Peak Filtering, Peak Annotation, and Wildcard Search for Glycoproteomics

Supplemental Information

Marshall Bern*, Abhishek Roushan, Doron Kletter, K. Ilker Sen, Eric Carlson

Protein Metrics Inc., Cupertino, CA 95014

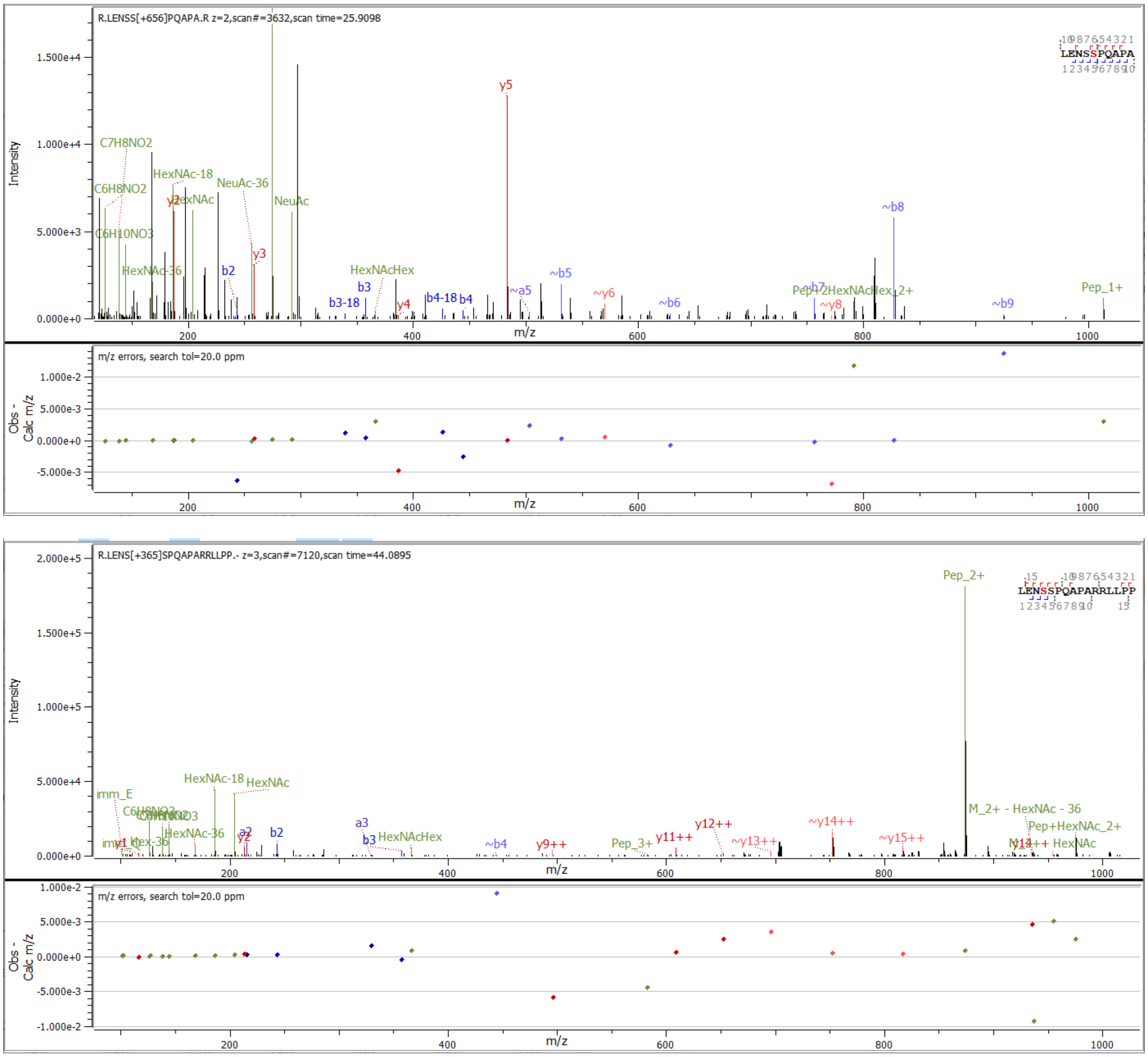
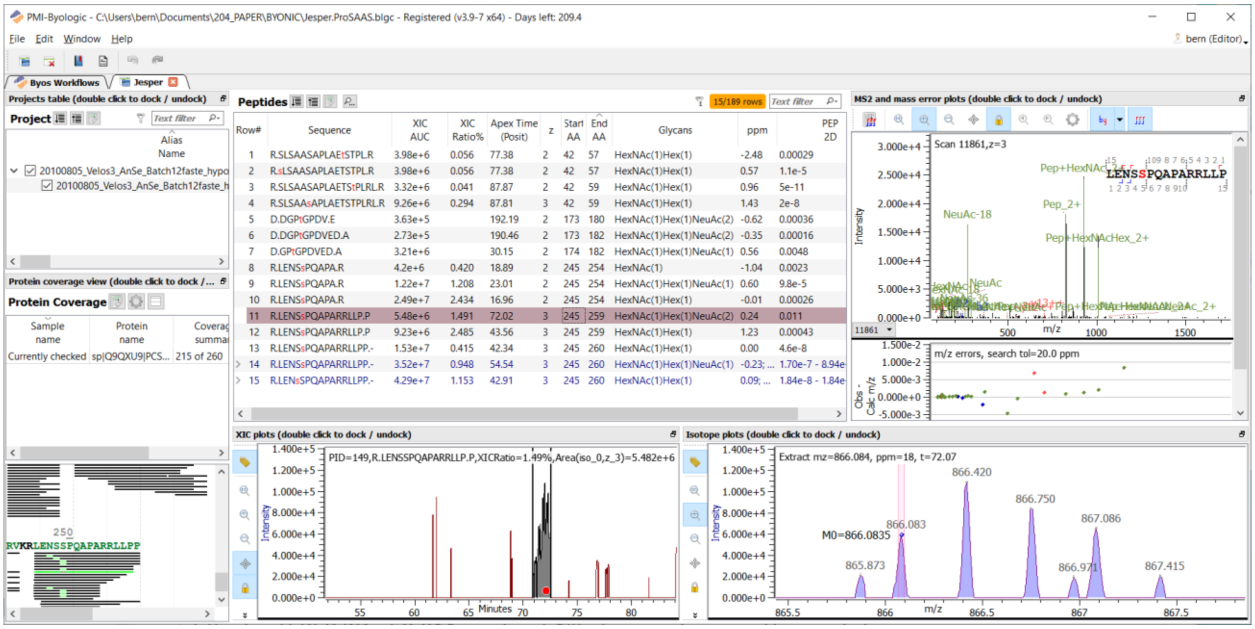


Figure S1. Little LEN and Big LEN glycopeptides from ProSAAS in the Copenhagen data set, ProteomeXchange PXD002431, 20100805_Velos3_AnSe_Batch12faste_hypot_S3.raw.



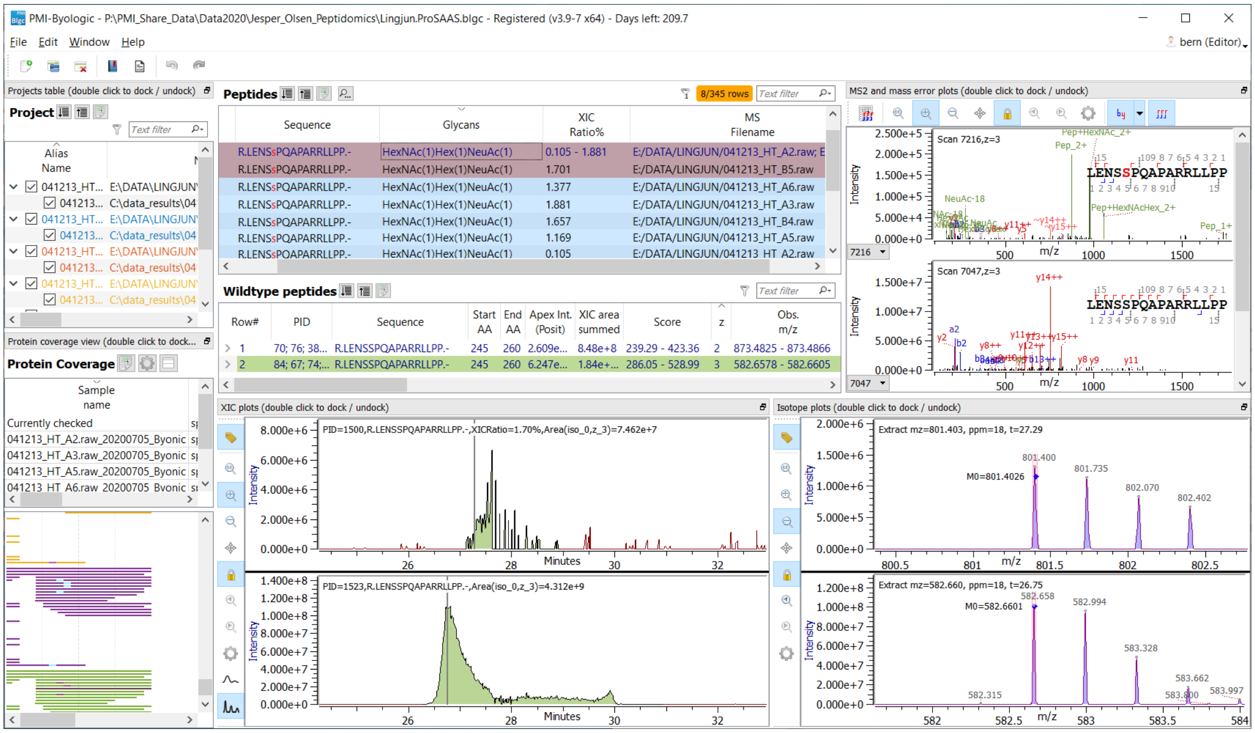
PMI-Byologic - C:\Users\bern\Documents\204_PAPER\BYONIC\Jesper.ProSAAS.blgc - Registered (v3.9-7 x64) - Days left: 209.4
File Edit Window Help
bern (Editor)
Byos Workflows
Jesper
Projects table (double click to dock / undock)
Project
Text filter
Alias Name
20100805_Velos3_AnSe_Batch12faste_hypo
20100805_Velos3_AnSe_Batch12faste_h
Protein coverage view (double click to dock / ...
Protein Coverage
250
RVKRLENSSPQAPARRLLPP
Peptides
15/189 rows
Text filter
MS2 and mass error plots (double click to dock / undock)
Scan 11861,z=3
Pep+HexNAc
LENSSPQAPARRLLP
Pep_2+
NeuAc-18
Pep+HexNAcHex_2+
Intensity
m/z
11861
m/z errors, search tol=20.0 ppm
Obs - Calc m/z
XIC plots (double click to dock / undock)
PID=149,R.LENSSPQAPARRLLP.P,XICRatio=1.49%,Area(iso_0,z_3)=5.482e+6
Intensity
Minutes
Isotope plots (double click to dock / undock)
Extract mz=866.084, ppm=18, t=72.07
M0=866.0835
865.873
866.083
866.420
866.750
866.971
867.086
867.415
m/z

Sample name	Protein name	Coverage summary
Currently checked	sp|Q9QXU9|PCS...	215 of 260

Row#	Sequence	XIC AUC	XIC Ratio%	Apex Time (Posit)	z	Start AA	End AA	Glycans	ppm	PEP 2D
1	R.SLSAASAPLAEtSTPL.R	3.98e+6	0.056	77.38	2	42	57	HexNAc(1)Hex(1)	-2.48	0.00029
2	R.sLSAASAPLAETSTPL.R	3.98e+6	0.056	77.38	2	42	57	HexNAc(1)Hex(1)	0.57	1.1e-5
3	R.SLSAASAPLAETStPLRL.R	3.32e+6	0.041	87.87	2	42	59	HexNAc(1)Hex(1)	0.96	5e-11
4	R.SLSAAsAPLAETSTPLRL.R	9.26e+6	0.294	87.81	3	42	59	HexNAc(1)Hex(1)	1.43	2e-8
5	D.DGPtGPDV.E	3.63e+5		192.19	2	173	180	HexNAc(1)Hex(1)NeuAc(2)	-0.62	0.00036
6	D.DGPtGPDVED.A	2.73e+5		190.46	2	173	182	HexNAc(1)Hex(1)NeuAc(2)	-0.35	0.00016
7	D.GPtGPDVED.A	3.21e+6		30.15	2	174	182	HexNAc(1)Hex(1)NeuAc(1)	0.56	0.0048
8	R.LENSsPQAPA.R	4.2e+6	0.420	18.89	2	245	254	HexNAc(1)	-1.04	0.0023
9	R.LENSsPQAPA.R	1.22e+7	1.208	23.01	2	245	254	HexNAc(1)Hex(1)NeuAc(1)	0.60	9.8e-5
10	R.LENSsPQAPA.R	2.49e+7	2.434	16.96	2	245	254	HexNAc(1)Hex(1)	-0.01	0.00026
11	R.LENSsPQAPARRLLP.P	5.48e+6	1.491	72.02	3	245	259	HexNAc(1)Hex(1)NeuAc(2)	0.24	0.011
12	R.LENSsPQAPARRLLP.P	9.23e+6	2.485	43.56	3	245	259	HexNAc(1)Hex(1)	1.23	0.00043
13	R.LENSsPQAPARRLLPP.-	1.53e+7	0.415	42.34	3	245	260	HexNAc(1)Hex(1)	0.00	4.6e-8
14	R.LENSsPQAPARRLLPP.-	3.52e+7	0.948	54.54	3	245	260	HexNAc(1)Hex(1)NeuAc(1)	-0.23; ...	1.70e-7 - 8.94e
15	R.LENsSPQAPARRLLPP.-	4.29e+7	1.153	42.91	3	245	260	HexNAc(1)Hex(1)	0.09; ...	1.84e-8 - 1.84e

Figure S2. ProSAAS glycoforms from 20100805_Velos3_AnSe_Batch12faste_hypot_S3.raw. The Little LEN peptide selected above has the worst (highest) PEP 2D of 0.011 or about 1% chance of error. We show this peptide to give a sense of the threshold for acceptance of a spectrum assignment.

XIC Ratio % reports the XIC integral of the peptide in the given modification state to the sum of XIC integrals of all "peptiforms" with the same peptide sequence. All of the glycosylated species are low-abundance. Glycosylated forms account for about 4% of Little LEN, 2.5% of Big LEN, and 4% of "Medium LEN", that is, LENSSPQAPARRLLP. The peptides containing the GPTG glycosylation site, rows 5 – 7 in the table, do not appear without glycosylation, probably because the data acquisition method did not select z=1+ precursors for MS2, so these peptides do not have an XIC Ratio %. Also notice the end-of-chromatography elution times for rows 5 and 6.



PMI-Byologic - P:\PMI_Share_Data\Data2020\Jesper_Olsen_Peptidomics\Lingjun.ProSAAS.blgc - Registered (v3.9-7 x64) - Days left: 209.7

File Edit Window Help

bern (Editor)

Projects table (double click to dock / undock)

Project

Alias Name	
☑ 041213_HT...	E:\DATA\LINGJUN\
☑ 041213...	C:\data_results\04
☑ 041213_HT...	E:\DATA\LINGJUN\
☑ 041213...	C:\data_results\04
☑ 041213_HT...	E:\DATA\LINGJUN\
☑ 041213...	C:\data_results\04
☑ 041213_HT...	E:\DATA\LINGJUN\
☑ 041213...	C:\data_results\04

Peptides 8/345 rows

Sequence	Glycans	XIC Ratio%	MS Filename
R.LENSsPQAPARRLLPP.-	HexNAc(1)Hex(1)NeuAc(1)	0.105 - 1.881	E:/DATA/LINGJUN/041213_HT_A2.raw; E
R.LENSsPQAPARRLLPP.-	HexNAc(1)Hex(1)NeuAc(1)	1.701	E:/DATA/LINGJUN/041213_HT_B5.raw
R.LENSsPQAPARRLLPP.-	HexNAc(1)Hex(1)NeuAc(1)	1.377	E:/DATA/LINGJUN/041213_HT_A6.raw
R.LENSsPQAPARRLLPP.-	HexNAc(1)Hex(1)NeuAc(1)	1.881	E:/DATA/LINGJUN/041213_HT_A3.raw
R.LENSsPQAPARRLLPP.-	HexNAc(1)Hex(1)NeuAc(1)	1.657	E:/DATA/LINGJUN/041213_HT_B4.raw
R.LENSsPQAPARRLLPP.-	HexNAc(1)Hex(1)NeuAc(1)	1.169	E:/DATA/LINGJUN/041213_HT_A5.raw
R.LENSsPQAPARRLLPP.-	HexNAc(1)Hex(1)NeuAc(1)	0.105	E:/DATA/LINGJUN/041213_HT_A2.raw

Wildtype peptides

Row#	PID	Sequence	Start AA	End AA	Apex Int. (Posit)	XIC area summed	Score	z	Obs. m/z
1	70; 76; 38...	R.LENSSPQAPARRLLPP.-	245	260	2.609e...	8.48e+8	239.29 - 423.36	2	873.4825 - 873.4866
2	84; 67; 74;...	R.LENSSPQAPARRLLPP.-	245	260	6.247e...	1.84e+...	286.05 - 528.99	3	582.6578 - 582.6605

Protein coverage view (double click to dock...

Protein Coverage

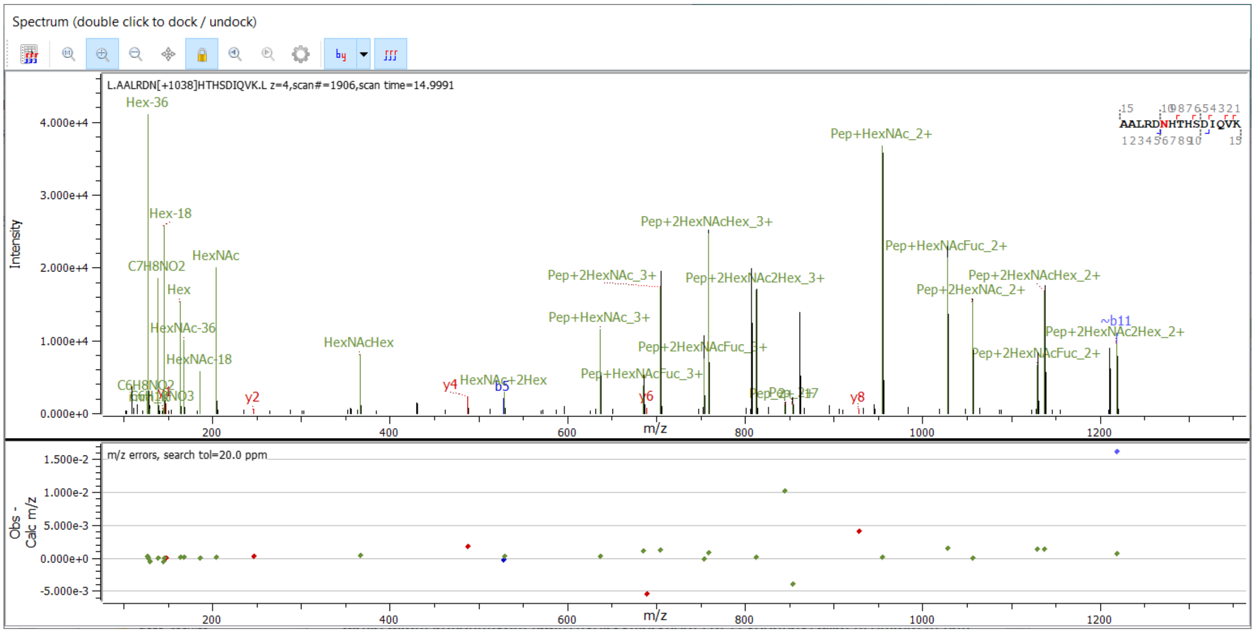
Sample name	
Currently checked	s
041213_HT_A2.raw_20200705_Byonic	s
041213_HT_A3.raw_20200705_Byonic	s
041213_HT_A5.raw_20200705_Byonic	s
041213_HT_A6.raw_20200705_Byonic	s

MS2 and mass error plots (double click to dock / undock)

XIC plots (double click to dock / undock)

Isotope plots (double click to dock / undock)

Figure S3. Big LEN glycopeptide compared across six Wisconsin data .raw files (all the files in ftp://massive.ucsd.edu/MSV000080106 with names starting “041213_HT”) from six different rat brains. The glycopeptide was identified by MS2 in only two of the data files but then transferred between runs using “Add Missing Via Existing Peptides” in Byos.

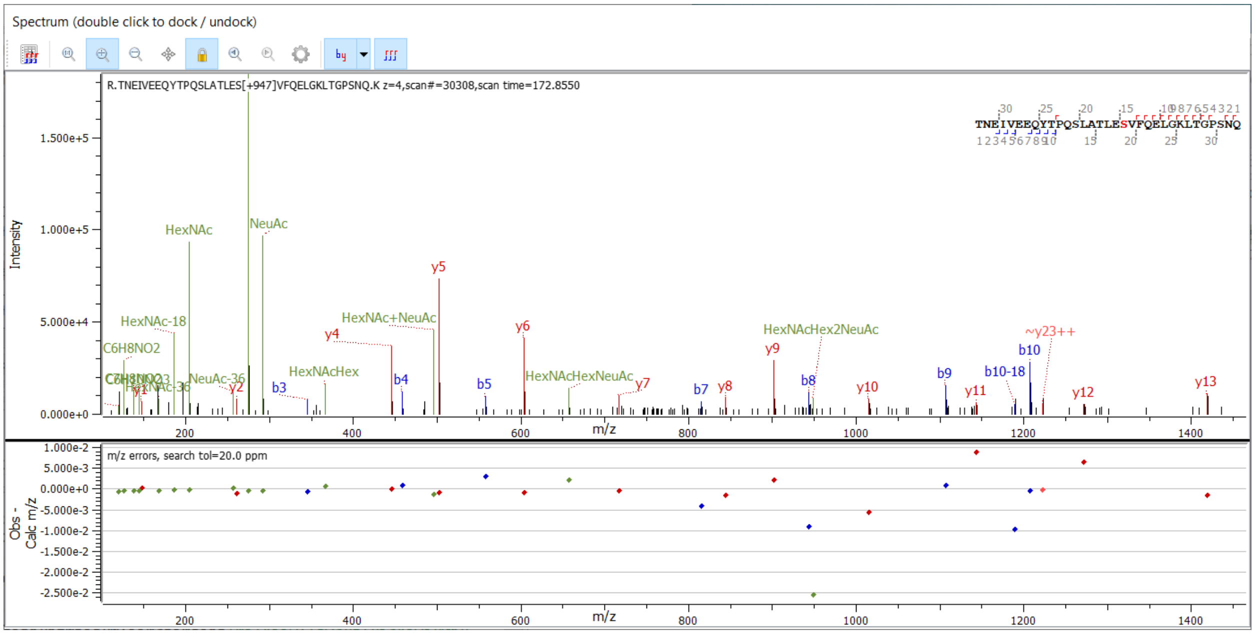


Spectrum (double click to dock / undock)

L.AALRDN[+1038]HTHSDIQVK.L z=4,scan#=1906,scan time=14.9991

m/z errors, search tol=20.0 ppm

Figure S4. Phospholipase D3 peptide, positions 380 – 393. Figures S4 – S11 show a variety of glycopeptide PSMs from the Copenhagen data set with “PEP 2D” error probabilities below 1%.

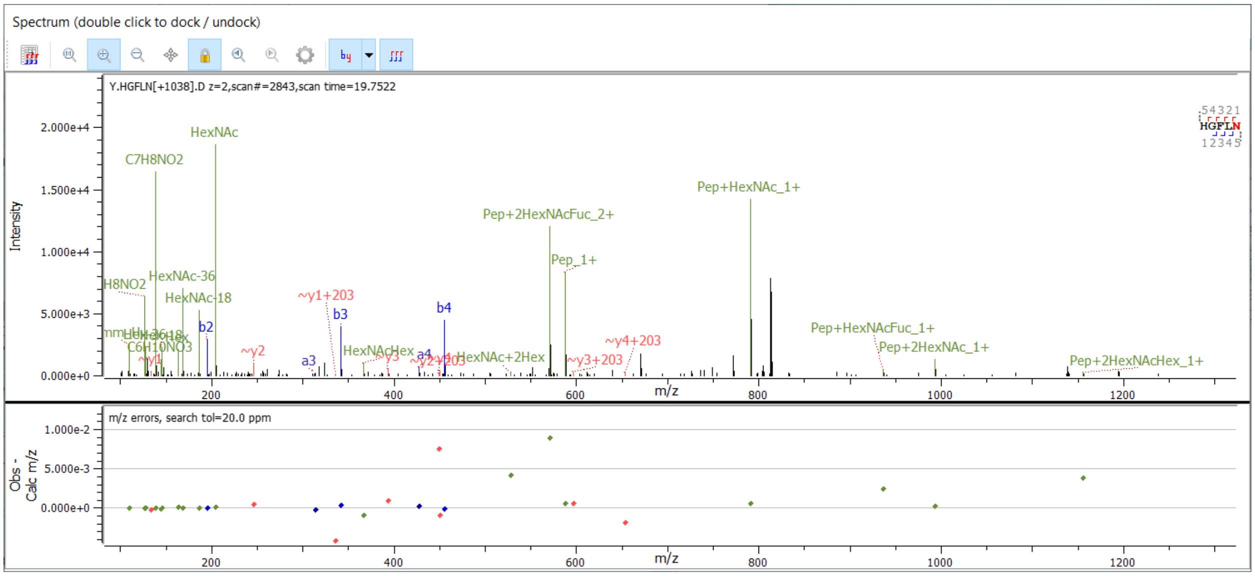


Spectrum (double click to dock / undock)

R.TNEIVEEQYTPQSLATLES[+947]VFQELGKLTGPSNQ.K z=4,scan#=30308,scan time=172.8550

m/z errors, search tol=20.0 ppm

Figure S5. Secretogranin-2 isoform X1.

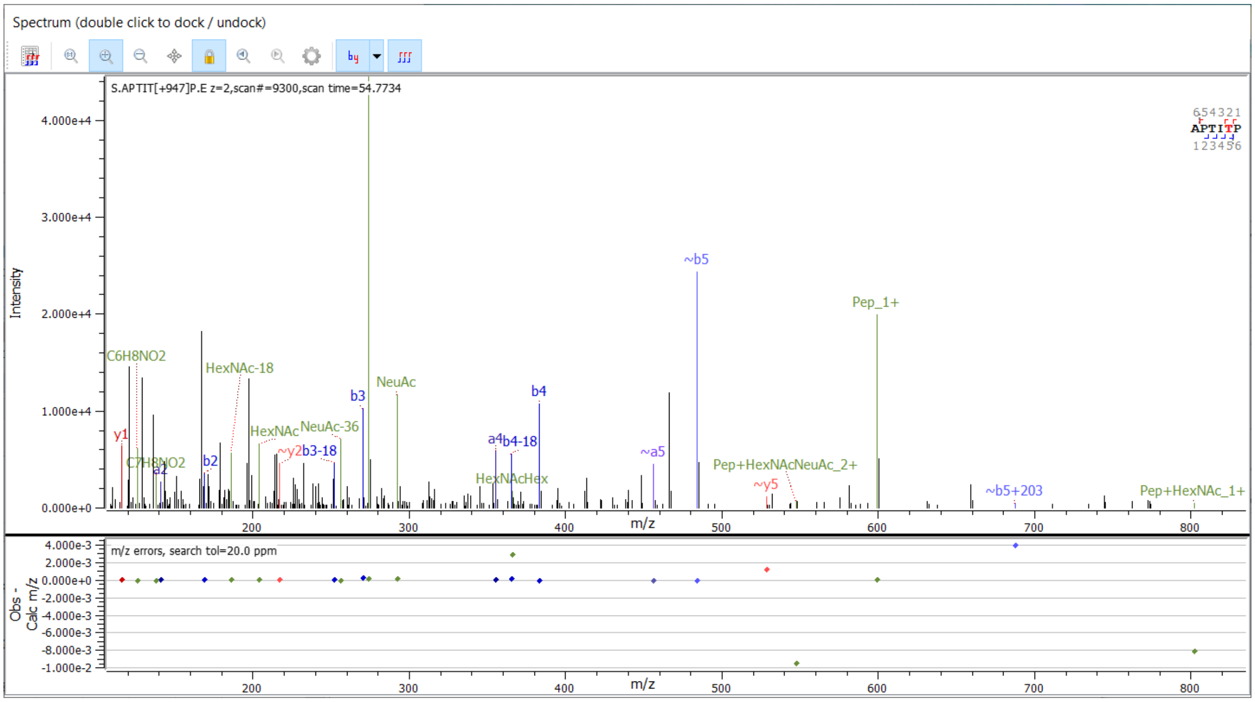


Spectrum (double click to dock / undock)

Y.HGFLN[+1038].D z=2,scan#=2843,scan time=19.7522

m/z errors, search tol=20.0 ppm

Figure S6. Peptide with an N-linked glycosylation (fucosylated N-glycan core). This short peptide appears in Lysosomal acid phosphatase isoforms with the C-terminal N in an N-glycosylation sequon, and in two other human proteins (furin and epoxide hydrolase 3) without the sequon. Short peptide matches such as those in S6, S7, and S9 must be considered somewhat ambiguous as to protein of origin.

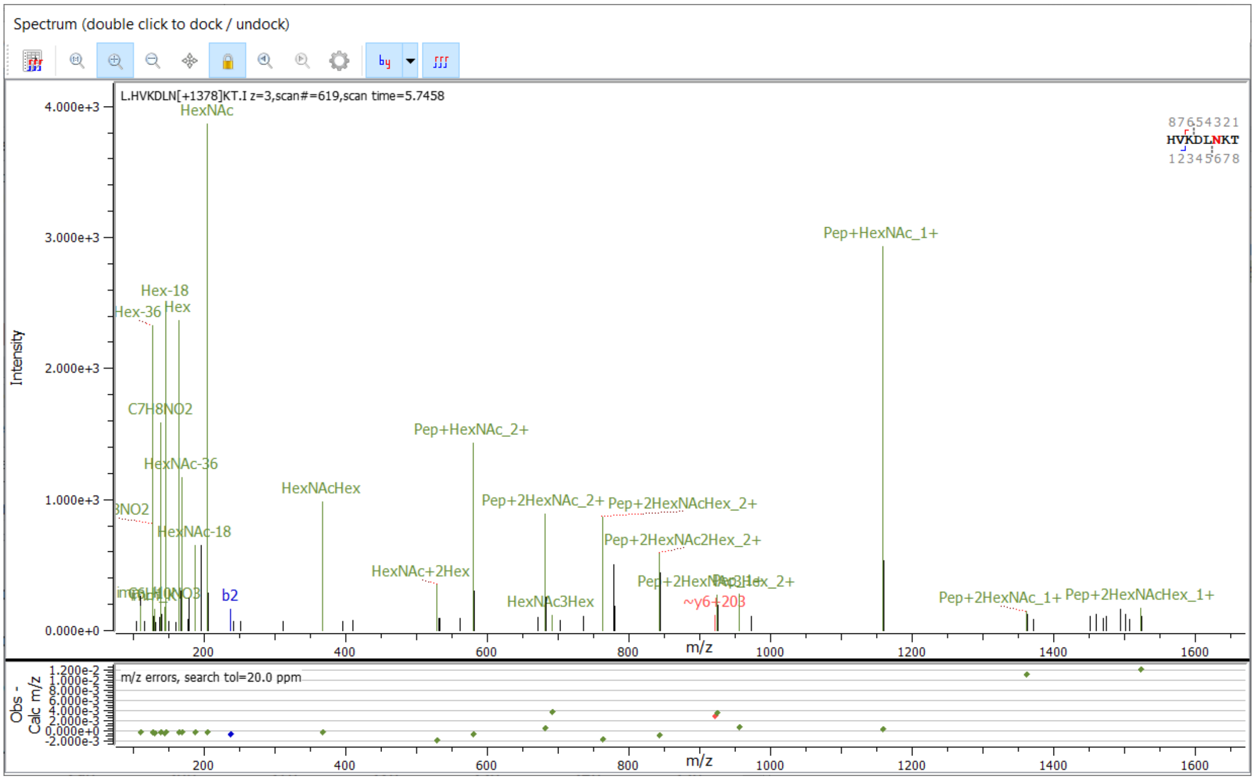


Spectrum (double click to dock / undock)

S.APTIT[+947]P.E z=2,scan#=9300,scan time=54.7734

m/z errors, search tol=20.0 ppm

Figure S7. Secretogranin-1 peptide.

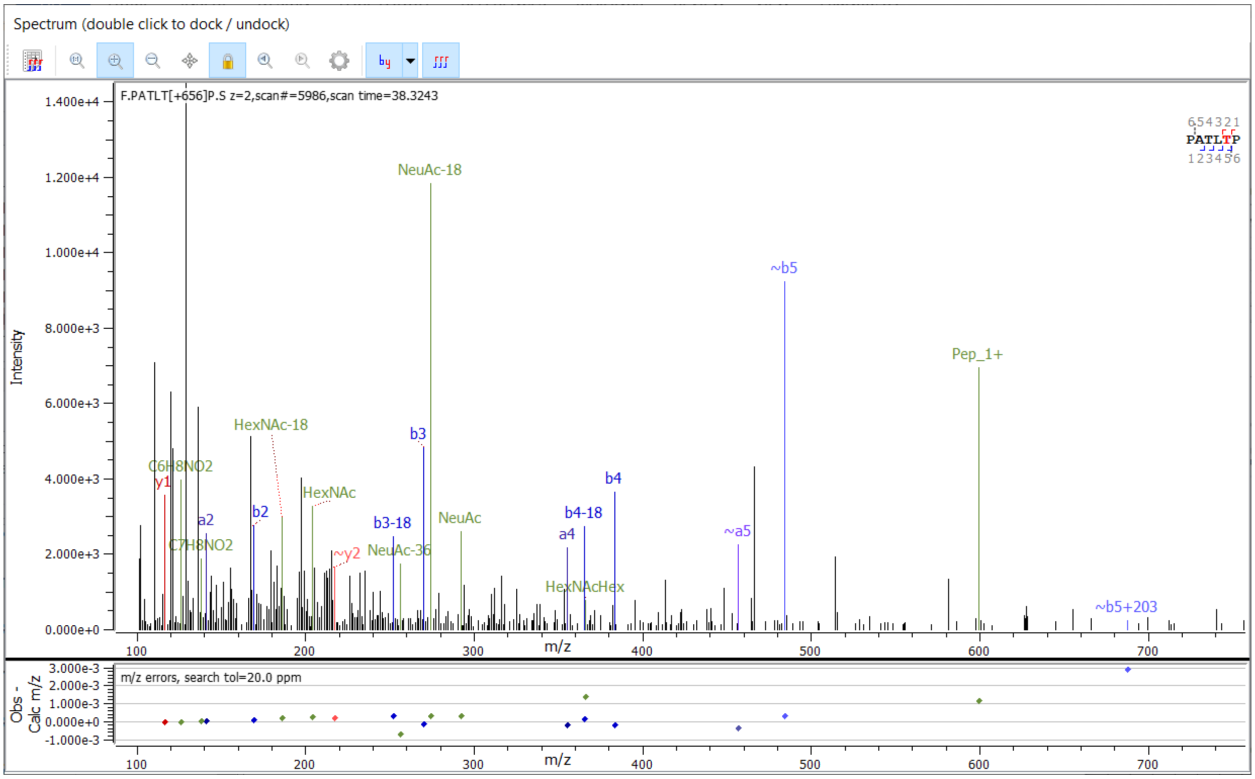


Spectrum (double click to dock / undock)

L.HVKDLN[+1378]KT.I z=3,scan#=619,scan time=5.7458

m/z errors, search tol=20.0 ppm

Figure S8. Legumain isoform X2.

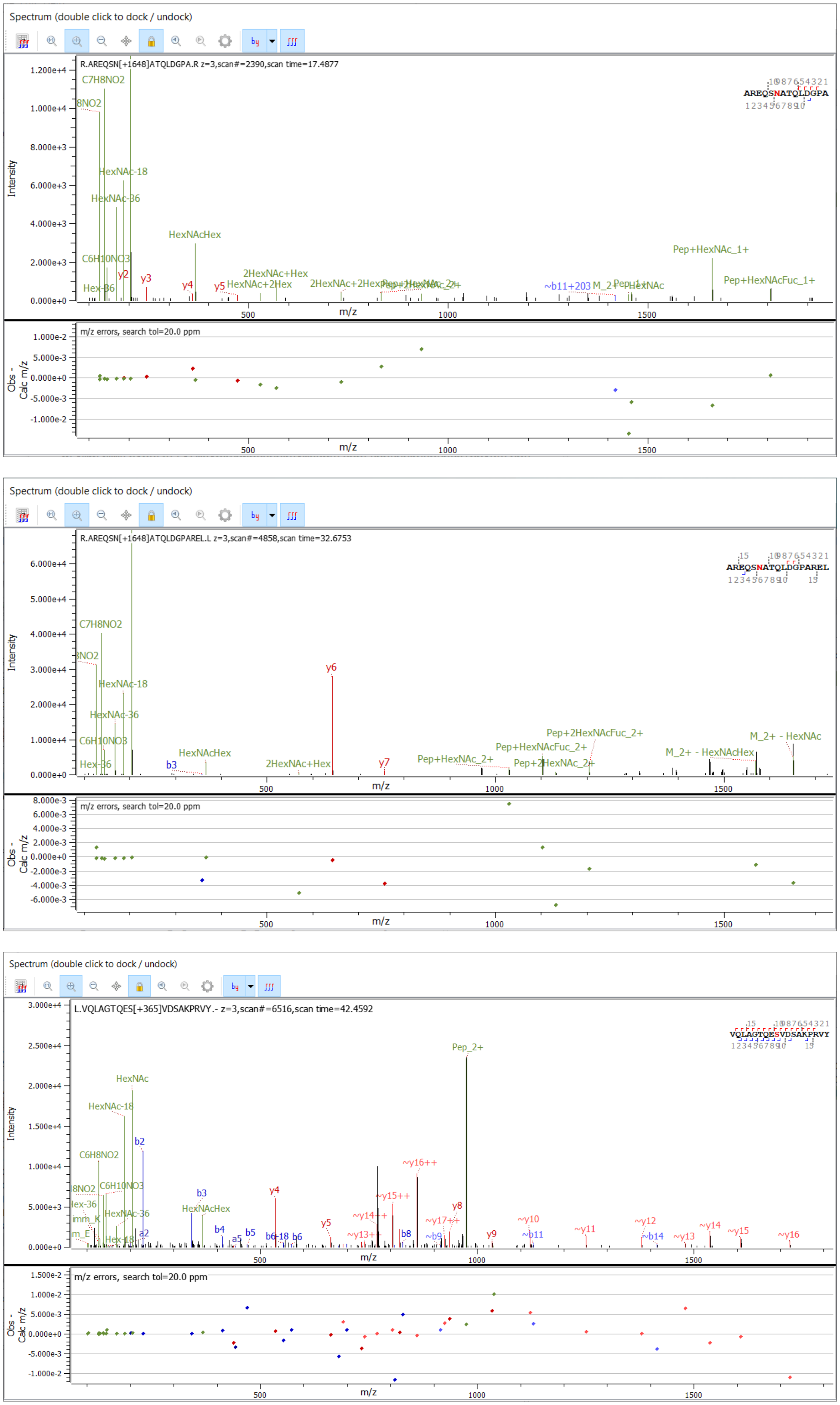


Spectrum (double click to dock / undock)

F.PATLT[+656]P.S z=2,scan#=5986,scan time=38.3243

m/z errors, search tol=20.0 ppm

Figure S9. Aftiphilin peptide.



Spectrum (double click to dock / undock)

R.AREQSN[+1648]ATQLDGPA.R z=3,scan#=2390,scan time=17.4877

m/z errors, search tol=20.0 ppm

Spectrum (double click to dock / undock)

R.AREQSN[+1648]ATQLDGPAREL.L z=3,scan#=4858,scan time=32.6753

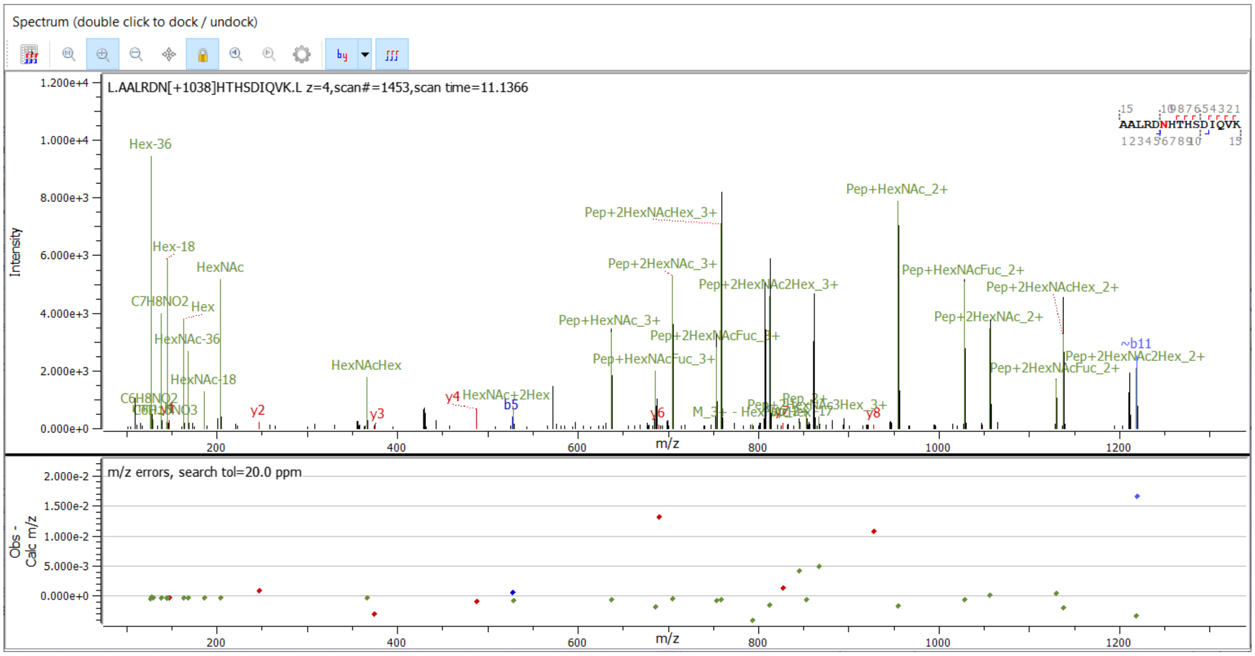
m/z errors, search tol=20.0 ppm

Spectrum (double click to dock / undock)

L.VQLAGTQES[+365]VDSAKPRVY.- z=3,scan#=6516,scan time=42.4592

m/z errors, search tol=20.0 ppm

Figure S10. Vasopressin-neurophysin 2-copeptin peptides with N- and O-linked glycosylation.

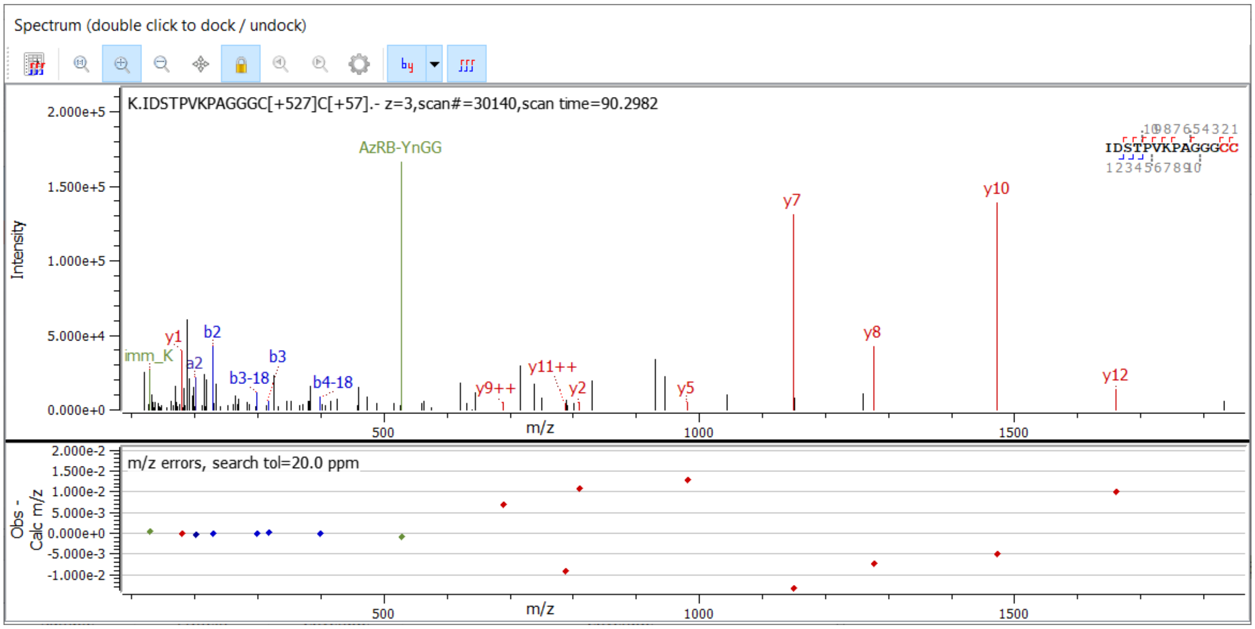


Spectrum (double click to dock / undock)

L.AALRDN[+1038]HTHSDIQVK.L z=4,scan#=1453,scan time=11.1366

m/z errors, search tol=20.0 ppm

Figure S11. 5'-3' exonuclease PLD3 with N-linked glycosylation.

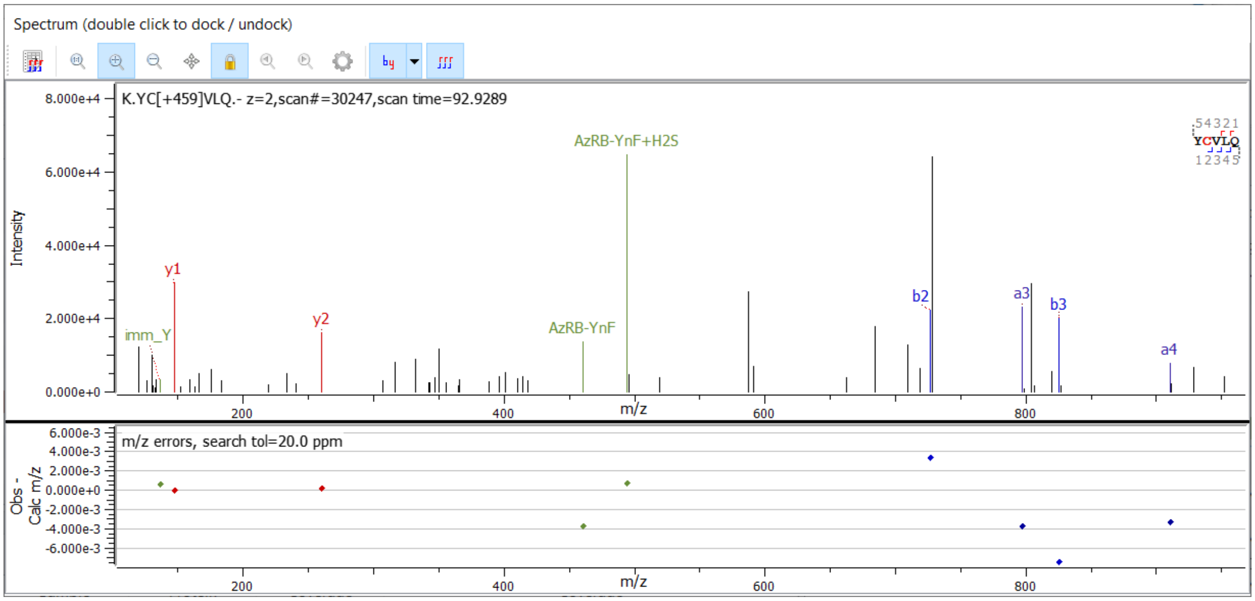


Spectrum (double click to dock / undock)

K.IDSTPVKPAGGGC[+527]C[+57].- z=3,scan#=30140,scan time=90.2982

m/z errors, search tol=20.0 ppm

Figure S12. Scan 30140 in JMS133H.raw from PXD009155 matched to the C-terminal peptide from RAB1B_HUMAN (Ras-related protein Rab-1B) with AzRB-captured YnGG. Uniprot KB lists the second cysteine as a geranylgeranylation site. The prenylated peptides in S12 – S15 were not reported in the Supplemental to the Storck et al publication (PMC6544531).

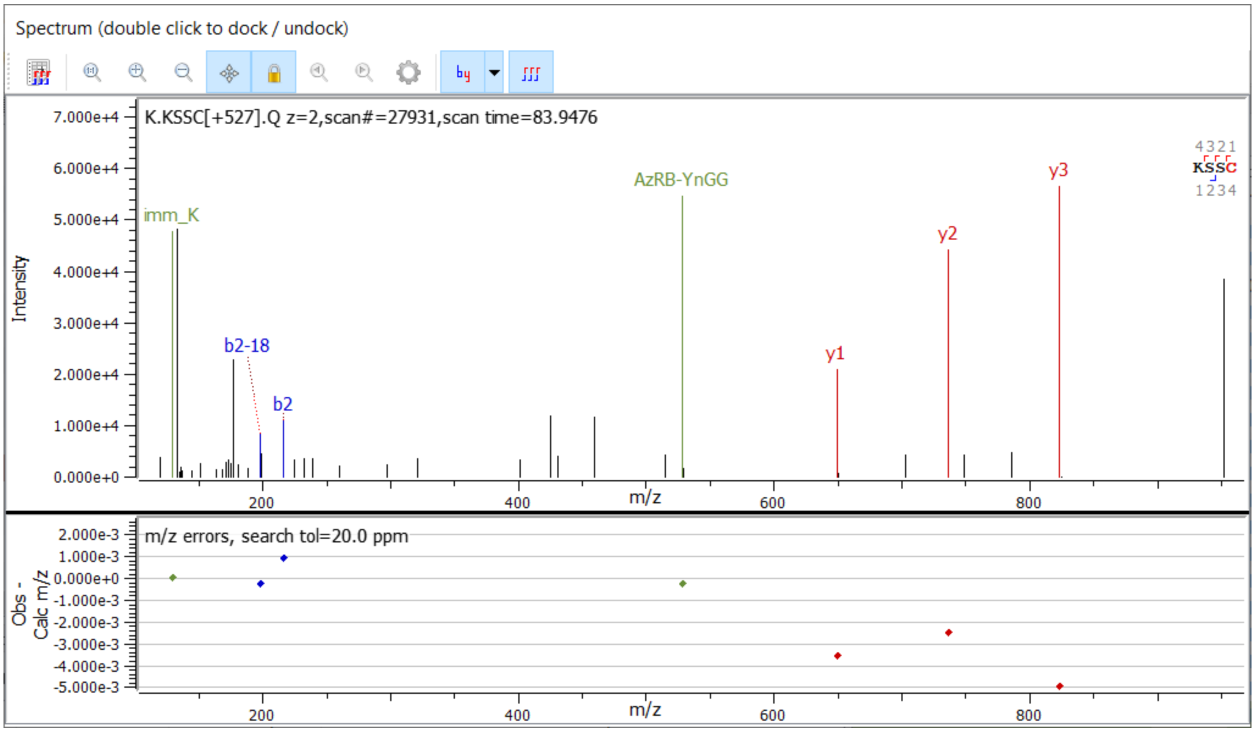


Spectrum (double click to dock / undock)

K.YC[+459]VLQ.- z=2,scan#=30247,scan time=92.9289

m/z errors, search tol=20.0 ppm

Figure S13. C-terminal peptide from UBP32_HUMAN (Ubiquitin carboxyl-terminal hydrolase 32) with AzRB-captured YnF (JMS133E.raw). Uniprot KB annotates this cysteine farnesylation site from "sequence analysis". As above, the protein of origin may be ambiguous for short peptide matches. Farnesylation, however, occurs near protein C-termini, usually with a Cxxx- motif, where – denotes protein C-terminus, which greatly limits the number of possible proteins.

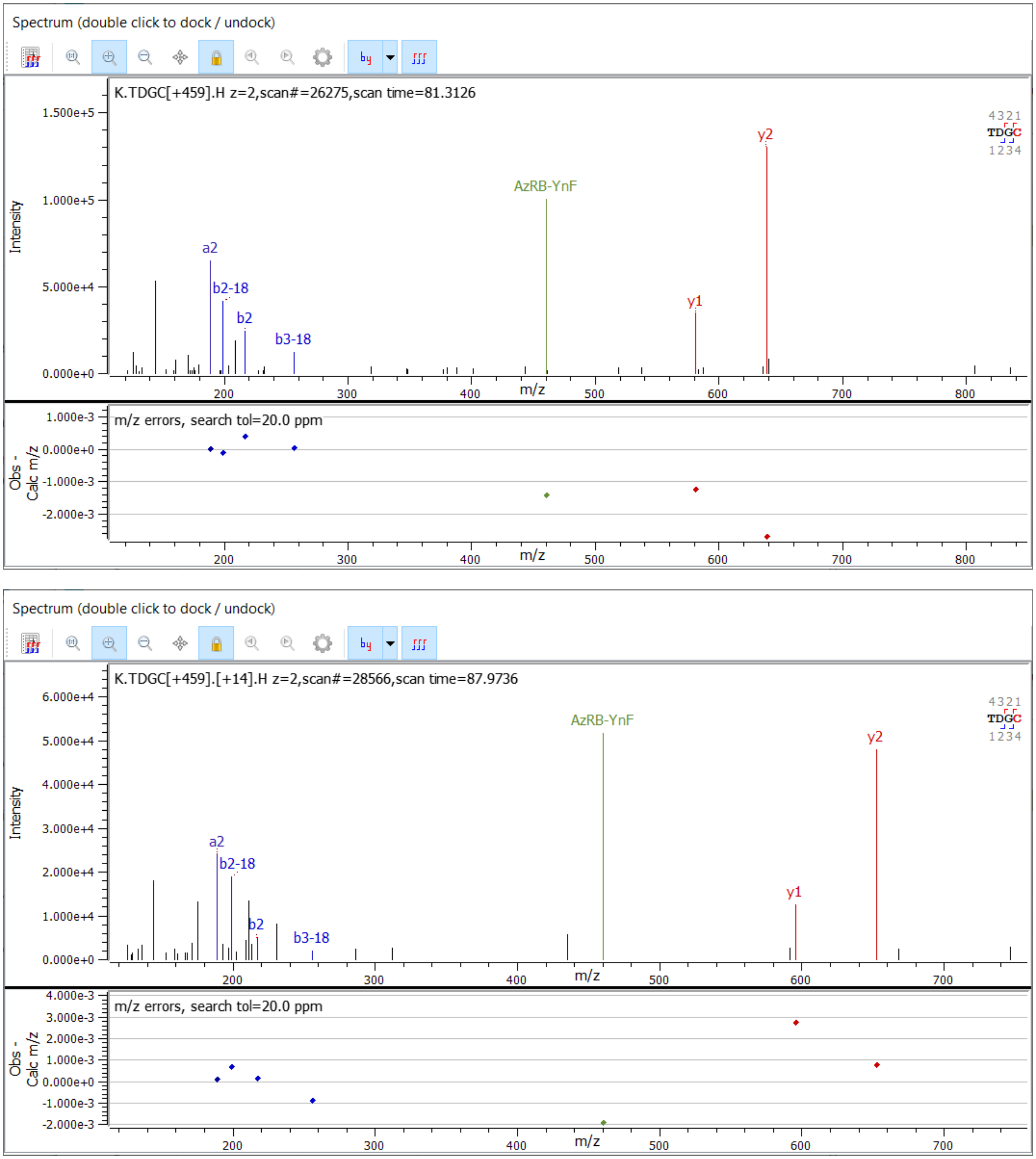


Spectrum (double click to dock / undock)

K.KSSC[+527].Q z=2,scan#=27931,scan time=83.9476

m/z errors, search tol=20.0 ppm

Figure S14. Peptide (C-terminal after processing) from RAP1B_HUMAN (Ras-related protein Rap-1b) with AzRB-captured YnGG (JMS133H.raw). Uniprot KB annotates this geranylgeranylation site. The unannotated peak at *m/z* 951.52 corresponds to a singly charged version of the peptide. Perhaps the precursor was a dimer?



Spectrum (double click to dock / undock)

K.TDGC[+459].H z=2,scan#=26275,scan time=81.3126

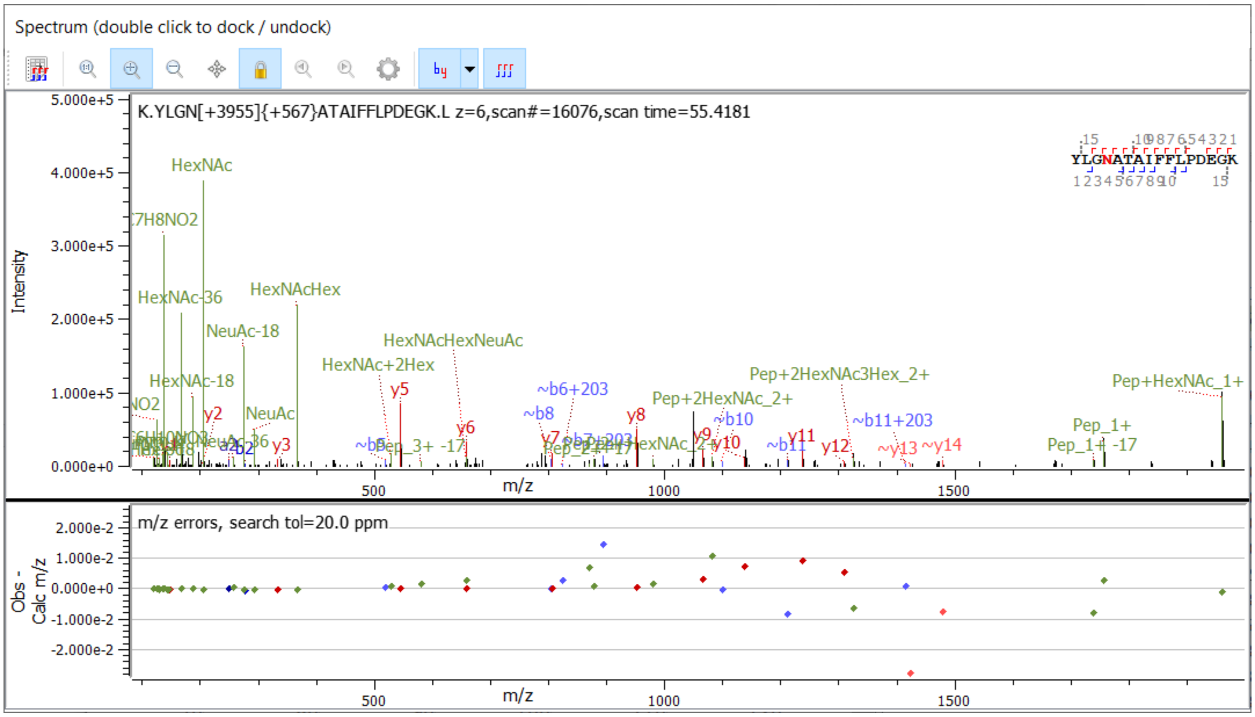
m/z errors, search tol=20.0 ppm

Spectrum (double click to dock / undock)

K.TDGC[+459].[+14].H z=2,scan#=28566,scan time=87.9736

m/z errors, search tol=20.0 ppm

Figure S15. Spectra of a peptide with AzRB-captured YnF (JMS133E.raw) with and without C-terminal methylation. Byonic matched these spectra to CAP1_HUMAN, but this is probably not the protein of origin, because the peptide is not very close to the C-terminus of CAP1_HUMAN.



Spectrum (double click to dock / undock)

K.YLGN[+3955]{+567}ATAIFFLPDEGK.L z=6,scan#=16076,scan time=55.4181

m/z errors, search tol=20.0 ppm

Figure S16. The peptide-spectrum match above is from a search of HUPO HGI data (file B_glycopepnew_HCDEThcDiTCIDpeptide.raw) that used MS2 Filtering Peaks to target a particular peptide in alpha-1-antitrypsin and Glycan Wildcard Search to find glycans not in the glycan database. The database glycan of mass 3955 Da is HexNAc(6)Hex(7)Fuc(3)NeuAc(4). The wildcard of mass 567 Da is probably HexNAc(2)Hex(1) with an off-by-one error in the monoisotopic mass call.

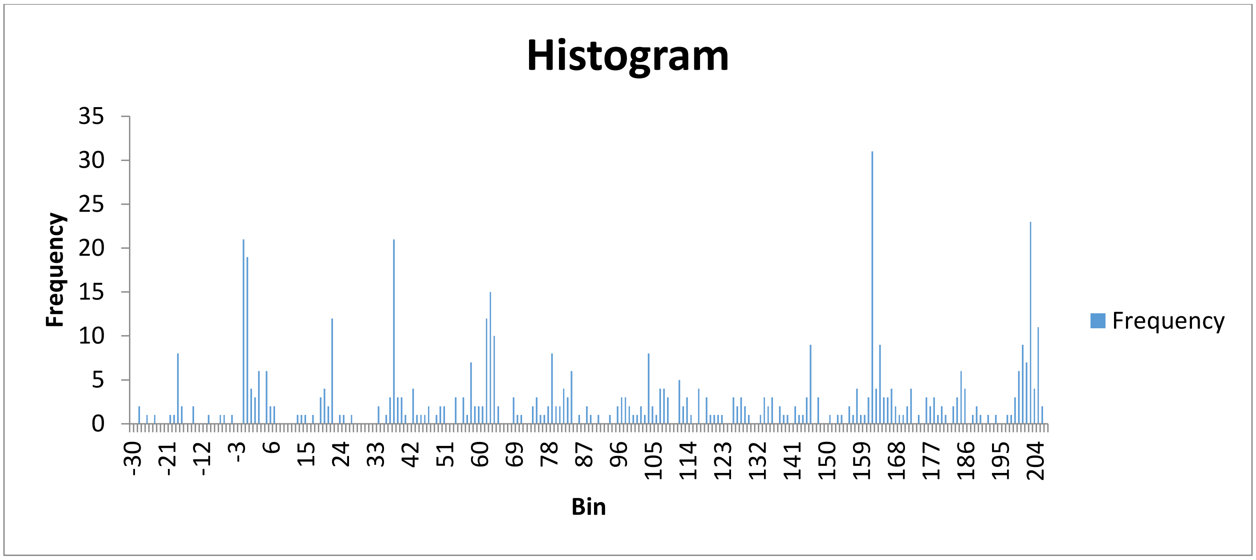
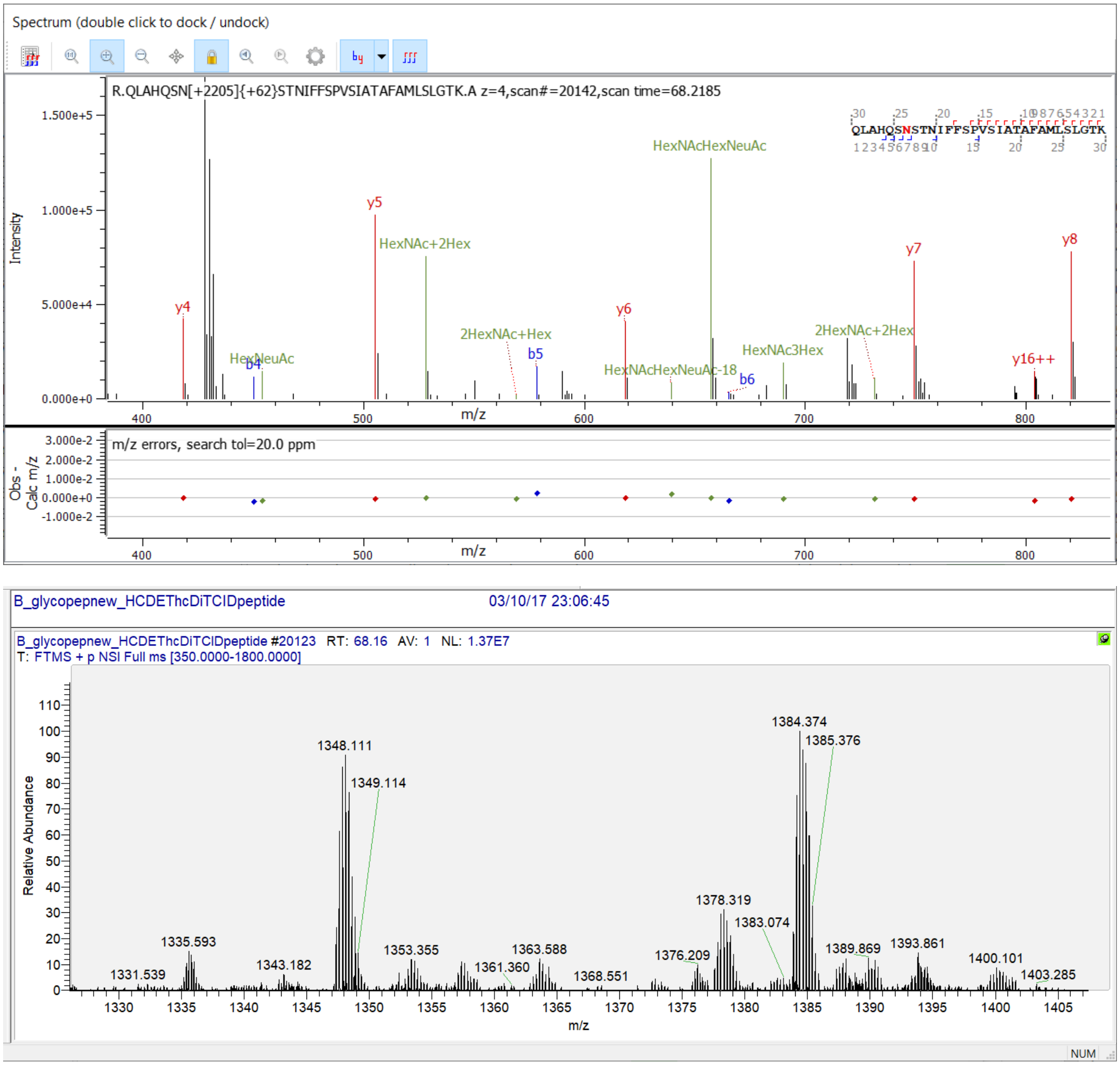


Figure S17. A histogram showing the masses of glycan wildcards on 547 high-scoring glycopeptide PSMs.



Spectrum (double click to dock / undock)

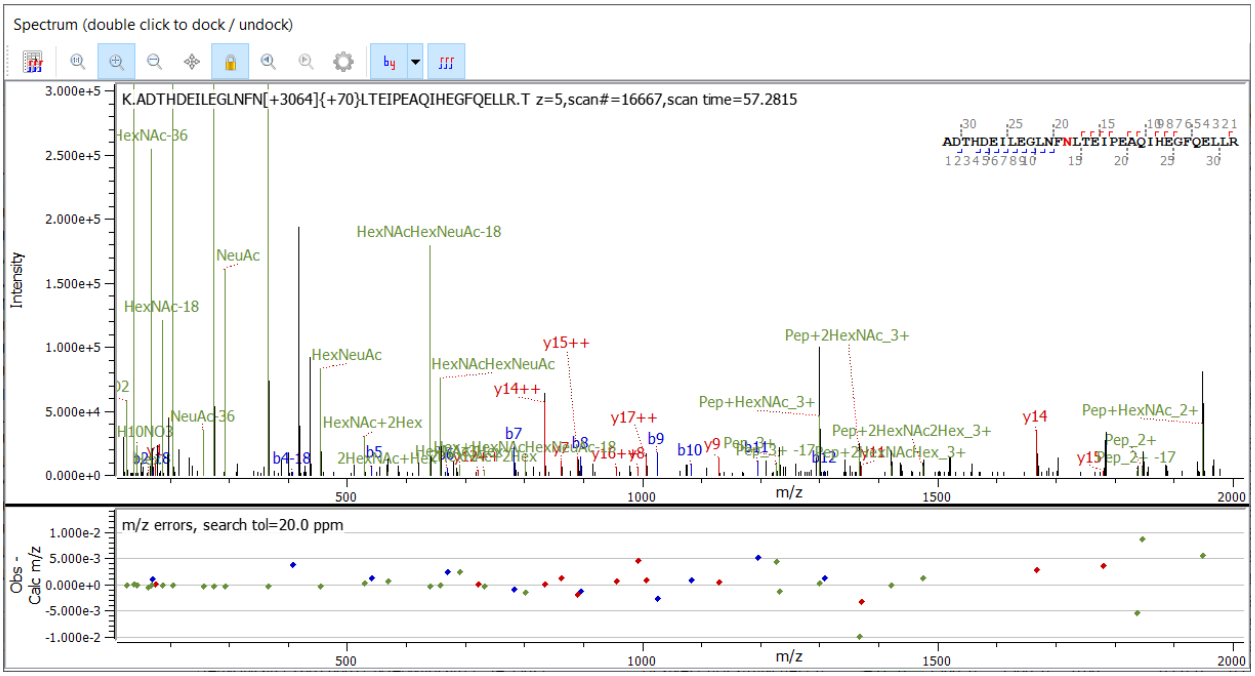
R.QLAHQSN[+2205]{+62}STNIFFSPVSIATAFAMLSLGTK.A z=4,scan#=20142,scan time=68.2185

m/z errors, search tol=20.0 ppm

B_glycopepnew_HCDEThcDiTCIDpeptide 03/10/17 23:06:45

B_glycopepnew_HCDEThcDiTCIDpeptide #20123 RT: 68.16 AV: 1 NL: 1.37E7
T: FTMS + p NSI Full ms [350.0000-1800.0000]

Figure S18. A glycan wildcard PSM with wildcard mass 61.916 Da. The database glycan of mass 2205 Da is HexNAc(4)Hex(5)NeuAc(2), one of the most common glycans in human plasma. We speculate that the wildcard represents Cu^+ replacing a proton, which has mass delta 61.922 Da. The precursor isotope cluster for the MS2 spectrum is the *z*=4+ group labeled 1363.588. The much larger cluster at 1348.11 is the same glycopeptide without copper, and 1384.374 is the same peptide with HexNAc(4)Hex(5)Fuc(1)NeuAc(2).



Spectrum (double click to dock / undock)

K.ADTHDEILEGLNFN[+3064]{+70}LTEIPEAQIHEGFQELLR.T z=5,scan#=16667,scan time=57.2815

m/z errors, search tol=20.0 ppm

Figure S19. A glycan wildcard PSM. The database glycan of mass 3064 Da is HexNAc(6)Hex(6)NeuAc(3), but we do not have an explanation for the extra 70.056 Da.