Article

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

ZAHN, Monique, *et al*.

Abstract

The neXtProt knowledgebase (https://www.nextprot.org) is an integrative resource providing both data on human protein and the tools to explore these. In order to provide comprehensive and up-to-date data, we evaluate and add new data sets. We describe the incorporation of three new data sets that provide expression, function, protein-protein binary interaction, post-translational modifications (PTM) and variant information. New SPARQL query examples illustrating uses of the new data were added. neXtProt has continued to develop tools for proteomics. We have improved the peptide uniqueness checker and have implemented a new protein digestion tool. Together, these tools make it possible to determine which proteases can be used to identify trypsin-resistant proteins by mass spectrometry. In terms of usability, we have finished revamping our web interface and completely rewritten our API. Our SPARQL endpoint now supports federated queries. All the neXtProt data are available via our user interface, API, SPARQL endpoint and FTP site, including the new PEFF 1.0 format files. Finally, the data on our FTP site is now CC BY 4.0 to [...]

Reference

ZAHN, Monique, *et al*. The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Research*, 2020, vol. 48, D1, p. D328-D334

PMID : 31724716
DOI : 10.1093/nar/gkz995

Nucleic Acids Research, 2019 **1**
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,*]

[1]CALIPHO group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland, [2]Department of microbiology and molecular medicine, Faculty of Medicine, University of Geneva, Geneva, Switzerland and [3]Haute école spécialisée de Suisse occidentale, Haute Ecole de Gestion de Genève, Carouge, Switzerland

Received September 11, 2019; Revised October 10, 2019; Editorial Decision October 11, 2019; Accepted October 18, 2019

ABSTRACT

The neXtProt knowledgebase (https://www.nextprot.org) is an integrative resource providing both data on human protein and the tools to explore these. In order to provide comprehensive and up-to-date data, we evaluate and add new data sets. We describe the incorporation of three new data sets that provide expression, function, protein-protein binary interaction, post-translational modifications (PTM) and variant information. New SPARQL query examples illustrating uses of the new data were added. neXtProt has continued to develop tools for proteomics. We have improved the peptide uniqueness checker and have implemented a new protein digestion tool. Together, these tools make it possible to determine which proteases can be used to identify trypsin-resistant proteins by mass spectrometry. In terms of usability, we have finished revamping our web interface and completely rewritten our API. Our SPARQL endpoint now supports federated queries. All the neXtProt data are available via our user interface, API, SPARQL endpoint and FTP site, including the new PEFF 1.0 format files. Finally, the data on our FTP site is now CC BY 4.0 to promote its reuse.

INTRODUCTION

Comprehensive, current, high quality data, as well as innovative and powerful tools are necessary for researchers to make the most of the ever-increasing data relevant to human biology. neXtProt (1), a knowledgebase focusing exclusively on human proteins, leverages the expert manual annotation carried out at specialist resources and in-house to provide a single point of reference. Information concerning human protein function, cellular localization, tissular expression, interactions, variants and their phenotypic effect, post-translational modifications (PTMs), as well as peptide identified in mass spectrometry experiments and epitopes recognized by antibodies have been integrated from a number of resources. By doing so, neXtProt extends the contents of UniProtKB/Swiss-Prot (2) to provide a more comprehensive data set.

However, data alone is not sufficient for scientists to comprehend complex information rapidly. For this reason, neXtProt organizes the information concerning an entry in several views, with interactive viewers that allow the user to select the data displayed. We also provide tools to analyze and explore the data. A basic, full text search, as well as an advanced, SPARQL-based search, allow users to search the data in neXtProt. Additional tools have been implemented. Users can store and compare private lists of entries. The peptide uniqueness checker (3) determines which peptides are unambiguous and can thus be used to confidently identify protein entries (4).

In this manuscript, we describe the latest progress on developing neXtProt. Since 2016, three major data sets have been integrated. Firstly, high quality, tissular expression data from the Human Protein Atlas (HPA) obtained by RNA-seq (5) has been added. Secondly, information annotated from the literature on the function, cellular localization, interactions and phosphorylations carried out by human protein kinases has been incorporated. Lastly, variant frequency data from the Genome Aggregation Database (gnomAD) (6) extends the information on sequence variations at the protein level. We also report on improvements made to the peptide uniqueness checker and the implementation of the new protein digestion tool. Finally, we present improvements to the web site and SPARQL endpoint to improve the accessibility and usability of the neXtProt data.

neXtProt data overview

The first neXtProt release in April 2011 contained data from UniProtKB, Ensembl, HPA, Bgee and GOA. Since then

*To whom correspondence should be addressed. Tel: +41 22 379 58 41; Email: lydie.lane@sib.swiss

neXtProt has been steadily incorporating new data from additional resources, with a particular emphasis on expression data, proteomics data and variant data. The current neXtProt release was built using human genome assembly GRCh38 (7). The data from UniProtKB (2) is currently supplemented with data from Bgee (8), HPA (5,9), PeptideAtlas (10), SRMAtlas (11), GOA (12), dbSNP (13), Ensembl (14), COSMIC (15), DKF GFP-cDNA localization (16,17), Weizmann Institute of Science's Kahn Dynamic Proteomics Database (18), IntAct (19), GlyConnect (20), gnomAD (6), as well as in-house curated data (21,22). Table 1 summarizes the changes in the content since our last neXtProt update (1).

The data in the UniProtKB/Swiss-Prot (Reviewed) entries for Homo sapiens (TaxID: 9606) having the keyword Complete proteome (KW-0181) provide the groundwork for neXtProt. In order to evaluate the improvement in coverage through the integration of data from sources other than UniProtKB, we determined the number of entries in neXtProt with data from UniProtKB with that having data from any source using SPARQL queries (Table 2). UniProtKB provides excellent coverage for a single resource; it thus provides a good foundation for the construction of neXtProt. The incorporation of data from additional sources considerably improves the coverage—over 78% of entries in neXtProt have information about the function, cellular localization, interactions, expression, post-translational modifications and variants.

RNA-seq

We incorporated the RNA-seq data from 37 different normal tissues from Human Protein Atlas. As RNA-seq data is highly accurate for quantifying expression levels with high reproducibility, this improved the expression data provided in neXtProt at the level of the transcript, which until then came from microarray and expressed sequence tag (EST) data. While RNA-seq is quantitative, the semi-quantitative expression values (undetected, low, medium and high) provided by HPA were taken and are displayed in the same manner as the other data in the *Expression view* to make for easier comparison (Figure 1). This also enables the RNA-seq data to be queried in the same manner using SPARQL.

Protein kinases

Another new large set that was integrated is a set of manual annotations that we have created to capture a wide range of published experimental results concerning 300 protein kinases. The proteins phosphorylated by these kinases, as well as whenever possible the specific amino acid residue which is phosphorylated, were annotated. This phosphorylation data set complements that of PeptideAtlas, which only provides phosphorylation sites. With this data, neXtProt now contains 10 725 entries (52%) with a phosphorylation on a serine, threonine or tyrosine residue. Of the 115 822 phosphorylation sites, only 5219 are associated with a specific protein kinase. The substrates of a specific protein kinase can be retrieved using a SPARQL query; for instance, the query NXQ_00069 retrieves all proteins phosphorylated by SYK. The protein's molecular function and its involvement in a biological process were captured using Gene Ontology (GO) terms (23,24). Binary interactions with human proteins were also annotated and are displayed in the *Interactions view*.

As with the phenotypic data described in our previous report (1), this protein kinase data is available in the new Protein kinase function portal. Accessible from the top menu 'Portals', the data are presented in tabular form, with each column being searchable and sortable. More details concerning the experimental context for the data in all portals is now provided. Two new columns, labeled Cell line / Tissue and Experimental details, can be used to filter the data. The data in the portals can be downloaded in CSV format, copied or printed. The entry accession (AC) corresponding to the annotation subject has also been added.

Variant frequency

To date the corpus of variant data in neXtProt covers variants observed in health and disease, as well as the phenotypic effect of the variants. The neXtProt database contains over six million single amino acid variations imported from UniProtKB, dbSNP, COSMIC and manually annotated from the literature, but it is difficult to make use of this variant data in the absence of information about their frequencies in human populations. The Genome Aggregation Database (gnomAD) (6) spans 126 216 exome sequences and 15 136 whole-genome sequences extracted from a variety of large-scale sequencing projects and provides computed allele frequencies for most of the reported variants. We have thus integrated variant frequency information from the gnomAD version 2.1.1.

neXtProt now contains 18 685 entries (92%) and 2 691 323 variants (45%) with frequency data from gnomAD. We display the number of times the allele was sequenced (allele count), the number of individuals homozygous for the allele (homozygous count), the total number of alleles sequenced (allele number) and the allele frequency in the evidence (Figure 2). SPARQL queries can be used to answer questions such as which variants have a frequency greater than 0.1 (NXQ_00255) or which variants are frequently found in a homozygous state (NXQ_00256).

Peptide uniqueness checker

The peptide uniqueness checker (3) allows scientists to define which peptides can be used to validate the existence of human proteins by determining whether a peptide maps uniquely versus multiply to human protein sequences taking into account isobaric substitutions, alternative splicing and single amino acid variants. It was adapted to take into account the entries with identical isoform sequences. Peptides matching such sequences were considered to be 'Found in other entries'; such peptides are now considered 'Pseudo-unique' and entries or isoforms having identical sequences are labelled with an asterisk in the results. We also added a brief description of what the tool does and examples covering all cases in the input section of the interface. In response to user requests, we added in the display of the results icons so as to help color-blind users to distinguish whether peptides are 'unique', 'pseudo-unique' or 'not unique'.

Table 1. Data content of neXtProt data release 2019-08-22

Entries	Statistics	Change since data release 2016-08-25	Source(s)
Entries	20 399	+338	UniProtKB
Isoforms (Sequences)	42 410	+386	UniProtKB
Binary interactions	240 010	+99 740	IntAct, neXtProt
Post-translational modifications (PTMs)	190 921	+48 468	UniProtKB, neXtProt, PeptideAtlas, GlyConnect
Variants (including disease mutations)	6 019 871	+1 075 957	UniProtKB, COSMIC, dbSNP, neXtProt, gnomAD
Phenotypic annotations	19 602	+11 588	neXtProt
Entries with a molecular function	17 177	+654	GOA, neXtProt
Entries with a biological process	16 964	+692	GOA, neXtProt
Entries with an expression profile	19 367	+1038	Bgee, HPA, neXtProt
Entries with a disease	4553	+637	UniProtKB
Entries with proteomics data	18 727	+1448	PeptideAtlas, neXtProt
Entries with an experimental 3D structure	6 505	+765	PDB via UniProtKB
Cited publications	115 935	+16 013	All resources

Table 2. Coverage in neXtProt data release 2019-08-22

Annotations	Entries with evidence from UniProtKB (%[a])	Entries with evidence from any source (%[a])
GO molecular function	12 360 (60%)	17 177 (84%)
GO biological process	11 665 (57%)	16 964 (83%)
GO cellular component	14 581 (71%)	18 129 (89%)
Subcellular location	16 590 (81%)	18 527 (91%)
Binary interactions	8653 (42%)	16 411 (80%)
Expression	9876 (48%)	19 527 (96%)
Peptide mapping	0 (0%)	18 727 (92%)
Antibody mapping	0 (0%)	16 423 (80%)
Post-translational modifications (PTMs)	14 000 (69%)	15 905 (78%)
Variants	12 919 (63%)	19 621 (96%)

[a]The total number of entries is 20 399.

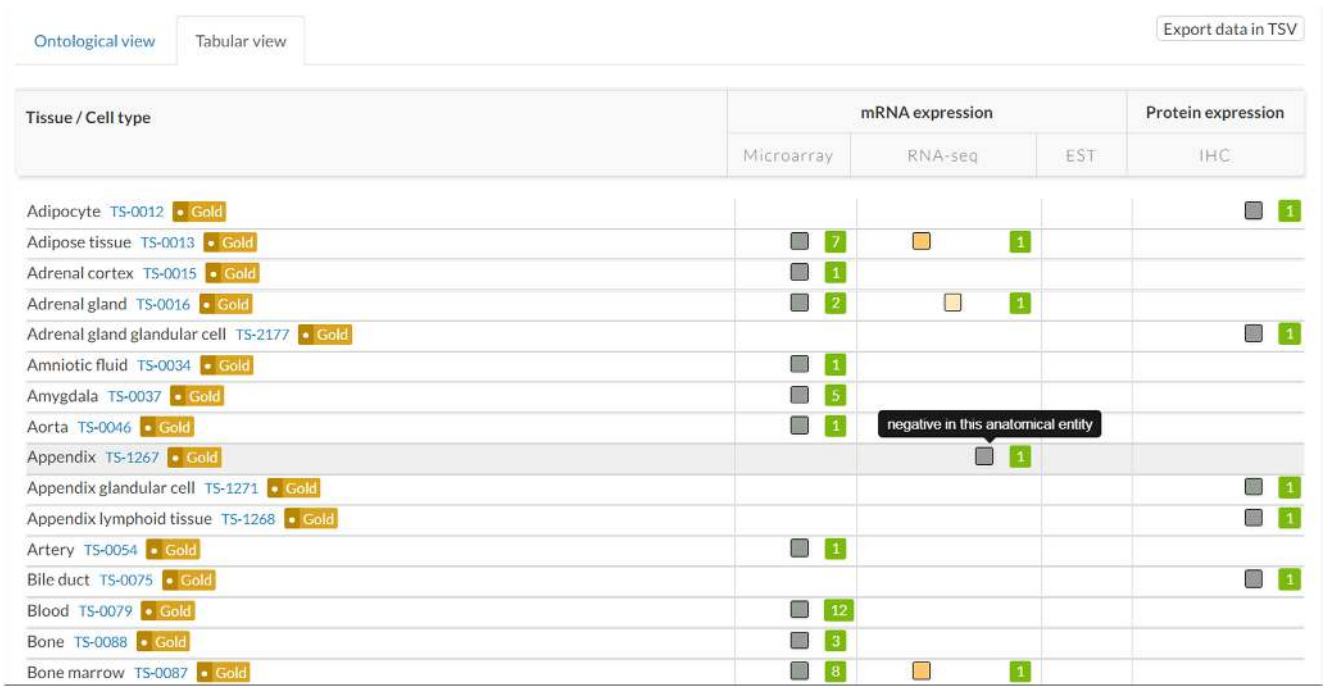

Figure 1. Tabular view showing the expression data for insulin (NX_P01308). Expression data at the mRNA and protein level are displayed in the same semi-quantitative manner for easier comparison. Four levels of expression (undetected, low, medium, high) are possible. Mousing-over the data point displays the expression level textually.

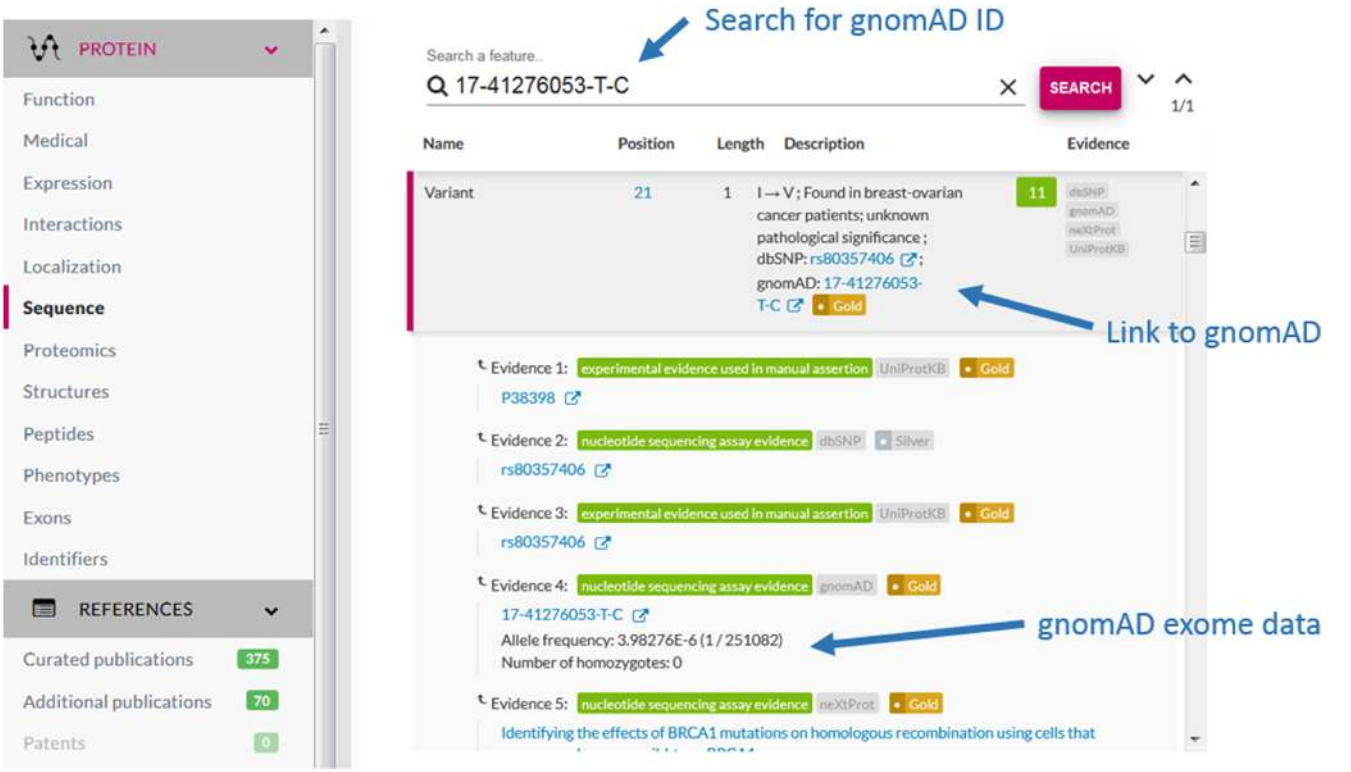

Figure 2. Allele frequency information in the *Sequence view* for BRCA1 (NX_P38398) variants. To find a gnomAD variant, search in the feature table with the gnomAD ID. A link to the corresponding variant in gnomAD is found in the description of the variant and the evidence. The allele frequency, with the allele count and allele number in brackets, as well as the homozygote count, are displayed in the evidence.

Protein digestion

A new tool that performs *in silico* protein digestion is now available at https://www.nextprot.org/tools/protein-digestion. This tool allows the user to input the neXtProt accession for a specific isoform of an entry and returns the peptide sequences obtained upon digestion. Combined with the peptide uniqueness checker, it can be used to determine which proteases can be used to identify trypsin-resistant proteins by mass spectrometry. For example, trypsin digestion of the cancer-testis antigen CTAGE1 isoform 1 (NX_Q9HC47-1) does not yield a peptide of 9–35 aa (Figure 3A and B); however digestion with high specificity cleavage with chymotrypsin (CHYMOTRYPSIN_HIGH_SPEC) results in two, non-overlapping proteotypic peptides (Figure 3C).

Web site and API

In our last update (1), we reported changes in the home page, the navigation and the documentation. Since then, we have completely revamped the neXtProt website. All entry, publication and controlled vocabulary term pages have been rewritten so that they load faster, thereby improving their usability. This was necessary, as the amount of data in neXtProt has increased considerably over the years. We also introduced a Tabular view in the *Expression view* (Figure 1). This new view shows the expression level for every tissue assayed. An export functionality allows all the experimental expression data for the entry to be downloaded in tab-delimited format. We also combined the Gene Identifiers and Protein Identifiers entry views in a single view. This allows users to find all the identifiers for an entry in the *Identifiers view*. Having completed the revamping of our web site, our original website and the associated API were disconnected. Users should now use the API at https://api.nextprot.org/.

Querying using SPARQL

In the last four years, neXtProt has been promoting the use of SPARQL, a semantic query language for databases, to explore human data. Semantic technologies can help to generate innovative hypotheses where classical data mining tools have failed (protein function prediction, drug repositioning, etc.). neXtProt provides over 160 pre-built queries in its Advanced search (https://www.nextprot.org/proteins/search?mode=advanced) and SNORQL (https://snorql.nextprot.org/) interfaces. The former retrieves entries, while the latter retrieves any data, meeting the defined criteria. The queries illustrate the types of questions that can be answered using SPARQL. The data model documentation is provided in the SNORQL Help and a user guide to help the user in his/her first steps in SPARQL has been published (25).

The use of SPARQL allows users to run federated queries across multiple resources relevant to human biology. Thus SPARQL allows queries to be carried out on data both in

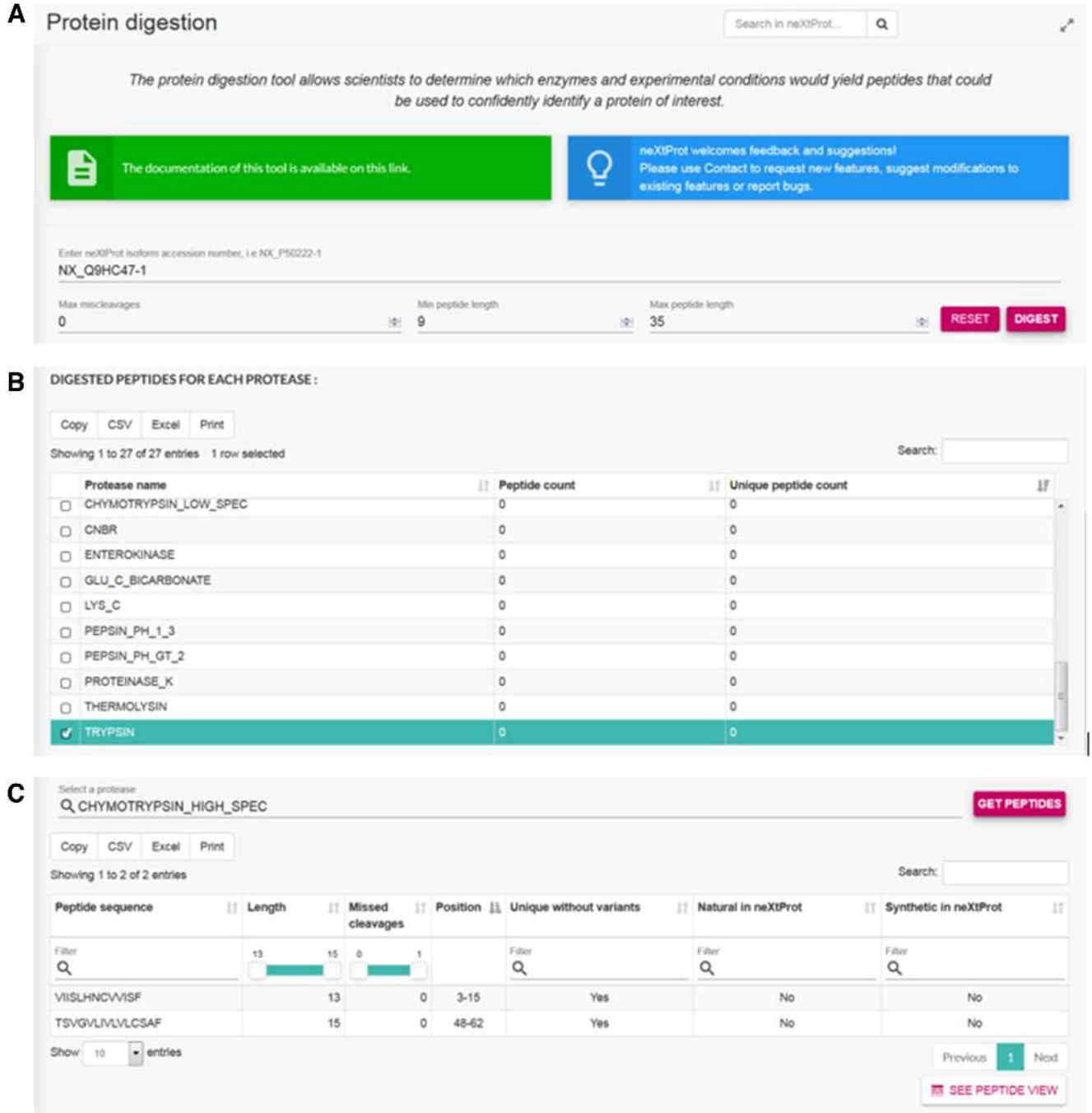

Figure 3. Protein digestion tool. (**A**) Input form requiring the neXtProt isoform accession number for the protein to be digested. Default digestion parameters (maximum number of miscleavages, minimum peptide length and maximum peptide length) can be modified by the user. (**B**) Peptide count and unique peptide count for the digestion with 27 proteases or conditions. Select a protease to see the peptides obtained. (**C**) Table displaying information about the peptides obtained with the selected digestion conditions. The peptide sequence, length, number of missed cleavages, position in the sequence, whether the peptide is unique or not (without taking into account variants) and whether the peptide is found in neXtProt, as a natural and synthetic (SRM peptide) are displayed. A link to the neXtProt *Peptide view* of the entry is provided.

and beyond neXtProt. Figure 4 shows examples of federated queries. Currently these also query data in ChEMBL, DrugBank, PDB, Rhea, UniProtKB and WikiPathways. The neXtProt SPARQL endpoint (https://api.nextprot.org/sparql) also supports federated queries.

Data availability

High-throughput identification and quantification of proteins, including sequence variants and post-translational modifications (PTMs) in biological samples by mass spectrometry-based proteomics is becoming commonplace.

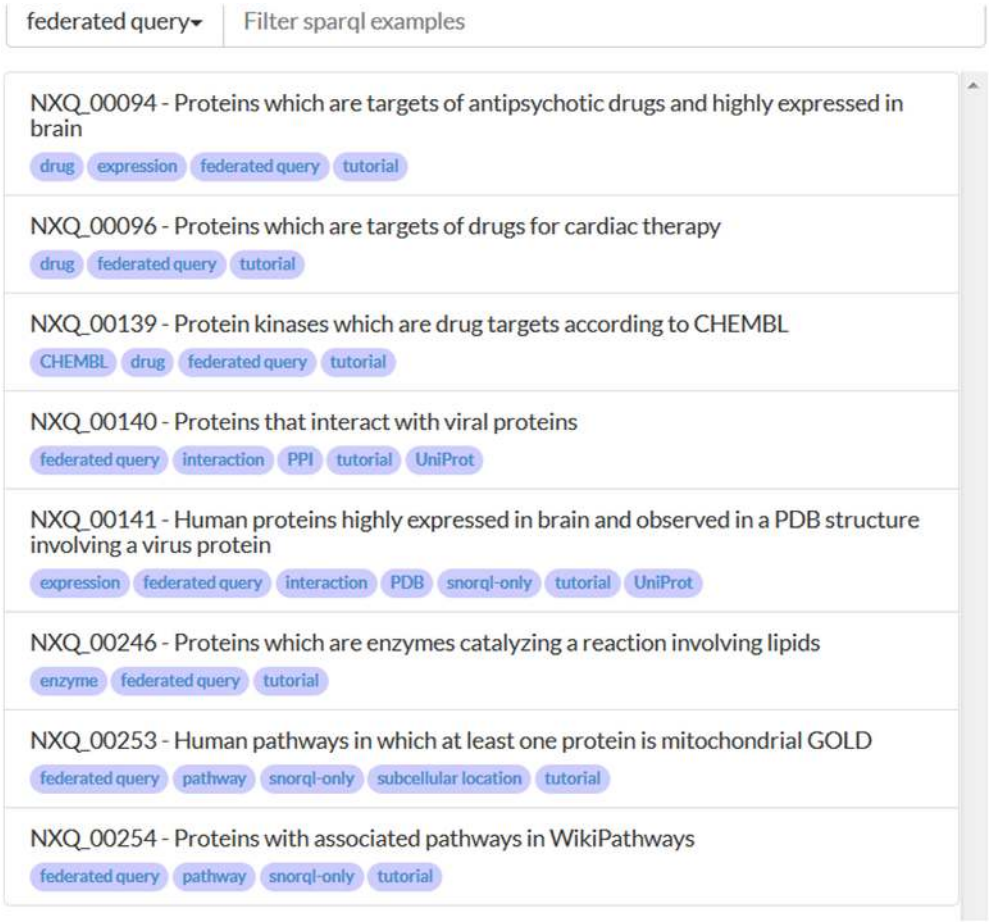

Figure 4. Federated SPARQL query examples. Screenshot showing all queries tagged 'federated' in the neXtProt SNORQL interface.

While sequence variations need to be taken into account in the search space used to analyze the data, doing so remains a challenge. The Proteomics Standards Initiative (PSI) has designed and implemented the PSI extended FASTA format (PEFF) (26) to facilitate the search for known sequence variants and PTMs. Based on the FASTA format, PEFF encodes substantially more metadata about the sequence collection as well as individual entries, including support for encoding known sequence variants, PTMs, and proteoforms. We have worked closely with PSI with the outcome that neXtProt was the first resource to implement the PEFF v1.0 format. This ensures the interoperability of neXtProt data with the sequence search engine Comet. We expect that the PEFF format will soon be adopted by other MS software tools.

All the neXtProt data, dating back to the first release in 2011, can be downloaded. In order to foster the reuse of the data in neXtProt, we have lifted the 'no derivatives' restriction applying to the data available from our FTP site (ftp://ftp.nextprot.org/pub/). As of 21 February 2018, the license applying to the use of our data available is CC BY 4.0.

CONCLUSION

The past years have seen numerous changes in neXtProt. We have incorporated new, high quality RNA-seq expression, protein kinase function and variant frequency data, and updated the data from practically all our sources, in order to provide an up-to-date, comprehensive data set. Some changes, such as the rewriting of our user interface, have been necessary to cope with the increase in data. Others, such as the modification to the peptide uniqueness checker, were prompted by changes in the data and feedback from users. We continue to support research in proteomics and have implemented the protein digestion tool to enable researchers to find alternatives to trypsin when planning experiments to validate the existence of a protein. We welcome feedback on our data, tools and website and encourage users to contact us using the e-mail support@nextprot.org, Twitter (@neXtProt_news) or ResearchGate (neXtProt project).

ACKNOWLEDGEMENTS

We thank the UniProtKB groups at SIB, EBI and PIR for their dedication in providing up-to-date high-quality annotations for the human proteins in UniProtKB/Swiss-Prot thus providing neXtProt with a solid foundation, as well as all the other resources from which we integrated data. We also thank Frédérique Lisacek and Eric Deutsch for stimulating discussions, advice and/or providing us data.

The neXtProt server is hosted by the SIB Swiss Institute of Bioinformatics' Core-IT facility.

FUNDING

Swiss State Secretariat for Education, Research and Innovation SERI funding to the SIB Swiss Institute of Bioinformatics.

Conflict of interest statement. None declared.

REFERENCES

1. Gaudet,P., Michel,P.-A., Zahn-Zabal,M., Britan,A., Cusin,I., Domagalski,M., Duek,P.D., Gateau,A., Gleizes,A., Hinard,V. *et al.* (2017) The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.*, **45**, D177–D182.
2. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
3. Schaeffer,M., Gateau,A., Teixeira,D., Michel,P.-A., Zahn-Zabal,M. and Lane,L. (2017) The neXtProt peptide uniqueness checker: a tool for the proteomics community. *Bioinformatics*, **33**, 3471–3472.
4. Deutsch,E.W., Overall,C.M., Van Eyk,J.E., Baker,M.S., Paik,Y.-K., Weintraub,S.T., Lane,L., Martens,L., Vandenbrouck,Y., Kusebauch,U. *et al.* (2016) Human proteome project mass spectrometry data interpretation guidelines 2.1. *J. Proteome Res.*, **15**, 3961–3970.
5. Uhlén,M., Fagerberg,L., Hallström,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,Å., Kampf,C., Sjöstedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
6. Lek,M., Karczewski,K.J., Minikel,E. V, Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60, 706 humans. *Nature*, **536**, 285–291.
7. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
8. Bastian,F.B., Parmentier,G., Roux,J., Moretti,S., Laudet,V. and Robinson-Rechavi,M. (2008) Bgee: Integrating and comparing heterogeneous transcriptome data among species. in *DILS: Data Integration in Life Sciences*. *Lect. Notes Comput. Sci.*, **5109**, 124–131.
9. Thul,P.J. and Lindskog,C. (2018) The Human Protein Atlas: A spatial map of the human proteome. *Protein Sci.*, **27**, 233–244.
10. Deutsch,E.W., Sun,Z., Campbell,D., Kusebauch,U., Chu,C.S., Mendoza,L., Shteynberg,D., Omenn,G.S. and Moritz,R.L. (2015) State of the human proteome in 2014/2015 as viewed through PeptideAtlas: enhancing accuracy and coverage through the AtlasProphet. *J. Proteome Res.*, **14**, 3461–3473.
11. Kusebauch,U., Campbell,D.S., Deutsch,E.W., Chu,C.S., Spicer,D.A., Brusniak,M.-Y., Slagel,J., Sun,Z., Stevens,J., Grimes,B. *et al.* (2016) Human SRMAtlas: a resource of targeted assays to quantify the complete human proteome. *Cell*, **166**, 766–778.
12. Huntley,R.P., Sawford,T., Mutowo-Meullenet,P., Shypitsyna,A., Bonilla,C., Martin,M.J. and O'Donovan,C. (2015) The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res.*, **43**, D1057–D1063.
13. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
14. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
15. Forbes,S.A., Beare,D., Boutselakis,H., Bamford,S., Bindal,N., Tate,J., Cole,C.G., Ward,S., Dawson,E., Ponting,L. *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.
16. Liebel,U., Starkuviene,V., Erfle,H., Simpson,J.C., Poustka,A., Wiemann,S. and Pepperkok,R. (2003) A microscope-based screening platform for large-scale functional protein analysis in intact cells. *FEBS Lett.*, **554**, 394–398.
17. Simpson,J.C., Wellenreuther,R., Poustka,A., Pepperkok,R. and Wiemann,S. (2000) Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.*, **1**, 287–292.
18. Frenkel-Morgenstern,M., Cohen,A.A., Geva-Zatorsky,N., Eden,E., Prilusky,J., Issaeva,I., Sigal,A., Cohen-Saidon,C., Liron,Y., Cohen,L. *et al.* (2010) Dynamic Proteomics: a database for dynamics and localizations of endogenous fluorescently-tagged proteins in living human cells. *Nucleic Acids Res.*, **38**, D508–D512.
19. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
20. Alocci,D., Mariethoz,J., Gastaldello,A., Gasteiger,E., Karlsson,N.G., Kolarich,D., Packer,N.H. and Lisacek,F. (2019) GlyConnect: glycoproteomics goes visual, interactive, and analytical. *J. Proteome Res.*, **18**, 664–677.
21. Hinard,V., Britan,A., Schaeffer,M., Zahn-Zabal,M., Thomet,U., Rougier,J.-S., Bairoch,A., Abriel,H. and Gaudet,P. (2017) Annotation of functional impact of voltage-gated sodium channel mutations. *Hum. Mutat.*, **38**, 485–493.
22. Cusin,I., Teixeira,D., Zahn-Zabal,M., Rech de Laval,V., Gleizes,A., Viassolo,V., Chappuis,P.O., Hutter,P., Bairoch,A. and Gaudet,P. (2018) A new bioinformatics tool to help assess the significance of BRCA1 variants. *Hum. Genomics*, **12**, 36.
23. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
24. The Gene Ontology Consortium. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
25. Zahn-Zabal,M. and Attwood,T.K. (2019) A critical guide to the neXtProt knowledgebase: querying using SPARQL [version 1; not peer reviewed]. *F1000Research*, **8**, 791.
26. Binz,P.-A., Shofstahl,J., Vizcaíno,J.A., Barsnes,H., Chalkley,R.J., Menschaert,G., Alpi,E., Clauser,K., Eng,J.K., Lane,L. *et al.* (2019) Proteomics standards initiative extended FASTA format. *J. Proteome Res.*, **18**, 2686–2692.