D1522–D1527 Nucleic Acids Research, 2022, Vol. 50, Database issue *Published online 6 December 2021*
https://doi.org/10.1093/nar/gkab1081

iProX in 2021: connecting proteomics data sharing with big data

Tao Chen[1,†], Jie Ma[1,†], Yi Liu[1], Zhiguang Chen[2], Nong Xiao[2], Yutong Lu[2], Yinjin Fu[2], Chunyuan Yang[1], Mansheng Li[1], Songfeng Wu[1], Xue Wang[1], Dongsheng Li[1], Fuchu He[1,*], Henning Hermjakob[1,3,*] and Yunping Zhu[1,4,*]

[1]State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China, [2]School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 26469, China, [3]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and [4]Basic Medical School, Anhui Medical University, Anhui 230032, China

Received September 15, 2021; Revised October 16, 2021; Editorial Decision October 18, 2021; Accepted October 22, 2021

ABSTRACT

The rapid development of proteomics studies has resulted in large volumes of experimental data. The emergence of big data platform provides the opportunity to handle these large amounts of data. The integrated proteome resource, iProX (https://www.iprox.cn), which was initiated in 2017, has been greatly improved with an up-to-date big data platform implemented in 2021. Here, we describe the main iProX developments since its first publication in *Nucleic Acids Research* in 2019. First, a hyperconverged architecture with high scalability supports the submission process. A hadoop cluster can store large amounts of proteomics datasets, and a distributed, RESTful-styled Elastic Search engine can query millions of records within one second. Also, several new features, including the Universal Spectrum Identifier (USI) mechanism proposed by ProteomeXchange, RESTful Web Service API, and a high-efficiency reanalysis pipeline, have been added to iProX for better open data sharing. By the end of August 2021, 1526 datasets had been submitted to iProX, reaching a total data volume of 92.42TB. With the implementation of the big data platform, iProX can support PB-level data storage, hundreds of billions of spectra records, and second-level latency service capabilities that meet the requirements of the fast growing field of proteomics.

INTRODUCTION

With the advance of high-throughput technologies, large-scale biological data has accumulated at an unprecedented rate. First, methods were established for the genome and transcriptome, followed by the proteome and metabolome, which may be collectively designated ‘multi-omics’ (1,2). Today, nearly all fields in life sciences can be connected by big data. As in other areas of big data science, the major challenges in the proteomics field include data storage, management, analyses, and open sharing.

The ProteomeXchange (PX, http://www.proteomexchange.org/) consortium (3,4) coordinates a stable, distributed infrastructure for effective proteomics data sharing through interaction with PX members, including PRIDE (http://www.ebi.ac.uk/pride/archive/, EMBL-EBI, Cambridge, UK) (5), PeptideAtlas with the PASSEL resource (http://www.peptideatlas.org/passel/, ISB, Seattle, WA, USA) (6), MassIVE (https://massive.ucsd.edu/, UCSD, San Diego, CA, USA), jPOST (https://jpostdb.org/, various institutions, Japan) (7), iProX (National Center for Protein Sciences, Beijing, China) (8), and Panorama Public (https://panoramaweb.org/, University of Washington, Seattle, WA, USA) (9). Driven by improvements in speed and resolution of mass spectrometry (MS), the scale and complexity of proteomics datasets are expanding, which has resulted in a rapid accumulation of data in almost all PX repositories. To move forward with sharing large proteomics datasets, both hardware and software must continue to improve; therefore, incorporating big data technology, framework and scalable cloud-based solutions will help to manage huge proteomics datasets.

iProX was launched in April 2017 and joined the PX consortium in November 2017. As a full member of the con-

[*]To whom correspondence should be addressed. Tel: +86 10 61777058; Fax: +86 10 61777058; Email: zhuyunping@gmail.com
Correspondence may also be addressed to Fuchu He. Email: hefc@nic.bmi.ac.cn
Correspondence may also be addressed to Henning Hermjakob. Email: hhe@ebi.ac.uk
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

sortium, iProX (https://www.iprox.cn/) has been updated significantly by implementing an up-to-date big data platform in 2021. It can support storage and rapid access to large amounts of proteomic data. Here, we describe the main developments in iProX since its first publication in *Nucleic Acids Research* in 2019. First, we summarize the overall data submission and data statistics to demonstrate the wide adoption of iProX. Then, we highlight the big data architecture and infrastructure of iProX, which can support PB-level data storage, hundreds of billions of spectra records, and second-level latency service capabilities to meet the requirements of rapidly accumulating proteomics data. Finally, we introduce the implementation of the Universal Spectrum Identifier (USI), RESTful Web Service API, and the reanalysis and visualization pipeline of public data in iProX.

CURRENT STATUS AND UPDATES OF REPOSITORY

By the end of August 2021, 1526 datasets had been submitted to iProX, consisting of 984 public released datasets (64%) and 542 to-be-released datasets (36%), for a total of 92.42TB of accumulated data. As shown in Figure 1A-C, there has been a rapid growth in the number and size of proteomics datasets generated from 2017 to 2021. This has been facilitated by the continuous improvements of mass spectrometry instruments and sample handling workflows (10). Significantly large (>500GB) and super large (>1TB) proteome studies have been completed in recent years (Figure 1B and Supplementary Figure S1). There are 19 datasets with more than 1 TB in data size submitted to iProX (Supplementary Figure S1) and 18 (94.7%) of them have been submitted within past 3 years (2019–2021). As shown in Figure 1D, the most frequent species in iProX include *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Escherichia coli*, *Saccharomyces cerevisiae*, *Triticum aestivum*, *Oryza sativa* and *Zea mays*. Moreover, the datasets generated from some species with few proteome studies have also been collected in iProX (Supplementary Figure S2). For example, datasets generated from studies on *Haemaphysalis longicornis* (10 datasets), *Anabaena* sp. *PCC 7120* (*Nostoc* sp. *PCC 7120,* 3 datasets), *Bambusa pervariabilis* (2 datasets), and *Yersinia pestis CO92* (7 datasets) in iProX, represent almost all the proteome datasets from these species in ProteomeXchange.

BIG DATA ARCHITECTURE AND INFRASTRUCTURE OF IPROX

First, a hyper-converged architecture with high scalability has been constructed to support the submission process. Next, a hadoop cluster (11) is used to store the large amounts of proteomics data. The storage capacity of iProX has expanded from 360TB to 1PB in 2021. Also, a distributed, RESTful-styled Elastic Search engine (12) was employed to retrieve millions of records within one second. As shown in Figure 2, for the submission process, the hyper-converged architecture integrates the server, storage, and network resources into a virtual pool, so that it can achieve high scalability using a distributed scale-out expansion for storage and computing resources. Re-analyzed datasets, such as that for proteins, peptides and spectra, are stored in a distributed column-oriented storage database known as HBase (13) in the Hadoop cluster. The indexes of the identifications are stored in an Elastic Search cluster for a distributed, high-scalable, real-time search and data analysis. A universal search interface for both web-based and APIs are provided for obtaining the metadata and reanalysis data of the original submitted datasets.

In order to achieve independent, high-speed data file transfer, both web-based and fast Asepra (https://asperasoft.com/) based upload and download steps were reconstructed into independent transferring sub-services through the RESTful API interface, which refers to the service-oriented architecture. To provide continuous and reliable high speed transfer services, we also upgraded Aspera into the latest release and integrated it into the iProX system. Searching metadata as well as identified proteins, peptides and spectra are also contained within sub-services to achieve a second level response without interrupting the regular submissions from individual submitters and large-scale projects. Thus, we have improved the performance of iProX in terms of high availability, high reliability and real-time response without changing the original submission process.

Notably, a disaster recovery system and full real-time backup site of iProX was designed and deployed to the National Supercomputing Center in Guangzhou (Guangdong, southern China), which can take over the service within several minutes when the main site in Beijing is unavailable. Thus, iProX can provide continuous uninterrupted service.

NEW FEATURES IN IPROX 2021

Based on the implementation of the hadoop-based big data platform for iProX, we developed several new features, including the implementation of Universal Spectrum Identifier (USI), the reanalysis and visualization pipeline for public data in iProX, and RESTful Web Service APIs, as shown in Figure 3.

UNIVERSAL SPECTRUM IDENTIFIER (USI)

The ability to refer to specific spectra of high importance and cite data in published manuscripts was done by implementing a new standardized Universal Spectrum Identifier (USI) proposed by PX (14,15). The USI is a multi-part key separated by colon characters; for example, mzspec:PXD006512:CNHPP_HCC_LC_profiling_L006_P_F1:scan:64442:VADALTNAVAHVDDMPNALSALSDLHAHK/3. This USI consists of five basic components, which indicates that the peptide VADALTNAVAHVDDMPNALSALSDLHAHK with a charge of 3 is identified in the raw file (CNHPP_HCC_LC_profiling_L006_P_F1.raw of PXD006512 dataset), with the scan number, 64 442. By this interpretation, it can be encoded to the specific spectrum contained in the dataset deposited to iProX. In iProX, USI locates the spectrum in the HBase cluster by using the Elastic Search engine. iProX enables USI lookup and display at http://www.iprox.cn/page/spectrum.html for 20 million spectra in the HBase. Users can also access

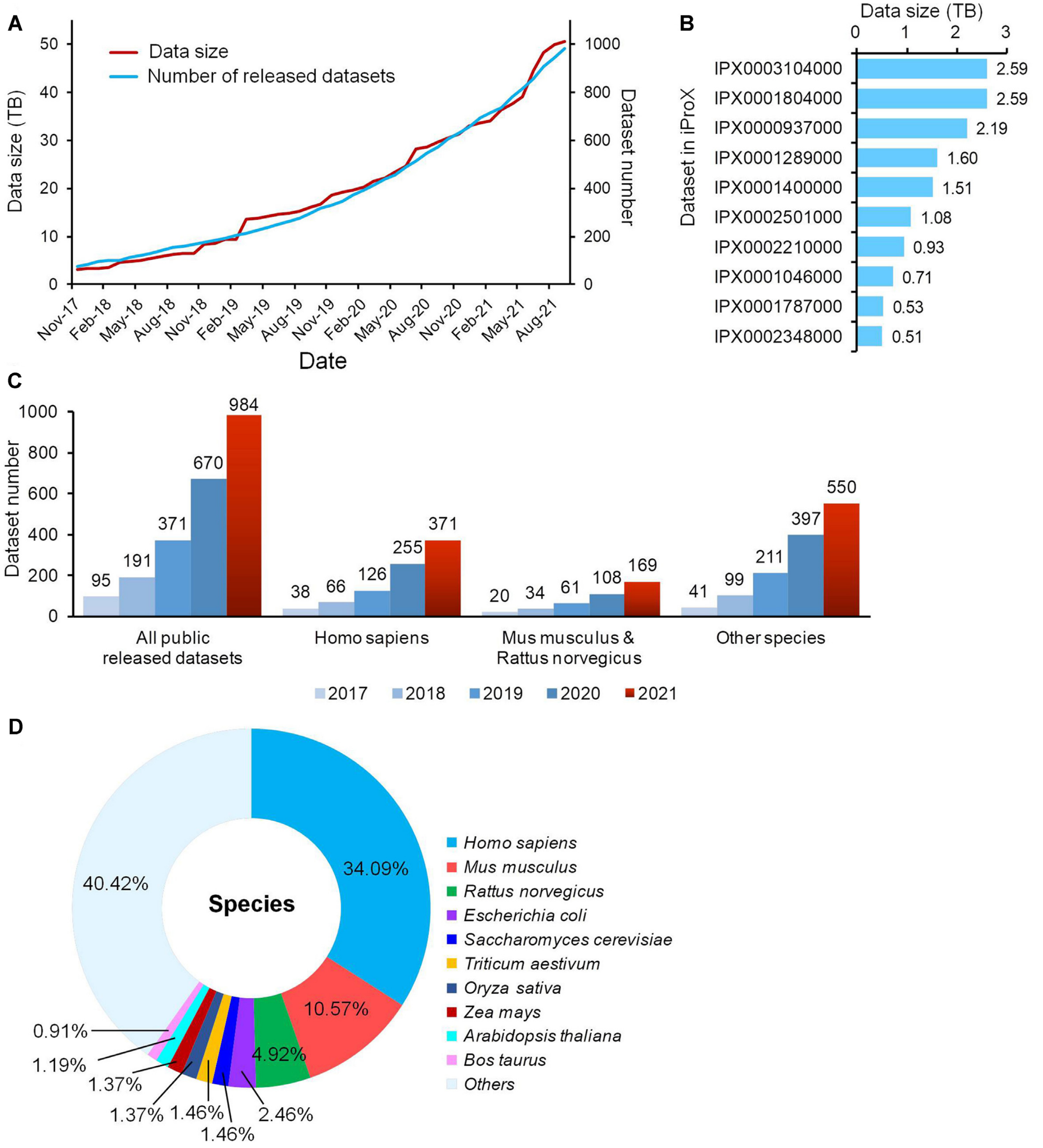


Figure 1. Summary of the datasets publicly released in iProX (as of the end of August 2021). (**A**) Cumulative data size and number of submitted datasets per month to (ranging from November 2017 to August 2021). (**B**) Top 10 released datasets with the largest size. (**C**) Cumulative numbers of submitted datasets per year. Some datasets in iProX were generated by the samples from multiple species, thus, the sum of the numbers of different species is a little higher than the number of all public datasets. (**D**) Distribution of the species of datasets publicly available in iProX.

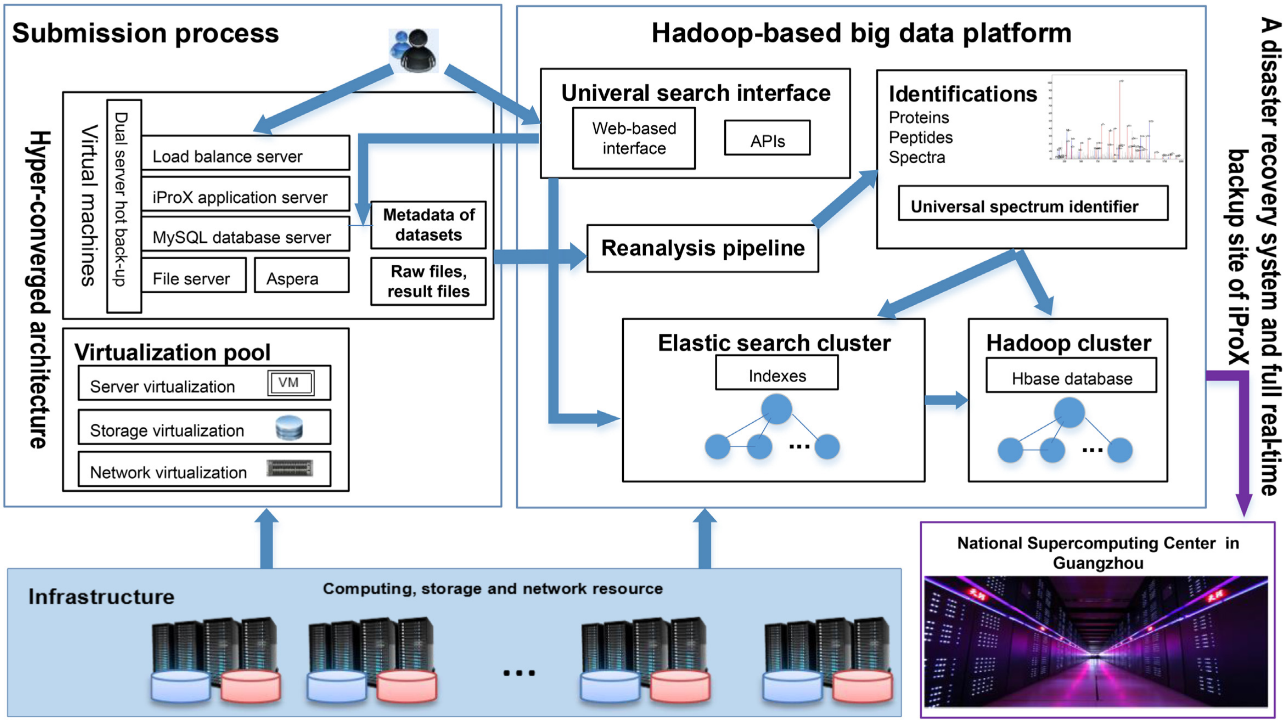
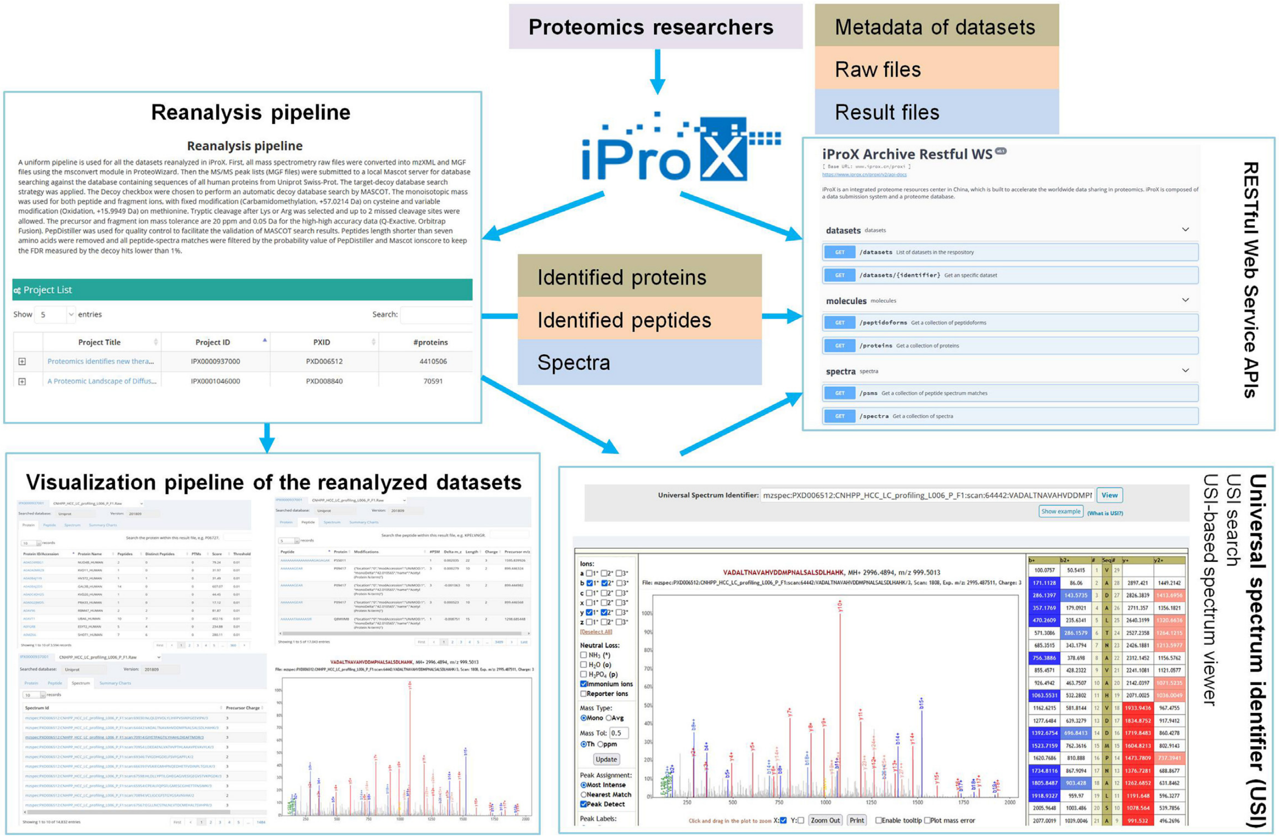


Figure 2. Hadoop-based big data architecture and infrastructure of iProX.



Reanalysis pipeline

Reanalysis pipeline

A uniform pipeline is used for all the datasets reanalyzed in iProX. First, all mass spectrometry raw files were converted into mzXML and MGF files using the msconvert module in ProteoWizard. Then the MS/MS peak lists (MGF files) were submitted to a local Mascot server for database searching against the database containing sequences of all human proteins from Uniprot Swiss-Prot. The target-decoy database search strategy was applied. The Decoy checkbox were chosen to perform an automatic decoy database search by MASCOT. The monoisotopic mass was used for both peptide and fragment ions, with fixed modification (Carbamidomethylation, +57.0214 Da) on cysteine and variable modification (Oxidation, +15.9949 Da) on methionine. Tryptic cleavage after Lys or Arg was selected and up to 2 missed cleavage sites were allowed. The precursor and fragment ion mass tolerance are 20 ppm and 0.05 Da for the high-high accuracy data (Q-Exactive, Orbitrap Fusion). PepDistiller was used for quality control to facilitate the validation of MASCOT search results. Peptides length shorter than seven amino acids were removed and all peptide-spectra matches were filtered by the probability value of PepDistiller and Mascot ionscore to keep the FDR measured by the decoy hits lower than 1%.

Project List

Show 5 entries Search:

	Project Title	Project ID	PXID	#proteins
⊞	Proteomics identifies new thera...	IPX0000937000	PXD006512	4410506
⊞	A Proteomic Landscape of Diffus...	IPX0001046000	PXD008840	70591

iProX Archive Restful WS

[Base URL: www.iprox.cn/proxi]

https://www.iprox.cn/proxi/v2/api-docs

iProX is an integrated proteome resources center in China, which is built to accelerate the worldwide data sharing in proteomics. iProX is composed of a data submission system and a proteome database.

datasets datasets

- GET /datasets List of datasets in the respository
- GET /datasets/{identifier} Get an specific dataset

molecules molecules

- GET /peptidoforms Get a collection of peptidoforms
- GET /proteins Get a collection of proteins

spectra spectra

- GET /psms Get a collection of peptide spectrum matches
- GET /spectra Get a collection of spectra

Universal Spectrum Identifier: mzspec:PXD006512:CNHPP_HCC_LC_profiling_L006_P_F1:scan:64442:VADALTNAVAHVDDMPI View

Show example (What is USI?)

VADALTNAVAHVDDMPNALSALSDLHAHK, MH+ 2996.4894, m/z 999.5013

Figure 3. New features implemented into iProX 2021.

this page by clicking the 'Resources' menu on the iProX main page and selecting the sub-menu 'Universal Spectrum Identifier (USI) Search'. Users paste a USI into a text box on the page and press the 'view' button to lookup the spectrum. The lookup result returned from the Elastic Search engine is visualized with an embedded Lorikeet spectrum viewer in the page. Instead, USI lookup service may also be triggered automatically by using the URL https://www.iprox.cn/page/spectrum.html?USI=⟨USI⟩. For example, the URL, https://www.iprox.cn/page/spectrum.html?USI = mzspec:PXD006512:CNHPP_HCC_LC_profiling_L006_P_F1:scan:64442:VADALTNAVAHVDDMPNALSALSDLHAHK/3, can be used to lookup the spectrum identified by the USI in the URL and visualize the spectrum, if available.

Along with the spectra identified by the USIs, the identified proteins and peptides are obtained by our designed reanalysis pipeline and stored in the HBase cluster. The results files of the 'complete submission' project in iProX are also parsed into spectra records and uniquely marked by USIs.

REANALYSIS AND VISUALIZATION PIPELINE OF PUBLIC DATA

Due to the unprecedented availability of proteomics data in the public domain, the need for data reuse is increasing (16,17). A high-efficiency reanalysis pipeline was built and applied to reuse the released data in iProX and identify evidence to analyze the publicly released datasets. This process generated millions of high-quality spectrum and identifications. The identified proteins are provided with UniProt accession numbers and associated URLs. At present, this reanalysis pipeline can handle the datasets generated from DDA (Data Dependent Acquisition) workflow. We applied the above reanalysis pipeline to the public dataset, IPX0000937000 (18), and derived 20 million new identifications at controlled false discovery rates. All of these identifications were parsed and stored into an HBase cluster. These reanalysis data are readily accessible at a new search interface based on the Elastic Search engine and can be traced back to the original datasets by IPX accession numbers. We will reanalyze all the public data, build a large-scale spectral library, and cross-reference MS identifications with other external datasets, such as UniProt.

Clicking the 'Resources' menu on the iProX main page and selecting the sub-menu 'Reanalyzed Datasets' reveals the list of the reanalyzed datasets with the number of identifications. By clicking the 'view' button along with the datasets, one can see the identification's visualization pipeline. The menu tabs 'Protein', 'Peptide,' and 'Spectrum' are linked to the details of the identifications for the chosen dataset. IPX accession numbers on the page can trace back to the metadata of the chosen dataset. Specifically, the link on the protein ID can be directed to all the identified peptides corresponding to the protein. Clicking 'Peptide' then links to the spectra related to the peptide. The link on the USI is directed to the visualization page for the spectrum. The 'Summary charts' menu links to the page with the statistics of these identifications.

IPROX RESTFUL WEB SERVICE API

Besides supporting human interactions to access the data, iProX provides a RESTful Web Service Application Programming Interface (API) presented by PX for automatically accessing proteomics results. It reports the metadata of datasets, or peptide, protein, and spectra data for reanalysis, including getting the metadata of a specific dataset or lists of datasets and collecting peptidoforms, proteins, and peptide spectrum matches (PSMs), or a list of spectra referred by USIs. These APIs are provided at https://www.iprox.cn/proxi/swagger-ui.html. Users can also access this page by clicking the 'Resources' menu on the main page and selecting the sub-menu 'Web Service APIs'. For example, 'https://www.iprox.cn/proxi/datasets/PXD008840' returns the json format of the details of the dataset, PXD008840, which can be used for reprogramming.

DISCUSSION AND FUTURE PLANS

The generation of tons of proteome data leads us towards the 'big data' era in proteomics. It is now easier than ever to produce large amounts of proteomics data. In a few days, a protein scientist can create a terabyte of data. In recent years, many datasets with continuously increasing data sizes have been submitted and deposited to iProX. There is an urgent need to bridge the connection between proteomics big data with its simple transfering, open sharing and efficient reuse in the scientific field. iProX can manage large and complex proteomics data by updating the software and hardware architecture using a big data platform, supporting PB-level data storage, hundreds of billions of spectra records, and second-level latency service capabilities.

Also, a high-efficiency reanalysis pipeline was built in iProX and used to analyze publicly available datasets and generate millions of high-quality spectra and identifications. Currently, we have primarily analyzed several large-scaled datasets [e.g. human proteome datasets generated from the Chinese Human Proteome Project (CNHPP)]. We will then transition to non-human organisms, particularly focusing on species with limited proteome information, in which MS-based proteomics data is mostly collected in iProX. These smaller datasets will provide useful resources for their specific fields of study.

At present, all PX resources are committed to completely open data, which means there are currently no limitations for data reuse by the community (19). One increasingly relevant topic is whether proteomics data from human samples can be used to identify individuals and whether proteomics data may be considered personally identifiable information (19,20). The privacy risks of proteomics data may emerge, and this issue will need to be properly assessed and managed. Currently, all PX resources have decided to phase the license for dataset sharing and reuse from a Creative Commons CC0 license as a minimum level initially and will likely move to CC-KY in the future (3). iProX has established its data license terms since it began in 2017, and we are moving towards the CC-BY license proposed by PX. Also, the management rule for human genetic resources in China was implemented in July 2019 and should be integrated into the improved data license for iProX.

In this study, with the implementation of a big data platform, iProX can support PB-level data storage, hundreds of billions of spectral records, and second-level latency service capabilities to meet the needs of the rapid growth proteomics field. iProX has an important role in facilitating the analysis and sharing of proteomics data worldwide.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The iProX team would like to thank all the data submitters and collaborators for their contributions, and the members of the PX consortium for their help and support.

FUNDING

The iProX team has been supported by the National Key Research Program of China [2020YFE0202200]; Innovation special zone [18-163-15-ZT-001-006-07]; Program for Guangdong Introducing Innovative and Entrepreneurial Teams [NO. 2016ZT06D211]; National Natural Science Foundation of China [U1811461]; Innovation project [16CXZ027]. Funding for open access charge: National Key Research Program of China [2020YFE0202200].
Conflict of interest statement. None declared.

REFERENCES

1. Marx,V. (2013) Biology: the big challenges of big data. *Nature*, **498**, 255–260.
2. Leonelli,S. (2019) The challenges of big data biology. *Elife*, **8**, e47381.
3. Deutsch,E.W., Bandeira,N., Sharma,V., Perez-Riverol,Y., Carver,J.J., Kundu,D.J., García-Seisdedos,D., Jarnuczak,A.F., Hewapathirana,S., Pullman,B.S. *et al.* (2020) The ProteomeXchange consortium in 2020: enabling 'big data' approaches in proteomics. *Nucleic Acids Res.*, **48**, D1145–D1152.
4. Vizcaíno,J.A., Deutsch,E.W., Wang,R., Csordas,A., Reisinger,F., Ríos,D., Dianes,J.A., Sun,Z., Farrah,T., Bandeira,N. *et al.* (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.*, **32**, 223–226.
5. Perez-Riverol,Y., Csordas,A., Bai,J., Bernal-Llinares,M., Hewapathirana,S., Kundu,D.J., Inuganti,A., Griss,J., Mayer,G., Eisenacher,M. *et al.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.
6. Farrah,T., Deutsch,E.W., Kreisberg,R., Sun,Z., Campbell,D.S., Mendoza,L., Kusebauch,U., Brusniak,M.Y., Hüttenhain,R., Schiess,R. *et al.* (2012) PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics*, **12**, 1170–1175.
7. Moriya,Y., Kawano,S., Okuda,S., Watanabe,Y., Matsumoto,M., Takami,T., Kobayashi,D., Yamanouchi,Y., Araki,N., Yoshizawa,A.C. *et al.* (2019) The jPOST environment: an integrated proteomics data repository and database. *Nucleic Acids Res.*, **47**, D1218–D1224.
8. Ma,J., Chen,T., Wu,S., Yang,C., Bai,M., Shu,K., Li,K., Zhang,G., Jin,Z., He,F. *et al.* (2019) iProX: an integrated proteome resource. *Nucleic Acids Res.*, **47**, D1211–D1217.
9. Sharma,V., Eckels,J., Schilling,B., Ludwig,C., Jaffe,J.D., MacCoss,M.J. and MacLean,B. (2018) Panorama Public: a public repository for quantitative data sets processed in skyline. *Mol. Cell. Proteomics*, **17**, 1239–1244.
10. Brenes,A., Afzal,V., Kent,R. and Lamond,A.I. (2018) The Encyclopedia of Proteome Dynamics: a big data ecosystem for (prote)omics. *Nucleic Acids Res.*, **46**, D1202–D1209.
11. Alnasir,J.J. and Shanahan,H.P. (2018) The application of Hadoop in structural bioinformatics. *Brief. Bioinform.*, **21**, 96–105.
12. Shah,N., Willick,D. and Mago,V. (2018) A framework for social media data analytics using Elasticsearch and Kibana. *Wireless Netw.*, 1–9.
13. Liu,J., Liu,Q., Zhang,L., Su,S. and Liu,Y. (2020) Enabling Massive XML-Based biological data management in HBase. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **17**, 1994–2004.
14. Schmidt,T., Samaras,P., Dorfer,V., Panse,C., Kockmann,T., Bichmann,L., van Puyvelde,B., Perez-Riverol,Y., Deutsch,E.W., Kuster,B. *et al.* (2021) Universal spectrum explorer: a standalone (web-)application for cross-resource spectrum comparison. *J. Proteome Res.*, **20**, 3388–3394.
15. Deutsch,E.W., Perez-Riverol,Y., Carver,J., Kawano,S., Mendoza,L., Van Den Bossche,T., Gabriels,R., Binz,P.A., Pullman,B., Sun,Z. *et al.* (2021) Universal Spectrum Identifier for mass spectra. *Nat. Methods*, **18**, 768–770.
16. Vaudel,M., Verheggen,K., Csordas,A., Raeder,H., Berven,F.S., Martens,L., Vizcaíno,J.A. and Barsnes,H. (2016) Exploring the potential of public proteomics data. *Proteomics*, **16**, 214–225.
17. Martens,L. and Vizcaíno,J.A. (2017) A golden age for working with public proteomics data. *Trends Biochem. Sci.*, **42**, 333–341.
18. Jiang,Y., Sun,A., Zhao,Y., Ying,W., Sun,H., Yang,X., Xing,B., Sun,W., Ren,L., Hu,B. *et al.* (2019) Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature*, **567**, 257–261.
19. Bandeira,N., Deutsch,E.W., Kohlbacher,O., Martens,L. and Vizcaíno,J.A. (2021) Data management of sensitive human proteomics data: current practices, recommendations, and perspectives for the future. *Mol. Cell. Proteomics*, **20**, 100071.
20. Mann,S.P., Treit,P.V., Geyer,P.E., Omenn,G.S. and Mann,M. (2021) Ethical principles, constraints and opportunities in clinical proteomics. *Mol. Cell. Proteomics*, **20**, 100046.