**Proteogenomic analysis to identify missing proteins from haploid cell lines**

Seung-Eun Lee1, 7, JongKeon Song1, Korbinian Bösl2, André C. Müller3, Dijana Vitko3, Keiryn L. Bennett3, Giulio Superti-Furga3,4, Akhilesh Pandey1,5, Richard K. Kandasamy3,\* and Min-Sik Kim1,6,7\*

1 Department of Applied Chemistry, Kyung Hee University, Republic of Korea

2 Centre of Molecular Inflammation Research (CEMIR), Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, 7491 Trondheim, Norway

3 CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, A-1090 Vienna, Austria

4 Center for Physiology and Pharmacology, Medical University of Vienna, Vienna, Austria

5 Department of Biological Chemistry, Johns Hopkins University School of Medicine, Baltimore, USAp

6 Global Center for Pharmaceutical Ingredient Materials, Kyung Hee University, Republic of Korea

7 Department of Biomedical Science and Technology, Kyung Hee Medical Science Research Institute, Kyung Hee University, Republic of Korea

\*Co-correspondence

To whom correspondence should be addressed:

Richard K. Kandasamy

Centre of Molecular Inflammation Research (CEMIR), Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, 7491 Trondheim, Norway; Tel: +47-728-24511; E-mail: richard.k.kandasamy@ntnu.no

Min-Sik Kim

Department of Applied Chemistry, Kyung Hee University, Yongin-si, Gyeonggi-do 446-701, South Korea ; Tel: 82-31-201-3899; Fax: 82-31-201-2340; E-mail: ms\_kim@khu.ac.kr

**Abbreviations**

C-HPP, Chromosome-based Human Proteome Project; PE, protein evidence; RT, room temperature; FASP, filter-aided sample preparation method; IAA, iodoacetamide; TEAB, triethyl ammonium bicarbonate; FPKM, fragments per kilobase million; UV, ultra-violet; AGC, automatic gain control; p.p.m., parts per million; PSM, peptide-spectrum match

**Keywords**

Missing protein, RNA-Seq, lncRNA, Proteogenomics, Haploid cell lines

**Total number of words:** 5,441

**Abstract**

Chromosome-centric Human Proteome Project aims at identifying and characterizing protein products encoded from all human protein-coding genes. As of early 2017, 19,837 protein-coding genes have been annotated in the neXtProt database including 2,691 missing proteins that have never been identified by mass spectrometry. Missing proteins may be low abundant in many cell types or expressed only in a few cell types in human body such as sperms in testis. In this study, we performed expression proteomics of two near haploid cell types such as HAP1 and KBM-7 to hunt for missing proteins. Proteomes from the two haploid cell lines were analyzed on an LTQ Orbitrap Velos, producing a total of 200 raw mass spectrometry files. After applying 1% false discovery rates at both levels of peptide-spectrum matches and proteins, more than ten thousand proteins were identified from HAP1 and KBM-7, resulting in the identification of nine missing proteins. Next, unmatched spectra were searched against protein databases translated in three frames from non-coding RNAs derived from RNA-Seq data, resulting in 6 novel protein-coding regions after careful manual inspection. This study demonstrates that expression proteomics coupled to proteogenomic analysis can be employed to identify many annotated and unannotated missing proteins.

**Significance**

A total of 200 raw mass spectrometry files were acquired from two haploid cell types (HAP1 and KBM-7) by an Orbitrap mass spectrometer to search missing proteins that have never been identified, resulting in the identification of ~10,000 proteins including nine missing proteins. The proteomes of the cell types have not yet been explored in the field of proteomics, thereby making it useful resource for the C-HPP community. In addition, we conducted the proteogenomic analysis by using specific RNA-seq data of the cell types. To do so, we collected unmatched spectra from the proteomic result and created custom protein databases translated in three frames from abundant non-coding RNAs of the haploid cells. This proteogenomic analysis resulted in the identification of 6 novel protein-coding regions in the current human genome. In conclusion, this comprehensive proteogenomic study identifies missing proteins from haploid cell lines that are annotated as well as unannotated in the human genome.

**Introduction**

The Human Proteome Project (HPP) aims to discover all the human proteins encoded by the human genome. Chromosome-based Human Proteome Project (C-HPP) is one of major initiatives in the international Human Proteome Organization (HUPO) and its main goal is to catalogue all proteins and their functions by separating tasks based on chromosome.[1,2] This type of initiatives need to be well-organized and spend a considerable amount of their efforts on maintaining the information exchange. To be coincided efforts on creating and maintaining web resources such as PeptideAtlas[3], GPMDB[4], and MissingProteinPedia[5], neXtProt[6] are being maintained to share information for all C-HPP teams to be at the same pages.[6,7] In this way, these databases help all the researchers to have their efforts non redundant.

About 20,000 representative proteins are expected to be encoded by the human genome, although there is no final number of proteins expressed in human. Information flow from gene to RNA to protein is complicated since there are diverse layers of regulation between molecules and their information.[8–11] This means the corresponding protein may not be expressed although a RNA for a protein-coding gene exists. Thus, it has become essential to directly sequence proteins themselves rather than to assume protein expression based on RNA expression. Despite many studies, there are still considerable number of proteins that were annotated as missed so called ‘missing proteins.’ Missing proteins have been continuously been found over years: 3,868 in 2013, and 2,949 in 2016.[5] As of mid-June 2017, 2,563 including 1,912 for PE2, 555 for PE3 and 96 for PE4) missing proteins are listed in neXtProt (https://www.nextprot.org/about/protein-existence). There are many reports that reasonably speculate why missing proteins have been difficult to identify in spite of a large number of proteomic studies.[12] First, some are very low abundant and/or transient in their expression although they are expressed in many cell types. Second, some probably are expressed only in a few cell types in the body while other missing proteins may be expressed only in fetus. Third, current bioinformatics tools may not be suitable to identify some missing proteins. According to these perspectives, we need to choose various approaches to identify missing proteins: more fractionation, higher scan speed, unexplored samples/cell types, and novel bioinformatics tools.

Despite such difficulties, many missing proteins were found in unexplored samples and cell types such as immune cells, testis and sperms. For example, testis tissue was reported to provide many missing proteins due to their own functions such as spermatogenesis.[13] Moreover sperms that are haploid cells were analyzed to identify missing proteins because it is expected germ-cell specific gene expressions.[14] We have also reported a large fraction of proteins in the initial draft of the human proteome in testis, ovary, and several hematopoietic cells.[14–16] So, we focused on HAP1 and KBM-7 cell lines whose proteomes are largely unexplored and may represent the haploid condition of blood cancer.[17,18] We hypothesis there may be some missing proteins expressed in haploid cells. KBM-7 was originally isolated from a chronic myeloid leukemia and has single copy of most chromosomes so called nearly haploid cell while chromosome 8 and part of chromosome 15 are disomic.[19] HAP1 derived from KBM-7 is a near-haploid cell line because it has further lost a copy of chromosome 8.[18] These two cell lines are often used in in biomedical research and genetic experiments such as targeted genetics screens due to their nature of haploidy.[20]

In this study, we analyzed the two haploid cell lines (i.e. HAP1 and KBM-7) whose proteomes have not been extensively explored to identify missing proteins. Since these cells are haploid, it may be expected that these cells would have experienced a different arrays of gene expression at the levels of proteome and transcriptome. Thus, we acquired new proteomics data on these cell types and collected RNA-Seq data of these cells for proteogenomics analysis. We identified >9,000 proteins and >10,000 transcripts for these cells, identifying 9 missing annotated proteins and discovering 6 novel protein-coding regions.

**Materials and Method**

**Cells**

HAP1 and KBM-7 cells were grown in IMDM (Gibco, Carlsbad, CA, USA) containing 10% fetal bovine serum (Gibco), 100 U/ml Penicillin (Sigma, St Louis, MO, USA), and 100 μg/ml streptomycin (Sigma) at 37 °C and 5% CO2.

**Sample preparation**

Approximately 2 × 107 HAP1 or KBM-7 cells were collected and washed 3 times with phosphate buffer saline (PBS), with a centrifugation steps in between to separate the supernatant containing residual cell culture media. Pellet was lysed in 500 µL 50 mM HEPES buffer (Merck, KgaA, Darmstadt, DE), pH 8.5, supplemented with 2% sodium dodecyl sulphate (SDS; SERVA Electrophoresis GmbH, DE) for 30 min at room temperature (RT). Additionally, DNA was sheared by sonication to reduce lysate viscosity and allow an accurate measurement of the protein concentration (Pierce™ Protein Assay Kit, Pierce Biotechnology, Rockford, IL). Aliquots containing 3 × 100 µg of protein was further digested following an adapted filter-aided sample preparation method (FASP).[21,22] Briefly, disulfide bridges were reduced with 100 mM dithiothreitol (DTT) at 99°C for 5 min and transferred into the VIVACON 500 filter units (Vivaproucts Inc., Littleton, MA). The SDS-containing buffer was removed by ultrafiltration facilitated by centrifugation and, in several washing steps, replaced with 8 M urea in 100 mM Tris-HCl. Proteins were alkylated on cysteine residues with 50 mM iodoacetamide (IAA) and washed with 50 mM triethyl ammonium bicarbonate (TEAB) to remove urea prior to protein digestion. DTT, IAA and TEAB were all purchased from SIGMA-Aldrich (SIGMA-Aldrich Chemie, GmbH, DE). The proteins were incubated overnight at 37°C with modified porcine trypsin (Promega Corp., Madison, WI) in enzyme to protein ratio of 1:100 w/w. The resultant peptides were collected as a filtrates and pooled together.

**RNA-Seq Analysis**

RNA-Seq data of HAP1 and KBM-7 cells is part of the publication by Essletzbichler et al 2014.[18] The associated FASTQ files were downloaded from NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) with SRA identification SRP044391. The sequence reads were aligned to human reference genome build hg19 using TopHat2[23] and Bowtie2[24]. FPKM (fragments per kilobase million) values were computed using Cufflinks[25].

**LC-MS/MS analysis**

Tryptic digest was concentrated by solid phase extraction (SPE) (MicroSpin columns 3-300 µg capacity, The Nest Group Inc., Southborough, MA, USA). Sample containing 100 µg of peptides was basified with 20 mM ammonium formate prior to injection onto a Phenomenex column (150×2.0 mm Gemini®NX-C18 3 µm 110Å; Phenomenex, Torrance, CA, USA) on an Agilent 1200 series HPLC (Agilent Biotechnologies, Palo Alto, CA) with UV detection at 214 nm. HPLC solvents consisted of 20 mM ammonium formate in 5% acetonitrile, pH 10 (solvent A) and 20 mM ammonium formate in 90% acetonitrile, pH 10 (solvent B). Peptides were separated at flow rate of 100 µL/min and eluted from the column with a non-linear gradient ranging from 0 to 100% solvent B. Seventy two time-based fractions were acidified and pooled into 50 HPLC vials. After removal of organic solvent in a vacuum centrifuge, sample was reconstituted with 5% formic acid to different volumes based on the intensity of the UV trace. Individual fractions were further separated and analysed at pH 2.4 by an Agilent 1200 nano-HPLC system (Agilent Biotechnologies, Palo Alto, CA) coupled to a hybrid linear trap quadrupole (LTQ) Orbitrap Velos mass spectrometer (ThermoFisher Scientific, Waltham, MA) utilising Xcalibur software version 2.1. for data acquisition. Briefly, single fractions were loaded onto a trap column (Zorbax 300SB-C18 5 μm, 5 × 0.3 mm, Agilent Biotechnologies, Palo Alto, CA) with a binary pump at a flow rate of 45 μL/min. Solvents for LCMS separation were composed of 0.1% trifluoracetic acid (TFA) in water (solvent A) and 0.1% TFA in 70% methanol and 20% isopropanol (solvent B). The peptides were eluted by back-flushing from the trap column onto a 16 cm fused silica analytical column with an inner diameter of 50 μm packed with C18 reversed-phase material (ReproSil-Pur 120 C18-AQ, 3 μm, Dr. Maisch GmbH, Ammerbuch-Entringen, Germany). Elution was achieved with a 27 min gradient ranging from 3 to 30% solvent B, followed by a 25 min gradient from 30 to 70% solvent B and, finally, a 7 min gradient from 70 to 100% solvent B at a constant flow rate of 100 nL/min.

The MS analysis was performed in a data-dependent acquisition mode using 15 most intense peaks for collision-induced dissociation (CID) and subsequent peptide identification. Dynamic exclusion for selected ions for fragmentation was set to 60 s and a single lock mass at m/z 445.120024 (Si(CH3)2O)6)20 was used for internal mass calibration with the target loss mass abundance of 0%. Maximal ion accumulation time allowed was 50 ms and overfilling of the C-trap was prevented by automatic gain control (AGC) set to 106 ions for a full FTMS scan and 5×104 ions for MS2 CID. Intact peptides were detected in the Orbitrap mass analyzer at resolution of 60,000 with the signal threshold of 2,000 counts for triggering an MS2 event.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD006614. To access the data please visit the PRIDE website by using username: reviewer67132@ebi.ac.uk and password: 1r2gjkWT.

**Database searching**

LC-MS/MS data were searched against neXtProt protein sequence database, containing 42,135 protein entries from http://neXtProt.org/; release date: 01-23-2017. Our overall protein database searches were performed using Proteome discoverer 2.1 (Thermo Scientific, SEQUEST HT search algorithm)[26] and MaxQuant (Max Plank Institute of biochemistry, Andromeda search algorithm)[27], and we used protein sequence database from neXtProt all chromosome protein sequence. Protein sequence database headers were parsed using Linux sed scripts (stream editor) to remove unnecessary information. We used following common SEQUEST search parameters: 10 p.p.m. and 0.6 Da for mass tolerances for precursor and fragments, respectively; fully tryptic peptides with up to two missed cleavages for enzymatic cleavage. Peptide N-terminal acetylation and methionine oxidation set as variable modifications while cysteine carbamidomethylation set as a fixed modification. The minimum peptide length was 6 amino acids, maximum peptide length was 30 amino acids for the initial identification prior to applying to the C-HPP guideline. Similarly, we set MaxQuant search parameters as follows: digestion enzyme, missed cleavage and variable modification were assigned to the same values as SEQUEST search parameters. The number of max modification set as 5. Label-free quantification was performed with iBAQ algorithm and re-quantify was checked for more accurate quantification. The data was searched decoy database (reversed sequence database) and the result using false discovery rate estimation. To peptide validation filtered using FDR and q-value, peptide spectrum match (PSM) were less than 1% FDR and PSMs q-value is greater than equal to 0.01 as a high confidence PSMs. Unmatched spectra from SEQUEST searches were collected and then searched against a protein database derived from RNA-Seq data (FPKM is greater than 0.1).

**Proteogenomics analysis**

To identify unannotated proteins by proteogenomics analysis, there were two required documents including sample-specific RNA data-based protein database and unmatched MS/MS spectra. To do so, we have used non-coding RNAs with ≥ 0.1 of FPKM value to create 3-frame translated protein database. The collected non-coding RNAs with relatively higher expression were searched for the corresponding RefSeq ID (i.e. XM, XR, or NR) from which RNA sequences were fetched to create FASTA formatted high abundant non-coding RNA sequences. This RNA sequences were translated in 3-frame using a ‘Six-frame translation’ tool with the ‘Forward’ option in MaxQuant platform. We also collected unmatched MS/MS spectra from the 200 raw MS files after database searching and created a peaklist file. Unmatched MS/MS spectra in the mgf file were searched against the proteins translated in 3-frame from high abundant non-coding RNAs.

**Results and Discussion**

Our goal is to identify missing proteins by analyzing transcriptomes and proteomes of two haploid cell lines, HAP1 and KBM-7. We hypothesize that some of proteins annotated as missing proteins may be expressed in these specialized cell lines. As shown in Figure 1, we have produced a total of 200 raw mass spectrometry files using an LTQ Orbitrap Velos high resolution mass spectrometer, while transcriptomics data were collected from the previous study by Essletzbichler et al.

**Transcriptome and proteome profiles of HAP1 and KBM-7**

We analyzed the RNA-Seq data collected from Essletzbichler et al. and computed the FPKM values for the transcripts. More than ten thousand genes were (with an FPKM value greater than 0.1) found to be expressed in HAP1 and KBM-7 (i.e. 14,975 genes in HAP1 replicate 1, 14,978 genes in HAP1 replicate 2, 14,577 genes in KBM-7 replicate 1 and 14,713 genes in KBM-7 replicate 2). This resulted in a total of 15,480 genes in HAP1 and a total of 15,162 genes in KBM-7, respectively. The full list of expressed genes of these two cell lines is presented in Supporting Information Table S1. A total of 14,473 genes (~93%**)** and 14,128 (~93%) genes were found to be overlapped between replicates in HAP1 and KBM-7. (Figure 2A and 2B)

A complex mixture of peptides were separated into 50 fractions by high pH RPLC, each of which was then analyzed on an Orbitrap mass analyzer. Mass spectrometry data were searched by two common database searching algorithms including Andromeda on MaxQuant and SEQUEST on Proteome Discoverer. When applying 1% FDR at both PSM and protein levels, Andromeda search resulted in the identification of 128,034 peptides mapped on 9,028 proteins - 8,301 in HAP1 and 8,624 KBM-7, respectively (Figure 2C and 2D) while SEQUEST search resulted in 124,782 peptides mapped on 10,578 proteins - 9,363 in HAP1 and 9,795 in KBM-7 (Figure 2E and 2F).

Of proteins identified, only 783 or 1,215 proteins were identified uniquely in HAP1 or in KBM-7, respectively. This may indicate that most of proteins identified were common between the two cell lines. A total of 7,704 (~93%) and 8,336 (~97%) proteins were found to be overlapped between replicates in HAP1 and KBM-7, respectively. The full list of proteins observed in these two cell lines is presented in Supporting Information Table S2 and peptides identified by MaxQuant and SEQUEST are presented in Supporting Information Table S3 and S4, respectively.

**Comparative analysis of transcripts and proteins between HAP1 and KBM-7**

Spearman’s correlation showed the very similar gene expression aspects between replicates while a little difference in gene expression between the two cell lines was observed. As shown in Figure 3A-3C, spearman’s correlation coefficient between HAP1 and KBM-7 was 0.85 while that for replicates of HAP1 and KBM-7 were 0.94 and 0.92, respectively. Although HAP1 and KBM-7 are closely related to each other, technical variation between duplicate experiments were found to be much less variable than biological variation, showing that alteration of proteome expression may be highly impacted by genetic alteration.

Correlation of the identified protein abundance between replicate data sets was higher than that for two different haploid cells. As shown in Figure 3D-3F, correlation between HAP1 and KBM-7 was 0.90 while that for replicates of HAP1 and KBM-7 were 0.97 and 0.98, respectively. Similar to transcriptome data, technical variation between duplicate proteome experiments were found to be much smaller than biological variation.

We compared FPKM-based transcript abundance with iBAQ-based corresponding protein abundance. The commonly expressed genes between transcriptome and proteome were 9,051. Scatter plots with transcript abundance verse corresponding protein abundance was depicted in Figure 3G and 3H. Spearman’s correlation of RNA and protein expressed in HAP1 and KBM-7 was 0.67 and 0.64, respectively, indicating that the genetic influence on expressed RNAs and proteins is similar.

Since HAP1 is a near-haploid cell lines because HAP1 further lost a copy of chromosome 8 from KBM7, we were curious whether or not protein expression on chromosome 8 were reduced. To see the effect of the deletion of a copy of chromosome 8 on protein abundance, we compared all proteins derived from genes on this chromosome. As shown in Figure 4, iBAQ values were generally skewed towards KBM-7 for chromosome 8 when compared to all other genes identified. That is, protein expression of genes on chromosome 8 was generally higher in KBM-7 than HAP1. This result may indicate that the genetic deletion of genes lead to decrease in corresponding protein expression.

**Identification of ‘missing proteins’**

The missing proteins were identified according to Human Proteome Project Data Interpretation Guidelines. The criteria threshold is stricter than before, stating that a protein should have at least two unique peptides of ≥ 9 amino acids identified based on ≤1% FDR at the PSM and protein levels. In this experiment, we identified 62 missing proteins which have only one unique peptide with ≥ 9 amino acids. When applying additional criteria (i.e. ≥ two peptides and manual inspection of each PSM), the number of identification of missing proteins was reduced to 9. All PSMs satisfying the filters are shown in Supporting Information Figure S1. The full list of missing proteins is listed in Table 1.

**Identification of novel protein-coding potentials from non-coding RNAs**

Previously, we have reported that over hundreds of novel protein-coding genes that were not annotated in the current human genome implying that many proteins may be identified in new types of biological samples[16]. Here, we have performed the similar proteogenomic analysis by collecting the unmatched mass spectra from the mass spectrometry data and by fetching the non-coding RNAs with an FPKM value higher than 0.1. When searching the unmatched spectra against a protein database of 3-frame translated non-coding RNAs, we identified 10 and 14 peptides from HAP1 and KBM-7, respectively. Manual inspection of the PSMs resulted in the identification of 6 novel protein-coding genomic regions. For example, the peptide GAPEPAQTQPQPQPQPAAPEGPEQPR identified from KBM-7 is shown in the Figure 5 (PSM, peptide, and BLAT search result). PSMs manually validated for these novel peptides are presented in (Supporting Information Figure S2). Although further study may have to be conducted to find their cellular functions, this analysis illustrates that proteogenomic analysis can help identify ‘unannotated’ missing proteins.

**Concluding remarks**

One of major goals of the C-HPP initiative is to identify all protein-coding genes on human chromosomes. To do so, researchers at the C-HPP group have tried to find missing proteins that may be low abundant and/or expressed only in a few cell types. As part of the tremendous efforts, we also have tried to seek for missing proteins in a new type of unexplored cells: haploid cells Although this study resulted in the identification of more than 10,000 proteins, we were only able to identify nine missing proteins, indicating that there may be more missing proteins in unexplored cell types in human body. When looking at FPKM values for all missing proteins, we found that corresponding mRNAs for more than 300 missing proteins have FPKM values with ≥ 1. Interestingly, seven of nine identified missing genes were in this category. To identify missing proteins, therefore, it may be important to search for samples with high expression of corresponding mRNAs. We were interested in what could be functions for these proteins and thus we have looked at GO molecular functions in the neXtProt database. Although it was not clearly defined where their functions are specifically related to the cell types, some of them might work as transcription factor or co-factor. Expected functions for all but *BEND4* gene were shown in Table 1. To further identify ‘unannotated’ missing proteins, we carried out proteogenomic analysis resulting in six novel protein-coding regions in the human genome. Of six, two were found in lncRNAs while the others in known genes including two peptides from ORF with >100 amino acids. This means that the two lncRNAs are protein-coding genes and their current classification should be revised. In conclusion, this study successfully demonstrates that expression proteomics coupled to proteogenomic analysis of unexplored cell types can be employed to identify both annotated and unannotated missing proteins in the human genome.

**References**

[1] W. Hancock, G. Omenn, P. Legrain, Y.-K. Paik, *J. Proteome Res.* **2011**, *10*, 210.

[2] Y.-K. Paik, S.-K. Jeong, G. S. Omenn, M. Uhlen, S. Hanash, S. Y. Cho, H.-J. Lee, K. Na, E.-Y. Choi, F. Yan, F. Zhang, Y. Zhang, M. Snyder, Y. Cheng, R. Chen, G. Marko-Varga, E. W. Deutsch, H. Kim, J.-Y. Kwon, R. Aebersold, A. Bairoch, A. D. Taylor, K. Y. Kim, E.-Y. Lee, D. Hochstrasser, P. Legrain, W. S. Hancock, *Nat. Biotechnol.* **2012**, *30*, 221.

[3] F. Desiere, E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich, R. Aebersold, *Nucleic Acids Res.* **2006**, *34*, D655.

[4] R. Craig, J. P. Cortens, R. C. Beavis, *J. Proteome Res.* **2004**, *3*, 1234.

[5] M. S. Baker, S. B. Ahn, A. Mohamedali, M. T. Islam, D. Cantor, P. D. Verhaert, S. Fanayan, S. Sharma, E. C. Nice, M. Connor, S. Ranganathan, *Nat. Commun.* **2017**, *8*, 14271.

[6] L. Lane, G. Argoud-Puy, A. Britan, I. Cusin, P. D. Duek, O. Evalet, A. Gateau, P. Gaudet, A. Gleizes, A. Masselot, C. Zwahlen, A. Bairoch, *Nucleic Acids Res.* **2012**, *40*, D76.

[7] P. Gaudet, G. Argoud-Puy, I. Cusin, P. Duek, O. Evalet, A. Gateau, A. Gleizes, M. Pereira, M. Zahn-Zabal, C. Zwahlen, A. Bairoch, L. Lane, *J. Proteome Res.* **2013**, *12*, 293.

[8] C. J. Mitchell, D. Getnet, M.-S. Kim, S. S. Manda, P. Kumar, T.-C. Huang, S. M. Pinto, R. S. Nirujogi, M. Iwasaki, P. G. Shaw, X. Wu, J. Zhong, R. Chaerkady, A. Marimuthu, B. Muthusamy, N. A. Sahasrabuddhe, R. Raju, C. Bowman, L. Danilova, J. Cutler, D. S. Kelkar, C. G. Drake, T. S. K. Prasad, L. Marchionni, P. N. Murakami, A. F. Scott, L. Shi, J. Thierry-Mieg, D. Thierry-Mieg, R. Irizarry, L. Cope, Y. Ishihama, C. Wang, H. Gowda, A. Pandey, *BMC Syst. Biol.* **2015**, *9*, 75.

[9] G. Menschaert, D. Fenyö, *Mass Spectrom. Rev.* **2017**, *36*, 584.

[10] O. A. Balbin, J. R. Prensner, A. Sahu, A. Yocum, S. Shankar, R. Malik, D. Fermin, S. M. Dhanasekaran, B. Chandler, D. Thomas, D. G. Beer, X. Cao, A. I. Nesvizhskii, A. M. Chinnaiyan, *Nat. Commun.* **2013**, *4*, 2617.

[11] R. Chen, G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. K. Lam, R. Chen, E. Miriami, K. J. Karczewski, M. Hariharan, F. E. Dewey, Y. Cheng, M. J. Clark, H. Im, L. Habegger, S. Balasubramanian, M. O’Huallachain, J. T. Dudley, S. Hillenmeyer, R. Haraksingh, D. Sharon, G. Euskirchen, P. Lacroute, K. Bettinger, A. P. Boyle, M. Kasowski, F. Grubert, S. Seki, M. Garcia, M. Whirl-Carrillo, M. Gallardo, M. A. Blasco, P. L. Greenberg, P. Snyder, T. E. Klein, R. B. Altman, A. J. Butte, E. A. Ashley, M. Gerstein, K. C. Nadeau, H. Tang, M. Snyder, *Cell* **2012**, *148*, 1293.

[12] T. Farrah, E. W. Deutsch, M. R. Hoopmann, J. L. Hallows, Z. Sun, C.-Y. Huang, R. L. Moritz, *J. Proteome Res.* **2013**, *12*, 162.

[13] W. Wei, W. Luo, F. Wu, X. Peng, Y. Zhang, M. Zhang, Y. Zhao, N. Su, Y. Qi, L. Chen, Y. Zhang, B. Wen, F. He, P. Xu, *J. Proteome Res.* **2016**, *15*, 3988.

[14] A. D. Rolland, B. Jégou, C. Pineau, *Adv. Exp. Med. Biol.* **2008**, *636*, 16.

[15] L. Li, J. Zhang, Q. Deng, J. Li, Z. Li, Y. Xiao, S. Hu, T. Li, Q. Tan, X. Li, B. Luo, H. Mo, *PloS One* **2016**, *11*, e0164538.

[16] M.-S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. N. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T.-C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. K. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, A. Pandey, *Nature* **2014**, *509*, 575.

[17] B. S. Andersson, V. P. Collins, R. Kurzrock, D. W. Larkin, C. Childs, A. Ost, A. Cork, J. M. Trujillo, E. J. Freireich, M. J. Siciliano, *Leukemia* **1995**, *9*, 2100.

[18] P. Essletzbichler, T. Konopka, F. Santoro, D. Chen, B. V. Gapp, R. Kralovics, T. R. Brummelkamp, S. M. B. Nijman, T. Bürckstümmer, *Genome Res.* **2014**, *24*, 2059.

[19] J. E. Carette, C. P. Guimaraes, M. Varadarajan, A. S. Park, I. Wuethrich, A. Godarova, M. Kotecki, B. H. Cochran, E. Spooner, H. L. Ploegh, T. R. Brummelkamp, *Science* **2009**, *326*, 1231.

[20] A. Wutz, *Dev. Camb. Engl.* **2014**, *141*, 1423.

[21] J. R. Wiśniewski, A. Zougman, N. Nagaraj, M. Mann, *Nat. Methods* **2009**, *6*, 359.

[22] M. Maurer, A. C. Müller, C. Wagner, M. L. Huber, E. L. Rudashevskaya, S. N. Wagner, K. L. Bennett, *J. Proteome Res.* **2013**, *12*, 1040.

[23] C. Trapnell, L. Pachter, S. L. Salzberg, *Bioinforma. Oxf. Engl.* **2009**, *25*, 1105.

[24] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, *Genome Biol.* **2009**, *10*, R25.

[25] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, L. Pachter, *Nat. Biotechnol.* **2010**, *28*, 511.

[26] J. K. Eng, A. L. McCormack, J. R. Yates, *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976.

[27] S. Tyanova, T. Temu, J. Cox, *Nat. Protoc.* **2016**, *11*, 2301.

**Acknowledgements**

We thank Professor Paik and Dr. Jeong for their helpful discussion. This study was supported by the International Research & Development Program of National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (Grant number: NRF-2016K1A3A1A47921601) and the Brain Research Program through the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) (Grant number: NRF-2017M3C7A1027472). R.K.K. is funded through the Research Council of Norway’s Centres of Excellence Funding Scheme Project (223255/F50); FRIPRO Young Research Talent grant (263168) and Onsager fellowship from NTNU.

Figures and Tables

Figure 1. Workflow. Proteins were extracted from the haploid cell types and subsequently processed for peptides prior to fractionation by high pH RPLC into 50 fractions per replicate per cell type. And each fraction was further analyzed on LTQ Orbitrap Velos mass spectrometer. The same samples have been used to analyze their RNA by RNA-seq and the results associated were acquired from the study by Essletzbichler et al.

Figure 2. Transcriptome and proteome profiles. Replicate RNA-seq data and mass spectrometry data were compared, showing good overlaps between replicates (i.e. 14,473 in HAP1 (A) and 14,128 in KBM-7 (B) for RNA, 7,704 in HAP1 (C) and 8,336 in KBM-7 (D) by MaxQuant search and 8,638 in HAP1 (E) and 9,177 in KBM-7 (F) by SEQUEST search).

Figure 3. Comparative analysis of transcriptome and proteome data. Replicate transcriptomic and proteomic data were compared. Spearman correlation of RNA-seq replicate data for HAP1 (A) and KBM-7 (B) cell lines were 0.94 and 0.92 while gene expression between HAP1 and KBM-7 cells were found to be 0.85 (C). And Spearman correlation of protein abundance for HAP1 (D) and KBM-7 (E) cell lines were 0.97 and 0.98, while a moderate correlation of 0.90 was found in between the two cell type (F). When looking at correlation between RNA and protein, Spearman correlation values were 0.67 for HAP1 and 0.64 for KBM-7.

Figure 4. Effect of the deletion of a copy of chromosome 8. Dot plot was depicted using iBAQ values between HAP1 and KBM-7 for genes on chromosome 8 (red) and all other genes (black).

Figure 5. Discovery of novel proteins. A novel peptide GAPEPAQTQPQPQPQPAAPEGPEQPR was Blat-searched against the human genome identifying the location of the peptide in the genome (A). The genomic location of this peptide was found on a non-coding RNA in Chromosome 20. The corresponding peptide-spectrum match was manually validated and its mass spectral assignment was depicted (B).

Table 1. List of missing proteins. Missing proteins identified in this study were listed along with other information such as gene name, GO molecular function, chromosomal location, and peptide sequences.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No.** | **neXtProt ID** | **neXtProt PE level** | **Gene name** | **GO Molecular Function in neXtProt** | **Chromosome location** | **Peptide Sequence identified** |
| 1 | NX\_Q6ZU67 | 2 | BEND4 | Unknown | 4p13 | AAAAASSSSPSCTPATSQGHLR |
| MEEEMQPAEEGPSVPK  |
| MVLQHHNQLLIPQPADQPTEGSK |
| VLSLLNCGGK |
| 2 | NX\_Q6P4F1 | 2 | FUT10 | Alpha-(1->3)-fucosyltransferase activity | 8p12 | LAPLVYVQSDCDPPSDR |
| SAILVSEFSHPR |
| 3 | NX\_Q8N5D6 | 2 | GBGT1 | Globoside alpha-N-acetylgalactosaminyltransferase activityMetal ion binding | 9q34.2 | EKPLQPVVWSQYPQPK |
| VHYYIFTDNPAAVPGVPLGPHR |
| VLSPEYLWDDR |
| 4 | NX\_A2RU54 | 2 | HMX2 | Sequence-specific DNA binding | 10q26.13 | APACFCPDQHGPK |
| LLPAGSPSPGSERPR |
| 5 | NX\_P20264 | 2 | POU3F3 | HMG box domain binding | 2q12.1 | DDLHAGTALHHRPSAQEITNLADSLQLEK |
| Sequence-specific DNA binding |
| Transcription factor activity, sequence-specific DNA binding |
| 6 | NX\_Q9HCL3 | 2 | ZFP14 | DNA binding | 19q13.12 | LLSQLTQHQSIHTGEKPYECKLYSFLTQHQR |
| Metal ion binding |
| Transcription factor activity, sequence-specific DNA binding |
| 7 | NX\_Q9Y2G7 | 2 | ZFP30 | DNA binding | 19q13.12 | KLTSLPLYQK |
| Metal ion binding | LFSQLTQHQSIHFGEKPFK |
| Transcription factor activity, sequence-specific DNA binding | LHSSLIQHQR |
| 8 | NX\_Q6ZN06 | 2 | ZNF813 | DNA binding | 19q13.42 | HESHHTGDFR |
| Metal ion binding | ISALVIHTAIHTGEKPYK |
| Transcription factor activity, sequence-specific DNA binding | TFSQTYSLTCHR |
| 9 | NX\_Q03936 | 2 | ZNF92 | DNA binding | 7q11.21 | AFNQSSIFTKFSTLITHQIIYTGEKPCK |
| Zinc ion binding |
| Transcription factor activity, sequence-specific DNA binding |

**Supporting information**

Supporting Information is available from the Wiley Online Library of from the author.