Progress Identifying and Characterizing the Human Proteome: HPP Metrics and Informatics Innovations for MPs and uPE1s

Gilbert S. Omenn, MD, PhD
Harold T. Shapiro Distinguished University Professor
Center for Computational Medicine and Bioinformatics
University of Michigan, Ann Arbor, MI, USA
Chair, HUPO Human Proteome Project

17th HUPO World Congress
Orlando, Florida
30 September, 2018
Goals of the Human Proteome Project

1. **Make proteomics a full counterpart to genomics:** Enhance the work of the entire biomedical research community with high-throughput instruments, reagents, specimens, pre-analytical preparation, and knowledge bases for identification, quantitation, and characterization of proteins in network context.

2. **Complete in stepwise fashion the Protein Parts List**—identifying and characterizing at least one protein product and as many PTM, SNP, and splice variant isoforms as possible from all of the 20,055 human protein-coding genes. Find missing proteins (next-50 MP); characterize known proteins (uPE1 challenge); predict undetectable proteins.
Searching for the real stuff of life

The discovery that humans have fewer genes than expected has thrust proteins into the research spotlight, says Victoria Griffith.

Now, companies are racing to decipher the human protein set.
Early Leaders of HUPO
The Vision of the HPP (2010-2018)

Future knowledge on proteins: structured

Biology driven projects

2015

2010

time

Present knowledge on proteins: random

KB

MS

Abs

« Adopt-a-chromosome » consortia

4th Resource Pillar added in 2018 = Pathology
HUPO Human Proteome Project Milestones

- HUPO announced the HPP at Sydney-2010, launch at Geneva-2011.
- 164 publications to date in 5 annual special issues of the *Journal for Proteome Research*; 6th in process for online, print to follow.
- ProteomeXchange; Guidelines (v2.1, 2016) for Interpretation of MS Data; HPA Antibody Validation Working Group
- SRM Atlas, PASSEL resource, synthetic peptides for identification and quantitation of all protein-coding genes with targeted proteomics
- C-HPP Next-50 Missing Proteins and uPE1 Challenges
- B/D-HPP Top 50 Popular Proteins for organ-specific research
- MS Pillar community sample with 96 phospho-peptides

See [https://hupo.org/human-proteome-project](https://hupo.org/human-proteome-project).
Overview of the HPP Data Workflow
Three Parts of this Introduction

1. Annual Update with HPP Metrics for Progress on the Human Proteome Parts List (Omenn et al, JPR, 2018, DOI:10.1021/acs.jproteome.8b00441) #127

2. Deep Dive on Chr 17 to understand how 43 MPs were added to neXtProt PE1 since the beginning of the next-50 MP Challenge in 2016 and to guide search for the 35 more of the remaining 105 MPs (Siddiqui et al, JPR 2018, online) poster #085

3. Application of I-TASSER & COFACTOR algorithms to predict functions of uPE1 proteins, starting with 66 on Chr 17 (Zhang et al, JPR 2018) Tues 11:34, TOC. All in C-HPP poster session.
## neXtProt Protein Existence Evidence Levels from 2012 to 2018 Showing Progress in Identifying PE1 Proteins and PeptideAtlas Canonical Proteins

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Evidence at protein level</td>
<td>13,975</td>
<td>15,646</td>
<td>16,491</td>
<td>16,518</td>
<td>17,008</td>
<td><strong>17,470</strong> a</td>
</tr>
<tr>
<td>2: Evidence at transcript level</td>
<td>5205</td>
<td>3570</td>
<td>2647</td>
<td>2290</td>
<td>1939</td>
<td>1660</td>
</tr>
<tr>
<td>3: Inferred from homology</td>
<td>218</td>
<td>187</td>
<td>214</td>
<td>565</td>
<td>563</td>
<td>452</td>
</tr>
<tr>
<td>4: Predicted</td>
<td>88</td>
<td>87</td>
<td>87</td>
<td>94</td>
<td>77</td>
<td>74</td>
</tr>
<tr>
<td>5: Uncertain or dubious</td>
<td>622</td>
<td>638</td>
<td>616</td>
<td>588</td>
<td>572</td>
<td>574</td>
</tr>
</tbody>
</table>

**Human PeptideAtlas canonical proteins**

12,509  13,377  14,928  14,629  15,173  15,798

---

a Percent of predicted proteins classified as PE1 by neXtProt = PE1/PE1+2+3+4 = **89%**.
b Missing Proteins PE 2+3+4 = 2186, down from 2579 in neXtProt v2017-01.

More stringent guidelines were imposed in 2016.
Detailed Accounting for Changes in neXtProt PE1,2,3,4,5 from 2017-01 to 2018-01

2017 release
Total: 20159

17008 PE1 → 431 → 2139 → 2186 MP
3 → 4 → 6

2579 MP

572 PE5

561

2018 release
Total: 20230

16995 → 44 → 17470 PE1

Deleted in 2018

40

7
neXtProt PE Classes 2018-01-17, with subgroups of PE1 and of MPs PE2,3,4

1378 PE1 non-MS: 530 PPI, 99 Edman, 170 PTMs, 176 disease mutations, 75 3D, 58 Ab, 270 other biochemical studies
The 8 Largest Contributors to Peptide Atlas Canonical Proteins Growth in 2017

- **PXD001194**: Harel et al. (2015): microparticles plasma biomarkers
- **PXD006482**: Peng et al. (2017): phosphoproteome of kidney cancer
- **PXD006465**: Wang et al. (2017): multi-protease in testis tissue
- **PXD006471**: Liu et al. (2017): pituitary and thyroid proteome
- **PXD006833**: Li et al. (2017): enrichment with ProteoMiner
- **PXD003431**: Schmitges et al. (2016): C2H2 zinc finger proteins
- **PXD004352**: Rieckmann et al. (2017): immune cells
- **PXD004927**: Hendriks et al. (2017): SUMO proteome
Progress on Finding Proteins in Six Largest Families/Groups
Fate of 73 Missing Proteins Nominated for neXtProt Review in JPR 2017

- Number promoted to PE1 in nextProt 2018-01-07
- ProteoMiner beads/Triton X-100/PRM (Li): 15
- Sperm proteome/PRM (Carapito): 12
- Multi-proteases/testis (Wang): 3
- Phosphoproteome of kidney (Peng): 1
- Chr Y protein in cardiac development (Meyfour): 1
- Stranded peptides strategy (Elguoshy): 41

Result in neXtProt: 43 new PE1 proteins, though 18 were based on other sources of data (Wang, Meyfour, Elguoshy)
Strategies for Accelerating the HPP

- Greatly expand analysis of splice variants, PTMs, sequence variants, and N-termini
- Use advanced instruments and targeted proteomics to elucidate biological networks, protein complexes, disease mechanisms
- Deepen proteogenomics analyses
- Collaborate on searches for “missing proteins” and uPE1 proteins lacking functional annotation
- Utilize “popular proteins” lists and SRMAAtlas for a wide range of organ-specific research
Chromosome 17 Missing Proteins Strategy focused on Annotation, following Chr 2/14 Consortium (Duek et al, 2016)

Summary: The PE2,3,4 Chr 17 Missing Proteins have been reduced from 148 to 105, based on the neXtProt version 2016-01. Thus, we have 43 new PE 1 proteins toward the next-50 MP goal of 50 officially announced by the C-HPP in September 2016 at the Sun Moon Lake HPP Workshop.

Progress for Chromosome 17 is as follows:

<table>
<thead>
<tr>
<th>neXtProt version</th>
<th>PE2+3+4</th>
<th>PE2</th>
<th>PE3</th>
<th>PE4</th>
<th>[PE5]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-01</td>
<td>148</td>
<td>123</td>
<td>19</td>
<td>6</td>
<td>23</td>
</tr>
<tr>
<td>2017-01</td>
<td>125</td>
<td>103</td>
<td>17</td>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>2017-08</td>
<td>114</td>
<td>98</td>
<td>12</td>
<td>4</td>
<td>23</td>
</tr>
<tr>
<td>2018-01</td>
<td>105</td>
<td>88</td>
<td>13</td>
<td>4</td>
<td>23</td>
</tr>
</tbody>
</table>
How 43 MPs were Upgraded to PE1 between 2016 and 2018 in neXtProt by MS +/- PPI

Omer Siddiqui, Hongjiu Zhang, Yuanfang Guan, Gil Omenn
Overall Strategy for Finding the Remaining 105 Chr 17 MPs with MS or PPI
For MS, 99/105 have 2 predicted proteotypic peptides; 29 have one annotated in neXtProt.
Among 29 with a single proteotypic peptide in PA/neXtProt, we found a second non-nested “stranded” peptide for 7 in GPMdb with PXD identifier and data in PRIDE.

<table>
<thead>
<tr>
<th>Missing.Protein</th>
<th>First.Peptide</th>
<th>Second.Peptide</th>
<th>log_{e}</th>
<th>PXD</th>
<th>PRIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD300C</td>
<td>MTVAGPVGGSLSVQCR</td>
<td>DSPEPSPHPGSLFSNVR</td>
<td>-8.0</td>
<td>PXD004034</td>
<td>Yes</td>
</tr>
<tr>
<td>CSHL1</td>
<td>STFTNNLNYDTSDDYHLLK</td>
<td>NYGLLHCFR</td>
<td>-11.2</td>
<td>PXD002391</td>
<td>Yes</td>
</tr>
<tr>
<td>EVPLL</td>
<td>VTQECAYCALYEK</td>
<td>MQASADQVER</td>
<td>-10.5</td>
<td>PXD000109</td>
<td>Yes</td>
</tr>
<tr>
<td>PIRT</td>
<td>VLEVDEKSPEAK</td>
<td>DLLPSQTASSLCISSR</td>
<td>-10.7</td>
<td>PXD005336</td>
<td>Yes</td>
</tr>
<tr>
<td>TMEM92</td>
<td>GPLELPSLPETP</td>
<td>CGLILACP</td>
<td>-6.9</td>
<td>PXD002121</td>
<td>Yes</td>
</tr>
<tr>
<td>RNF222</td>
<td>RSRALLLITLIAVVAVVAAILPWVLLVR</td>
<td>HGMLPGEQDSVLPR</td>
<td>-10.3</td>
<td>PXD003028</td>
<td>Yes</td>
</tr>
<tr>
<td>SLC16A5</td>
<td>QAVAADALERDLFLEAK</td>
<td>ECPPPPPETPALGCLAACGR</td>
<td>-8.5</td>
<td>PXD005748</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Spectral Match of Observed and Synthetic DLLPSQTASSLCISSR Peptide for PIRT
Summary of 35 High-Priority Chr 17 MPs

35 High-priority Chromosome 17 Missing Proteins

105 MPs On Chr17

35 Prioritized MPs

62 MPs for MS consideration/ 25 MPs to search for

25 MPs
- Transcription of at least 10 TPM median in at least one tissue

37 MPs
- Low Transcription/ Not PE2

10 MPs
- One Proteotypic Peptide of Length at least 9 aa already detected

16 MPs
- No Proteotypic Peptides of Length at least 9 aa already detected

43 MPs for PPI consideration (not amenable to MS)/ 10 MPs with promising baits

13 Olfactory Receptors

10 Membrane-Embedded Proteins

5 KRTAPs

13 TBC1Ds

8 MPs
- 4 Olfactory Receptors (OR3A1, OR1E2, OR3A2, OR1A2)
- 4 Membrane-Embedded Proteins (TMEM220, TMEM92, TMEM99, TMEM95) with assorted baits

2 other MPs lacking two proteotypic peptides

2 MPs
- KRTAP9-1, KRTAP9-7 with other KRTAPs as baits

88 PE2

13 PE3

4 PE4
Predicting uPE1 Functions as GO Terms with I-TASSER and COFACTOR Algorithms

Chengxin Zhang, GS Omenn, Yang Zhang, U Michigan
Predicting uPE1 Functions as GO Terms with I-TASSER and COFACTOR Algorithms

uPE1 protein sequence

I-TASSER protein structure prediction

Query structure model

Global and local structure alignment

Structure analogs with known functions

Human chromosome 17

(PSI-)BLAST homologous sequence search

STRING protein interaction database mapping

Protein-Protein Interactions (PPI)

Weighted average

Functions from PPI

Functions from structure

Functions from sequence

Sequence homologs

GO:003674

GO:0003824

GO:0016787

GO:0140096

GO:0017171

GO:0008233

GO:0070011

GO:0008236

Protein Functions by COFACTOR
GO Term Prediction Accuracy (Fmax) with several methods on 100 random PE1 Chr 17 Proteins as Benchmark Analysis
Prediction of GO Terms for MF, BP, CC with I-TASSER/COFACTOR Pipeline on Benchmark Set
Results for GO Terms MF, BP, CC for the 66 uPE1 Chromosome 17 Proteins (13,33, 49 exceed thresholds)
Summary of Progress Reported Today

- Metrics: Now 17,470 PE1 (89% of total predicted proteins) and 15,798 canonical proteins. There are now 2186 PE2,3,4 MPs.

- Very close to next-50 MP Challenge goal for Chr 17, with 43 new PE1, an excellent MS candidate, 35 high-priority PPI (10) or MS (25) targets.

- At least 13 uPE1 Chr 17 proteins with high-confidence functional annotations using I-TASSER/COFACTOR and Gene Ontology.
<table>
<thead>
<tr>
<th>NeXtProt ID (Gene Name)</th>
<th>Molecular Function (MF)</th>
<th>Biological Process (BP)</th>
<th>Cellular Component (CC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 * NX_Q8TBR7-2 (FAM57A)</td>
<td>GO:0016740 (1.00) transferase activity GO:0050291 (0.99) sphingosine N-acyltransferase activity</td>
<td>GO:0032502 (0.69) developmental process GO:0007420 (0.54) brain development</td>
<td>GO:0005887 (1.00) integral component of plasma membrane GO:0005886 (1.00) plasma membrane</td>
</tr>
<tr>
<td>2 NX_Q12767-1 (TMEM94)</td>
<td>GO:0022892 (0.91) substrate-specific transporter activity GO:0046873 (0.57) metal ion transmembrane transporter activity</td>
<td>GO:0065008 (0.80) regulation of biological quality GO:0030001 (0.56) metal ion transport</td>
<td>GO:0005654 (1.00) nucleoplasm</td>
</tr>
<tr>
<td>3 NX_Q5BKU9-1 (OXLD1)</td>
<td>GO:0016491 (0.87) oxidoreductase activity GO:0004128 (0.73) cytochrome-b5 reductase activity, acting on NAD(P)H</td>
<td>GO:0015701 (0.90) bicarbonate transport GO:0008652 (0.53) cellular amino acid biosynthetic process</td>
<td>GO:0005739 (0.90) mitochondrion GO:0005737 (0.66) cytoplasm</td>
</tr>
<tr>
<td>4 * NX_A6NGC4-1 (TLCD2)</td>
<td>GO:0016740 (0.86) transferase activity GO:0050291 (0.76) sphingosine N-acyltransferase activity</td>
<td>GO:0006643 (0.76) membrane lipid metabolic process GO:0006672 (0.73) ceramide metabolic process</td>
<td>GO:0016021 (1.00) integral component of membrane GO:0005783 (0.75) endoplasmic reticulum</td>
</tr>
<tr>
<td>5 * NX_O43934-1 (MFSD11)</td>
<td>GO:0005215 (0.85) transporter activity GO:0005351 (0.66) sugar:proton symporter activity</td>
<td>GO:0006810 (0.82) transport GO:0008643 (0.68) carbohydrate transport</td>
<td>GO:0016021 (1.00) integral component of membrane GO:0005887 (0.77) integral component of plasma membrane</td>
</tr>
<tr>
<td>6 NX_Q9P298-1 (HIGD1B)</td>
<td>GO:0016740 (0.79) transferase activity GO:0061630 (0.71) ubiquitin protein ligase activity</td>
<td></td>
<td>GO:0043234 (0.88) protein complex GO:0005634 (0.71) nucleus</td>
</tr>
<tr>
<td>NeXtProt ID (Gene Name)</td>
<td>Molecular Function (MF)</td>
<td>Biological Process (BP)</td>
<td>Cellular Component (CC)</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------------------------</td>
<td>-------------------------</td>
<td>-------------------------</td>
</tr>
</tbody>
</table>
| 7 NX_Q2TAL 5-1 (SMTNL2) | GO:0008092 (0.77) cytoskeletal protein binding | GO:0016043 (0.70) cellular component organization  
  GO:0048856 (0.59) anatomical structure development | GO:0005737 (0.66) cytoplasm  
  GO:0044430 (0.50) cytoskeletal part |
| 8 NX_Q9BQS 6-1 (HSPB9) | GO:0042802 (0.76) identical protein binding  
  GO:0051082 (0.52) unfolded protein binding | GO:0050896 (0.82) response to stimulus  
  GO:0042981 (0.51) regulation of apoptotic process | GO:0005634 (0.97) nucleus  
  GO:0005737 (0.96) cytoplasm |
| 9 NX_Q96LD 4-1 (TRIM47) | GO:0004842 (0.76) ubiquitin-protein transferase activity | GO:0031323 (0.54) regulation of cellular metabolic process  
  GO:0019538 (0.54) protein metabolic process | GO:0005737 (0.57) cytoplasm |
| 10 NX_Q8N7B 9-1 (EFCAB3) | GO:0043169 (0.74) cation binding | GO:0019538 (0.58) protein metabolic process | GO:0016020 (0.82) membrane  
  GO:0005737 (0.68) cytoplasm |
| 11 * NX_Q6AI12 1-1 (ANKRD40) | GO:0008092 (0.62) cytoskeletal protein binding  
  GO:0030507 (0.57) spectrin binding | GO:0060255 (0.62) regulation of macromolecule metabolic process  
  GO:0016043 (0.60) cellular component organization | GO:0005737 (0.77) cytoplasm  
  GO:0043234 (0.51) protein complex |
| 12 NX_Q6UX5 2-1 (C17orf99) | GO:0004872 (0.63) receptor activity  
  GO:0019199 (0.50) transmembrane receptor protein kinase activity | GO:0032502 (0.68) developmental process  
  GO:0030030 (0.54) cell projection organization | GO:0031224 (1.00) intrinsic component of membrane  
  GO:0005887 (0.63) integral component of plasma membrane |
| 13 NX_Q3MH D2-1 (LSM12) | GO:0003723 (0.59) RNA binding | GO:0090304 (0.79) nucleic acid metabolic process  
  GO:0016070 (0.73) RNA metabolic process | GO:0005576 (0.55) extracellular region |