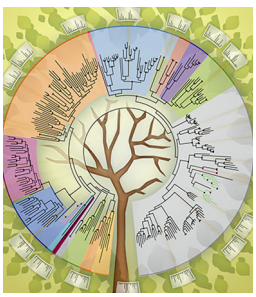
J. Am. Soc. Mass Spectrom. (2016) 27:1752–1757
DOI: 10.1007/s13361-016-1448-3

FOCUS: BIOINFORMATICS, SOFTWARE, AND MS-BASED "OMICS": RESEARCH ARTICLE

A Skyline Plugin for Pathway-Centric Data Browsing

Michael G. Degan,[1] Lillian Ryadinskiy,[1] Grant M. Fujimoto,[1] Christopher S. Wilkins,[1] Cheryl F. Lichti,[2,3] Samuel H. Payne[1]

[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99354, USA
[2]Department of Pharmacology and Toxicology, The University of Texas Medical Branch, Galveston, TX 77555, USA
[3]Mitchell Center for Neurodegenerative Diseases, The University of Texas Medical Branch, Galveston, TX 77555, USA



Abstract. For targeted proteomics to be broadly adopted in biological laboratories as a routine experimental protocol, wet-bench biologists must be able to approach selected reaction monitoring (SRM) and parallel reaction monitoring (PRM) assay design in the same way they approach biological experimental design. Most often, biological hypotheses are envisioned in a set of protein interactions, networks, and pathways. We present a plugin for the popular Skyline tool that presents public mass spectrometry data in a pathway-centric view to assist users in browsing available data and determining how to design quantitative experiments. Selected proteins and their underlying mass spectra are imported to Skyline for further assay design (transition selection). The same plugin can be used for hypothesis-driven data-independent acquisition (DIA) data analysis, again utilizing the pathway view to help narrow down the set of proteins that will be investigated. The plugin is backed by the Pacific Northwest National Laboratory (PNNL) Biodiversity Library, a corpus of 3 million peptides from >100 organisms, and the draft human proteome. Users can upload personal data to the plugin to use the pathway navigation prior to importing their own data into Skyline.
Keywords: Targeted proteomics, Bioinformatics, Systems biology, Data analysis

Received: 25 April 2016/Revised: 8 July 2016/Accepted: 9 July 2016/Published Online: 16 August 2016

Introduction

Targeted proteomics experiments focus on the measurement of an explicit subset of peptides instead of the broader measurements employed in global proteomics experiments [1]. The most common method for targeted proteomics is selected reaction monitoring (SRM), where peptides are detected and measured via the precursor *m/z* and a limited set of fragment ions. Parallel reaction monitoring (PRM), a high resolution analog of SRM in which all fragment ions are measured simultaneously, has recently become popular. Unlike SRM, fragment ions used for quantification are chosen post hoc, based upon relative intensity and lack of interferences. Designing a targeted proteomics experiment, the process of identifying the most responsive peptides and transitions for a given protein and sample of interest is time-consuming, and numerous publications have reported important considerations for assay design [2–4]. Most of these methods use previously identified peptide/spectrum matches, both to determine which peptides are most reliably observed for a given protein and to identify highly abundant fragment ion peaks. Although several large repositories store tandem mass spectrometry data for proteomics [5–7], they are typically built from a mass spectrometry perspective, without higher-order biological context.

Data-independent acquisition (DIA) is an untargeted data acquisition method [8]. For DIA, a MS scan for the entire mass range of interest is followed by a series of MS/MS scans that fragment all peptides within a specific *m/z* window (typically 10–25 Da). Successive MS/MS spectra tile consecutive *m/z* windows to cover the entire range. Thus, all ions are fragmented and measured. To analyze the data, a spectral library is leveraged to generate extracted ion chromatograms for *m/z* values corresponding to parent and fragment ions of expected peptides. Therefore, DIA data files provide a digital fingerprint for samples, allowing for repeated hypothesis-driven querying of data without requiring reacquisition of MS data.

Skyline is a robust and powerful software tool for analyzing mass spectrometry data. Although originally designed for targeted proteomics measurements [9], it has expanded to become a general tool box for SRM, PRM

Correspondence to: Samuel H. Payne; *e-mail:* samuel.payne@pnnl.gov

[10], data dependent acquisition (DDA) [11], and DIA [12, 13] mass spectrometry data. An important feature of Skyline is the intuitive visualization of results to assist users in manual data curation. These key strengths have led to the broad adoption of Skyline within the community. In both SRM/PRM assay design and DIA data analysis, users benefit greatly from working in Skyline with a spectral library. Currently, however, users manually enter peptide and protein data and upload a spectral library. The PNNL Biodiversity Plugin provides an intuitive interface for uploading data in a pathway-oriented interface. In addition to the vast spectral libraries for >100 organisms, users can point the plugin to additional libraries and use the same pathway navigation to push their own personal data into Skyline.

Experimental

Methods

Software and Availability The PNNL Biodiversity Plugin is programmed in C# using .NET 4.5.2 framework. The software is open source and is available at our group's GitHub account, https://github.com/PNNL-Comp-Mass-Spec/PNNL-Biodiversity-Library-Plugin. The plugin executable is available through Skyline as an external tool and can be found through the tool store or online at https://skyline.gs.washington.edu/. Due to changes in the Skyline API for external tools, the PNNL Biodiversity Library plugin does not work with Windows XP. The version of Skyline used with the plugin must be version 3.1.1.7490 or greater.

Kyoto Encyclopedia of Genes and Genomes (KEGG) Mapping The mapping of protein identifications from the PNNL Biodiversity Library and the Draft Human Proteome are as previously described [14]. This mapping is kept within the plugin as a SQLite file noting observed proteins. Mapping of user uploaded data is done as follows. The Uniprot identifiers present in the Bibliospec library uploaded by users can be easily indexed for uploads of type 'supplement' or 'replace' by updating the SQLite file. For adding a new organism, we first obtain all the genes for a specific organism using KEGG's REST API (http://rest.kegg.jp/link/ko/<org_code>). We then update the SQLite tables with this information and the protein identifiers from the Bibliospec library.

Spectral Data All spectral data presented by the plugin and transferred to Skyline is stored in Bibliospec libraries. There are two sources of data: the PNNL Biodiversity Library and the Pandey Lab Draft Human Proteome. All PNNL data were aggregated and searched as described [14], and are posted at the UCSD server MassIVE, http://massive.ucsd.edu/, ProteomeXchange identifier PXD001860, MassIVE identifier MSV000079053. Data for each organism is stored separately and contain raw, mzML, mzIdentML, and Bibliospec formats.

The human data available in the plugin is the Pandey Lab Draft Human Proteome. We downloaded the raw spectral files from ProteomeXhange (data set identifier PXD000561). These RAW files were converted to mzML and searched with MSGF+ using the following parameters: semi-tryptic protease specificity, 10 ppm parent ion tolerance, HCD fragmentation type, cysteine alkylation as a static modification, and oxidized methionine as a dynamic modification. The database for the search was the human reference proteome from Uniprot downloaded on 05_20_2015. The resulting .mzIdentML files were converted to the Bibliospec library format using the following command line parameters:

```
~>BlibBuild.exe –c 0.9999 C:\path_to\file1.mzid Library.blib
~>BlibBuild.exe –c 0.9999 C:\path_to\file2.mzid Library.blib
#as many BlibBuild commands as there are identification files
~>BlibFilter.exe –b 1 Library.blib FinalLibrary.blib
```

Because of the interest in proteomes of specific human tissues, we have included with the plugin a separate bibliospec library for each of the 30 tissues studied in the Pandey Lab manuscript. In addition, we have created a single library for 'Homo sapiens all tissues'.

User uploaded data is taken as a Bibliospec file. It is important to use the version of Bibliospec that embeds protein identifications within the file. We suggest the following command line for file creation:

```
~>BlibBuild.exe –c 0.9999 C:\path_to\file1.mzid Library.blib
~>BlibBuild.exe –c 0.9999 C:\path_to\file2.mzid Library.blib
#as many BlibBuild commands as there are identification files
~>BlibFilter.exe –b 1 Library.blib FinalLibrary.blib
```

The plugin does no additional filtering of this file. While data from the PNNL Biodiversity Library have been filtered to $q < 0.0001$, users are free to upload any quality of data and are responsible for understanding the limitations of their data quality. The plugin only cross-references protein identifications with KEGG orthologs for convenient mapping on the interface.

FASTA File Creation The Skyline program organizes peptides around their parent protein, as specified in a protein sequence. To obtain a user's specific subset of proteins, we dynamically create a FASTA file via a web API from uniprot.org. This customized FASTA file is transferred to Skyline utilizing the Tool Client object API.

Proteotypic Peptide Viewer This software program is also available at our group's GitHub account and at the Skyline tool store. After users have chosen proteins and organisms, we dynamically create a FASTA file as described above

and use the EBI web-API for the MUSCLE alignment program. Peptides are then mapped onto the aligned sequences and displayed.

Results and Discussion

The primary goal of the Biodiversity plugin is to help users get MS/MS data into Skyline for review and use during targeted proteomics assay design or hypothesis-driven DIA data analysis. There are numerous public resources with proteomics data, but they are not indexed in an easily navigable interface. To provide a more intuitive navigation, we created a pathway-oriented interface that allows users to think first and foremost about the biological question at hand and be presented with relevant resources. We seeded the plugin with data from the PNNL Biodiversity Library, containing 3 million peptides from 230,000 proteins in 112 bacteria and archaea [14]. Numerous biomedical and environmental organisms are contained in the library, including most model microbial organisms. Additionally, we included the draft human proteome, which has spectra collected from 30 different tissues [15]. Thus, the plugin comes prepackaged with an incredible breadth of publicly available data.

Pathway-Centric Data Browsing

The plugin walks users through data selection in a wizard to promote intuitive and efficient browsing (Figure 1). The plugin is designed around the concept of biological pathways and networks. Since most biological investigations focus on cellular functions that are typically described and categorized as pathways, it is a natural entry point for researchers when they begin to set up their assay. The first step is to select an organism. The second tab assists users in their selection of pathway(s) of interest. Pathways are listed in categories as organized by KEGG, which contains hundreds of pathways annotated across thousands of organisms, including pathways for metabolism, signaling, regulation, and disease-related processes [16]. The pathway coverage, calculated as the number of proteins in a pathway for which proteomics data is present in the Library, is provided alongside the name to help indicate the depth of proteomics data for a specific pathway.

The pathway visualization tab puts the peptide and protein data on top of KEGG pathway maps. This dynamically created image quickly conveys the proteomic coverage of specific proteins in the pathway (Figure 2). The goal of the plugin is to rapidly get users to this view, where they can determine the extent of proteome coverage for their biological question of choice. Here users can deselect individual proteins in the pathway image to remove them from consideration.

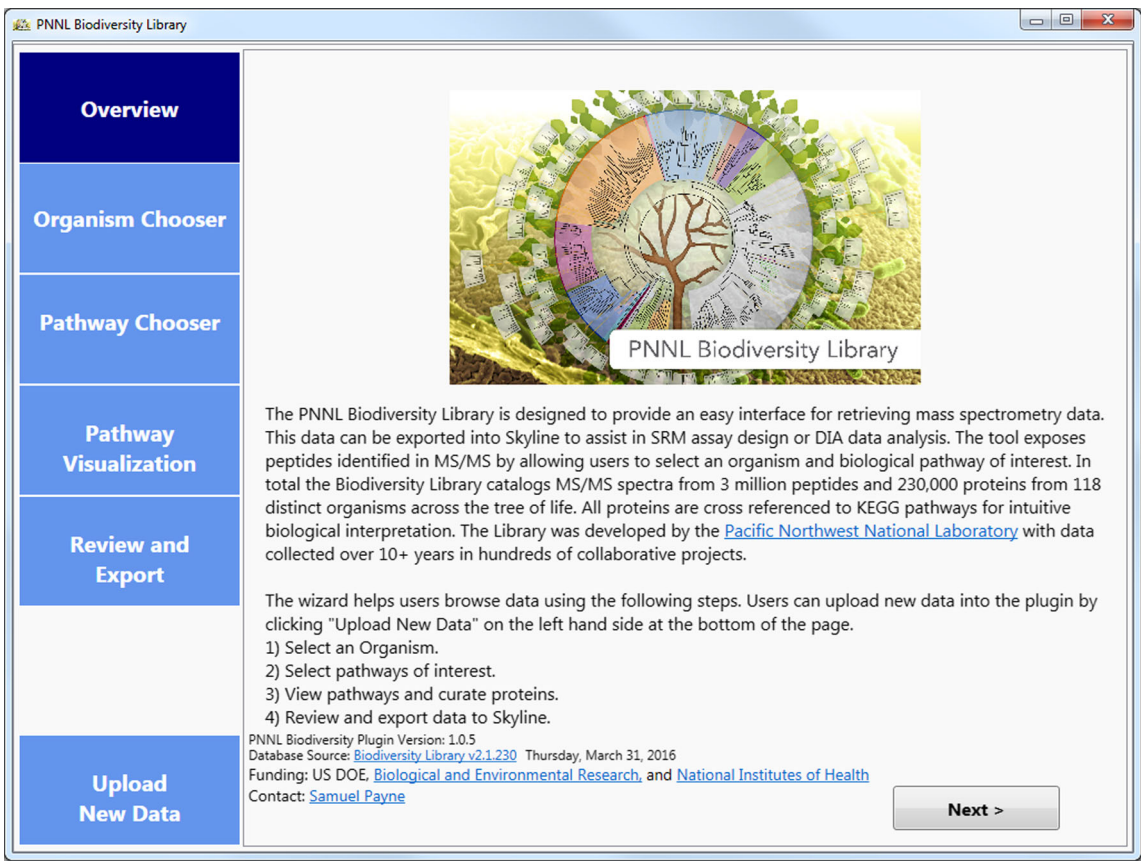


Figure 1. Biodiversity Library Plugin. The wizard interface of the plugin walks users through the steps of choosing data and importing it into Skyline

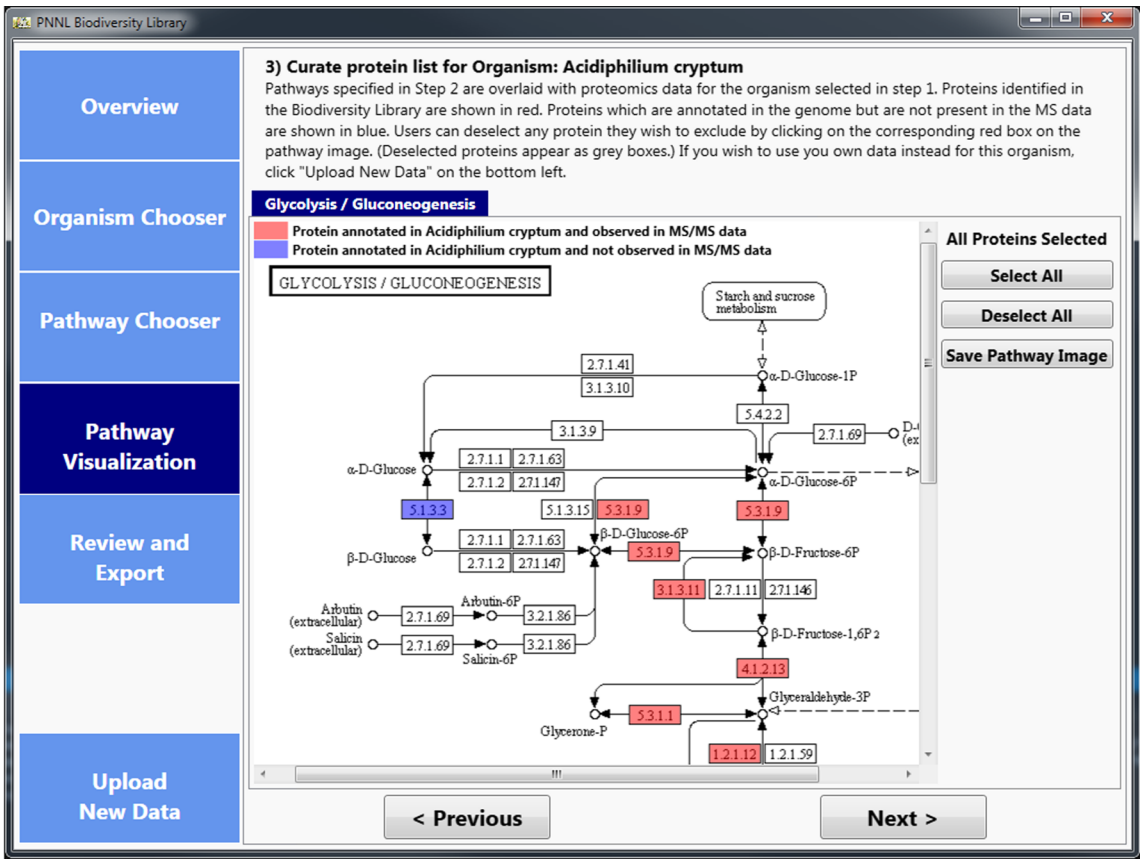


Figure 2. Pathway display. Pathway images put the proteomics data in the library into biological context. Proteins for which there exists data in the library are shown in red. Proteins that are annotated in the organism's genome but lack proteomics data are shown in blue. Proteins that are not annotated in the organism's genome are left white. All annotations come from KEGG

The Review and Export tab gives a quick overview of the selected data, listing the proteins in specified pathways along with the accessions and identifiers. By confirming this selection, the plugin will load the spectral library into Skyline and create a custom FASTA file corresponding to the selected proteins. Using Skyline, they can now browse spectra and identify the best transitions for SRM assay design or use the Bibliospec library for DIA data analysis.

Using Personal Data

By uploading personal data into the plugin, users have the ability to use the same pathway-oriented interface to load their data directly into Skyline. Given a properly formatted Bibliospec file using Uniprot identifiers, any local data can leverage the tool and be imported into Skyline. There are three options for customization: to replace data, to supplement data for an existing organism, or to add an entirely new organism. When replacing, all data for an existing organism will be removed and replaced with the custom data provided by the user. This would be appropriate if, for example, the user needed data from a different instrument type than is currently available in the library. In addition, using a local spectral library provides the retention time information that is critical for DIA data analysis. For supplementing, the data provided by the Biodiversity Plugin is combined with the custom data that the user provides. Supplementing is an appropriate choice when users have an experimental condition or sample with proteins that were not seen in the library. The final option of adding a new organism simply adds a new organism with the data provided by the user. The plug-in wizard walks through these options, verifies the input data and gives an overview of the changes that will take place. Once the user confirms the changes, their personal copy of the plugin database will be updated. Any changes made will persist on the user's machine only, ensuring that other users' data are not modified.

Hypothesis of DIA Data

Many investigators are interested in designing targeted assays for specific biochemical or disease pathways. However, SRM and PRM assays can be limited in terms of the number of proteins/peptides that can be analyzed in a single injection. This number varies based upon the scan speed of the instrument and also in the ability to perform scheduled analysis for peptides of interest. In addition, if the analysis of data from a targeted experiment implicates other proteins/pathways, additional injections of the same sample are required in order to quantify new proteins of interest. DIA does not suffer from these limitations, since all peptide species are fragmented in a single

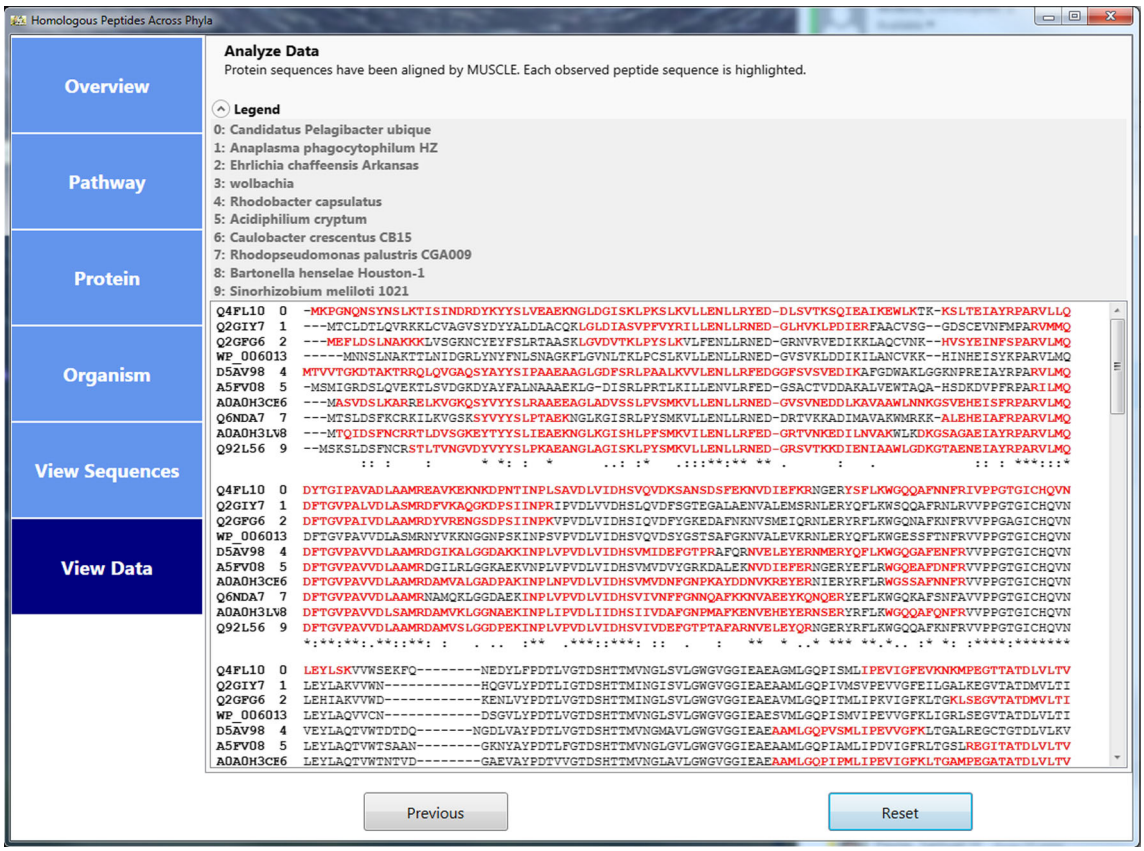


Figure 3. Peptide homology. Using the Proteotypic Peptide Viewer plugin, users can view peptides observed from a single protein across several organisms. In the image, the aconitate hydratase protein is shown for nine alpha proteobacteria. Eight organisms have proteomics data from the Biodiversity Library. The Wolbachia protein was uploaded separately for comparison

injection and therefore can be repeatedly inspected to verify new hypotheses. The Biodiversity Plugin can be very useful in facilitating repeated, hypothesis-driven interrogation of DIA data files. To accomplish this analysis, users would simply need to select a different set of proteins with the plugin and re-import them into Skyline. In those cases when follow-up studies are needed in order to analyze a greater number of samples with the higher sensitivity provided by SRM, the DIA data can be used to facilitate SRM experimental design [17].

For this type of analysis to be most effective, a custom in-house spectral library is necessary in order to provide retention time information for peptides corresponding to proteins of interest. However, the iRT algorithm [18], using synthetic or common internal [19] retention time standard peptides, can provide a mechanism for inclusion of peptides from the external spectral library by estimating a predicted retention time window. This approach could be valuable in those instances where additional peptides are needed for a protein of interest, or when proteins of interest are not represented in the user's spectral library.

Identifying Proteotypic Peptides Across Species

Given the wide diversity of organisms used in various biological experiments, it is impossible to have available mass spectrometry data for each. Biomedical and environmental research frequently identifies novel strains or species. To assist such investigations in understanding what peptides are likely to be observed in proteomics data of a new organism, we have created an additional plugin, the Proteotypic Peptide Viewer, which helps compare a given protein sequence with the data present in the Biodiversity Library. The plugin aligns homologous proteins from the Biodiversity Library with a user specified protein and then displays the observed peptides (Figure 3).

One use case for this viewer is to help identify peptides likely to be observed in a sample consisting of a multi-organism consortium. In such a metaproteomics experiment, the exact species present in the consortium may not be known beforehand. In this scenario, one may wish to use a marker protein that is reliably observed in many organisms and build a targeted assay to determine the presence/abundance of an organism of interest. The proteotypic peptide viewer helps to facilitate this design by showing which regions of a protein are reliably observed as peptides across taxa.

Conclusions

For mass spectrometry to be broadly adopted in biological laboratories as a routine experimental protocol, we must improve the ease of experimental design and the utility of the results. Targeted proteomics is an appealing approach for

proteomic characterization, from the standpoint of always returning a quantified abundance for every requested target. This is in sharp contrast to global or shotgun proteomics, which all too often fails to quantify one or more of the proteins that are critical to the hypotheses under investigation. One of the significant drawbacks to SRM-based proteomics is the complexity and time involved in designing assays before running the experiment. For non-mass spectrometry experts, this barrier is still too high. We have attempted to address this usability problem with the PNNL Biodiversity Plugin for Skyline, which summarizes available mass spectrometry data in a familiar pathway-centric visualization and seamlessly interacts with Skyline so that users can approach the data in the same way that they approach the biological experiments.

Acknowledgments

This research was supported by the NIH National Institute of General Medical Sciences (GM103493), and by the Department of Energy Office of Biological and Environmental Research Genome Sciences Program under the Pan-omics project. Work was performed in the Environmental Molecular Science Laboratory, a U.S. Department of Energy (DOE) national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, WA. Battelle operates PNNL for the DOE under contract DE-AC05-76RLO01830.

References

1. Liebler, D.C., Zimmerman, L.J.: Targeted quantitation of proteins by mass spectrometry. Biochemistry **52**(22), 3797–3806 (2013)
2. Prakash, A., Tomazela, D.M., Frewen, B., Maclean, B., Merrihew, G., Peterman, S., Maccoss, M.J.: Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. J. Proteome Res. **8**(6), 2733–2739 (2009)
3. Picotti, P., Aebersold, R.: Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat. Methods **9**(6), 555–566 (2012)
4. Wu, C., Shi, T., Brown, J.N., He, J., Gao, Y., Fillmore, T.L., Shukla, A.K., Moore, R.J., Camp II, D.G., Rodland, K.D., Qian, W.J., Liu, T., Smith, R.D.: Expediting SRM assay development for large-scale targeted proteomics experiments. J. Proteome Res. **13**(10), 4479–4487 (2014)
5. Vizcaino, J.A., Cote, R.G., Csordas, A., Dianes, J.A., Fabregat, A., Foster, J.M., Griss, J., Alpi, E., Birim, M., Contell, J., O'Kelly, G., Schoenegger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R., Hermjakob, H.: The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. **41**(Database issue), D1063–D1069 (2013)
6. Vizcaino, J.A., Deutsch, E.W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J.A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L.: ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. **32**(3), 223–226 (2014)
7. Craig, R., Cortens, J.P., Beavis, R.C.: Open source system for analyzing, validating, and storing protein identification data. J. Proteome Res. **3**(6), 1234–1242 (2004)
8. Egertson, J.D., Kuehn, A., Merrihew, G.E., Bateman, N.W., MacLean, B.X., Ting, Y.S., Canterbury, J.D., Marsh, D.M., Kellmann, M., Zabrouskov, V., Wu, C.C., MacCoss, M.J.: Multiplexed MS/MS for improved data-independent acquisition. Nat. Methods **10**(8), 744–746 (2013)
9. MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern, R., Tabb, D.L., Liebler, D.C., MacCoss, M.J.: Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics (Oxford, Engl.) **26**(7), 966–968 (2010)
10. Schilling, B., MacLean, B., Held, J.M., Sahu, A.K., Rardin, M.J., Sorensen, D.J., Peters, T., Wolfe, A.J., Hunter, C.L., MacCoss, M.J., Gibson, B.W.: Multiplexed, scheduled, high-resolution parallel reaction monitoring on a full scan QqTOF Instrument with integrated data-dependent and targeted mass spectrometric workflows. Anal. Chem. **87**(20), 10222–10229 (2015)
11. Schilling, B., Rardin, M.J., MacLean, B.X., Zawadzka, A.M., Frewen, B.E., Cusack, M.P., Sorensen, D.J., Bereman, M.S., Jing, E., Wu, C.C., Verdin, E., Kahn, C.R., Maccoss, M.J., Gibson, B.W.: Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. Mol. Cell. Proteom. **11**(5), 202–214 (2012)
12. Egertson, J.D., MacLean, B., Johnson, R., Xuan, Y., MacCoss, M.J.: Multiplexed peptide analysis using data-independent acquisition and Skyline. Nat. Protoc. **10**(6), 887–903 (2015)
13. Rardin, M.J., Schilling, B., Cheng, L.Y., MacLean, B.X., Sorensen, D.J., Sahu, A.K., MacCoss, M.J., Vitek, O., Gibson, B.W.: MS1 peptide ion intensity chromatograms in MS^2 (SWATH) data independent acquisitions. improving post acquisition analysis of proteomic experiments. Mol. Cell. Proteom. **14**(9), 2405–2419 (2015)
14. Payne, S.H., Monroe, M.E., Overall, C.C., Kiebel, G.R., Degan, M., Gibbons, B.C., Fujimoto, G.M., Purvine, S.O., Adkins, J.N., Lipton, M.S., Smith, R.D.: The Pacific Northwest National Laboratory library of bacterial and archaeal proteomic biodiversity. Scientific Data **2**, 150041 (2015)
15. Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., Thomas, J.K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N.A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L.D., Patil, A.H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S.K., Marimuthu, A., Sathe, G.J., Chavan, S., Datta, K.K., Subbannayya, Y., Sahu, A., Yelamanchi, S.D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K.R., Syed, N., Goel, R., Khan, A.A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.C., Zhong, J., Wu, X., Shaw, P.G., Freed, D., Zahari, M.S., Mukherjee, K.K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C.J., Shankar, S.K., Satishchandra, P., Schroeder, J.T., Sirdeshmukh, R., Maitra, A., Leach, S.D., Drake, C.G., Halushka, M.K., Prasad, T.S., Hruban, R.H., Kerr, C.L., Bader, G.D., Iacobuzio-Donahue, C.A., Gowda, H., Pandey, A.: A draft map of the human proteome. Nature **509**(7502), 575–581 (2014)
16. Kanehisa, M., Goto, S.: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28**(1), 27–30 (2000)
17. Searle, B.C., Egertson, J.D., Bollinger, J.G., Stergachis, A.B., MacCoss, M.J.: Using data independent acquisition (DIA) to model high-responding peptides for targeted proteomics experiments. Mol. Cell. Proteom. **14**(9), 2331–2340 (2015)
18. Escher, C., Reiter, L., MacLean, B., Ossola, R., Herzog, F., Chilton, J., MacCoss, M.J., Rinner, O.: Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics **12**(8), 1111–1121 (2012)
19. Parker, S.J., Rost, H., Rosenberger, G., Collins, B.C., Malmstrom, L., Amodei, D., Venkatraman, V., Raedschelders, K., Van Eyk, J.E., Aebersold, R.: Identification of a set of conserved eukaryotic internal retention time standards for data-independent acquisition mass spectrometry. Mol. Cell. Proteom. **14**(10), 2800–2813 (2015)