Developing Reproducible Methods in Site-Specific Glycosylation Analysis

Zachary Goecker, Meghan Burke, Concepcion Remoroza, Yi Liu, Yuri Mirokhin, Sergey Sheetlin, Guanghui Wang, Dmitrii Tchekhovskoi, Xiaoyu Yang, and Stephen Stein

WBMSDG Seminar

Feb 27, 2023

National Institute of Standards and Technology
U.S. Department of Commerce

Mass Spectrometry
Data Center

NIST Mass Spectrometry Data Center

NIST20

NIST/EPA/NIH EI Library
NIST Tandem Library

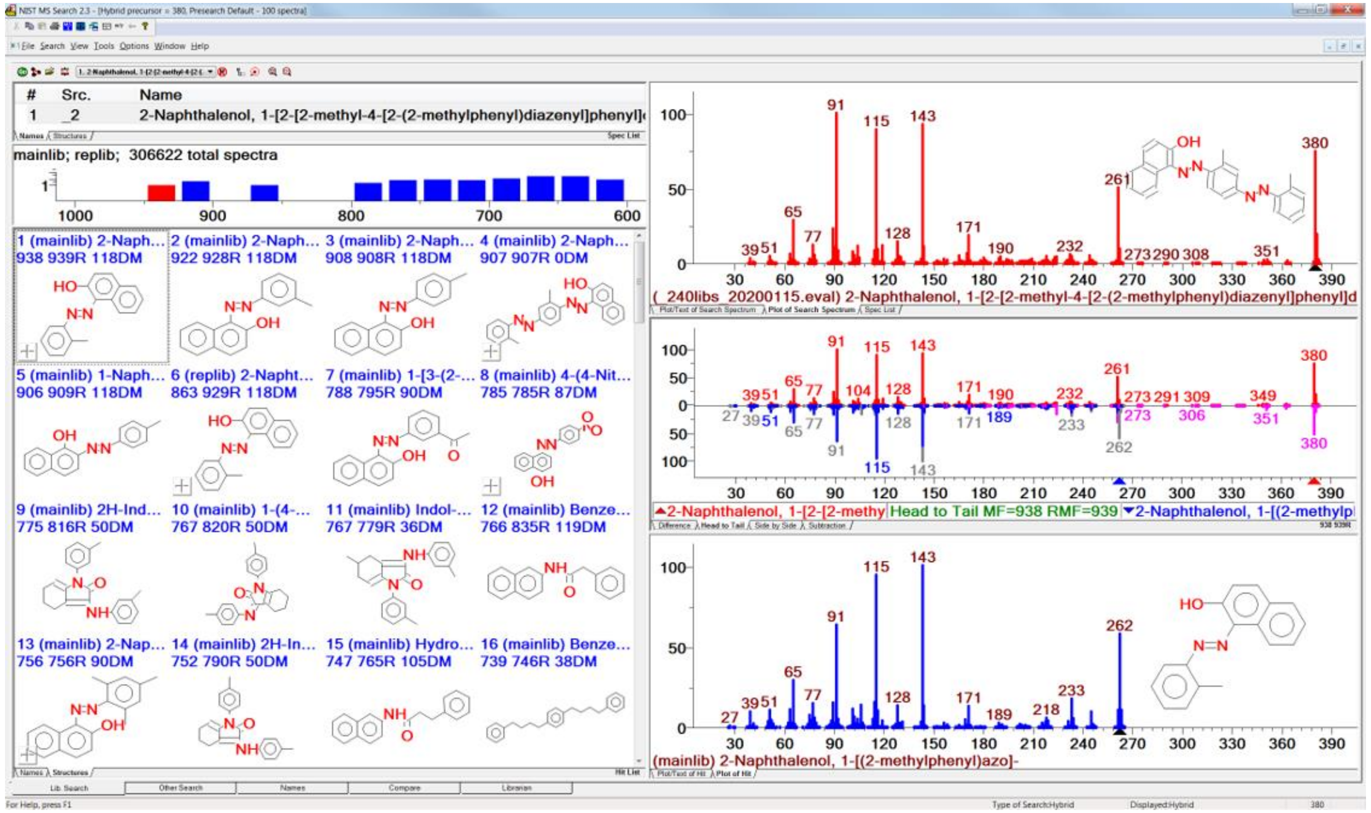
Search/Analysis Programs and Data
for Microsoft Windows XP (SP3), Vista,
Windows 7, 8, 8.1, 10

- Search Program
- Combined EI and Tandem Libraries
- Mass Spec Interpreter
- AMDIS

EI: 351K EI Mass Spectra of 307K
RI: 447K Retention Indices of 140K Compounds
Tandem: 1.3M Spectra, 186K Precursors, 31K Compounds

NIST20

June 2020



NIST MS Search 2.3 - [Hybrid precursor = 380, Presearch Default - 100 spectra]

#	Src.	Name
1	_2	2-Naphthalenol, 1-[2-[2-methyl-4-[2-(2-methylphenyl)diazenyl]phenyl]

mainlib; replib; 306622 total spectra

1 (mainlib) 2-Naph... 938 939R 118DM
2 (mainlib) 2-Naph... 922 928R 118DM
3 (mainlib) 2-Naph... 908 908R 118DM
4 (mainlib) 2-Naph... 907 907R 0DM
5 (mainlib) 1-Naph... 906 909R 118DM
6 (replib) 2-Napht... 863 929R 118DM
7 (mainlib) 1-[3-(2-... 788 795R 90DM
8 (mainlib) 4-(4-Nit... 785 785R 87DM
9 (mainlib) 2H-Ind... 775 816R 50DM
10 (mainlib) 1-(4-... 767 820R 50DM
11 (mainlib) Indol-... 767 779R 36DM
12 (mainlib) Benze... 766 835R 119DM
13 (mainlib) 2-Nap... 756 756R 90DM
14 (mainlib) 2H-In... 752 790R 50DM
15 (mainlib) Hydro... 747 765R 105DM
16 (mainlib) Benze... 739 746R 38DM

(240libs_20200115.eval) 2-Naphthalenol, 1-[2-[2-methyl-4-[2-(2-methylphenyl)diazenyl]phenyl]d

2-Naphthalenol, 1-[2-[2-methy Head to Tail MF=938 RMF=939 2-Naphthalenol, 1-[(2-methylp

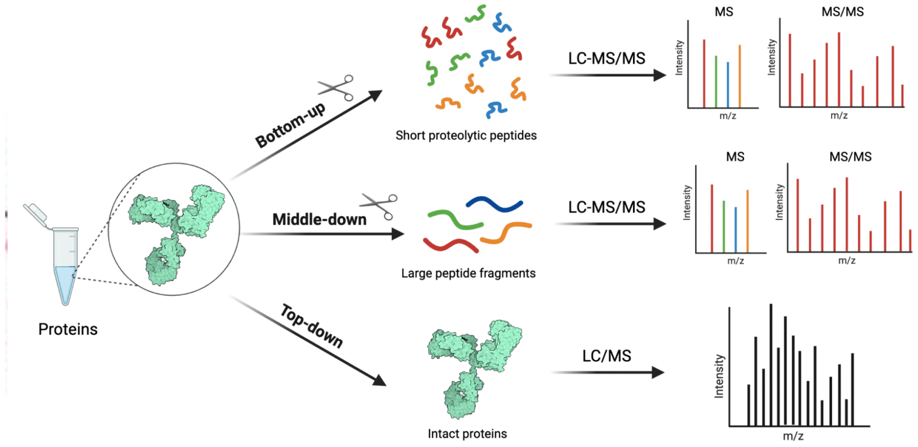
(mainlib) 2-Naphthalenol, 1-[(2-methylphenyl)azo]-

For Help, press F1

Mass Spectrometry of Glycoproteins

Analysis methods

- NMR
  - Glycan structure
  - Segmental dynamics
- Mass Spectrometry
  - Released Glycan
  - Intact glycopeptides (site-specific)



Human Milk Proteins

- Lactoferrin
- IgA1, IgA2, IgJ
- PIGR
- Tenascin

Virus Proteins

- Influenza HA + NA
- Sars-CoV-2 Spike

Plasma Proteins

- Transferrin
- Haptoglobin
- 30+ other serum glycoproteins

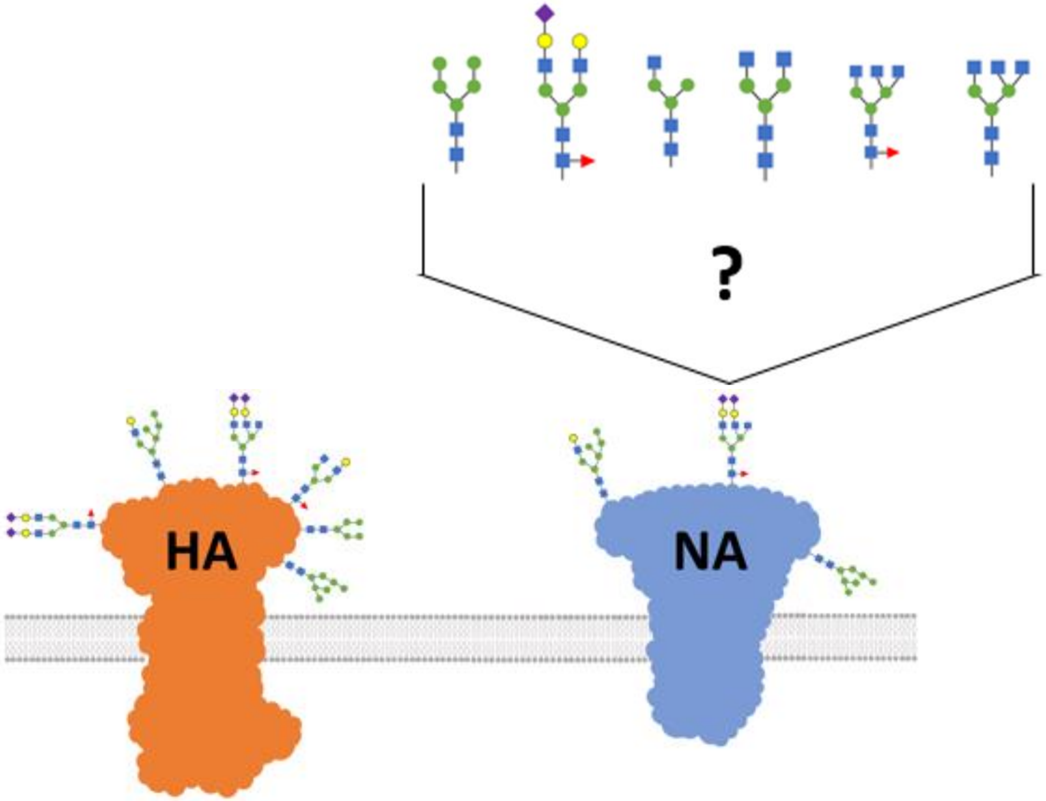
Other Relevant Proteins

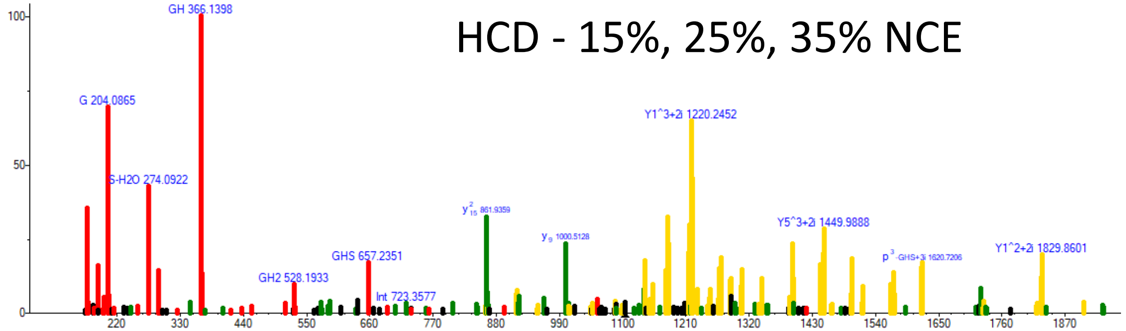
- Prostate specific antigen

Biologics (mAb)

- Repatha
- Remicade
- Embrel
- 3+ other mAb

Site-Specific Glycosylation



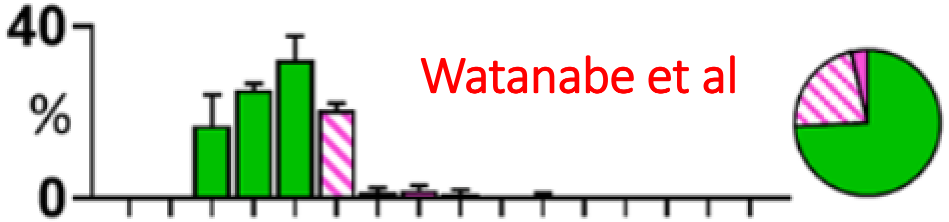


HCD - 15%, 25%, 35% NCE

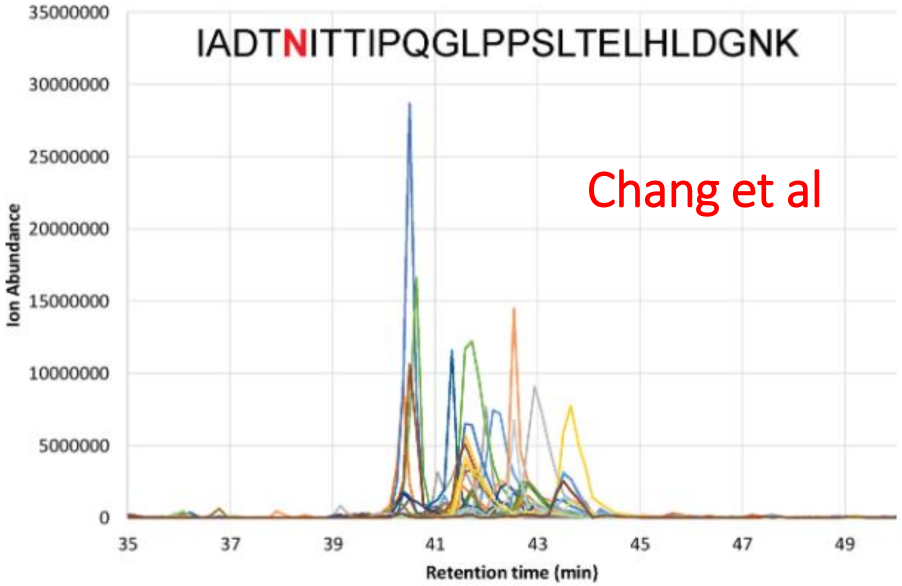
Oxonium ions
Peptide backbone ions
Glycopeptide ions

Visualizing Glycan Distributions

1. Glycopeptide search using Byonic software
2. Tandem library creation and spectral validation via create_glycopeptide_lib.exe and MS_Piano.exe.
3. GADS creation using make-gads.exe



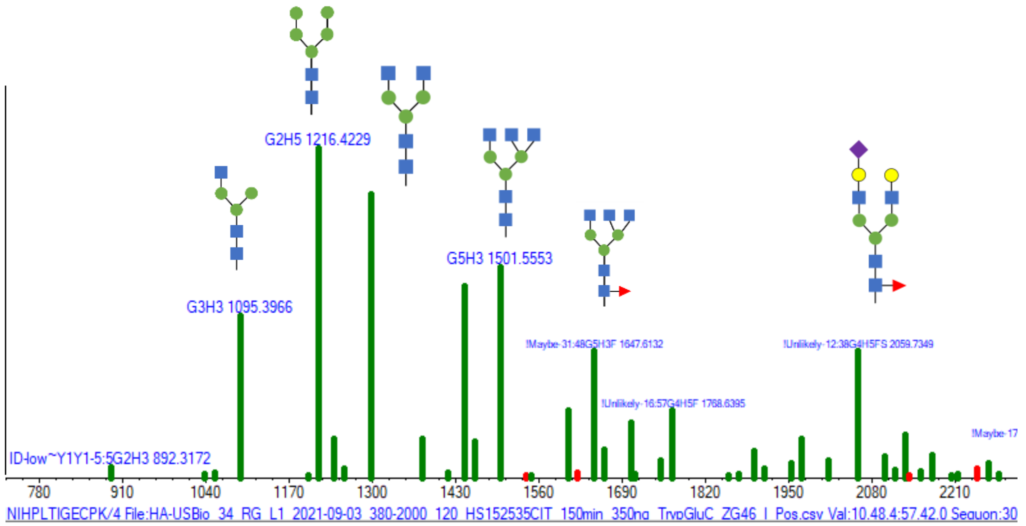
Watanabe et al



Chang et al

Visualizing Glycan Distributions

1. Glycopeptide search using Byonic software
2. Tandem library creation and spectral validation via create_glycopeptide_lib.exe and MS_Piano.exe.
3. GADS creation using make-gads.exe



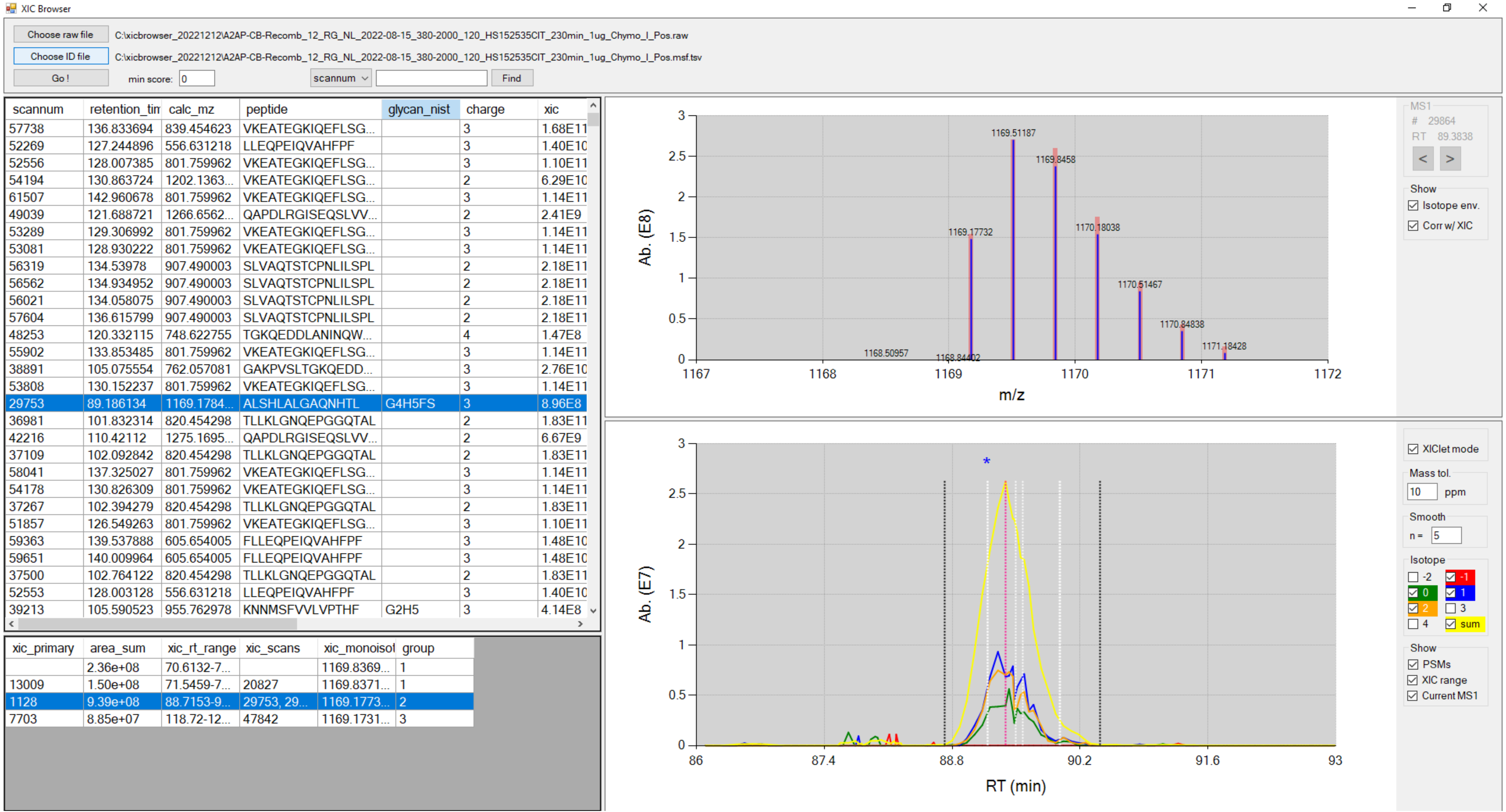
Glycopeptide Abundance Distribution Spectrum

Data Validation

- Hyperscore
- Retention time
- Contingent ion trap scan
- Oxonium ions
- Number of instances
- Glycopeptide (Y) ion series
- MS1 XIC
- MS1 Isotopic envelope
- MS2 purity
- Precursor mass overlap

$$\text{hyperscore} = \log\left(N_b!\,N_y!\sum_{i=1}^{N_b} I_{\mathrm{b,i}} \sum_{i=1}^{N_y} I_{\mathrm{y,i}} \right)$$

XIC Browser



XIC Browser

Choose raw file C:\xicbrowser_20221212\A2AP-CB-Recomb_12_RG_NL_2022-08-15_380-2000_120_HS152535CIT_230min_1ug_Chymo_I_Pos.raw

Choose ID file C:\xicbrowser_20221212\A2AP-CB-Recomb_12_RG_NL_2022-08-15_380-2000_120_HS152535CIT_230min_1ug_Chymo_I_Pos.msf.tsv

Go ! min score: 0 scannum Find

scannum	retention_tir	calc_mz	peptide	glycan_nist	charge	xic
57738	136.833694	839.454623	VKEATEGKIQEFLSG...		3	1.68E11
52269	127.244896	556.631218	LLEQPEIQVAHFPF		3	1.40E10
52556	128.007385	801.759962	VKEATEGKIQEFLSG...		3	1.10E11
54194	130.863724	1202.1363...	VKEATEGKIQEFLSG...		2	6.29E10
61507	142.960678	801.759962	VKEATEGKIQEFLSG...		3	1.14E11
49039	121.688721	1266.6562...	QAPDLRGISEQSLVV...		2	2.41E9
53289	129.306992	801.759962	VKEATEGKIQEFLSG...		3	1.14E11
53081	128.930222	801.759962	VKEATEGKIQEFLSG...		3	1.14E11
56319	134.53978	907.490003	SLVAQTSTCPNLILSPL		2	2.18E11
56562	134.934952	907.490003	SLVAQTSTCPNLILSPL		2	2.18E11
56021	134.058075	907.490003	SLVAQTSTCPNLILSPL		2	2.18E11
57604	136.615799	907.490003	SLVAQTSTCPNLILSPL		2	2.18E11
48253	120.332115	748.622755	TGKQEDDLANINQW...		4	1.47E8
55902	133.853485	801.759962	VKEATEGKIQEFLSG...		3	1.14E11
38891	105.075554	762.057081	GAKPVSLTGKQEDD...		3	2.76E10
53808	130.152237	801.759962	VKEATEGKIQEFLSG...		3	1.14E11
29753	89.186134	1169.1784...	ALSHLALGAQNHTL	G4H5FS	3	8.96E8
36981	101.832314	820.454298	TLLKLGNQEPGGQTAL		2	1.83E11
42216	110.42112	1275.1695...	QAPDLRGISEQSLVV...		2	6.67E9
37109	102.092842	820.454298	TLLKLGNQEPGGQTAL		2	1.83E11
58041	137.325027	801.759962	VKEATEGKIQEFLSG...		3	1.14E11
54178	130.826309	801.759962	VKEATEGKIQEFLSG...		3	1.14E11
37267	102.394279	820.454298	TLLKLGNQEPGGQTAL		2	1.83E11
51857	126.549263	801.759962	VKEATEGKIQEFLSG...		3	1.10E11
59363	139.537888	605.654005	FLLEQPEIQVAHFPF		3	1.48E10
59651	140.009964	605.654005	FLLEQPEIQVAHFPF		3	1.48E10
37500	102.764122	820.454298	TLLKLGNQEPGGQTAL		2	1.83E11
52553	128.003128	556.631218	LLEQPEIQVAHFPF		3	1.40E10
39213	105.590523	955.762978	KNNMSFVVLVPTHF	G2H5	3	4.14E8

xic_primary	area_sum	xic_rt_range	xic_scans	xic_monoisol	group
	2.36e+08	70.6132-7...		1169.8369...	1
13009	1.50e+08	71.5459-7...	20827	1169.8371...	1
1128	9.39e+08	88.7153-9...	29753, 29...	1169.1773...	2
7703	8.85e+07	118.72-12...	47842	1169.1731...	3

MS1 # 29864 RT 89.3838

Show ☑ Isotope env. ☑ Corr w/ XIC

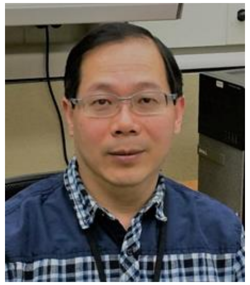
☑ XIClet mode

Mass tol. 10 ppm

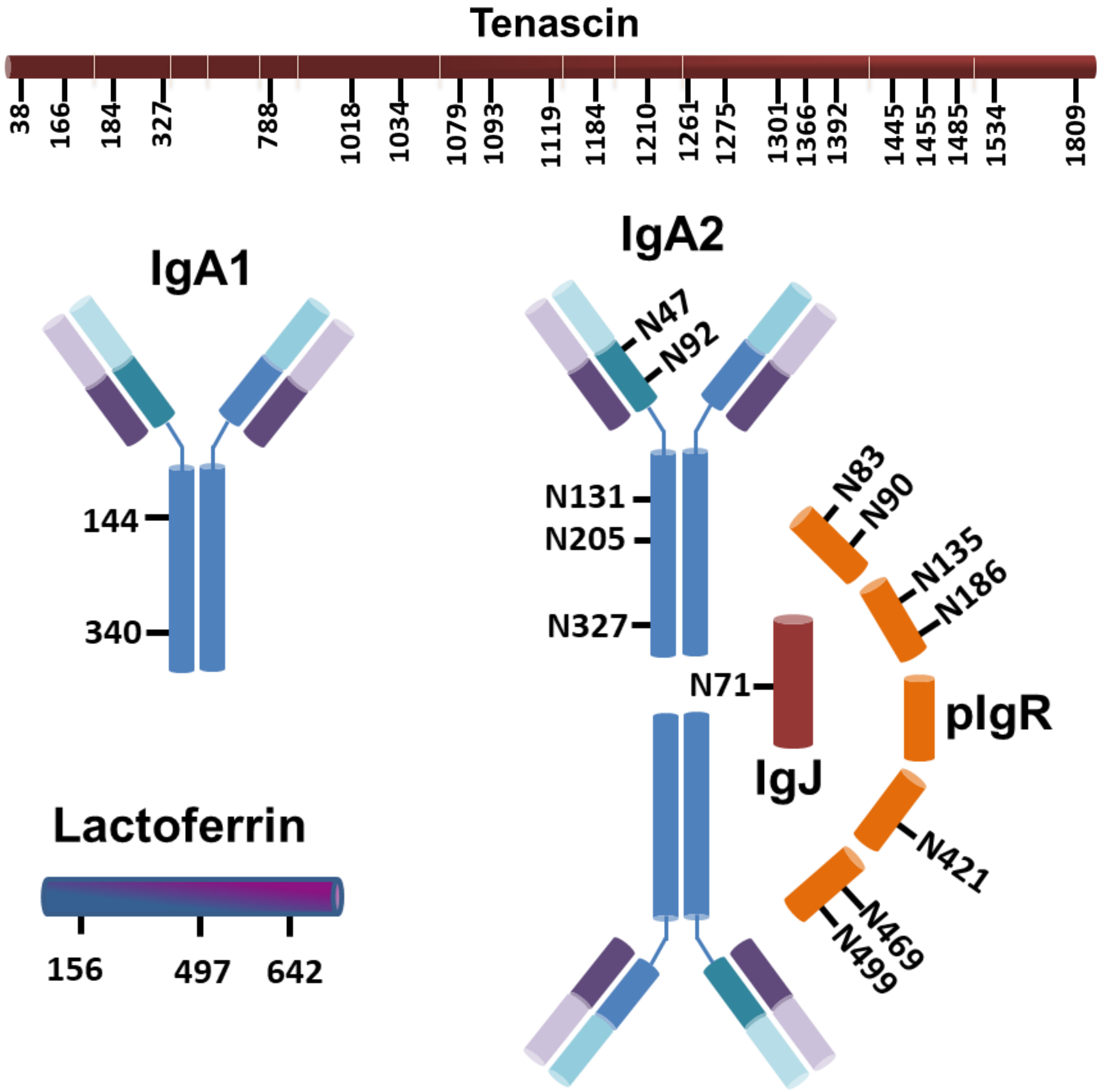
Smooth n = 5

Isotope ☐ -2 ☑ -1 ☑ 0 ☑ 1 ☑ 2 ☐ 3 ☐ 4 ☑ sum

Show ☑ PSMs ☑ XIC range ☑ Current MS1



Milk Glycoproteins



Purified proteins

Human milk (NIST SRM 1953)

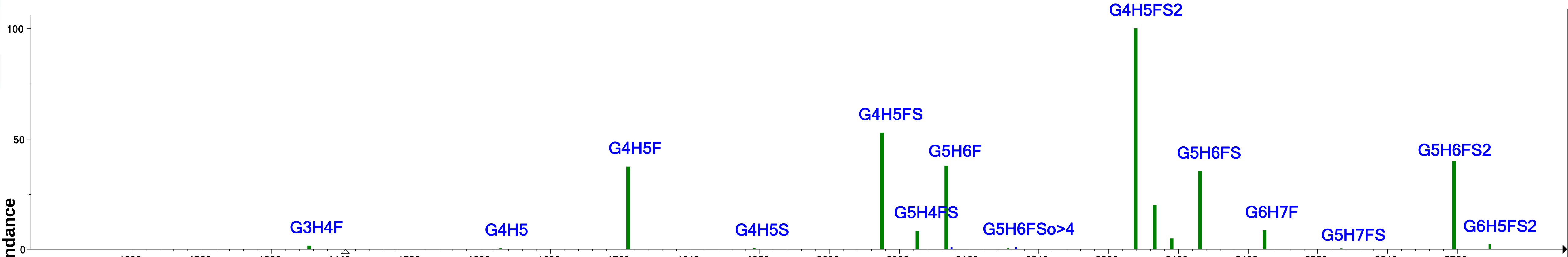
Interlaboratory Studies

Objectives of the study

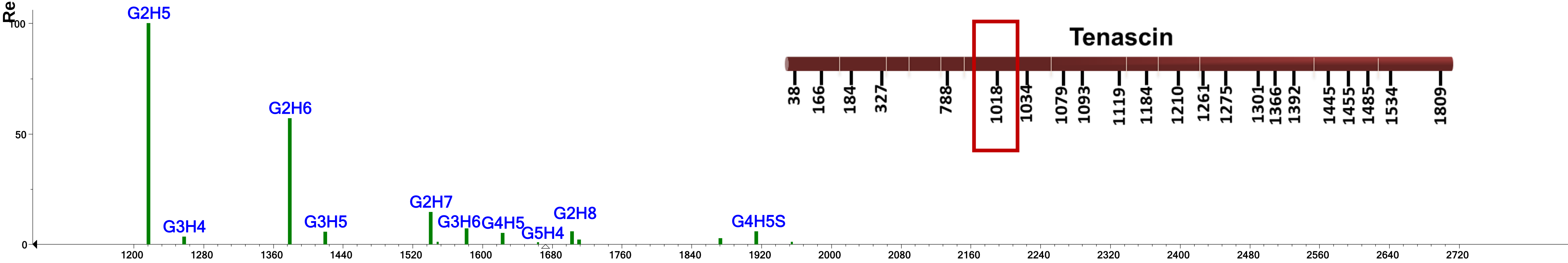
☐ To build a searchable MS library format of MS1 data to compare relative abundances

☐ To develop a Tandem MS Library of annotated human milk N-glycopeptides

Site-specific Glycosylation Analysis- GADS of Tenascin

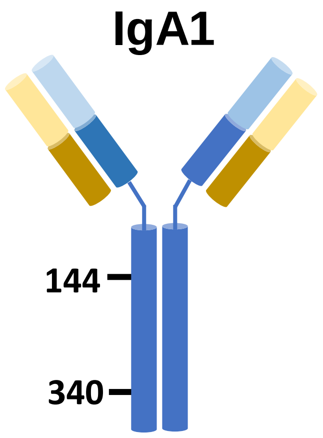


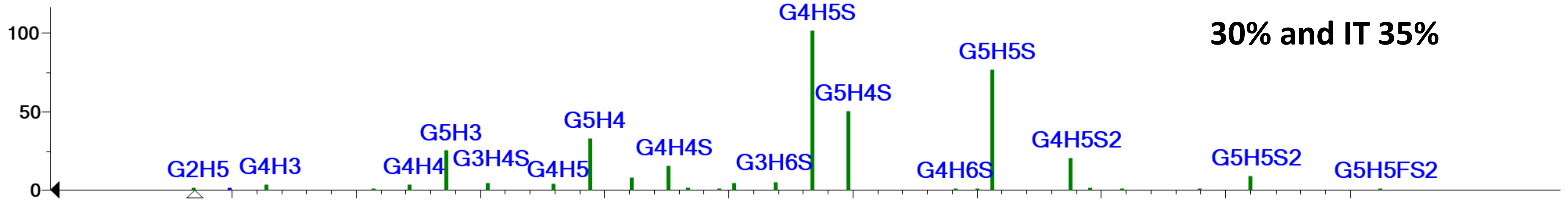
(tenascin) EVGWDGLRLnWT/+2+3+4 Class:HiMan=0.000,Hybrid=0.000,Complex=0.964,Fucose=0.997,Sialyl=0.790 Sequon:1093 File:TNC_06_urea_L1_2021-08-04_380-2000_



(tenascin) KFDRYRLnYSLPT/+3+4 Class:HiMan=0.703,Hybrid=0.070,Complex=0.043,Fucose=0.000,Sialyl=0.051 Sequon:1018 File:Tenascin_3_urea_NLumos_2020-09-16_alphalyt_120_NM_H

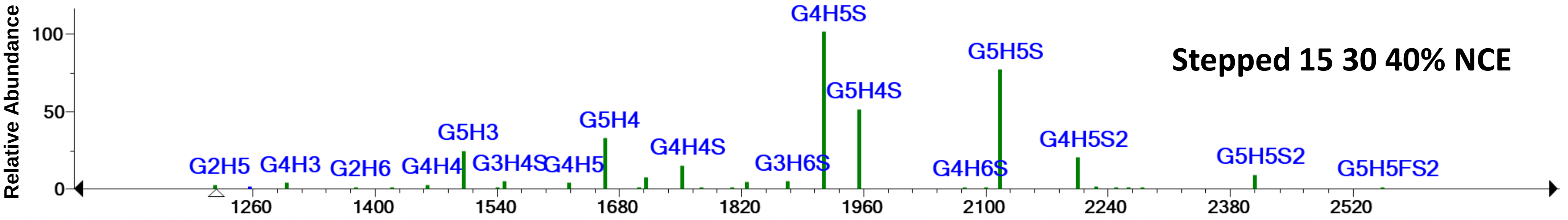
Measurement Reproducibility and Variability- GADS of IgA1





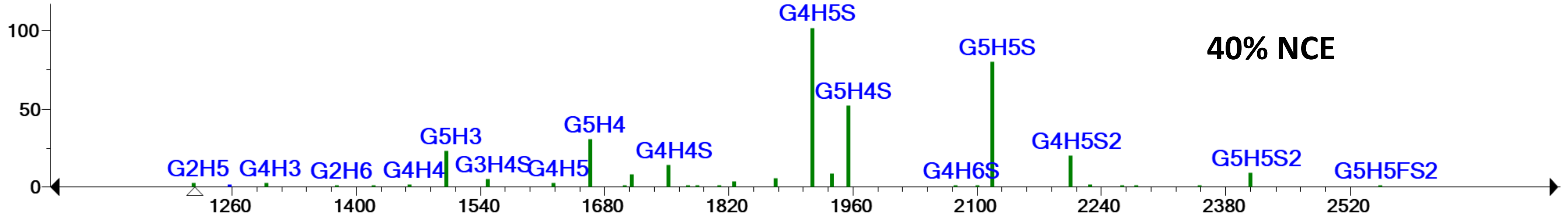
30% and IT 35%

(nist-iga1) AnLTCTLTGLR/+2+3+4 Class:HiMan=0.004,Hybrid=0.034,Complex=0.592,Fucose=0.009,Sialyl=0.797 Sequon:144 File:IGHA1_35_NLumos_2020-12-21_380-2000_120_NM_HCD30CIT



Stepped 15 30 40% NCE

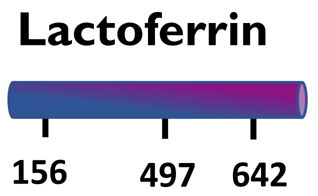
(nist-iga1) AnLTCTLTGLR/+2+3+4 Class:HiMan=0.008,Hybrid=0.032,Complex=0.588,Fucose=0.008,Sialyl=0.793 Sequon:144 File:IGHA1_34_NLumos_2020-12-21_380-2000_120_NM_SH153040

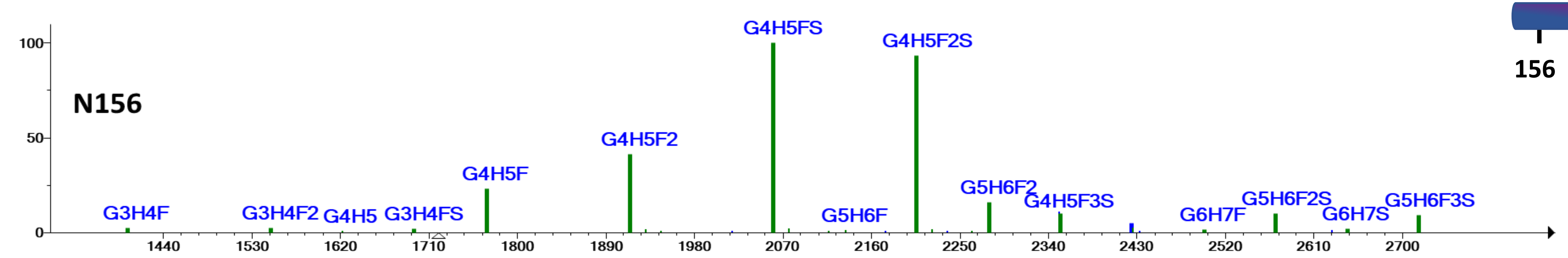


40% NCE

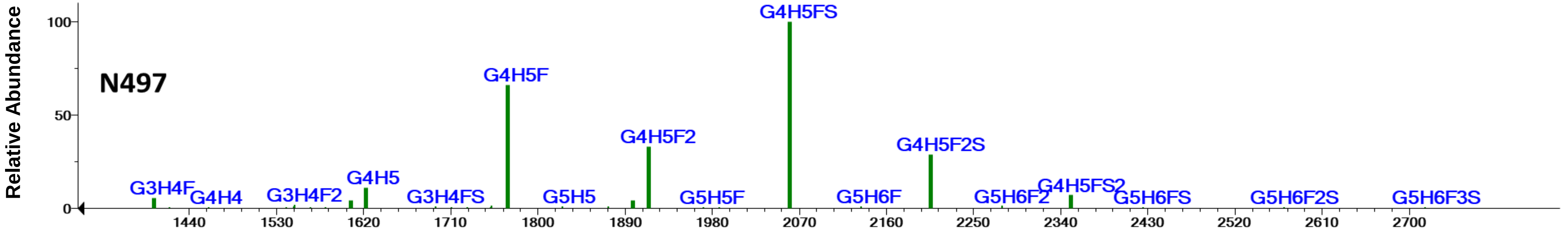
(nist-iga1) AnLTCTLTGLR/+2+3+4 Class:HiMan=0.008,Hybrid=0.034,Complex=0.610,Fucose=0.010,Sialyl=0.818 Sequon:144 File:IGHA1_05_urea_NLumos_2020-11-20_380-2000_120_NM_HCD40

GADS of Lactoferrin (Human Milk)

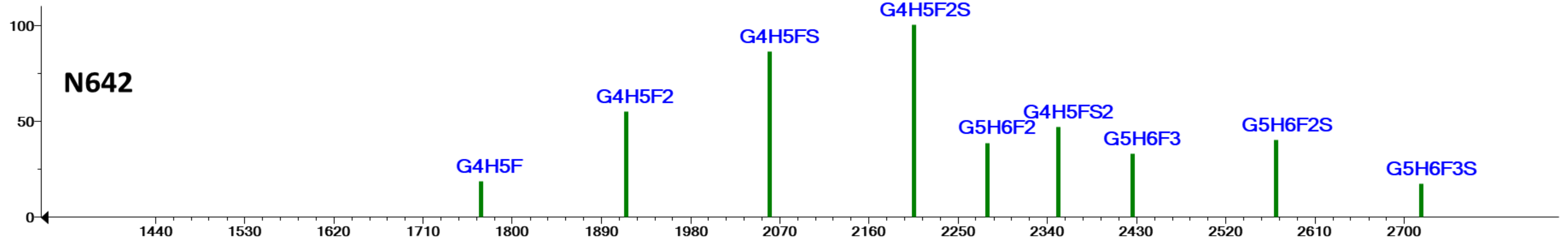




(srm1953-nist) LnWTGPPEPIEAAVAR/+2+3+4 Class:HiMan=0.000,Hybrid=0.002,Complex=0.983,Fucose=0.900,Sialyl=0.692 Sequon:156 File:SRM1953_03_urea_NL_2021-04-06_HILIC-H_SH153040

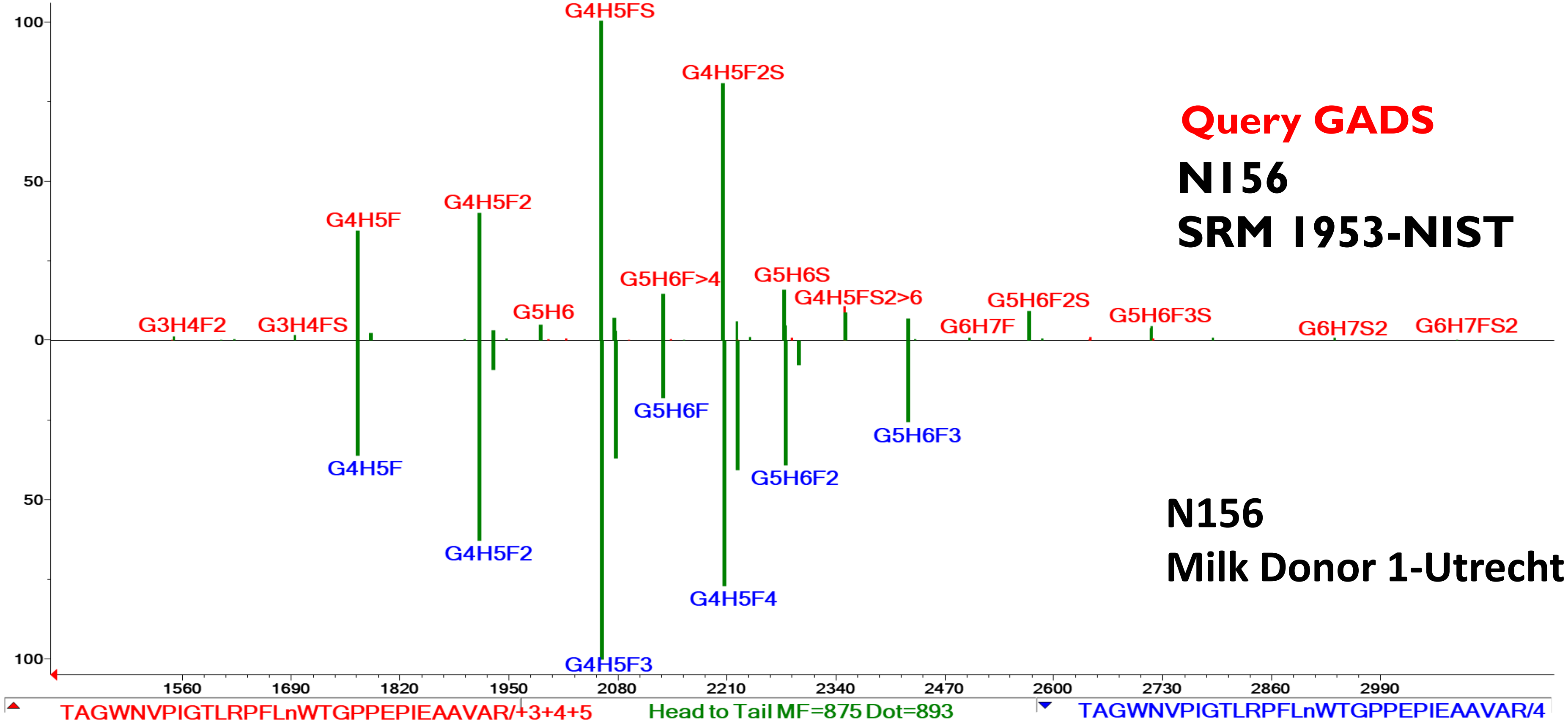


(srm1953-nist) LFnQTGSCK/+2+3+4 Class:HiMan=0.004,Hybrid=0.007,Complex=0.913,Fucose=0.872,Sialyl=0.544 Sequon:497 File:SRM1953_03_urea_NL_2021-04-06_HILIC-H_380-2000_SH153040



(srm1953-nist) GRnGSDCPDKF/3-Consensus(2) Class:HiMan=0.000,Hybrid=0.000,Complex=0.999,Fucose=1.000,Sialyl=0.617 Sequon:642 File:SRM1953-HILIC_02_urea_NL_2021-05-17_HS154040CIT

Human milk (Interlab)-Application of GADS



TAGWNVPIGTLRPFLnWTGPPEPIEAAVAR/+3+4+5 Head to Tail MF=875 Dot=893 TAGWNVPIGTLRPFLnWTGPPEPIEAAVAR/4

Query GADS

N156

SRM 1953-NIST

N156

Milk Donor 1-Utrecht U.

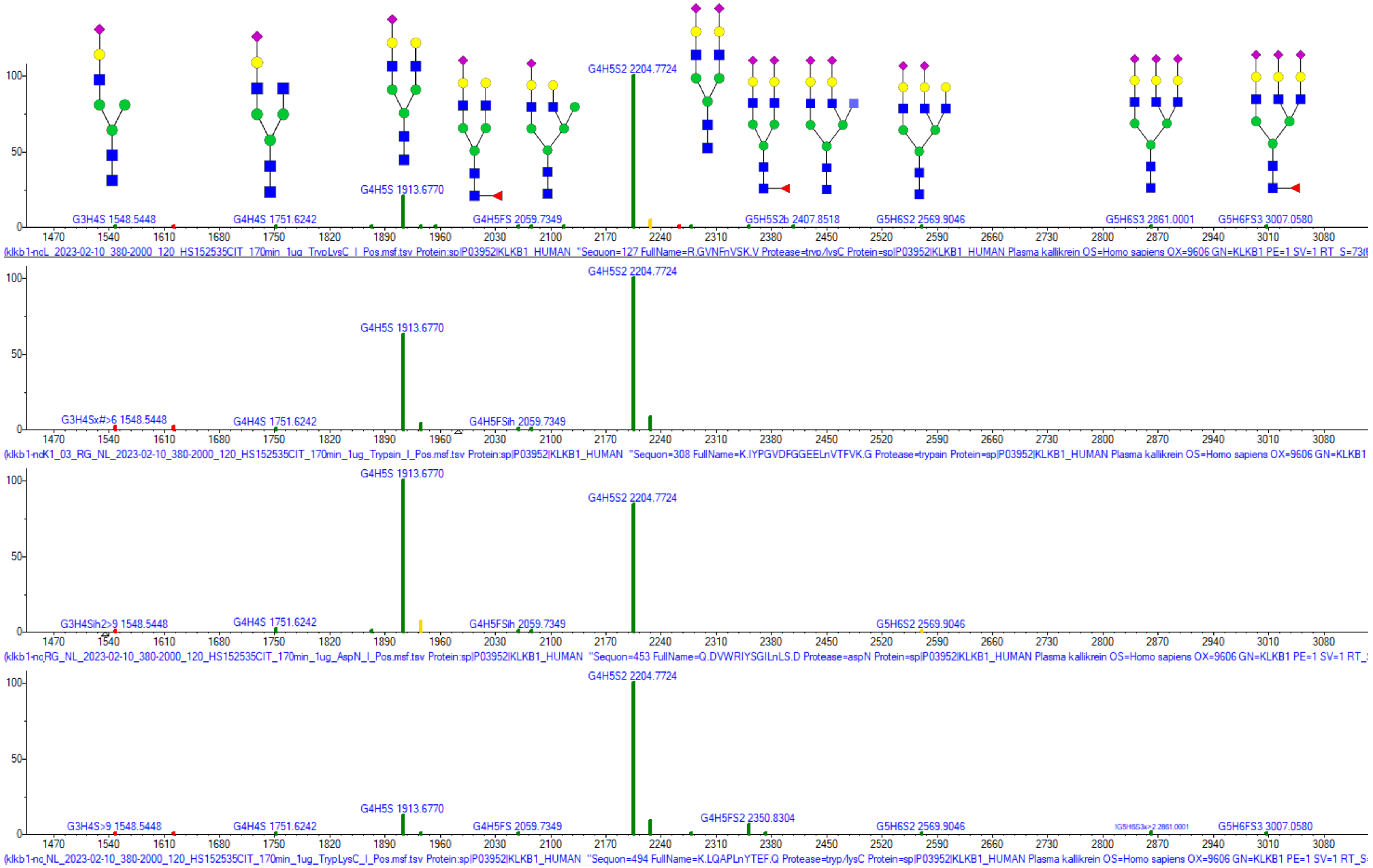
#	Library	Score	Dot-Prod	Rev-Dot	PSS-Dot	Name	Sample
1	**hlf-donor1-utrech**	**875**	**893**	**969**	**893**	**TAGWNVPIGTLRPFLnWTGPPEPIEAAVAR/4**	**Week 10**
2	hlf-donor1-utrecht	870	891	966	891	TAGWNVPIGTLRPFLnWTGPPEPIEAAVAR/+4+5	Week 10
3	hlf-donor1-utrecht	856	876	941	875	TAGWNVPIGTLRPFLnWTGPPEPIEAAVAR/+4	Week 8
4	hlf-donor1-utrecht	849	874	940	874	TAGWNVPIGTLRPFLnWTGPPEPIEAAVAR/4-Consensus (27)	Week 10
5	hlf-donor1-utrecht	831	859	959	863	TAGWNVPIGTLRPFLnWTGPPEPIEAAVAR/+4+5	Week 12
6	hlf-donor1-utrecht	824	863	952	849	TAGWNVPIGTLRPFLnWTGPPEPIEAAVAR/+4+5	Week 8

Zhu et al, J. Nutrition 2021

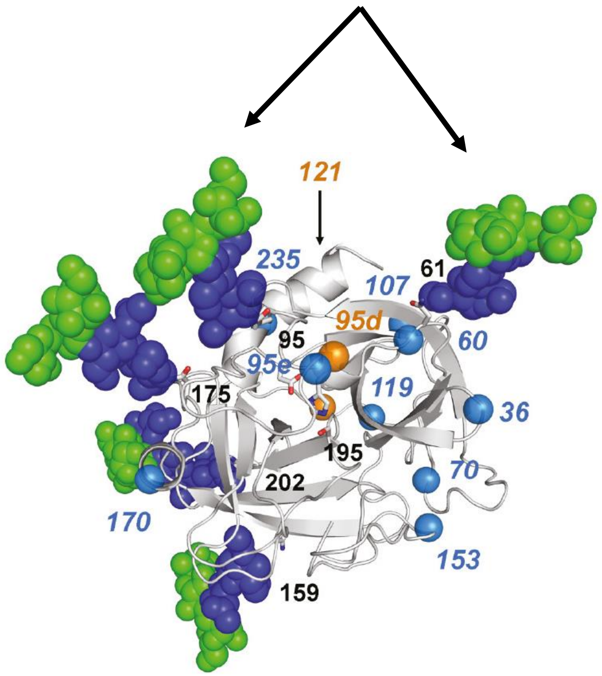
Utrecht University (PXD014917)

Lessons learned from Kallikrein

(Yi Liu)



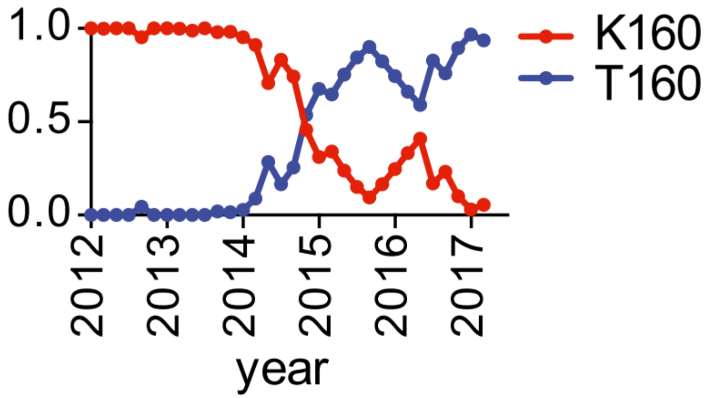
Metaheterogeneity



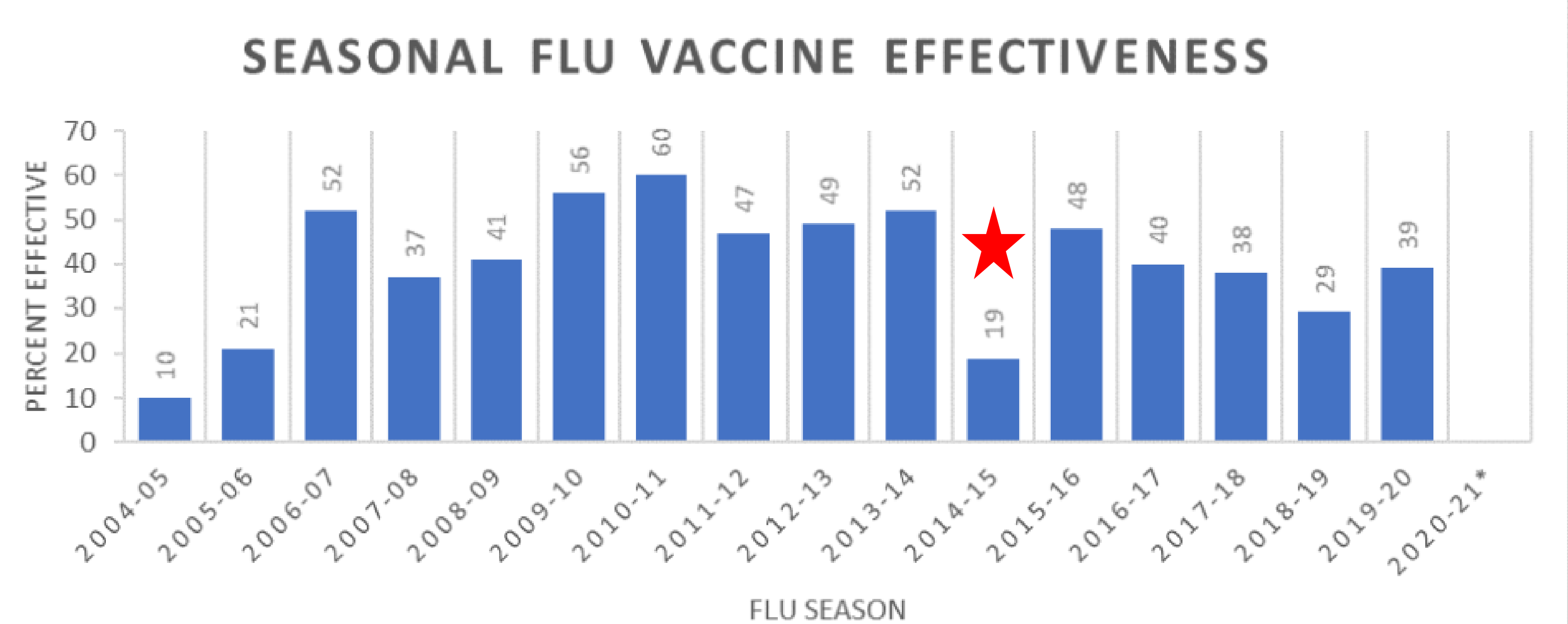
Why Glycosylation Matters For Vaccines

…THLNFKYPAL…
…THLNFTYPAL…

Sequon motif
NXT/S X≠P

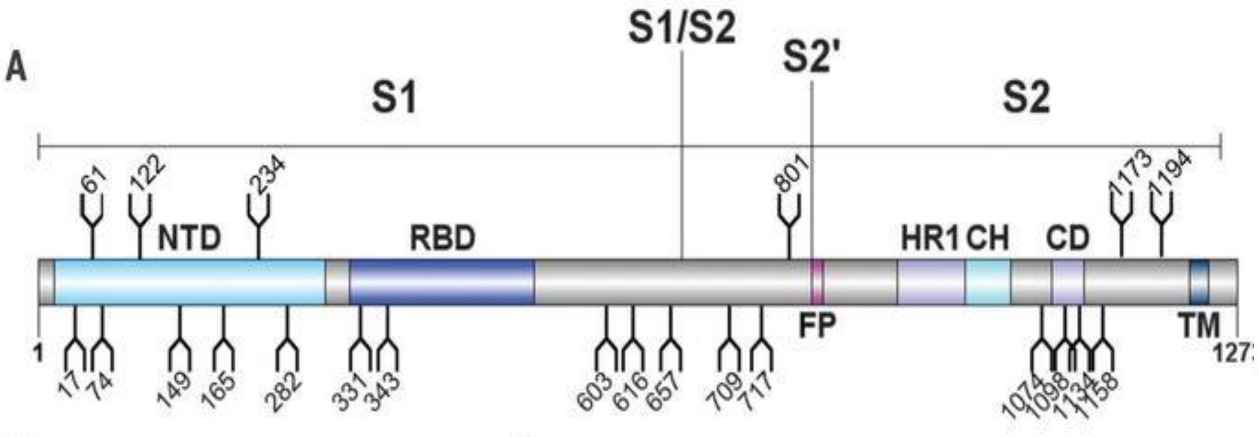


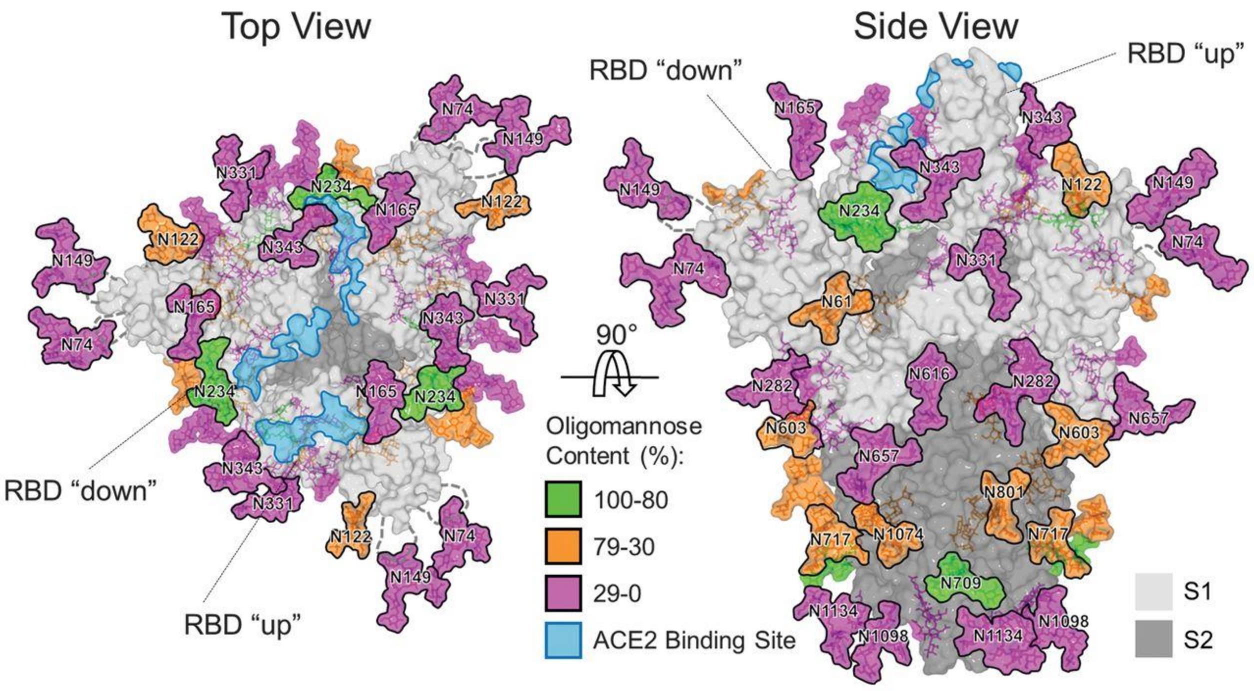
Zost et al., 2017



www.cdc.gov

SARS CoV-2 Spike Protein



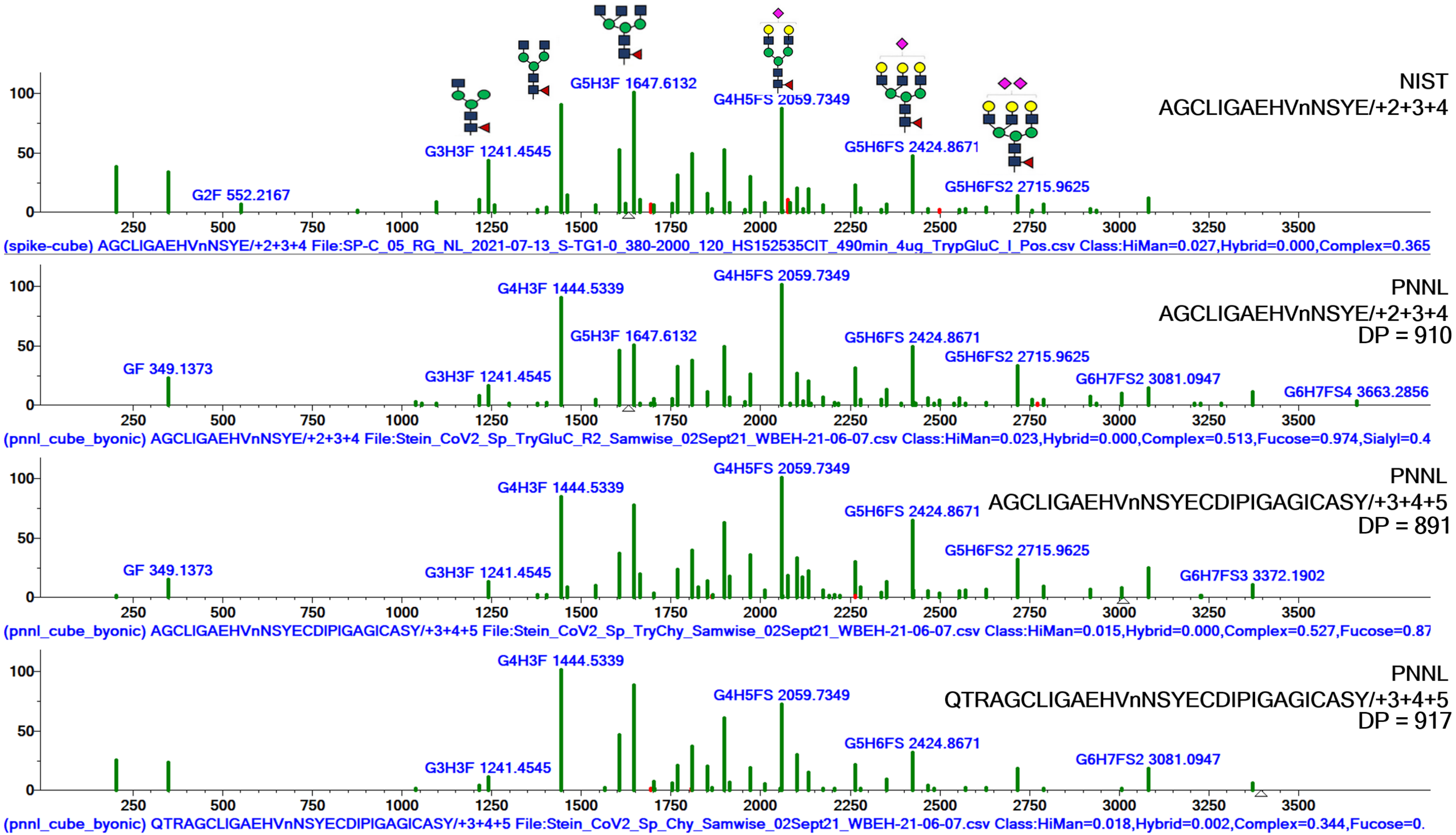


Site-specific glycan analysis of the SARS-CoV-2 spike, Volume: 369, Issue: 6501, Pages: 330-333, DOI: (10.1126/science.abb9983)

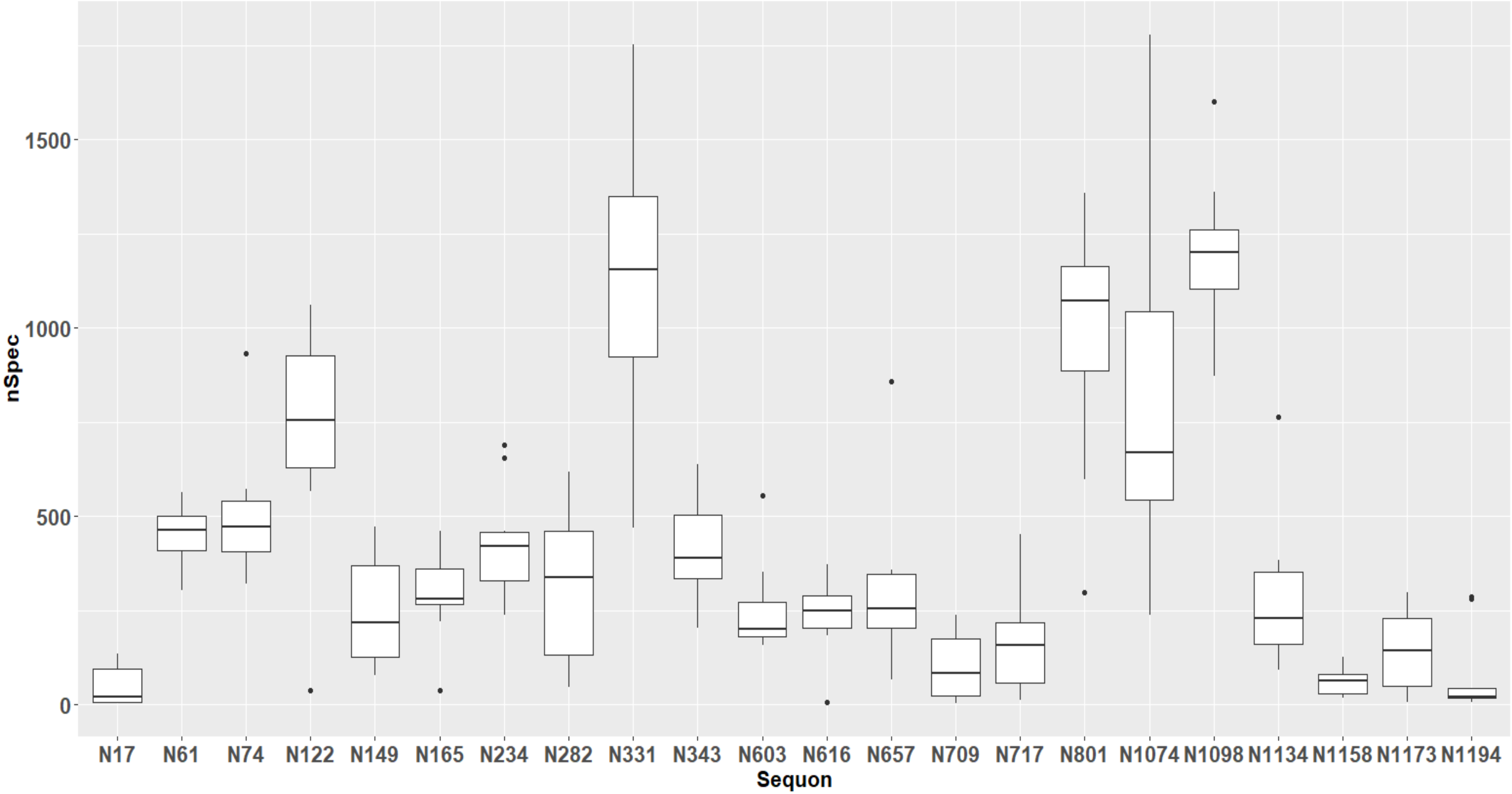
11 Sources of Recombinant Spike

Source	Sequence	Cells	Furin	Pro Substitution	Mutations	Tag
A	16-1213	HEK293	RAAA			T4, 10xHis
B	16-1213	HEK293	RAAA	F817, A892, A899, A942, K986, K87		T4, 10xHis
C	16-1213	HEK293	RAAA			His
D	1-1273	HEK293 Expi	GSAG	K986, V987		Rho 1D4
E	1-1273	HEK293 Expi	GSAG	K986, V987	del 69-70 & 144, N501Y, A570 D, D614G, P681H, T716I	Rho 1D4
F	15-1208	HEK293	GSAS	K986, K987		6xHis
G	1-1208	CHOExpress	GSAS	F817, A892, A899, A942, K986, V987	del 69-70 & 144-145, N501Y, A570D, D614G, P681H	8xHis
H	1-1208	CHOExpress	GSAS	K986, K987		8xHis
I	15-1208	HEK293	GSAS	K986, K987		His
J	16-1188	HEK	RAAA			T4, His+Avi
K	16-1213	HEK293	RAAA			T4, 10xHis

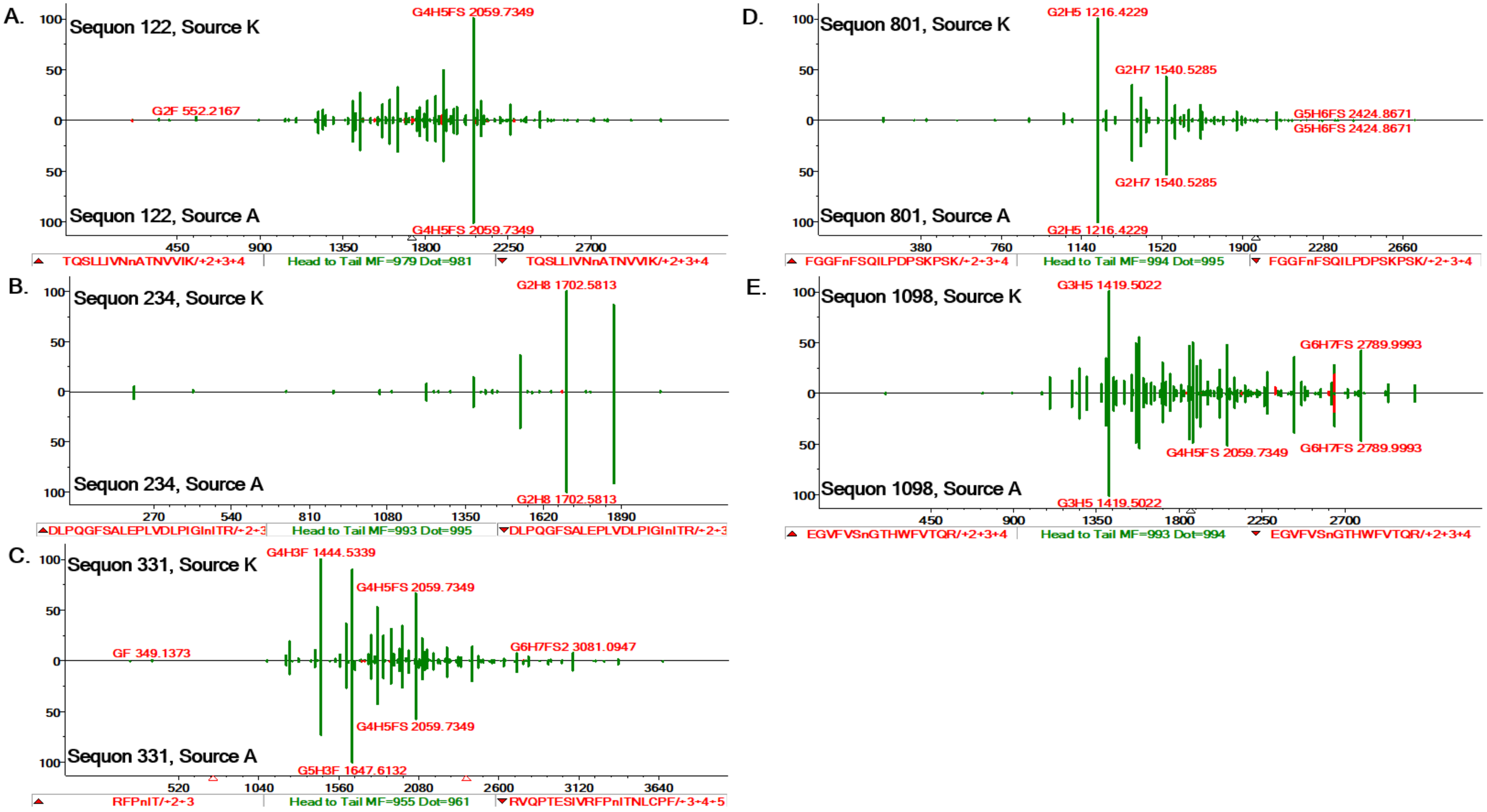
GADS Are Reproducible

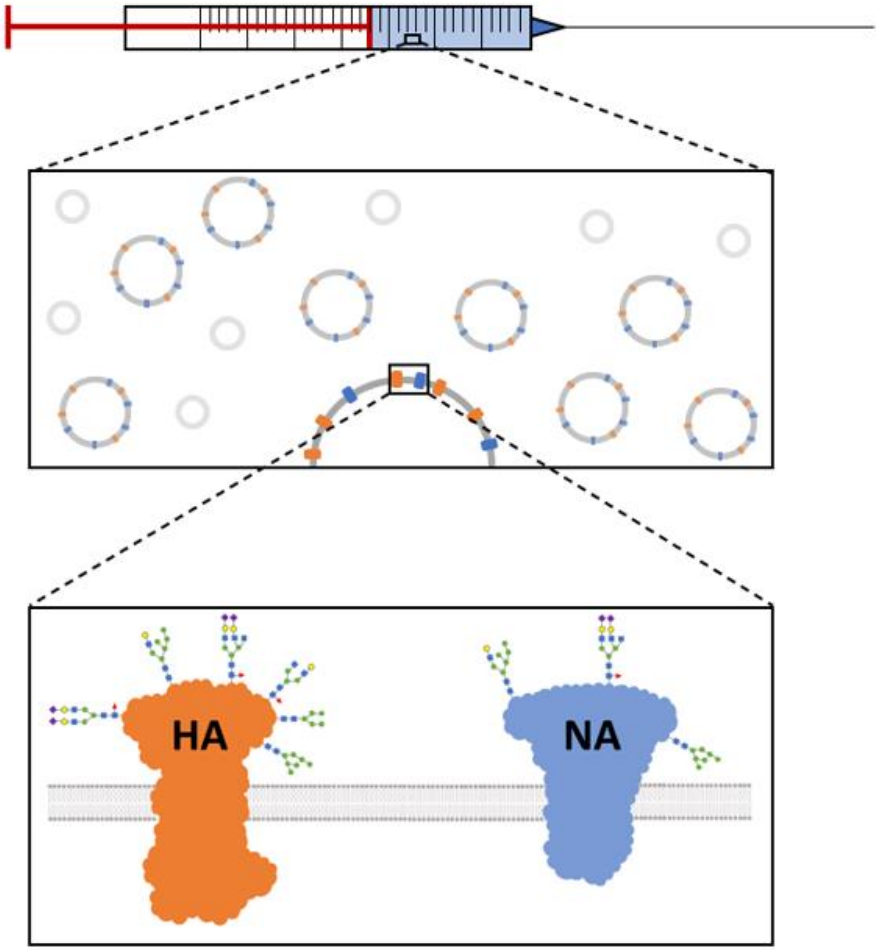


Glycosylation Sites Vary in Abundance

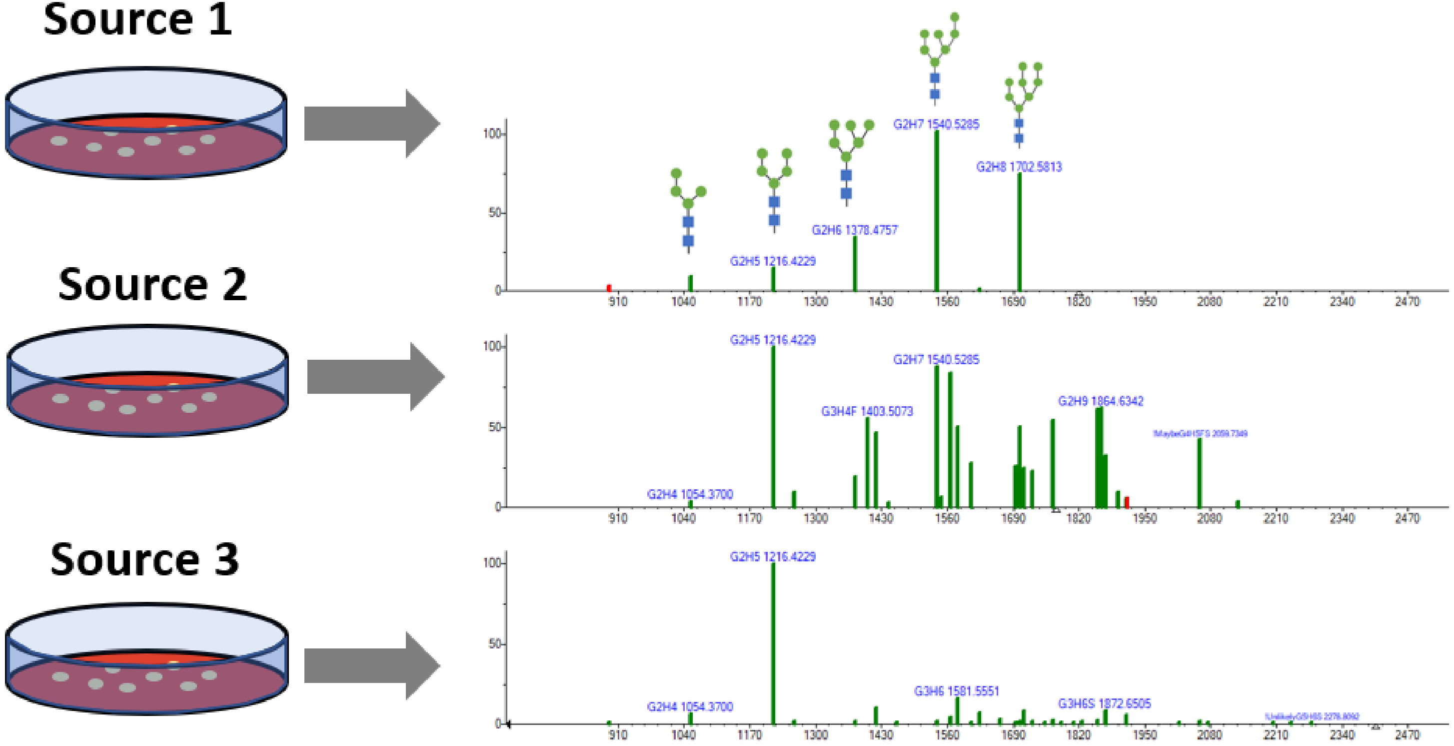


Similar Preparations Produce Similar Glycan Profiles





Glycopeptide Abundance Distribution Spectrum



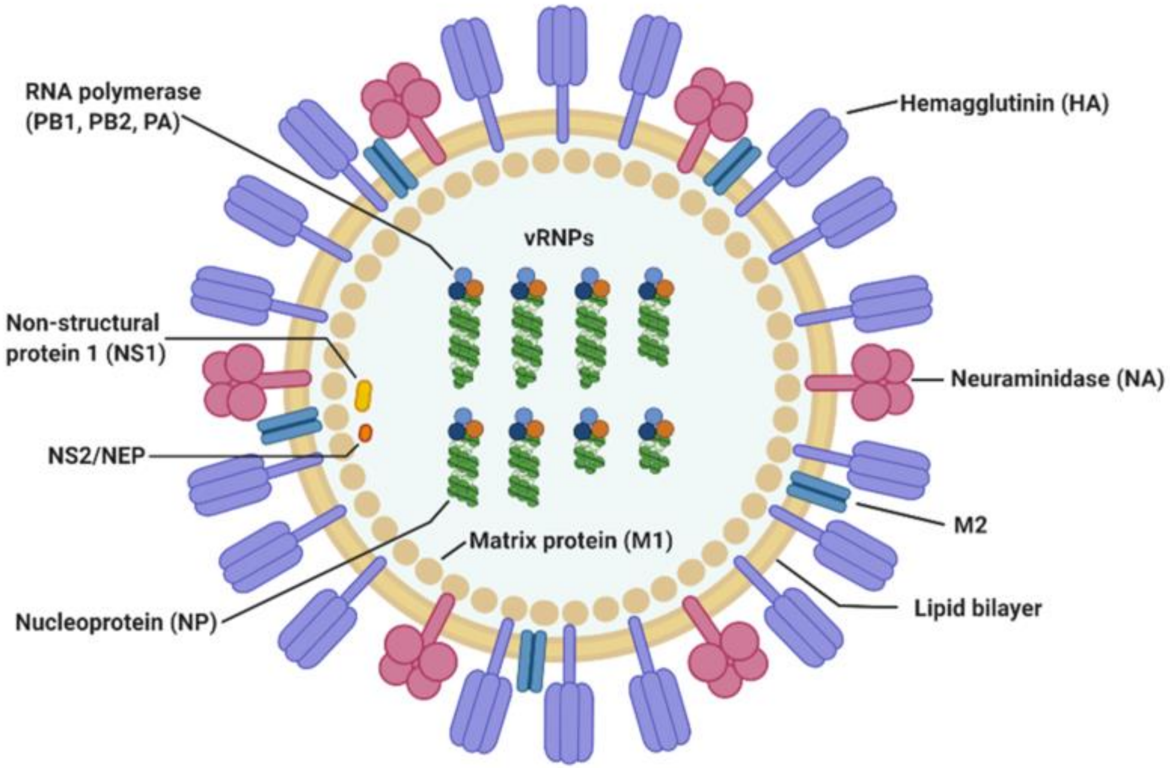
Source 1

Source 2

Source 3

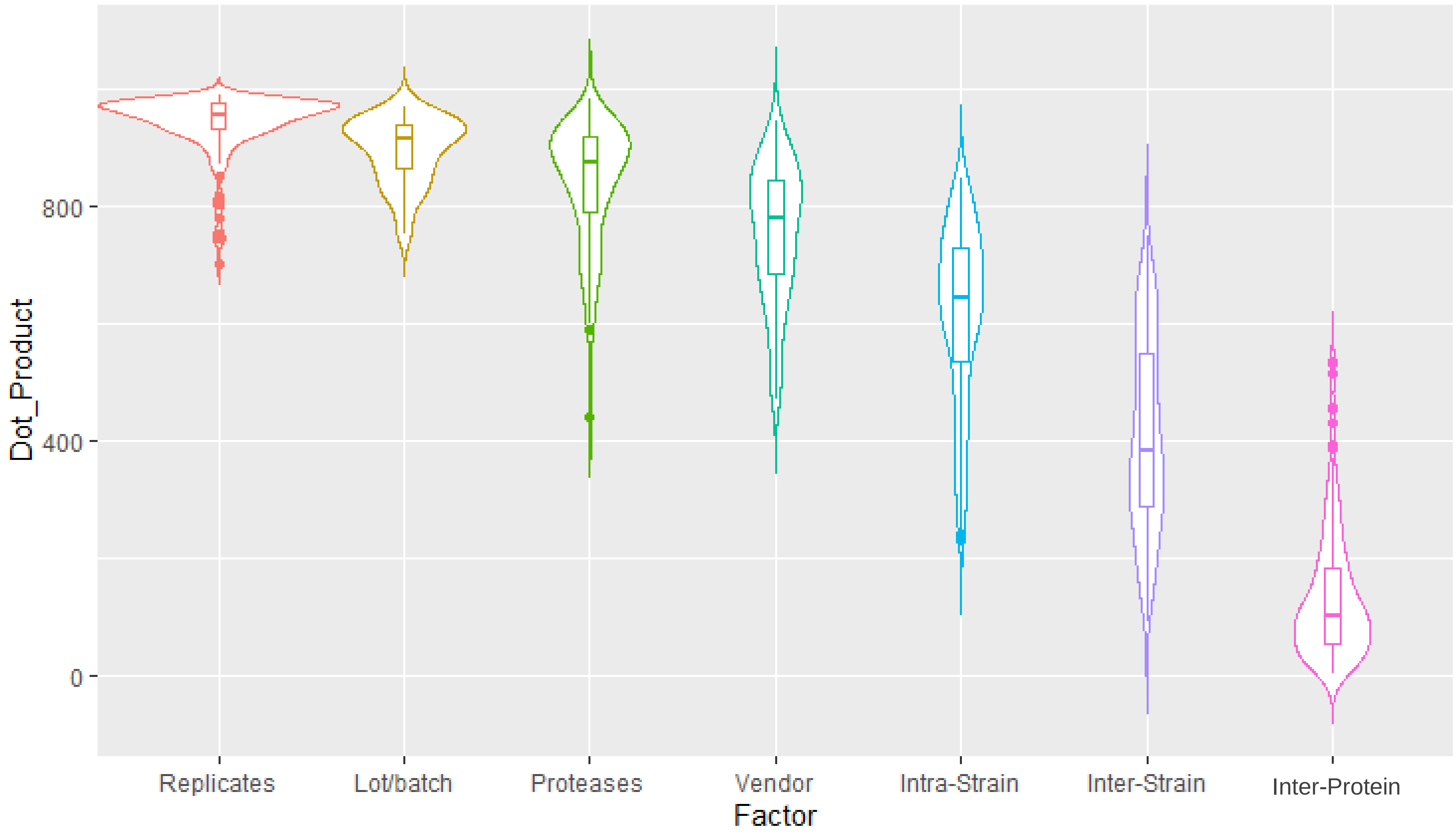
Influenza Virus

- 10 proteins
- Hemagglutinin (HA) and neuraminidase (NA) transmembrane proteins.
- Diversity of strains arise through point mutations or reassortment.

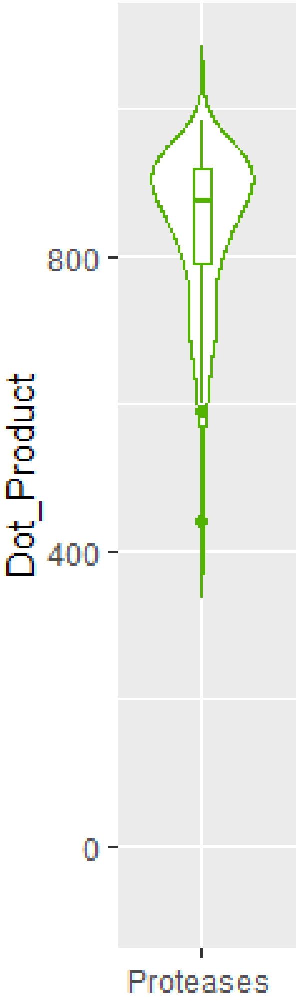


Jung et al., 2020

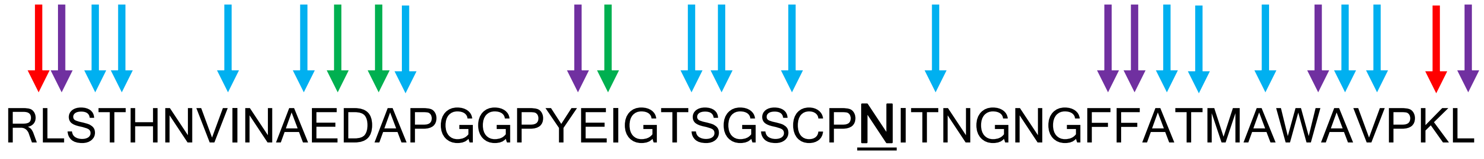
Variation in glycosylation profile



Variation From Different Protease Digestions

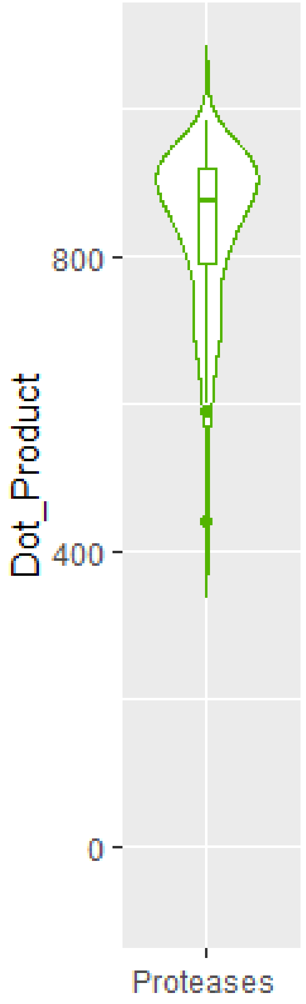


Protease Combination	Amino Acid Cleavage Sites
Trypsin + Lys-C	KR
Trypsin + Glu-C	KRED
Trypsin + Chymotrypsin	KRFWYL
Chymotrypsin + Glu-C	FWYLED
Chymotrypsin	FWYL
Alpha-lytic	TASV

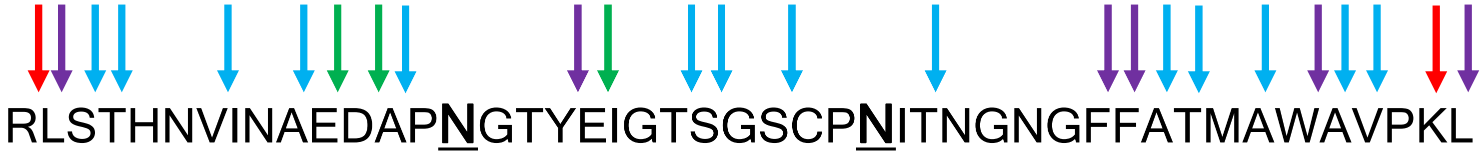


RLSTHNVINAEDAPGGPYEIGTSGSCPNITNGNGFFATMAWAVPKL

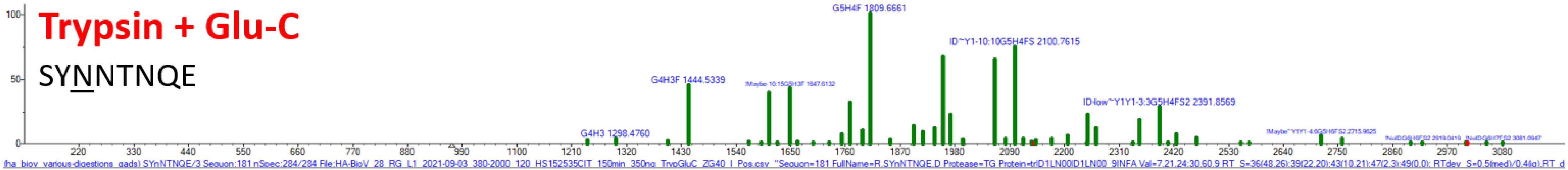
Variation From Different Protease Digestions



Protease Combination	Amino Acid Cleavage Sites
Trypsin + Lys-C	KR
Trypsin + Glu-C	KRED
Trypsin + Chymotrypsin	KRFWYL
Chymotrypsin + Glu-C	FWYLED
Chymotrypsin	FWYL
Alpha-lytic	TASV

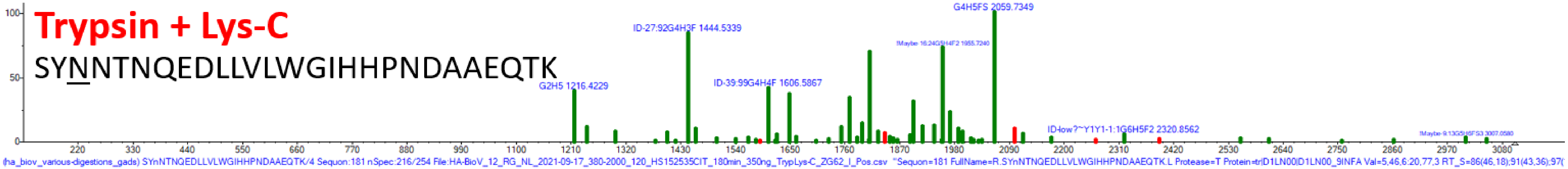


Variation From Different Protease Digestions



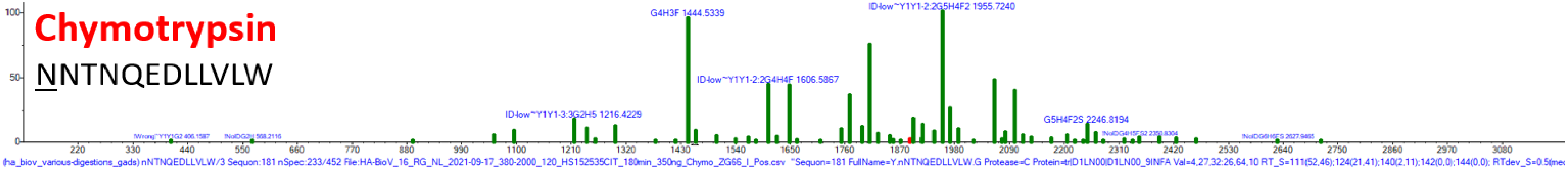
Trypsin + Glu-C

SYNNTNQE



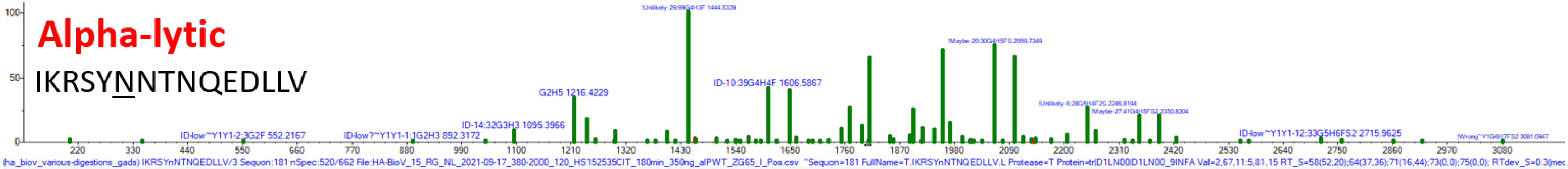
Trypsin + Lys-C

SYNNTNQEDLLVLWGIHHPNDAAEQTK



Chymotrypsin

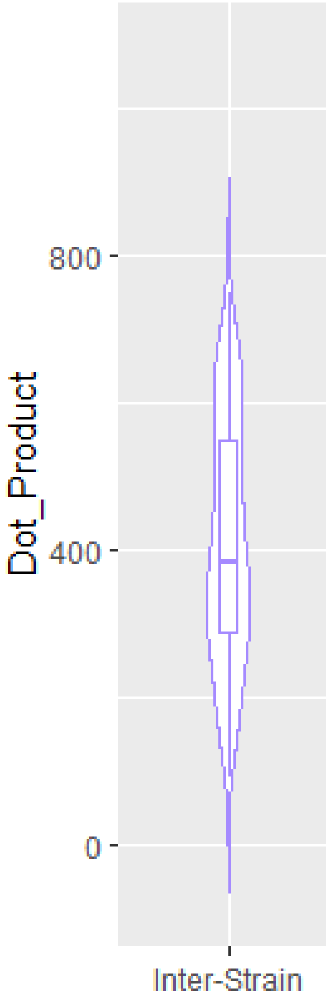
NNTNQEDLLVLW



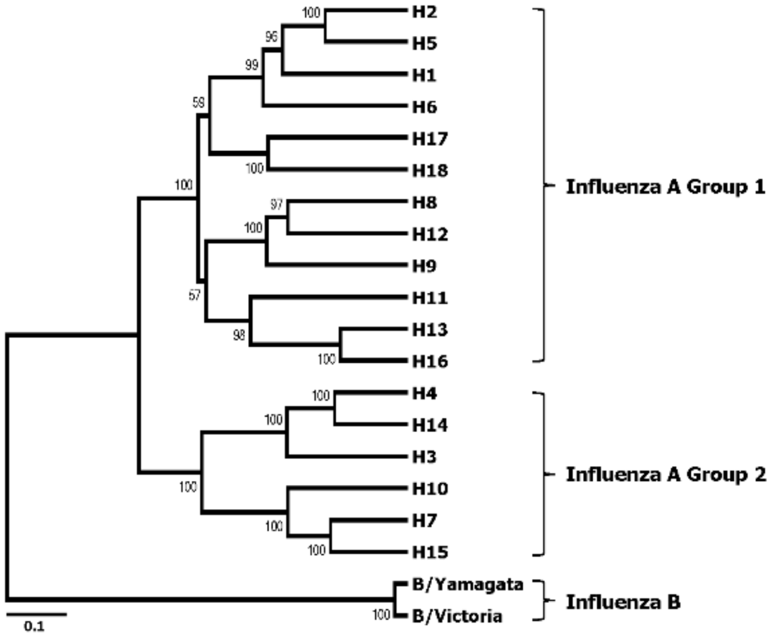
Alpha-lytic

IKRSYNNTNQEDLLV

Variation From Different Strains



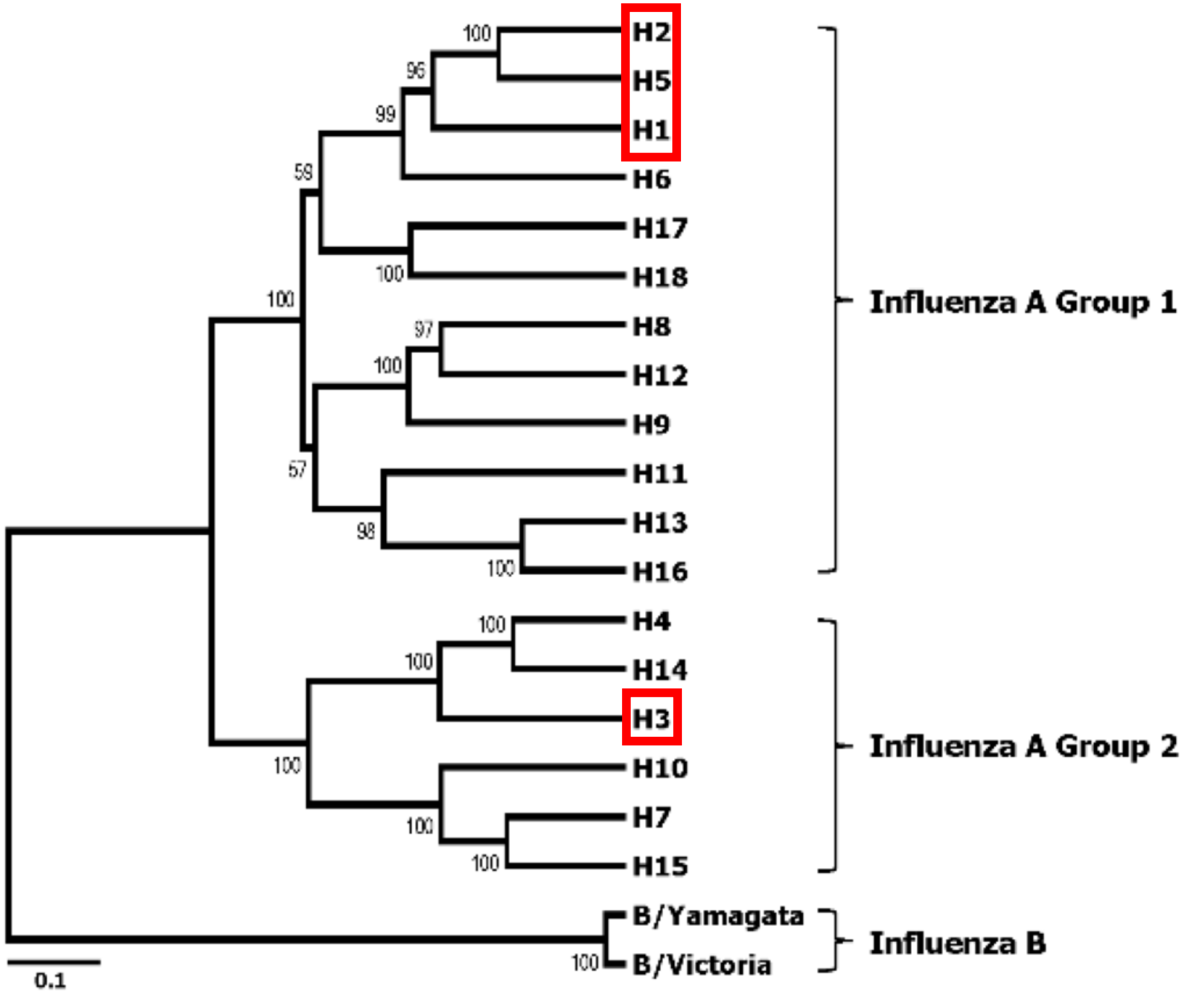
Protein	Strain	Subtype
HA	**A/California/04/2009**	**H1N1**
HA	**A/New Caledonia/20/1999**	**H1N1**
HA	**A/Japan/305/1957**	**H2N2**
HA	**A/Hong Kong/485197/2014**	**H3N2**
HA	**A/Hong Kong/483/1997**	**H5N1**
NA	**A/Arizona/13/2008**	**H1N1**
NA	**A/Thailand/1(KAN-1)/2004**	**H5N1**
NA	**A/Netherlands/219/2003**	**H7N7**



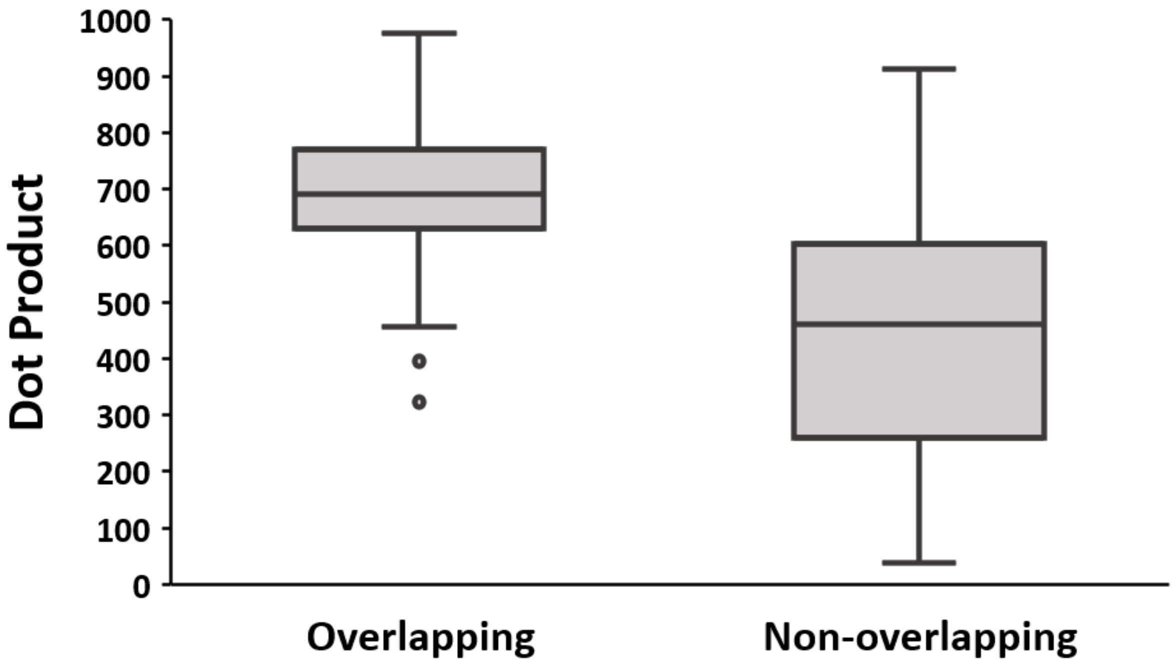
```
MK-TIIALSYILCLVFAQKIPGNDNSTATLCLGHHAVPNGTIVKTITNDRIEVTNATELV
MKAKLLVLLCTFT----------ATYADTICIGYHANNSTDTVDTVLEKNVTVTHSVNLL
MKAILVVLLYTFA----------TANADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLL
-MEKIVLLLATVS----------LVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDIL
--MAIIYLILLFT----------AVRGDQICIGYHANNSTEKVDTILERNVTVTHAKDIL
    :: :   .                 :* *  :  .   *.*  : .: **    :

QNSSIGEICDSPH------QILDGENCTLIDALLGDPQCDGFQN-KKWDLFVERS-KAYS
EDSHNGKLCLLKGI-----APLQLGNCSVAGWILGNPECELLISKESWSYIVETPNPENG
EDKHNGKLCKLRGV-----APLHLGKCNIAGWILGNPECESLSTASSWSYIVETPSSDNG
ERTHNGKLCDLNGV-----KPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVEKASPAND
EKTHNGKLCKLNGI-----PPLELGDCSIAGWLLGNPECDRLLSVPEWSYIMEKENPRDG
     . :.                .*.  .  *. * *         . : *      .

NCYPYDVPDYASLRSLVATSGTLE---FNNE----SFNWTGVTQ-NGTSSACIRR-SSSS
TCYPGYFADYEELREQLSSVSSFERFEIFPK----ESSWPNHTV-TGVSASCSHN-GKSS
TCYPGDFIDYEELREQLSSVSSFERFEIFPK----TSSWPNHDSNKGVTAACPHA-GAKS
LCYPGNFNDYEELKHLLSRINHFEKIQIIPK-----SSWSNHDASSGVSSACPYL-GKSS
LCYPGSFNDYEELKHLLSSVKHFEKVKILPK-----DRWTQHTTT-GGSRACAVS-GNPS
 *:*    :  .::         *   :  .               * : :*    .  .

FFSRLNWLTHLNYTY-----PALNVTMPNNEQFDKLYIWGVHHPGTDKDQIFLYAQSSGR
FYRNLLWLTGKNGLY-----PNLSKSYVNNKEKEVLVLWGVHHPPNIGNQRALYHTENAY
FYKNLIWLVKKGNSY-----PKLSKSYINDKGKEVLVLWGIHHPSTSADQQSIYQNADTY
FFRNVVWLIKKNSTY-----PTIKRSYNNTNQEDLLVLWGIHHPNDAAEQTKLYQNPTTY
FFRNMVWLTKKGSDY-----PVAKGSYNNTSGEQMLIIWGVHHPNDETEQRTLYQNVGTY
*:  : *                .         : : :**.*          ::

IT-VSTKRSQQAVIPNIGSRPRI-----RDIPSRISIYWTIVKPGDILLINSTGNLIAPR
VS-VVSSHYSRRFTPEIAKRPKV-----RDQEGRINYYWTLLEPGDTIIFEANGNLIAPW
VF-VGSSRYSKKFKPEIAIRPKV-----RDQEGRMNYYWTLVEPGDKITFEATGNLVVPR
IS-VGTSTLNQRLVPEIATRPKV-----NGQSGRIEFFWTILKPNDAINFESNGNFIAPE
VS-VGTSTLNKRSTPEIATRPKV-----NGQGGRMEFSWTLLDMWDTINFESTGNLIAPE
     :.        :*.  *           .*:   : : .  .   :    ..: *
```

```
GYFKI-RSGKSSIMRSDAPIGKCKSECITPNGSIPNDKPF-QNVNRITYGACPRYVKHST
YAFALSRGFGSGIITSNAPMDECDAKCQTPQGAINSSLPF-QNVHPVTIGECPKYVRSAK
YAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPF-QNIHPITIGKCPKYVKSTK
YAYKIVKKGDSTIMKSELEYGNCNTKCQTPMGAINSSMPF-HNIHPLTIGECPKYVKSNR
YGFKISKRGSSGIMKTEGTLENCETKCQTPLGAINTTLPF-HNVHPLTIGECPKYVKSEK
  :       * ::  .              *.: .  *:  . :  : * ** :*:

LKLATGMRNVPEKQ----TRGIFGAIAGFIENGWEGMVDGWYGFRHQNSEGRGQAADLKS
LRMVTGLRNIPSIQ----SRGLFGAIAGFIEGGWTGMVDGWYGYHHQNEQGSGYAADQKS
LRLATGLRNIPSIQ----SRGLFGAIAGFIEGGWTGMVDGWYGYHHQNEQGSGYAADLKS
LVLATGLRNAPQRERRRKKRGLFGAIAGFIEGGWQGMVDGWYGYHHSNEQGSGYAADQES
LVLATGLRNVPQIE----SRGLFGAIAGFIEGGWQGMVDGWYGYHHSNDQGSGYAADKES
* :..* :  *  :     **:*******:*.** **: **:*:   . .* . *** :*

TQAAIDQINGKLNRLIGKTNEKFHQIEKEFSEVEGRIQDLEKYVEDTKIDLWSYNAELLV
TQNAINGITNKVNSVIEKMNTQFTAVGKEFNKLERRMENLNKKVDDGFLDIWTYNAELLV
TQNAIDEITNKVNSVIEKMNTQFTAVGKEFNHLEKRIENLNKKVDDGFLDIWTYNAELLV
TQKAIDGVTNKVNSIIDKMNTQFEAVGREFNNLERRIENLNKKMEDGFLDVWTYNAELLV
TQKAFDGITNKVNSVIEKMNTQFEAVGKEFSNLERRLENLNKKMEDGFLDVWTYNAELLV
** *:: :. ::* :      ::  :   :..:. .: :*:: ::*   *  : : ** *

ALENQHTIDLTDSEMNKLFEKTKKQLRENAEDMGNGCFKIYHKCDNACIGSIRNGTYDHN
LLENERTLDFHDSNVKNLYEKVKSQLKNNAKEIGNGCFEFYHKCNNECMESVKNGTYDYP
LLENERTLDYHDSNVKNLYEKVRSQLKNNAKEIGNGCFEFYHKCDNTCMESVKNGTYDYP
LMENERTLDFHDSNVKNLYDKVRLQLRDNAKELGNGCFEFYHKCDNECMESVKNGTYDYP
LMENERTLDFHDSNVKNLYDKVRMQLRDNVKELGNGCFEFYHKCDDECMNSVKNGTYDYP
 :.*:  ::  *..:  *  * :  *  .. ::*****:  ***:: *:  :   *::

VYRDEALNNRFQIKGVELK--SGYKDWILWI-SFAISCFLLCVALLGFIMWACQKGNIRC
KYSEESKLNREKIDGVKLE--SMGVYQILAIYSTVASSLVLLVSLGAISFWMCSNGSLQC
KYSEEAKLNREEIDGVKLE--STRIYQILAIYSTVASSLVLVVSLGAISFWMCSNGSLQC
QYSEEARLNREEISGVKLE--SMGTYQILSLYSTVASSLALAIMVAGLSLWMCSNGSLQC
KYEEESKLNRNEIKGVKLS--SMGVYQILAIYATVAGSLSLAIMMAGISFWMCSNGSLQC
 :   :     :* ...*.  .     **   : . ..: : : :  : .:  ....: *
```

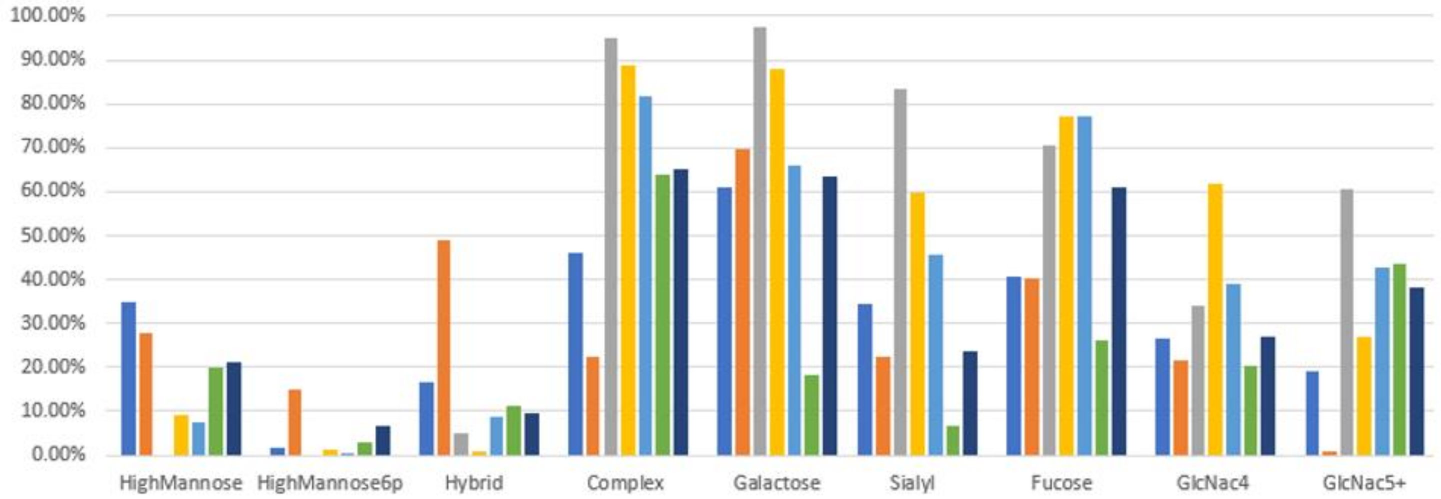
Conserved Regions Have Conserved Glycosylation Distribution



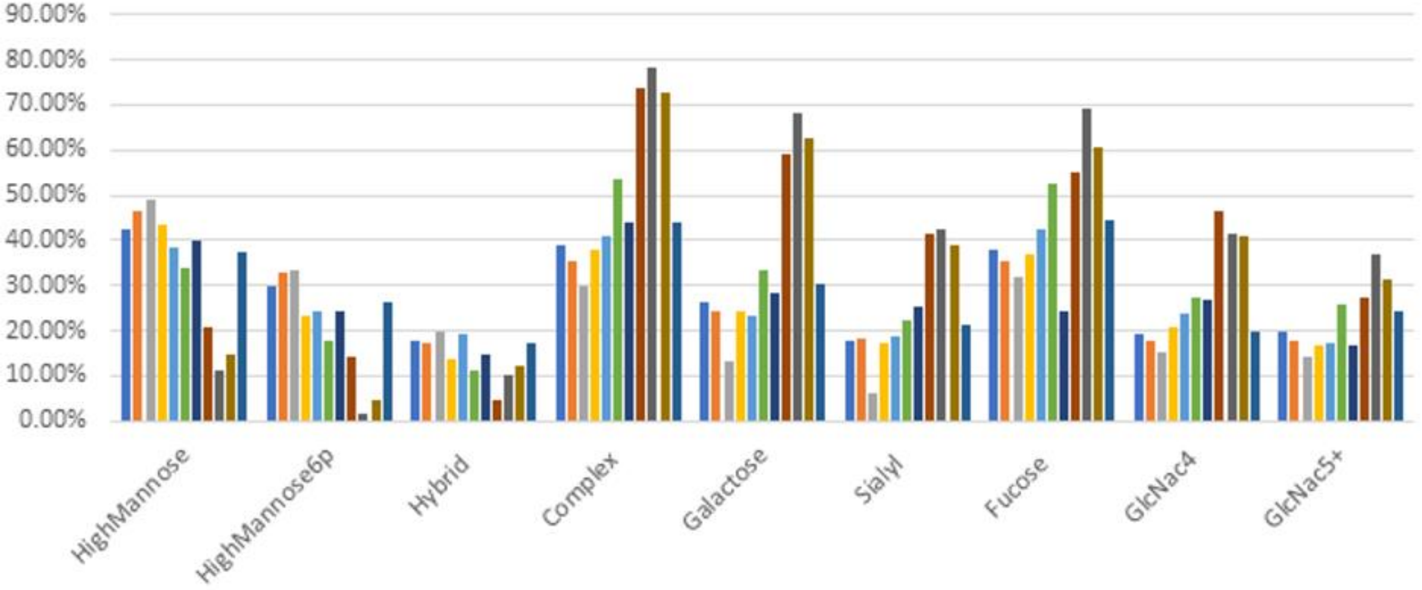
Jang et al., 2014



Inter-Protein Comparison



Influenza Hemagglutinin

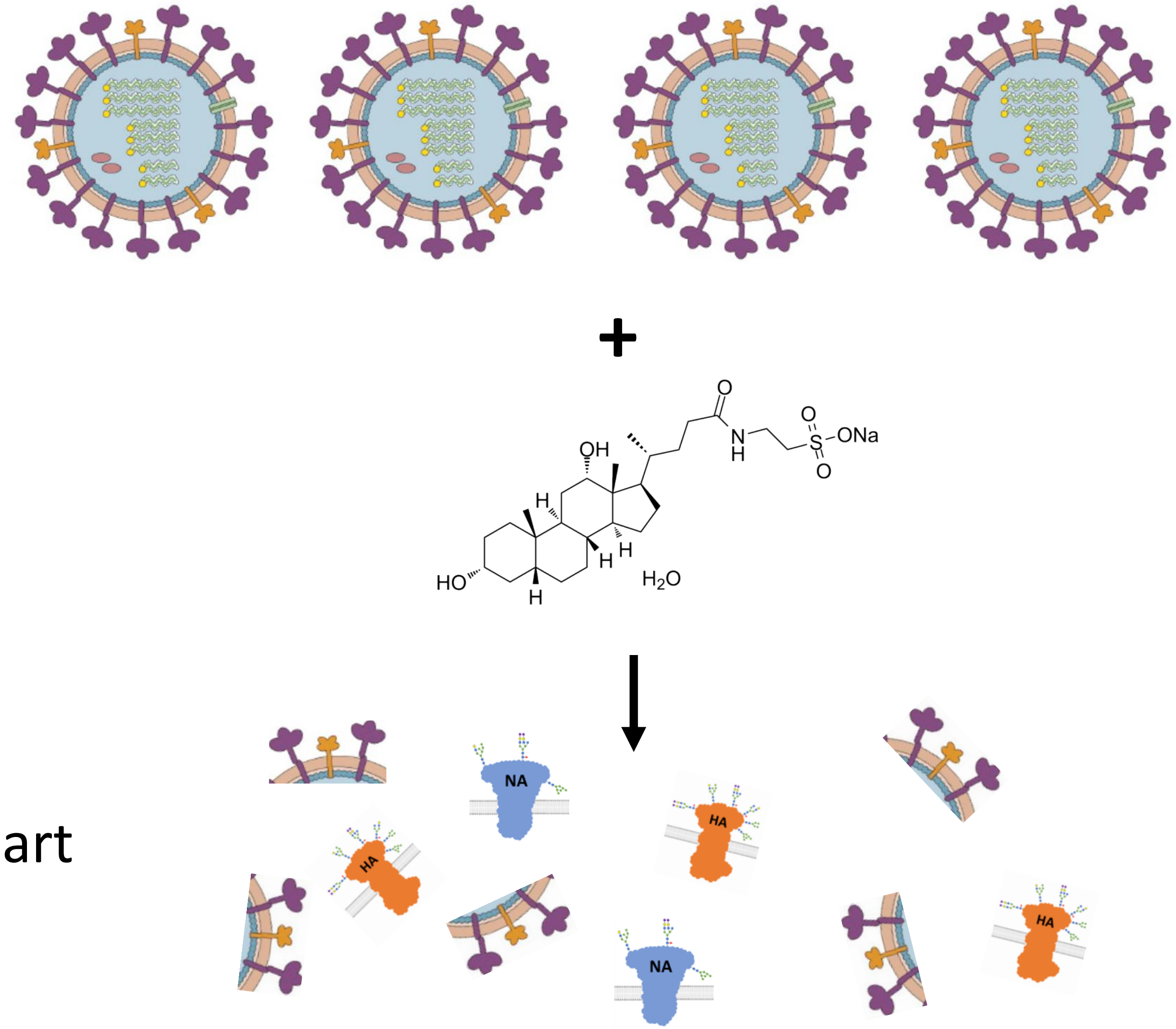


Sars-CoV-2 Spike protein

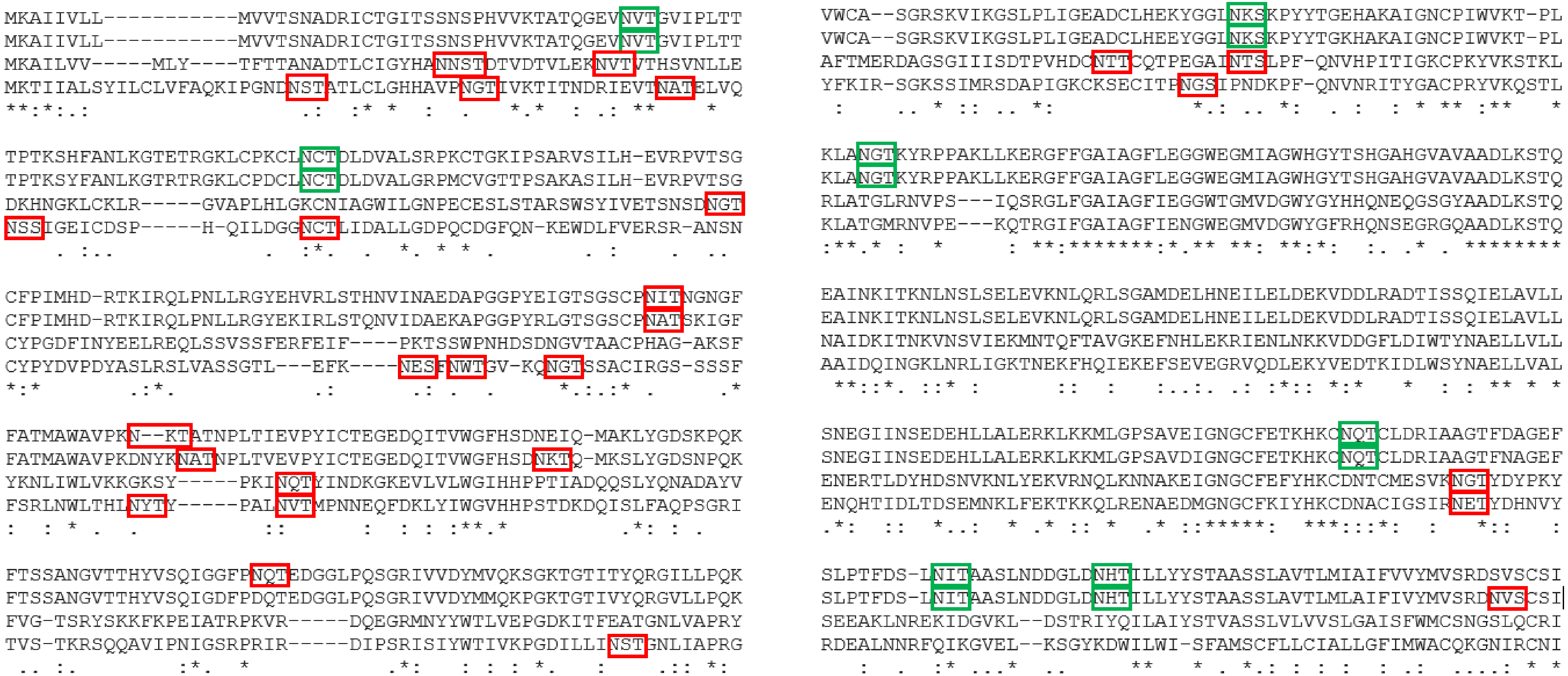
Influenza Vaccines

- Contents
    - Split virion
    - 60 mg HA per dose (15 mg per strain)
    - Thimerosal
    - Mercury
    - Sodium Chloride
    - Sodium Phosphate
    - Sodium Taurodeoxycholate
    - Ovalbumin
    - Sucrose………

- Afluria Quadrivalent
- 3 Monovalent vaccines from NIBSC
- 3 Monovalent vaccines from Creative Biomart



Quadrivalent Vaccine Challenges

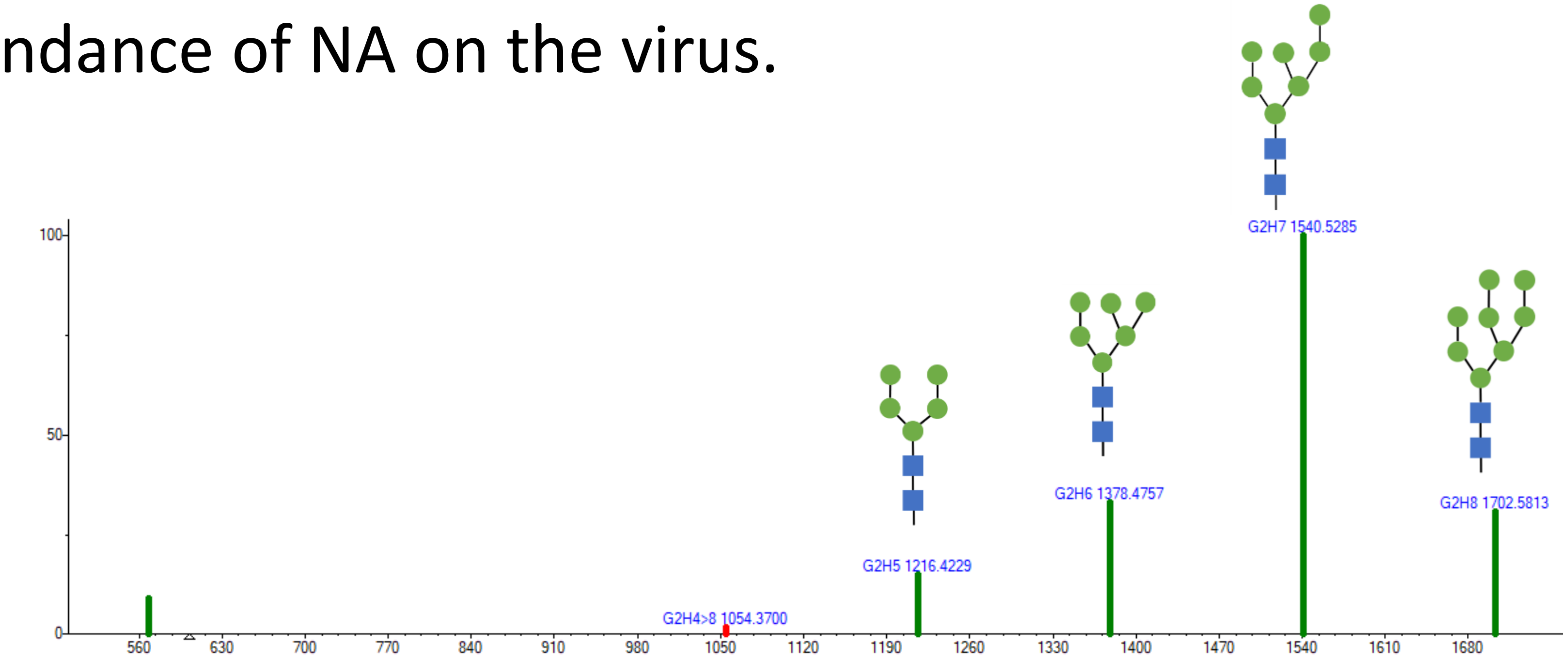


```
MKAIIVLL-----------MVVTSNADRICTGITSSNSPHVVKTATQGEVNVTGVIPLTT
MKAIIVLL-----------MVVTSNADRICTGITSSNSPHVVKTATQGEVNVTGVIPLTT
MKAILVV-----MLY----TFTTANADTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLLE
MKTIIALSYILCLVFAQKIPGNDNSTATLCLGHHAVPNGTIVKTITNDRIEVTNATELVQ
**:*:.:                 .:  :* *  :  .   *.*  : .: **    *

TPTKSHFANLKGTETRGKLCPKCLNCTDLDVALSRPKCTGKIPSARVSILH-EVRPVTSG
TPTKSYFANLKGTRTRGKLCPDCLNCTDLDVALGRPMCVGTTPSAKASILH-EVRPVTSG
DKHNGKLCKLR-----GVAPLHLGKCNIAGWILGNPECESLSTARSWSYIVETSNSDNGT
NSSIGEICDSP-----H-QILDGGNCTLIDALLGDPQCDGFQN-KEWDLFVERSR-ANSN
    . :..            .  :*.  .  *. * * .       . :    .  ..

CFPIMHD-RTKIRQLPNLLRGYEHVRLSTHNVINAEDAPGGPYEIGTSGSCPNITNGNGF
CFPIMHD-RTKIRQLPNLLRGYEKIRLSTQNVIDAEKAPGGPYRLGTSGSCPNATSKIGF
CYPGDFINYEELREQLSSVSSFERFEIF----PKTSSWPNHDSDNGVTAACPHAG-AKSF
CYPYDVPDYASLRSLVASSGTL---EFK----NESFNWTGV-KQNGTSSACIRGS-SSSF
*:*       .:*.           .:      .: .  .     *.:.:* .     .*

FATMAWAVPKN--KTATNPLTIEVPYICTEGEDQITVWGFHSDNEIQ-MAKLYGDSKPQK
FATMAWAVPKDNYKNATNPLTVEVPYICTEGEDQITVWGFHSDNKTQ-MKSLYGDSNPQK
YKNLIWLVKKGKSY-----PKINQTYINDKGKEVLVLWGIHHPPTIADQQSLYQNADAYV
FSRLNWLTHLNYTY-----PALNVTMPNNEQFDKLYIWGVHHPSTDKDQISLFAQPSGRI
:  : * .  .          ::      :  : : :**.*         .*: : .

FTSSANGVTTHYVSQIGGFPNQTEDGGLPQSGRIVVDYMVQKSGKTGTITYQRGILLPQK
FTSSANGVTTHYVSQIGDFPDQTEDGGLPQSGRIVVDYMMQKPGKTGTIVYQRGVLLPQK
FVG-TSRYSKKFKPEIATRPKVR-----DQEGRMNYYWTLVEPGDKITFEATGNLVAPRY
TVS-TKRSQQAVIPNIGSRPRIR-----DIPSRISIYWTIVKPGDILLINSTGNLIAPRG
 .. :.        :*.  *           .*:   : : : *.   :    .:: *:
```

```
VWCA--SGRSKVIKGSLPLIGEADCLHEKYGGLNKSKPYYTGEHAKAIGNCPIWVKT-PL
VWCA--SGRSKVIKGSLPLIGEADCLHEEYGGLNKSKPYYTGKHAKAIGNCPIWVKT-PL
AFTMERDAGSGIIISDTPVHDCNTTCQTPEGAINTSLPF-QNVHPITIGKCPKYVKSTKL
YFKIR-SGKSSIMRSDAPIGKCKSECITPNGSIPNDKPF-QNVNRITYGACPRYVKQSTL
 :    .. * :: .. *:           *.: .. *:  . :  : * ** :**   *

KLANGTKYRPPAKLLKERGFFGAIAGFLEGGWEGMIAGWHGYTSHGAHGVAVAADLKSTQ
KLANGTKYRPPAKLLKERGFFGAIAGFLEGGWEGMIAGWHGYTSHGAHGVAVAADLKSTQ
RLATGLRNVPS---IQSRGLFGAIAGFIEGGWTGMVDGWYGYHHQNEQGSGYAADLKSTQ
KLATGMRNVPE---KQTRGIFGAIAGFIENGWEGMVDGWYGFRHQNSEGRGQAADLKSTQ
:**.* :  *     :  **:*******:*.** **: **:*:  :. .* . ********

EAINKITKNLNSLSELEVKNLQRLSGAMDELHNEILELDEKVDDLRADTISSQIELAVLL
EAINKITKNLNSLSELEVKNLQRLSGAMDELHNEILELDEKVDDLRADTISSQIELAVLL
NAIDKITNKVNSVIEKMNTQFTAVGKEFNHLEKRIENLNKKVDDGFLDIWTYNAELLVLL
AAIDQINGKLNRLIGKTNEKFHQIEKEFSEVEGRVQDLEKYVEDTKIDLWSYNAELLVAL
 **::*. ::* :      ::  :   :..:. .: :*:: *:*    * : : ** * *

SNEGIINSEDEHLLALERKLKKMLGPSAVEIGNGCFETKHKCNQTCLDRIAAGTFDAGEF
SNEGIINSEDEHLLALERKLKKMLGPSAVDIGNGCFETKHKCNQTCLDRIAAGTFNAGEF
ENERTLDYHDSNVKNLYEKVRNQLKNNAKEIGNGCFEFYHKCDNTCMESVKNGTYDYPKY
ENQHTIDLTDSEMNKLFEKTKKQLRENAEDMGNGCFKIYHKCDNACIGSIRNETYDHNVY
.*:  ::  *..:  * .* :: *  .* ::*****:  ***:::*:  :   *::   :

SLPTFDS-LNITAASLNDDGLDNHTILLYYSTAASSLAVTLMIAIFVVYMVSRDSVSCSI
SLPTFDS-LNITAASLNDDGLDNHTILLYYSTAASSLAVTLMLAIFIVYMVSRDNVSCSI
SEEAKLNREKIDGVKL--DSTRIYQILAIYSTVASSLVLVVSLGAISFWMCSNGSLQCRI
RDEALNNRFQIKGVEL--KSGYKDWILWI-SFAMSCFLLCIALLGFIMWACQKGNIRCNI
   :  .  :* ...*  ..     **   * . *.: : : :  : .:  ....: * *
```

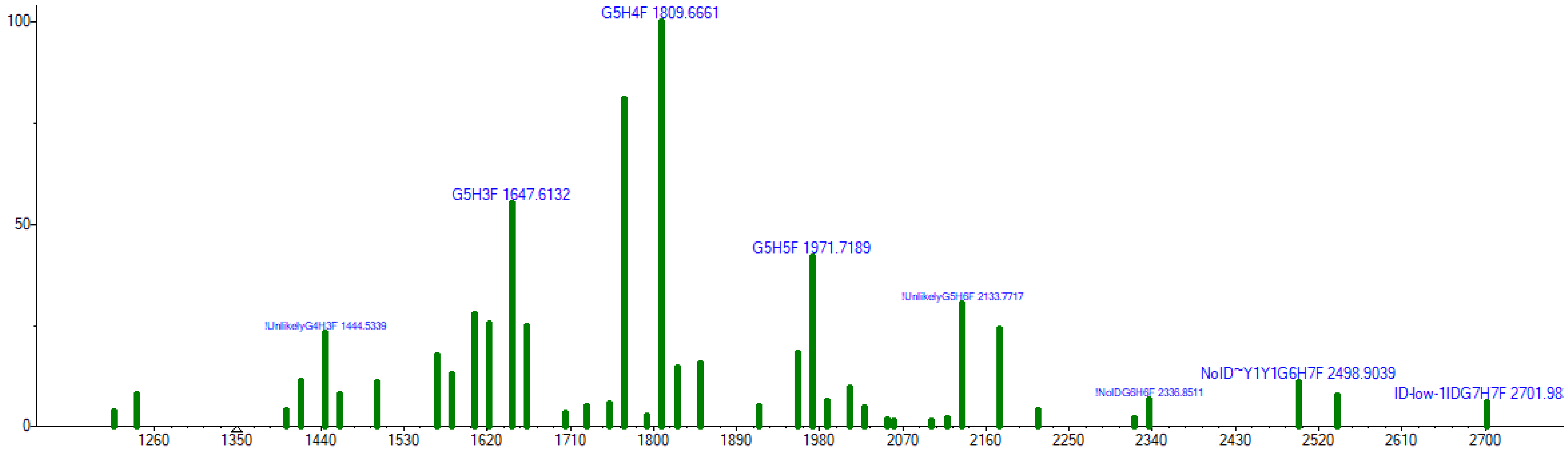
Afluria GADS

- 8→23 of the 40 sequons for the four hemagglutinin proteins identified. 27 are non-overlapping
- Three major GADS classes: high mannose, G4F complex, G5F complex
- No GADS were identified for NA proteins, probably due to a lower abundance of NA on the virus.



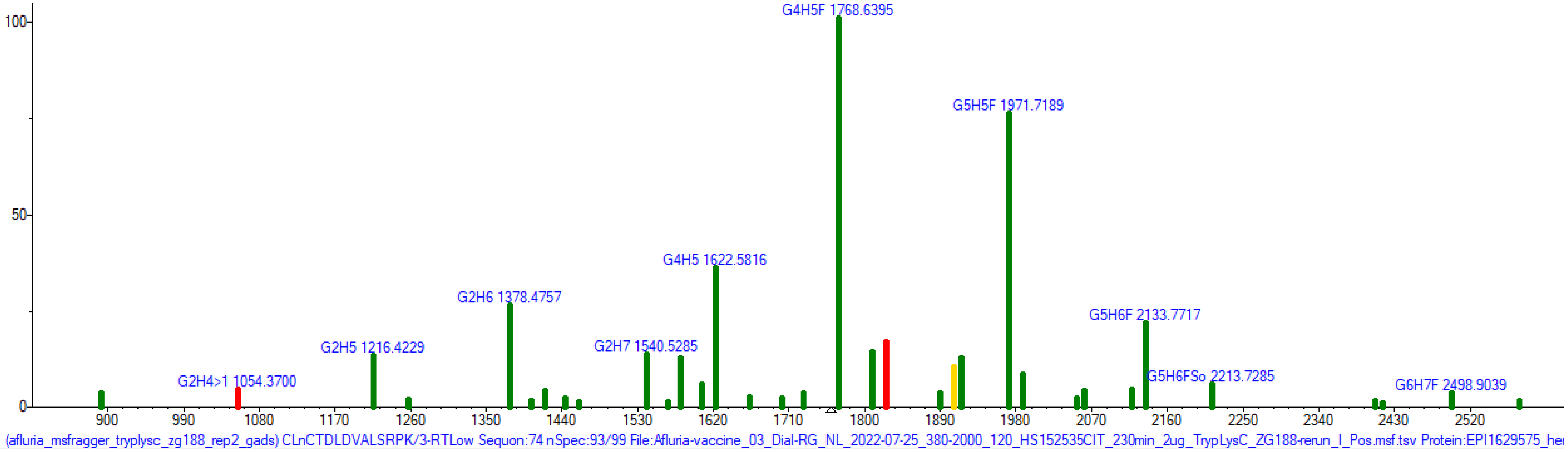
Afluria GADS

- 8→23 of the 40 sequons for the four hemagglutinin proteins identified. 27 are non-overlapping
- Three major GADS classes: high mannose, G4F complex, G5F complex
- No GADS were identified for NA proteins, probably due to a lower abundance of NA on the virus.



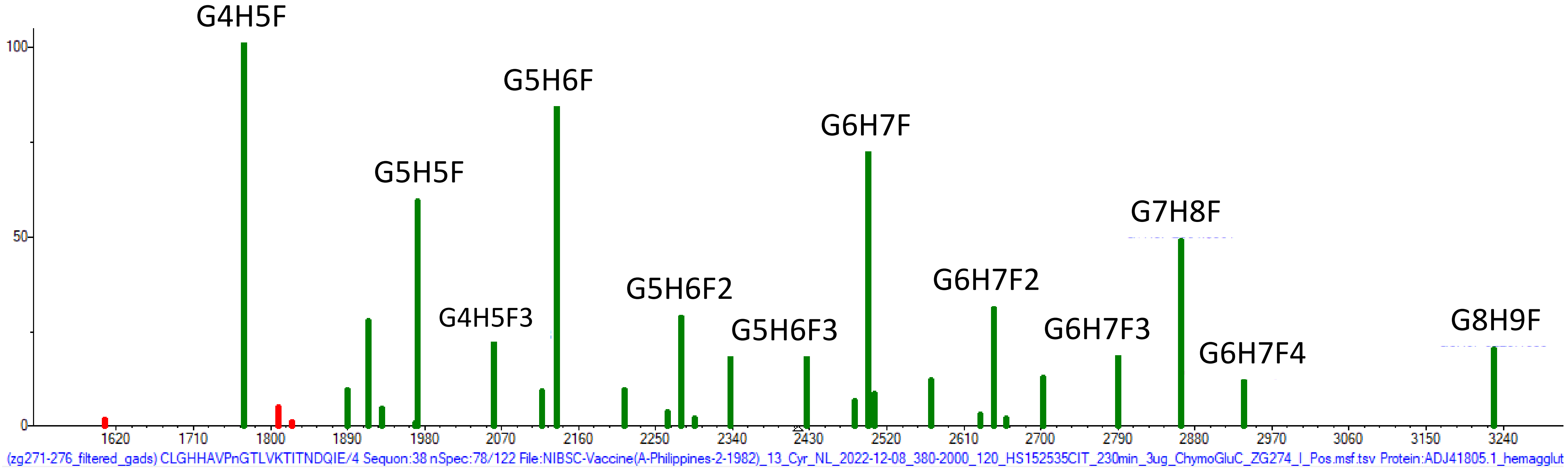
Afluria GADS

- 8→23 of the 40 sequons for the four hemagglutinin proteins identified. 27 are non-overlapping
- Three major GADS classes: high mannose, G4F complex, G5F complex
- No GADS were identified for NA proteins, probably due to a lower abundance of NA on the virus.



Monovalent vaccine GADS

- 90% sequons detected
- Three major GADS classes: high mannose, G4/G5, G4/G5 + high mannose
- Again, no major GADS identified for NA.



(zg271-276_filtered_gads) CLGHHAVPnGTLVKTITNDQIE/4 Sequon:38 nSpec:78/122 File:NIBSC-Vaccine(A-Philippines-2-1982)_13_Cyr_NL_2022-12-08_380-2000_120_HS152535CIT_230min_3ug_ChymoGluC_ZG274_I_Pos.msf.tsv Protein:ADJ41805.1_hemagglut

Open Search for Afluria Vaccine Modifications

Mass shift (Da)	Modification
-2.01	Val → Pro
14.01	Methylation
15.99	Oxidation
52.91	Fe(III)
57.02	+Gly
72.02	Carboxyethyl
113.08	+Leu/Ile
128.09	+Lys
129.04	+Glu

Formalin

B-Propiolactone

Next Steps

- Three more commercial vaccines were purchased for the 2022-2023 flu season
    - Flublok – Recombinant vaccine
    - Afluria – Egg based vaccine
    - Flucelvax – Cell culture based vaccine
    - How does glycosylation change… year to year
      … strain to strain
      … between sources

Conclusions

- Milk glycoproteins exhibit diverse GADS between sites, but are reproducible between labs and MS methods.
- Recombinant Sars-CoV-2 spike protein is highly reproducible in glycosylation distribution between vendor and processing method, but is very different in glycan abundance depending on the site.
- Hemagglutinin replicates are most similar in glycosylation profile and digests from different proteins are least similar in glycosylation profile.
- Homologous sequence regions between different influenza strains have similar glycosylation distribution compared to non-conserved regions.

Acknowledgments

Steve Stein
Bill Wallace

Meghan Burke
Connie Remoroza
Yi Liu
Yuri Mirokhin
Sergey Sheetlin
Dmitrii Tchekhovskoi
Guanghui Wang
Sara Yang

References

Jung, H. E., & Lee, H. K. (2020). Host protective immune responses against influenza A virus infection. *Viruses*, *12*(5), 504.

Remoroza, C. A., Burke, M. C., Liu, Y., Mirokhin, Y. A., Tchekhovskoi, D. V., Yang, X., & Stein, S. E. (2021). Representing and Comparing Site-Specific Glycan Abundance Distributions of Glycoproteins. Journal of Proteome Research, 20(9), 4475-4486.

https://www.cdc.gov/flu/vaccines-work/effectiveness-studies.htm

Zost, S. J., Parkhouse, K., Gumina, M. E., Kim, K., Perez, S. D., Wilson, P. C., ... & Hensley, S. E. (2017). Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains. *Proceedings of the National Academy of Sciences*, *114*(47), 12578-12583.

Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S., & Crispin, M. (2020). Site-specific glycan analysis of the SARS-CoV-2 spike. Science, 369(6501), 330-333.

Chang, D., & Zaia, J. (2019). Why glycosylation matters in building a better flu vaccine. Molecular & Cellular Proteomics, 18(12), 2348-2358.

Jang, Y. H., & Seong, B. L. (2014). Options and obstacles for designing a universal influenza vaccine. Viruses, 6(8), 3159-3180.

Questions?

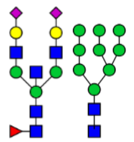
zachary.goecker@nist.gov

Isolation of Adjacent Sequons

- Site-specific glycosylation methods require a single sequon per peptide. Different proteases are used to maximize isolation.
- Influenza, HIV, Ebola, Herpesvirus, and MERS have adjacent sequons

NXT/S X≠P

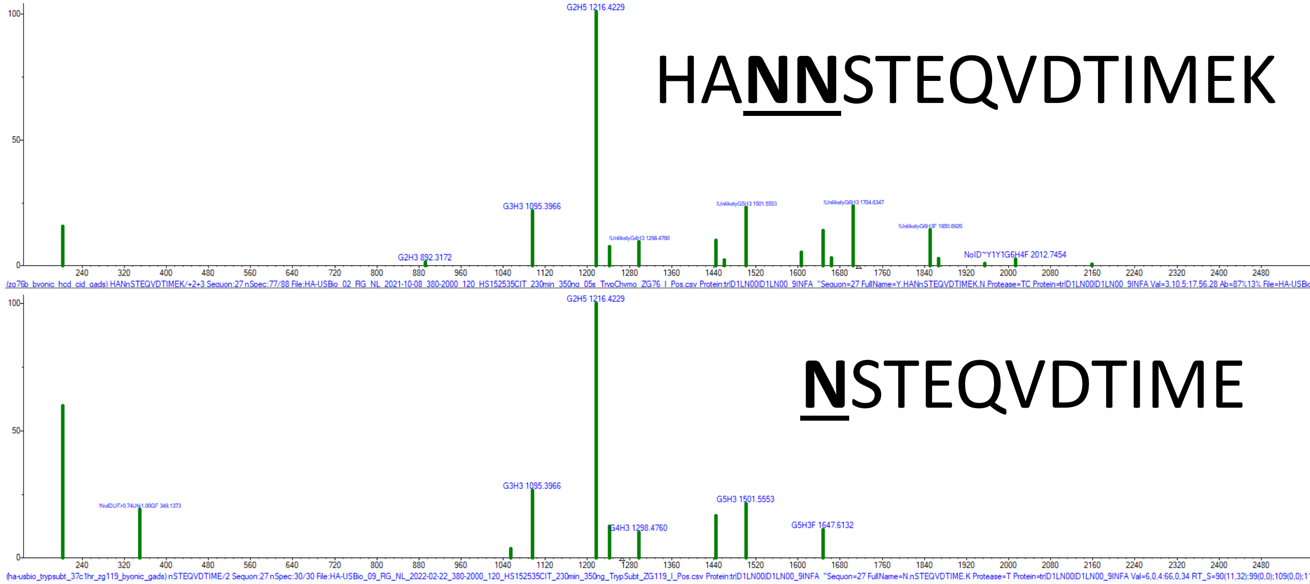
NNSS, NNTT, NNST, NNTS



…PIDETEQGSYNNTSGEQMLIIWGVHHR…

Isolation of Adjacent Sequons - Subtilisin

- *Bacillus subtilis* – extracellular serine endopeptidases



HA**NN**STEQVDTIMEK

NSTEQVDTIME

DQICIGYHA**N** ???

Isolation of Adjacent Sequons - Subtilisin

- *Bacillus subtilis* – extracellular serine endopeptidases

- Three more proteases from *Bacillus subtilis* were obtained
  - Esperase
  - Savinase
  - Neutrase

Supplemental Slides

Dot Product

Cosine of the angle between spectra represented as vectors

$$\frac{(\Sigma\, W_L W_U)^2}{\Sigma\, W_L^2\ \Sigma\, W_U^2}$$

W_L = Weighted intensity of library

W_U = Weighted intensity of unknown

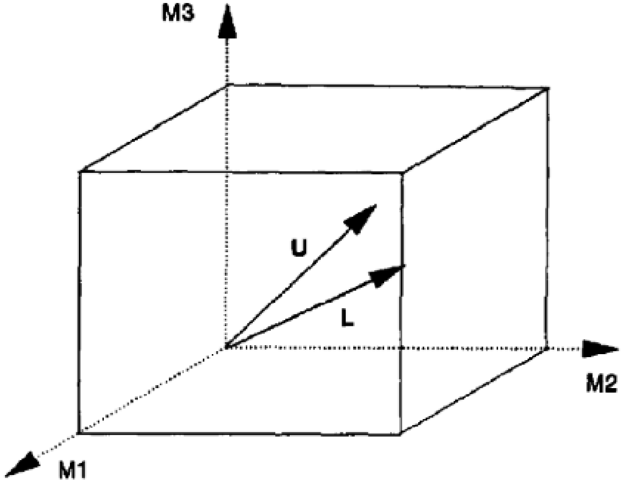


Figure 1. Vector representation of a hypothetical three-peak unknown (U) and library (L) mass spectrum in three-dimensional space (peaks have mass M1, M2, and M3).

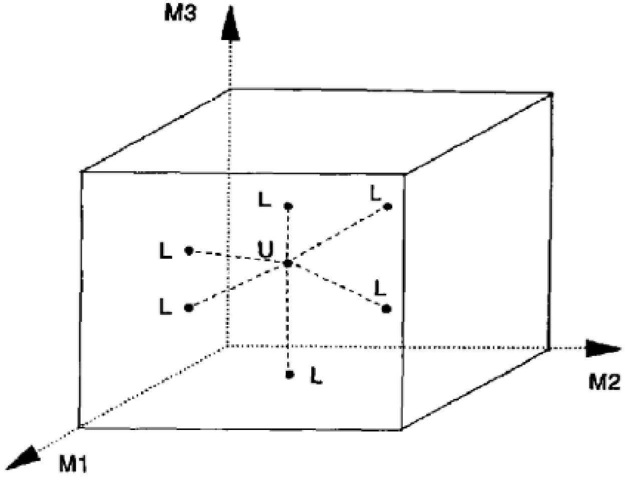
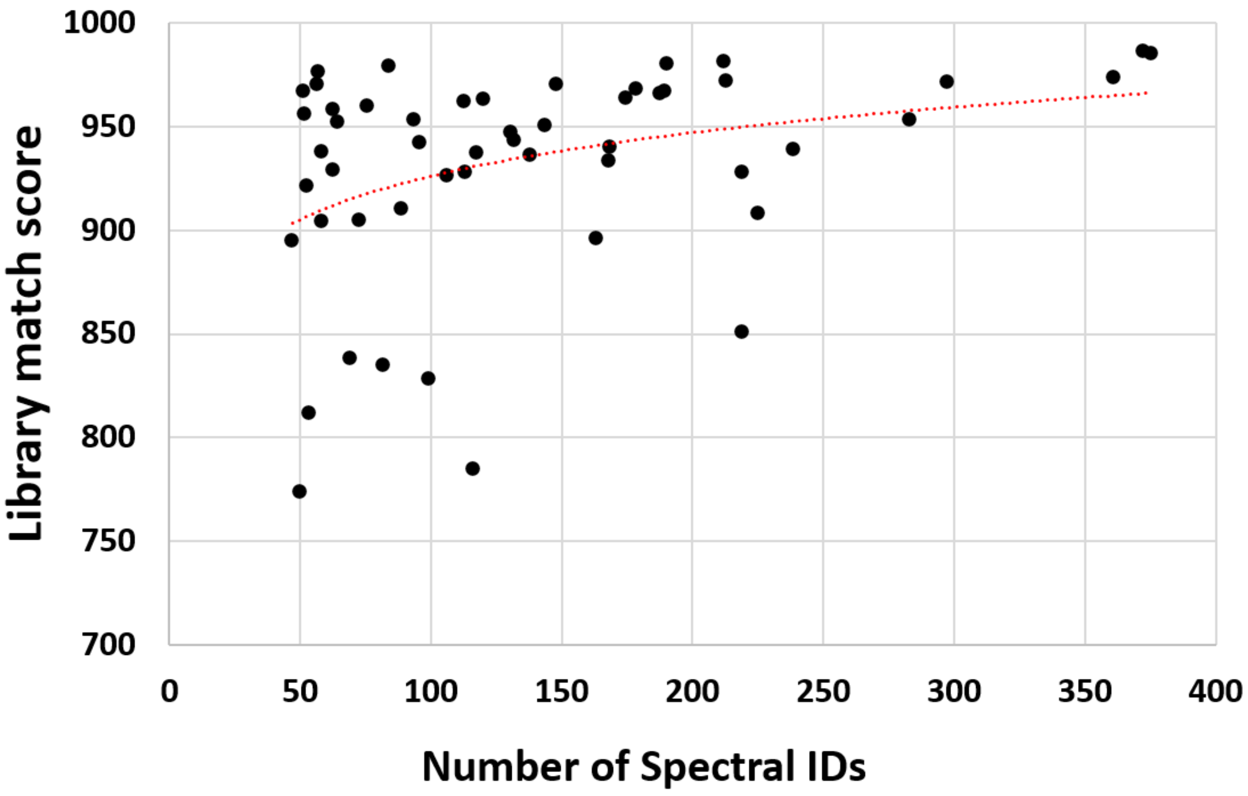


Figure 2. Point representation of library search results (L) for a hypothetical three-peak unknown (U) spectrum (masses M1, M2, and M3)

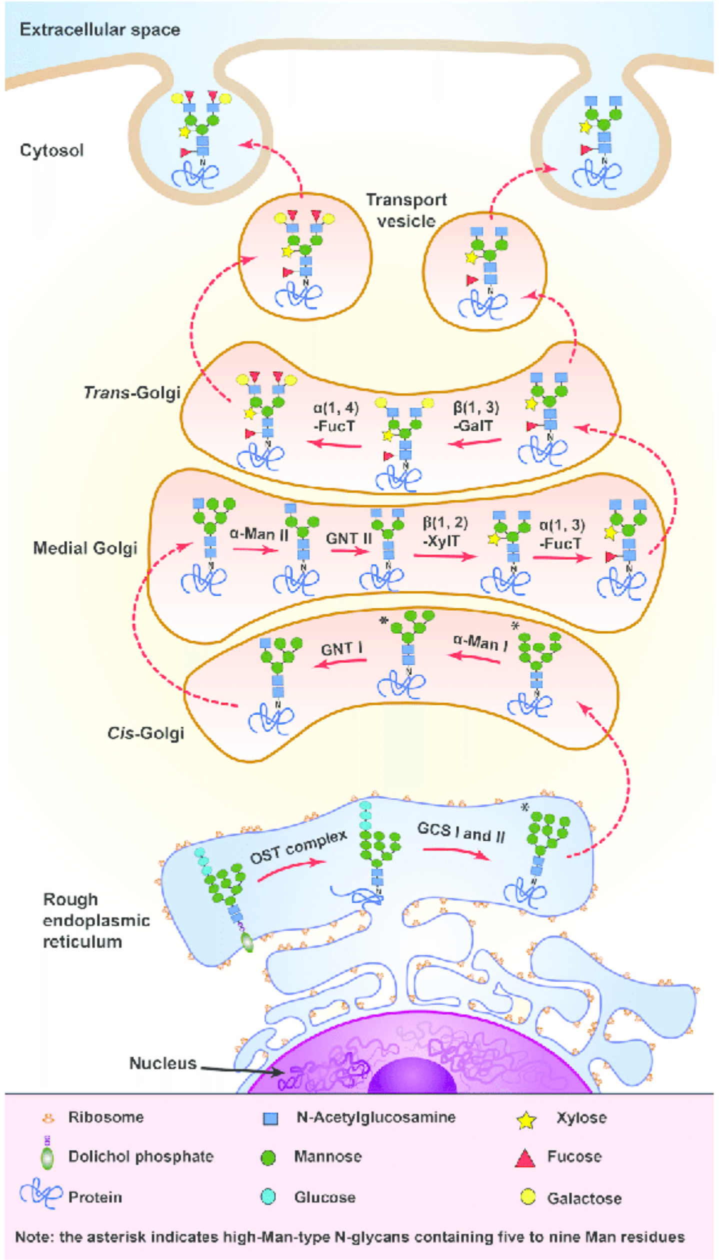
Supplemental Slides



Supplemental Slides

Glycobiology

- In viruses
  - Immune evasion
  - Host cell attachment
- In mammalian cells
  - Protein folding
  - Protein stabilization
  - Communication
  - Function

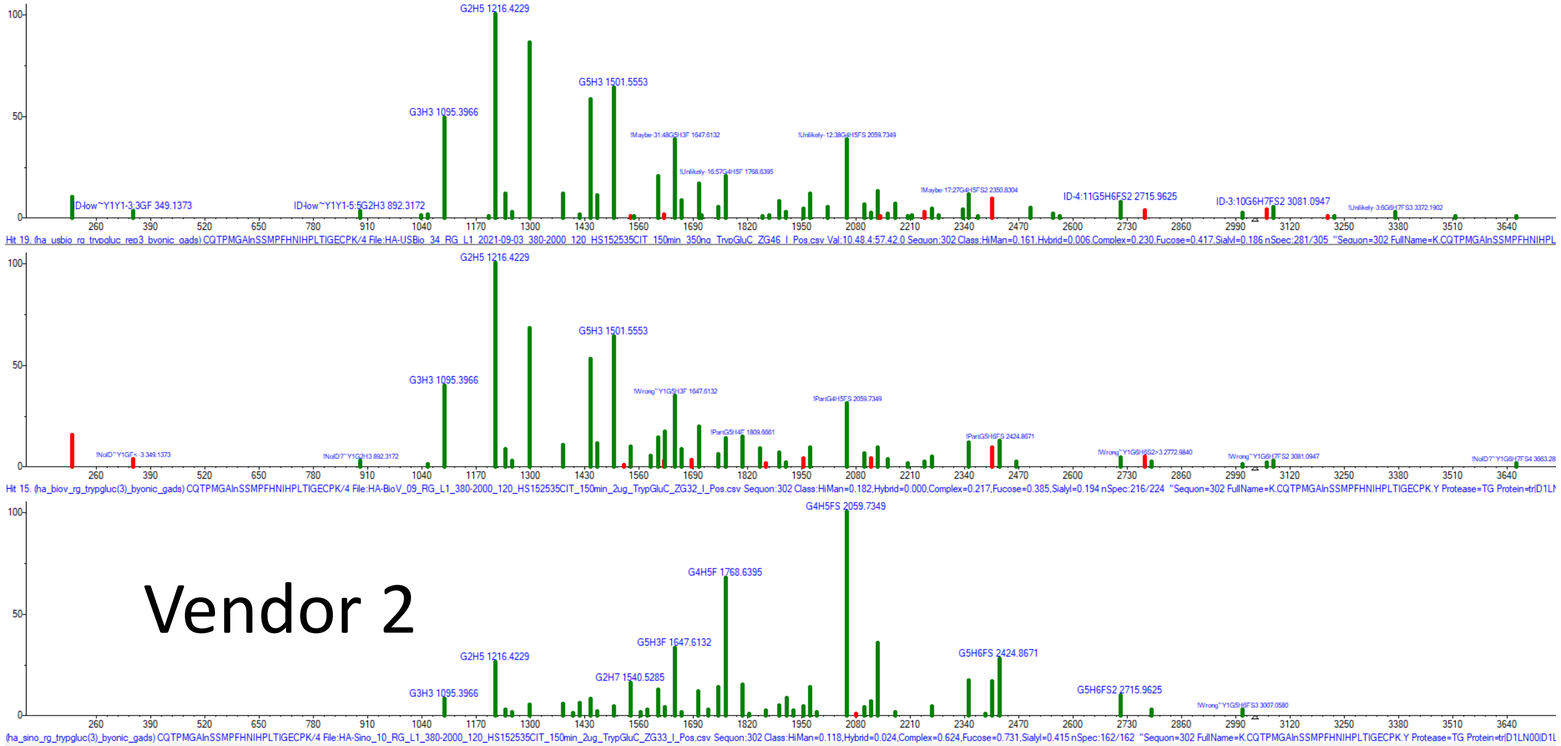


Note: the asterisk indicates high-Man-type N-glycans containing five to nine Man residues

Supplemental Slides

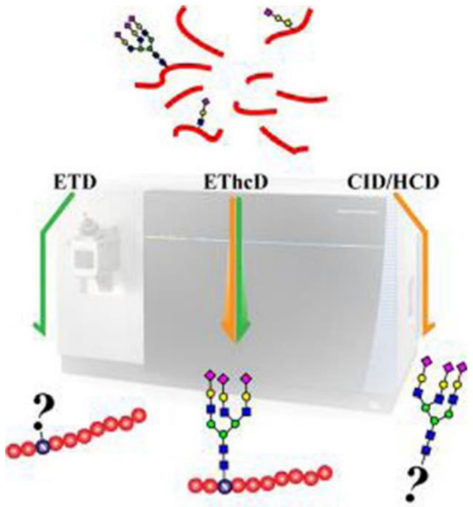
Vendor Comparison

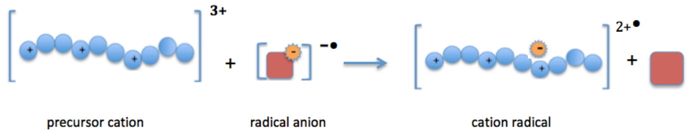
Glycan distribution was skewed toward higher mass (sialylated complex) glycans for vendor 2 and toward lower mass glycans (oligomannose and hybrid) for vendor 1 and vendor 3.



Supplemental Slides

EThcD





Fluoranthene

Stepped HCD

Peptide bond
Glycosidic bond
N-glycosidic bond