# In quest of SPARQLing lights in the dark human proteome

Lydie Lane

CASSS Mass spectrometry symposium
Long Beach, CA, Sept 28-30 2022

Swiss Institute of Bioinformatics

UNIVERSITÉ DE GENÈVE

# Why is it important to characterize the human proteome?

- Proteins (not genes) are the active components of cells
- Defects in proteins can cause diseases (e.g. genetic diseases or cancers)
- Proteins can be drug targets
- Proteins can be biomarkers

# What is “the” human proteome?


## human genome

**~ 20’000 protein coding-genes**

**alternative splicing of mRNA**

2-5 fold increase

**~ 50 to 100’000 transcripts (mRNAs)**

## human proteome

**~ 5'000'000 different proteoforms**

**post-translational modifications of proteins (PTMs)**

50-100 fold increase

ıte, (3S)-3-hydroxyasparagine, 1'-histidyl-3'-tyrosine, 1-nine, 3'-(S-cysteinyl)-tyrosine, 3-hydroxyproline, 3-4-carboxyglutamate, 4-hydroxyproline, 5-glutamyl, 5-lroxylysine, 5-imidazolinone, ADP-ribosylasparagine, ADP-ribosylcysteine, ADP-ribosylserine, Allysine, Arginine amide, Asparagine amide, Aspartate 1-(chondroitin 4-sulfate)-ester, Asymmetric dimethylarginine, Beta-decarboxylated aspartate, Cholesterol glycine ester, Citrulline, Cysteine methyl ester, Cysteine sulfenic acid, Cysteinyl-selenocysteine, Deamidated asparagine, Deamidated glutamine, Dimethylated arginine, Diphthamide, Disulfide bond, GPI-anchor amidated alanine, GPI-anchor amidated asparagine, GPI-anchor amidated aspartate, GPI-anchor amidated cysteine, GPI-anchor amidated glycine, GPI-anchor amidated serine, Glutamic acid 1-amide, Glutamine amide, Glycine amide, Glycyl adenylate, Glycyl lysine isopeptide, Hydroxyproline, Hydroxyproline, Hypusine, Isoglutamyl cysteine thioester, Isoglutamyl lysine isopeptide, Isoleucine amide, Leucine amide, Leucine methyl ester, Lysine amide, Lysine tyrosylquinone, Methionine amide, N,N,N-trimethylalanine, N-acetylalanine, N-acetylaspartate, N-acetylcysteine, N-acetylglutamate, N-acetylglycine, N-acetylmethionine, N-acetylproline, N-acetylserine, N-acetylthreonine, N-acetylvaline, N-myristoyl glycine, N-palmitoyl cysteine, N-palmitoyl glycine, N-pyruvate 2-iminyl-valine, N4,N4-dimethylasparagine, N6,N6,N6-trimethyllysine, N6,N6-dimethyllysine, N6-(pyridoxal phosphate)lysine, N6-(retinylidene)lysine, N6-1-carboxyethyl lysine, N6-acetyllysine, N6-biotinyllysine, N6-carboxylysine, N6-lipoyllysine, N6-methylated lysine, N6-methyllysine, N6-myristoyl lysine, Nitrated tyrosine, O-(pantetheine 4'-phosphoryl)serine, O-AMP-threonine, O-AMP-tyrosine, O-acetylserine, O-acetylthreonine, O-decanoyl serine, O-palmitoyl serine, Omega-N-methylarginine, Omega-N-methylated arginine, Omega-hydroxyceramide glutamate ester, Phenylalanine amide, Phosphatidylethanolamine amidated glycine, Phosphohistidine, Phosphoserine, Phosphothreonine, Phosphotyrosine, PolyADP-ribosyl glutamic acid, Proline amide, Pyrrolidone carb
FAD cysteine, S-Lysyl-methion
cysteine, S-glutathionyl cystein
cysteine, Sulfoserine, Sulfotyrosine, Symmetric dimethylarginine, Tele-8alpha-FAD histidine, Tele-methylhistidine, Thyroxine, Triiodothyronine, Tyrosine amide, Valine amide

**+ all the individual variations (SNPs, disease mutations, etc)**

**+ all the different glycosylation forms and the processing events !**

# The main tasks of protein knowledgebases

- provide a reference set of protein sequences predicted from genome(s) analysis

- gather experimental validations for the predicted proteins

- keep information up to date about the predicted/confirmed functions of proteins and proteoforms in health and disease

# Curating the human proteome

## Known knowns

- *Well characterized proteins*
- *Known PTMs and variants regulating function or localization of proteins*

MANUAL CURATION OF PAPERS

## Known unknowns

- *Expected proteins*

**GENOMIC/TRANSCRIPTOMIC DATA ANALYSIS**

- *Uncharacterized proteins identified in human samples*

**PROTEOMIC DATA ANALYSIS**

- *Predicted PTMs*

**AUTOMATIC PREDICTIONS**

- *Genomic variants of unknown significance*

**GENOMIC DATA ANALYSIS**

# Tracking the known unknowns in the human proteome

- The **Human Proteome Project (HPP)** (www.thehpp.org), launched in 2010, is an international project organized by the Human Proteome Organization (HUPO) *that aims to revolutionize our understanding of the human proteome via a coordinated effort by many research laboratories around the world.*
- *It is designed to map the entire human proteome in a systematic effort using currently available and emerging techniques. Completion of this project will enhance understanding of human biology at the cellular level and lay a foundation for development of diagnostic, prognostic, therapeutic, and preventive medical applications.*

  - **Goal 1: validate the existence of 1 protein product per gene**
  - **Goal 2: find at least one biological function for every gene**

Human Proteome Project

HUPO HUMAN PROTEOME ORGANIZATION | 2001 2021 celebrating 20 years

# neXtProt, the human protein knowledge platform

- neXtProt (www.nextprot.org) was created in 2011 in order to organize human protein knowledge

- It has been chosen as **the reference database for the HUPO Human Proteome Project** in 2013

- neXtProt was the first resource to implement the HUPO Protein Standard Initiative (PSI) **PEFF format**, an enriched FASTA format allowing MS search engines and other tools to easily and consistently access sequence variations and PTM data.

Journal of proteome research
Technical Note
pubs.acs.org/jpr

**neXtProt: Organizing Protein Knowledge in the Context of Human Proteome Projects**

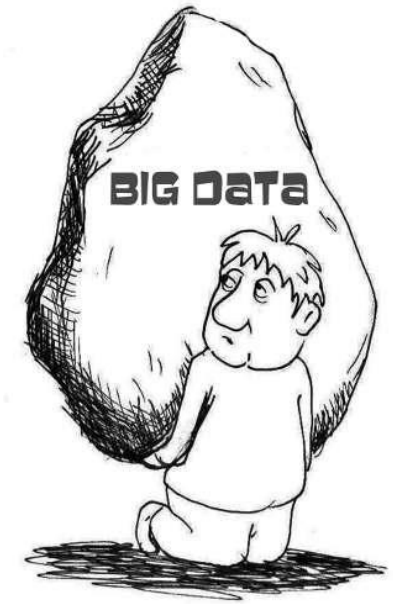
Pascale Gaudet,[†] Ghislaine Argoud-Puy,[†] Isabelle Cusin,[†] Paula Duek,[†] Olivier Evalet,[†] Alain Gateau,[†] Anne Gleizes,[†] Mario Pereira,[†] Monique Zahn-Zabal,[†] Catherine Zwahlen,[†] Amos Bairoch,[†,‡] and Lydie Lane*[,†,‡]

dx.doi.org/10.1021/pr300830v | *J. Proteome Res.* 2013, 12, 293−298

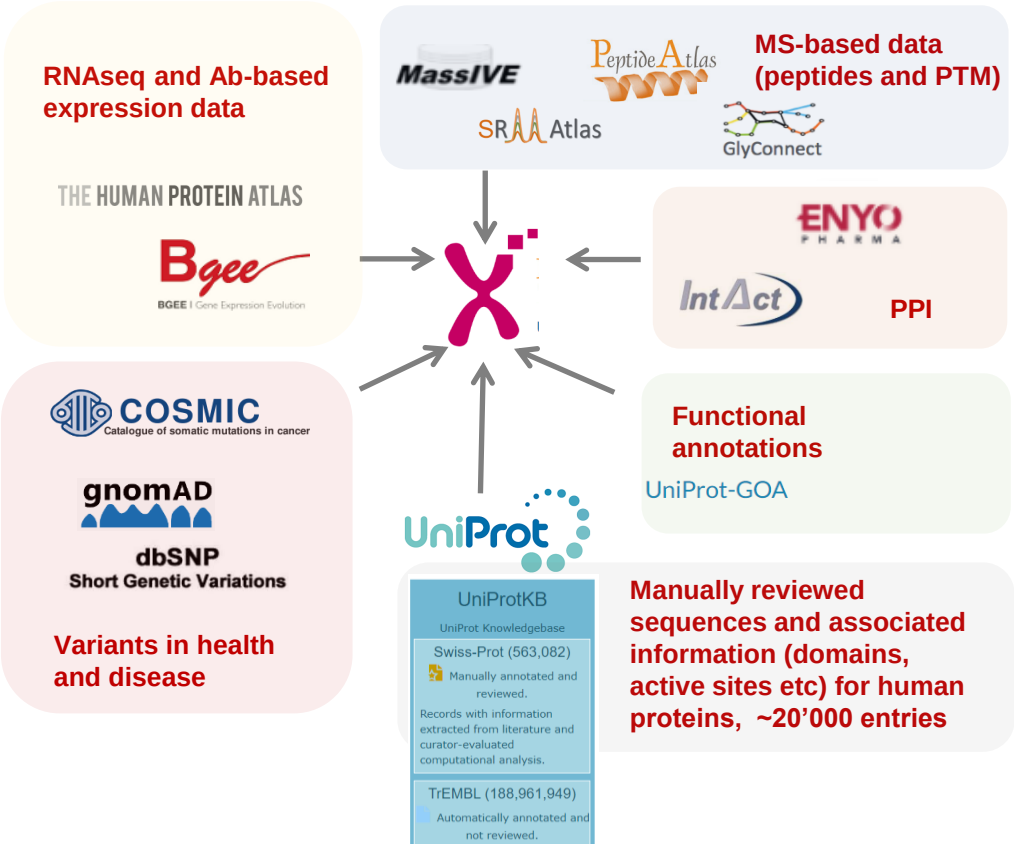
# neXtProt, the human protein knowledge platform

***Huge amounts of data on the ~20,300 predicted protein-coding human genes constantly being generated by the life science community***

***[genomic variations in health and disease, mRNA expression in tissues, protein identification by antibody-based or mass spectrometry-based techniques, protein-protein interactions, PTMs, etc.]***




neXtprot
Exploring the universe of human proteins




*D328–D334 Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 14 November 2019*
*doi: 10.1093/nar/gkz995*

**The neXtProt knowledgebase in 2020: data, tools and usability improvements**

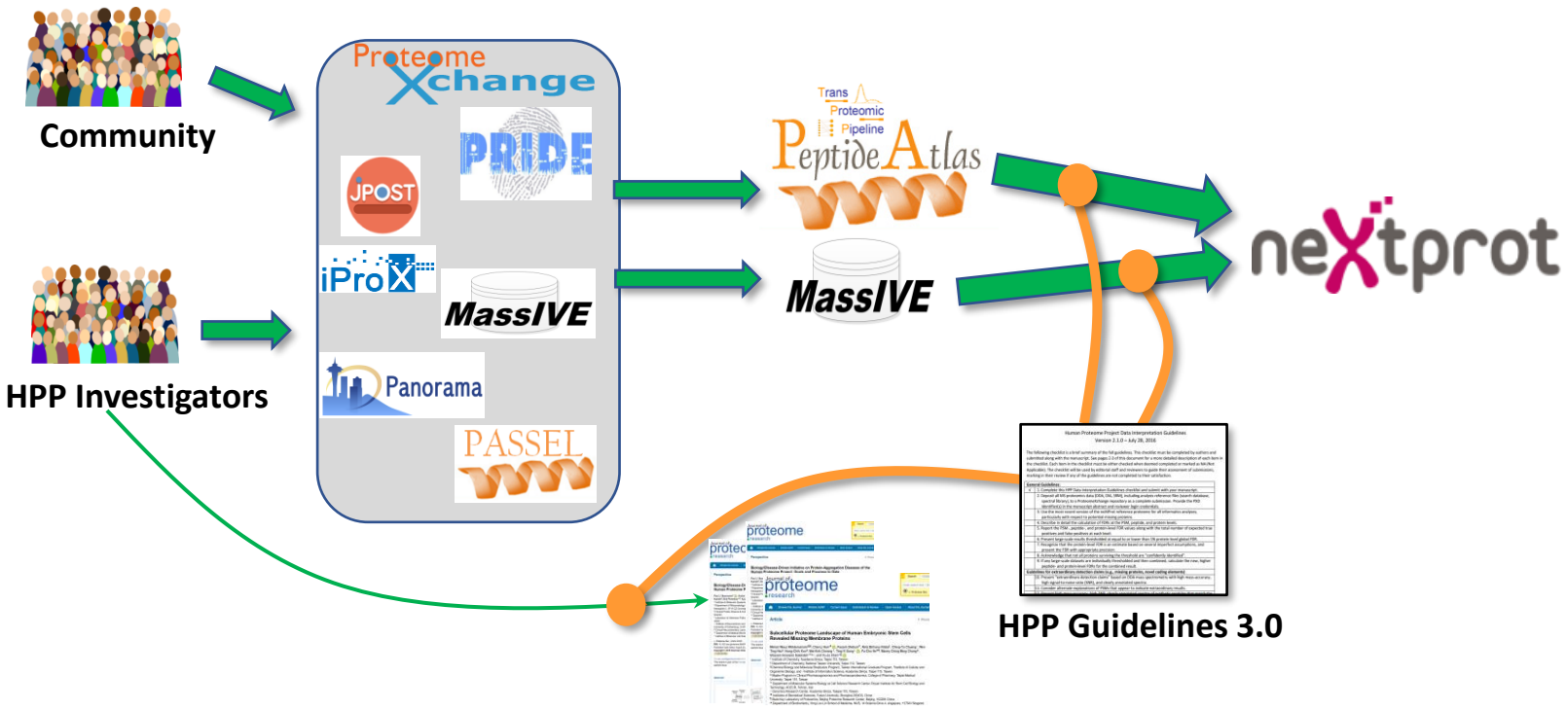
Monique Zahn-Zabal[1], Pierre-André Michel[1], Alain Gateau[1], Frédéric Nikitin[1], Mathieu Schaeffer[1,2], Estelle Audot[1], Pascale Gaudet[1], Paula D. Duek[1], Daniel Teixeira[1], Valentine Rech de Laval[1,2,3], Kasun Samarasinghe[1,2], Amos Bairoch[1,2] and Lydie Lane[1,2,*]

# HPP Goal 1: validate the existence of 1 protein product per gene

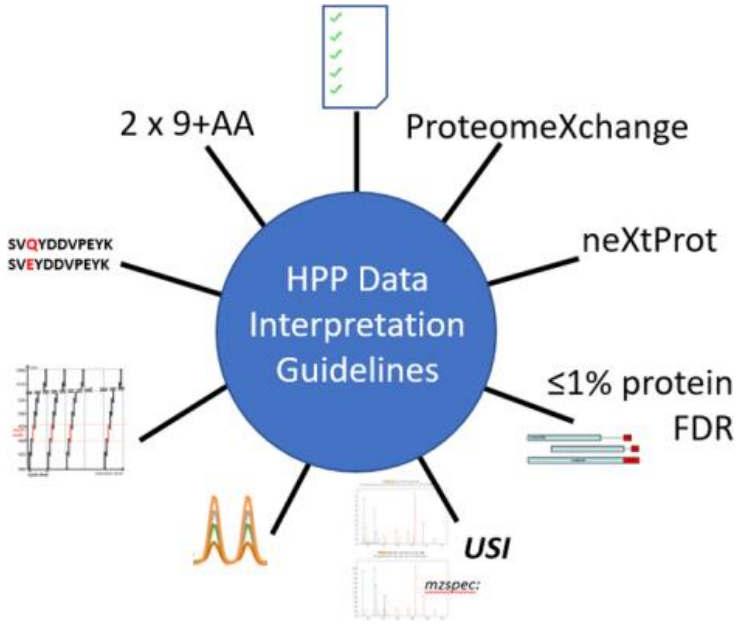
| Exptal. Evidence | PE value |
|---|---|
| Protein | PE1 |
| Transcript | PE2 |
| Non-human species | PE3 |
| Predicted gene model | PE4 |
| Uncertain | PE5 |

Mass spectrometry data

*"Missing proteins"*



# Guidelines for Mass Spectrometry Data Interpretation



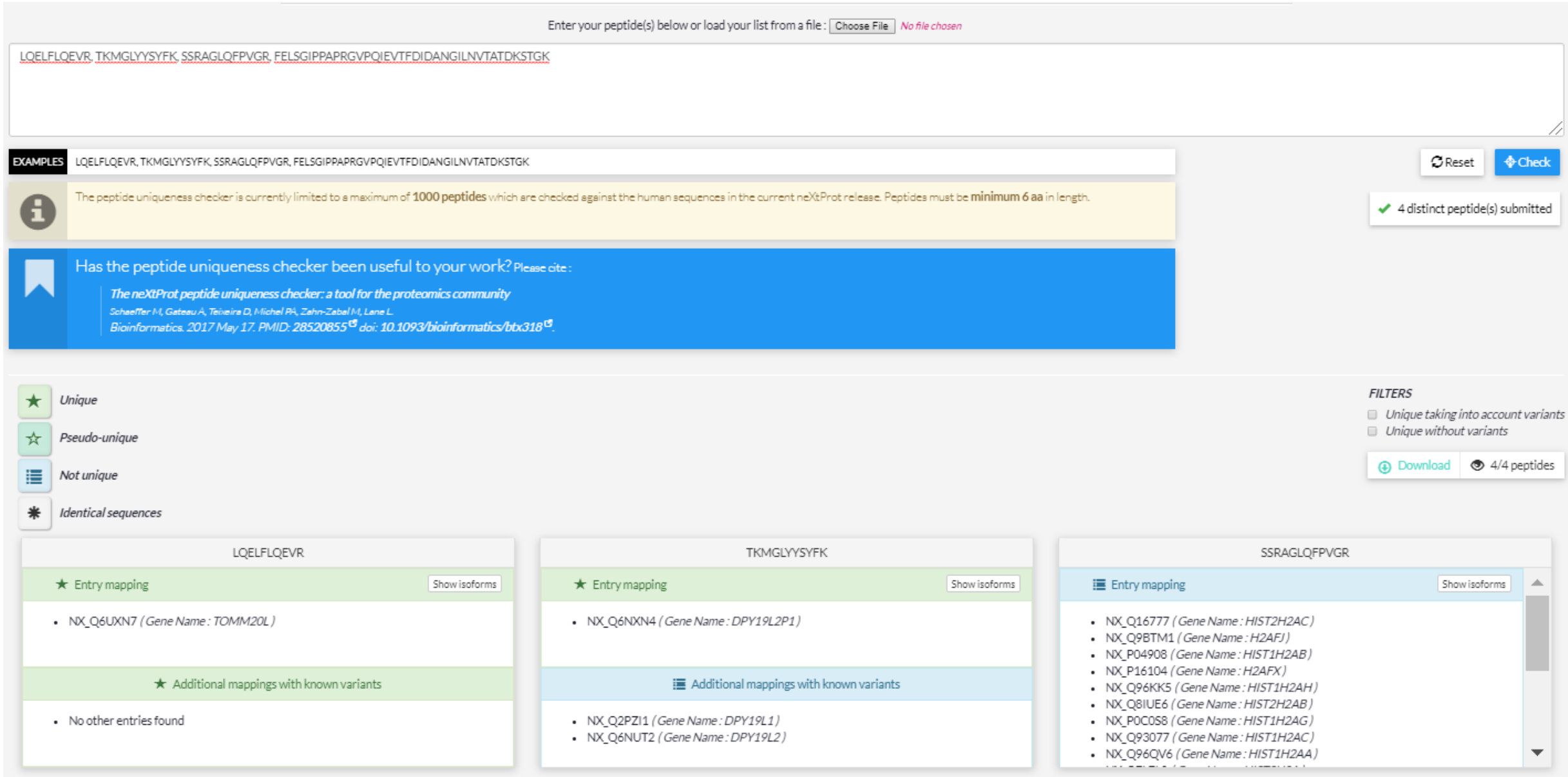

Deutsch et al, J. Proteome Res. 2019, 18, 12, 4108–4116

Information (p12L20 for page 12 line 20, S12 for supplementary table 2, etc.)

| | | |
|---|---|---|
| **General guidelines for all manuscripts:** | | |
| √ | Loc | 1. Complete this HPP MS Data Interpretation Guidelines checklist and submit with your manuscript. |
| 2. Data deposition guidelines | | |
| | | 2a. Deposit all MS proteomics data to a ProteomeXchange repository as a "complete" submission. |
| | | 2b. Include analysis reference files (search database, spectral library, transition list, etc.) in submission. |
| | | 2c. Provide the PXD identifier(s) in the manuscript abstract. |
| | | 2d. Provide the reviewer login credentials if the dataset is not yet public. |
| | | 3. Use the most recent version of the neXtProt reference proteome for all informatics analyses, particularly with respect to new PE1 protein detection claims. |
| 4. FDR-related guidelines | | |
| | | 4a. Describe in detail the calculation of FDRs at the PSM, peptide, and protein levels. |
| | | 4b. Report the PSM-, peptide-, and protein-level FDR values along with the total number of expected false positives at each level, using precision appropriate to the uncertainty in computed FDR. |
| | | 4c. Present large-scale results thresholded at equal to or lower than 1% protein-level global FDR. |
| | | 4d. If any large-scale datasets are individually thresholded and then combined, calculate the new, higher peptide- and protein-level FDRs for the combined result. |
| **Guidelines for claims of new PE1 protein detections (i.e., presenting evidence to categorize a protein to PE1)** | | |
| | | 5a. If using DDA mass spectrometry for such claims, present high mass-accuracy, high signal-to-noise ratio (SNR), and clearly annotated spectra. Scrutinize spectra for missing and extra peaks. |
| | | 5b. Present high mass-accuracy, high-SNR, clearly annotated spectra of synthetic peptides that match the spectra supporting the claims. Peptides from recombinant proteins are acceptable synthetics. |
| | | 5c. Provide Universal Spectrum Identifiers (USIs) for all natural and synthetic peptide spectra that support such claims, ideally as a supplementary data table. |
| | | 6. If using SRM verification for such claims, present target traces alongside synthetic heavy-labeled peptide traces, demonstrating co-elution and closely matching fragment mass intensity patterns. |
| | | 7. If using DIA MS, then, if the data are analyzed with XICs, apply the above SRM guidelines (6); if the data are analyzed by extracting deconvoluted spectra, apply the above DDA guidelines 5a-5c. |
| | | 8. Even when very high confidence peptide identifications are demonstrated, consider alternate mappings of the peptide to proteins other than the claimed one. Consider isobaric sequence/mass modification variants, all known SAAVs, and unreported SAAVs. |
| | | 9. Support such claims by two or more distinct uniquely-mapping, non-nested peptide sequences of length ≥9 amino acids with the above evidence in the same paper. When 2 peptides overlap, the total extent must be ≥18 amino acids. When weaker evidence is offered for such a claim, justify that other peptides cannot be expected by any common digestion proteases. When 2 or more proteins are *exactly* sequence identical (irrespective of SAAVs), peptides are considered uniquely mapping if they map only to the group, and such proteins will have the same PE level and be counted. |

Author comments (use this space and extra pages to explain any nonadherence [NA/NC] in the above checklist):

# The neXtProt peptide uniqueness checker



# Validation of “missing proteins” 2011-2021

nature COMMUNICATIONS

PERSPECTIVE

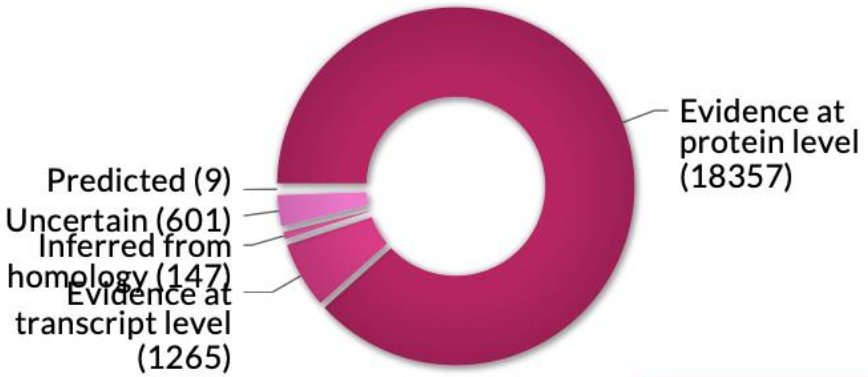
https://doi.org/10.1038/s41467-020-19045-9 OPEN

A high-stringency blueprint of the human proteome

Subash Adhikari et al.#

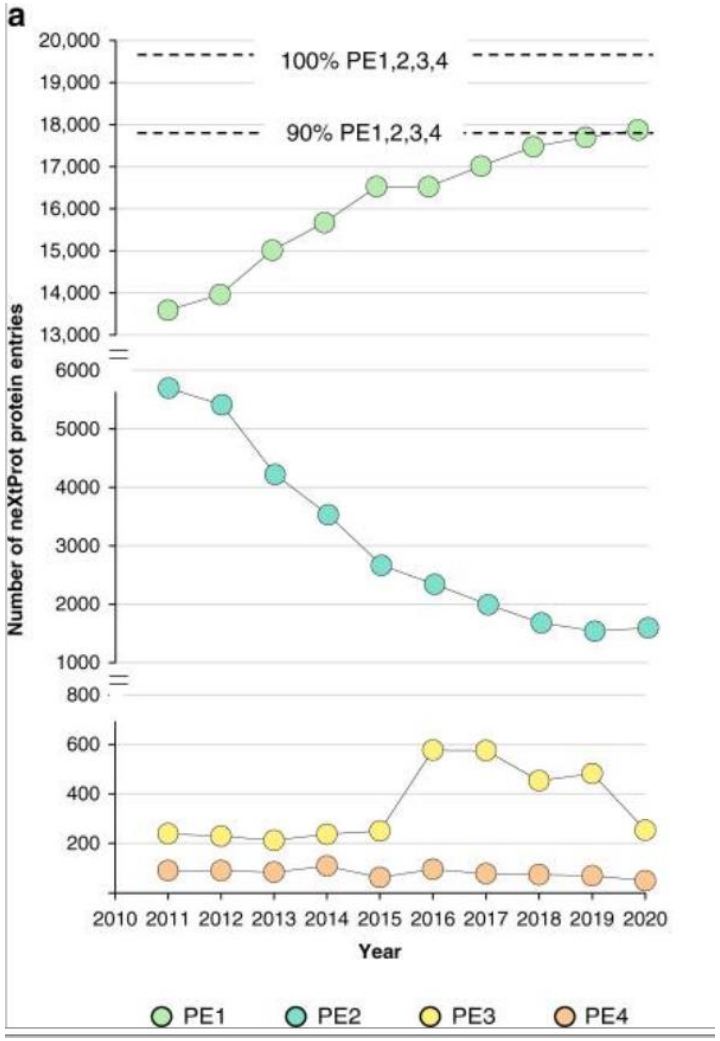
Release 2021-02-15

Protein existence in neXtProt




Release statistics

90.4% complete high-stringency human proteome



# Display of proteomics information in neXtProt

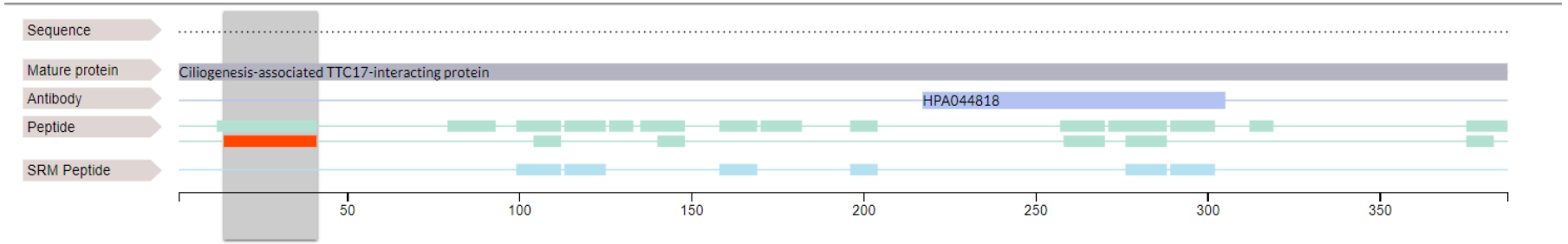
- ✓ ***Which (unique) MS/MS peptides from my protein have been found, and where ?***
- ✓ ***What peptides can I use for targeted proteomics studies?***

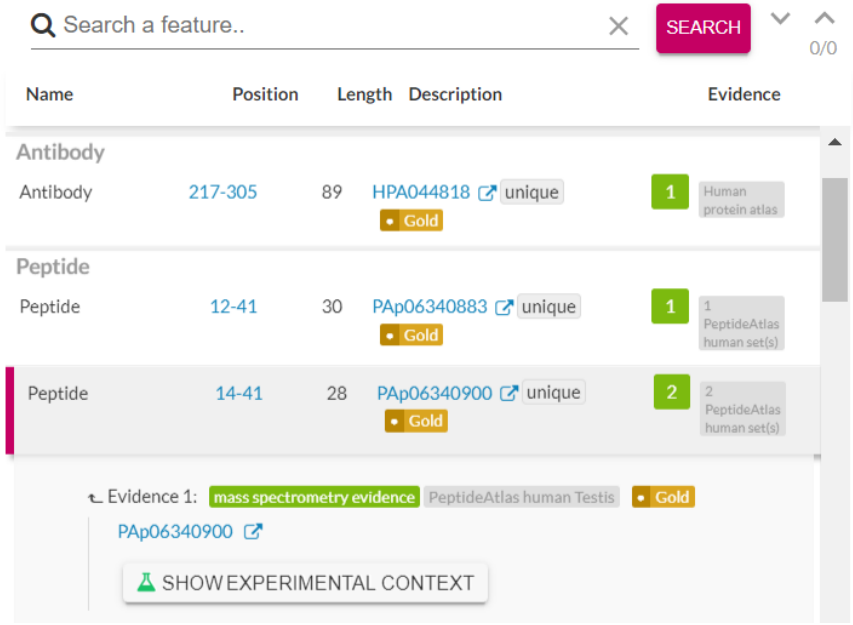
THE HUMAN PROTEIN ATLAS

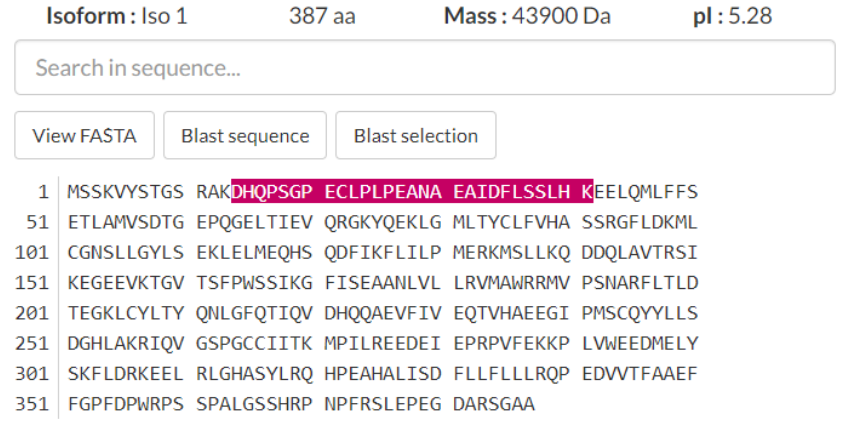
PeptideAtlas

SRM Atlas

MassIVE



# Are there still missing proteins to validate?

1343 proteins are still “missing” (PE2-4) in neXtProt

- Difficult to detect with conventional mass spectrometry?
  - Low abundant proteins, or with a expression pattern restricted in time and/or space (incl. 392 olfactory/taste receptors)
  - Highly hydrophobic or hard to extract from cells?
  - Digestion with usual enzymes not leading to 2 unique peptides?
- Wrong predictions?

## Potential solutions

- Use different samples (selected using RNAseq data)
- Improve sample preparation
- Design targeted assays, use affinity-enrichment approaches
- Improve the sensitivity of the detection
- Use different digestion enzymes
- Adapt validation criteria?

# The neXtProt protein digestion tool

Enter neXtProt isoform accession number, i.e NX_P50222-1

NX_Q9HC47-1

Max miscleavages: 0 | Min peptide length: 9 | Max peptide length: 40 | RESET | DIGEST

**DIGESTED PEPTIDES FOR EACH PROTEASE :**

Copy | CSV | Excel | Print

Showing 1 to 27 of 27 entries  1 row selected  Search:

| Protease name | Peptide count | Unique peptide count |
|---|---|---|
| GLU_C_BICARBONATE | 0 | 0 |
| LYS_C | 0 | 0 |
| PEPSIN_PH_1_3 | 0 | 0 |
| PEPSIN_PH_GT_2 | 0 | 0 |
| PROTEINASE_K | 0 | 0 |
| THERMOLYSIN | 0 | 0 |
| TRYPSIN | 0 | 0 |
| ASP_N | 2 | 2 |
| CHYMOTRYPSIN_HIGH_SPEC | 2 | 2 |
| GLU_C_PHOSPHATE | 2 | 2 |

Select a protease: GLU_C_PHOSPHATE | GET PEPTIDES

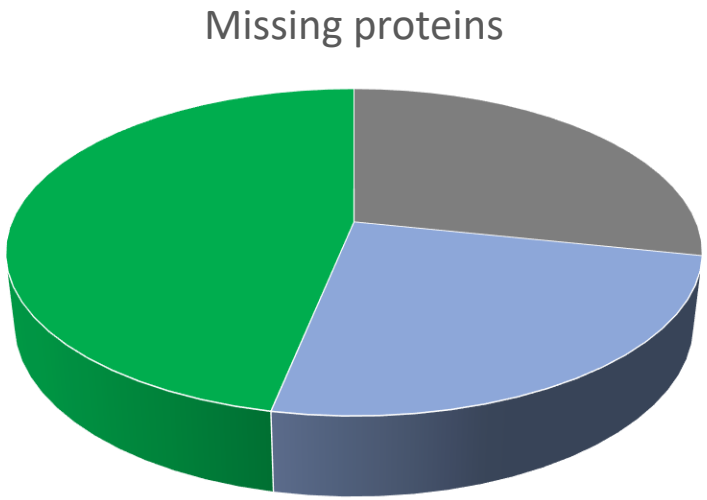
Copy | CSV | Excel | Print

Showing 1 to 2 of 2 entries  Search:

| Peptide sequence | Length | Missed cleavages | Position | Unique without variants | Natural in neXtProt | Synthetic in neXtProt |
|---|---|---|---|---|---|---|
| Filter | 32 37 | 0 1 | | Filter | Filter | Filter |
| MFVIISLHNCVVISFVLFLFGGNNFIQNFYLPQNYID | 37 | 0 | 1-37 | Yes | No | No |
| QFLLTSFPTFTSVGVLIVLVLCSAFLLLWQGE | 32 | 0 | 38-69 | Yes | No | No |

Show 10 entries  Previous 1 Next

**Example: TMEM213**
**(2 tryptic peptides)**




## TMEM213 → Transmembrane protein 213

Gene name : TMEM213

Entry whose protein(s) existence is based on evidence at transcript level.

*Annotations in this section apply to all the isoforms if not specified otherwise.*

EXPRESSION

Expression in kidney and salivary gland. • Gold

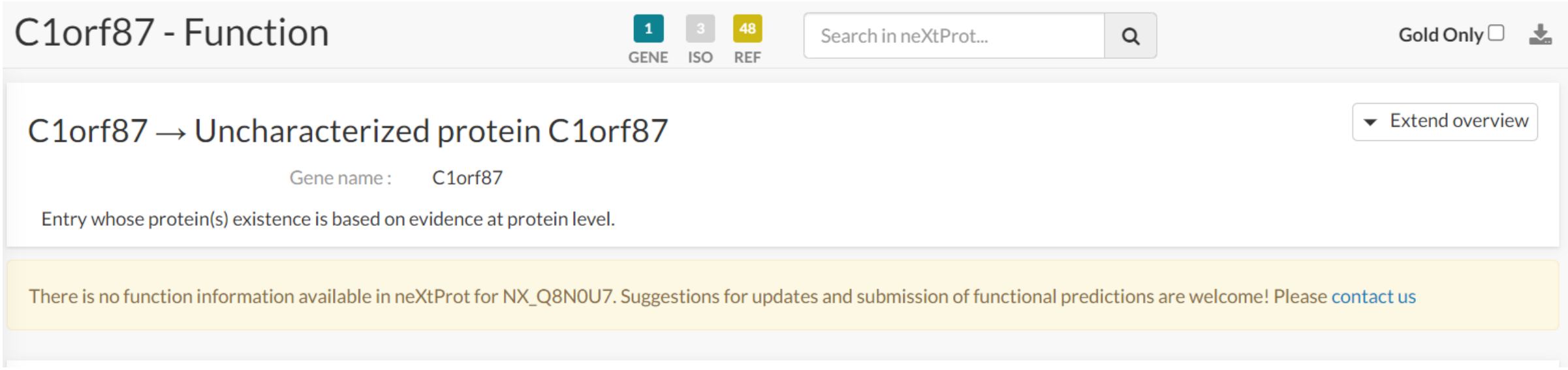
RNA tissue specificity: Group enriched (kidney, salivary gland). RNA tissue distribution: Detected in some. • Gold

Single cell type specificity: Cell type enhanced (Collecting duct cells, Distal tubular cells, Melanocytes). • Gold

# HPP Goal 2: “A Function for Every Protein”

- Most research on human genes only concentrates on approximately 2,000 of the 19,000 genes of the human genome

- In neXtProt 1191 proteins are validated at protein level but do not have any predicted or experimentally validated function



# The CALIPHO laboratory (2009-2018)

Classical approach (hypothesis-driven, single gene)

With a focus on:

- Atypical (small) mitochondrial proteins
- Proteins with a typical “ciliary” profile
- Enzymatic pathway holes

Function prediction by data analysis

↓

Gene knock-out in human cell lines or model organisms (zebrafish)

↓

Assay of the predicted function

1 PhD or 1 post doc for 3-4 years for 1 gene

# The CALIPHO laboratory (2009-2018)

PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

## Functional Identification of APIP as Human mtnB, a Key Enzyme in the Methionine Salvage Pathway

Camille Mary, Paula Duek, Lisa Salleron, Petra Tienz, Dirk Bumann, Amos Bairoch, Lydie Lane

Published: December 28, 2012 • https://doi.org/10.1371/journal.pone.0052877

MCB
Journals.ASM.org

## C11orf83, a Mitochondrial Cardiolipin-Binding Protein Involved in $bc_1$ Complex Assembly and Supercomplex Stabilization

Marjorie Desmurs,[a] Michelangelo Foti,[b] Etienne Raemy,[c] Frédéric Maxime Vaz,[d] Jean-Claude Martinou,[c] Amos Bairoch,[a,e] Lydie Lane[a,e]

frontiers in Cellular Neuroscience

ORIGINAL RESEARCH
published: 16 March 2021
doi: 10.3389/fncel.2021.653075

## C21orf91 Regulates Oligodendroglial Precursor Cell Fate—A Switch in the Glial Lineage?

*Laura Reiche[1], Peter Göttle[1], Lydie Lane[2,3], Paula Duek[2,3], Mina Park[1], Kasum Azim[1], Jana Schütte[1], Anastasia Manousi[1], Jessica Schira-Heinen[1] and Patrick Küry[1]**

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbamcr

## DERA is the human deoxyribose phosphate aldolase and is involved in stress response

Lisa Salleron [a,*], Giovanni Magistrelli [b], Camille Mary [a], Nicolas Fischer [b], Amos Bairoch [a,c], Lydie Lane [a,c,**]

[a] Department of Human Protein Sciences, Faculty of Medicine, University of Geneva, Geneva, Switzerland
[b] Novimmune SA, Plan-les-Ouates, Switzerland
[c] CALIPHO GroupSIB-Swiss Institute of Bioinformatics, Geneva, Switzerland

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-14999-2 OPEN

## Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly

Shan Zhang[1], Boris Reljić[2,12], Chao Liang[1,12], Baptiste Kerouanton[1,12], Joel Celio Francisco[3], Jih Hou Peh[1], Camille Mary[4], Narendra Suhas Jagannathan[5], Volodimir Olexiouk[6], Claire Tang[1], Gio Fidelito[1], Srikanth Nama[7], Ruey-Kuang Cheng[8], Caroline Lei Wee[9], Loo Chien Wang[9], Paula Duek Roggli[10], Prabha Sampath[11], Lydie Lane[4], Enrico Petretto[5], Radoslaw M. Sobota[9], Suresh Jesuthasan[8,9], Lisa Tucker-Kellogg[3,5], Bruno Reversade[7,9], Gerben Menschaert[6], Lei Sun[1], David A. Stroud[2] & Lena Ho[1,7]✉

PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

## C2orf62 and TTC17 Are Involved in Actin Organization and Ciliogenesis in Zebrafish and Human

Franck Bontems, Richard J. Fish, Irene Borlat, Frédérique Lembo, Sophie Chocu, Frédéric Chalmel, Jean-Paul Borg, Charles Pineau, Marguerite Neerman-Arbez, Amos Bairoch, Lydie Lane

Published: January 27, 2014 • https://doi.org/10.1371/journal.pone.0086476

Lessons learned:

- Very risky
- Difficult to fund
- Lack of available reagents

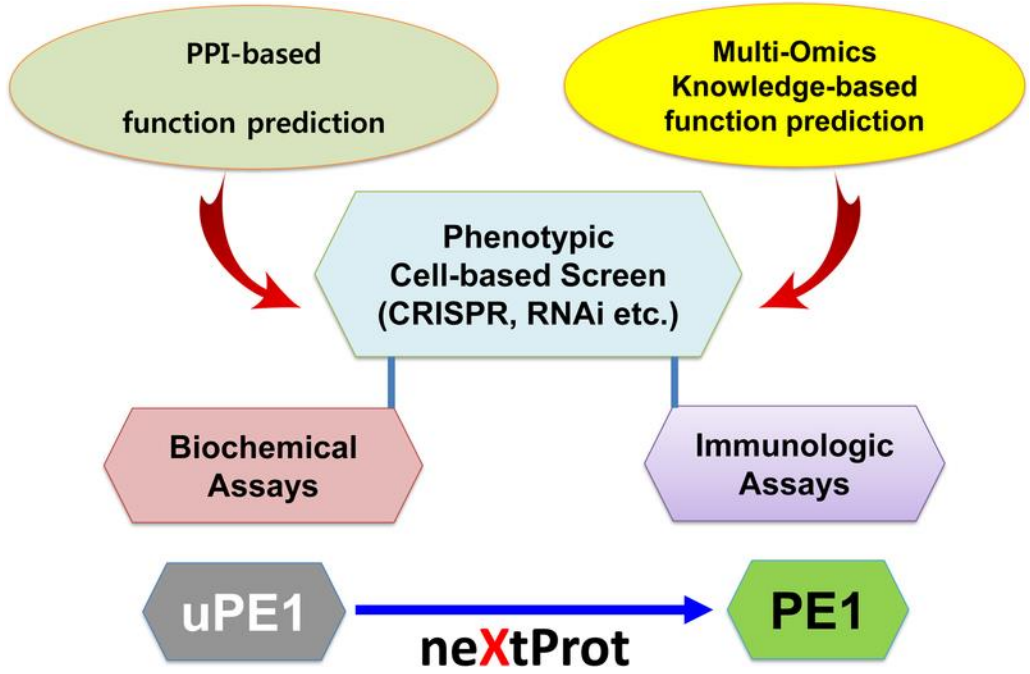
# The HPP next-CP50 initiative (2018-2021)

*On March 1, 2018, HUPO C-HPP announced launching the neXt-CP50 where CP stands for "characterization of proteins". This pilot project aims to* ***characterize the function of 50 identified proteins with unknown function*** *during ~3 years (2018-2021). This challenge is to test the feasibility of the functional characterization of large numbers of dark proteins*

Journal of proteome research
Perspective
Cite This: J. Proteome Res. 2018, 17, 4042−4050
pubs.acs.org/jpr

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

Young-Ki Paik, Lydie Lane, Takeshi Kawamura, Yu-Ju Chen, Je-Yoel Cho, Joshua LaBaer, Jong Shin Yoo, Gilberto Domont, Fernando Corrales, Gilbert S. Omenn, Alexander Archakov, Sergio Encarnación-Guevara, Siqi Lui, Ghasem Hosseini Salekdeh, Jin-Young Cho, Chae-Yeon Kim, and Christopher M. Overall




Comparable issues:

- Very risky
- Difficult to fund
- Lack of available reagents

# The Human Proteome Grand Project 2022-2032

**aims to:**

- Build upon the large resources constructed through the significant efforts of the HPP
- Understand the Proteome in the context of Networks of Networks
- Partner with complementary consortia to expand the reach of the HPP and increase its relevance
- Develop forward thinking strategies to provide functional information of each protein
- Be a focal point for translational researchers for mechanisms, diagnostics, and therapeutics to improve human health

**is:**

- Open to all interested groups or individuals
- Agnostic to disease and biological context

- **will:**
  Utilize resources already familiar to the research groups
- Advance the knowledge of systems being worked on through the access to resources to perturb these systems

# Network-based exploratory approach:

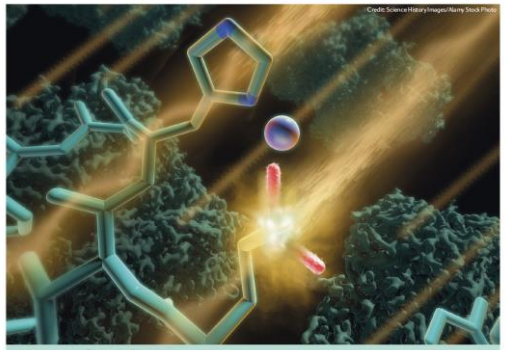
Chemical Probes.org

## WHY TARGET 2035

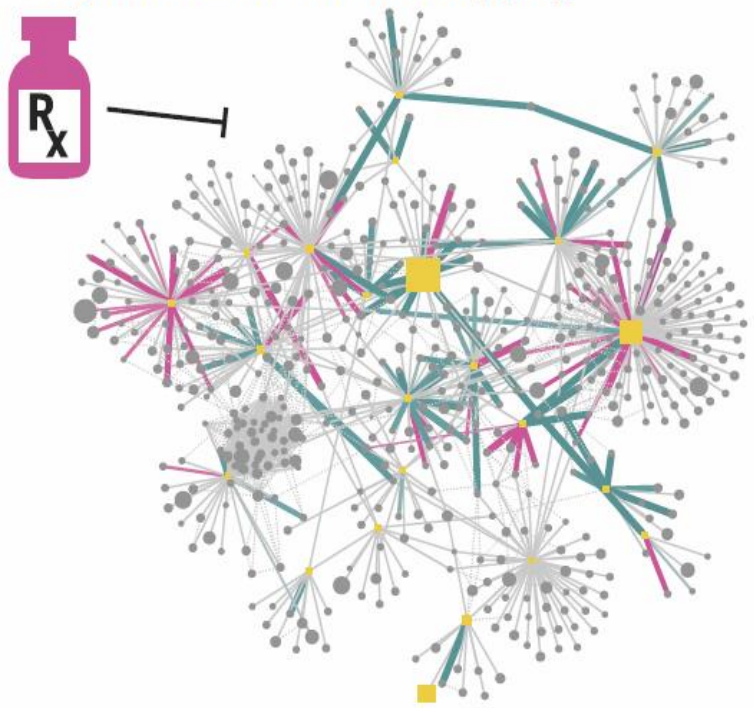
Almost 20 years after the human genome was sequenced, many of the genes linked to disease phenotypes or those associated with specific disease traits by genome-wide association studies remain severely understudied: the so-called ‘dark genome’.

Generating chemical and biological reagents for these understudied proteins will enable research, lead to the validation of novel therapeutic targets, and the discovery of better medicines.

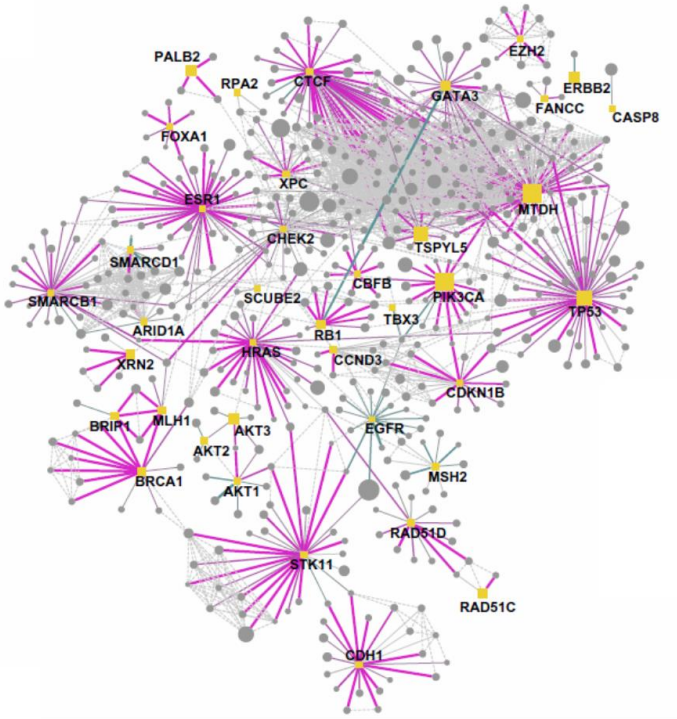
NEWS & ANALYSIS



A probe for every protein




**Comparative mapping**




**Define interactors and perturbers (variations amongst individuals, drugs…) to understand impact and function**

# The unknown unknowns in the human proteome

## Known unknowns

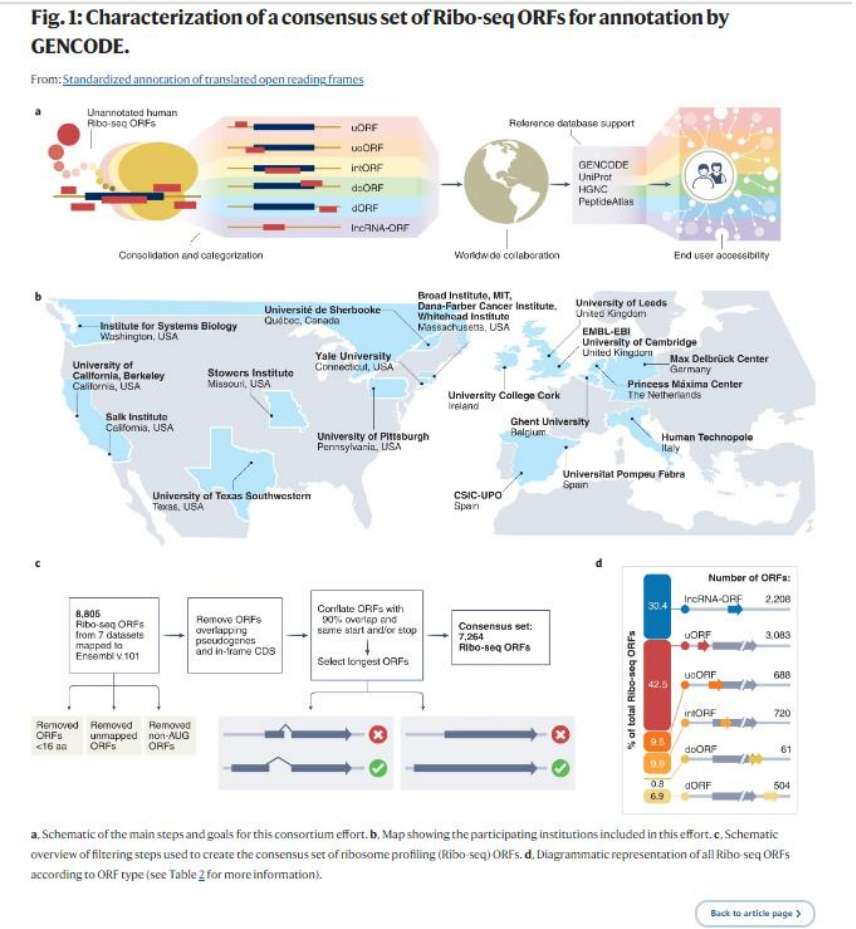
- *Expected proteins*
- *Uncharacterized proteins identified in human samples*
- *Predicted  PTMs*
- *Variants of unknown significance*

## Unknown unknowns

- *Are there coding elements in the genome outside the 1.2% predicted (e.g. smORFs and lncRNAs)?*
- *Novel functions to discover even for well-known proteins (one protein can have multiple functions!)*
- *Novel PTMs to discover*
- *…?*

# How to track the unknown unknowns ?

- Validation of novel coding elements will require new reference databases and new guidelines

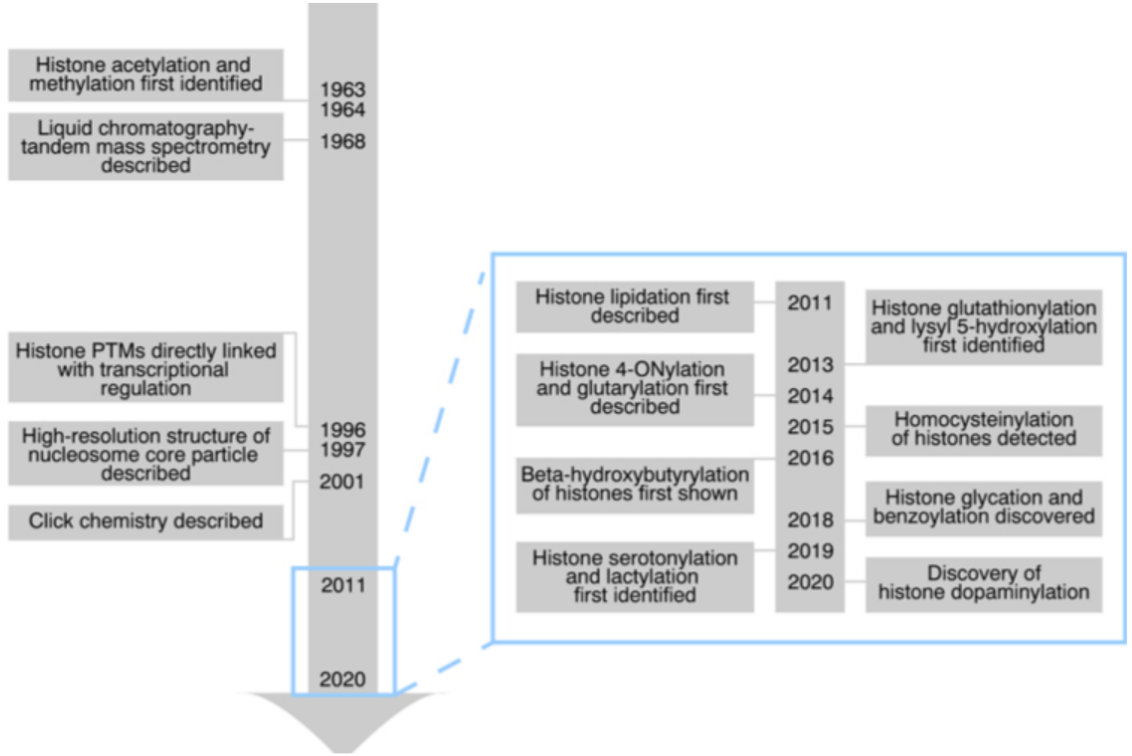


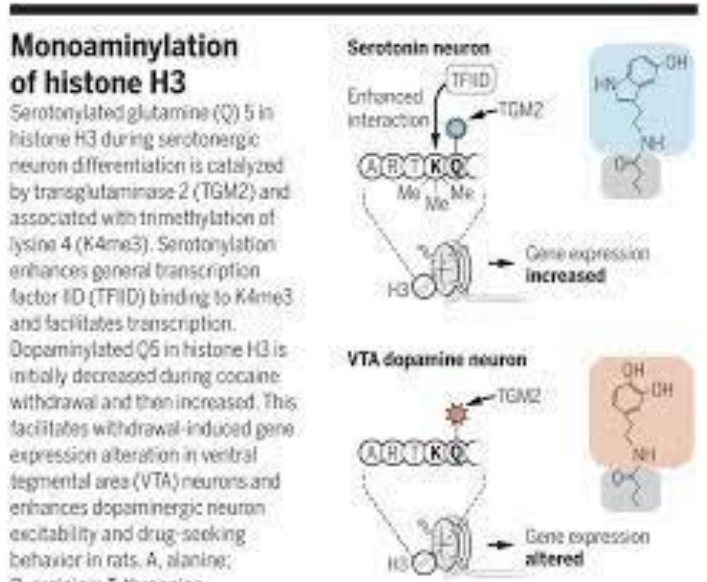

Example: A catalog of 7264 ORFs, some as short as 16 aa in length, with substantial evidence of ribosome translation activity based on Ribo-seq experiments currently analysed for MS evidence by PeptideAtlas

Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, Gonzalez JM, Magrane M, Martinez TF, Schulz JF, Yang YT, Albà MM, Aspden JL, Baranov PV, Bazzini AA, Bruford E, Martin MJ, Calviello L, Carvunis AR, Chen J, Couso JP, Deutsch EW, Flicek P, Frankish A, Gerstein M, Hubner N, Ingolia NT, Kellis M, Menschaert G, Moritz RL, Ohler U, Roucou X, Saghatelian A, Weissman JS, van Heesch S. Standardized annotation of translated open reading frames. Nat Biotechnol. 2022 Jul;40(7):994-999. doi: 10.1038/s41587-022-01369-0. PMID: 35831657.

# How to track the unknown unknowns ?

New PTMs and new functions are constantly discovered




**Selected Milestones in Histone Post-Translational Modification (PTM) Discovery.**

Chan JC, Maze I. Nothing Is Yet Set in (Hi)stone: Novel Post-Translational Modifications Regulating Chromatin Function. Trends in Biochemical Sciences. 2020 Oct;45(10):829-844. DOI: 10.1016/j.tibs.2020.05.009. PMID: 32498971; PMCID: PMC7502514.

# These newly discovered PTMs are now found in neXtProt

# Are there unknown knowns in the human proteome?

## Known knowns

- *Well characterized proteins*
- *Known PTMs and variants regulating function or localization of proteins*

## Unknown knowns

- *Unannotated characterization data*
- *Unpublished/hidden characterization data*
- *Data from large scale studies that could be combined into functional knowledge*
- *…*

# Unannotated and unpublished data

- Curators @neXtProt @UniProt @GOA @KEGG @Reactome monitor the literature and update proteins' functional information every day, but there might be some gaps/delays!

-> *Don't hesitate to send update requests if you spot missing annotations*

Unfortunately, negative results or results difficult to interpret are often unpublished

- In the last five years, we estimate that 4 - 5 papers describe newly characterized human proteins each month

-> *~20 years to be completed! How can this be accelerated?*

# Combining scattered data from large scale studies

Most proteins with no function annotated do have:

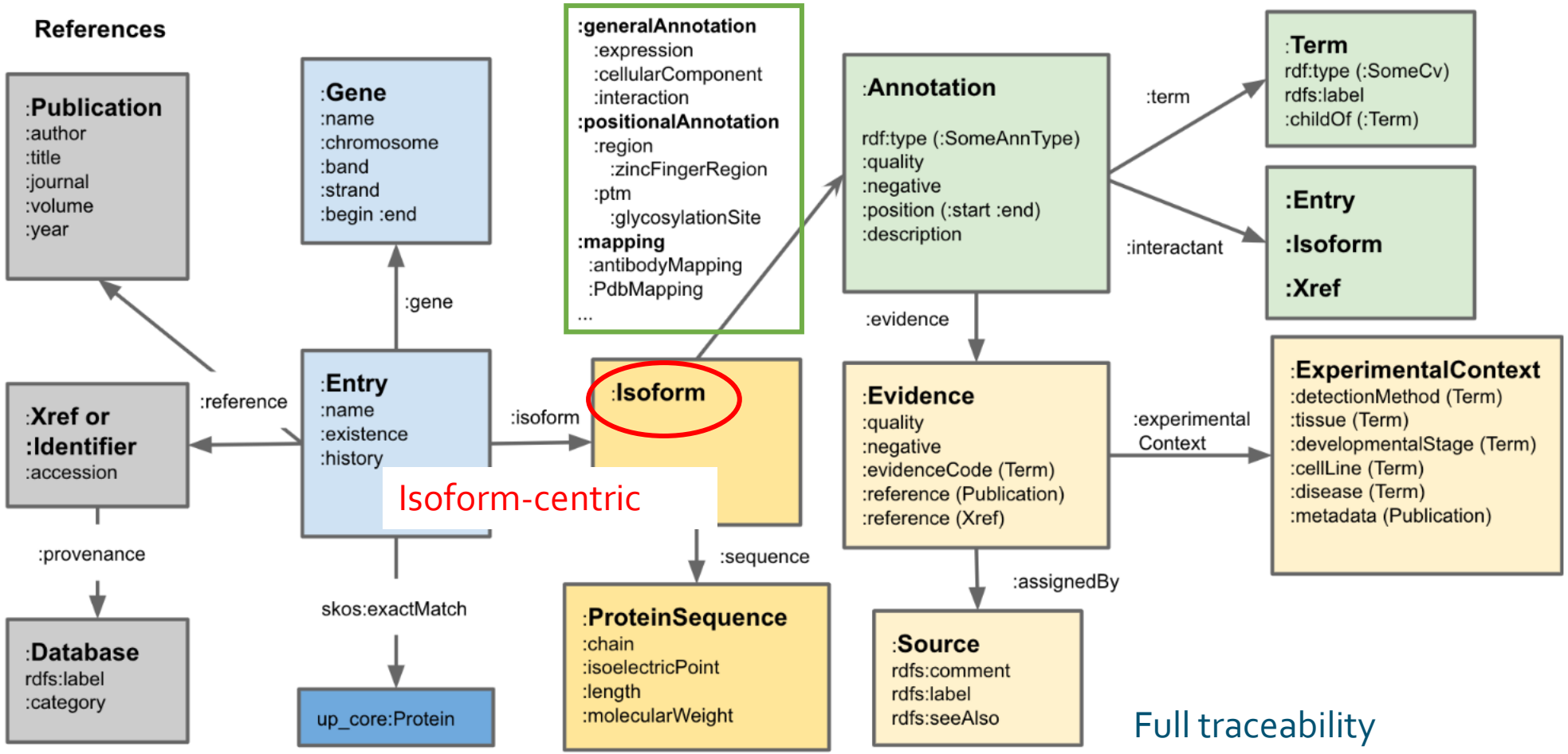
- Expression data
- Subcellular location information
- PPI data
- Orthologs in model organisms
- Predicted 3D structures

-> Can we retrieve this information from neXtProt, combine it, and deduce possible function(s)?

# www.nextprot.org RDF data model

A wide variety of experimental and predicted annotations

Extensive use of CV




Full traceability

# RDF data can be queried using SPARQL queries

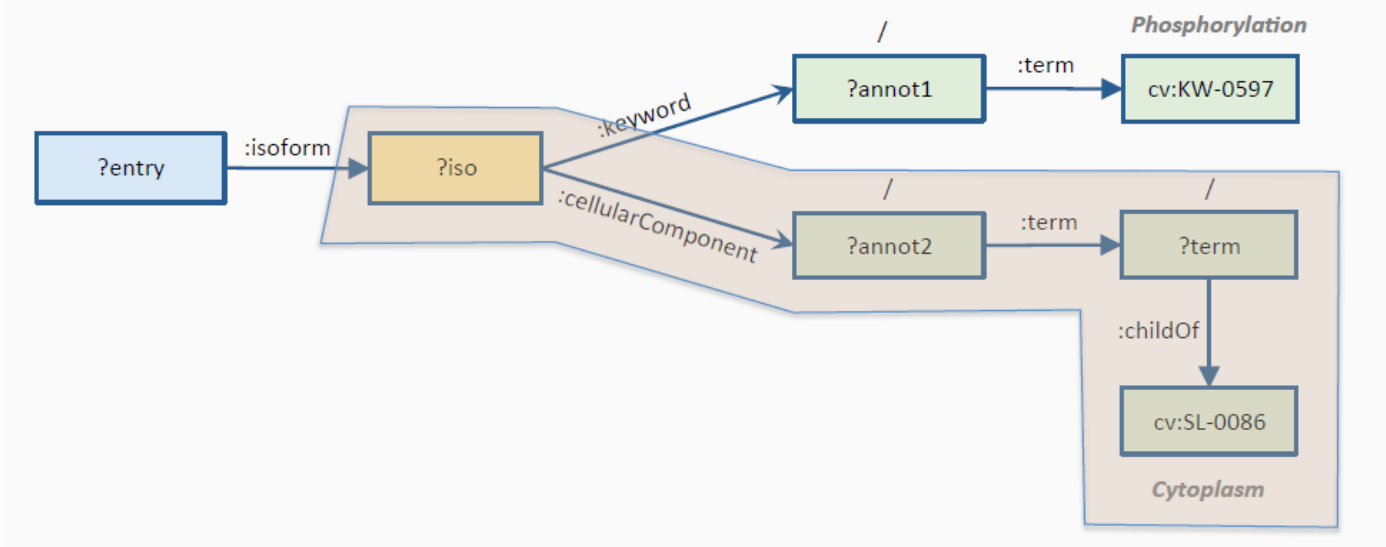
Simple search  Advanced search

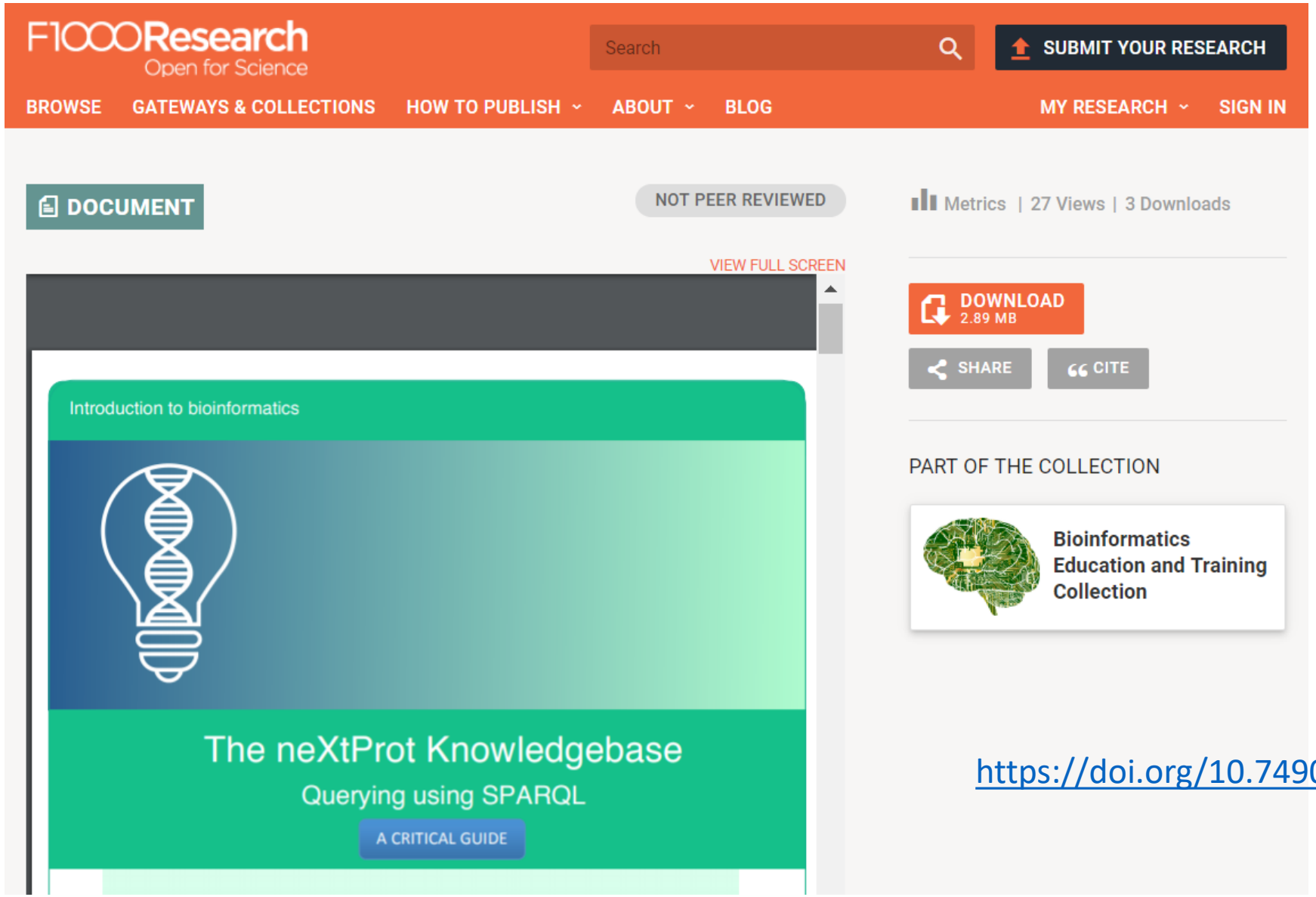
```
1 #Proteins phosphorylated and located in
2 #the cytoplasm
3
4 select distinct ?entry where {
5   ?entry :isoform ?iso.
6   ?iso :keyword / :term cv:KW-0597.
7   ?iso :cellularComponent /:term /:childOf cv:SL-0086.
8 }
```

Search

SPARQL queries consist of: 1) The query result variables – these will be shown when the query is executed; 2) the query pattern – query constructs defining the data selection, and generating any variables required; and 3) the optional query modifiers – used to modify the results.

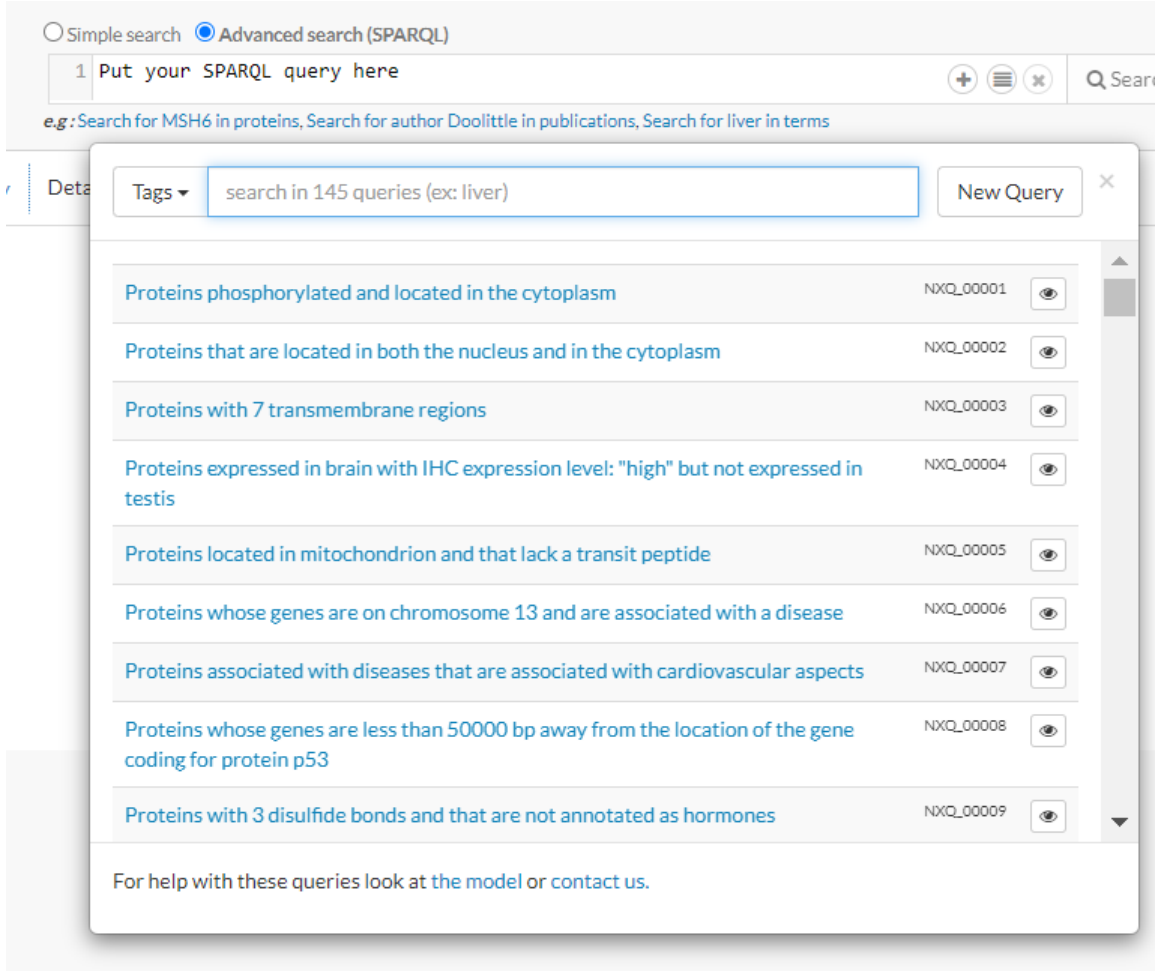
*SELECT ... query result variables ... WHERE {*
*... query pattern ...*
*}*
*... optional query modifiers ...*





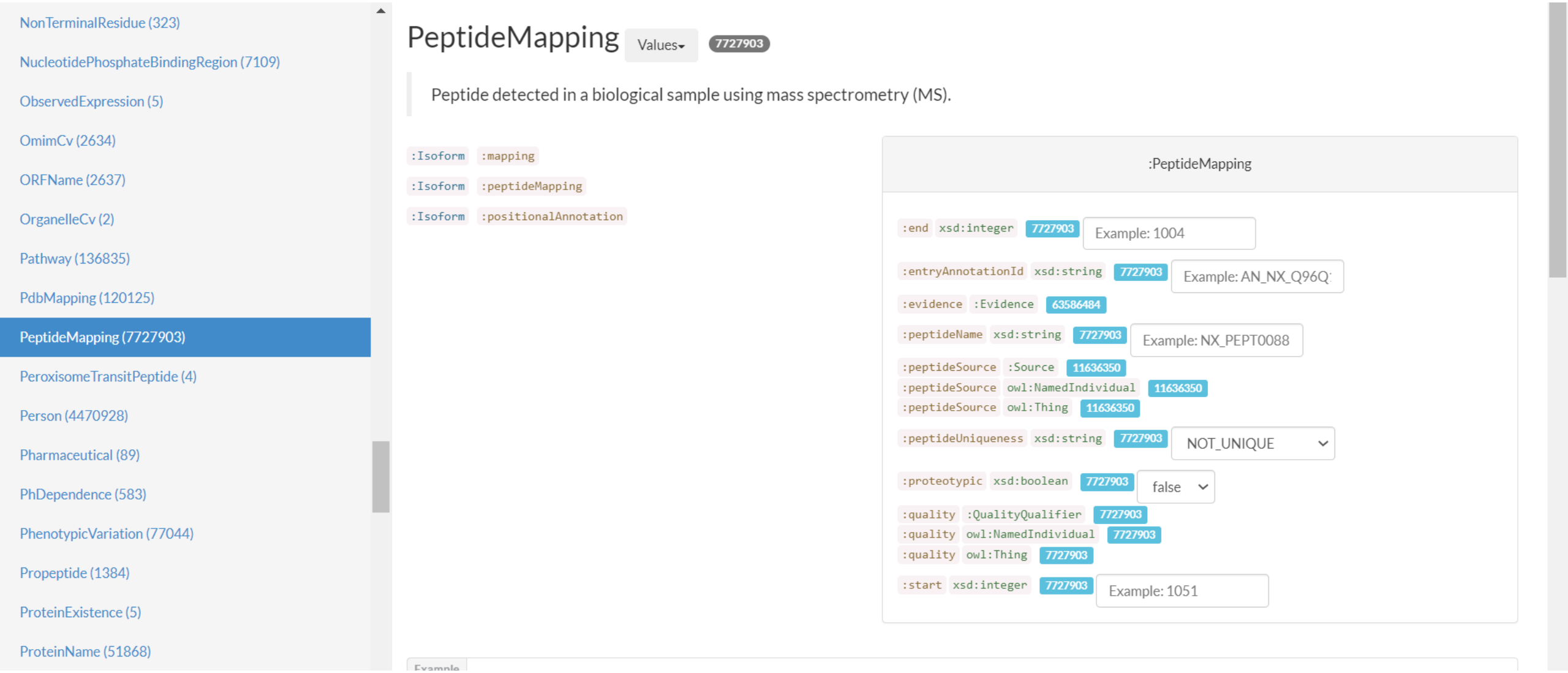

https://doi.org/10.7490/f1000research.1116829.1

# Almost 200 premade queries to explore the human proteome



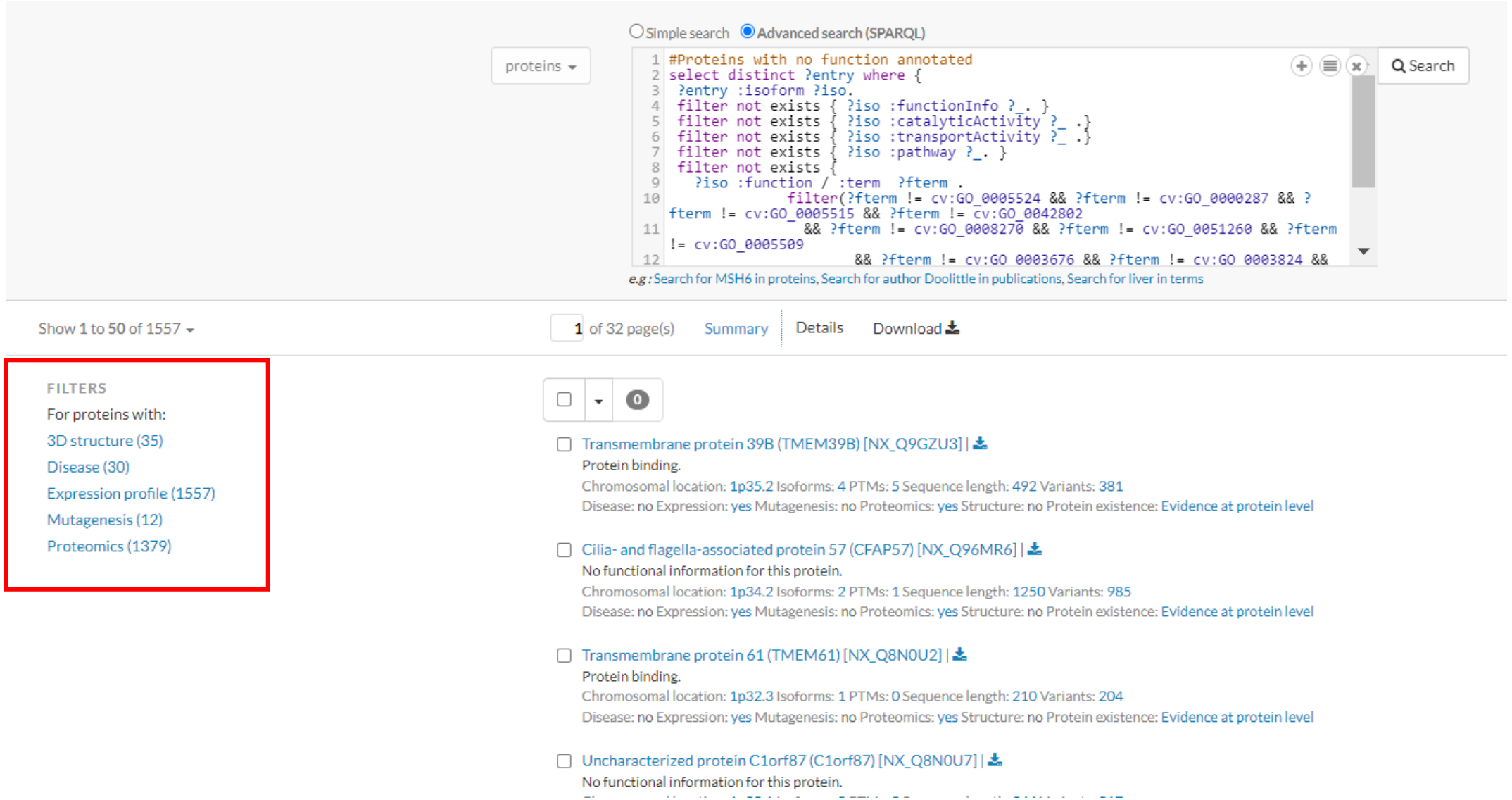

Missing proteins (NXQ_00204)
Proteins with no function (NXQ_00022)
Proteins with specific expression patterns (NXQ_00004)
Proteins associated with diseases (NXQ_00007)
Proteins with specific PTMs…

# A detailed help on neXtProt RDF entities is available

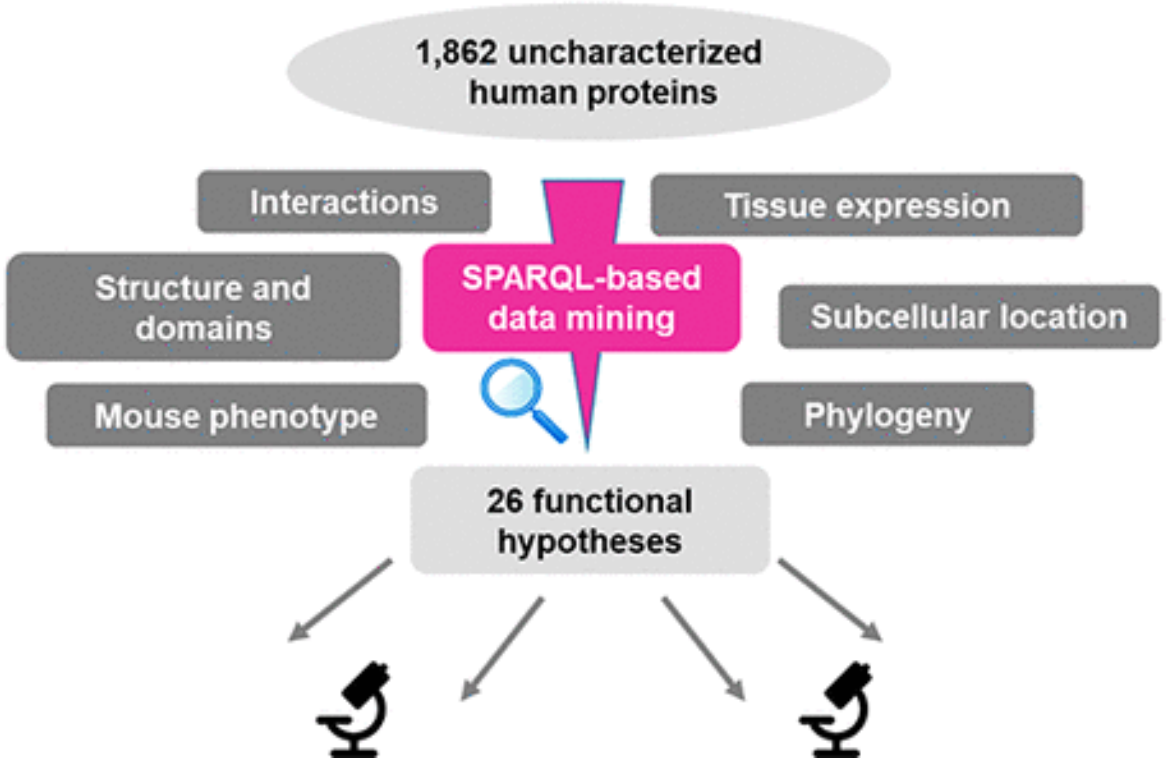


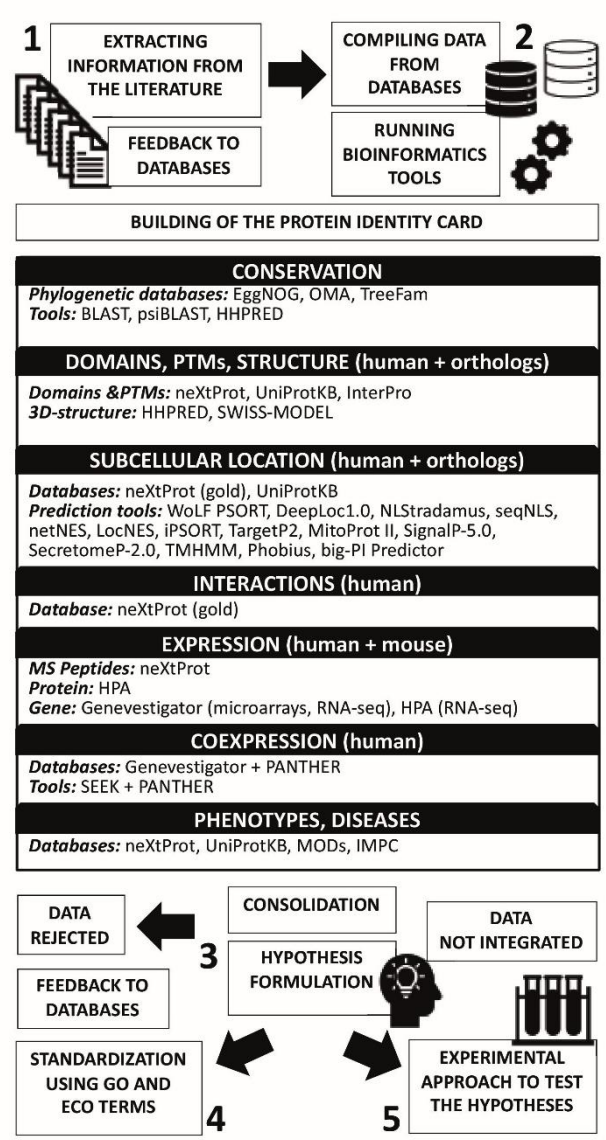

https://snorql.nextprot.org/help/doc/introduction

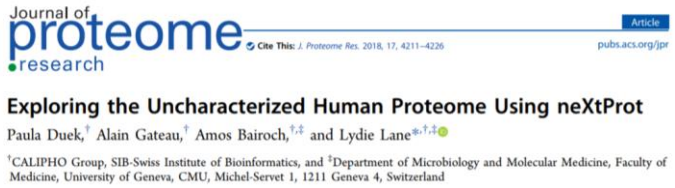
# Proteins with no function annotated (NXQ_00022)



# Manual datamining on a selection of uncharacterized proteins





Journal of proteome research

Article

Cite This: J. Proteome Res. 2018, 17, 4211−4226

pubs.acs.org/jpr

**Exploring the Uncharacterized Human Proteome Using neXtProt**

Paula Duek,† Alain Gateau,† Amos Bairoch,†,‡ and Lydie Lane*,†,‡

†CALIPHO Group, SIB-Swiss Institute of Bioinformatics, and ‡Department of Microbiology and Molecular Medicine, Faculty of Medicine, University of Geneva, CMU, Michel-Servet 1, 1211 Geneva 4, Switzerland

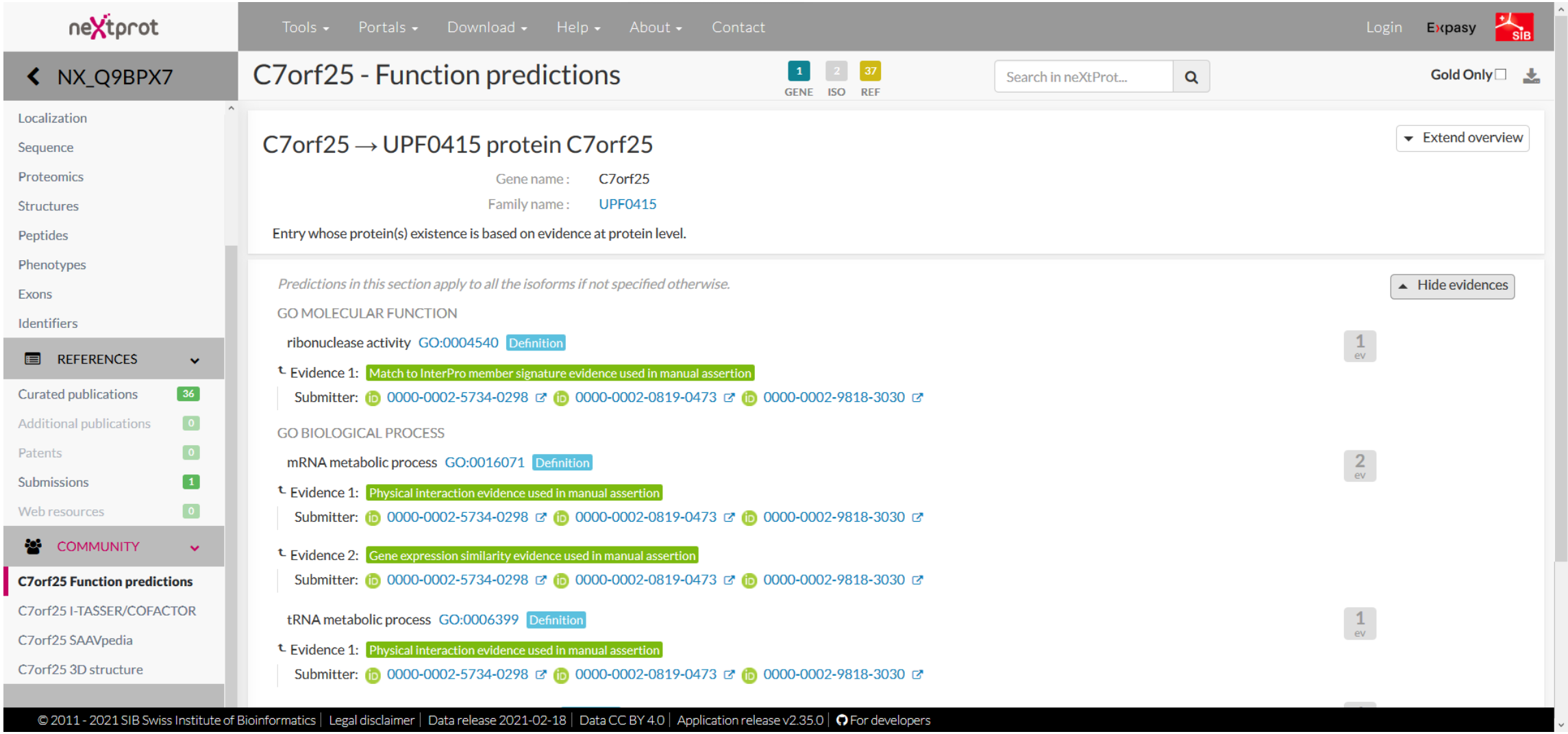

Database, 2021, 00(0), 1–13
DOI: https://doi.org/10.1093/database/baab046
Original article

DATABASE

**Functionathon: a manual data mining workflow to generate functional hypotheses for uncharacterized human proteins and its application by undergraduate students**

Paula Duek[1,2], Camille Mary[2], Monique Zahn-Zabal[1], Amos Bairoch[1,2] and Lydie Lane[1,2,*]

# Standardization of the hypotheses as GO/ECO terms and integration in the neXtProt function prediction pages



- We currently provide predictions for **204 uncharacterized proteins** in the neXtProt community pages

- Feel free to give us some feedback us and/or start experiments to test them!

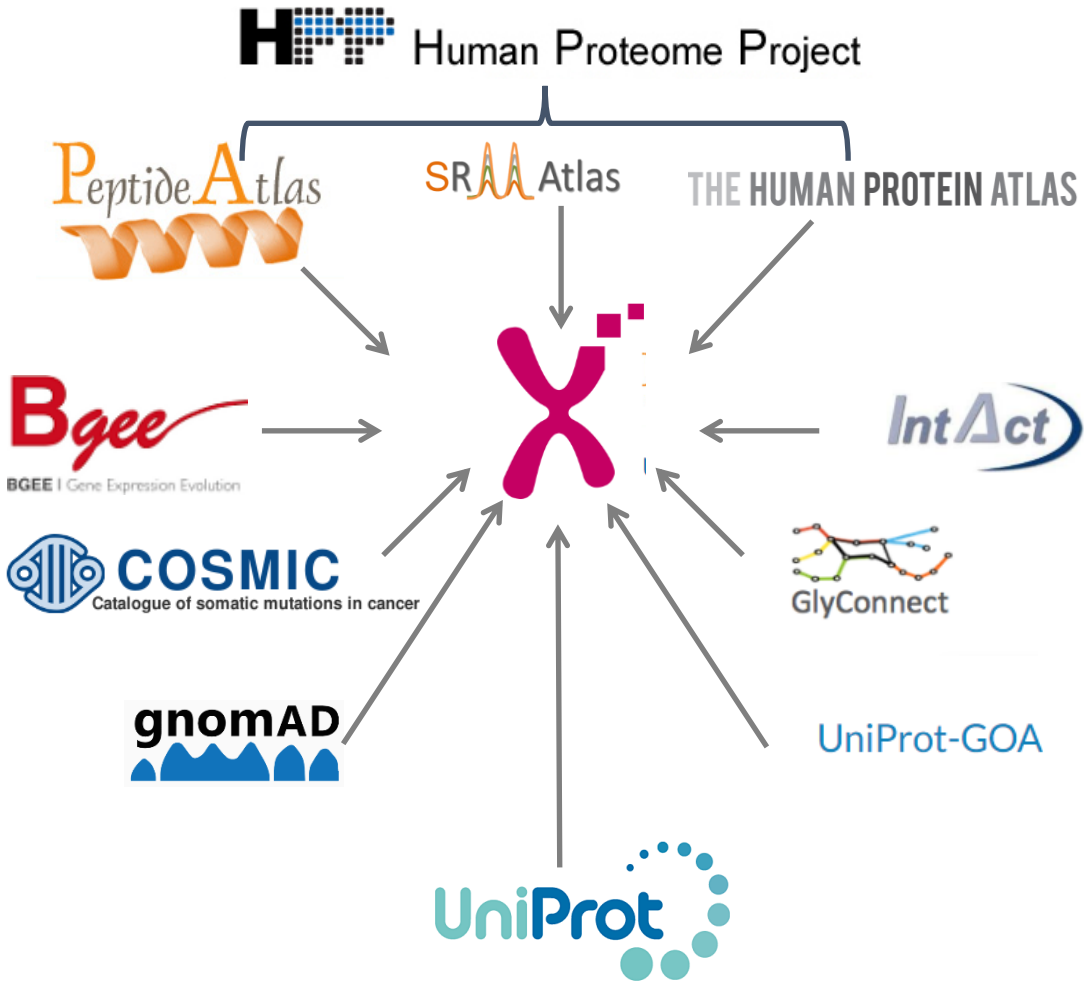
- Publish your own hypotheses (derived from manual or automated tools)

Please use the **Contact** option in neXtProt & provide:

1. Your **ORCID** *but you can choose not to have it displayed*
2. Protein **entry AC** *from query NXQ_00022*
3. **GO term** *describing molecular function / biological process*
4. **ECO term** *describing the type of supporting evidence*
5. **PubMed ID** *for publication with the function prediction*


https://www.nextprot.org/about/functional-proteome-project

# www.nextprot.org does not contain everything….




- ✓ MS-data
- ✓ Antibody-based data
- ✓ RNA-seq data
- ✓ PTMs
- ✓ Variants (genomic and somatic)
- ✓ PPIs
- ✓ Functional annotations

- ○ Phylogenetic information
- ○ Data on model organisms
- ○ Pharmacology data
- ○ Clinical proteogenomic data
- ○ Structural data
- ○ Protein interactions with pathogens
- ○ …

# Extending neXtProt contents and empowering network-based approaches by federating other life science SPARQL endpoints

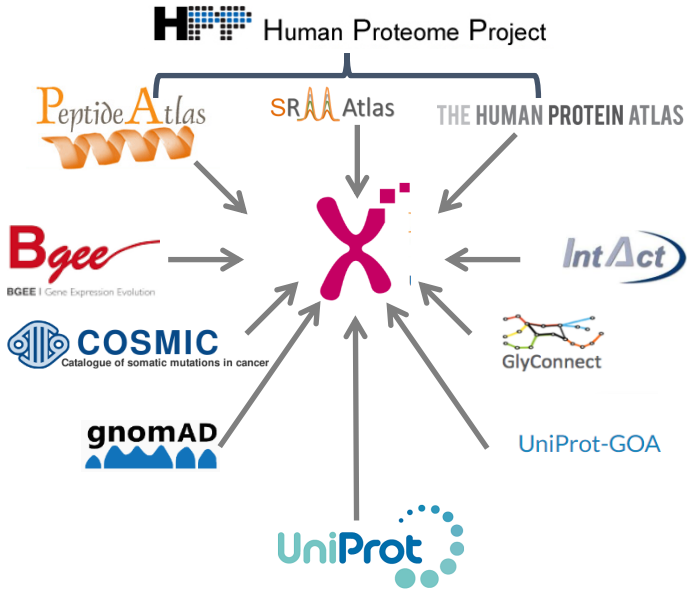
Pharmacology/Clinical data

DisGeNET

ChEMBL

Systems biology

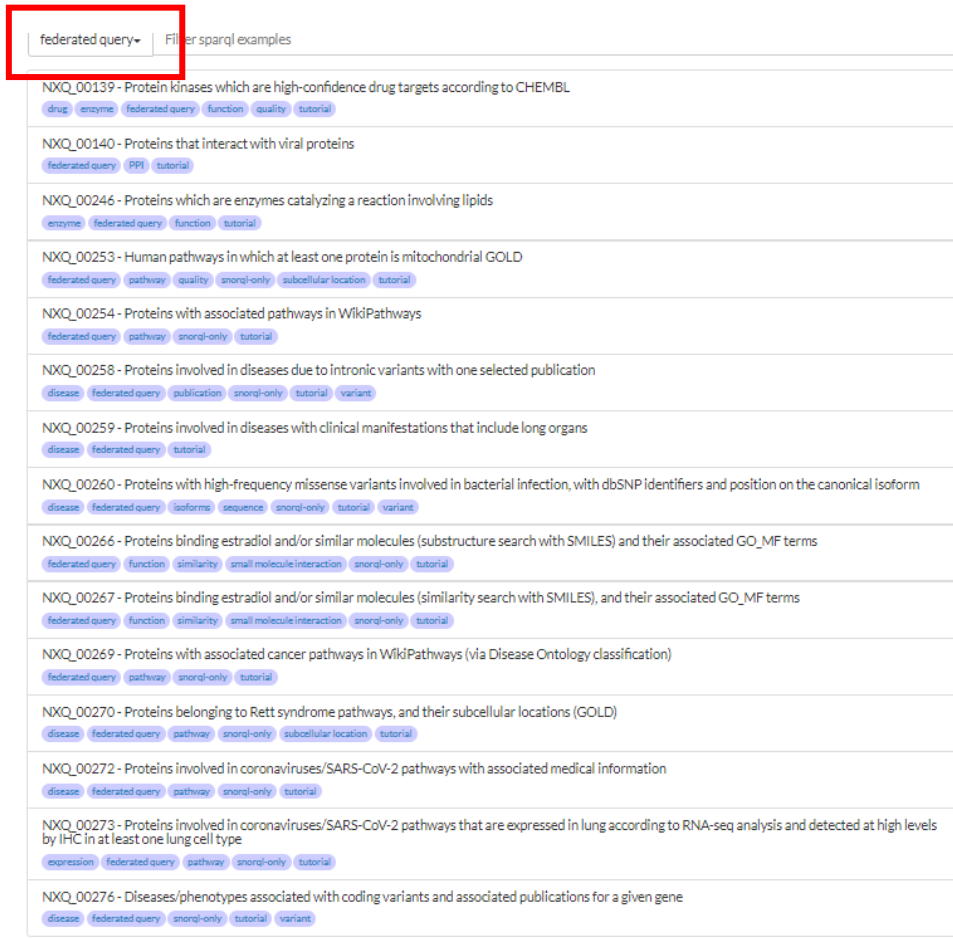
WikiPathways
Pathways for the People




PDBj
Protein Data Bank Japan

Structural biology

15 federated queries already available



# Example: Proteins binding estradiol and/or similar molecules (federated query with IDSM/SACHEM)

```
PREFIX sachem: <http://bioinfo.uochb.cas.cz/rdf/v1.0/sachem#>
PREFIX idsm: <https://idsm.elixir-czech.cz/sparql/endpoint/>
PREFIX chembl: <http://rdf.ebi.ac.uk/terms/chembl#>

SELECT distinct ?entry (group_concat(distinct str(?gomflab); SEPARATOR = ",") as ?gomfx) WHERE {
        SERVICE <https://idsm.elixir-czech.cz/sparql/endpoint/idsm> {
        SERVICE <https://idsm.elixir-czech.cz/sparql/endpoint/chembl> {
            ?compound sachem:similarCompoundSearch [ sachem:query "CC12CCC3C(C1CCC2O)CCC4=C3C=CC(=C4)O" ] . # smiles chain for estradiol
          }
                ?ACTIVITY rdf:type chembl:Activity;
        chembl:hasMolecule ?compound;
            chembl:hasAssay ?ASSAY.
          ?ASSAY chembl:hasTarget ?TARGET.
          ?TARGET chembl:taxonomy <http://identifiers.org/taxonomy/9606> . # human protein target
          ?TARGET chembl:hasTargetComponent ?COMPONENT.
          ?COMPONENT chembl:targetCmptXref ?UNIPROT.
          filter(contains(str(?UNIPROT),"uniprot"))
          }
    ?entry skos:exactMatch ?UNIPROT.
    ?entry :isoform ?iso.
    ?iso :goMolecularFunction / :term ?gomf .
    ?gomf rdfs:label ?gomflab .
}
group by ?entry
```

Proteins binding estradiol and/or similar molecules (similarity search with SMILES), and their associated GO_MF terms

html Go Reset eg. peroxisome, liver

Found a bug? Improve this query!

Query time is 10.989[s] for 240 rows

| entry | gomfx |
|---|---|
| entry:NX_Q9HAW8 -- (neXtProt link) | "enzyme binding,glucuronosyltransferase activity,protein heterodimerization activity,protein homodimerization activity,protein kinase C binding,retinoic acid binding" |
| entry:NX_P00918 -- (neXtProt link) | "arylesterase activity,carbonate dehydratase activity,hydro-lyase activity,protein binding,zinc ion binding" |
| entry:NX_P20813 -- (neXtProt link) | "anandamide 11,12 epoxidase activity,anandamide 14,15 epoxidase activity,anandamide 8,9 epoxidase activity,arachidonic |

NXQ_00139 - Protein kinases which are high-confidence drug targets according to CHEMBL
drug enzyme federated query function quality tutorial

NXQ_00140 - Proteins that interact with viral proteins
federated query PPI tutorial

NXQ_00246 - Proteins which are enzymes catalyzing a reaction involving lipids
enzyme federated query function tutorial

NXQ_00253 - Human pathways in which at least one protein is mitochondrial GOLD
federated query pathway quality snorql-only subcellular location tutorial

NXQ_00254 - Proteins with associated pathways in WikiPathways
federated query pathway snorql-only tutorial

NXQ_00258 - Proteins involved in diseases due to intronic variants with one selected publication
disease federated query publication snorql-only tutorial variant

NXQ_00259 - Proteins involved in diseases with clinical manifestations that include long organs
disease federated query tutorial

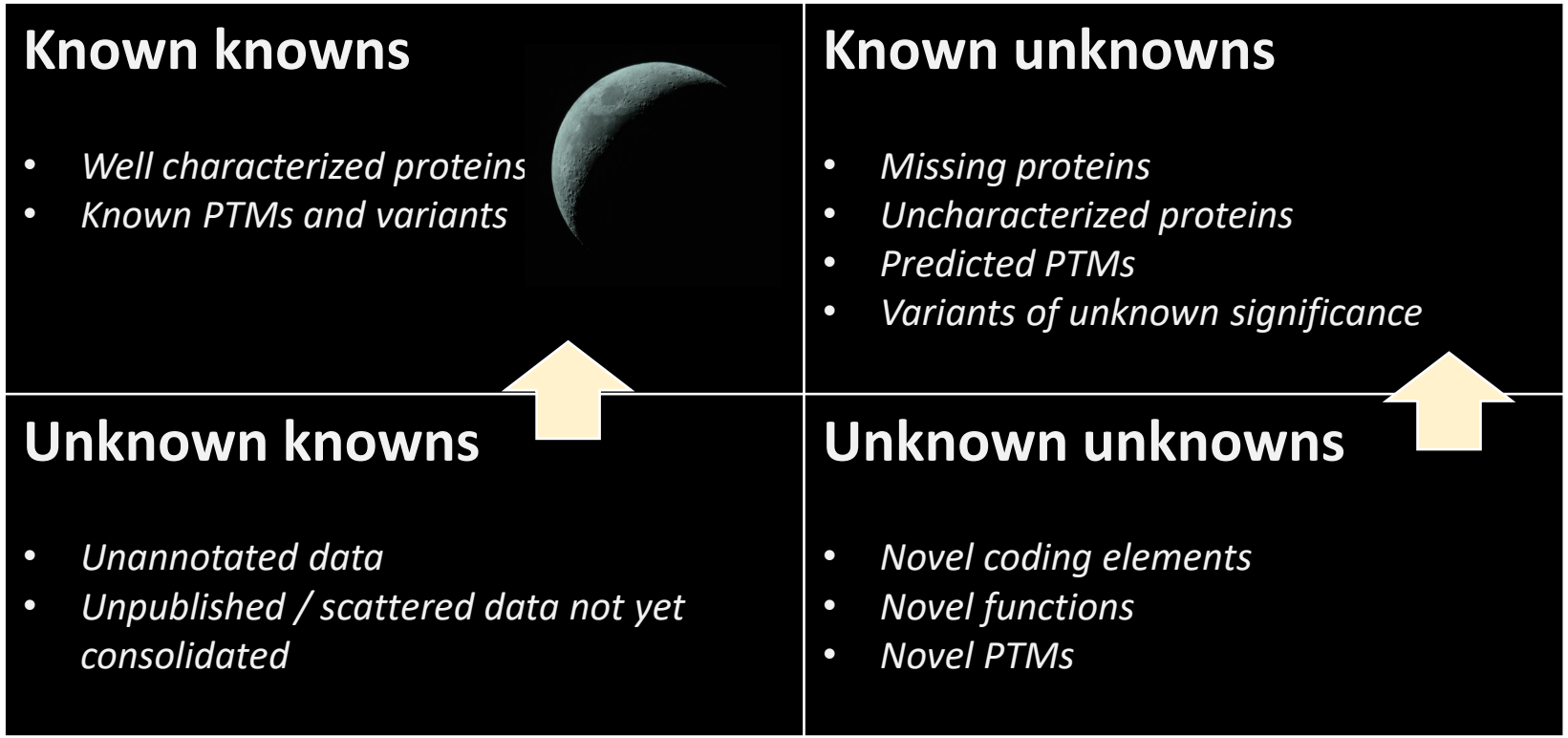
NXQ_00260 - Proteins with high-frequency missense variants involved in bacterial infection, with dbSNP identifiers and position on the canonical isoform
disease federated query isoforms sequence snorql-only tutorial variant

NXQ_00266 - Proteins binding estradiol and/or similar molecules (substructure search with SMILES) and their associated GO_MF terms
federated query function similarity small molecule interaction snorql-only tutorial

NXQ_00267 - Proteins binding estradiol and/or similar molecules (similarity search with SMILES), and their associated GO_MF terms
federated query function similarity small molecule interaction snorql-only tutorial

NXQ_00269 - Proteins with associated cancer pathways in WikiPathways (via Disease Ontology classification)
federated query pathway snorql-only tutorial

# Let’s complete the functional proteome together!




**Follow us!**

**Twitter:** @neXtProt_news
**ResearchGate:** neXtProt project

**Interested in**

- **- Integrating your data in neXtProt ?**
- **- Federating your data with neXtProt data using semantic technologies ?**
- **- Building personal, customized SPARQL queries for your own research ?**

**Let’s discuss!**

**lydie.lane@sib.swiss**

**support@nextprot.org**

# Special thanks

The neXtProt current team members

Paula Duek
Pierre-André Michel
Valentine Rech de Laval
Kasun Samarasinghe
Monique Zahn
Amos Bairoch

...and all the past members!

Young-Ki Paik

Chris Overall

Charles Pineau

Gil Omenn

Eric Deutsch

and all C-HPP collaborators