neXt-CP50 Challenge

Progress Report: neXt-CP50 Challenge

1. Background

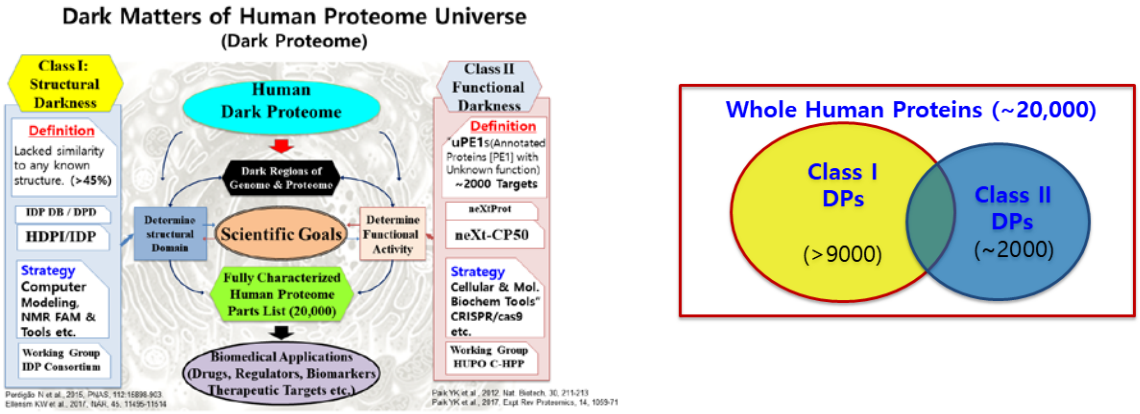


Figure 1. A proposed classification of dark proteins

(1) Classification of Dark Proteins (Tentatively Proposed) (Fig. 1)

To distinguish the uPE1 proteins from other structural dark proteins, C-HPP leadership tentatively proposed at St. Malo C-HPP Workshop in May 2019 that whole dark proteins may be grouped into two classes according to the nature of information lacking structure or function (Fig. 1). This proposal needs to be further discussed in Adelaide.

A. Class I Dark-proteins (>9000 proteins)
• Regions of proteins never observed by experimental structure determination and inaccessible to homology modelling.
• The majority of sequence (up to 100%) lacked similarity to any known structure.
• For 546,000 Swiss-Prot proteins, it was found that 44-54% of the proteome in eukaryotes and viruses still remained in dark side (Perdigao et al., 2015).
• 15% of the all Swiss-Prot is composed of dark proteins (Perdigao et al., 2017).

References

1. Unexpected features of the dark proteome. Perdigão et al., *PNAS* (2015).
2. The Dark Proteome Database. Perdigão et al., *BioData Mining* (2017).

B. Class II Dark Proteins (~2000 uPE1 Proteins) (HUPO C-HPP)
neXtProt DB (Advanced search: NXQ_00022, query)
• Class II Dark proteins, uPE1, can be defined as a group of proteins that do not have any annotated function information such as 'functionInfo', 'catalyticActivity', 'transportActivity', and 'pathway information' (neXtProt DB)

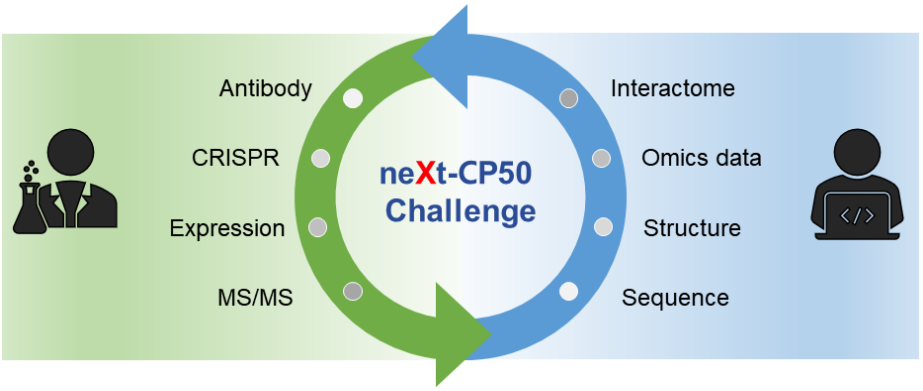
• neXt-CP50 challenges
A specific C-HPP initiative that aims to characterize some cellular function(s) of 50 uPE1 proteins within 3 years by C-HPP working groups (Paik et al., 2018, JPR)

• Teams: As of March 1, 2018, 15 international teams have joined this challenges.
They are: **Chr 2 (Swiss), 3 (Japan), 4 (Taiwan), 9 (Korea), 10 (USA), 11 (Korea), 13 (Korea), 14 (France), 15 (Brazil), 16 (Spain), 17 (USA), 18 (Russia), 19 (Mexico), 20 (China) and Y (Iran)**.

2. Progress Report (Period: 10/1/2018*-8/31/2019)

*(*An official publication date of JPR on announcement of neXt-CP50 challenge; Paik et al., 2018, JPR, **17**, 4042−4050)*

Tools and methods used for this challenge



Chr 2: Lydie Lane, Switzerland		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments	*THEM6 (chr 8)*	Grant from "Ligue Suisse contre le Cancer" Mary et al, manuscript in preparation
	C12orf73 (chr 12)	No funding Collaboration with Lena Ho's group in Singapore Zhang S, Liang C, Mary C, Kerouanton B, Francisco JC, Suhas Jagannathan N, Olexiouk V, Peh JH, Tang C, Fidelito G, Nama S, Cheng R-K, Wee C, Wang LC, Duek P, Sampath P, Lane L, Petretto E, Sobota R, Jesuthasan S, Sun L, Tucker-Kellogg L, Reversade B, Menschaert G, Stroud D, Ho L, BRAWNIN: A sORF-encoded Peptide Essential for Vertebrate Mitochondrial Complex III Assembly, submitted
	C15orf61 (chr 15)	No funding Mary et al, manuscript in preparation
	FAM205A (chr 9)	Seed money grant from St Gall university in 2018 Collaboration with a M. Vazquez' group in Buenos-Aires Work in progress

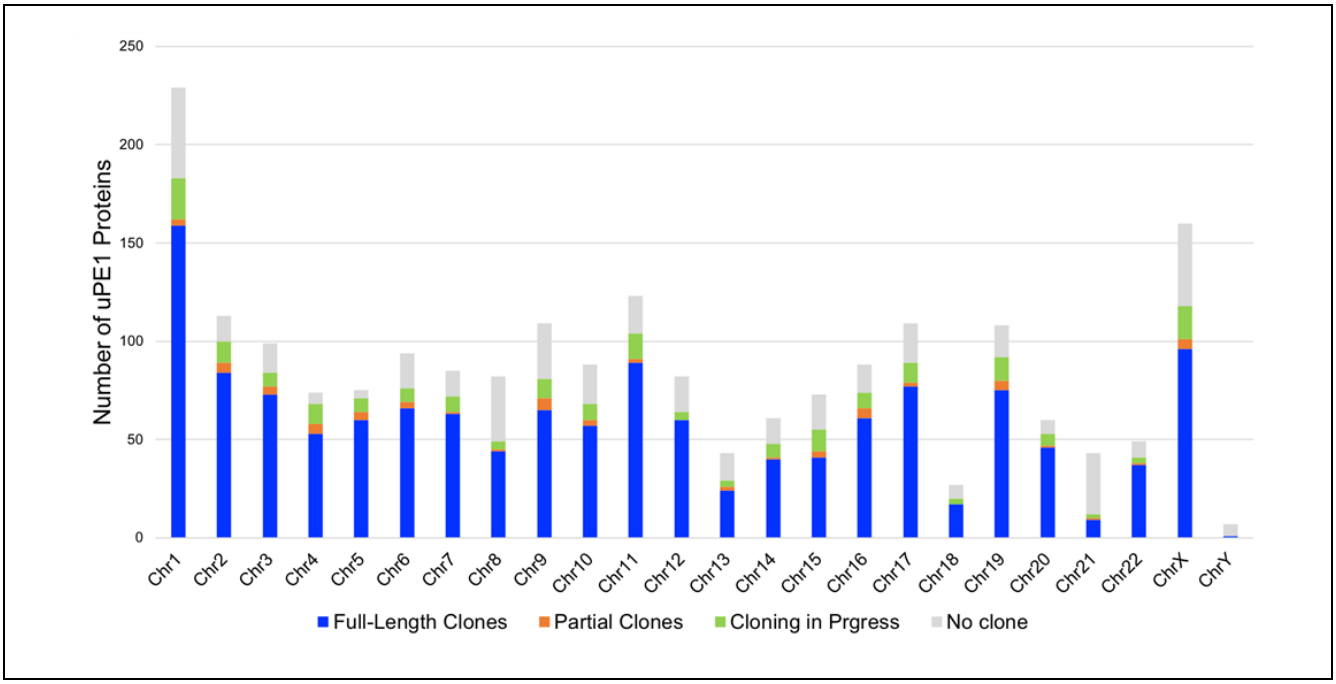
Chr 3: Takeshi Kawamura, Japan		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments	*ZCWPW2*	We have not detected this protein in our any data. We are plan**n**ing to analyze this protein through protein complex analyses after immunoprecipitation.

	TMA7	We have detected this protein in several cancers. During considering how to analyze it, we found the latest NeXtProt excluded this protein from the uPE1 list. The function was assigned to cytoplasmic translation.

Chr 4: Yu-Ju Chen, Taiwan		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments	*SEL1L3* *C1QTNF7*	**Work Plan** (1) Examine the expression levels of SEL1L3 and C1QTNF7 in the lung cancer tissues. (2) Establish of stable cell lines with high expression and knockdown of uPE1s. (3) Character cancer related phenotypes with stable cell lines including invasiveness, tumorigenesis and stemness in the in vitro cell models and in vivo mouse models. (4) Elucidate the uPE1s involved pathways by RNAseq and quantitative proteomics analyses in uPE1s manipulated cell lines. (5) Correlate the uPE1s expression and clinicopathological features. **Progress** (1) Among the 96 pairs of tumor and adjacent normal lung tissue samples from lung cancer patient, SEL1L3 showed significant up-regulation in 42% of patients while C1QTNF7 showed down-regulation in 87% of patients. (2) SE1L3 & C1QTNF7 mRNA expression was measured in 18 lung cancer cell lines. (3) The SE1L3 silencing cell lines are establishing by lentivirus with SE1L3 shRNA constructs. (4) The C1QTNF7 expressing vector has been constructed and the stably C1QTNF7 expressing lung cancer cell lines are under establishment. (5) The SE1L3 expressing vector has been constructed and a stable strains of lung cancer cell lines with knocked-down SE1L3 are under construction.

Chr 9: Je-Yoel Cho, Korea		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments	*NIPSNAP3A* *TSTD2* *FOXD4 (MPs)* *ARID3C (MPs)* *OR1J1 (MPs)* *ANKRD18A (MPs)* *ZNF510 (MPs)*	(1) Selected proteins for characterization include five MPs. (2) Human cell-based overexpression has been applied for characterization study. (3) Target proteins fused to halo-tag and/or his-tag. (4) IP-MS analysis for interactome discovery. (5) Confocal microscope observation to identify subcellular location. (6) Two proteins (FOXD4, ARID3C) have been discovered its subcellular localization and binding partner proteins via our IP-MS strategy and manuscript in preparation.

Chr 10: Joshua LaBaer, USA		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments	>90% of all uPE1 proteins (by the end of 2019)	We performed genome-wide CRISPR-based function genomics screen to identify mutations that can promote cancer progression, especially invasion, in breast epithelial cells expressing different mutant p53 proteins. From the screens, a few hundred hits were identified for 2 different p53 mutants, and we are currently down-selecting the top candidates, including several uPE1 proteins, for individual validation. We are aiming to submit the manuscript describing the screening results in combination with RNA-Seq and ChIP-Seq data within 2019. We are also producing more full-length plasmid clones for uPE1 proteins for functional studies, and the current clone coverage is shown below. Currently, we have full-length plasmids for around 67% of ~2,070 uPE1 proteins (shown below) and aim to reach >90% by the end of 2019, which is available to the entire C-HPP team.



Chr 11: Jong Shin Yoo, Korea		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments	*C11orf96* *SMAP* *FNBP4* *C11orf52* *CCDC90B*	Non-specific funding **Aim:** To define the implication of target proteins in cholangiocarcinoma or bile duct tissue **Work done or in progress:** Currently proteome characterization by labelled, shotgun proteomics analysis is ongoing in human tissue samples.

Chr 13: Young-Ki Paik, Korea		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments And Progress	*UBL3, KCTD4, TMTC4, PROSER1, CCDC122, SPRADY7, CCDC70 CCDC122, FAM124A ERICH6B etc.*	(1) Priority was given to those genes having no known isoforms for convenient gene editing study. (2) Cell-based screening of k.d. target gene may be the first step for functional study. (3) Multiple model organisms and their mutants may be useful for cross validating the functions of uPE1s (e.g., C. elegans, yeast, mice etc.) (4) Precautions for studying uPE1 (dark proteins) While working on a few dark proteins (5 uPE1s) encoded by Chr 13, it was found that there are several points that should be taken seriously before moving to the next level.

		We would like to share with you on some of precautions as listed below. Precaution 1: The importance of a highly scrutinized survey of all available public DBs and literatures when exploring Dark Proteins Functions. (neXtProt, 2017-8-1 vs. 2019-1-11 release). Two of our initial 10 targets have been published in highly cited journals such as Nature Comm and J. Clin. Invest during our initial investigation. Precaution 2: Inconsistency in the transcript Information between public DBs The notable inconsistency in the presence of transcript variants among the public DBs (e.g., NCBI, Ensemble, UniProt, PeptideAtlas, and neXtProt) brought about difficulties on the experimental designs and verification of data. This inconsistency might also have been associated with the time between an initial identification of the candidate proteins and their functional characterization. Precaution 3: Selection of Gene Deletion Sites for CRISPR/cas9 genome editing Due to the presence of isoforms (e.g., variants), caution should be given when we select the target sites for functional verification. Sometimes gene knock-down by deleting redundant gene segment may not give a right answer. Precaution 4: Antibody Issue In general, most of commercial antibodies corresponding to some uPE1 proteins appear to be either not high-quality ones or limited only to IHC or WB, which are not much useful for detection of targets in whole tissues or cells when the they are low abundance proteins. Thus, it would be better off to have enrichment process and tagged antibodies. We need some helps from HPA groups. In summary, we ended up narrowing down to a few of 10 targets, due to those unexpected problems as described above. Hopefully this survived one may produce some meaningful output.

Chr No. 14	**Charles Pineau, France**	**Work plan and progress**
uPE1s under Investigation	*MAGEC1, OOSP2, ERICH6, DCAF4L1, CTAG1A, CFAP45, EFCAB1, RSPH10B2, TMEM210, SEL1L2, SPATC1L, BEND2, FAM209B, AXDND1, LRRC23, TEX55, C16orf71, C7orf61*	(1) Priority of study is given to those genes with distinct testicular protein expression patterns. (2) Antibody-based screening of 512 protein candidates including 88 uPE1 and 60 MPs (trans-chromosomes). **->** Interesting clues provided on the possible function of uPE1 proteins and other MPs during spermatogenesis. For several uPE1 candidates, expression restricted to elongated/late spermatids in the testis, cilia of Fallopian tubes and airway epithelia suggests a role in the formation and/or motility of sperm. Just published in: *Pineau C, Hikmet Noraddin F, Zhang C, Oksvold P, Chen S, Fagerberg L, Uhlén M, Lindskog C. Cell type-specific expression of testis elevated genes based on transcriptomics and antibody-based proteomics. J. Proteome Res. 2019, In Press. PMID: 31429579*

Chr 15: Gilberto Domont, Brazil		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments	*Q9NYA3* *P0C870* *Q6ZRI6*	Collection of raw data files of our bioinformatics search.

Chr 16: Fernando Corrales, Spain		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments	APIP	Non-specific funding **Aim:** To define the implication of APIP in liver biology and disease **Work done or in progress:** Stable HepG2 and Huh7 deficient clones have been obtained. Currently characterizing the proliferative capacity and sensitivity to pro-apoptotic stimuli Proteome characterization by label free, shotgun proteomics analysis is ongoing in both cell lines.
	ARM5 *METTL26* *METTL9*	No funding allocated to this study yet **Aim:** To define the molecular function of these proteins and their role in liver biology and disease

		Work in progress: We are currently producing deficient cell clones to perform the functional studies.

Chr 17: Gilbert S. Omenn, USA		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Computation	All 66 (now 65) in neXtProt 2018-01 (now 2019-01) for Chr 17.	The Chromosome 17 team has created an analytical pipeline with I-TASSER and COFACTOR based on sequence, structure, and interactions to predict Gene Ontology terms for functional annotation of uPE1 proteins. We demonstrated the method on chromosome 17 with a benchmark set of 100 proteins and then predictions for all the 66 Chr 17 uPE1 proteins as of neXtProt release 2018-01 (Zhang C, Wei X, Omenn GS, Zhang Y, JPR 2018). During 2019 we have conducted blinded analyses of the predicted function for unannotated proteins using I-TASSER/COFACTOR on (1) 25 neXtProt uPE1 entries that were then annotated by SwissProt for the 2019-01 neXtProt release, and (2) on 267 proteins with Molecular Function-based predictions and 912 with Biological Process predictions from the 2017-2019 CAFA3 Challenge (Zhang C, Lane L, Omenn GS, Zhang Y, JPR 2019, under review). We have created a link for neXtProt users to access this automated I-TASSER/COFACTOR analysis of any uPE1 protein of interest to that user. The link has been incorporated into neXtProt for the upcoming release timed for the HUPO 2019 Congress. This service will be announced and demonstrated at the Bioinformatics Hub and C-HPP Poster Session in Adelaide. We encourage all C-HPP teams and also B/D-HPP teams to examine this pipeline and consider using it to assist their search for evidence of function(s) for uPE1 proteins. **<Reference>** Zhang, C.; Wei, X.; Omenn, G. S.; Zhang, Y., Structure and Protein Interaction-Based Gene Ontology Annotations Reveal Likely Functions of Uncharacterized Proteins on Human Chromosome 17. Journal of Proteome Research 2018, 17 (12), 4186-4196. Zhang, C.; Lane, L.; Omenn, G. S.; Zhang, Y., A Blinded Testing of Function Annotation for uPE1 Proteins by the I-TASSER/COFACTOR Pipeline Using the 2018-2019 Additions to neXtProt and CAFA3 Challenge. Journal of Proteome Research 2019, under review.

Chr 18: Alexander Archakov, Russia		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Computation	All 14 in neXtProt 2019-01 for Chr 18.	At the first stage we made retrospective analysis of Nextprot database (Gaudet et al., 2017) to reveal the most popular way of experiments for protein function validation (for different functions-different experimental approaches). After that we decided to focus our efforts on the functional annotation of chromosome 18 uPE1 proteins (uPE1 protein – proteins without known function). We decided to perform text-mining and meta-analysis. Search queries - the names of this protein in the PubMed does not give results. PRIDE contained 23 datasets with this protein. For the further analysis we have chosen 16 datasets created after 2016 (when HPP Data Interpretation Guidelines version 2.0 were published). These datasets were described in 12 articles respectively. Analysis of their MeSH-terms allowed us to form primary hypothesis about the Q68DL7 protein functional role. At the next stage we analyzed co-occurrence of this protein with other proteins in the same articles and experimental datasets. We used COFACTOR (Zhang et al., 2017) and I-TASSER (Yang et al., 2015) algorithms for protein function prediction based on protein structure. Basing on the principle “guily-by-association” the hypothesis about the role of this protein in different metabolic pathways.

Chr 19: Sergio Encarnación-Guevara, Mexico		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments	*CCDC97*	(1) In order the CCDC97 protein functional study and the relationship with viral proteins (HPV18 and HPV16), we cell-based screening of target gen on cervical cancer cells. New insights into cervical cancer biology may also have great implications for finding new treatment strategies. (2) We have obtained evidence that CCDC97 mRNA is expressed in cervical cancer lines (HeLa [HPV-18 positive], SiHa [HPV-16 positive] and C33A [HPV negative]) and HaCaT cells (transformed keratinocyte cells line, as control). (3) Also, we have obtained evidence that CCDC97 protein is differentially express in cervical cancer lines (HeLa, SiHa and C33A) and HaCaT cells.

		(4) Immunofluorescence microscopic analysis of CCDC97 in cervical cancer lines, showed a differential cellular distribution both nuclear and cytoplasmic. (5) Knockout (KO) was generated by the CRISPR/Cas9 system without any off-target effect detected. Western blot results showed successful validation of the CCDC97 knockout in the cervical cancer lines (HeLa, SiHa, C33A) and HaCaT cells (6) The screen also revealed a potential role for CCDC97, in many cellular functions (cytoskeleton arrangement, adhesion, migration or proliferation), since we observe a different morphology in KO cells. However, relevant assays are required to assign a protein function and to identify if CCDC97 loss conferred a selective disadvantage or vantage on cells. (7) Identifying the partners of a given protein (the interactome) may provide leads about the protein function and the molecular mechanisms in which it is involved. To identify proteins interacted with CCDC97, we have made an immunoprecipitation with specific antibodies and soon we will do a mass spectrometry assay to characterize protein interactomes obtains of co-immunoprecipitation from each cellular line.
Experiments	*CCDC61* *TMEM160* *C19orf47* *LENG8*	(1) In order to describe the proteins functional study in a model of cervical cancer and establish the relationship these molecules and viral proteins (HPV18 and HPV16), we will follow the same strategy describe to CCDC97. (2) Until the moment, we have obtained evidence that CCDC61, TMEM160, C19orf47 and LENG8 mRNA are expressed in cervical cancer lines (HeLa [HPV-18 positive], SiHa [HPV-16 positive] and C33A [HPV negative]) and HaCaT cells (transformed keratinocyte cells line). (3) CRISPR-Cas9 will use for the knockout of individual genes in genome-scale functional screens.

Chr 20: Siqi Lui, China		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments	*MANBAL* *FNDC11*	(1) Construction of stable cell lines with high expression of uPE1s. (2) Knocking down the uPE1s in cells using siRNAs.

		(3) Quantitative proteomic study of the constructed cells to discover the related pathways. (4) Function validation in cells and animal models. **Progress** (1) Got the stable strains of HeLa cell lines with higher expression of MANBAL and FNDC11. (2) The stable strains of HepG2 cell lines with higher expression of MANBAL and FNDC11 are under construction.

Chr Y: Ghasem Hosseini Salekdeh, Iran		
Methods	**uPE1s Under Investigation**	**Work plan and progress**
Experiments	**PRY protein**  	**PRY protein:** PRY is a multi-copy of chromosome Y, which is expressed in testis at protein level [PMID: 14665702; 11420382]. It was also shown to be expressed at RNA level in epididymis, testis, and blood T-cells [Human Protein Atlas]. However, there is no information about its cellular localization as well as interactions with other proteins. The peptide sequence of this protein is highly unique having only about 10% identity with HLA class I histocompatibility antigen according to NeXtprot blasting tool. It has two isoforms with molecular weights of 16.5 and 7.9 KDa. **PRY gene cloning:** The designed primers for PRY full gene cloning was as below:

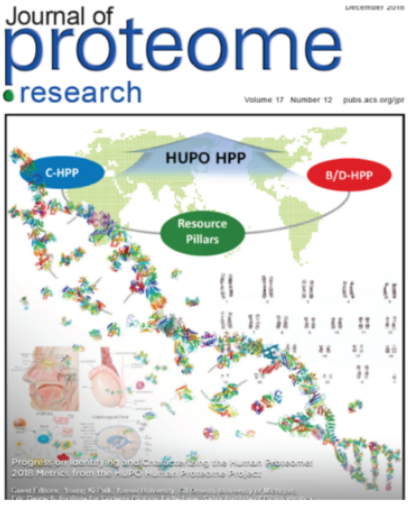
Gene name		**Sequence**	**RE**
PRY	F W	***ATTA****GGATCC***ATG** CTGGAGACAAGACAATTTG	Bam HI
466bp	R E	***ATTAAAGCTT***AGTAGGCTTAG TCTTCTTC	Hind III

For cloning, the PRY gene was PCR amplified from cDNA of a normal testis tissue using primers containing sequences specific for restriction sites of BamHI enzyme (in forward primer), and HindIII enzyme (in reverse primer). The PCR product was enzymatically digested using the restriction enzymes (Invitrogen), and subsequently cloned into the same restriction sites of multiple cloning sites from pET-28a vector. Target gene expression was under direct control of T7 promoter and terminator.

		Recombinant PRY protein expression: After sequence confirmation, validated vector was transformed to E. coli BL21 (DE3) strain. The recombinant E. coli cells were cultured in 10 ml of Luria Bertani medium containing 80 µg/ml kanamycin at 37°C and shaken at 180 rpm overnight. The next day, the Luria Bertani medium containing 50 µg/ml kanamycin was maintained at 37°C until cells reached the log phase (OD600 ~0.6). The expression of recombinant proteins was induced by 1 mM isopropyl-β-thiogalactopyranoside for 5 h. Recombinant protein was extracted by sonication and lysed by a urea buffer (pH 8.8; 8 M urea and 0.1 M sodium phosphate) and purified by 6×His-Ni-NTA chromatography. Obtained recombinant proteins were desalted and concentrated using Amicon columns (Millipore, USA). The purified recombinant protein was run on SDS- PAGE, stained with Coomassie brilliant blue, and then the corresponding protein band was cut and confirmed by a Bruker Ultraflex III MALDI TOF/TOF mass spectrometer. **Antibody production against PRY:** Desalted recombinant PRY proteins (400 µg) was emulsified with Freund's complete adjuvant and injected to young female New Zealand white rabbits (Albino). The rabbit was boosted after one month. Two other boosters were given at three-week intervals, and bleeding was done two weeks after last booster. Antisera were used for antibody titration and immunodetections. The produced antibody was confirmed by western blotting against recombinant PRY protein (figure 1).

International Working Groups for neXt-CP50 Initiative

Our 15 C-HPP Teams from 12 countries are now working on this project in collaboration with B/D-HPP and Resource Pillars in HUPO community.



2018 JPR Special Issue