Vol. 12, No. 3, 2021
ISSN 2233-4203/ e-ISSN 2093-8950
REVIEW www.msletters.org | Mass Spectrometry Letters

Chromosome-Centric Human Proteome Study of Chromosome 11 Team

Heeyoun Hwang[1], Jin Young Kim[1], and Jong Shin Yoo[1,2]*

[1]*Research Center for Convergence Analysis, Korea Basic Science Institute, Cheongju 29118, Korea*
[2]*Graduate School of Analytical Science and Technology, Chungnam National University, Daejeon 34134, Korea*

Received August 10, 2021, Accepted August 19, 2021
First published on the web September 30, 2021; DOI: 10.5478/MSL.2021.12.3.60

Abstract : As a part of the Chromosome-centric Human Proteome Project (C-HPP), we have developed a few algorithms for accurate identification of missing proteins, alternative splicing variants, single amino acid variants, and characterization of function unannotated proteins. We have found missing proteins, novel and known ASVs, and SAAVs using LC-MS/MS data from human brain and olfactory epithelial tissue, where we validated their existence using synthetic peptides. According to the neXtProt database, the number of missing proteins in chromosome 11 shows a decreasing pattern. The development of genomic and transcriptomic sequencing techniques make the number of protein variants in chromosome 11 tremendously increase. We developed a web solution named as SAAvpedia for identification and function annotation of SAAVs, and the SAAV information is automatically transformed into the neXtProt web page using REST API service. For the 73 uPE1 in chromosome 11, we have studied the function annotaion of CCDC90B (NX_Q9GZT6), SMAP (NX_O00193), and C11orf52 (NX_Q96A22).

Key words : C-HPP, proteomics, missing protein, protein variants, protein function annotation, LC-MS/MS

How many proteins (exactly protein-coding genes) are in human, and what function they have?"[1] In order to answer to this question, the Human Proteome Organization (HUPO) has organized the Human Proteome Project (HPP) since the end of the Human Genome Project. The HPP aims to map, annotate, and functionally characterize the entire human proteome supported by four resource pillars such as mass spectrometry (MS), affinity reagent (Ab), knowledge base (Kb), and pathology.[1] In addition, two complimentary initiatives of the Chromosome-centric Human Proteome Project (C-HPP) and Biology/Disease driven Human Proteome Project (B/D-HPP) focus on the completion of the entire proteoforms and aim to make proteomics an integral part of multi-omics research of the life sciences and biomedical research, respectively.[1,2] A total of 25 international teams of C-HPP has found the evidence of the protein existence and its function annotation to map it in chromosome 1 – 22, X, Y, and mitochondrial DNA (Figure 1).[3,4] Since 2012, the chromosome 11 (Chr. 11) group which is led by Dr. Jong Shin Yoo (Korea Basic Science Institute, Korea) have focused on the proteoforms in the chromosome 11 with study of finding 'missing proteins' and protein variants using LC-MS/MS and bioinformatics, and characterizing the function unannotated proteins.

The human protein coding genes are classified as proteins with the evidence at protein level (PE1), evidence at transcript level (PE2), homology (PE3), a predicted by gene model (PE4) and uncertain (PE5), where the proteins of PE2, 3, and 4 are defined as 'missing proteins (MP)'.[4] Since the launch of the C-HPP in 2012, remarkable progression has been reported, and only 10% of human coding genes is remained as the MP (Figure 2A), whereas there is little change in total protein entries of 20k. From the latest version of neXtProt database (2021-Feb-18; https://www.nextprot.org), 1,421 out of 20,379 proteins still remains as MP. In details, however, 19,684 protein entries has consistently existed for all neXtProt databases of nine years, about 5% of them has appeared or disappeared (Figure 2B). In particular, 87 new protein entries have been joined and 32 have been removed in 2021.02 version of neXtProt database against 2020.02 (Figure 2C). It is indicated that we should keep eyes on the latest version of database for human proteomic studies. According to the HPP data interpretation guideline (version 3.0), informatics analysis for human proteome should always be presented in comparison with the most recent version of reference database from neXtProt.[5] The Proteomics Standards Initiatives Extended FASTA Format

*Reprint requests to Jong Shin Yoo
https://orcid.org/0000-0002-8588-3310
E-mail: jongshin@kbsi.re.kr

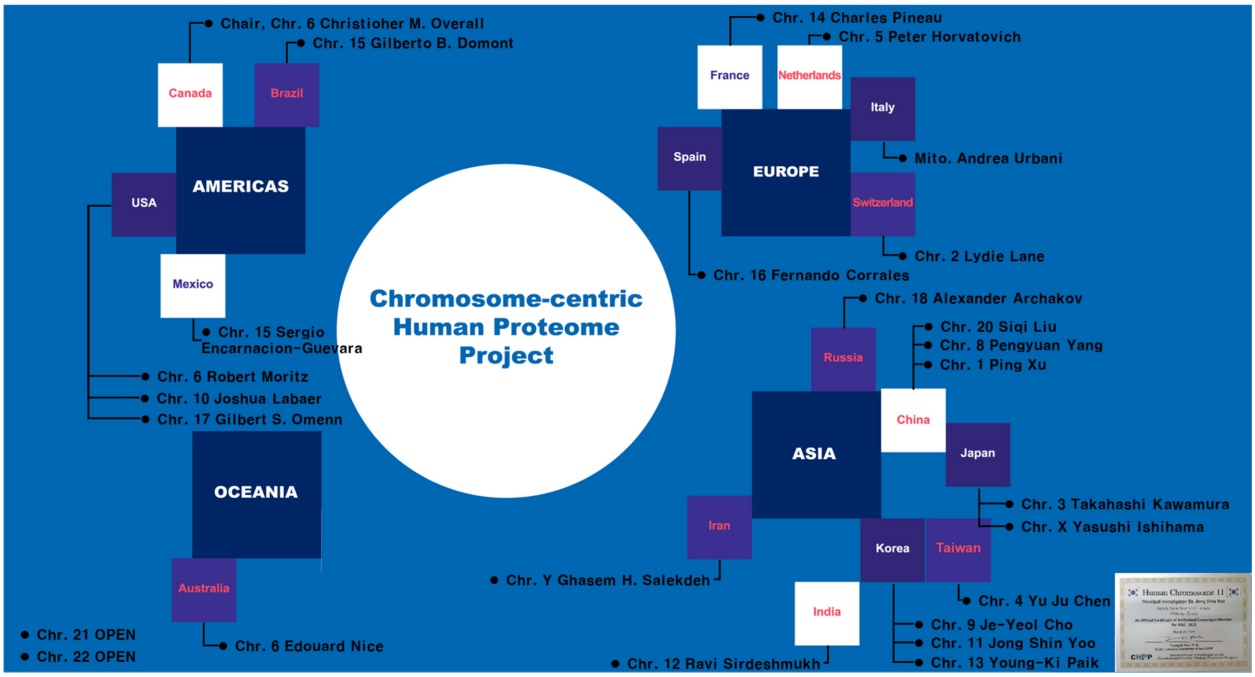

Figure 1. International Participants of Chromsome-centric Human Proteome Project. The leader name of each team is indicated. Original image is downloaded from https://www.hupo.org/C-HPP.

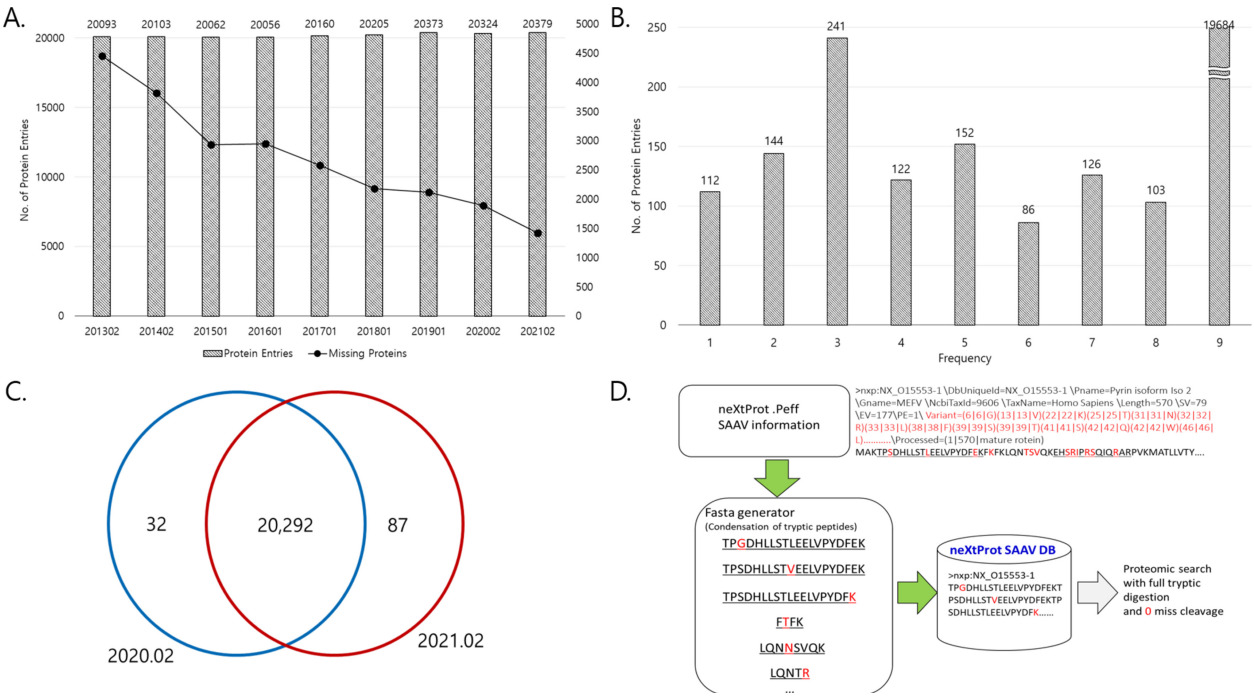

Figure 2. Status of neXtProt database from 2013.02 to 2021.02. (A) Status of protein entries and missing proteins (B) A histogram of protein entries in neXtProt databases from 2013.02 to 2021.02 (C) Comparison of protein entries between 2020.02 and 2021.02, and (D) An exmaple of PEFF file format and a method for generation **of** a SAAV Fasta file.

(PEFF) provided by neXtProt database, is very similar to FASTA format (Figure 2D). The title of protein entry includes more information on protein modification and known sequence variants, and altered protein isoforms generated from alternative splicing. The title of PEFF is more informative but, too long to use for proteomic search

engines. To solve this problem, we provide a simple program to convert from PEFF to FASTA file format in Github (https://github.com/heeyounh/pefftofasta). The program is coded by python 3.8 in Anaconda 3 environment.

Missing Proteins in Chromosome 11

According to the neXtProt database, the number of protein entries in Chr. 11 shows very little change from 2013.02 to 2021.02, whereas approximately the half of missing proteins in Chr. 11 has been discovered (Figure 3A). It has similar pattern with whole protein entries of human shown in Figure 2A. In terms of protein entries, the net change in Chr. 11 is zero between 2020.02 and 2021.02. In details, C11orf44 (NX_Q8N8P7, located at 11q24.3: 130672956 ~ 130717352) is disappeared, but BET1L (NX_Q9NYM9, located at 11p15.5: 167784 ~ 207428) has joined in Chr. 11. The advance of next generation sequencing technique might help to find BET1L located at low region of the Chr. 11. The number of missing proteins in Chr. 11 at the latest version of database is 222 out of 1,331 protein entries, where the percentage of missing proteins to protein entries in Chr. 11 is 16.67%. It is more than twice than those in whole proteins entries of human (approx. 7%) (Figure 3A). We found two missing proteins of

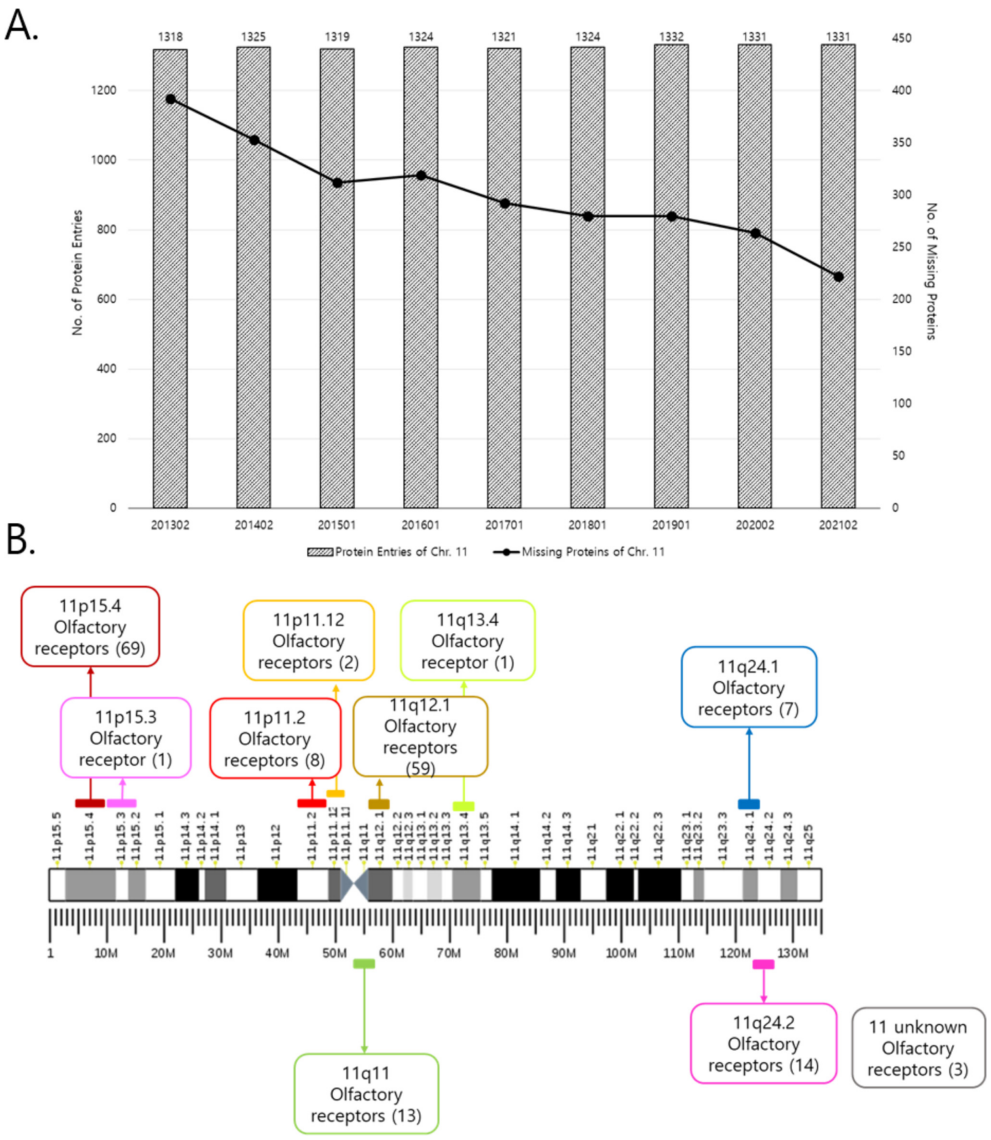

Figure 3. Status of Chromosome 11 in neXtProt database from 2013.02 to 2021.02. (A) Status of protein entries and missing proteins in chromosome 11, and (B) Location of olfactory receptors in chromosome 11. The number of protein coding genes of olfactory receptor is indicated in bracket.

GRIK4 (NX_Q16478) and EVI2A (NX_P22794) from human hippocampal tissues using a merge method of multiple proteome search results from SEQUEST, MASCOT, and MS-GF+.[8] In this method, E-values of true target PSMs were used in the least-squares line against decoy PSMs.[6,7]

In Chr. 11, there are 171 olfactory receptors, where a total of 409 olfactory receptors exists in human protein coding genes (Figure 3B). Baker et al. collected 23 peptides of olfactory receptors from public proteomic repositories, however, no olfactory receptor was identified under the HPP guidelines.[8] We tried to find the olfactory receptors from human olfactory epithelial tissue, whereas five other MPs (ACSM4, NX_P0C7M7; SAXO2, Q658L1; SLCO1A2, P46721; SVOPL, Q8N434; BPIFB3, P59826) out of 4,295 proteins were identified using LC-MS/MS.[6] From the latest version of neXtProt, 162 olfactory receptors of Chr. 11 are still remained as missing proteins. In addition, we have validated the missing proteins with scrutiny of MS/MS spectra comparison between experimentally identified peptides and their corresponding synthetic peptides. According to the HPP guideline 3.0, the use of synthetic peptides is powerful for determining the correct identification of spectra.[5]

Protein Variants in Chromosome 11

From liquid chromatography and tandem mass spectrometry data, we can identify novel peptides using novel protein sequence database such as alternative splicing variants (ASV; protein isoforms) and single amino acid variants (SAAV).[9] Proteogenomics could be performed against customized protein variant sequences generated from genomic and transcriptomic sequencing data. In cancer research field, tremendous proteogenomic studies were mainly reported by Clinical Proteomic Tumor Analysis Consortium (CPTAC) led by the National Cancer Institute of USA. Including genomic, transcriptomic, proteomic and copy number variation data, the results from protein variants have been used to classify unknown subclass of cancer patient groups or to present new druggable target molecules.

Nesvizhskii previously noted that the large scale of the proteogenomic studies needs more accurate process of novel peptide identification using estimated false discovery rate (FDR) against the customized sequence database.[9] In 2015, we also suggested a method to identify the proteoforms from long non-coding RNA, alternative splicing variant, and single amino acid variant using 3-frame translated transcript sequences from GENCODE database or protein variant information from neXtProt database.[10] With estimated FDR < 1% at spectra and protein level, we identified four novel ASV and 128 SAAVs from human hippocampus tissue, where we validated them using their corresponding synthetic peptides. In addition, we suggested a merge method of multiple proteome search results using an E-value calculated by the reversed decoy sequence, and also identified 477 known ASVs from the hippocampus data.[8] We suggested another method named as “next-generation proteomic pipeline (nextPP)” using a concatenated sequence database with neXtProt, 3-frame translated GENCODE, and SAAV database.[11] From the five different and public LC-MS/MS data of human brain tissue, we commonly found seven novel ASVs from 3-frame translated GENCODE database in two or more datasets with estimated FDR < 1% at spectra and protein level. The identified novel peptides of them were mapped onto the exon insertion, 5’-untranslated region or novel protein-coding sequences. In particular, a novel ASV of NCAM-013 coded in Chr. 11 were identified with novel peptides mapped at the exon insertion and validated with their corresponding synthetic peptides. Then we applied this method to human epithelial tissue, and we found 49 and 50 ASVs from neXtProt and 3-frame translated GENCODE databases, respectively.[7]

Technical development of genomic and transcriptomic sequencing is tremendously increasing the number of protein variants in the neXtProt databases. (Figure 4A) From the 2020.02, a new protein variant database of gnomAD (version 2.1.1) is added in the neXtProt variant information from dbSNP and COSMIC. According to the latest version of neXtProt, 9.5 million of protein variants and 42.5K ASVs are curated and mounted. Chr. 11 includes 588,540 protein variants and 2,635 known ASVs in 1,331 protein entries, where their pattern of difference shows very similar with the whole human data (Figure 4B).

In order to browse the SAAVs from genomic, transcriptomic and proteomic data, we developed a comprehensive platform named as “Encyclopedia of Single Amino Acid Variants: SAAVpedia”.[12] The SAAVpedia consists of four modules: SAAVidentifier, SAAVannotator, SNV/SAAVretriever and SAAVvisualizer. A total of 18,206,090 of SAAVs, their information of biological, clinical and pharmacological annotation is included. A webpage for user are mounted in a Linux server from Microsoft Azure clouding computing service (http://saavpedia.org). We also provide the REST API service of SAAVpedia via standard HTML (Figure 4C). Against the request of SAAV information, the API returns the output as XML and JSON formats. As a third party tool of neXtProt, SAAVpedia supplies the information of SAAV annotation back to neXtProt webpage via the REST API. All users of neXtProt can easily check out the SAAV information of a protein entry.

uPE1 in Chromosome 11

The net change of the number of function unannotated proteins with evidence at protein level (uPE1) from 2020.02 (1,254) to 2021.02 (1,273) is 19 plus in whole chromosome. In Chr. 11, nine uPE1s are charaterized, but other 10 new and one old members are joined into uPE1 in

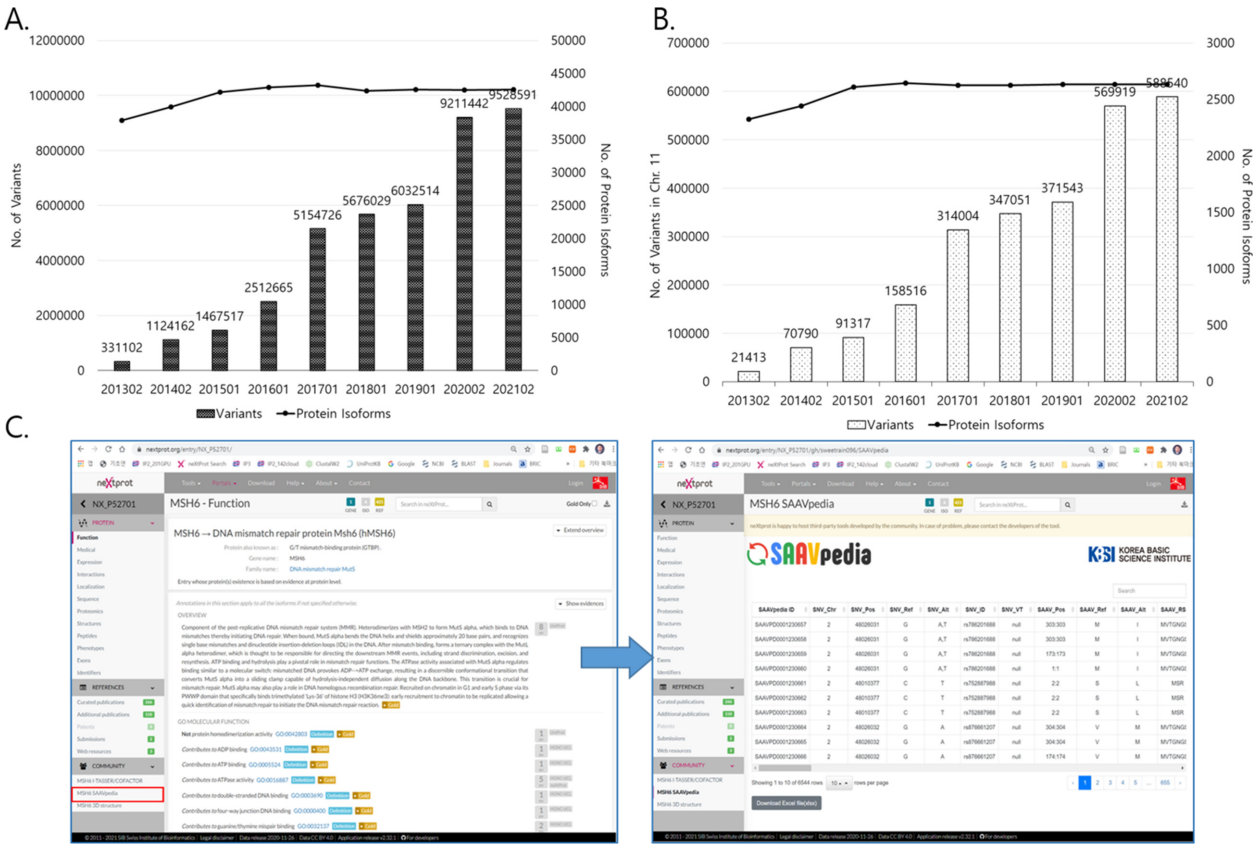

Figure 4. Status of protein variants and isoforms in neXtProt database from 2013.02 to 2021.02 (A) Status of protein variants and isoforms in whole chromosome (B) Status of protein variants and isoforms in chromosome 11, and (C) An example of using SAAVpeida in neXtProt web page. REST API service of SAAVpedia was provided via standard HTML.

the latest version of neXtProt database. Therefore, the net change is two plus from 71 to 73. Because the research for function annotation is often time and labour consuming, bioinformatic technique is required. ProteoRE (http://www.proteore.org) and UPEFinder (https://upefinder.proteored.org) are recently reported that they are able to charaterize uPE1s using the expanding knowledgebase and public RNA-Seq data.[13,14] I-TASSER and COFACTOR predicts the protein 3D structure and its gene ontology term, respectively.[15,16] The I-TASSER/COFACTOR is also available in a webpage of neXtProt database using independent server system. We performed the I-TASSER/COFACTOR with 66 uPE1s in Chr. 11 from the neXtProt webpage. We selected three protein-coding genes of CCDC90B (NX_Q9GZT6), SMAP (NX_O00193), and C11orf52 (NX_Q96A22) with predicted GO terms of high potential. We have validated their function through wet-lab experiments such as immunohistochemistry with HEK293T cells.[17]

Conclusions

As a part of C-HPP, we have developed a bioinformatic mathod for accurate identification, and discovery of missing proteins, ASVs, SAAVs as well as function annotation of uPE1 in chromosome 11. However, 222 missing proteins and 73 uPE1 are still remained in the darkside of proteome. We will further explore them with our colaborators.

Acknowledgments

This research was supported by the Korea Basic Science Institute (research grant C122000 and C170100).

References

1. 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.; Uhlén, M.; Wu, C. H.; Yamamoto, T.; Paik, Y.-K.; Omenn, G. S. *Mol. Cell. Proteom.* **2011**, 10, M111.009993, DOI: 10.1074/mcp.M111.009993.
2. Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J. E.; Kussman, M.; Qin, J.; Omenn, G. S. *J. Proteome Res.*

2013, 12, 23, DOI: 10.1021/pr301151m.
3. 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. *Nat. Biotechnol.* **2012,** 30, 221, DOI: 10.1038/nbt.2152.
4. Paik, Y.-K.; Lane, L.; Kawamura, T.; Chen, Y.-J.; Cho, J.-Y.; LaBaer, J.; Yoo, J. S.; Domont, G.; Corrales, F.; Omenn, G. S.; Archakov, A.; Encarnación-Guevara, S.; Lui, S.; Salekdeh, G. H.; Cho, J.-Y.; Kim, C.-Y.; Overall, C. M. *J. Proteome Res.* **2018**, 17, 4042, DOI: 10.1021/acs.jproteome.8b00383.
5. Deutsch, E. W.; Lane, L.; Overall, C. M.; Bandeira, N.; Baker, M. S.; Pineau, C.; Moritz, R. L.; Corrales, F.; Orchard, S.; Van Eyk, J. E.; Paik, Y.-K. Weintraub, S. T.; Vandenbrouck, Y.; Omenn, G. S. *J. Proteome Res.* **2019**, 18, 4108, DOI: 10.1021/acs.jproteome.9b00542.
6. Hwang, H.; Jeong, J. E.; Lee, H. K.; Yun, K. N.; An, H. J.; Lee, B.; Paik, Y.-K.; Jeong, T. S.; Yee, G. T.; Kim, J. Y.; Yoo, J. S. *J. Proteome Res.* **2018**, 17, 4320, DOI: 10.1021/acs.jproteome.8b00408.
7. Park, G. W.; Hwang, H.; Kim, K. H.; Lee, J. Y.; Lee, H. K.; Park, J. Y.; Ji, E. S.; Park, S.-K. R.; Yates, III, J. R.; Kwon, K.-H.; Park, Y. M.; Lee, H.-J.; Paik, Y.-K.; Kim, J. Y.; Yoo, J. S. *J. Proteome Res.* **2016**, 15, 4082, DOI: 10.1021/acs.jproteome.6b00376.
8. Baker, M. S.; Ahn, S. B.; Mohamedali, A.; Mohammad, T. I.; Cantor, D.; Verhaert, P. D.; Fanayan, S.; Sharma, S.; Nice, E. C.; Connor, M.; Ranganathan, S. *Nat. Commun.* **2017**, 8, 14271. DOI: 10.1038/ncomms14271.
9. Nesvizhskii, A. I. *Nat. Methods* **2014**, 11, 1114, DOI: 10.1038/nmeth.3144.
10. Hwang, H.; Park, G. W.; Kim, K. H.; Lee, J. Y.; Lee, H. K.; Ji, E. S.; Park, S.-K. R.; Xu, T.; Yates, III, J. R.; Kwon, K.-H.; Park, Y. M.; Lee, H.-J.; Paik, Y.-K.; Kim, J. Y.; Yoo, J. S. *J. Proteome Res.* **2015**, 14, 5028, DOI: 10.1021/acs.jproteome.5b00472.
11. Hwang, H.; Park, G. W.; Park, J. Y.; Lee, H. K.; Lee, J. Y.; Jeong, J. E.; Park, S.-K. R.; Yates, III, J. R.; Kwon, K.-H.; Park, Y. M.; Lee, H.-J.; Paik, Y.-K.; Kim, J. Y.; Yoo, J. S. *J. Proteome Res.* **2017**, 16, 4425, DOI: 10.1021/acs.jproteome.7b00223.
12. Lee, S. Y.; Hwang, H.; Kang, Y.-M.; Kim, H.; Kim, D. G.; Jeong, J. E.; Kim, J. Y.; Yoo, J. S. *J. Proteome Res.* **2019**, 18, 4133, DOI: 10.1021/acs.jproteome.9b00366.
13. Vandenbrouck, Y.; Pineau, C.; Lane, L. *J. Proteome Res*. **2020**, 19, 4782, DOI: 10.1021/acs.jproteome.0c00516.
14. González-Gomariz, J.; Serrano, G.; Tilve-Álvarez, C. M.; Corrales, F. J.; Guruceaga, E.; Segura, V. *J. Proteome Res*. **2020**, 19, 4795, DOI: 10.1021/acs.jproteome.0c00364.
15. Zhang, C.; Wei, X.; Omenn, G. S.; Zhang, Y. *J. Proteome Res*. **2018**, 17, 4186, DOI: 10.1021/acs.jproteome.8b00453.
16. Zhang, C.; Lane, L.; Omenn, G. S.; Zhang, Y. *J. Proteome Res*. **2019**, 18, 4154, DOI: 10.1021/acs.jproteome.9b00537.
17. Hwang, H.; Im, J. E.; Yang, Y.; Kim, H.; Kwon, K.-H.; Kim, Y.-H.; Kim, J. Y.; Yoo, J. S. *J. Proteome Res.* **2020**, 19, 4907, DOI: 10.1021/acs.jproteome.0c00482.