Published online 11 November 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D1523–D1528***
doi: 10.1093/nar/gkaa1012

GlycoPOST realizes FAIR principles for glycomics mass spectrometry data

Yu Watanabe[1], Kiyoko F. Aoki-Kinoshita[2], Yasushi Ishihama[3,4] and Shujiro Okuda[1,*]

[1]Division of Bioinformatics, Niigata University Graduate School of Medical and Dental Sciences, 1–757 Asahimachi-dori, Chuo-ku, Niigata 951–8510, Japan, [2]Faculty of Science and Engineering, Soka University, 1–236 Tangi-machi, Hachioji, Tokyo 192-8577, Japan, [3]Department of Molecular and Cellular BioAnalysis, Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan and [4]Department of Proteomics and Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

Received August 13, 2020; Revised October 13, 2020; Editorial Decision October 13, 2020; Accepted October 14, 2020

ABSTRACT

For the reproducibility and sustainability of scientific research, FAIRness (Findable, Accessible, Interoperable and Re-usable), with respect to the release of raw data obtained by researchers, is one of the most important principles underpinning the future of open science. In genomics and transcriptomics, the sharing of raw data from next-generation sequencers is made possible through public repositories. In addition, in proteomics, the deposition of raw data from mass spectrometry (MS) experiments into repositories is becoming standardized. However, a standard repository for such MS data had not yet been established in glycomics. With the increasing number of glycomics MS data, therefore, we have developed GlycoPOST (https://glycopost.glycosmos.org/), a repository for raw MS data generated from glycomics experiments. In just the first year since the release of GlycoPOST, 73 projects have already been registered by researchers around the world, and the number of registered projects is continuously growing, making a significant contribution to the future FAIRness of the glycomics field. GlycoPOST is a free resource to the community and accepts (and will continue to accept in the future) raw data regardless of vendor-specific formats.

INTRODUCTION

For reproducibility and sustainability of scientific research, the public sharing of raw data obtained by researchers is of paramount significance (1). The FAIRness (Findable, Accessible, Interoperable and Re-usable) of datasets is the most important principle that supports open science in the future (1–3). For genomics, the sharing of raw data from next generation sequencers (NGS) is implemented through public repositories (4). In addition, registration of gene expression data such as RNAseq into data repositories is becoming standardized (5,6). Furthermore, mass spectrometry (MS) has become the method of choice for the qualitative and quantitative characterization of complex protein and glycan mixtures (7–11), and thus a need for a repository for sharing such data has been recognized. For data in the field of proteomics, qualitative and quantitative mass spectrometry-based analyses are performed and reported. These studies may characterize relatively simple systems, such as protein complexes or much more complex mixtures, such as cell organelles, full cell lysates or different organs. Thus standards such as for data processing, common data formats and issuance of common accession numbers for submitting raw data to repositories are being promoted under the global activity called the ProteomeXchange (PX) consortium (12). In the proteome field, there are several repositories approved by this PX Consortium all over the world (13–16), and each of them operates its own repository according to their respective region and specific data format. As a result, all proteome MS data are accessible from the ProteomeCentral portal site managed by the PX Consortium, where over 20,000 projects are currently registered (17).

Proteomics analysis may also include the characterization of post-translational modifications (PTMs) including glycosylation. However, in most cases such PTMs are simply added in the annotations as text. With the recent development of a glycan structure repository GlyTouCan (18), this information should also be linked with glycan and glycomics data. Therefore, in the glycomics field, the Minimum Information Required for A Glycomics Experiment (MIRAGE) initiative began with the recommendation of minimum information required to be reported when publicizing glycomics experiments (19). MIRAGE standards for 'minimum information required for a glycomics experiment' and proposes guidelines for many of the experimental techniques used when working with glycans. The first of these

*To whom correspondence should be addressed. Tel: +81 25 227 0390; Fax: +81 25 227 0393; Email: okd@med.niigata-u.ac.jp

guidelines was for MS experiments, where the minimum information needed to be reported was delineated including the type of instrument used, its parameters, peak lists with characterized structures and raw data. This guideline helped standardize the metadata required for the registration of MS data in a repository. UniCarb-DR was recently announced as a repository for characterized glycans by MS, storing peak list information, and GlycoPOST was mentioned as the raw data repository (20). In this manuscript, we describe the details on the usage of GlycoPOST.

We believe that there is an urgent need for an official repository for glycomics MS raw data, so we have developed and since operated a repository called GlycoPOST. GlycoPOST has been made available for over a year so far, and through our efforts to approach the glycomics community, 50 users have registered, >70 projects have been created, and over 2000 files have now been deposited in GlycoPOST, totalling 700 GB of data. As the use of MS in the glycoscience field is expected to grow further in the future, the number of projects registered with GlycoPOST is very likely to increase.

DATABASE DESCRIPTION

GlycoPOST accepts MS data from glycomics experiments and issues an accession number to provide traceability for reuse and reanalysis of the data. This system is an adaptation of the jPOST repository system (14), which has already proven to be a stable MS data repository for proteomics. Basically, the technology implemented in the jPOST repository has been implemented in GlycoPOST as well, and it inherits the usability of the jPOST repository. In addition, the GlycoPOST system has been designed to make it easy to input various metadata such as experimental conditions and instrument settings specific to glycomics. Metadata such as experimental conditions are set to comply with the MIRAGE guidelines, and thus we can claim that GlycoPOST contributes to standardization in the glycomics field (Figure 1). As illustrated in this figure, GlycoPOST is a part of the GlyCosmos portal (21), which also includes UniCarb-DR and GlyTouCan (18) as fellow repository systems. GlyTouCan is the international glycan structure repository, and it assigns accession numbers to individual glycans. UniCarb-DR is a repository of peak lists, and the raw data is registered in GlycoPOST. Due to this relationship between UniCarb-DR and GlycoPOST, we have implemented a combined user registration system that handles user information for both repositories.

MIRAGE guidelines

MIRAGE (Minimum Information Required for A Glycomics Experiment) is a set of guidelines established by the MIRAGE committee to specify the minimum information required for reporting glycan-related experiments, such as sample preparation (doi:10.3762/mirage.1), mass spectrometry (doi:10.3762/mirage.2), glycan arrays (doi:10.3762/mirage.3), and liquid chromatography (doi:10.3762/mirage.4) (19). MIRAGE is supported by the Beilstein Institute in Germany, and the MIRAGE committee is made up of renowned glycomics scientists and glyco-informaticians from around the world.

GlycoPOST has adopted the guidelines for the portion of MIRAGE that is relevant to mass spectrometry experiments for glycomics. To make it easier for users to enter and manage metadata, the input section for metadata has been divided into the following five sections, 'Sample preparation', 'General features', 'Ion sources', 'Ion transfer optics', and 'Spectrum and peak list generation and annotation', each of which can be registered in GlycoPOST as a reusable 'preset'. As long as the experimental conditions are the same, the user can use a previously created preset as is, or they can change some parts of it and create another preset. Note that the content of each of the following presets are all based on the current version of the MIRAGE guidelines and are subject to change based on any updates to these guidelines. The latest version of the details of this information is available at https://glycopost.glycosmos.org/help#mirage.

Preset 1: Sample preparation

The sample preparation section is designed to include all aspects of sample generation, purification and modifications of the biological and/or synthetic material analyzed. Users input biologically derived material and/or chemically derived material as sample origin, and enzymatic and/or chemical treatments as sample processing for isolation. In addition, enzymatic and/or chemical modifications, and purification steps are needed to be registered.

Preset 2: General features

In this preset, global descriptions such as the used instrumentation, any particular customizations, and general instrument control parameters such as instrument control software. This includes the software name and version information.

Preset 3: Ion sources

This preset is used for summarizing all the parameters for ion generation including controls of in-source fragmentation, the degree of fragmentation, as well as other more common parameters such as capillary voltage or laser intensity settings.

Preset 4: Ion transfer optics

This preset requires instrumental details related to the processes after ions are generated such as transport, gas phase reactions and detection of ions.

Preset 5: Spectrum and peak list generation and annotation

The software used to generate peak list files from mass spectrometry raw data files and software and/or databases used to annotate each spectrum are needed to be input. This category is optional because it is often not possible to obtain this data.

In addition, UniCarb-DR (https://unicarb-dr.glycosmos.org/) provides a web tool that allows users to enter MIRAGE-related information for their experiments, which

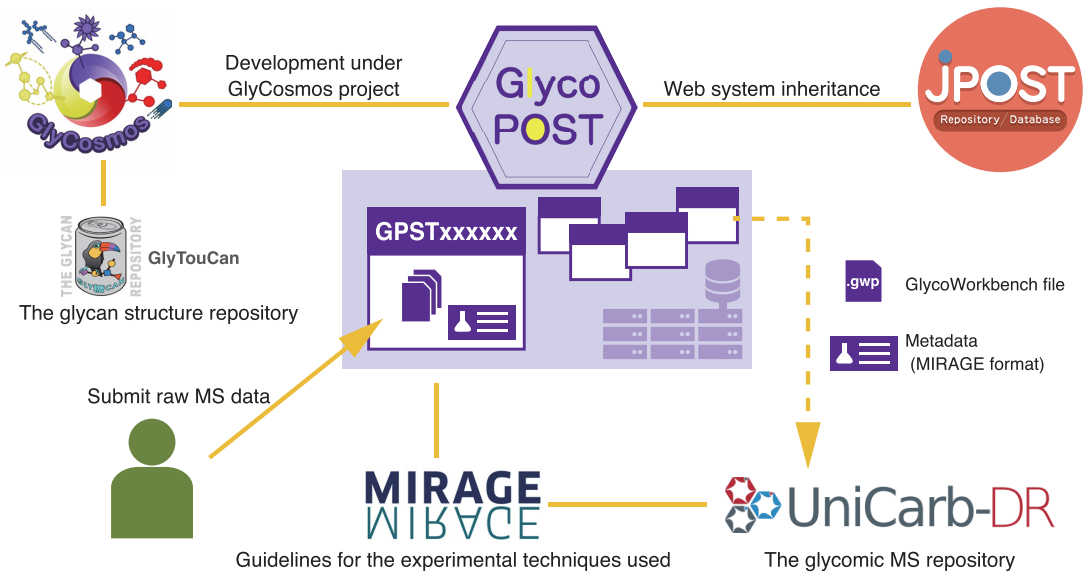

Figure 1. Schematic representation of the GlycoPOST environment. GlycoPOST has been developed under the GlyCosmos project, and the GlycoPOST system was adapted from the repository system of the jPOST project. The metadata to be registered follows the MIRAGE guidelines and the Excel format for the metadata input used by UniCarb-DR is also importable.

produces an Excel file formatted in a specific format with the required information (20). GlycoPOST has a function to import the data from this Excel spreadsheet and automatically create presets. Conversely, users can also export the preset data from GlycoPOST and download an Excel spreadsheet in the same format.

Thus, we made efforts to ensure compatibility with other glycomics data repositories. Furthermore, by adopting the MIRAGE guidelines, we can ensure the quality of the metadata registered in GlycoPOST and that it is compatible with databases in related fields.

Project creation and file upload

In general, users would register their data as a single project, which receives a unique accession number. Each project can contain one or more raw data and must be linked with metadata as defined by the presets described previously. Each project is required to be linked with at least Presets 1–4 for describing samples, experiments, and instruments, and once registered, and one accession number will be issued. After a project is generated, any metadata for sample preparation, general features, ion sources, ion transfer, spectrum and peak list registered as a preset will be linked to the MS raw data files to be registered (Figure 2). The same metadata information can be linked to all files at once by drag-and-drop of the raw data files into the browser with presets selected beforehand, greatly reducing the registration operation for users with many files to register.

After linking the metadata profiles as presets with the deposited data files, the user can upload the files to the repository. GlycoPOST utilizes the PRESTO system (https://prestotools.github.io/) for uploading data. The upload process of this system splits the file into smaller pieces, called 'chunks', which are then uploaded to the repository in parallel. In the process of data transmission over the Internet, it is known that the longer the distance of the data communication route, the greater the delay (the delay before the data actually starts to be transferred). This often results in very slow data transfer rates between physically distant locations. This delay problem can be remedied by uploading small chunks in parallel.

This data transfer system is already implemented in the jPOST repository, and it has shown to have a positive correlation between file size and transfer time, with an average transfer rate of about 5MB/s, which is fast enough to take only about four minutes to upload a 1GB file (14). Thus, it has been found that the file transfer speed is, in most cases, independent of the distance from where the user deposits the data in this system.

Data publication and download

When the deposited dataset is determined to be valid against the MIRAGE guideline criteria, the users can lock and exit the submission process, at which time a GlycoPOST identifier is generated as an accession number. The submitted data and metadata are automatically checked by our system and if the dataset is determined to be incomplete, an accession number will not be assigned and the dataset cannot be announced. At this stage of a 'project' submission, datasets submitted to the repository are in 'embargo', meaning it is set as private, and it will be automatically published on a 'publication date' set by the users themselves. During this embargo period, users will be issued a dedicated URL and password that will allow anyone with this information, such as journal editors and peer reviewers, to access the project. Users can also revise a temporarily locked project in response to reviewers' comments and revised data will be assigned a revision number.

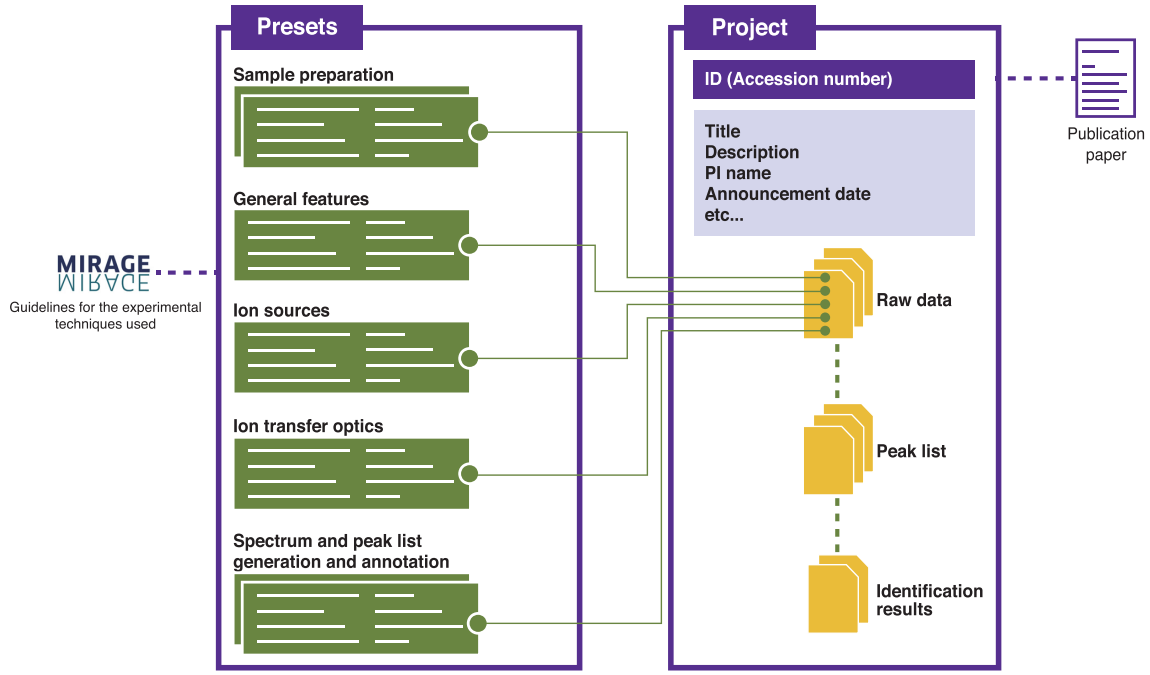

Figure 2. Relationships between data files and metadata. Presets that follow the MIRAGE guidelines are registered in GlycoPOST, and each preset is linked to raw data files obtained from mass spectrometry. In addition to the raw data, a project is created that contains the peak list and result files, and an accession number is issued for that project.

Datasets of published projects can be downloaded without restrictions. Users can also search for keywords found in any of the fields registered under presets or projects.

System implementation

The web application for GlycoPOST was built using the React framework (https://reactjs.org/), and the proprietary PRESTO system is used for file uploads. This eliminates the need for FTP and external software for file uploads, allowing the entire process from project creation to file uploads to be completed within a single web browser, contributing to an improved user experience.

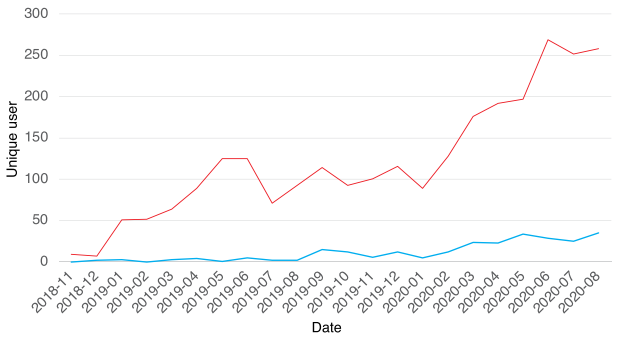

Figure 3. History of user access. Red line indicates the number of unique users accessed to the top page of GlycoPOST and blue line indicates the number of unique users downloaded the data stored in GlycoPOST.

DISCUSSION

The alpha version of GlycoPOST was launched in December 2018, with the beta version released in April 2019, and its official release in March 2020. During this time, it has been used by many users, with over 70 projects registered, of which >20 are in the public domain. Over half of the registered projects are based on ESI-MS/MS analysis, but others include ESI-MS, MALDI-MS and MALDI-MS/MS, which are the instrumentation listed by the MIRAGE guidelines. However, other technologies can be selected under 'Not specified' for the time being. As more data using these other technologies are deposited, they will be added to the predefined list. The numbers of datasets deposited using positive and negative mode were about half and half. Regarding glycan labeling and derivatization, currently there is no controlled vocabulary, so users have entered free text to describe this under the sample processing section. Those who used glycan labeling will have specified this information, but unlabeled glycans would not be mentioned. All the required metadata have been specified by all users since the official release of GlycoPOST in April, 2020. The number of accesses and downloads have increased in general, as shown in Figure 3. Although the server itself is located in Japan, it has attracted a lot of attention as it is used by researchers worldwide not only Asia but also North America and Europe. We assume that this is due in large part to the need that it fulfills and its usability.

All metadata submitted to this repository, especially the experimental procedures described in the current five presets, are not currently represented as any ontology or controlled vocabulary. This is due to the fact that although the MIRAGE guidelines exist, no repository could fully ac-

commodate the guidelines until now. Because many other glycomics-related databases already use ontologies to represent glycan-related information (18,21), by increasing the use of ontologies in GlycoPOST in the future and expressing them in a unified framework such as the Resource Description Framework (RDF) data model, it should become possible to integrate the data in GlycoPOST with other glycan-related databases (2,3,22). Moreover, MIRAGE has yet to publish a glycoproteomics guideline, but there are plans to make one available soon in collaboration with the HUPO Proteomics Standards Initiative (PSI) (23). As soon as these guidelines are complete and announced, we plan on implementing functionality to accept the relevant metadata in GlycoPOST to be able to accept glycoproteomics data. This will prepare us to apply to the ProteomeXchange as a fellow member. The MIRAGE guidelines will delineate the metadata for glycan structure information, which will be linked with GlyTouCan and other related glycan resources; this is currently lacking in proteomics repositories.

Currently, UniCarb-DR and GlycoPOST are independent systems except for user information. In the near future, the data between these repositories will be shared, so that the raw data registered in GlycoPOST can be mapped to the peak lists and glycans registered in UniCarb-DR, and vice-versa. Moreover, these glycan data will be linked to GlyTouCan accession numbers. As a result, the raw data in GlycoPOST can be visualized with the spectra of the glycan fragments registered in UniCarb-DR. Moreover, this workflow can be made more seamless by having users first register the raw data, peak lists and glycan data in GlycoPOST, which then automatically registers the glycan data into UniCarb-DR to take advantage of the latter's connection with GlyTouCan. Then, by integrating all this information under a common framework, glycans can be searched throughout UniCarb-DR and GlycoPOST in the future. This will make re-analysis of the glycomics MS data in GlycoPOST easier, by allowing users to search for raw data containing a particular glycan.

Making data findable, accessible, interoperable and reusable will not only enhance its value as a public asset, but also contribute to the many studies that will help us create new value. In the field of glycomics, this concept of FAIRness is an important common philosophy that needs to be realized in the same way as in genomics and proteomics. This will allow for re-analysis of the data as detection algorithms and technologies improve, especially considering that de novo analysis is currently extremely difficult. Moreover, by working with journals to require the submission of raw data to a repository, the data will be more accessible for other users, where metadata will aid in searching for the most appropriate datasets and the guidelines ensure that the data is accessible in a standard format. We believe that GlycoPOST will be a major contributor to these public roles.

ACKNOWLEDGEMENTS

The GlycoPOST team would like to thank all the data submitters and collaborators for their contributions, the members of the carbohydrate research community and the jPOST team for their support.

FUNDING

Database Integration Coordination Program of the National Bioscience Database Center (NBDC), Japan Science and Technology Agency [17934031, 18063028]. Funding for open access charge: Database Integration Coordination Program of the National Bioscience Database Center (NBDC), Japan Science and Technology Agency.
Conflict of interest statement. None declared.

This paper is linked to: doi:10.1093/nar/gkaa947.

REFERENCES

1. Wilkinson,M.D., Dumontier,M., Aalbersberg,Ij.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.W., da Silva Santos,L.B., Bourne,P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
2. Katayama,T., Wilkinson,M.D., Vos,R., Kawashima,T., Kawashima,S., Nakao,M., Yamamoto,Y., Chun,H.-W., Yamaguchi,A., Kawano,S. *et al.* (2011) The 2nd DBCLS BioHackathon: Interoperable bioinformatics Web services for integrated applications. *J. Biomed. Semantics*, **2**, 4.
3. Katayama,T., Wilkinson,M.D.M.D., Aoki-Kinoshita,K.F.K.F., Kawashima,S., Yamamoto,Y., Yamaguchi,A., Okamoto,S., Kawano,S., Kim,J.-D.J.-D., Wang,Y. *et al.* (2014) BioHackathon series in 2011 and 2012: penetration of ontology and linked data in life science domains. *J. Biomed. Semantics*, **5**, 5.
4. Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2016) GenBank. *Nucleic. Acids. Res.*, **44**, D67–D72.
5. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
6. Kolesnikov,N., Hastings,E., Keays,M., Melnichuk,O., Tang,Y.A., Williams,E., Dylag,M., Kurbatova,N., Brandizi,M., Burdett,T. *et al.* (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113–D1116.
7. Shukla,A. (2017) In: *Proteomics in Biology, Part A, Volume 585 - 1st Edition*. Academic Press.
8. Shukla,A. (2017) In: *Proteomics in Biology, Part B, Volume 586 - 1st Edition*. Academic Press.
9. Chen,Z., Huang,J. and Li,L. (2019) Recent advances in mass spectrometry (MS)-based glycoproteomics in complex biological samples. *TrAC Trends Anal. Chem.*, **118**, 880–892.
10. Yang,X. and Bartlett,M.G. (2019) Glycan analysis for protein therapeutics. *J. Chromatogr. B*, **1120**, 29–40.
11. Li,Q., Xie,Y., Wong,M. and Lebrilla,C. (2019) Characterization of Cell Glycocalyx with Mass Spectrometry Methods. *Cells*, **8**, 882.
12. 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.
13. Vizcaíno,J.A., Csordas,A., Del-Toro,N., Dianes,J.A., Griss,J., Lavidas,I., Mayer,G., Perez-Riverol,Y., Reisinger,F., Ternent,T. *et al.* (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.*, **44**, D447–D456.
14. Okuda,S., Watanabe,Y., Moriya,Y., Kawano,S., Yamamoto,T., Matsumoto,M., Takami,T., Kobayashi,D., Araki,N., Yoshizawa,A.C. *et al.* (2017) JPOSTrepo: an international standard data repository for proteomes. *Nucleic. Acids. Res.*, **45**, D1107–D1111.
15. Kusebauch,U., Deutsch,E.W., Campbell,D.S., Sun,Z., Farrah,T. and Moritz,R.L. (2014) Using PeptideAtlas, SRMAtlas, and PASSEL: comprehensive resources for discovery and targeted proteomics. In: *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ. Vol. **46**, pp. 13.25.1–13.25.28.
16. 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.

17. 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.
18. Aoki-Kinoshita,K., Agravat,S., Aoki,N.P., Arpinar,S., Cummings,R.D., Fujita,A., Fujita,N., Hart,G.M., Haslam,S.M., Kawasaki,T. *et al.* (2016) GlyTouCan 1.0 - the international glycan structure repository. *Nucleic. Acids. Res.*, **44**, D1237–D1242.
19. York,W.S., Agravat,S., Aoki-Kinoshita,K.F., McBride,R., Campbell,M.P., Costello,C.E., Dell,A., Feizi,T., Haslam,S.M., Karlsson,N. *et al.* (2014) MIRAGE: the minimum information required for a glycomics experiment. *Glycobiology*, **24**, 402–406.
20. Rojas-Macias,M.A., Mariethoz,J., Andersson,P., Jin,C., Venkatakrishnan,V., Aoki,N.P., Shinmachi,D., Ashwood,C., Madunic,K., Zhang,T. *et al.* (2019) Towards a standardized bioinformatics infrastructure for N- and O-glycomics. *Nat. Commun.*, **10**, 3275.
21. Yamada,I., Shiota,M., Shinmachi,D., Ono,T., Tsuchiya,S., Hosoda,M., Fujita,A., Aoki,N.P., Watanabe,Y., Fujita,N. *et al.* (2020) The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences. *Nat. Methods*, **17**, 649–650.
22. Aoki-Kinoshita,K.F., Bolleman,J., Campbell,M.P., Kawano,S., Kim,J.-D., Lütteke,T., Matsubara,M., Okuda,S., Ranzinger,R., Sawaki,H. *et al.* (2013) Introducing glycomics data into the semantic web. *J. Biomed. Semantics*, **4**, 39.
23. Deutsch,E.W., Orchard,S., Binz,P.-A., Bittremieux,W., Eisenacher,M., Hermjakob,H., Kawano,S., Lam,H., Mayer,G., Menschaert,G. *et al.* (2017) Proteomics standards initiative: fifteen years of progress and future work. *J. Proteome Res.*, **16**, 4288–4298.