's exact test) and DNA replication (normal = 15 proteins, tumor = 106 proteins ; $p < 0.01$)

A Protein level = No

neXtProt	This Study	un-identified		Number of Proteins
633	125	508	Chr.1	
329	77	252	Chr.2	
294	64	230	Chr.3	
207	53	154	Chr.4	
250	62	188	Chr.5	
467	58	409	Chr.6	
315	50	265	Chr.7	
222	51	171	Chr.8	
266	52	214	Chr.9	
223	40	183	Chr.10	
480	51	429	Chr.11	
250	55	195	Chr.12	
86	14	72	Chr.13	
173	42	131	Chr.14	
195	38	157	Chr.15	
243	53	190	Chr.16	
309	58	251	Chr.17	
67	13	54	Chr.18	
518	98	420	Chr.19	
160	32	128	Chr.20	
101	8	93	Chr.21	
132	19	113	Chr.22	
297	31	266	Chr.X	
29	1	28	Chr.Y	
6247	1145	5101		

B Proteomics = No

neXtProt	This Study	un-identified		Number of Proteins
1027	327	700	Chr.1	
574	195	379	Chr.2	
512	179	333	Chr.3	
404	124	280	Chr.4	
377	124	253	Chr.5	
837	161	676	Chr.6	
471	153	318	Chr.7	
376	118	258	Chr.8	
426	132	294	Chr.9	
364	111	253	Chr.10	
720	166	554	Chr.11	
475	154	321	Chr.12	
157	50	107	Chr.13	
288	105	183	Chr.14	
284	92	192	Chr.15	
396	144	252	Chr.16	
567	186	381	Chr.17	
135	50	85	Chr.18	
707	190	517	Chr.19	
273	88	185	Chr.20	
162	25	137	Chr.21	
202	65	137	Chr.22	
426	93	333	Chr.X	
43	1	42	Chr.Y	
10203	3033	7170		

Figure 4. Chromosomal distribution of the identified proteins (gray) and total registered proteins (black) in the neXtProt database with no evidence of expression at the protein level (A) and at the proteomic level (B).
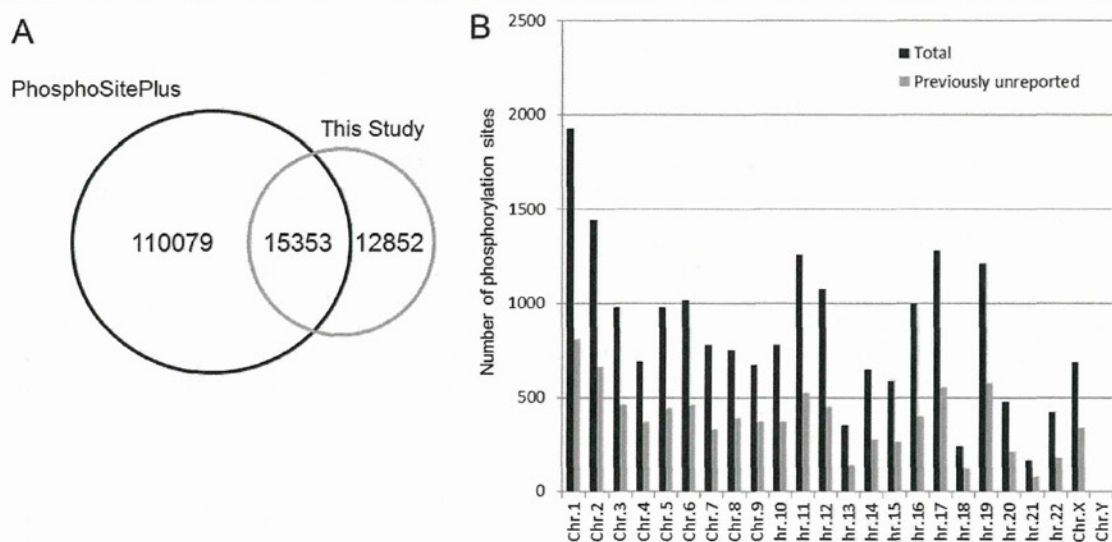

Figure 5. (A) Overlap between phosphorylation sites identified in this study and those registered in the PhosphoSitePlus database. (B) Chromosomal distribution of the identified previously unreported phosphorylation sites (gray) and total registered sites (black).

functions were more abundant in cancer tissues than in normal tissues (Supplementary Figure 2).

■ DISCUSSION

The objective of C-HPP is to map and annotate all protein-coding genes on each human chromosome. C-HPP also prioritizes particular protein subsets such as post-translational modifications (PTMs) and low-abundance proteins. Thus, we have integrated proteomic and phosphoproteomic data obtained from a shotgun analysis using human cancer tissue and cell lines prepared for various purposes. We have integrated quantitative and non-quantitative data; quantitative analysis was

performed for the relative quantification between metastatic and non-metastatic colorectal carcinoma samples, while non-quantitative analysis was performed to compare the tumor and normal tissues. As a result, we identified 11,278 proteins, 8,305 phosphoproteins, and 28,205 phosphorylation sites, and their chromosomal locations were defined using the neXtProt database. Furthermore, we were able to identify 3,033 missing proteins that currently lack evidence by mass spectrometry and 12,852 unknown phosphorylation sites that are not in the PhosphoSitePlus database.

Currently, the research group with the most advanced mass analysis system can identify over 10,000 proteins in a single

F

analysis run and identify about 50% of the proteins in their comprehensive analyses using multiple cell lines.[23] Additionally, the number of phosphorylation sites identified has exponentially increased,[24] largely due to improvements in phosphopeptide enrichment methods such as IMAC[15] and TiO_2 affinity chromatography.[8] A phosphoproteomic study of HeLa cells identified more than 65,000 phosphopeptides using a combination of phosphopeptide enrichment and SCX chromatography.[25] Several phosphoproteomic studies using tissue samples have been reported and have identified 5,195 phosphopeptides from human dorsolateral prefrontal cortex[26] and 5,698 phosphorylation sites from tumor tissues of melanoma model mice.[27] In our study, we identified 11,278 proteins and 28,205 phosphorylation sites; some had been identified in previous reports, but a number of the proteins and phosphorylation sites are not listed in the neXtProt or PhosphoSitePlus databases. Since mass analysis systems are rapidly becoming more powerful, in the future an individual research group may be able to identify all the proteins in the human genome in one analysis. However, in order to build an extensively annotated proteome database, which is one purpose of the C-HPP project, it is necessary to combine the analyzed data of various samples from many research groups.

Our analysis increased the number of identified proteins by combining the results of proteome analysis and phosphoproteome analysis on identical samples. Even using commonly studied cell lines, combining the results of post-translational modification analysis and analysis of fractionated samples will increase the number of identified missing proteins. The data presented here are based on relative quantification, and thus to confirm protein expression and examine protein abundance and localization, validation using antibodies or selected reaction monitoring (SRM) is required. Such validations will benefit from information on the identified cell line, sample preparation methods, MS analysis data, and the sequences of the identified peptides/phosphopeptides. We and other researchers, including Muraoka and colleagues,[28] Narumi and colleagues (unpublished data), and Kume and colleagues (unpublished data), are currently using a strategy for large-scale proteomics and SRM-based validation to discover biomarkers for various diseases and aim to obtain additional proteomics data by SRM validation and quantitation that will be integrated into the C-HPP project.

■ ASSOCIATED CONTENT

ⓢ Supporting Information

Annotated mass spectra and retention time data from liquid chromatography; results of phosphoproteomic analysis in normal and carcinoma tissues; lists of identified proteins and peptides by phosphoproteomic and proteomic analysis; list of identified phosphorylation sites. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository[29] with the data set identifier PXD000089. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +81-72-641-9862. Fax: +81-72-641-9861. E-mail: tomonaga@nibio.go.jp.

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

C-HPP, Chromosome-Centric Human Proteome Project; CRC, colorectal cancer; PTMs, post-translational modifications; IMAC, immobilized metal ion affinity chromatography; FASP, filter-assisted sample preparation; SCX, strong cation-exchange; FDR, false discovery rate; SRM, selected reaction monitoring; CID, collision-induced dissociation; HCD, higher energy collision-induced dissociation; LC−MS/MS, liquid chromatography tandem mass spectrometry; CE, collision energy; LTQ, linear ion trap; fwhm, full wide at half-maximum

■ REFERENCES

(1) Hancock, W.; Omenn, G.; Legrain, P.; Paik, Y. K. Proteomics, human proteome project, and chromosomes. *J. Proteome Res.* **2011**, *10* (1), 210.

(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), 221−223.

(3) 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.

(4) Hanahan, D.; Weinberg, R. A. The hallmarks of cancer. *Cell* **2000**, *100* (1), 57−70.

(5) Kaminska, B. MAPK signalling pathways as molecular targets for anti-inflammatory therapy–from molecular mechanisms to therapeutic benefits. *Biochim. Biophys. Acta* **2005**, *1754* (1−2), 253−62.

(6) Peifer, C.; Wagner, G.; Laufer, S. New approaches to the treatment of inflammatory disorders small molecule inhibitors of p38 MAP kinase. *Curr. Top. Med. Chem.* **2006**, *6* (2), 113−49.

(7) White, M. F. Regulating insulin signaling and beta-cell function through IRS proteins. *Can. J. Physiol. Pharmacol.* **2006**, *84* (7), 725−37.

(8) Larsen, M. R.; Thingholm, T. E.; Jensen, O. N.; Roepstorff, P.; Jorgensen, T. J. Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Mol. Cell. Proteomics* **2005**, *4* (7), 873−86.

(9) Collins, M. O.; Yu, L.; Coba, M. P.; Husi, H.; Campuzano, I.; Blackstock, W. P.; Choudhary, J. S.; Grant, S. G. Proteomic analysis of in vivo phosphorylated synaptic proteins. *J. Biol. Chem.* **2005**, *280* (7), 5972−82.

(10) Molina, H.; Horn, D. M.; Tang, N.; Mathivanan, S.; Pandey, A. Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (7), 2199−204.

(11) Wissing, J.; Jansch, L.; Nimtz, M.; Dieterich, G.; Hornberger, R.; Keri, G.; Wehland, J.; Daub, H. Proteomics analysis of protein kinases by target class-selective prefractionation and tandem mass spectrometry. *Mol. Cell. Proteomics* **2007**, *6* (3), 537−47.

(12) Villen, J.; Beausoleil, S. A.; Gerber, S. A.; Gygi, S. P. Large-scale phosphorylation analysis of mouse liver. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (5), 1488−93.

(13) Ballif, B. A.; Villen, J.; Beausoleil, S. A.; Schwartz, D.; Gygi, S. P. Phosphoproteomic analysis of the developing mouse brain. *Mol. Cell. Proteomics* **2004**, *3* (11), 1093−101.

(14) Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P. Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (33), 12130−5.

(15) Ficarro, S. B.; McCleland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat. Biotechnol.* **2002**, *20* (3), 301−5.

(16) Lee, J.; Xu, Y.; Chen, Y.; Sprung, R.; Kim, S. C.; Xie, S.; Zhao, Y. Mitochondrial phosphoproteome revealed by an improved IMAC method and MS/MS/MS. *Mol. Cell. Proteomics* **2007**, *6* (4), 669−76.

(17) Moser, K.; White, F. M. Phosphoproteomic analysis of rat liver by high capacity IMAC and LC−MS/MS. *J. Proteome Res.* **2006**, *5* (1), 98−104.

(18) Trinidad, J. C.; Specht, C. G.; Thalhammer, A.; Schoepfer, R.; Burlingame, A. L. Comprehensive identification of phosphorylation sites in postsynaptic density preparations. *Mol. Cell. Proteomics* **2006**, *5* (5), 914−22.

(19) Li, X.; Gerber, S. A.; Rudner, A. D.; Beausoleil, S. A.; Haas, W.; Villen, J.; Elias, J. E.; Gygi, S. P. Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. *J. Proteome Res.* **2007**, *6* (3), 1190−7.

(20) Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **2009**, *6* (5), 359−62.

(21) Matsumoto, M.; Oyamada, K.; Takahashi, H.; Sato, T.; Hatakeyama, S.; Nakayama, K. I. Large-scale proteomic analysis of tyrosine-phosphorylation induced by T-cell receptor or B-cell receptor activation reveals new signaling pathways. *Proteomics* **2009**, *9* (13), 3549−63.

(22) Taus, T.; Kocher, T.; Pichler, P.; Paschke, C.; Schmidt, A.; Henrich, C.; Mechtler, K. Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* **2011**, *10* (12), 5354−62.

(23) Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **2012**, *11* (3), M111 014050.

(24) Lemeer, S.; Heck, A. J. The phosphoproteomics data explosion. *Curr. Opin. Chem. Biol.* **2009**, *13* (4), 414−20.

(25) Dephoure, N.; Zhou, C.; Villen, J.; Beausoleil, S. A.; Bakalarski, C. E.; Elledge, S. J.; Gygi, S. P. A quantitative atlas of mitotic phosphorylation. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (31), 10762−7.

(26) Martins-de-Souza, D.; Guest, P. C.; Vanattou-Saifoudine, N.; Rahmoune, H.; Bahn, S. Phosphoproteomic differences in major depressive disorder postmortem brains indicate effects on synaptic function. *Eur. Arch. Psychiatry Clin. Neurosci.* **2012**, *262*, 657−666.

(27) Zanivan, S.; Gnad, F.; Wickstrom, S. A.; Geiger, T.; Macek, B.; Cox, J.; Fassler, R.; Mann, M. Solid tumor proteome and phosphoproteome analysis by high resolution mass spectrometry. *J. Proteome Res.* **2008**, *7* (12), 5314−26.

(28) Muraoka, S.; Kume, H.; Watanabe, S.; Adachi, J.; Kuwano, M.; Sato, M.; Kawasaki, N.; Kodera, Y.; Ishitobi, M.; Inaji, H.; Miyamoto, Y.; Kato, K.; Tomonaga, T. Strategy for SRM-based verification of biomarker candidates discovered by iTRAQ method in limited breast cancer tissue samples. *J. Proteome Res.* **2012**, *11* (8), 4201−10.

(29) Vizcaino, J. A.; Cote, R.; Reisinger, F.; Barsnes, H.; Foster, J. M.; Rameseder, J.; Hermjakob, H.; Martens, L. The Proteomics Identifications database: 2010 update. *Nucleic Acids Res.* **2010**, *38* (Database issue), D736−742.

In-depth Membrane Proteomic Study of Breast Cancer Tissues for the Generation of a Chromosome-based Protein List

Satoshi Muraoka,[†] Hideaki Kume,[†] Jun Adachi,[†] Takashi Shiromizu,[†] Shio Watanabe,[†] Takeshi Masuda,[‡] Yasushi Ishihama,[§] and Takeshi Tomonaga*,[†]

[†]Laboratory of Proteome Research, National Institute of Biomedical Innovation, Ibaraki, Osaka, Japan
[‡]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, Japan
[§]Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto, Japan

ⓈSupporting Information

ABSTRACT: The Chromosome-centric Human Proteome Project (C-HPP) aims to define all proteins encoded in each chromosome and especially to identify proteins that currently lack evidence by mass spectrometry. The C-HPP also prioritizes particular protein subsets such as membrane proteins, post-translational modifications, and low-abundance proteins. In this study, we aimed to generate deep profiling of the membrane proteins of human breast cancer tissues on a chromosome-by-chromosome basis using shotgun proteomics. We identified 7092 unique proteins using membrane fractions isolated from pooled breast cancer tissues with high confidence. A total of 3282 proteins were annotated as membrane proteins by Gene Ontology analysis, which covered 45% of the membrane proteins predicted in 20 859 protein-coding genes. Furthermore, we were able to identify 851 membrane proteins that currently lack evidence by mass spectrometry in neXtProt. Our results will contribute to the accomplishment of the primary goal of the C-HPP in identifying so-called "missing proteins" and generating a whole protein catalog for each chromosome.

KEYWORDS: *missing protein, shotgun proteomics, membrane protein, neXtProt, chromosome, C-HPP*
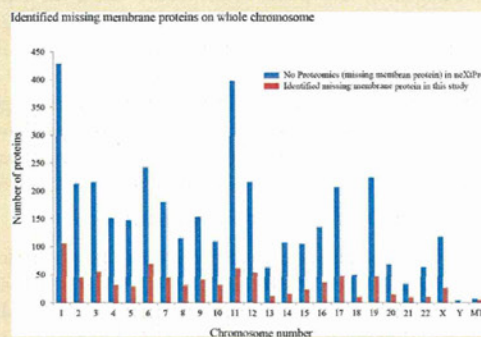
■ INTRODUCTION

Completed in 2003, the Human Genome Project (HGP) was a 13-year project coordinated by the U.S. Department of Energy and the National Institutes of Health.[1] The project was to identify all of the approximately 20 000–25 000 genes in human DNA.[1,2] Results were published as the human genome database. The age of whole-genome sequencing has made the research field of proteomics possible. In 2008, the Human Proteome Organization (HUPO) developed a strategy for the first phase of the human proteome project (HPP). The C-HPP is one component of the HPP and focuses on constructing a proteomic catalog in a chromosome-by-chromosome fashion and aims to define the full set of proteins encoded in whole-chromosomes.[3–5] The initial goal of the C-HPP is to identify and characterize proteins that currently lack MS evidence, referred to as "missing proteins", in neXtProt, a new human protein-centric knowledge platform.[6] "Missing proteins" are likely due to their very low-abundance and/or absence of expression in given cells or tissues. Thus, more in-depth proteomic studies of cell lines and patient tissues are needed.

The C-HPP also underscores the mapping of particular protein subsets such as membrane proteins and/or post-translational modifications. Membrane proteins are of great interest, particularly because they could be key biomarkers for

early diagnosis, progression of diseases, and suitable drug targets; however, there have been difficulties in enrichment/solubilization and also subsequent protease digestion in membrane proteome analysis.[7–9] Recently, several protocols have been reported to increase the solubilization and digestion of proteins, which has greatly improved membrane proteomic analysis of cells and tissues.[10,11]

In this study, to generate a chromosome-based membrane protein list, we integrated membrane proteomic analysis data from human breast cancer tissues with previous data[12] and analyzed with Proteome Discoverer and Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics Resources followed by chromosome-based categorization using the neXtProt database.

■ MATERIALS AND METHODS

Human Tissue Samples

Tissue samples were obtained from 18 patients with high-risk or low-risk MammaPrint breast cancer who underwent surgery at

the Osaka Medical Center for Cancer & Cardiovascular Diseases (Supplementary Figure 1, Supporting Information). All samples were frozen by liquid nitrogen and were stored at −80 °C until analysis. Written informed consent was obtained from all subjects. The Ethics Committee of our institute and the Osaka Medical Center for Cancer & Cardiovascular Diseases approved the protocol.

Enrichment of Membrane Proteins

For enrichment of membrane proteins, frozen tissue samples were homogenized in PBS containing a protease inhibitor mixture (Complete; Roche, Mannheim, Germany) using a Dounce homogenizer (WHEATON, Millville, NJ) following centrifugation (1000× g) for 10 min at 4 °C. The postnuclear supernatant was centrifuged at 100 000× g for 1 h at 4 °C. The pellet was suspended in ice-cold 0.1 M Na_2CO_3 solution following centrifugation (100 000× g) for 1 h at 4 °C. After centrifugation, the pellet was treated using an MPEX PTS reagent kit (GL sciences, Tokyo, Japan) as follows.[10] Briefly, the pellet was solubilized with PTS B buffer at 95 °C for 5 min followed by sonication for 5 min using a Bioruptor sonicator (Cosmo Bio, Tokyo, Japan). The solution was centrifuged at 100 000× g for 30 min at 4 °C. Supernatant containing membrane proteins was stored at −80 °C. Protein concentration was determined using a DC protein assay kit (Bio-Rad, USA).

In Solution Digestion and iTRAQ Labeling

Membrane proteins from pooled high-risk (n = 9) or low-risk (n = 9) breast cancer tissue samples were digested with Lys-C (Wako Pure Chemical Industries, Osaka, Japan), followed by trypsin (Proteomics grade; Roche, Swiss). Tryptic digests were treated according to the PTS protocol and desalted using C18 StageTips.[13] Briefly, a sample of 90 μg of pooled membrane proteins was reduced with 10 mM dithiothreitol (DTT), alkylated with 20 mM iodoacetamide (IAA), and sequentially digested by 1:100 (w/w) LysC (Wako Pure Chemical Industries, Osaka, Japan) for 8 h at 37 °C and 1:100 (w/w) trypsin (proteomics grade; Roche) for 12 h at 37 °C. An equal volume of an organic solvent, ethyl acetate, was added to digested samples, the mixtures were acidified by 1% trifluoroacetic acid (TFA), and vortexed to transfer detergents to the organic phase. After centrifugation, the aqueous phase containing peptides was collected. BSA (0.45 μg) was spiked into membrane protein samples as a quality control for iTRAQ labeling. The tryptic digest sample was desalted using C18 stage Tips. Desalted samples were dissolved in 30 μL of dissolution buffer and labeled with two different iTRAQ reagents at room temperature for 1h and quenched by Milli-Q water. Sample labeling was as follows: high-risk breast cancer tissue samples with 114 tag and low-risk breast cancer tissue samples with 115 tag. Labeled samples were mixed and dried by a Speed-Vac concentrator, dissolved in 100 μL of 2% acetonitrile (ACN), 0.1% formic acid (TFA), and desalted with C18 stage Tips.

Separation with Strong Cation Exchange Chromatography (SCX)

The tryptic peptide sample was fractionated using a HPLC system (Shimadzu prominence UFLC) fitted with a SCX column (50 mm × 2.1 mm, 5 μm, 300 Å, ZORBAX 300SCX, Agilent technology). The mobile phases consisted of (A); 25% ACN with 10 mM KH_2PO_4 (pH 3.0) and (B); (A) containing 1 M KCl. The mixed sample was separated at a flow rate of 200 μL/min using a four-step linear gradient; 0% B for 30 min, 0 to 10% B in 15 min, 10 to 25% B in 10 min, 25 to 40% B in 5 min, and 40 to 100% B in 5 min, and 100% B in 10 min.

NanoLC–MS/MS

NanoLC–MS/MS analysis was conducted by an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a nanoLC interface (AMR, Tokyo, Japan), a nanoHPLC system (Michrom Paradigm MS2), and an HTC-PAL autosampler (CTC, Analytics, Zwingen, Switzerland). L-column2 C18 particles (3 μm) (Chemicals Evaluation and Research Institute (CERI), Japan) were packed into a self-pulled needle (200 mm length × 100 μm inner diameter) using a Nanobaume capillary column packer (Western Fluids Engineering). Mobile phases consisted of (A) 0.1% FA and 2% ACN and (B) 0.1% FA and 90% ACN. SCX-fractionated peptides dissolved in 2% ACN and 0.1% TFA were loaded onto a trap column (0.3 × 5 mm, L-column ODS; CERI). The nanoLC gradient was delivered at 500 nL/min and consisted of a linear gradient of mobile phase B developed from 5 to 30% B in 135 min. A spray voltage of 2000 V was applied.

Data Acquisition with LTQ-Orbitrap Velos

Full MS scans were performed in the orbitrap mass analyzer of LTQ-Orbitrap Velos (scan range 350–1500 m/z, with 30K fwhm resolution at 400 m/z). In MS scans, the ten most intense precursor ions were selected for MS/MS scans of LTQ-Orbitrap Velos respectively, in which a dynamic exclusion option was implemented with a repeat count of one and exclusion duration of 60 s. This was followed by collision-induced dissociation (CID) MS/MS scans of selected ions performed in the linear ion trap mass analyzer, and further followed by higher energy collision-induced dissociation (HCD) MS/MS scans of the same precursor ions performed in the orbitrap mass analyzer with 7500 fwhm resolution at 400 m/z. The values of automated gain control (AGC) were set to $1.00 \times 10^{+06}$ for full MS, $1.00 \times 10^{+04}$ for CID MS/MS, and $5.00 \times 10^{+04}$ for HCD MS/MS. Normalized collision energy values were set to 35% for CID and 50% for HCD. CID, also known as collision-activated dissociation, is performed in the linear ion trap. It is able to increase the number of peptide identifications, and, thus, is applied to obtain peptide sequence information. HCD is performed in the C-trap of the LTQ Orbitrap and is a useful tool for elucidating the structure of small molecules, metabolites, peptides, and PTM peptides, and for de novo sequencing of peptides. It allows quantitative information to be obtained from iTRAQ ions in the lower mass area. By analyzing the sample using a combination of CID with HCD, we are able to obtain the best conditions for both peptide sequencing and iTRAQ quantitation.

Identification and Quantification of Membrane Proteins

CID and HCD raw spectra were extracted and searched separately against UniProtKB/Swiss-Prot (release-2010_05) containing 20 295 sequences of Homo sapiens using Proteome Discoverer (Thermo Fisher Scientific, Beta Version 1.3) and Mascot v2.3.1. Search parameters included trypsin as the enzyme with one missed cleavage allowed; Carbamidomethylation at cysteine and iTRAQ labeling at lysine and the N-terminal residue were set as fixed modifications while oxidation at methionine and iTRAQ labeling at tyrosine were set as variable modifications. Precursor mass tolerance was set to 7 ppm and a fragment mass tolerance was set to 0.6 Da for CID and 0.01 Da for HCD. Protein identification required at least one unique peptide and quantification required at least two peptides. FDR was calculated

by enabling peptide sequence analysis using Percolator. High confidence peptide identification was obtained by setting a target FDR threshold of ≤1.0% at the peptide level. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral. proteomexchange.org) via the PRIDE partner repository (http://www.ebi.ac.uk/pride/) with the data set identifier PXD000066.

Bioinformatics Analysis

The subcellular locations of identified proteins were annotated by DAVID Bioinformatics Resources 6.7, available at http://david.abcc.ncifcrf.gov/home.jsp.[14] The chromosomal locations and missing protein analysis of identified proteins were elucidated by neXtProt, available at http://www.nextprot.org/db/. The function of identified missing membrane proteins was elucidated by the Ingenuity system, available at www.ingenuity.com.

■ RESULTS

C-HPP is collecting protein data identified by the chromosome-independent shotgun approach and then sharing this data

Table 1. Comparison of the Number of Identified Membrane Protein with Our Result and Previously Reported Results

|	Muraoka et al.	Polisetty et al.[7]	Han et al.[17]
Protein identified	7092	1834	1482
Membrane protein	3282	1027	642

according to the chromosome number to ensure a complete parts list.[15] In this study, we integrated membrane proteomic analysis data from human breast cancer tissues and analyzed with Proteome Discoverer and DAVID Bioinformatics Resources and

characterized them on a chromosome-by-chromosome basis using the neXtProt database.

A total of 7092 unique proteins were identified with high confidence. A list of proteins and peptides are presented in Supplementary Tables 1 and 2, Supporting Information. Identified unique proteins were examined with respect to subcellular localization using Gene Ontology annotation analysis in DAVID Bioinformatics Resources. It revealed that 3282 (46%) were annotated to membrane proteins (Supplementary Table 3, Supporting Information), 692 (10%) proteins were extracellular space, 4030 (57%) proteins were cytoplasm proteins, and 1782 (25%) proteins were nucleus proteins by GO analysis. As shown in Table 1, this number of identified membrane proteins is much greater than previously reported.
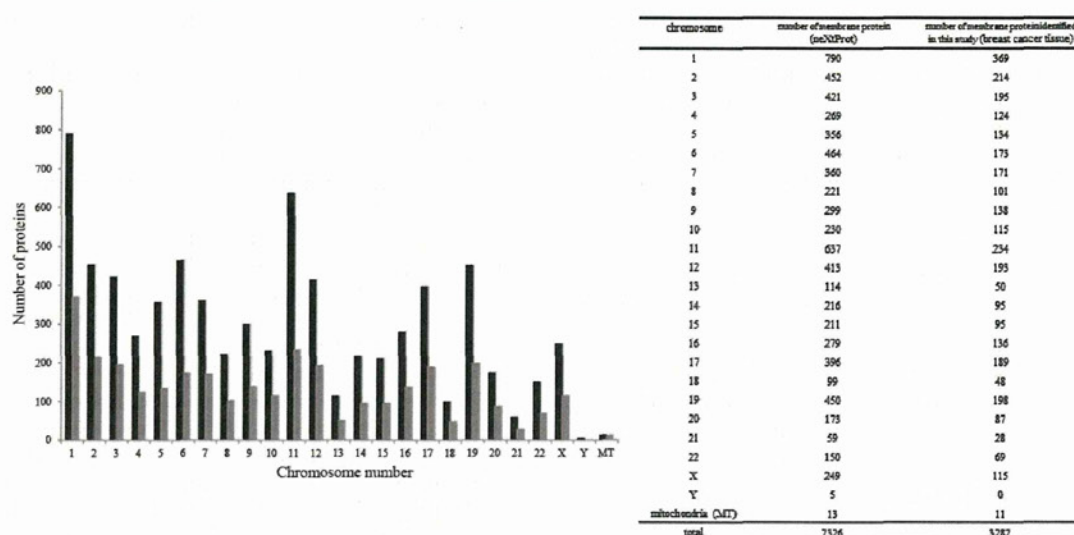
To generate a chromosome-based membrane protein list, the identified 3282 membrane proteins were examined with respect to chromosomal location using the neXtProt database. The chromosomal distribution of protein-coding genes in the neXtProt database, identifying total proteins, and membrane proteins are shown in Figure 1. The neXtProt database annotates 7326 proteins as membrane proteins in the 20 859 protein-coding genes, and surprisingly, we identified 45% of them in this study (Figure 2). These results support the effectiveness of the method to solubilize and digest integral membrane proteins, allowing large-scale detection and identification of this protein class with no bias against membrane proteins.

A primary goal of the C-HPP is to identify and characterize proteins that currently lack MS evidence and are referred to as "missing proteins". Thus, we examined how many missing proteins were identified in this study. We compared our membrane protein list with a list of no proteomic proteins (missing proteins) in the neXtProt database. Surprisingly, 851 membrane missing proteins (22.7%) were identified in this study (Figure 3 and Supplementary Table 3, Supporting Information).
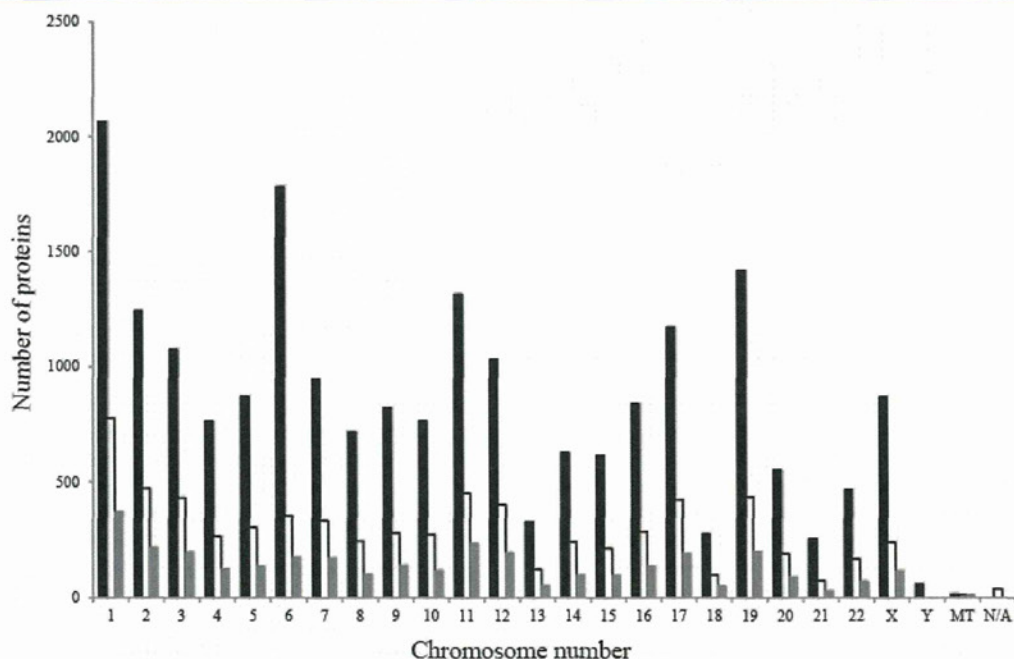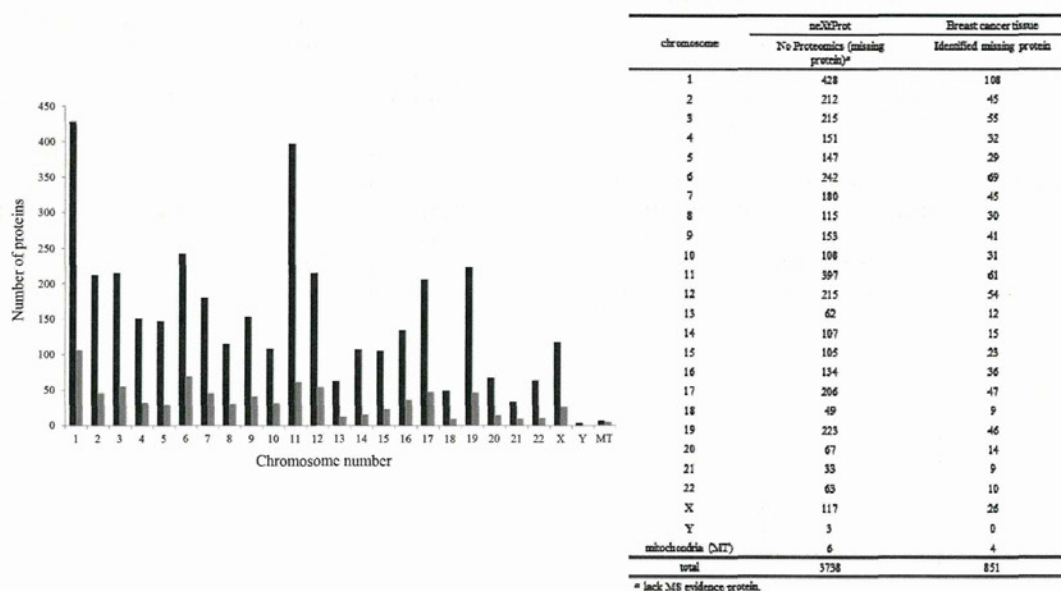

Figure 1. Distribution of identified total and membrane proteins on a whole-chromosomal location. Black bar, neXtProt database proteins; white bar, identified total proteins; gray bar, identified membrane proteins; N/A, no protein in the neXtProt database; MT, mitochondria.

chromosome	number of membrane protein (neXtProt)	number of membrane protein identified in this study (breast cancer tissue)
1	790	369
2	452	214
3	421	195
4	269	124
5	356	134
6	464	173
7	360	171
8	221	101
9	299	138
10	230	115
11	657	234
12	413	193
13	114	50
14	216	95
15	211	95
16	279	136
17	396	189
18	99	48
19	450	198
20	173	87
21	59	28
22	150	69
X	249	115
Y	5	0
mitochondria (MT)	13	11
total	7326	3282

Figure 2. Comparison of chromosome-based membrane proteins annotated by the neXtProt database with identified membrane proteins in this study. (Left) Black bar, neXtProt database membrane proteins; gray bar, identified membrane proteins. (Right) Number of identified and neXtProt database membrane proteins on a whole-chromosomal location.

chromosome	neXtProt No Proteomics (missing protein)[a]	Breast cancer tissue Identified missing protein
1	428	108
2	212	45
3	215	55
4	151	32
5	147	29
6	242	69
7	180	45
8	115	30
9	153	41
10	108	31
11	397	61
12	215	54
13	62	12
14	107	15
15	105	23
16	134	36
17	206	47
18	49	9
19	223	46
20	67	14
21	33	9
22	63	10
X	117	26
Y	3	0
mitochondria (MT)	6	4
total	3738	851

[a] lack MS evidence protein.

Figure 3. Identified and missing membrane proteins on a whole-chromosomal location. (Left) Black bar, no proteomics (missing membrane proteins) in neXtProt; gray bar, identified missing membrane proteins in this study. (Right) Number of identified and missing membrane proteins on a whole chromosome.

Lipid metabolism, small molecule biochemistry, cell-to-cell signaling and interaction, hematological system development and function, and immune cell trafficking were the major molecular and cellular processes identified by IPA (Figure 4). These results indicate that our in-depth membrane proteomic study of breast cancer tissue samples was able to identify and characterize a number of low-abundance missing proteins.

■ DISCUSSION

The objective of C-HPP is to map and annotate all protein-coding genes on each human chromosome, especially so-called "missing proteins", which only have transcriptomic evidence and

a predicted sequence. To accomplish this, deep profiling for low-abundance proteins and subcellular proteins such as membrane proteins is needed. In this study, we performed an in-depth membrane proteomic study of breast cancer tissues. A total of 7092 proteins were identified, of which 3282 proteins were annotated as membrane proteins by Gene Ontology analysis. Furthermore, we could identify not only nearly 50% of the membrane proteins mapped on the whole chromosome but also 851 proteins among the 3738 missing membrane proteins.

Several previously published reports have described membrane proteome analysis.[8,16,17] Polisetty and co-workers recently performed a large-scale proteomic study utilizing shotgun
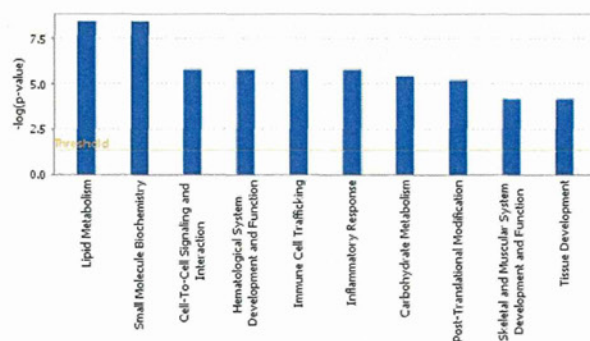
Figure 4. Bar chart indicating the cellular function of proteins found with missing proteins determined using Ingenuity software.

technology and identified 1834 distinct proteins from membrane fractions of glioblastoma multiforme patient specimens, with 56% of them (1027) being annotated as membrane proteins.[7] In this study, we identified a total of 7092 proteins in the membrane fraction; with 46% of them (3282) being known membrane proteins associated with major cellular processes. This number of membrane proteins is much greater than those previously reported. Moreover, we were able to identify a number of missing proteins that currently lack MS evidence. This is probably due to utilization of the PTS method-based isolation of membrane proteins and SCX fractionation before LC−MS/MS analysis. Efficient isolation and solubilization of membrane proteins can be achieved with PTS by allowing the use of a high detergent concentration while avoiding interference with tryptic digestion before LC−MS/MS analysis.[18] SCX prefractionation is also able to improve the number of proteins identified by reducing the complexity of clinical samples and consequently avoiding ion suppression. We have succeeded in large scale identification of membrane proteins and phosphoproteins using the above technique.[12,19]

In conclusion, a subcellular fractionation of membrane proteins would improve low-abundance proteome coverage for identification of missing proteins. Our in-depth membrane proteomic studies of human cancer tissue will greatly contribute to the progression of the C-HPP.

■ ASSOCIATED CONTENT

ⓢ Supporting Information

Supplementary figure and tables. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Laboratory of Proteome Research, National Institute of Biomedical Innovation 7-6-8 Saito-Asagi, Ibaraki City, Osaka 567-0085, Japan. Tel.: +81-72-641-9862. Fax: +81-72-641-9861. E-mail: tomonaga@nibio.go.jp.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ ABBREVIATIONS

C-HPP, The Chromosome-Centric Human Proteome Project; PTS, phase-transfer surfactants; CID, collision-induced dissociation; HCD, higher energy collision-induced dissociation; LC−MS/MS, Liquid chromatography−tandem mass spectrometry; LTQ, linear ion Trap; fwhm, Full Wide at Half Maximum; FDR, false discovery rate

■ REFERENCES

(1) Collins, F. S.; Green, E. D.; Guttmacher, A. E.; Guyer, M. S. A vision for the future of genomics research. *Nature* **2003**, *422* (6934), 835−47.

(2) Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J. P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J. C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R. H.; Wilson, R. K.; Hillier, L. W.; McPherson, J. D.; Marra, M. A.; Mardis, E. R.; Fulton, L. A.; Chinwalla, A. T.; Pepin, K. H.; Gish, W. R.; Chissoe, S. L.; Wendl, M. C.; Delehaunty, K. D.; Miner, T. L.; Delehaunty, A.; Kramer, J. B.; Cook, L. L.; Fulton, R. S.; Johnson, D. L.; Minx, P. J.; Clifton, S. W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J. F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R. A.; Muzny, D. M.; Scherer, S. E.; Bouck, J. B.; Sodergren, E. J.; Worley, K. C.; Rives, C. M.; Gorrell, J. H.; Metzker, M. L.; Naylor, S. L.; Kucherlapati, R. S.; Nelson, D. L.; Weinstock, G. M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D. R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H. M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R. W.; Federspiel, N. A.; Abola, A. P.; Proctor, M. J.; Myers, R. M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D. R.; Olson, M. V.; Kaul, R.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G. A.; Athanasiou, M.; Schultz, R.; Roe, B. A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W. R.; de la Bastide, M.; Dedhia, N.; Blocker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J. A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D. G.; Burge, C. B.; Cerutti, L.; Chen, H. C.; Church, D.; Clamp, M.; Copley, R. R.; Doerks, T.; Eddy, S. R.; Eichler, E. E.; Furey, T. S.; Galagan, J.; Gilbert, J. G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L. S.; Jones, T. A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W. J.; Kitts, P.; Koonin, E. V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T. M.; McLysaght, A.; Mikkelsen, T.; Moran, J. V.; Mulder, N.; Pollara, V. J.; Ponting, C. P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A. F.; Stupka, E.; Szustakowski, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y. I.; Wolfe, K. H.; Yang, S. P.; Yeh, R. F.; Collins, F.; Guyer, M. S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K. A.; Patrinos, A.; Morgan, M. J.; de Jong, P.; Catanese, J. J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y. J. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409* (6822), 860−921.

(3) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; Deutsch, E. W.; Domon, B.; Hancock, W.; He, F.; Hochstrasser, D.; Marko-Varga, G.; Salekdeh, G. H.; Sechi, S.; Snyder, M.; Srivastava, S.; Uhlen, M.; Wu, C. H.; Yamamoto, T.; Paik, Y. K.; Omenn, G. S. The

human proteome project: current state and future direction. *Mol. Cell. Proteomics* **2011**, *10* (7), M111 009993.

(4) Hancock, W.; Omenn, G.; Legrain, P.; Paik, Y. K. Proteomics, human proteome project, and chromosomes. *J. Proteome Res.* **2011**, *10* (1), 210.

(5) 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.

(6) 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.

(7) Polisetty, R. V.; Gautam, P.; Sharma, R.; Harsha, H. C.; Nair, S. C.; Gupta, M. K.; Uppin, M. S.; Challa, S.; Puligopu, A. K.; Ankathi, P.; Purohit, A. K.; Chandak, G. R.; Pandey, A.; Sirdeshmukh, R. LC-MS/MS analysis of differentially expressed glioblastoma membrane proteome reveals altered calcium signaling and other protein groups of regulatory functions. *Mol. Cell. Proteomics* **2012**, *11* (6), M111 013565.

(8) Josic, D.; Clifton, J. G. Mammalian plasma membrane proteomics. *Proteomics* **2007**, *7* (16), 3010−29.

(9) Russell, W. K.; Park, Z. Y.; Russell, D. H. Proteolysis in mixed organic-aqueous solvent systems: applications for peptide mass mapping using mass spectrometry. *Anal. Chem.* **2001**, *73* (11), 2682−5.

(10) Masuda, T.; Tomita, M.; Ishihama, Y. Phase transfer surfactant-aided trypsin digestion for membrane proteome analysis. *J. Proteome Res.* **2008**, *7* (2), 731−40.

(11) Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **2009**, *6* (5), 359−62.

(12) Muraoka, S.; Kume, H.; Watanabe, S.; Adachi, J.; Kuwano, M.; Sato, M.; Kawasaki, N.; Kodera, Y.; Ishitobi, M.; Inaji, H.; Miyamoto, Y.; Kato, K.; Tomonaga, T. Strategy for SRM-based Verification of Biomarker Candidates Discovered by iTRAQ Method in Limited Breast Cancer Tissue Samples. *J. Proteome Res.* **2012**, *11* (8), 4201−10.

(13) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2007**, *2* (8), 1896−906.

(14) Huang da, W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2009**, *4* (1), 44−57.

(15) 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.

(16) Chen, J. S.; Chen, K. T.; Fan, C. W.; Han, C. L.; Chen, Y. J.; Yu, J. S.; Chang, Y. S.; Chien, C. W.; Wu, C. P.; Hung, R. P.; Chan, E. C. Comparison of membrane fraction proteomic profiles of normal and cancerous human colorectal tissues with gel-assisted digestion and iTRAQ labeling mass spectrometry. *FEBS J.* **2010**, *277* (14), 3028−38.

(17) Han, C. L.; Chen, J. S.; Chan, E. C.; Wu, C. P.; Yu, K. H.; Chen, K. T.; Tsou, C. C.; Tsai, C. F.; Chien, C. W.; Kuo, Y. B.; Lin, P. Y.; Yu, J. S.; Hsueh, C.; Chen, M. C.; Chan, C. C.; Chang, Y. S.; Chen, Y. J. An informatics-assisted label-free approach for personalized tissue membrane proteomics: case study on colorectal cancer. *Mol. Cell. Proteomics* **2011**, *10* (4), M110 003087.

(18) Iwasaki, M.; Masuda, T.; Tomita, M.; Ishihama, Y. Chemical cleavage-assisted tryptic digestion for membrane proteome analysis. *J. Proteome Res.* **2009**, *8* (6), 3169−75.

(19) Narumi, R.; Murakami, T.; Kuga, T.; Adachi, J.; Shiromizu, T.; Muraoka, S.; Kume, H.; Kodera, Y.; Matsumoto, M.; Nakayama, K.; Miyamoto, Y.; Ishitobi, M.; Inaji, H.; Kato, K.; Tomonaga, T. A strategy for large-scale phosphoproteomics and SRM-based validation of human breast cancer tissue samples. *J. Proteome Res.* **2012**, *11* (11), 5311−22.

Integrated View of the Human Chromosome X-centric Proteome Project

Tadashi Yamamoto,*,[†] Keiichi Nakayama,[‡] Hisashi Hirano,[§] Takeshi Tomonaga,[||] Yasushi Ishihama,[⊥] Tetsushi Yamada,[#] Tadashi Kondo,[#] Yoshio Kodera,[□] Yuichi Sato,[□] Norie Araki,[¶] Hiroshi Mamitsuka,[○] and Naoki Goshima[●]

[†]Institute of Nephrology, Graduate School of Medical and Dental Sciences, Niigata, Japan
[‡]Medical Institute of Bioregulation, Kyushu University, Kyushu, Japan
[§]Graduate School of Nanobioscience, Yokohama City University, Yokohama, Japan
[||]Laboratory of Proteome Research, National Institute of Biomedical Innovation, Ibaraki, Japan
[⊥]Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan
[#]National Cancer Center Research Institute, Tokyo, Japan
[□]Kitasato University, Kanagawa, Japan
[¶]Graduate School of Medical Sciences, Kumamoto University, Kumamoto, Japan
[○]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Japan
[●]National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

ABSTRACT: This article introduces how the human chromosome X-centric proteome project is carried out by the Japan Chromosome X Project Consortium. The inactivation of one of two chromosomes in female mammals and accumulation of genes related to neural/immune systems/tumor/testis are characteristic of chromosome X. In this Chromosome X Project, information on proteins translated from genes on chromosome X is collected by both mass spectrometry- and antibody-based proteomics. Information on the following resources is also provided: antibodies to proteins translated and full-length cDNAs transcribed from the chromosome X genes for recombinant proteins. The consortium aims to provide the following tools to search useful antibodies in the literature (Antibody Ranker), to find gene expression sites in microarray databases (Transcript Localizer) and to do advanced MRM analysis (information-based MRM).

KEYWORDS: *chromosome X, mass spectrometry, antibody, MRM*
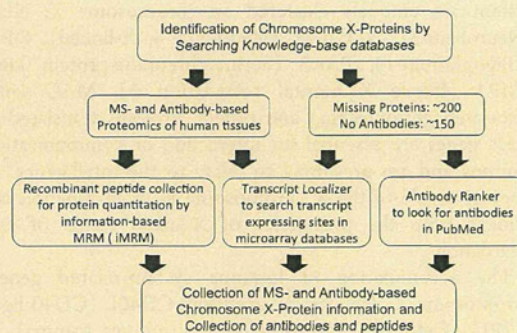
■ INTRODUCTION

Chromosome X is one of the two sex-determining chromosomes (the other is the Y chromosome) in many animal species, including mammals, and is found in both males and females. This chromosome has the following characteristics, which are not found in other chromosomes: presence of sex-determining genes such as androgen receptor gene, inactivation of one of two chromosomes (X-inactivation), a large number of genes associated with human hereditary diseases, and significant accumulation of genes for networks in neuronal, immune and tumor-related systems.

The Japanese Proteomics Society (JHUPO) was chosen to participate in the Chromosome-centric Human Proteome Project and was asked to be in charge of chromosome X. The Japan Chromosome X Project Consortium (JCXPC) was organized to complete the project by collecting information and resources of all proteins translated from genes located on

chromosome X (chromosome X-proteins). Knowledge of the human chromosome X-proteins and the current activities and aims of the chromosome X project are briefly described.

■ KNOWLEDGE OF HUMAN CHROMOSOME X

1. Characteristics

The human chromosome X spans about 153 million base pairs (4.94%) out of 3100 million base pairs of total human DNA length. The number of genes on chromosome X is presumed to be 888 genes (4.37%) out of total human genes of 20300.[1] The other sex-determining chromosome, Y chromosome, has only 68 genes and one of these genes, named SRY (Sex-determining