The neXt-50 Challenge

Status report and acceptance of next-50 Challenge by individual Chromosome groups February 24, 2017.

Chromosome 1

(Ping Xu)

PIC Leaders: Ping Xu, Fuchu He

Contributing labs:

Ping Xu, Beijing Proteome Research Center

Fuchu He, Beijing Proteome Research Center

Dong Yang, Beijing Proteome Research Center

Wantao Ying, Beijing Proteome Research Center

Pengyuan Yang, Fudan University

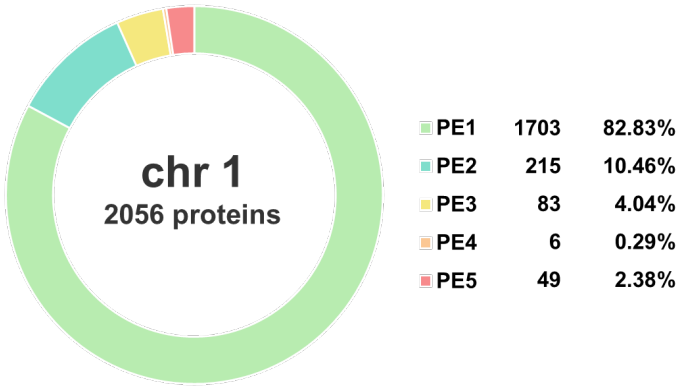
Siqi Liu, Beijing Genome Institute

Qinyu He, Jinan University

Major lab members or partners contributing to the neXt50:

Yao Zhang (Beijing Proteome Research Center), Yihao Wang (Beijing Proteome Research Center), Cuitong He (Beijing Proteome Research Center), Wei Wei (Beijing Proteome Research Center), Yanchang Li (Beijing Proteome Research Center), Feng Xu (Beijing Proteome Research Center), Xuehui Peng (Beijing Proteome Research Center).

Status of the Chromosome "parts list":



Step by step milestone plan to find, identify and validate MPs

1. We try to identify more missing proteins through high coverage proteomics technology with testis samples. We will finish this project before May.
2. We tested and confirmed that PTM approach could identify significant number of missing proteins. We will finalize the data sets before May.

Chromosome 2

See combined response and work plan with Chromosome 14.

Chromosome 3

(Takeshi Kawamura)

PIC Leaders: Takeshi Kawamura (since January 1)

Contributing labs: Major lab members or partners contributing to the neXt50:

Takeshi Kawamura: Associate Professor, Proteomics Laboratory, Isotope Science Center, The University of Tokyo, Tokyo, Japan.

Toshihide Nishimura: Professor, Department of Translational Medicine Informatics, St. Marianna University School of Medicine, Kanagawa, Japan.

Hiromasa Tojo: Associate Professor, Department of Biophysics and Biochemistry, Osaka University; Graduate School of Medicine, Osaka, Japan.

Tadashi Imanishi: National Institute of Advanced industrial Science and Technology (AIST), Tokyo, Japan.

Thomas P. Conrads: Adjunct Associate Professor, Chief Scientific Officer, Gynecologic Cancer Center of Excellence Women's Health Integrated Research Center at Inova Health System, USA.

Siu Kwan Sze: Assistant Professor, Director of Proteomics Core of Bioscience Research Centre, Singapore.

Status of the Chromosome "parts list":
In preparation

Confirmation that PIC and C-HPP lab members have read:

Yes

i) Deutsch et al 2016: Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1.

ii) Duek et al 2016: Missing Protein Landscape of Human Chromosomes 2 and 14: Progress and Current Status.

iii) Omenn et al 2016: Metrics for the Human Proteome Project 2016: Progress on Identifying and Characterizing the Human Proteome, Including Post-Translational Modifications.

iv) Vandenbrouck et al 2016: Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update.

Step by step milestone plan to find, identify and validate MPs:

Milestone dates:

PIC is changing from Prof. Nishimura to Kawamura and the project is under taking over. Therefore, we cannot set milestones at the present time. Members will gather at the February and plan to consider. I think that milestones can be decided before the JPR special issue in May.

Chromosome 4

(Yu-Ju Chen)

Apologized for delay, expected by Jan 27.

Reminder sent February 19.

Leader	Yu-Ju Chen
Contributing labs	1. 2. 3.

	4. Institute of Chemistry, Academia Sinica PI: Yu-Ju Chen Members: Reta Birhanu Kitata, Mehari Muuz, C Institute of Information Science, Academia Sinica PI: Ting-Yi Sung PI: Wen-Lian Hsu Members: Wai-Ko Master Program for Clinical Pharmacogenomics and Pharmacoproteomics, Taipei Medical University PI: Chia-Li Han Department of Biochemistry and Molecular Biology, Chang Gung University PI: Jau-Song Yu PI: Chia-Ju
Address	Institute of Chemistry, Academia Sinica 128 Academia Road Sec. 2, Nankang Taipei 115 Taiwan, ROC
Contact E-mail	yujuchen@gate.sinica.edu.tw
Telephone and Fax (with country code)	Phone: +886-2-27898660; Fax: +886-2-27831237

1. **Status of the Chromosome "parts list"**
2. **Milestone plan to find, identify and validate MPs**

2-1. Investigating the possibility of identifying 82 missing proteins via informatics approach

The unknown abundance of the 82 missing proteins in the biological sample present a great challenge for identifying these missing proteins. To aid experimental design of using proper material, we will develop an informatics ranking system for different cell/tissue types based on their gene expression data and mass spectrometry- based post-translational modification database. We will also perform prediction on subcellular localization and biochemical properties for each missing protein. Finally, the detectivity and properties of the enzymatic peptides of each missing proteins, such as peptide uniqueness, peptide length, and genome-based variants frequency (i.e., the amino acid length per sequence variant, LSV) will be investigated.
Expected milestone: complete analysis in 2017

2-2. Optimized Multiprotease Shotgun Strategy for Identifying 82 MPs

Chromosome 4 contains 761 protein entries of which 659 have protein level evidence while 23 at transcript level, 2 evidence from homologous species and 20 uncertain. At the updated version, there are 82 missing proteins (PE2-PE4) awaiting protein-level identification and validation (neXtProt 2017-01-23).

Based on the above informatics analysis, we will develop and apply an optimized shotgun strategy that integrates multiple enzymatic digestions to increase protein and peptide coverage, peptide fractionation to maximize identified protein number, LC- MS/MS analysis and protein identification by multiple search engines for comprehensive identification of missing proteins. Different sample types (organs/tissue/cells) will be analyzed for facilitating identification of tissue-specific missing proteins.
Expected milestone: complete analysis in 2018

2-3.「Cancer Moonshot as a Resource toward Comprehensive Human Proteome Library

To meet urgent need of finding new solution for cancer detection and therapy, Academia Sinica, Chang Gang University have signed Memoranda of Understanding (MOUs) for international collaboration under the *Cancer Moonshot* project in 2016. Specifically, the global efforts will jointly develop state-of-art proteogenomic technology for large-scale analysis of cancer patients and make available an unprecedented international dataset to advance the cancer research and care. Under the global efforts, we expect that the in-depth tissue proteomic and phosphoproteomic dataset for different cancer types will provide a rich resource for searching missing proteins.
Expected milestone: 100 pairs of cancerous and adjacent normal tissue from lung cancer will be expected to be finished in 2017.

2-4. MRM-MS for Validation of Missing Proteins

Multiple reaction monitoring (MRM) mass spectrometry using synthetic heavy isotope labeled peptide will be performed to validate missing proteins identified by the discovery method. The multiprotease approach is expected to provide additional unique peptides for those proteins with few or none detectable tryptic peptides. Informatics analysis of similarity score of spectral feature will be designed along with rigorous analytical assay evaluation to reduce the false positive result during validation.

Chromosome 5

(Peter Horatovich)

PIC Leaders: Peter Horvatovich and Rainer Bischoff

Contributing labs (Lab Heads named with affiliation University/Institute/Company):

- Rainer Bischoff and Peter Horvatovich, University of Groningen, Department of Pharmacy, Analytical Biochemistry
- Gyorgy Marko-Varga and Peter Horvatovich, University of Lund, The Center of Excellence in Biological and Medical Mass Spectrometry

Status of the Chromosome "parts list":

According to NextProt (release: 2016-12-02):

PE1	PE2	PE3	PE4	PE5	Total
738	99	18	5	10	870

Confirmation that PIC and C-HPP lab members have read:

Peter Horvatovich read all papers and lab members will do it until end of January 2017.

Step-by-step milestone plan to find, identify and validate MPs:

- Participation in IVTT cluster, which target the 50 missing proteins in chromosomes 5, 10, 15, 16, and 19 that has the highest probability for identification in COV318, PANC1, DFCI024, D341MED, ST486, KLE cell line determined using mRNA array expression.
- Close collaboration with chromosome 19 team in Moonshot project performing deep profiling of 4000 melanoma samples and sharing bioinformatics expertise. In this project missing proteins can be obtained from deep profiling of large number of primary and metastatic melanoma samples.
- European trasncan-2 project on ovarian cancer will allow to obtain primary and metastatic protein profiles on large number of ovarian cancer samples, where missing proteins is expected to be identified.
- Local project on proteogenomics profiling of human lung tissue in context of COPD. Human lung tissue is highly complex in term of cell types, which can be the source of multiple missing proteins.

Milestone dates (corresponding to the project listed above):

- We are currently analyzing COV318 cell line using MRM approach. We expect to analyze additional cell lines and human tissue (human lung, tissues from oral cavities) and report it until next HUPO in Ireland.
- This project is just started, but we expect to have the first results in the end of 2017.
- This project is just started, but we expect to have the first results in the end of 2017.
- This project already provided protein profile and is currently under bioinformatics evaluation. Identification of missing proteins is expected to be performed within three months.

Chromosome 6

(Christoph Borchers)

Christoph Borchers is new PIC as of January 20.

Reminder sent Feb 19. Report expected after Grant interviews in February.

Chromosome 7

(Ed Nice)

The Chr 7 NeXT 50 challenge will continue along the lines presented for the Missing Proteins at the recent C-HPP workshops. The Top 50 proteins to be pursued will be identified from the current NeXt Prot release and will be driven around perceived feasibility. A Biology/Disease driven approach will be used to populate the project. This could result in missing proteins from other chromosomes also being identified.

In brief the following steps will be used:

1. Check the Chr 7 missing proteins against MissingProteinPedia

2. Identify proteins/protein families where credible non-MS based evidence exists for their existence (we are already aware this is the case for more than 50 proteins including a number of olfactory receptors)

3. Identify "champions" for these proteins/protein families who can be recruited into HUPO if not already members. This will include biochemists, physicians, pathologists.

4. Develop SOPs and build up a Biobank meeting best practice for available tissues or cell lines.

5. Use state of the art MS technologies for deep mining of the associated proteomes.

It is anticipated that steps 1 -3 will be complete for presentation at the Dublin meeting.

Chromosome 8

(Pengyuan Yang)

No response.

Reminder sent Feb 19.

Chromosome 9

(Je-Yoel Cho)

PIC leaders: Je-Yoel Cho[1]*, Soo-Yeon Lee[2], Jin-Hwan Lee[3]

Contributing Members: Yong-in Kim[1], Hyung-Min Park[1], Rooa Yoon[1], Ji-Sook Park[2],

1. Department of Biochemistry, BK21 PLUS Program for Creative Veterinary Science Research and Research Institute for Veterinary Science, College of Veterinary Medicine, Seoul National University, South Korea

2. Samsung Medical Center, Seoul, South Korea

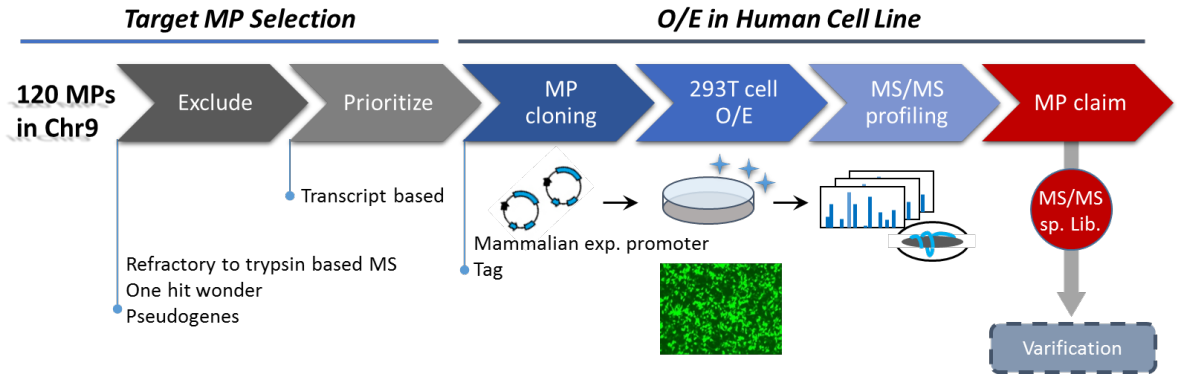
3. Korea Research Institute of Standards and Science, KRISS, Daejeon, South Korea

Milestone Dates (Expected):

Step 1: March 2017

Step 2: March 2018

Step 3: November 2018



STEP 1. TARGET MP SELECTION

Targeting to missing proteins (MPs), which are theoretically easy to detect.

1. Refractory to trypsin based MS/MS → exclude

2. One hit to wonder proteins → exclude

3. Transcript expression database (e.g. HPA, GTEx, BGee…) → prioritize

*Chr9 status: 120 proteins are remained missing (neXtProt rel. 2016-12-02)

4. Select top 60 lists

STEP 2. OVER-EXPRESSION of TOP 50 MP IN HUMAN CELL LINE

1. CDS of top 60 missing proteins insert to plasmid vector that contain mammalian cell expressed promoter and affinity/fluorescence tag.

2. Transfection to 293T cell.

3. If available, utilize tag for transfection check, subcellular localization observation and affinity-based MP enrich.

4. MS/MS profiling

5. For top 50, MP claim and MS/MS spectra library construction for further verification step.

STEP 3. VERIFICATION: NATURALLY EXPRESSED MPs IN HUMAN CELL/TISSUE

1. Select top 50 target proteins from Step 2 that are show clear MS spectrum and expressed in target tissues/cells samples such as testis using transcript expression database (e.g. HPA, GTEx, BGee etc).

2. Targeted analysis by MRM or PSM (MS inclusion list) with synthetic peptides spikes.

Chromosome 10

(Josh LaBear)

Updating by Feb 7 after grants submitted.

Reminder sent 19 Feb. Report promised soon.

Several months ago my lab sent out the lysates from several selected cell lines – chosen based on mRNA expression of the missing proteins – to the members of our mini-consortium. They have probed these lysates looking for evidence of the missing proteins using the optimized spectra that they determined were ideal for these individual proteins – learned by producing them in cell free lysates and testing them on MS.

Chromosome 11

(Jong Shin)

PIC Leaders: Jong Shin Yoo, Ph.D.

Contributing labs: Jong Shin Yoo/Biomedical Omics Research Group/Korea Basic Science Institute

Major lab members or partners contributing to the neXt50: Jin Young Kim/KBSI, Bong Hee Lee/Gacheon Univ.

Status of the Chromosome "parts list":

PE2- 213,

PE3- 98,

PE4- 6,

Total 317 are missing proteins in chromosome 11.

(from https://www.nextprot.org/about/protein-existence)

Confirmation that PIC and C-HPP lab members have read: *OK, we have checked.*

Step by step milestone plan to find, identify and validate MPs:

Step. 1 20 missing proteins in all Chr. mapped

Step. 2 20 more missing proteins in all Chr. mapped

Step. 3 10 more missing proteins in all Chr. mapped

Total missing proteins: 50 in all Chr. Mapped

Milestone dates:

Step. 1 Apr. 2017

Step. 2 Dec. 2017

Step. 3 Jun. 2018

Chromosome 12

(Ravi Sirdeshmukh)

No response.

Reminder sent February 19.

Chromosome 13

(Young Ki Paik)

PIC Leaders: Young-Ki Paik, Ph.D.

Young-Ki Paik/Yonsei Proteome Research Center/Yonsei University

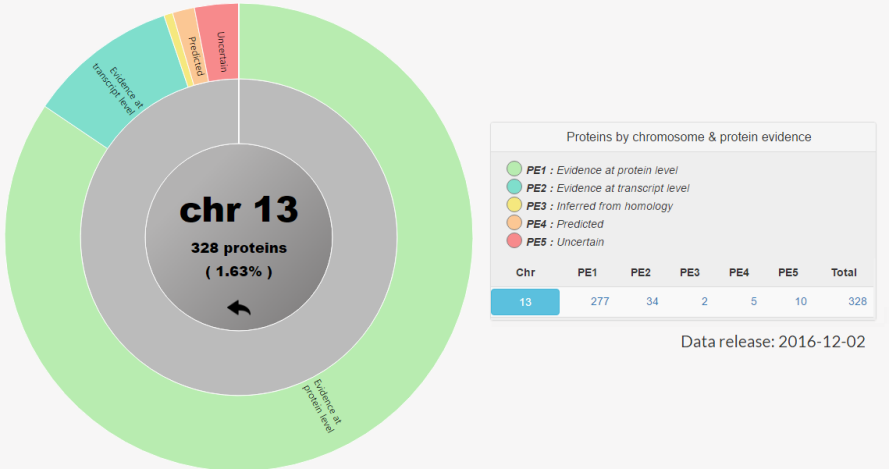
Major lab members or partners contributing to the neXt50:

Seul-Ki Jeong/YPRC/Yonsei University.

Jin-Young Cho/YPRC/Yonsei University.

Chae-Yeon Kim/YPRC/ Integrated Omics for Biomedical Sciences, Yonsei University.

Status of the Chromosome "parts list":



(from https://www.nextprot.org/about/protein-existence)

Confirmation that PIC and C-HPP lab members have read:

OK, we have checked.

Step by step milestone plan to find, identify and validate MPs:

Top-Down Approaches

Step. 1 20~30 missing proteins from liver (normal-like, tumor, and/or cell lines) in all Chr. mapped

Step. 2 20~30 missing proteins from pancreas (normal-like, tumor, and/or cell lines) in all Chr. mapped

Step. 3 10~20 missing proteins from placenta (normal, pre-eclampcia) in all Chr. Mapped

Total missing proteins: 50 in all Chr. mapped

Milestone dates:

Step. 1 Jul. 2017

Step. 2 Feb. 2018

Step. 3 Aug. 2018

Step by step milestone plan to find, identify and validate MPs:

Bottom-Up Approaches

Step 1 Select target missing proteins under criteria (available/target sample, transcript expression level, number of proteotypic peptides, etc.)

Step 2 Makes synthetic peptides and MS spectra of target proteins.

Step 3 Discovering and validating missing proteins.

Milestone dates:

Step. 1 Feb. 2017

Step. 2 Feb. 2018

Step. 3 Aug. 2018

Chromosome 14

(Charles Pineau)

PIC Leaders:

PI: Charles Pineau; Co-PI: Yves Vandenbrouck

Contributing labs:

- Charles Pineau

PROTIM - Inserm U1085 - Irset, Campus de Beaulieu, 35042 Rennes cedex France

- Jérôme Garin

Proteomics French Infrastructure (ProFi). http://www.profiproteomics.fr - MS/MS platform members:

- Virginie Brun

EDyP - CEA, DRF, BIG, Laboratoire de Biologie a Grande Echelle, Inserm U1038, 17 rue des martyrs, Grenoble F-38054, France

- Sarah Cianferani

Laboratoire de Spectrométrie de Masse BioOrganique (LSMBO), IPHC, Université de Strasbourg, CNRS UMR7178, 25 Rue Becquerel, 67087 Strasbourg, France

- Odile Burlet-Schiltz

Institut de Pharmacologie et de Biologie Structurale, Université de Toulouse, CNRS, UPS, 31062 Toulouse, France

- Thomas Fréour

Service de Médecine de la Reproduction, Inserm U1064, CHU de Nantes, 38 boulevard Jean Monnet, 44093 Nantes cedex, France

Major lab members or partners contributing to the neXt50:

- **Major lab members *(alphabetical order)***

  Christine Carapito *(Targeted mass spectrometry)*

  Anne Gonzalez de Peredo *(Discovery mass spectrometry)*

  Thomas Fréour *(Clinical male infertility)*

  Charles Pineau *(Male Reproduction, clinical proteomics)*

  Yves Vandenbrouck *(Bioinformatics)*

- **Partners:**
  - Lydie Lane (CALIPHO, SIB, Geneva, Switzerland)
  - Cecilia Lindskog (HPA, Uppsala, Sweden)

Status of the Chromosome "parts list":

Current statistics in neXtProt indicate that 93 entries are still not experimentally validated and considered as missing proteins (PE2-4) on chromosome 14. There are also 17 PE5 remaining to be eventually identified as a product of a protein-coding gene and validated.

Chromosome 14 report in neXtProt release 2016-12-02:

PE1	*PE2*	*PE3*	*PE4*	*PE5*	*Total*
531	*65*	*23*	*5*	*17*	*641*

Two years ago, the Chromosome 14 consortium has selected the testis as a major source for searching Missing Proteins. We consider our work as a trans-chromosome initiative based on a biological function that is spermatogenesis and sperm maturation. Mass spectrometry analyses are performed by the French Proteomics Infrastructure (ProFI) for which the C-HPP chromosome 14 (France) and 2 (Switzerland) initiatives represent a flagship project. Of note is that no dedicated funding could be obtained in France to tackle C-HPP objectives.

According to our calculation and based on our in house RNAseq transcriptomes datasets, only 18 missing proteins correspond to genes on chromosome 14 that are either preferentially or exclusively expressed in the testis. Interestingly, most of Missing Proteins that are testis-specific are expected to be present in meiotic, postmeiotic germ cells and spermatozoa. These correspond to genes that are statistically more represented on the longer chromosomes (*e.g.,* chromosomes 1, 2). As a consequence, we consider our work as a trans-chromosome initiative and will indeed contribute significantly to other chromosomes with newly found MPs.

Based on mRNA distribution (ArrayExpress, RNAseq Atlas and in house RNAseq datasets), the estimated number of remaining Missing Proteins to be found in reproductive organs/cells is close to 800. As much as 650 could be found in meiotic and postmeiotic germ cells. Approximately 100 could be found in the epididymis, most of which correspond to small peptides (*e.g.*, defensins...) thus calling for specific mass spectrometry validation guidelines. Smaller numbers of missing proteins could be found in the ovary, placenta and embryos. Thus, members of the C-HPP Reproductive cluster will contribute, on the basis of their savoir-faire, to the following steps of our global project by providing news rare and valuable biological materials in which to search for additional MPs (*e.g.,* purified ovarian cells).

PIC and major lab members are aware and strictly follow the HPP metrics and most up to date HPP guidelines. The Vandenbrouck et al., 2016 was produced by our consortium in collaboration with L. Lane (Chr 2) and C. Lindskog (HPA), and all members are co-authors. In addition, Y. Vandenbrouck has participated to the Duek et al., 2016 paper and has been involved in the definition of the latest version of HPP guidelines (Deutsch *et al.*, 2016).

Step by step milestone plan to find, identify and validate MPs:

In a first ongoing step, we have decided to chose the spermatozoa as a source for Missing Proteins. Ongoing projects are:

1) The search and identification of **the top 40 MPs** proposed in Duek *et al.*, 2016.
2) The identification, characterization and validation of novel protein-coding genes in the human testis using a proteogenomics approach.
3) The search, identification and validation of **rare MPs** in the human spermatozoa. That specific project involves prefractionation protocols of the human sperm cell into enriched head, intermediate pieces and flagellar fractions, then preparation of membranes and solubilization of lipophilic transmembrane proteins.

In parallel , an ongoing project aims at demonstrating that some MPs identified in the human spermatozoa proteome (Vandenbrouck *et al.,* 2016) originate from other organs. The corresponding genes are not expressed in the germ cell lineage but in the epididymis head or seminal vesicles. These proteins will integrate the spermatozoa during its transit through the epididymis. This is an interesting new concept that will help understanding the post-testicular maturation of the male gamete.

Milestone dates:

Q2-2017

Identification of the top 40 MPs proposed in Duek *et al.*, 2016.
To be submitted to the 2017 Special Issue C-HPP of J. Prot Res.

Q4-2017

Identification, characterization and validation of novel protein-coding genes in the human testis using a proteogenomics approach. This ongoing work contributes to the validation of additional proteins that are not considered to date as part of the human proteome. *Work will be submitted to a specialized journal in the field of Reproductive Biology*

Q2_2018

Identification, characterization and validation of rare MPs in the human spermatozoa. *To be submitted to the 2018 Special Issue C-HPP of J. Prot Res.*

Identification, characterization and validation of rare MPs in male and female reproductive tracts.

To be submitted to the 2018 Special Issue C-HPP of J. Prot Res.

Chromosome 15

(Gilberto Domont)

PIC Leaders:

Gilberto B Domont, Fabio CS Nogueira, Paulo C Carvalho, Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil

Contributing labs:

Gilberto B Domont, Lab Head, Proteomics Unit, Federal University of Rio de Janeiro, RJ, Brazil, gilbertodomont@gmail.com

Fabio Cesar Sousa Nogueira, Proteomics Unit, Lab Head, Proteomics Unit, Federal University of Rio de Janeiro, RJ, Brazil, fabiocsn@gmail.com

Paulo C Carvalho, Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil

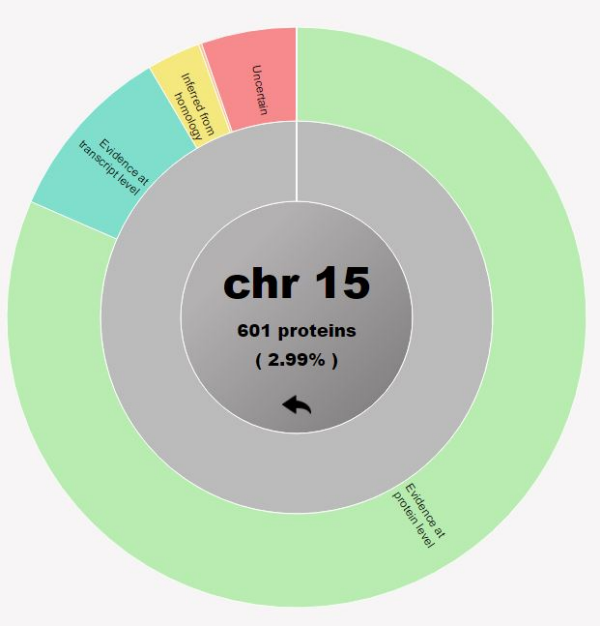
Rafael Melani, PhD, Proteomics Unit, Federal University of Rio de Janeiro, RJ, Brazil

Erika Velasquez, PhD student, Proteomics Unit, Federal University of Rio de Janeiro, RJ, Brazil

Clarissa Ferolla, PhD student, Proteomics Unit, Federal University of Rio de Janeiro, RJ, Brazil

Jimmy Murillo, PhD student, Proteomics Unit, Federal University of Rio de Janeiro, RJ, Brazil

Status of the Chromosome "parts list":



Proteins by chromosome & protein evidence

PE1 : Evidence at protein level
PE2 : Evidence at transcript level
PE3 : Inferred from homology
PE4 : Predicted
PE5 : Uncertain

Chr	PE1	PE2	PE3	PE4	PE5	Total
14	531	65	23	5	17	641
15	490	60	18	1	32	601
16	713	76	21	3	23	836
17	997	122	18	6	23	1166
18	244	18	2	2	9	275
19	1138	222	29	7	32	1428
20	456	62	15	3	13	549
21	179	29	20	1	24	253
All	16618	2247	573	96	574	20108
	82.65%	11.17%	2.85%	0.48%	2.85%	100%

Confirmation that PIC and C-HPP lab members have read:

Confirmed

Step by step milestone plan to find, identify and validate MPs:

STEP 1.

Part of the IVTT cluster, which targets 50 missing proteins in C5 10, 15, 16, and 19 that have high probability of identification in cell lines COV318, PANC1, DFCI024, D341MED, ST486, KLE. COV318 and KLE cell lines are under analysis. MRM approach.

STEP 2:

PRM of 65 missing proteins. First classical PRM using 65 synthetic peptides that map for missing proteins. Almost all initial adjusting experiments were done. Literature search showed that thyroid, brain and testis express the proteins to be assayed. Also we have already collected the thyroid normal and cancerous lobes as the samples to be tested first. Next we will use sample from fetuses brain. This project just started and we expect to present results during Dublin HUPO.

Milestone dates:

STEP 1 - Apr. 2017

STEP 2 - Sept 2017

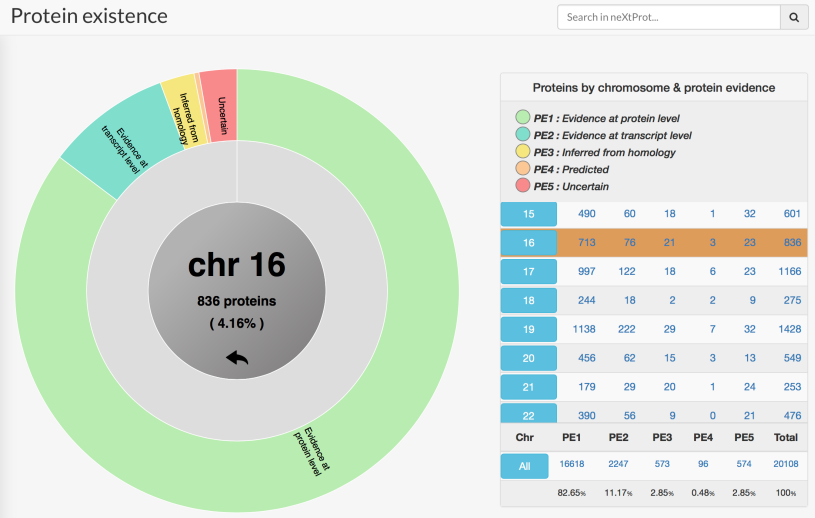
Chromosome 16

(Fernando Corrales)
PIC Leaders: Fernando J. Corrales, Concha Gil

Contributing labs:

Major lab members or partners contributing to the neXt50: ProteoRed (Coordinator Fernando J. Corrales; Centro Nacional de Biotecnología (CNB-CSIC), Madrid, Spain).

neXtProt v2.3.0



Chr	PE1	PE2	PE3	PE4	PE5	Total
15	490	60	18	1	32	601
16	713	76	21	3	23	836
17	997	122	18	6	23	1166
18	244	18	2	2	9	275
19	1138	222	29	7	32	1428
20	456	62	15	3	13	549
21	179	29	20	1	24	253
22	390	56	9	0	21	476
All	16618	2247	573	96	574	20108
	82.65%	11.17%	2.85%	0.48%	2.85%	100%

**The numbers in this schema are not the same in the downloadable table.

Confirmation Reading key papers:
Done.

Step by step milestone plan to find, identify and validate MPs:

- Prioritization of the 50 chr16 MPs to be targeted. Annotation with Missing ProteinPedia, HPA, Human MRM Atlas and Chr16 already generated information. February 2017.
- List of chr16 MPs that would require non standard procedures (non-trypsin digestion sites, lack of unique peptides...). March 2017.
- Preparation of SRM methods and identification of the biological sources with highest probability of finding the selected chr16 MPs. June 2017
- SRM analysis, 10 MPS/lab. Starting on June-July 2017).
- IVTT initiative (5 chromosomes participating). 5-10 proteins/year. (2 MPs already identified by SRM and 2 aditional are pending on validation)

IVTT initiative (5 chromosomes participating). 5-10 proteins/year.

Targeted search of Ch.16 MPs identified in our extensive shotgun datasets (compliant with HUPO guidelines) in sperm and HEK293 cells. 10-15 proteins/year.

Identification of those MPs that would require non standard procedures (non-trypsin digestion sites, lack of unique peptides...). March 2017.

Chromosome 17

(Gil Omenn)

PIC Leaders: New Gil Omenn, Mike Snyder, and Ron Beavis

Contributing labs: Guan Lab, University of Michigan; Eric Deutsch, ISB.

Major lab members or partners contributing to the neXt50: Bharat Panwar, Hongjiu Zhang, Yuanfang Guan

Status of the Chromosome "parts list":

We have expression profiles from GTEx and from Cancer Genome Cloud.

We have initiated analysis of the Chr 17 full set of 1167 predicted proteins following the Duek et al Chr 2/Chr 14 pipeline, as reported to C-HPP leadership and Chr 17 colleagues on 23 Dec 2016.

Step by step milestone plan to find, identify and validate MPs:

(a) 23 December 2016 email to C-HPP (and B/D-HPP) leaders and Chr 17 team:

"Of the 1167 protein-coding genes on Chr 17, there are 148 "missing proteins", which are neXtProt 2016-02 PE2,3,4 (excluding PE5). The C-HPP goal is to find credible evidence for at least 50 from each chromosome. I propose that we carefully annotate the missing proteins to identify those most likely to be detectable by mass spectrometry and to identify those not (sufficiently) expressed as transcripts or not suitable for solubilization and tryptic digestion. We need guidance on the use of HPA immunohistochemistry findings (illustrated by SMIM6 and SMIM5, below); at present, I believe neXtProt uses HPA findings for corroboration but not for primary PE1 identification.

We also need to know which of these 148 moves into the neXtProt to SwissProt curation pipeline with pending approval as PE1.

I will make inquiries about securing SRM analyses for our top candidates if appropriate tissues can be obtained; sharing of key specimens across C-HPP teams would be highly desirable.

Ron and Mike: we welcome your guidance and participation, especially with algorithms for characterizing these protein sets. My colleague, Bharat Panwar, has linked his MI-Proteome Visualization Tool for quantitative tissue expression of transcripts and proteins to searches of neXtProt, GPMdb, GTEx, Genomic Data Commons, PeptideAtlas, and Protein Atlas.

Fernando and Lydie: We are keen to interact with the Spanish Chr 16 group and the Chr 2/14 consortium on applying each other's methods and tools.

Mike, Chris, et al: I think it would be desirable to have a de minimus median level for transcript expression in each tissue (probably 1.0 RPKM) for the whole project, to avoid vast confusion across the C-HPP teams.

Jenny and Fernando: Hopefully, the B/D-HPP teams will contribute actively, too, with tissue-based orientation.

1. neXtProt 485 nearly-PE1 proteins excluded under Guidelines v2.1 (from Metrics 2016 Omenn et al, Supplementary table): 33 are on Chr 17, including 28 PE2, 4 PE3, 1PE4.

- We will determine where is the deficiency in qualifying for PE1 status in each case.

- We can determine tissues with transcript expression across multiple data resources (neXtProt/PeptideAtlas, GPMdb/green, Pandey/human proteome map, GTEx, Genomic Data Commons, and Protein Atlas) using MI-Proteome Visualization Tool MI-PVT (Panwar, JPR 2015).

- Last month I inquired of Lydie whether they had analyzed needed data for the 485; she responded "I would suggest to wait for our new release in February, since many of those proteins will probably be upgraded to PE1 due to the integration of all the HPP data from 2016 JPR SI."

2. While awaiting nX 2017-02, we searched nX2016-02 from the Supplementary Table of 2949 missing proteins in the Metrics 2016 paper: There are 148 Chr 17 missing proteins = 123 PE2, 19 PE3, 6 PE4. All of the 33/485 above are included in the 148.

We have started to perform Duek et al-like analysis for Chr 17.

There are 14 olfactory receptor coding genes, with 11 in a cluster, 10 PE2 and 4 PE3. Specifically:

- OR1A1 is listed as PE2, but in GTEx shows expression only in endocervix (N=5), with max RKPM value 0.04 and median value 0.00!

- OR1A2 listed as PE2, but GTEx shows expression only in testis (N=172) with median 0.00, 75th percentile 0.05, and a handful of points higher, with highest outlier 0.23 RPKM.

- OR1D4, PE3 in neXtProt apparently based on Zebrafish! However, nX gives a citation to Ben Arie, Lancet, et al, Hum Mol Genetics 3:229, 1994 (a top group in Israel's Weizmann Institute; developers of Gene Cards and many studies of olfactory receptor genes. Meanwhile, GTEx shows only negligible activity for one isoform in human testis; median RPKM 0.00 (N=172). Glusman/Lancet in Genomics 2000;63:227 described extremely complex sequences in the 17p13.3 cluster of 17 OR coding regions (6 pseudogenes; various interspersed

repetitive elements/LINE repeats; mapped to two families and seven subfamilies in seemingly random orientations; sounds like non-functional sequences!).

- We have long discussed in HPP potential strategies for analyses of gene families; those sequences that are highly homologous will be hard to distinguish from a few peptide sequences, thereby contributing to non-detectable protein products. The Lancet Lab published DEFOG: A practical scheme for deciphering families of genes; Genomics 2002;80:295. Using degenerate primers for PCR, DEFOG tripled the initial repertoire of human ORs in a single experiment. There were a billion combinations of distinct PCR primer pairs.

- I think we can exclude the OR genes from our analysis, as Duek et al did, given the failure of essentially all previous OR protein expression claims to survive independent reanalyses. This example shows a limitation of the PE2 strategy, namely reliance on extremely low RPKM reports.

- **Mark:** Are you proceeding with OR analyses across all chromosomes? We are interested to collaborate. I highly recommend the large body of work from Doron Lancet (with refs above).

- There are also 14 keratin-associated proteins at 17q21.2 For example, KRTAP2-2 (PE2) in GTEx shows expression only in skin, with a few very high values (RPKM 500-1000) but most very close to zero. KRTAP9-1 (PE3): also limited to skin, with only a few values above 0.5 RPKM.

- We will create a table showing not just PE2 or PE3, but highest expressed tissue(s) and median RPKM for the whole 148 predicted proteins. It is important to use multiple sources.

- For example, SMIM6 (P0D180) is “small integral membrane protein 6”. It is PE4 in nX. But in GTEx, there is quite high expression in testis (median 20 RPKM) and in stomach (median 7 RPKM) and Kidney cortex (median 6) across 172 samples; this would not qualify as testis-enriched, since the testis level is <5x other tissues. We even have a project using Drop-Seq for single cell analyses of spermatogenesis stages in mouse testis; that could yield some interesting findings for homology to human (PE3). For MS/MS, SMIM6 has only 62 amino acids with a single K at position 7 and no R residue! No tryptic peptides of 9-30 aa in length can be generated; ProteomeMap is blank by MS for all tissues (even though Pandey’s Nature paper used a single peptide of 6aa as sufficient; we checked that paper. Protein Atlas shows highest IHC in testis (along with RPKM 22, great agreement with GTEx).

- SMIM5 (77aa): ProteomeMap shows highest activity in platelets; none in testis; low in placenta, frontal cortex, spinal cord. This protein sequence has 6 R’s and 2 K’s, but they are clustered such that only the N-terminal and C-terminal tryptic peptides might be useful. In fact, Pandey reported only the N-terminal (10 aa/16 aa with a missed cleavage) and two modified N-terminal peptides (acetylated/oxidized). This is expected for an integral membrane protein. GTEx shows transcript expression medians >5 for several tissues, not including testis. In Protein Atlas, SMIM5 is detected as highly expressed in essentially all tissues by immunohistochemistry, but is highly variable and not well correlated in RNA expression. Might be a good one to subject to validation in the Antibody Validation Work Group project.

Response from Ron Beavis:

I've attached my analysis of the list.

The spreadsheet has 5 categories, broken out as separate tabs.

As of 1/16/17, we have added the numbers in each category and searched Vandenbrouck et al for their missing protein findings of Chr 17 proteins (see (c), below):

1. high priority (n=44): proteins that should be observable and for which there is some reasonably good evidence at the moment. I have indicated the tissues/cell types where they should be present. The (11) KRTAP proteins have considerable sequence homology, but I think there are enough differences to tease them out of good data.

2. peptide overlap (n=25): proteins that are observable, but very difficult (some impossible) to distinguish from other proteins at the level of tryptic peptides because of sequence identity. In these cases the problem is biology, not technique.

3. low priority (n=69: not enough information to know where to look for these proteins, if they exist at all.

4. revised genes (n=3): proteins associated with genes that have been significantly revised over the last few

years. The extensive revisions may indicate a problem with the DNA sequences.

5. retired+non-coding (n=28): genes that have either been retired from the current genome assembly (GRCh38) or are now considered either lincRNA or pseudogenes.

(c) Additional analyses as of 16 January 2017

Eric and Lydie:

Compared the 14 Chr 17 missing proteins reported by Vandenbrouck et al with the latest PeptideAtlas 148 proteins still missing (link below).

Eric, 10 of the 14 failed to pass your filters: FAM1887A (0 distinct pep); LRRC37A2 (representative); C17orf105 (0 pep); C17orf 74 (1 pep, subsumed); TMEM95 (0 pep); FBXW10 (2 pep, but marginally distinguished); FBX039 (1 pep); PROCA1 (0 pep); C17orf50 (1 pep); SLC25A39 (2 pep, marginally distinguished). That leaves only LRRC37A1, C1yorf987, SPEMC17orf83, and EFCAB3. Vandenbrouck et al report PRM assays for the one-hit wonders, including FBX039 (positive PRM); C17orf50 (PRM weak); TMEM95 (PRM missing, I think).

Ron, of these 14 you had LRRC37A2, LRRC37A1, C17orf98, C17orf105, C17orf74, PROCA1, C17orf50, and SLC25A39 as “high priority” (of your 44 Chr 17 missing proteins with high priority), with the rest ranked lower.

Another striking finding involves the 11 KRTAPs. Several have 30-60 peptides, but all end up as subsumed or indistinguishable in PeptideAtlas. Among the 25 from families with peptide overlap, there are 10 of the TBC1D3 family; there are several of these on Eric’s list for Chr 17.

An interesting individual low-priority protein is testis-expressed sequence 19 protein (Q8NA77) which generates only 6 peptides, two of which are <9 aa, 1 C-terminal, 1 extremely long, and 2 that seemed well-suited to MS; however, this protein has not been detected. PeptideAtlas shows this protein as negative by proteomics, but positive by antibodies.

MP Status:

Ron, of these 14 you had LRRC37A2, LRRC37A1, C17orf98, C17orf105, C17orf74, PROCA1, C17orf50, and SLC25A39 as “high priority” (of your 44 Chr 17 missing proteins with high priority), with the rest ranked lower.

Another striking finding involves the 11 KRTAPs; Ron, you suggested these might be differentiable. Several have 30-60 peptides, but all end up as subsumed or indistinguishable in PeptideAtlas. Among the 25 from families with peptide overlap, there are 10 of the TBC1D3 family; there are several of these on Eric’s list for Chr 17.

An interesting individual low-priority protein is testis-expressed sequence 19 protein (Q8NA77) which generates only 6 peptides, two of which are <9 aa, 1 C-terminal, 1 extremely long, and 2 that seemed well-suited to MS; however, this protein has not been detected. PeptideAtlas shows this protein as negative by proteomics, but positive by antibodies.

Chromosome 18

(Andrey Archakov)

At our page of Chr-18 we do not have so much "completely undiscovered" proteins, but we will be looking forward to verify the most unclear ones.

Reminder sent February 19 for a more detailed plan to complete Ch18.

Response:

In version neXtProt v.2.3.0. Chr18 contains 23 missing proteins (PE2-19; PE3-2 and PE4-2) from 275 entries. During our previous work we did transcriptome and proteome analysis of liver tissue and HepG2 cell line and found 4 missing proteins on transcriptomic level (1 for Liver and 3 for HepG2) and none at the protein level in these types of biomaterial. Now we are using the same transcripto-proteomic approach (RNASeq and SRM) for the study of sperm, as has been shown (Vandenbrouck et al., 2016) that this type of biomaterial contains the greatest diversity of expressed proteins.

As for the missing proteins there are some questions about status of proteins coded Chr18 according neXtProt. For example, GALR1 (https://www.nextprot.org/entry/NX_P47211/proteomics) has protein existence PE1 however there are no references to any experiments PeptideAtlas or Human Protein Atlas. And opposite situation for some PE2 proteins, which have proteomic experimental data from PeptideAtlas and HPA (https://www.nextprot.org/entry/NX_Q8TB69/proteomics). We wonder how it can be explanation of?

Chromosome 19

Ch19 PIC and group is being merged with another Ch group.

The C-HPP are still seeking a group willing to take over from Gyorgy who have officially pulled out from the C-HPP.

Chromosome 20

(Siqi Liu)

PIC Leaders: Siqi Liu, Qingyu He, Gong Zhang

Contributing labs:

Center of Proteomics Analysis, BGI, China

Institute of Biotechnology, Jinan University, China

Major lab members or partners contributing to the neXt50:

Yan Ren, Qidan Li, Shaohang Xu, Quanhui Wang (BGI)

Tong Wang, Gong Zhang (Jinan University)

Status of the Chromosome "parts list":

In neXtprot, data release: 2016-12-02, application release: v2.3.0

Chr	PE1	PE2	PE3	PE4	PE5	Total
20	456	62	15	3	13	549

According to CHPP definition to missing proteins, the proteins in PE2+PE3+PE4 are defined as missing proteins. Therefore, the missing proteins in chromosome 20 in the neXtprot database are 80 (62+15+3).

Confirmation that PIC and C-HPP lab members have read:

We have read the articles and understood the core concepts of missing proteins.

Step by step milestone plan to find, identify and validate MPs:

In the project of "Cancer Proteogenomics of the Esophageal Squamous Cell Carcinoma tissues", we are going to identify more than 40 missing proteins in chr. 20.

In the project of "Cancer Proteogenomics of the colorectal cancer", we are going to find more than 15 missing proteins in chr. 20.

We are able to search for missing proteins for all chromosomes, not only for just one chromosome. This may be even easier for us and can contribute more to the community.

To tackle the NeXt-50 challenge, we developed a new method, "high-throughput de novo proteome identification aided by translatome sequencing" to maximize the utilization of mass spectra. Using this method, we have identified altogether 840 missing proteins (755 PE2, 58 PE3, 7 PE4 and 26 PE5). Among them, we have identified 519 missing proteins with ≥ 2 unique peptides. We have verified some of them using MRM and antibody. We presented this strategy in AOHUPO 2016 as an oral presentation.

Milestone dates:

We will try the best to find over 50 missing proteins in ch 20 within 2017.

Chromosome 21

(Albert Sickmann)

Newly appointed PIC for Ch21 in February 2017. Plan expected soon.

Chromosome 22

(Akhilesh Pandy)

No response.

Chromosome X

(Tadashi Yamamoto)

Pending ChX meeting on Jan 20 regarding new leadership, team organization and plan.

Reminder sent February 19.

As of 21 Feb new leadership:

PI	Yasushi Ishihama	Kyoto University
Co-PI?	Tadashi Yamamoto	Niigata University

Chromosome Y

(Hosseini Salekdeh)

No response.

Reminder sent Feb 19.

Less than 20 unique missing proteins on Ch Y. We are focusing to identify them in various tissues and look for the function of missing proteins. Report expected soon.

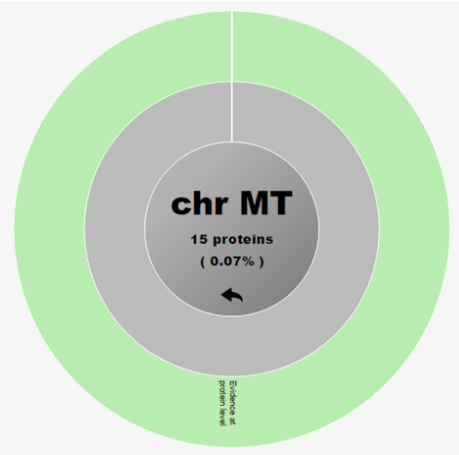
Chromosome Mt

(Andrea Urbani)

***Contributing labs:**

- Andrea Urbani, Catholic University of Sacred Hearth;
- Paola Roncada, Istituto Sperimentale Italiano L. Spallanzani, Milan, Italy;
- Mauro Fasano, University of Insubria, Busto Arsizio, Italy;
- Maurizio Ronci, University G. d'Annunzio of Chieti-Pescara, Chieti, Italy;
- Tiziana Alberio, University of Insubria, Busto Arsizio, Italy;
- Luisa Pieroni, Santa Lucia Foundation, IRCCS, Rome, Italy
- Alessio Soggiu, University of Milan, Milan, Italy;
- Antonio Lucacchini, University of Pisa, Pisa, Italy;
- Italia Bongarzone, IRCCS National Cancer Institute (INT) Milan
- Luigi Palmieri, University of Bari and CNR - IBBE – Bari, Italy;
- Cristina Banfi, Cardiological center, Monzino, Milan;
- Fulvio Magni, University of Milano-Bicocca, Milan, Italy;
- Gabriella Tedeschi, University of Milan, Milan, Italy;
- Salvatore Foti, University of Catania;
- Simona Fontana, University of Palermo, Palermo, Italy;
- Pietro Pucci, University of Naples 'Federico II', Naples, Italy.
- Albert Sickmann, ISAS-Dortmund, DE
- Emma Lundberg, SciLifeLab, KTH Royal Institute of Technology, Stockholm, SW
- Mohan Babu, University of Regina, CA
- Peipei Ping, UCLA, California, USA
- Chris Borchers, University of Victoria Genome British Columbia Proteomics Centre, CA

***Status of the Chromosome "parts list":**



Chr	PE1	PE2	PE3	PE4	PE5	Total
MT	15	0	0	0	0	15

***Confirmation that PIC and C-HPP lab members have read:**

Yes, we have read them.

***Step by step milestone plan to find, identify and validate MPs:**

For the Mitochondrial Proteome Initiative, Italian consortium purpose is to investigate the protein mitochondrial repertoire in 10 cell lines, which include the following cell lines: HeLa, Hek293, U2OS, BJ, SH-SY5Y, MDA-MB-231, NCI-H28, HUVEC, THP1, HepG2. It has been necessary and indispensable develop a well-standardized method for each cellular model and for the specific biological issue faced.

In brief the following steps have been done:

- Mitochondria have been isolated by three different methods (differential centrifugation, sucrose gradient separation and Mitochondrial Isolation Kit MITOISO2 (Sigma-Aldrich)) in order to evaluate the best preparation;
- Integrity of mitochondrial preparations have been assessed by measuring the oxygen consumption rate and the activity of selected enzymes. They have been evaluated with microscale oxygraphy (Seahorse Bioscience XF96 Analyzer);
- After digestion, mitochondrial peptides have been analyzed using several MS platforms and chromatographic conditions (short list of instrument currently employed: nanoAcquity M Class coupled to Synapt G2 Si (waters); nanoEASY II coupled to Bruker maXis HD; Dionex UltiMate 3000 coupled to Bruker Impact HD; Dionex UltiMate 3000 coupled to Orbitrap Velos ETD; Dionex UltiMate 3000 coupled to Orbitrap Fusion; Ekspert nanoLC 400 coupled to TripleTOF 5600+; nanoEASY II coupled to LTQ-Orbitrap-XL);
- Raw data have been processed using Peaks 7.5 (Bioinformatics Solutions Inc., Waterloo, ON Canada) and searched without taxonomy restriction against a custom database containing human proteins, common contaminants and yeast enolase (P00924) downloaded from Uniprot repository (20,442 total entries);
- In the light of the initiative of TAMPA we are checking the presence of missing proteins following the updated Guidelines 2.1.0 evaluating the coding in any chromosome. Open data sharing available before publication following data transfer agreement acceptance.

***Milestone dates:**

- May 2017: special issue JPR, standardization paper on mitochondria preparation

- September 2017: Dublin, HUPO 2017