neXtProt - 2019 update

Lydie Lane
www.nextprot.org is the HPP reference knowledge base for human protein curation collects and represents knowledge on human proteins at genomic, transcriptomic and proteomic levels.
neXtProt release 2019-01-11

• Reference for the JPR 2019 HPP SI

• Based on UniProtKB release 2018_11

• Includes the Jan 2019 “Human build” of PeptideAtlas (1,772,990 natural peptides in total) and the latest SRMAAtlas data

• PEFF export format adapted to the most recent specifications from HUPO PSI
neXtProt and the MP50 project

neXtProt release 2019-01-11

2129 missing proteins (MP)
(2186 MP in 2018)
Changes of PE status of neXtProt entries from 2018 to 2019

2018 release
Total: 20230

17470 PE1
2186 MP
574 PE5

2019 release
Total: 20399

17694 PE1
2129 MP
576 PE5

Deleted in 2019

New

2019 release
Total: 20399

17694 PE1
2129 MP
576 PE5

2018 release
Total: 20230

17470 PE1
2186 MP
574 PE5

Deleted in 2019

78 new
116
Mass spectrometry information in neXtProt release 2019-01-11

- PE1, MS-validated: 1688
- PE1, validated with non-MS criteria: 441
- Missing (PE2+PE3+PE4), insufficient MS info: 1096
- Missing (PE2+PE3+PE4), no MS info: 576
- Dubious/uncertain: 16598
pe1 in neXtProt and the next-CP50 project

Launching the C-HPP neXt-CP50 Pilot Project for Functional Characterization of Identified Proteins with No Known Function

Termed “uPE1” by C-HPP

Unknown function??
Types and sources of functional information in neXtProt

- “Free-text” UniProtKB function summaries and associated KWs
- Structured pathway annotations
- GO-MF and GO-BP annotations
Example: AMOTL1 (NX_Q8IY63)

**Free text summary of the literature (Swiss-Prot curators)**

**GO MOLECULAR FUNCTION**
- Identical protein binding: GO:0042802
- Protein binding: GO:0005515

**Generic terms, not indicative of a function**

**GO BIOLOGICAL PROCESS**
- Actin cytoskeleton organization: GO:0003036
- Angiogenesis: GO:0001525
- Establishment of cell polarity involved in ameboidal cell migration: GO:0003365
- Hippo signaling: GO:0035329
- **Positive regulation of blood vessel endothelial cell migration**: GO:0043536
- Regulation of cell migration: GO:0030334
- Wnt signaling pathway: GO:0016055

**Specific GO terms assigned by a variety of sources (Gold/Silver)**

**KEGG PATHWAYS**
- Tight junction: hsa04530-154810

**Pathway annotations**

**REACTOME PATHWAYS**
- Signaling by Hippo: R-HSA-2028269
Entries without or only “generic” functional annotation (NXQ_00022)

->2222 “uncharacterized proteins” in neXtProt release 2019-01-11

- 1254 uPE1
- 638 uMP
- 330 uPE5
Curation is a dynamic process!

2018-01 release

With function annotated: 2323 (333 dubious)
Without function annotated: 2041

2019-01 release

With function annotated: 2222
(330 dubious)

New: 58

17

Deleted in 2019
neXtProt release 2019-08-22

• No major change in the data and annotations

• Integration of gnomAD frequency data (derived from 125,748 exomes) for 2,691,323 neXtProt variants (45%)

• New protein cleavage tool

• Integration of I-TASSER/COFACTOR as “community tool”
Integration of variant frequencies from gnomAD

NEW frequency data with link to gnomAD
Queries to explore the new variant frequency data (on SNORQL interface only)
New protein digestion tool
### DIGESTED PEPTIDES FOR EACH PROTEASE:

<table>
<thead>
<tr>
<th>Protease name</th>
<th>Peptide count</th>
<th>Unique peptide count</th>
</tr>
</thead>
<tbody>
<tr>
<td>LYS_C</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PRPS1_P1_1_1_C</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PRPS1_P1_1_H_2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PROTEINASE_K</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>THERMOLYSIN</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TRYPsin</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GIU_C_PHOSPHATE</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CHYMOTRYPSIN_High_Spec</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

---

**Feedback and suggestions welcome!**
Integration of the 1st community tool in neXtProt: I-TASSER/COFACTOR
More details at the Bioinformatics hub session on Tuesday at 4 pm
More about neXtProt... on Tuesday, Sept 17:

• Keynote talk at 10:40 am, session: Bioinformatics and Statistics
  “neXtProt: a SPARQLing light in the dark human proteome”

• Hands-on session on SPARQL at 2 pm at the Bioinformatics Hub

• Session on I-TASSER/COFACTOR with Gil Omenn at 4 pm at the Bioinformatics Hub

... any question now?