Protocol for analysis of glycoproteomics LC-MS data using GlycReSoft

Joshua A. Klein[1] and Joseph Zaia[1,2]

Bioinformatics Program[1], Dept. of Biochemistry[2] Boston University

Contact information:

Joseph Zaia

Boston University Medical Campus

670 Albany St., Rm. 509

Boston, MA 02119

USA

(v) 1-617-638-6762

(e) jzaia@bu.edu

Running Head

GlycReSoft

Summary/Abstract

The GlycReSoft software tool allows users to process glycoproteomics LC-MS data sets. The tool accepts proteomics database search results or a user-defined list of proteins in the sample. GlycReSoft processes LC-MS data to yield deconvoluted exact mass values. The user has the option to import a list of theoretical glycans from an external database, a curated glycan list, or a measured glycome. The tool assembles a list of theoretical glycopeptides from the lists of theoretical glycans and proteins, respectively. The program then scores the tandem mass spectra in the LC-MS data files and provides graphical views of the identified glycopeptides for each protein in the sample, and the set of glycoforms identified for each peptide sequence.

Key words

Glycomics, glycoproteomics, glycoinformatics, bioinformatics, mass spectrometry

1. Introduction

Conventional bottom-up proteomics workflows apply most directly to identification of unmodified peptides or those with relatively small chemical or enzymatic post-translational modifications. For communication of database searching results, the mzIdentML format was developed (1,2) and includes such small PTMs that fit with the original proteomics use cases. Complex glycosylation does not fit within the scope of bottom-up proteomics as originally defined. One problem is that glycosylation, resulting from ER and Golgi-mediated biosynthetic reactions, is heterogeneous as a rule, necessitating the definition of a range of post-translationally modified forms of a given modified peptide sequence. Another problem is that the glycans undergo

dissociation during the tandem MS experiment, and there is no way to specify the product ions using existing data standards.

The proteoform concept was formulated in response to the observation of protein heterogeneity using top-down MS (3,4). Proteoforms, as different modified forms of a given gene product, diversify the functions that can be attributed to a protein through allosteric regulation and/or activation of interactions with binding partners or adapter molecules. For complex glycosylation, it is not uncommon to observe more than 30 glycoforms at a glycosite. The number of theoretical glycoforms thus multiplies as the number of protein glycosites increases, rapidly exceeding the number of proteoforms that could be made by a cell (5). Therefore, the goal of glycoproteomics is to identify the subset of the theoretical glycoforms that exist in a given biological context.

As summarized in recent reviews, glycopeptides are assigned in mass spectrometry experiments using features including mass, elemental composition, tandem mass spectra, and time of elution or migration from the separation system (liquid chromatography or capillary electrophoresis) (6-8). In principle, the confidence of an assignment improves as the mass spectrometer accuracy increases. For the tandem MS step, three types of product ions define the glycopeptide in terms of peptide sequence, glycan composition and glycosylation site(s). Glycopeptides dissociate to form low mass oxonium ions corresponding to saccharide fragments. Neutral losses of saccharide units give rise to a series of peptide + Y_n ions. The presence of peptide backbone product ions is particularly useful for identifying the peptide sequence. Collisional dissociation defines the peptide sequence and glycan composition. In favorable cases, the site of glycosylation can be defined for singly glycosylated peptides. For multiply glycosylated peptides, collisional dissociation defines the total glycan composition but often does not define the glycosylation at individual peptide sites. Electron activated

dissociation methods produce preferential cleavage of the peptide backbone and are therefore more likely to succeed in assigning multiply glycosylated peptides.

GlycReSoft is an open-source software program for processing glycomics and glycoproteomics LC-MS data (9,10) (Figure 1, see Note 1). The program uses a deconvoluter based on DeconTools (11) and MD-DeconV (12) that employs both peptide and glycopeptide averagine values that identifies glycopeptide precursor ion elemental compositions based on isotope pattern matching. The program uses the list proteins identified using a proteomics search engine in the form of an export in mzIdentML or FASTA format. The user can specify the range of theoretical glycan compositions by importing from an external database or defining a theoretical list from algebraic combinations of glycans. The program also accepts measured glycan profiles from experimental data. Users can also import curated glycan lists. The program then calculates the theoretical glycopeptide precursor ions using the theoretical glycan lists and the observed sample proteome.

We provide here a method for using GlycReSoft for analysis of a human α1-acidglycoprotein tryptic digest from a research publication (13). For this, the user is directed to download data files from the Pride archive. These consist of the reversed phase LC-MS data acquired using a Thermo-Fisher Scientific Q-Exactive plus mass spectrometry system, see Figure 2.

2. Materials

2.1. Installation

2.1.1. Windows. Download and install the Windows Graphical Installer (14). GlycReSoft will prompt the user to create a working folder in which results files will be stored.

2.1.2. LINUX. A Windows command line interface is also available along with technical documentation (14).

2.2. Quick tour

GlycReSoft implements algorithms for:

- Generation of lists of theoretical glycan compositions using algebraic rules, the glySpace database network (15), or from text files;
- Generation of lists of theoretical glycopeptide masses using protein sequences in FASTA format or proteomics search results in Human Proteome Organization (HUPO) Proteome Standards Initiative (PSI) (16) mzIdentML format (1,2), combined with a glycan search space;
- Support for *N*-linked, *O*-linked, or GAG-linker glycopeptides;
- Deisotoping and charge state deconvolution of glycan and glycopeptide mass spectra;
- Identification and quantification of glycans by MS and glycopeptides by MS/MS.

GlycReSoft reads LC-MS data in mzML (17) and mzXML (18) formats. It also reads Thermo-Fisher RAW file format. It is available as a Windows precompiled build (described here), compatible with Windows 8 and 10, and a UNIX/LINUX command line interface (14).

3. Methods

3.1. Download public proteomics and glycoproteomics data on human alpha-1-acid glycoprotein digests (13). File names indicate protease used and whether or not samples were deglycosylated using peptide N-glycosidase F (PNGaseF) before analysis. The files are available through the following PRIDE repository (19):
Project Name: Influenza A virus- integrated glycomics, proteomics and glycoproteomics
Project accession: PXD003498

Project DOI: 10.6019/PXD003498

Keys to data:

Tryp- tryptic digest

Chymo- Chymotryptic digest

O16- Samples subjected to PNGaseF deglycosylation in presence of regular water (H216O).

GP- No deglycosylation.

3.1.1. Download the files “AGP-tryp-GP.raw” and “AGP_O16_tryp_peptides_1_1_0_new.mzid.gz” (see Note 2).

3.1.2. Convert the “AGP-tryp-GP.raw” file to mzML format using ProteoWizard MS_convert (20) (see Note 4) using the default parameters (Figure 3).

3.2. Create a glycan search space. Click on “BUILD A GLYCAN SEARCH SPACE” (Figure 4). In the window, give the glycan search space a name by clicking on “Hypothesis Name”. Select Reduction type “native” and Derivatization Type “native”. Click “COMBINATORIAL HYPOTHESIS” and use the default monosaccharide and algebraic rules (see Note 3). Click “GENERATE”. The bar on the left indicates the task status.

3.3. Create a glycoproteomics search space. Click “BUILD A GLYCOPEPTIDE SEARCH SPACE” (Figure 5). Click “Hypothesis Name” and enter a name. Click “SELECT FILE”, then click on the protein list file “AGP_O16_tryp_peptides_1_1_0_new.mzid”. Under “Select a Glycan Hypothesis or Sample Analysis” select the name of your glycan hypothesis. Leave all other parameters at their default values (see Note 5). Click “GENERATE”. The bar on the left indicates the task status.

3.4. Run LC-MS preprocessing steps. Click “ANALYZE SAMPLE”. Click “SELECT MZML FILE” (Figure 6). Provide a sample name. Under “Preset Configurations” select “LC-MS/MS Glycoproteomics”. Under “MS1 Parameters” “Averagine” select “glycopeptide”.

Under “MSn Parameters” “Averagine” select “glycopeptide”. Use the default settings for all other parameters. Click “SUBMIT”. The bar on the left indicates the task status.

3.5. Search glycopeptide sequences. Click “SEARCH GLYCOPEPTIDE SEQUENCES” (Figure 7). Select a sample from the menu under “Select One or More Samples”. The samples correspond to pre-processed mzML data files stored in the working directory. Under “Choose a Hypothesis” select the name of the glycopeptide search space file created in step 3.3. Leave all other parameters as their default values. Click “SUBMIT”. Progress is shown on the task window on the left side of the screen.

3.6. View results. Under “Analyses” click on the results that you wish to view. Clicking on the “OVERVIEW” tab, GlycReSoft displays results for each protein in the proteome. Clicking on the “GLYCOPEPTIDES” tab displays a table of all glycopeptides detected for the selected protein (Figure 8). Clicking on individual rows jumps to display of the extracted ion chromatogram, tandem mass spectrum, and product ion table. Scrolling down, the program displays a pileup diagram of the glycopeptide glycoforms identified for each glycosite in the protein sequence. Mousing over the pileup diagram displays glycopeptide mass, sequence, glycan composition and MS2 statistics. Clicking on an individual glycopeptide bar will display the extracted ion chromatogram, tandem mass spectrum, and product ion table (Figure 9). Clicking on the “GLYCOPEPTIDES” tab displays a table of all glycopeptides identified (Figure 10). Clicking on the “SITE DISTRIBUTION” tab displays bar plots of the glycoforms identified for each peptide sequence (Figure 11).

3.7. Export results. Click on the disk icon to display the results export options. GlycReSoft exports results to the as comma separated value files. It generates annotated pdf files for each tandem mass spectrum. It generates a HTML report that is viewable using a web browser.

4. Notes

Note 1. The use of GlycReSoft for glycoproteomics is described here.

Note 2. Unzip the *.gz directory using 7-Zip (https://wcww.7-zip.org/).

Note 3. The monosaccharide list can be modified to include NeuGc or other monosaccharides as desired.

Note 4. The ProteoWizard package is available for download at:

http://www.proteowizard.org/download.html

Note 5. When using mzIdentML files the fixed modifications, proteases, and number of missed cleavages are read from the mzIdentML file.

Note 6. For MSn spectra, the peptide averagine works best higher normalized collision energy (>25) dissociation. The reason is that most b/y ions are almost entirely peptide, and the peptide+Y ions dominated by the intact peptide most of the time. When lower normalized collision energy is used, peptide+Y ions with more glycan are present, when a combination of peptide and glycopeptide averagines works better.

5. Acknowledgements

The development of GlycReSoft was supported by NIH grants P41GM104888 and U01CA221234.

Figures

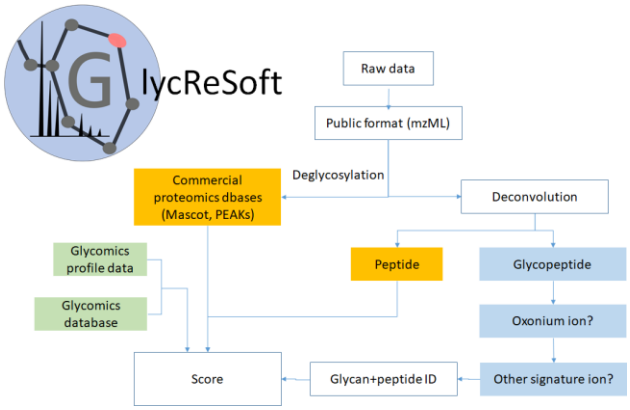

Figure 1. Glycresoft pipeline for processing glycopeptide LC-MS data

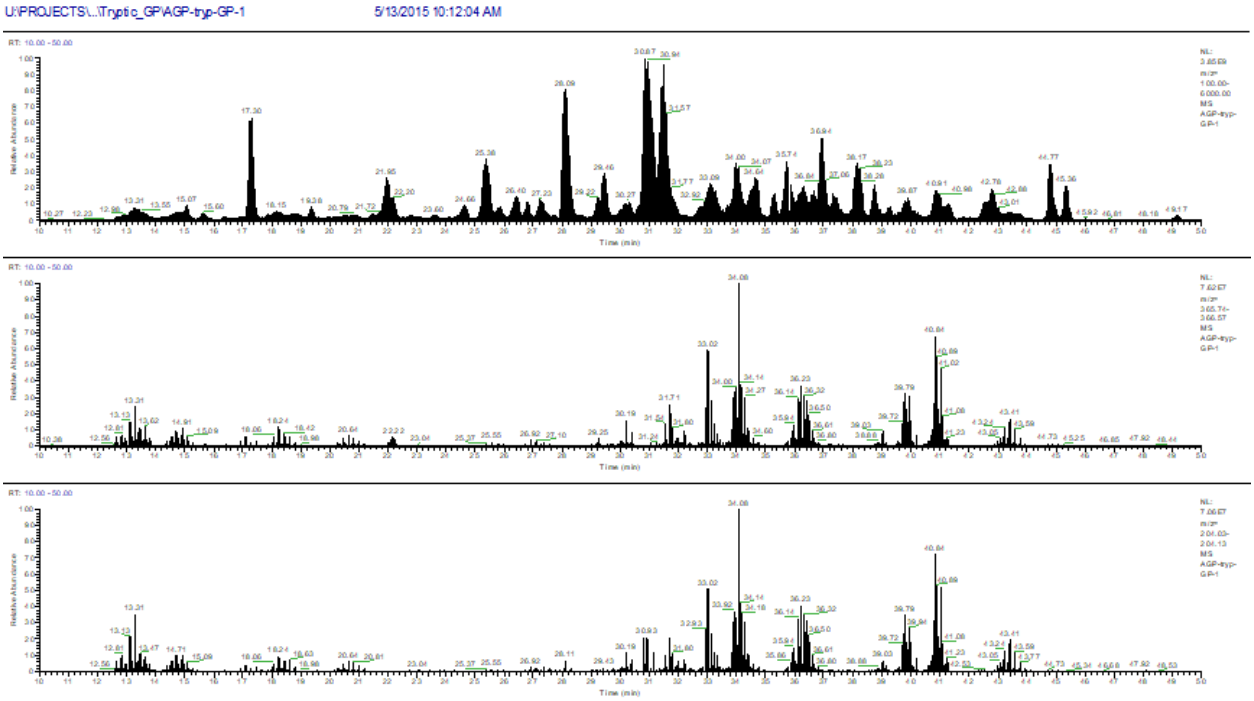

Figure 2. Human α1-acidglycoprotein tryptic peptides analyzed using reversed phase LC-MS (13). The top panel shows the extracted total ion chromatogram. The middle panel shows the extracted ion chromatogram (EIC) for the Hex-HexNAc oxonium ion. The bottom panel shows the EIC for the HexNAc oxonium ion.

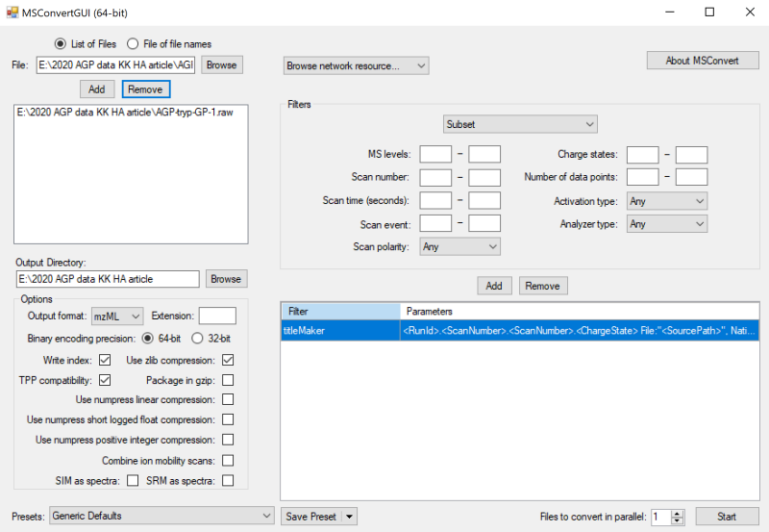

Figure 3. Screen shot of the MS_convert utility

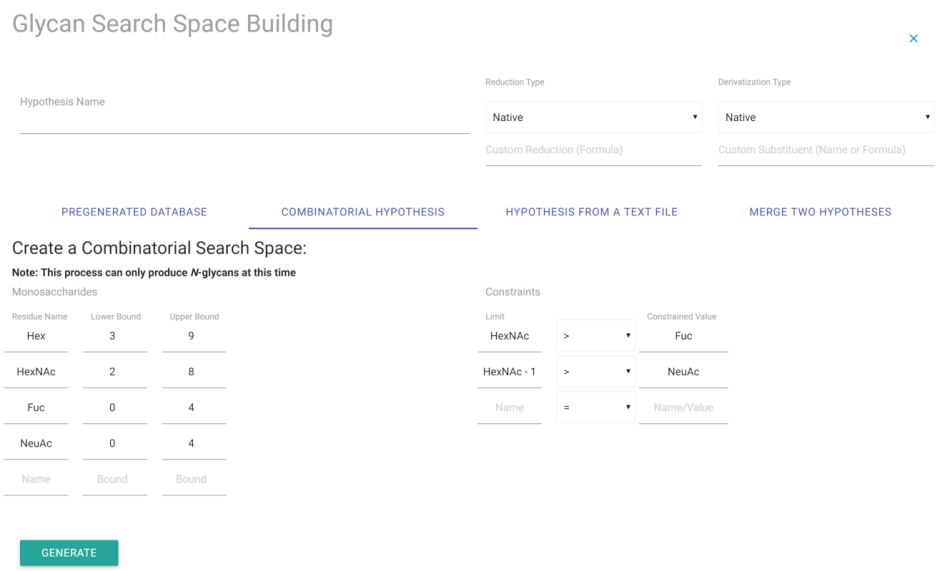

Figure 4. Glycan Search Space Building window

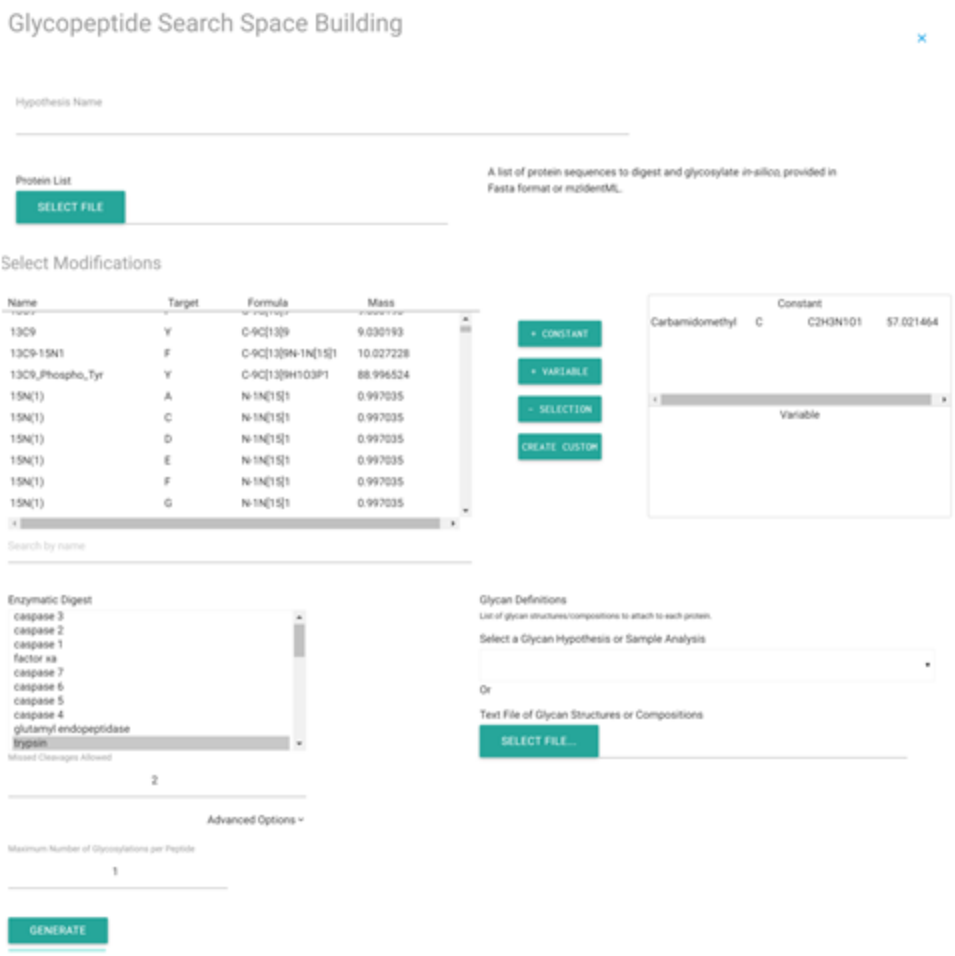

Figure 5. Glycopeptide Search Space Building window

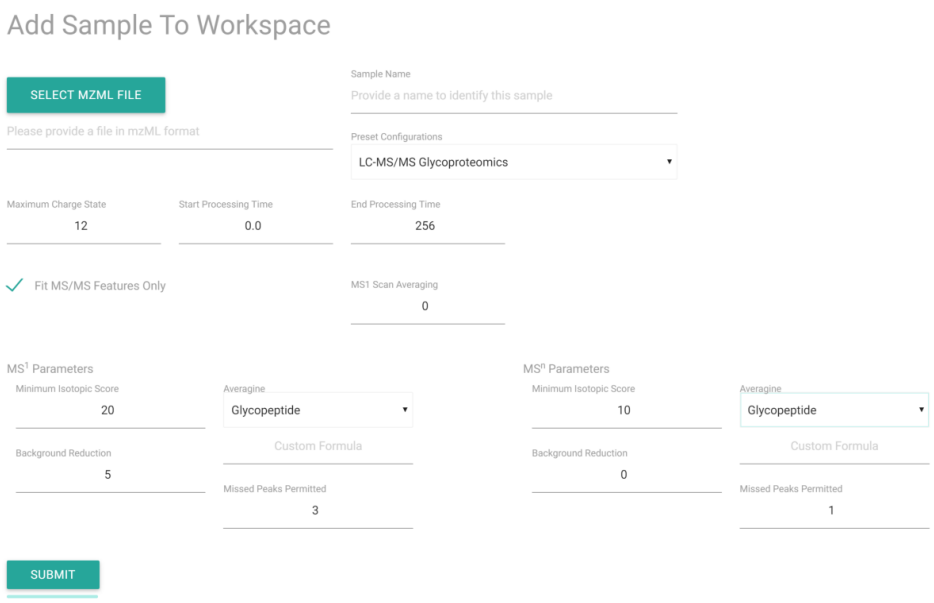

Figure 6. ANALYZE SAMPLE window

Figure 7. Search Glycopeptide Sequences window

Figure 8. Glycopeptide pileup diagram for AGP1.

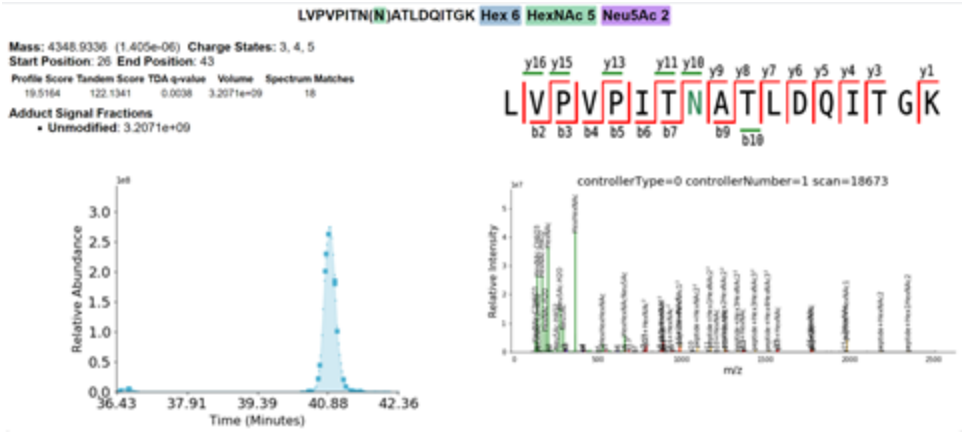

Figure 9. Glycopeptide extracted ion chromatogram, annotated tandem mass spectrum, fragmentation diagram, and tandem MS statistics.

Observed Mass	Sequence	Tandem Score	q-value	Profile Score	Volume
2709.0190	NEEYN(N)K Hex 5 HexNAc 4 NeuSAc 1	174.3502	0.045	17.2993	8.99898e+08
3365.2465	NEEYN(N)K Hex 6 HexNAc 5 NeuSAc 2	174.2526	0.045	17.7857	6.39060e+08
3000.1158	NEEYN(N)K Hex 5 HexNAc 4 NeuSAc 2	173.9959	0.045	18.9811	1.32985e+09
3220.2078	NEEYN(N)K Fuc 1 Hex 6 HexNAc 5 NeuSAc 1	158.0731	0.045	17.7644	1.18188e+08
3074.1503	NEEYN(N)K Hex 6 HexNAc 5 NeuSAc 1	150.8929	0.045	16.2362	2.25395e+08
3656.3429	NEEYN(N)K Hex 6 HexNAc 5 NeuSAc 3	150.8922	0.045	16.9519	8.49359e+08
4057.8338	LVPVPITN(N)ATLDQITGK Hex 6 HexNAc 5 NeuSAc 1	126.7192	0.004	20.0030	1.43114e+09
4348.9301	LVPVPITN(N)ATLDQITGK Hex 6 HexNAc 5 NeuSAc 2	122.1341	0.004	19.5164	3.20713e+09
3802.4009	NEEYN(N)K Fuc 1 Hex 6 HexNAc 5 NeuSAc 3	119.2988	0.045	17.4281	4.07150e+08
4640.0285	LVPVPITN(N)ATLDQITGK Hex 6 HexNAc 5 NeuSAc 3	116.1081	0.004	18.2666	2.19433e+09
4074.8550	LVPVPITN(N)ATLDQITGK Fuc 1 Hex 7 HexNAc 5	114.2240	0.004	17.1595	3.31680e+08
3692.6956	LVPVPITN(N)ATLDQITGK Hex 5 HexNAc 4 NeuSAc 1	113.0208	0.004	19.2424	5.71748e+08
4365.9494	LVPVPITN(N)ATLDQITGK Fuc 1 Hex 7 HexNAc 5 NeuSAc 1	112.3450	0.004	14.8225	6.46612e+08
4203.8870	LVPVPITN(N)ATLDQITGK Fuc 1 Hex 6 HexNAc 5 NeuSAc 1	110.9910	0.004	18.5219	4.55767e+08
3983.7899	LVPVPITN(N)ATLDQITGK Hex 5 HexNAc 4 NeuSAc 2	110.5868	0.004	18.0868	9.98522e+08

Figure 10. Example table of identified glycopeptides (partial listing)

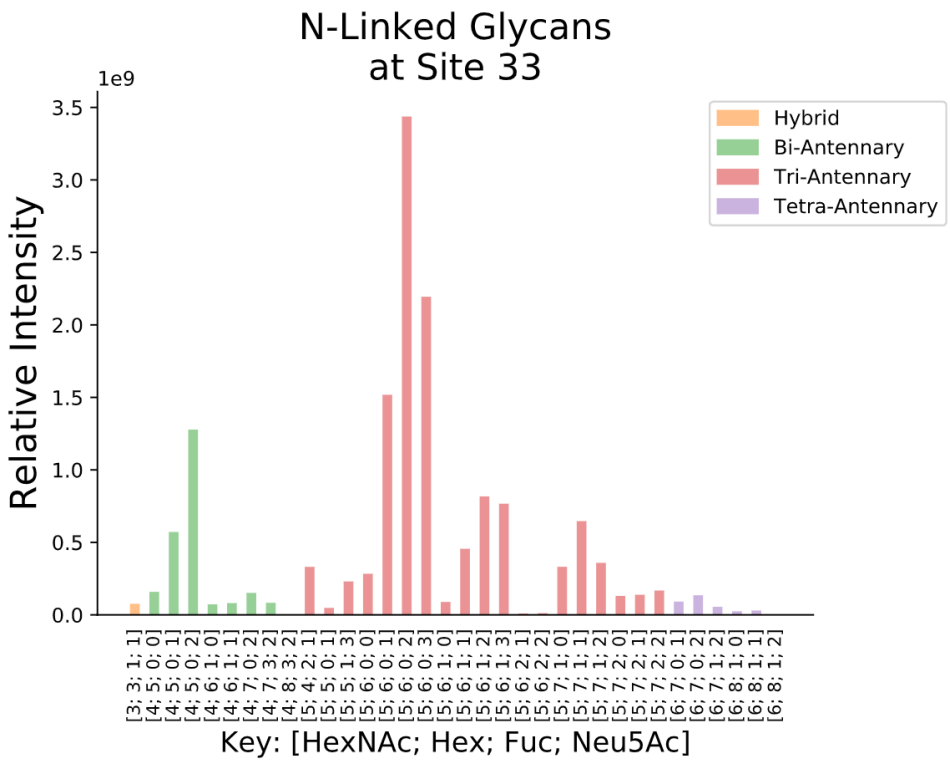

Figure 11. Example bar plot showing all glycoforms identified for a peptide sequence.

References

1. Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S. J., Selley, J. N., Searle, B. C., Shofstahl, J., Seymour, S. L., Julian, R., Binz, P. A., Deutsch, E. W., Hermjakob, H., Reisinger, F., Griss, J., Vizcaino, J. A., Chambers, M., Pizarro, A., and Creasy, D. (2012) *Molecular & cellular proteomics : MCP* **11**, M111 014381
2. Eisenacher, M. (2011) *Methods Mol Biol* **696**, 161-177
3. Smith, L. M., and Kelleher, N. L. (2013) *Nat. Methods* **10**, 186-187
4. Smith, L. M., Thomas, P. M., Shortreed, M. R., Schaffer, L. V., Fellers, R. T., LeDuc, R. D., Tucholski, T., Ge, Y., Agar, J. N., Anderson, L. C., Chamot-Rooke, J., Gault, J., Loo, J. A., Pasa-Tolic, L., Robinson, C. V., Schluter, H., Tsybin, Y. O., Vilaseca, M., Vizcaino, J. A., Danis, P. O., and Kelleher, N. L. (2019) *Nat Methods*
5. Aebersold, R., Agar, J. N., Amster, I. J., Baker, M. S., Bertozzi, C. R., Boja, E. S., Costello, C. E., Cravatt, B. F., Fenselau, C., Garcia, B. A., Ge, Y., Gunawardena, J., Hendrickson, R. C., Hergenrother, P. J., Huber, C. G., Ivanov, A. R., Jensen, O. N., Jewett, M. C., Kelleher, N. L., Kiessling, L. L., Krogan, N. J., Larsen, M. R., Loo, J. A., Ogorzalek Loo, R. R., Lundberg, E., MacCoss, M. J., Mallick, P., Mootha, V. K., Mrksich, M., Muir, T. W., Patrie, S. M., Pesavento, J. J., Pitteri, S. J., Rodriguez, H., Saghatelian, A., Sandoval, W., Schluter, H., Sechi, S., Slavoff, S. A., Smith, L. M., Snyder, M. P., Thomas, P. M., Uhlen, M., Van Eyk, J. E., Vidal, M., Walt, D. R., White, F. M., Williams, E. R., Wohlschlager, T., Wysocki, V. H., Yates, N. A., Young, N. L., and Zhang, B. (2018) *Nat Chem Biol* **14**, 206-214
6. Klein, J. A., and Zaia, J. (2019) *Biochemistry*
7. Hu, H., Khatri, K., Klein, J., Leymarie, N., and Zaia, J. (2016) *Glycoconj J* **33**, 285-296
8. Hu, H., Khatri, K., and Zaia, J. (2017) *Mass Spectrom Rev* **36**, 475-498

9. Maxwell, E., Tan, Y., Tan, Y., Hu, H., Benson, G., Aizikov, K., Conley, S., Staples, G. O., Slysz, G. W., Smith, R. D., and Zaia, J. (2012) *PLoS ONE* **7**, e45474

10. Klein, J., and Zaia, J. (2020) *J Proteome Res* **doi: 10.1021/acs.jproteome.0c00051**

11. Jaitly, N., Mayampurath, A., Littlefield, K., Adkins, J. N., Anderson, G. A., and Smith, R. D. (2009) *BMC Bioinformatics* **10**, 87

12. Liu, X., Inbar, Y., Dorrestein, P. C., Wynne, C., Edwards, N., Souda, P., Whitelegge, J. P., Bafna, V., and Pevzner, P. A. (2010) *Molecular & Cellular Proteomics* **9**, 2772-2782

13. Khatri, K., Klein, J. A., White, M. R., Grant, O. C., Leymarie, N., Woods, R. J., Hartshorn, K. L., and Zaia, J. (2016) *Mol Cell Proteomics* **15**, 1895-1912

14. GlycReSoft software for glycomics and glycoproteomics, http://www.bumc.bu.edu/msr/glycresoft/

15. Aoki-Kinoshita, K. F., Lisacek, F., Mazumder, R., York, W. S., and Packer, N. H. (2020) *Glycobiology* **30**, 70-71

16. HUPO Proteome Standards Initiative, http://www.psidev.info/

17. Deutsch, E. (2008) *Proteomics* **8**, 2776-2777

18. Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004) *Nat Biotechnol* **22**, 1459-1466

19. Deutsch, E. W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., Campbell, D. S., Bernal-Llinares, M., Okuda, S., Kawano, S., Moritz, R. L., Carver, J. J., Wang, M., Ishihama, Y., Bandeira, N., Hermjakob, H., and Vizcaino, J. A. (2017) *Nucleic Acids Res* **45**, D1100-D1106

20. Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) *Bioinformatics* **24**, 2534-2536