Identification, Mapping and Relative Quantitation of SARS-CoV-2 Spike Glycopeptides by Mass-Retention Time Fingerprinting

Chalk, R.[1], Greenland, W.[2], Moreira, T.[1], Coker, J.[1], Mukhopadhyay S.M.M[1], Williams, E.[1], Manning, C.[1], Bohstedt, T.[1], McCrorie, R.[1], Fernandez-Cid, A.[1], Burgess-Brown, N.A.[1]

[1]Centre for Medicines Discovery, ORCRB, Oxford University, OX3 7DQ, UK

[2]Agilent Technologies, Lakeside, Cheadle Royal Business Park, Cheadle, Cheshire, SK8 3GR, UK

Abstract

We describe a novel analytical method for rapid and robust identification, mapping and relative quantitation of glycopeptides from SARS-CoV-2 Spike protein. The method may be executed using any LC-TOF mass spectrometer, requires no specialised knowledge of glycan analysis and makes use of the differential resolving power of reversed phase HPLC. While this separation technique resolves peptides with high efficiency, glycans are resolved poorly, if at all. Consequently, glycopeptides consisting of the same peptide bearing different glycan structures will all possess very similar retention times and co-elute. While this has previously been viewed as a disadvantage, we show that shared retention time can be used to map multiple glycan species to the same peptide and location. In combination with MSMS and pseudo MS3, we have constructed a detailed mass-retention time database for Spike. This database allows any ESI-TOF equipped lab to reliably identify and quantify spike glycans from a single overnight elastase protein digest in less than 90 minutes.

Key words:

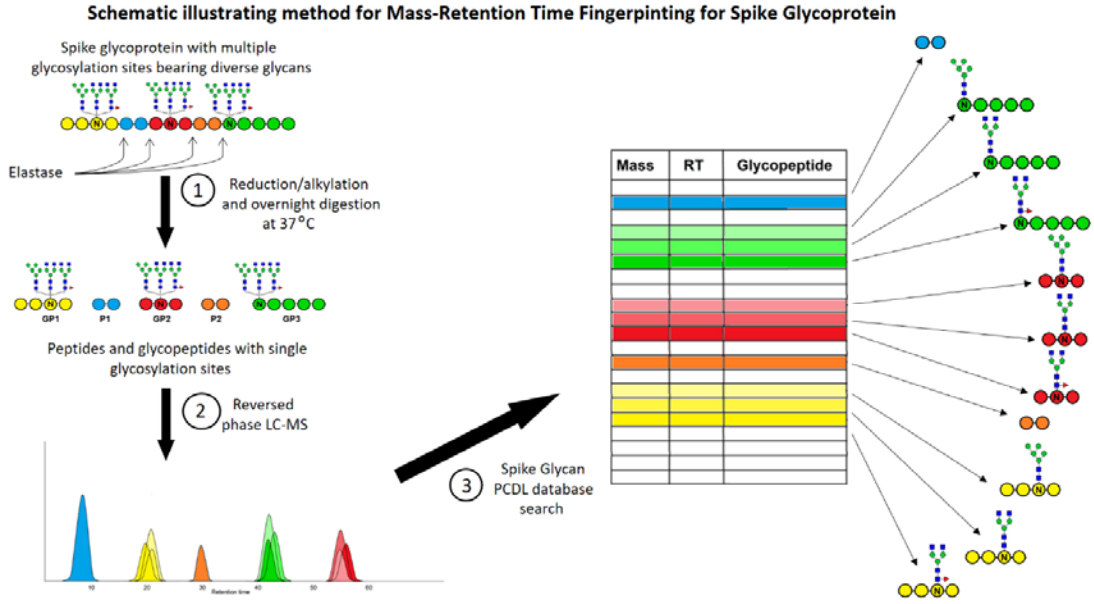
SARS-CoV-2, Spike, RBD, Glycoprotein, Glycopeptide, Glycan, Mass Spectrometry, HPLC, Database

Introduction

Glycosylation is known to play an important role in the efficacy and antigenicity of therapeutic proteins [1-3]. The current SARS-CoV-2 pandemic has spurred urgent research, much of it devoted to preparing vaccines, therapeutic antibodies or antibody tests based on Spike protein, the virus's primary surface antigen [4]. This 145 kDa protein forms a trimer [5] with each subunit bearing twenty-two potential N-linked glycosylation sites and two O-linked sites of which approximately seventeen are occupied [5]. The unusually heavy and complex glycosylation observed in Spike protein is believed to play an important role in the pathogenicity of SARS-CoV-2 by mimicking host cell glycans and allowing the virus to evade the normal immune response [6]. Analysis of expressed Spike protein by mass spectrometry presents unique challenges in terms of its size and the number and complexity of its glycans. These challenges have been commendably met to date by laboratories with wide experience in glycan analysis and access to very sensitive, high-end nano-LC-MSMS mass spectrometers [1, 7-9]. However, in our laboratory and in others a rapid and more robust methodology is needed for routine analysis of different batches of expressed Spike protein. In addition, any method which is reliant on LC-MSMS of glycopeptides may not necessarily detect specific glycans which fail to fragment under the conditions selected. LC-MS, by contrast, generates a mass, retention time and relative abundance for all ionizable species. We have developed a simple Mass-Retention Time Fingerprinting (MRTF) method for rapid and robust identification, mapping and relative quantitation of Spike glycans. Overnight digestion using a single enzyme followed by a 65-minute LC-MS run using any accurate mass instrument are the only experimental requirements. The resulting LC-MS data contains accurate mass, retention time and relative abundance values for each glycopeptide component. This dataset needs only to be matched against the pre-existing Spike glycopeptide database reported here, as shown in Figure 1. We describe this method as "analytical mode", which is both conceptually simple to understand, and straightforward to implement in a typical mass spectrometry laboratory. For scientific completeness, we also describe the "discovery mode" which we have used to generate the data for our Mass-Retention Time Fingerprinting

database. We stress, however, that there is no requirement for users to duplicate this discovery mode. The analytical mode in conjunction with the database we have provided is all that is necessary to characterise Spike glycans by MRTF.

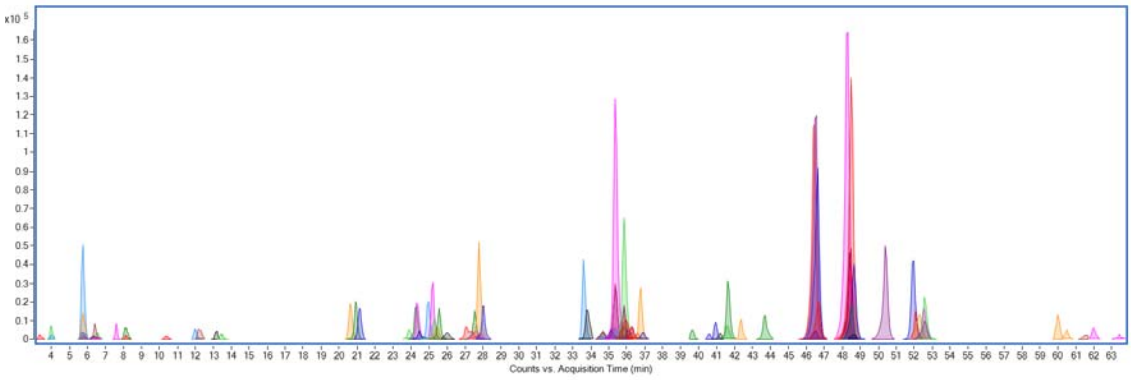
Figure 1. Mass-Retention Time Fingerprint analysis of Spike glycans



Results

Combined extracted ion chromatograms are illustrated in Figure 2. It can be seen that some species (peptides) are completely resolved, while many other species (glycans) co-elute.

Figure 2. Combined Extracted Ion Chromatogram (EIC) for 140 observed Spike glycopeptides separated by reversed phase LC-MS and identified by Mass-Retention Time Fingerprinting (MRTF)



We observed one hundred and forty glycopeptides by LC-MS. These are recorded grouped by ascending retention time in Tables 1a and 1b along with accurate masses, peptide sequences, glycan assignments and a key to the glycan structures. The location of each glycopeptide series on Spike is indicated in the first column. It may be seen that all observed glycans for the same peptide occur within a four minute retention time window. Accurate mass and estimated retention time are included for a further three hundred and six glycopeptides.

Table 1a. Spike elastase glycopeptide mass retention time database (PCDL) containing data for 140 observed glycopeptides and data for a further 306 inferred glycopeptides (RT 2-32 min)

Glycan posn.	RT (min)	Mass	Glycopeptide	Observed/ Inferred
N343	2.4	2151.9206	VFNATRF G0	Inf
	2.4	2297.9785	VFNATRF G0F	Inf
	2.4	1421.6562	VFNATRF Man1	Inf
	2.4	1583.709	VFNATRF Man2	Inf
	2.4	1907.8146	VFNATRF Man4	Inf
	2.4	2069.8674	VFNATRF Man5	Inf
	2.4	2231.9202	VFNATRF Man6	Inf
	2.4	2393.973	VFNATRF Man7	Inf
	2.4	2556.0258	VFNATRF Man8	Inf
	2.4	2718.0786	VFNATRF Man9	Inf
	2.464	1745.7618	VFNATRF Man3	Obs
N801	3.3	1792.6885	NFSQ G0	Inf
	3.3	1938.7464	NFSQ G0F	Inf
	3.3	1872.6881	NFSQ Man6	Inf
	3.3	2034.7409	NFSQ Man7	Inf
	3.3	2196.7937	NFSQ Man8	Inf
	3.3	2358.8465	NFSQ Man9	Inf
	3.364	1710.6353	NFSQ Man5	Obs
	3.8	1062.4241	NFSQ Man1	Inf
	3.8	1224.4769	NFSQ Man2	Inf
	3.888	1548.5825	NFSQ Man4	Obs
	3.898	1386.5297	NFSQ Man3	Obs
N17/234/331 ambiguous assignment (isobaric peptides)	5.636	2210.8192	NL/IT Man9	Obs
	5.7	752.344	NL/IT (GlcNAc)2	Inf
	5.7	549.2646	NL/IT GlcNAc stump	Inf
	5.7	1644.6612	NL/IT G0	Inf
	5.7	1790.7191	NL/IT G0F	Inf
	5.7	914.3968	NL/IT Man1	Inf
	5.7	1076.4496	NL/IT Man2	Inf
	5.7	1238.5024	NL/IT Man3	Inf
	5.7	346.1852	NL/IT peptide	Inf
	5.737	2048.7664	NL/IT Man8	Obs
	5.751	1886.7136	NL/IT Man7	Obs
	5.761	1724.6608	NL/IT Man6	Obs
	5.821	1562.608	NL/IT Man5	Obs
	6.549	1400.5552	NL/IT Man4	Obs
N801	6.4	2187.8412	DFGGFNFS G0	Inf
	6.4	2333.8991	DFGGFNFS G0F	Inf
	6.4	1457.5722	DFGGFNFS Man1	Inf
	6.4	1619.625	DFGGFNFS Man2	Inf
	6.4	1781.6778	DFGGFNFS Man3	Inf
	6.4	1943.7352	DFGGFNFS Man4	Inf
	6.4	2267.8408	DFGGFNFS Man6	Inf
	6.4	2429.889	DFGGFNFS Man7	Inf
	6.4	2591.9418	DFGGFNFS Man8	Inf
	6.4	2753.9946	DFGGFNFS Man9	Inf
	6.454	2105.7834	DFGGFNFS Man5	Obs
N149	6.537	1725.6561	NKS Man6	Obs
	6.554	1563.6033	NKS Man5	Obs
	7.2	1645.6565	NKS G0	Inf
	7.2	1791.7144	NKS G0F	Inf
	7.2	915.3921	NKS Man1	Inf
	7.2	1077.4449	NKS Man2	Inf
	7.2	1239.4977	NKS Man3	Inf
	7.2	1401.5505	NKS Man4	Inf
	7.2	1887.7089	NKS Man7	Inf
	7.2	2049.7617	NKS Man8	Inf
	7.2	2211.8145	NKS Man9	Inf
	7.271	1563.6033	NKS Man5	Obs
N282	7	2323.9062	YNENGTITD G0	Inf
	7	2469.9641	YNENGTITD G0F	Inf
	7	2485.959	YNENGTITD G1	Inf
	7	2632.0169	YNENGTITD G1F	Inf
	7	2794.0697	YNENGTITD G2F	Inf
	7	1593.6418	YNENGTITD Man1	Inf
	7	1755.6946	YNENGTITD Man2	Inf
	7	1917.7474	YNENGTITD Man3	Inf
	7	2079.8002	YNENGTITD Man4	Inf
	7	2241.853	YNENGTITD Man5	Inf
	7	2403.9058	YNENGTITD Man6	Inf
	7	2565.9586	YNENGTITD Man7	Inf
	7	2728.0114	YNENGTITD Man8	Inf
	7	2890.0642	YNENGTITD Man9	Inf
	7.054	2648.0118	YNENGTITD G2	Obs
N282	8.082	2120.8366	NGTITDAVD Man5	Obs
	8.144	2444.9422	NGTITDAVD Man7	Obs
	8.2	2202.8898	NGTITDAVD G0	Inf
	8.2	2348.9477	NGTITDAVD G0F	Inf
	8.2	1472.6254	NGTITDAVD Man1	Inf
	8.2	1634.6782	NGTITDAVD Man2	Inf
	8.2	1796.731	NGTITDAVD Man3	Inf
	8.2	1958.7838	NGTITDAVD Man4	Inf
	8.2	2606.995	NGTITDAVD Man8	Inf
	8.2	2769.0478	NGTITDAVD Man9	Inf
	8.239	2282.8894	NGTITDAVD Man6	Obs

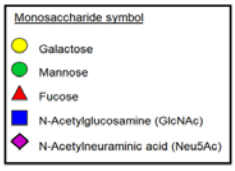
Glycan posn.	RT (min)	Mass	Glycopeptide	Observed/ Inferred
N717	10.3	1876.746	PTNFT G0	Inf
	10.3	2022.8039	PTNFT G0F	Inf
	10.3	2038.7988	PTNFT G1	Inf
	10.3	1146.4816	PTNFT Man1	Inf
	10.3	1308.5344	PTNFT Man2	Inf
	10.3	1470.5872	PTNFT Man3	Inf
	10.3	1632.64	PTNFT Man4	Inf
	10.3	1794.6928	PTNFT Man5	Inf
	10.3	1956.7456	PTNFT Man6	Inf
	10.3	2118.7984	PTNFT Man7	Inf
	10.3	2280.8512	PTNFT Man8	Inf
	10.3	2442.904	PTNFT Man9	Inf
	10.365	2184.8567	PTNFT G1F	Obs
	10.4	2200.8516	PTNFT G2	Inf
	10.448	2346.9095	PTNFT G2F	Obs
N657	11.995	2533.9915	CLIGAEHVNNS Man6	Obs
	12.1	2453.9919	CLIGAEHVNNS G0	Inf
	12.1	2600.0498	CLIGAEHVNNS G0F	Inf
	12.105	2371.9387	CLIGAEHVNNS Man5	Obs
	12.204	2209.8859	CLIGAEHVNNS Man4	Obs
N165	13.161	2825.0954	EFRVYSSANNCT Man6	Obs
	13.4	2745.0958	EFRVYSSANNCT G0	Inf
	13.4	2891.1537	EFRVYSSANNCT G0F	Inf
	13.4	2014.8314	EFRVYSSANNCT Man1	Inf
	13.4	2176.8842	EFRVYSSANNCT Man2	Inf
	13.4	2338.937	EFRVYSSANNCT Man3	Inf
	13.4	2500.9898	EFRVYSSANNCT Man4	Inf
	13.4	2987.1482	EFRVYSSANNCT Man7	Inf
	13.4	3149.201	EFRVYSSANNCT Man8	Inf
	13.4	3311.2538	EFRVYSSANNCT Man9	Inf
	13.446	2663.0426	EFRVYSSANNCT Man5	Obs
N165	19.8	3001.1428	YSSANNCTFEYVS G1	Inf
	19.852	3147.2007	YSSANNCTFEYVS G1F	Obs
	20.2	2839.09	YSSANNCTFEYVS G0	Inf
	20.2	2108.8256	YSSANNCTFEYVS Man1	Inf
	20.2	2270.8784	YSSANNCTFEYVS Man2	Inf
	20.2	2432.9312	YSSANNCTFEYVS Man3	Inf
	20.2	2594.984	YSSANNCTFEYVS Man4	Inf
	20.2	2757.0368	YSSANNCTFEYVS Man5	Inf
	20.2	2919.0896	YSSANNCTFEYVS Man6	Inf
	20.2	3081.1424	YSSANNCTFEYVS Man7	Inf
	20.2	3243.1952	YSSANNCTFEYVS Man8	Inf
	20.2	3405.248	YSSANNCTFEYVS Man9	Inf
	20.29	2985.1479	YSSANNCTFEYVS G0F	Inf
GP1	20.61	2744.1259	GP1+Hex2	Inf
	20.926	2780.0707	GP1+360.1	Obs
	21.144	2420.0129	GP1	Obs
	21.41	2456.9655	GP1+37.0	Obs
N165	20.8	2676.0267	SSANNCTFEYVS G0	Inf
	20.8	1945.7623	SSANNCTFEYVS Man1	Inf
	20.8	2107.8151	SSANNCTFEYVS Man2	Inf
	20.8	2269.8679	SSANNCTFEYVS Man3	Inf
	20.8	2431.9207	SSANNCTFEYVS Man4	Inf
	20.8	2593.9735	SSANNCTFEYVS Man5	Inf
	20.8	2756.0263	SSANNCTFEYVS Man6	Inf
	20.8	2918.0791	SSANNCTFEYVS Man7	Inf
	20.8	3080.1319	SSANNCTFEYVS Man8	Inf
	20.8	3242.1847	SSANNCTFEYVS Man9	Inf
	20.881	2822.0846	SSANNCTFEYVS G0F	Obs
N616	20.9	2457.9939	QDVNCTEVPV G0	Inf
	20.9	1727.7295	QDVNCTEVPV Man1	Inf
	20.9	1889.7823	QDVNCTEVPV Man2	Inf
	20.9	2051.8351	QDVNCTEVPV Man3	Inf
	20.9	2213.8879	QDVNCTEVPV Man4	Inf
	20.9	2375.9407	QDVNCTEVPV Man5	Inf
	20.9	2537.9935	QDVNCTEVPV Man6	Inf
	20.9	2700.0463	QDVNCTEVPV Man7	Inf
	20.9	2862.0991	QDVNCTEVPV Man8	Inf
	20.9	3024.1519	QDVNCTEVPV Man9	Inf
	20.982	2604.0518	QDVNCTEVPV G0F	Obs
N282	21.2	2677.0431	NENGTITDAVDCA G0	Inf
	21.2	1946.7787	NENGTITDAVDCA Man1	Inf
	21.2	2108.8315	NENGTITDAVDCA Man2	Inf
	21.2	2270.8843	NENGTITDAVDCA Man3	Inf
	21.2	2432.9371	NENGTITDAVDCA Man4	Inf
	21.2	2594.9899	NENGTITDAVDCA Man5	Inf
	21.2	2757.0427	NENGTITDAVDCA Man6	Inf
	21.2	2919.0955	NENGTITDAVDCA Man7	Inf
	21.2	3081.1483	NENGTITDAVDCA Man8	Inf
	21.2	3243.2011	NENGTITDAVDCA Man9	Inf
	21.236	2823.101	NENGTITDAVDCA G0F	Obs
N165	21.7	3255.2204	NCTFEYVSQPF Man9	Inf
	21.758	3093.1676	NCTFEYVSQPF Man8	Obs
	22	2689.0624	NCTFEYVSQPF G0	Inf
	22	2835.1203	NCTFEYVSQPF G0F	Inf
	22	1958.798	NCTFEYVSQPF Man1	Inf
	22	2120.8508	NCTFEYVSQPF Man2	Inf
	22	2282.9036	NCTFEYVSQPF Man3	Inf
	22	2444.9564	NCTFEYVSQPF Man4	Inf
	22	2607.0092	NCTFEYVSQPF Man5	Inf
	22	2769.062	NCTFEYVSQPF Man6	Inf
	22.044	2931.1148	NCTFEYVSQPF Man7	Obs

Glycan posn.	RT (min)	Mass	Glycopeptide	Observed/ Inferred
N343	23.295	2438.9204	GEVFNAT Man8	Obs
	23.399	2838.1073	GEVFNAT Complex NeuAc (F)2	Obs
	23.762	2276.8676	GEVFNAT Man7	Obs
	23.8	2600.9732	GEVFNAT Man9	Inf
	23.94	2692.0536	GEVFNAT Complex NeuAc F	Obs
	23.962	2114.8148	GEVFNAT Man6	Obs
	24.203	1628.6564	GEVFNAT Man3	Obs
	24.206	2488.9838	GEVFNAT G1(F)2	Obs
	24.272	2504.9787	GEVFNAT G2F	Obs
	24.3	1304.5508	GEVFNAT Man1	Inf
	24.3	1466.6036	GEVFNAT Man2	Inf
	24.349	1790.7092	GEVFNAT Man4	Obs
	24.352	1952.762	GEVFNAT Man5	Obs
	24.618	2383.9525	GEVFNAT G0F+GlcNAc	Obs
	24.7	2196.868	GEVFNAT G1	Inf
	24.7	2358.9208	GEVFNAT G2	Inf
	24.713	2342.9259	GEVFNAT G1F	Obs
	24.9	2034.8152	GEVFNAT G0	Inf
	24.967	2180.8731	GEVFNAT G0F	Obs
	25	2237.8946	GEVFNAT G0+GlcNAc	Inf
	25.323	2383.9525	GEVFNAT G0F+GlcNAc	Obs
	25.4	2650.0162	GEVFNAT A1	Inf
	25.4	2941.1116	GEVFNAT A2	Inf
	25.476	2983.1586	GEVFNAT A1(F)2-Gal+GlcNAc	Obs
	25.492	2796.0741	GEVFNAT A1F	Obs
	25.734	2837.1007	GEVFNAT A1F-Gal+GlcNAc	Obs
	25.771	3087.1695	GEVFNAT A2F	Obs
	25.9	1142.49791	GEVFNAT (GlcNAc)2	Inf
	25.9	939.4186	GEVFNAT GlcNAc stump	Inf
	25.9555	736.33917	GEVFNAT peptide	Obs
	26.023	2634.0213	GEVFNAT A1F-Gal	Obs
	26.161	3109.1451	GEVFNAT Very Complex	Obs
GP2	25.198	1545.693	GP2	Obs
	25.587	2173.9303	GP2+628.0	Obs
N343	25.2	2123.8516	(G)EVFNAT G0F	Inf
	25.2	1895.7405	(G)EVFNAT Man5	Inf
	25.203	1977.7937	(G)EVFNAT G0	Obs
	26.5	1933.7675	GEVFNA(T) G0	Inf
	26.5	2079.8254	GEVFNA(T) G0F	Inf
	26.566	1851.7143	GEVFNA(T) Man5	Obs
N603	27.039	3295.3341	GGVSVITPGTNTSNQ Man9	Obs
	27.6	2729.1761	GGVSVITPGTNTSNQ G0	Inf
	27.6	2875.234	GGVSVITPGTNTSNQ G0F	Inf
	27.6	1998.9117	GGVSVITPGTNTSNQ Man1	Inf
	27.6	2160.9645	GGVSVITPGTNTSNQ Man2	Inf
	27.6	2323.0173	GGVSVITPGTNTSNQ Man3	Inf
	27.6	2485.0701	GGVSVITPGTNTSNQ Man4	Inf
	27.6	2647.1229	GGVSVITPGTNTSNQ Man5	Inf
	27.6	2809.1757	GGVSVITPGTNTSNQ Man6	Inf
	27.6	2971.2285	GGVSVITPGTNTSNQ Man7	Inf
	27.627	3133.2813	GGVSVITPGTNTSNQ Man8	Obs
N165	27.7	3092.2109	MESEFRVYSSANNCT G0	Inf
	27.7	3238.2688	MESEFRVYSSANNCT G0F	Inf
	27.7	2361.9465	MESEFRVYSSANNCT Man1	Inf
	27.7	2523.9993	MESEFRVYSSANNCT Man2	Inf
	27.7	2686.0521	MESEFRVYSSANNCT Man3	Inf
	27.7	2848.1049	MESEFRVYSSANNCT Man4	Inf
	27.7	3172.2105	MESEFRVYSSANNCT Man6	Inf
	27.7	3334.2633	MESEFRVYSSANNCT Man7	Inf
	27.7	3496.3161	MESEFRVYSSANNCT Man8	Inf
	27.7	3658.3689	MESEFRVYSSANNCT Man9	Inf
	27.778	3010.1577	MESEFRVYSSANNCT Man5	Obs
GP3	27.785	2971.2343	GP3+GlcNAc	Obs
	28.022	2768.1567	GP3	Obs
N801	29.861	2703.1208	PPIKDFGGFNFS Man6	Obs
	30.5	2623.1212	PPIKDFGGFNFS G0	Inf
	30.5	2769.1791	PPIKDFGGFNFS G0F	Inf
	30.5	2054.9096	PPIKDFGGFNFS Man2	Inf
	30.5	2216.9624	PPIKDFGGFNFS Man3	Inf
	30.5	2379.0152	PPIKDFGGFNFS Man4	Inf
	30.5	2541.068	PPIKDFGGFNFS Man5	Inf
	30.5	2865.1736	PPIKDFGGFNFS Man7	Inf
	30.5	3027.2264	PPIKDFGGFNFS Man8	Inf
	30.5	3189.2792	PPIKDFGGFNFS Man9	Inf
	31.037	2703.1208	PPIKDFGGFNFS Man6	Obs
N331	31.2	2206.9363	RFPNIT G1	Inf
	31.204	2352.9942	RFPNIT G1F	Obs
	32	2044.8835	RFPNIT G0	Inf
	32	1314.6191	RFPNIT Man1	Inf
	32	1476.6719	RFPNIT Man2	Inf
	32	1638.7247	RFPNIT Man3	Inf
	32	1800.7775	RFPNIT Man4	Inf
	32	1962.8303	RFPNIT Man5	Inf
	32	2124.8831	RFPNIT Man6	Inf
	32	2286.9359	RFPNIT Man7	Inf
	32	2448.9887	RFPNIT Man8	Inf
	32	2611.0415	RFPNIT Man9	Inf
	32.044	2190.9414	RFPNIT G0F	Obs

Table 1b. Spike elastase glycopeptide mass retention time database (PCDL) containing data for 140 observed glycopeptides and data for a further 306 inferred glycopeptides (RT 32-60 min and key)

Glycan posn.	RT (min)	Mass	Glycopeptide	Observed/ Inferred
N149	32.4	3333.3262	VYYHKNNKSWM Man9	Inf
	32.463	2847.1678	VYYHKNNKSWM Man6	Obs
	32.463	3171.2734	VYYHKNNKSWM Man8	Obs
	32.6	2767.1682	VYYHKNNKSWM G0	Inf
	32.6	2913.2261	VYYHKNNKSWM G0F	Inf
	32.6	2036.9038	VYYHKNNKSWM Man1	Inf
	32.6	2198.9566	VYYHKNNKSWM Man2	Inf
	32.6	2361.0094	VYYHKNNKSWM Man3	Inf
	32.6	2523.0622	VYYHKNNKSWM Man4	Inf
	32.6	2685.115	VYYHKNNKSWM Man5	Inf
	32.682	3009.2206	VYYHKNNKSWM Man7	Obs
GP4	33.611	2694.1185	GP4 Man6	Obs
	33.8	2760.1768	GP4 G0F	Inf
	33.819	2532.0657	GP4 Man5	Obs
N17	32.463	2828.1148	CVNLTTRT Man9	Obs
	33.8	2261.9568	CVNLTTRT G0	Inf
	33.8	2408.0147	CVNLTTRT G0F	Inf
	33.8	1531.6924	CVNLTTRT Man1	Inf
	33.8	1693.7452	CVNLTTRT Man2	Inf
	33.8	1855.798	CVNLTTRT Man3	Inf
	33.8	2017.8508	CVNLTTRT Man4	Inf
	33.8	2179.9036	CVNLTTRT Man5	Inf
	33.8	2341.9564	CVNLTTRT Man6	Inf
	33.8	2504.0092	CVNLTTRT Man7	Inf
	33.841	2666.062	CVNLTTRT Man8	Obs
N343	34.6	1848.7511	VFNAT G0	Inf
	34.6	1994.809	VFNAT G0F	Inf
	34.6	1118.4867	VFNAT Man1	Inf
	34.6	1280.5395	VFNAT Man2	Inf
	34.6	1442.5923	VFNAT Man3	Inf
	34.6	2252.8563	VFNAT Man8	Inf
	34.6	2414.9091	VFNAT Man9	Inf
	34.656	1928.7507	VFNAT Man6	Obs
	34.674	2090.8035	VFNAT Man7	Obs
	34.682	1766.6979	VFNAT Man5	Obs
	34.688	1604.6451	VFNAT Man4	Obs
N717	35.101	2195.8349	NFTI Man8	Obs
	35.353	1709.6765	NFTI Man5	Obs
	35.373	1385.5709	NFTI Man3	Obs
	35.374	2033.7821	NFTI Man7	Obs
	35.8	2357.8877	NFTI Man9	Inf
	35.86	1871.7293	NFTI Man6	Obs
	36.2	2406.9307	NFTI A1	Inf
	36.2	2552.9886	NFTI A1F	Inf
	36.2	2698.0261	NFTI A2	Inf
	36.2	2844.084	NFTI A2F	Inf
	36.2	1791.7297	NFTI G0	Inf
	36.2	1994.8091	NFTI G0 +GlcNAc	Obs
	36.2	2140.867	NFTI G0F +GlcNAc	Obs
	36.2	1953.7825	NFTI G1	Inf
	36.2	2115.8353	NFTI G2	Inf
	36.2	2261.8932	NFTI G2F	Inf
	36.2	1061.4653	NFTI Man1	Inf
	36.2	1223.5181	NFTI Man2	Inf
	36.295	1547.6237	NFTI Man4	Obs
	36.599	2099.8404	NFTI G1F	Obs
	36.77	2089.8308	NFTI Man5+380.2	Obs
	36.896	1937.7876	NFTI G0F	Obs
N61	35.4	2765.11	PFFSNVTW G2F	Inf
	35.4	1564.6821	PFFSNVTW Man1	Inf
	35.4	1726.7349	PFFSNVTW Man2	Inf
	35.4	1888.7877	PFFSNVTW Man3	Inf
	35.4	2050.8405	PFFSNVTW Man4	Inf
	35.4	2212.8933	PFFSNVTW Man5	Inf
	35.4	2536.9989	PFFSNVTW Man7	Inf
	35.4	2699.0517	PFFSNVTW Man8	Inf
	35.4	2861.1045	PFFSNVTW Man9	Inf
	35.437	2619.0521	PFFSNVTW G2	Obs
	35.601	2456.9993	PFFSNVTW G1	Obs
	36	2498.0259	PFFSNVTW G0 +GlcNAc	Inf
	36	2441.0044	PFFSNVTW G0F	Inf
	36	2644.0838	PFFSNVTW G0F +GlcNAc	Inf
	36	2603.0572	PFFSNVTW G1F	Inf
	36.025	2294.9465	PFFSNVTW G0	Obs
	36.854	2374.9461	PFFSNVTW Man6	Obs
N343	36.5	2181.8836	FGEVFNAT G0	Inf
	36.5	2327.9415	FGEVFNAT G0F	Inf
	36.5	1451.6192	FGEVFNAT Man1	Inf
	36.5	1613.672	FGEVFNAT Man2	Inf
	36.5	1775.7248	FGEVFNAT Man3	Inf
	36.5	1937.7776	FGEVFNAT Man4	Inf
	36.5	2261.8832	FGEVFNAT Man6	Inf
	36.5	2423.936	FGEVFNAT Man7	Inf
	36.5	2585.9888	FGEVFNAT Man8	Inf
	36.5	2748.0416	FGEVFNAT Man9	Inf
	36.528	2099.8304	FGEVFNAT Man5	Obs

Glycan posn.	RT (min)	Mass	Glycopeptide	Observed/ Inferred
N343	36.8	2552.0511	LCPFGEVFNAT G0	Inf
	36.8	2698.109	LCPFGEVFNAT G0F	Inf
	36.8	1821.7867	LCPFGEVFNAT Man1	Inf
	36.8	1983.8395	LCPFGEVFNAT Man2	Inf
	36.8	2145.8923	LCPFGEVFNAT Man3	Inf
	36.8	2307.9451	LCPFGEVFNAT Man4	Inf
	36.8	2469.9979	LCPFGEVFNAT Man5	Inf
	36.8	2794.1035	LCPFGEVFNAT Man7	Inf
	36.8	2956.1563	LCPFGEVFNAT Man8	Inf
	36.8	3118.2091	LCPFGEVFNAT Man9	Inf
	36.864	2632.0507	LCPFGEVFNAT Man6	Obs
GP6	40.948	1570.7208	GP6	Obs
	41.645	1881.7363	GP6 +311.0	Obs
N331	42.6	1855.7569	PNITN G0	Inf
	42.6	2001.8148	PNITN G0F	Inf
	42.6	1125.4925	PNITN Man1	Inf
	42.6	1287.5453	PNITN Man2	Inf
	42.6	1611.6509	PNITN Man4	Inf
	42.6	1773.7037	PNITN Man5	Inf
	42.6	1935.7565	PNITN Man6	Inf
	42.6	2097.8093	PNITN Man7	Inf
	42.6	2259.8621	PNITN Man8	Inf
	42.6	2421.9149	PNITN Man9	Inf
	42.631	1449.5981	PNITN Man3	Obs
N801	46.236	3854.7283	KQIYKTPPIKDFGGFNFS G2F	Obs
	46.369	3788.67	KQIYKTPPIKDFGGFNFS Man8	Obs
	46.424	3626.6226	KQIYKTPPIKDFGGFNFS G0F +96.0	Obs
	46.5	3546.6176	KQIYKTPPIKDFGGFNFS G1	Inf
	46.5	3692.6755	KQIYKTPPIKDFGGFNFS G1F	Inf
	46.5	3708.6704	KQIYKTPPIKDFGGFNFS G2	Inf
	46.5	3950.7228	KQIYKTPPIKDFGGFNFS Man9	Inf
	46.506	3464.5644	KQIYKTPPIKDFGGFNFS Man6	Obs
	46.506	3626.6172	KQIYKTPPIKDFGGFNFS Man7	Obs
	46.509	3530.6227	KQIYKTPPIKDFGGFNFS G0F	Obs
	46.509	3733.7021	KQIYKTPPIKDFGGFNFS G0F +GlcNAc	Obs
	46.538	3140.4588	KQIYKTPPIKDFGGFNFS Man4	Obs
	46.568	2978.406	KQIYKTPPIKDFGGFNFS Man3	Obs
	46.575	2654.3004	KQIYKTPPIKDFGGFNFS Man1	Obs
	46.581	2816.3532	KQIYKTPPIKDFGGFNFS Man2	Obs
	46.601	3302.5116	KQIYKTPPIKDFGGFNFS Man5	Obs
	46.668	3384.5648	KQIYKTPPIKDFGGFNFS G0	Obs
	46.966	3587.6442	KQIYKTPPIKDFGGFNFS G0 +GlcNAc	Obs
	47.5	4145.8237	KQIYKTPPIKDFGGFNFS A1F	Inf
	47.5	4290.8612	KQIYKTPPIKDFGGFNFS A2	Inf
	47.5	4436.9191	KQIYKTPPIKDFGGFNFS A2F	Inf
	48.492	3999.7658	KQIYKTPPIKDFGGFNFS A1	Obs
N801	48.1	3630.5282	YKTPPIKDFGGFNFS A1	Inf
	48.1	3776.5861	YKTPPIKDFGGFNFS A1F	Inf
	48.1	3921.6236	YKTPPIKDFGGFNFS A2	Inf
	48.1	4067.6815	YKTPPIKDFGGFNFS A2F	Inf
	48.121	3485.4907	YKTPPIKDFGGFNFS G2F	Obs
	48.27	4145.822	YKTPPIKDFGGFNFS Very complex	Obs
	48.332	3419.4324	YKTPPIKDFGGFNFS Man8	Obs
	48.339	3323.4379	YKTPPIKDFGGFNFS G1F	Obs
	48.408	3257.3796	YKTPPIKDFGGFNFS Man7	Obs
	48.453	3364.4645	YKTPPIKDFGGFNFS G0F+GlcNAc	Obs
	48.462	3161.3851	YKTPPIKDFGGFNFS G0F	Obs
	48.499	3983.7666	YKTPPIKDFGGFNFS Very complex	Obs
	48.518	3095.3268	YKTPPIKDFGGFNFS Man6	Obs
	48.6	1716.8512	YKTPPIKDFGGFNFS	Inf
	48.6	2123.01	YKTPPIKDFGGFNFS (GlcNAc)2	Inf
	48.6	3015.3272	YKTPPIKDFGGFNFS G0	Inf
	48.6	3177.38	YKTPPIKDFGGFNFS G1	Inf
	48.6	3339.4328	YKTPPIKDFGGFNFS G2	Inf
	48.6	1919.9306	YKTPPIKDFGGFNFS GlcNAc stump	Inf
	48.6	2285.0628	YKTPPIKDFGGFNFS Man1	Inf
	48.6	2447.1156	YKTPPIKDFGGFNFS Man2	Inf
	48.6	2609.1684	YKTPPIKDFGGFNFS Man3	Inf
	48.6	2771.2212	YKTPPIKDFGGFNFS Man4	Inf
	48.6	3581.4852	YKTPPIKDFGGFNFS Man9	Inf
	48.623	3218.4066	YKTPPIKDFGGFNFS G0+GlcNAc	Obs
	48.647	2933.274	YKTPPIKDFGGFNFS Man5	Obs
N343	50.406	3614.5291	NLCPFGEVFNAT Complex	Obs
	51.783	2908.1464	NLCPFGEVFNAT Man7	Obs
	51.942	3136.2575	NLCPFGEVFNAT G2F	Obs
	52.073	2131.0262	NLCPFGEVFNAT glyco	Obs
	52.167	2421.988	NLCPFGEVFNAT Man4	Obs
	52.185	2259.9352	NLCPFGEVFNAT Man3	Obs
	52.229	2666.094	NLCPFGEVFNAT G0	Obs
	52.265	2746.0936	NLCPFGEVFNAT Man6	Obs
	52.3	3281.295	NLCPFGEVFNAT A1	Inf
	52.3	3572.3904	NLCPFGEVFNAT A2	Inf
	52.3	2828.1468	NLCPFGEVFNAT G1	Inf
	52.3	2990.1996	NLCPFGEVFNAT G2	Inf
	52.32	2974.2047	NLCPFGEVFNAT G1F	Obs
	52.445	3015.2313	NLCPFGEVFNAT G0F+GlcNAc	Obs
	52.5	1935.8296	NLCPFGEVFNAT Man1	Inf
	52.5	2097.8824	NLCPFGEVFNAT Man2	Inf
	52.5	3070.1992	NLCPFGEVFNAT Man8	Inf
	52.5	3232.252	NLCPFGEVFNAT Man9	Inf
	52.575	2584.0408	NLCPFGEVFNAT Man5	Obs
	52.632	2812.1519	NLCPFGEVFNAT G0F	Obs
	53.304	3427.3529	NLCPFGEVFNAT A1F	Obs
	53.961	3718.4483	NLCPFGEVFNAT A2F	Obs
GP5	60.483	1861.7463	GP5	Obs



The complete Spike PCDL database is available to download in .cdb or .xlsx format here: https://zenodo.org/record/3958218#.Xxn_BChKhoY

Figure 3. Combined Extracted Ion Chromatogram (EIC) for 27 isoforms of glycopeptide GEVFNAT (N343) within +/- 2 min retention time window from RBD. Only three glycans are labelled, the remainder are listed in the accompanying table 2, below.

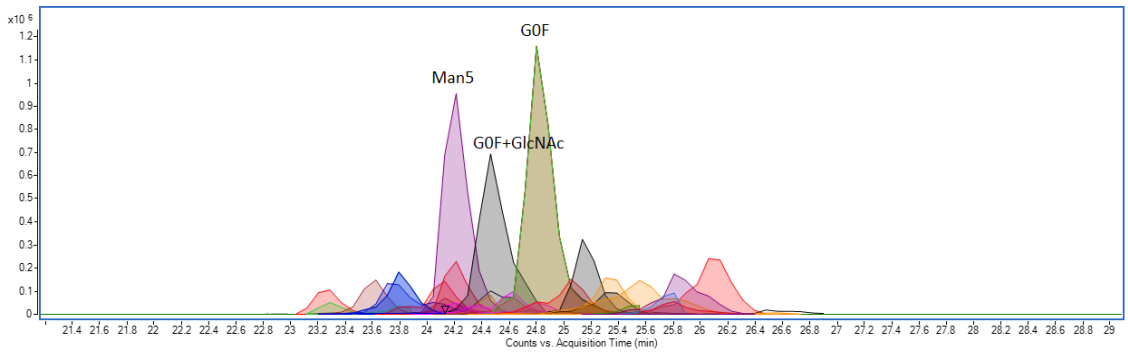


Table 2. GEVFNAT glycopeptide (N343) isoforms from RBD shown in Figure 3

Name	Mass	RT	Volume	ppm error		Name	Mass	RT	Volume	ppm error
GEVFNAT Complex NeuAc (F)2	2838.1078	23.26	733674	-0.2		GEVFNAT G1F	2342.9180	24.62	1090359	3.4
GEVFNAT Man8	2438.9113	23.30	744169	3.7		GEVFNAT G0	2034.8089	24.80	950868	3.1
GEVFNAT Man7	2276.8615	23.61	1867525	2.7		GEVFNAT G0F	2180.8688	24.82	12980259	2.0
GEVFNAT Complex NeuAc F	2692.0523	23.77	1266219	0.5		GEVFNAT A1(F)2-Gal+GlcNAc	2983.1427	24.91	732781	5.3
GEVFNAT Man6	2114.8091	23.80	2432950	2.7		(G)EVFNAT G0	1977.7887	25.05	2947447	2.5
GEVFNAT G1(F)2	2488.9768	24.09	1292274	2.8		GEVFNAT G0F+GlcNAc	2383.9464	25.16	3836871	2.6
GEVFNAT G2F	2504.9755	24.14	1007345	1.3		GEVFNAT A1F	2796.0641	25.33	1512501	3.6
GEVFNAT Man4	1790.7051	24.20	3040952	2.3		GEVFNAT A1(F)2-Gal+GlcNAc	2983.1444	25.34	1320994	4.8
GEVFNAT Man5	1952.7583	24.20	12492713	1.9		GEVFNAT A1F-Gal	2634.0105	25.51	608358	4.1
GEVFNAT Man3	1628.6504	24.20	666739	3.7		GEVFNAT A2F	3087.1629	25.77	805142	2.2
GEVFNAT G1F	2342.9183	24.37	1333546	3.3		GEVFNAT A1F	2796.0644	25.80	541989	3.5
GEVFNAT G0F	2180.8638	24.45	953034	4.3		GEVFNAT A1F-Gal	2634.0125	25.86	3692097	3.4
GEVFNAT G0F+GlcNAc	2383.9436	24.47	9038631	3.7						

Figure 4 illustrates a complete glycan fragmentation series for RBD glycopeptide GEVFNAT-Man5 showing the peptide stump (GEVFNAT-GlcNAc) and mannose ladders. Calculated mass errors are shown in table 3.

Figure 4. Complete glycan fragmentation series for RBD glycopeptide GEVFNAT-Man5 (N343)

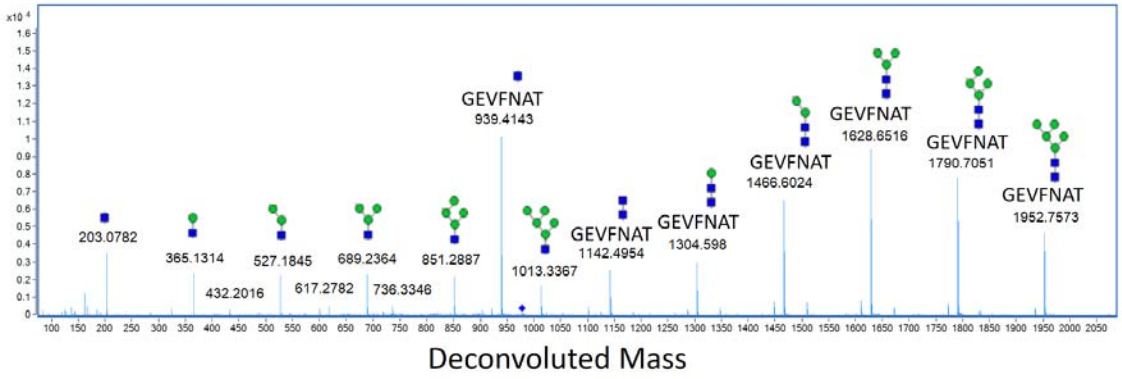


Table 3. Glycan assignment and mass errors (parts per million) for RDB glycopeptide GEVFNAT-Man5

Deconvoluted Mass Obs	Formula	Mass Calc	ppm	Assignment
203.0782	$C_8 H_{13} N O_5$	203.0794	-5.9	GlcNAc
365.1314	$C_8 H_{13} N O_5 (C_6 H_{10} O_5)1$	365.1322	-2.2	GlcNAc(Man)1
527.1845	$C_8 H_{13} N O_5 (C_6 H_{10} O_5)2$	527.1850	-0.9	GlcNAc(Man)2
689.2364	$C_8 H_{13} N O_5 (C_6 H_{10} O_5)3$	689.2364	0.0	GlcNAc(Man)3
851.2887	$C_8 H_{13} N O_5 (C_6 H_{10} O_5)4$	851.2907	-2.3	GlcNAc(Man)4
1013.3367	$C_8 H_{13} N O_5 (C_6 H_{10} O_5)5$	1013.3435	-6.7	GlcNAc(Man)5
939.4143	$C_{32}H_{48}N_8O_{12} (C_8 H_{13} N O_5)$	939.4185	-4.5	GEVFNAT (GlcNAc)
1142.4954	$C_{32}H_{48}N_8O_{12} (C_8 H_{13} N O_5)2$	1142.4979	-2.2	GEVFNAT (GlcNAc)2
1304.5498	$C_{32}H_{48}N_8O_{12} (C_8 H_{13} N O_5)2 (C_6 H_{10} O_5)1$	1304.5507	-0.7	GEVFNAT (GlcNAc)2 (Man)1
1466.6024	$C_{32}H_{48}N_8O_{12} (C_8 H_{13} N O_5)2 (C_6 H_{10} O_5)2$	1466.6036	-0.8	GEVFNAT (GlcNAc)2 (Man)2
1628.6516	$C_{32}H_{48}N_8O_{12} (C_8 H_{13} N O_5)2(C_6 H_{10} O_5)3$	1628.6564	-2.9	GEVFNAT (GlcNAc)2 (Man)3
1790.7051	$C_{32}H_{48}N_8O_{12} (C_8 H_{13} N O_5)2(C_6 H_{10} O_5)4$	1790.7092	-2.3	GEVFNAT (GlcNAc)2 (Man)4
1952.7573	$C_{32}H_{48}N_8O_{12} (C_8 H_{13} N O_5)2(C_6 H_{10} O_5)5$	1952.7620	-2.4	GEVFNAT (GlcNAc)2 (Man)5

In the pseudo MS3 experiment glycans were lost by in-source decay. GEVFNAT-GlcNAc was isolated in the quadrupole and fragmented in the collision cell. Sequence confirmation for the peptide stump GEVFNAT-GlcNAc is shown in Figure 5 with mass errors calculated in Table 4.

Figure 5. Pseudo MS3 fragmentation analysis of RBD glycopeptide stump GEVFNAT-GlcNAc (N343)

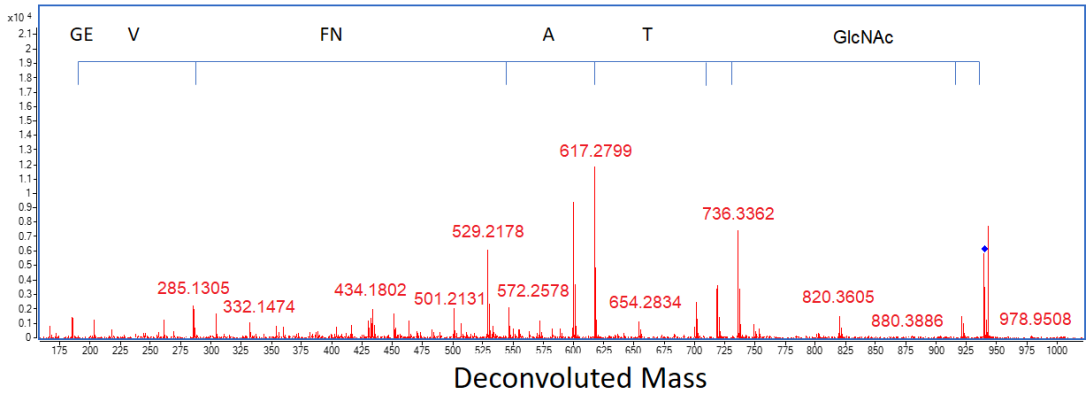
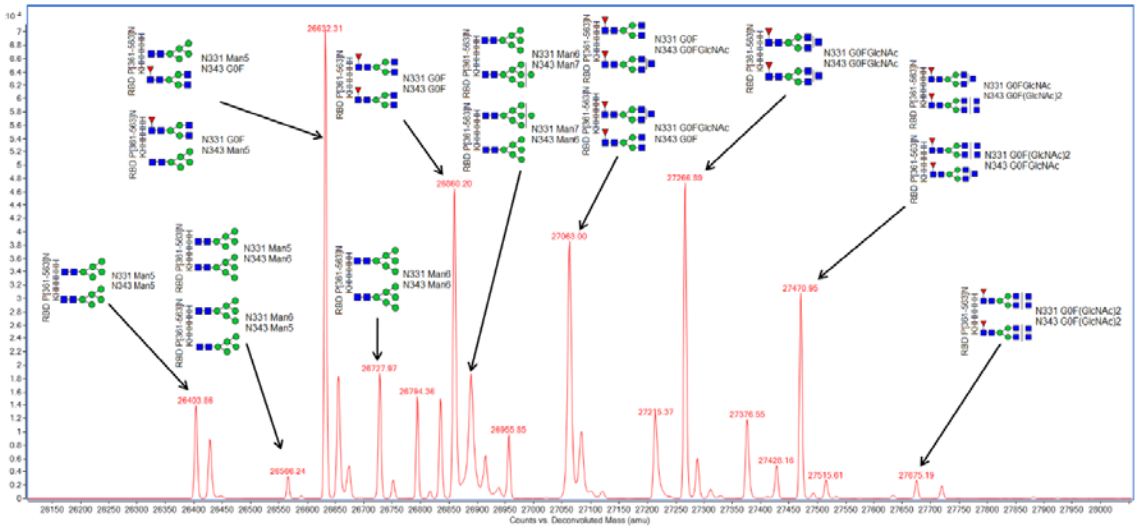


Table 4. Pseudo MS3 fragment ion assignment and mass errors (parts per million) for RBD glycopeptide stump GEVFNAT-GlcNAc (N343)

Deconvoluted Observed Mass	Formula	Calculated Mass	Error (ppm)	Fagment ion assignment	Peptide Sequence
186.0645	$C_7 H_{10} N_2 O_4$	186.0641	2.1	b2	GE
285.1305	$C_{12} H_{19} N_3 O_5$	285.1325	-7.0	b3	GEV
546.2410	$C_{25} H_{34} N_6 O_8$	546.2438	-5.1	b5	GEVFN
617.2799	$C_{28} H_{39} N_7 O_9$	617.2809	-1.6	b6	GEVFNA
718.3257	$C_{32} H_{46} N_8 O_{11}$	718.3286	-4.0	b7	GEVFNAT
736.3362	$C_{32} H_{48} N_8 O_{12}$	736.3392	-4.1	y7	GEVFNAT
939.4174	$C_{32} H_{48} N_8 O_{12} C_8 H_{13} N O_5$	939.4185	-1.2	M GlcNAc	GEVFNAT GlcNAc
921.4064	$C_{32} H_{46} N_8 O_{11} C_8 H_{13} N O_5$	921.408	-1.7	M GlcNAc - H_2O	GEVFNAT GlcNAc
820.3605	$C_{28} H_{39} N_7 O_9 C_8 H_{13} N O_5$	820.3603	0.2	b6 GlcNAc - H_2O	GEVFNA GlcNAc
749.3213	$C_{25} H_{34} N_6 O_8 C_8 H_{13} N O_5$	749.3232	-2.5	b5 GlcNAc - H_2O	GEVFN GlcNAc
701.3021	$C_{32} H_{43} N_7 O_{11}$	701.3021	0.0	b7 - NH_3	GEVFNAT
600.2529	$C_{28} H_{36} N_6 O_9$	600.2544	-2.5	b6 - NH_3	GEVFNA
529.2178	$C_{25} H_{31} N_5 O_8$	529.2173	0.9	b5 - NH_3	GEVFN

Intact mass measurement of fully glycosylated Spike was unsuccessful due to the polydispersity of its innumerable glycoforms and the resulting dilution of ion signal. However, the smaller receptor binding domain, bearing only two glycosylation sites did prove amenable to intact mass analysis. Figure 6 shows twenty-one glycoforms for intact RBD, of which ten major glycoforms could be assigned. This showed that the principal glycan species were Man5, G0F and G0F+GlcNAc which was in agreement with the glycopeptide analysis.

Figure 6. Intact mass analysis of RBD showing the principal glycan species Man5, G0F and G0F+GlcNAc in agreement with glycopeptide analysis. (This method cannot differentiate individual glycosylation sites, hence when two structures are possible, both are shown)



Elastase was chosen as a single digestion enzyme because it was judged to give the best chance of generating glycopeptides with a single NXS/T motif, essential for unambiguous glycan mapping. For non-glycosylated Spike peptides, elastase generated 63 high quality MSMS hits and 26% coverage allowing for five missed cleavages. The same data searched for non-specific cleavage gave 135 high quality MSMS hits and 48% coverage allowing for twenty missed cleavages. Elastase itself contains 2 NXS/T motifs. We therefore prepared elastase only, at x10 the usual concentration, searched the resulting LC-MS data using the PCDL as a control, and no hits were found. The Spike protein LC-MS data did contain a small number of elastase autodigestion peptides.

Methods

Cloning, expression and purification of Spike

The gene encoding amino acids 1-1208 of the SARS-CoV-2 Spike glycoprotein ectodomain (S), with mutations of RRAR > GSAS at residues 682-685 (to remove the furin cleavage site) and KV > PP at residues 986-987 (to stabilise the protein), was synthesised with a C-terminal T4 fibritin trimerization domain, HRV 3C cleavage site, 8xHis tag, and Twin-Strep-tag [5]. The construct was sub-cloned into pHL-sec [10] using the AgeI and XhoI restriction sites and the sequence was confirmed by sequencing. Recombinant Spike was produced in *Expi293F*^{TM}cells by transient transfection with purified DNA (0.5 mg/L cells) using a 1:6 DNA:L-PEI ratio, mixed in minimal medium, and sodium butyrate as an additive. Cells were grown in suspension in *FreeStyle293*^{TM} medium with shaking at 150 rpm in 2 L smooth roller bottles, filled with 0.5 L cells at 2 e^6/mL per bottle at 30°C with 8% CO_2 and 75% humidity. Supernatants from transfected cells were harvested 3-days post-transfection by centrifugation. Clarified supernatant was mixed with Ni^{2+} IMAC *Sepharose® 6 Fast Flow* (*GE*; 2 mL bed volume per L of supernatant) at room temperature for 2 h. Using a gravity flow column, resin

was collected and washed stringently with 50 CV each of base buffer (1X PBS), WB25 (BB + 25 mM imidazole), and WB40 (BB+ 40 mM imidazole), followed by elution with EB (0.30 M imidazole in 1X PBS). Protein was dialyzed into 1X PBS using *SnakeSkin™* 3,500 MWCO dialysis tubing, concentrated to 1 mg/mL using a 100,000 MWCO *VivaSpin* centrifugal concentrator (*GE*), and centrifuged at 21,000 x *g* for 30 min to remove aggregates. The trimeric Spike protein was flash frozen in LN_2 and stored at -80°C until use. Final purified yield was 1 mg of Spike protein per L of transfected cells.

Cloning, expression and purification of Receptor Binding Domain

The receptor binding domain (RBD; aa 330-532) of SARS-CoV-2 Spike (Genbank MN908947) was inserted into the pOPINTTGneo expression vector fused to an N-terminal signal peptide and a C-terminal 6xHis tag [11]. RBD was produced by transient transfection in *Expi293F™* cells (*ThermoFisher Scientific*, UK) using purified DNA (1.0 mg/L cells), a 1:3 DNA:L-PEI ratio, and sodium butyrate as an additive. Cells were grown in suspension in *FreeStyle293™* expression medium at 37°C with 8% CO_2 and 75% humidity. Supernatants from transfected cells were harvested 3-days post-transfection and the supernatant was collected by centrifugation. Clarified supernatant was incubated with 5 mL of Ni^{2+} IMAC *Sepharose® 6 Fast Flow* (*GE*) at room temperature for 2 h. Using gravity flow, resin was washed with 50 CV of base buffer (1X PBS) and 50 CV of WB (1X PBS + 25 mM imidazole) before elution with EB (0.5 M imidazole in 1X PBS). Protein was concentrated using a 10,000 MWCO *Amicon Ultra-15* before application to a *Superdex 75* 16/600 column pre-equilibrated with 1X PBS pH 7.4. Peak monomeric fractions were pooled and concentrated to 2 mg/mL, flash frozen in LN_2, and stored at -80°C until use. Final purified yield was >15 mg RBD per L of transfected cells.

Sample preparation

SARS-CoV-2 Spike or RBD-6H at 1 mg/mL in PBS were prepared in aliquots of either 20 µL or 80 µL and diluted 1 in 3 in 100 mM ammonium bicarbonate, pH 8.0, followed by reduction by addition of 1, 4 Dithiothreitol (DTT) to 5 mM and incubation 37°C for 1 h. Next, the protein was alkylated by addition of iodoacetamide (IAA) to 15 mM and incubation in the dark for 30 min. This was followed by overnight digestion using elastase (*Promega*) at a ratio of 1:20 (w/w). The following day, the supernatant was dried using a rotary evaporator, and re-suspended in 60 µL of 0.1% formic acid for injection into the LC-MS.

'Analytical mode' LC-MS glycopeptide data acquisition

LC-MS 'analytical mode' was performed using a *1290 Infinity* UHPLC coupled to a *G6530A* ESI QTOF mass spectrometer (*Agilent Technologies*). TOF and quadrupole were calibrated prior to analysis and the reference ion 922.0098*m/z* was used for continuous mass correction. Sample was introduced using a 50 µL full-loop injection. Reversed phase chromatographic separation was achieved using an *AdvancedBio Peptide* reversed phase 2.7 µm particle, 2.1 mm x 100 mm column 655750-902 (*Agilent Technologies*). Mobile phase A was 0.1% formic acid in water and mobile phase B 0.1% formic acid in methanol (*Optima* LC-MS grade, *Fisher*). Initial conditions were 5% B and 0.200 mL/min flow rate. A linear gradient from 5% B - 60% B was applied over 60 min, followed by isocratic elution at 100% B for 2 min returning to initial conditions for a further 2 min. Post time was 10 min. MS source parameters were drying gas temperature 350°C, drying gas 8 L/min, nebulizer 30 psi, capillary 4000 V, fragmentor 150 V. MS spectrum range was 100 – 3200 *m/z* (centroid only), 2 GHz Extended Dynamic range, with the instrument in positive ion mode.

LC-MSMS glycopeptide data acquisition 'discovery mode'

LC-MSMS 'discovery mode' was performed as described above, with the following changes: Soft CID collision energy parameters for MSMS were slope 1.0, intercept 0 using argon as the collision gas (if using nitrogen slope 2.0, intercept 0) were used to favour glycan fragmentation over peptide

fragmentation for glycopeptides. Sufficient non-glycosylated peptides were fragmented to give reasonable sequence coverage. Care was taken to reduce sodium and potassium contamination where possible and Tris buffers were avoided as these adducts interfere with glycopeptide analysis.

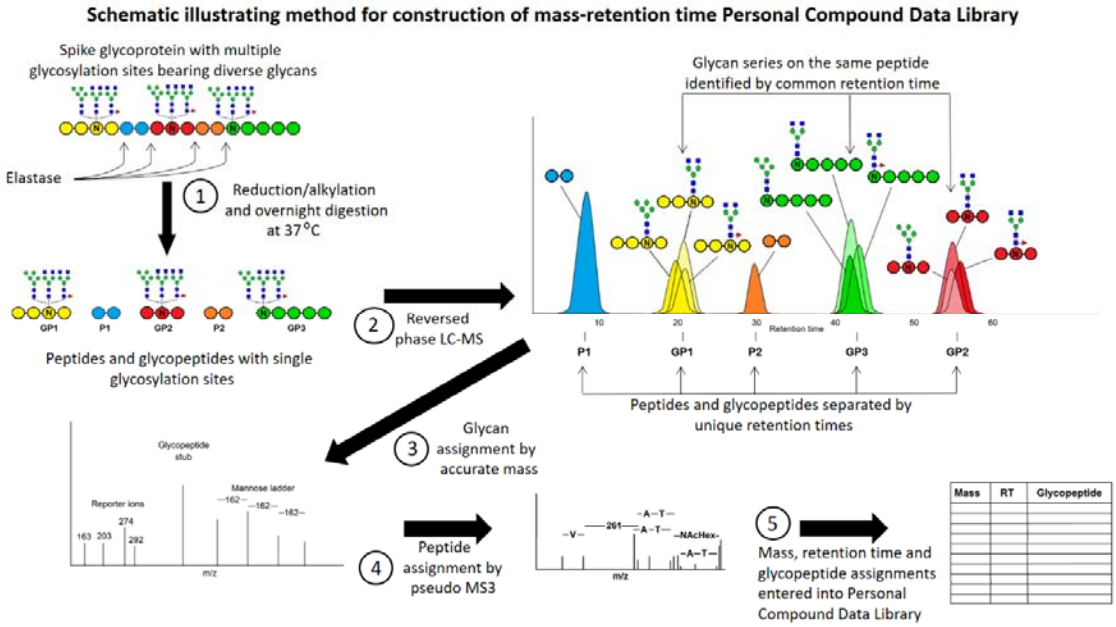
LC-MS glycopeptide data analysis 'analytical mode'

Analysis only required retention time and accurate mass data using the Spike PCDL database created as described below. This is possible either using the *Agilent* software described, software provided by other vendors, or by manual inspection. In our case, we used *Masshunter Qualitative Analysis* version B.07 (*Agilent Technologies*) and the Molecular Feature Extraction tool to extract H+, Na+ and K+ adducts and charge states +1 to +5. Briefly, this tool identifies and associates common spectral features such as carbon isotopes, adducts and multiple charge states as belonging to same Compound (peptide) by virtue of sharing the same accurate mass and retention time, then combines these features together to give a mass, retention time and volume for each compound. Compounds were then searched against Spike PCDL using a mass error window +/- 10 ppm and a retention time window +/- 2 min. Some filtering of the data was used to reduce the number of compounds and thence speed-up the PCDL search. Relative quantitation of each glycan on a particular glycopeptide could then be assessed.

LC-MSMS glycopeptide analysis "discovery mode"

Construction of the Spike glycopeptide mass-retention time database ("discovery mode") was more complex and time-consuming, but once constructed and made available to the scientific community, there is no further need to repeat this step. By using reverse phase HPLC, glycopeptides are separated by the relatively hydrophobic peptide moiety, whereas the associated hydrophilic glycans are grouped together by retention time as illustrated in figure 7.

Figure 7. LC-MSMS "discovery" mode used to generate the Spike glycopeptide Mass-Retention Time PCDL database



Initial LC-MSMS discovery mode data for incorporation into a glycopeptide PCDL was performed using *Masshunter Qualitative Analysis* with *Bioconfirm* B.07.00 (*Agilent Technologies*). Compounds were identified using the Find by Molecular Feature (MFE) tool looking for H+, Na+ and K+ adducts

and charge states +2 to +5. The results were filtered to remove compounds <1000 Da (too small to be glycopeptides). Compound MSMS spectra were screened manually for the following oxonium reporter ions: Hex *m/z* 163.0601, HexNAc *m/z* 204.0866, HexHexNAc *m/z* 366.1395, Neu5Ac *m/z* 274.0921/ *m/z* 291.0949 and/or a Hexose ladder $_{\delta}$M 162.0528 Da. High quality *m/z* spectra were deconvoluted to neutral mass spectra with glycan *de novo* interpretation performed manually. Once a glycopeptide had been identified, it was entered into a personal compound data library database (PCDL, *Agilent Technologies*) as a mass and retention time. In addition, the database made use of known mammalian N-linked glycan processing. After the initial glycopeptide identification, other processed glycopeptides, which were considered likely to also be present, were added to the database at the same retention time and with a calculated mass. For example, if a glycopeptide with Man5 was identified by MSMS, Man1-9 and G0/F were added at the same retention time. If these glycans were subsequently found in the data, their actual retention times were updated, and the next round of processing to more complex glycans was added, in order to produce the most comprehensive PCDL possible, while still being manageable. Processing order:

Man(n) → G0/F → G1/F → G2/F → A1/F → A2/F → Very Complex

Valid glycan identifications resulted in a calculated peptide mass that could be matched to the sequence. Where high quality spectra were present, a peptide-GlcNAc stump was observed (Figure 4). This was used in a pseudo MS3 experiment with manual peptide *de novo* interpretation to confirm the peptide sequence (Figure 5). Mass data adjacent to the glycopeptide retention time was then searched for neutral differences corresponding to glycans, for example, Man5 → G0F or Man7 → G2F has a neutral delta mass of 228.1111 Da.

As expected, not all species could be matched to the sequence, presumably due to unexpected modifications. In this case, they were added to the database as 'GP' with an identifying number and as much information as could be extracted. Data for the most likely glycan was added to the PCDL, including a deconvoluted mass MSMS spectra were available, using nomenclature generating the most easily readable format.

A second round of glycopeptide discovery used *Bioconfirm* v10.0 data analysis software (*Agilent Technologies*). Sequences were matched by peptide accurate mass using the following parameters: peptide cleavage nonspecific, number of missed cleavages 20, N-linked modifications Man3, Man5-9, G0, G0F, G0F GlcNAc, G1, G1F, G2, G2F. Any peptide bearing the glycosylation motif NXS/T with two or more glycan hits within a retention time window +/-2 min was added to the PCDL, excepting missed cysteine alkylations.

In-source fragmentation due to glycopeptide ions absorbing excess energy could be identified in the MS by searching extracted ion chromatograms (EICs) of the oxonium reporter ions and also by related glycopeptides appearing with exactly the same retention times. Both were observed infrequently and at manageable levels.

Intact mass analysis

Concentrated protein samples were diluted to 0.02 mg/mL in 0.1% formic acid and 50 µL was injected on to a 2.1 mm x 12.5 mm *Zorbax* 5 µm *300SB-C3* guard column (*Agilent Technologies*) housed in a column oven set at 40°C. The solvent system used consisted of 0.1% formic acid (solvent A) and 0.1% formic acid in methanol (solvent B). Chromatography was performed as follows: Initial conditions were 90% A and 10% B and a flow rate of 1.0 mL/min. A linear gradient from 10% B to 80% B was applied over 35 seconds. Elution then proceeded isocratically at 95% B for 40 seconds followed by equilibration at initial conditions for a further 15 seconds. The mass spectrometer was configured with the standard ESI source and operated in positive ion mode. The ion source was operated with the capillary voltage at 4000 V, nebulizer pressure at 60 psig, drying gas at 350°C and

drying gas flow rate at 12 L/min. The instrument ion optic voltages were as follows: fragmentor 250 V, skimmer 60 V and octopole RF 250 V.

Discussion

Glycoprotein analysis is difficult. It is either performed in biopharmaceutical laboratories with proprietary expertise of glycan analysis on simple glycoproteins, such as immunoglobulins, or performed by a handful of academic labs with experience of glycan discovery from complex glycoproteins. Many protein researchers choose to ignore it, manipulating cell lines such that they cannot process beyond Man5, or to remove glycans entirely by mutation at the glycosylation motif or enzymatically [12]. While this approach has its merits, it has exposed a serious weakness in analytical capability when faced with a pathogen such as SARS-CoV-2 whose ability to evade the immune system is dependent upon heavy and complex glycosylation.

We have chosen an approach relying on elastase digestion to generate glycopeptides bearing a single glycan but with a sufficient number of amino acid residues to enable chromatographic separation by reversed-phase HPLC, as well as confident identification by accurate mass or *de novo* sequencing. Our choice of reversed phase HPLC has excellent discrimination for short elastase peptides, whereas glycans show little or no interaction with the column. Thus, species originating from a single glycosylation site with the same peptide sequence but several different glycans, eluted with the same retention time and could be discriminated by mass spectrometry. We used reversed phase HPLC and MSMS to characterise as many glycopeptides as possible. Although this required complex and time-consuming data analysis, it needed only be performed once, with the goal of building an accurate mass-retention time database for all observed Spike glycopeptides. Provided the same HPLC column and mobile phase conditions are used, retention times should not vary significantly. Thus, working in the analytical mode we describe, glycan structure and peptide sequence is assigned confidently, by accurate mass and retention time alone. LC-MS data need only to be searched against the mass-retention time database, and peak areas recorded, to generate a complete characterisation of Spike glycans.

We believe the MRTF method described here has advantages over other approaches to Spike glycan analysis. Previous studies relied upon very expensive equipment and software unavailable in most analytical laboratories. Working in 'analytical' mode, all that is necessary is to reproduce the chromatography, hence our method is a generic one, which can be run using any HPLC coupled to any accurate mass instrument and is not restricted to specific proprietary data analysis software. We used PCDL and *Masshunter*, but MRTF analysis can be performed on any vendor software or manually. Moreover, it demands no specialised expertise in glycobiology, and is thus accessible to many more researchers. Some published methods require multiple specific endoproteases, some of which cannot be readily sourced. Our method uses a single enzyme, elastase, which is inexpensive and widely available. Nor does it rely on glycosidases, which may not work efficiently and do not cleave O-linked glycans.

Our data contains an excess of glycopeptides with the motif (y)nNxS/T. This appears to be a very convenient function of elastase on glycopeptides, because the presence of the motif at the C-terminus facilitates *de novo* sequencing. We would be interested to know if this cleavage bias towards the C-terminus of the glycan motif is reproducible in other labs and whether it indicates steric hindrance within the elastase enzyme structure. If such bias is real, then these peptides are less likely to be a false positive result.

Receptor binding domain (RBD) from Spike protein is of interest in many labs for development of serological tests or neutralising antibodies. Because the yield of RBD was five times higher than Spike and more was initially available, we used it for method optimisation, and since it bears only two glycosylation sites which are also present on Spike, it functioned as a useful model. Consequently, N343 on glycopeptide GEVFNAT is over-represented in our demonstration PCDL. We consistently

observed the same three major glycans (Man5, G0F and G0F+GluNAc) on this peptide and these were also in agreement with intact mass analysis of RBD protein as shown in Figure 6. On closer inspection, glycans up to A2F could also be observed at lower levels. We suspect that sufficiently detailed analysis may reveal all possible glycan structures with low abundance at all available sites. The most important would therefore be the top three to five glycans. If the complete complement of Spike protein glycoforms proves too challenging for a single analysis, this site, which is the most complete, would make a good proxy for total glycosylation.

We acknowledge that the mass-retention time fingerprinting method described, like all database searching methods, is dependent on the reproducibility of the enzyme digestion and both the quality and the completeness of database being searched. The example PCDL database reported here is provided as a demonstration. Due to glycan complexity and the likely absence of specific glycans within the Spike batches prepared by us, it will always be incomplete. Moreover, individual glycopeptides were identified with variable degrees of certainty, and we recommend that they should be validated by the user. As with all glycan analysis methods, there is a bias towards glycopeptides that are easiest to identify by the techniques used, and such bias will also be reflected within the database. Once the PDCL has been created, it must be refined and extended over time to improve data quality, and it is our intention to do so.

References

1. Zhang, Y., et al., *Site-specific N-glycosylation Characterization of Recombinant SARS-CoV-2 Spike Proteins using High-Resolution Mass Spectrometry.* bioRxiv, 2020.
2. Solá, R.J. and K. Griebenow, *Glycosylation of therapeutic proteins.* BioDrugs, 2010. **24**(1): p. 9-21.
3. Vugmeyster, Y., et al., *Pharmacokinetics and toxicology of therapeutic proteins: advances and challenges.* World journal of biological chemistry, 2012. **3**(4): p. 73.
4. Tortorici, M.A. and D. Veesler, *Structural insights into coronavirus entry.* Advances in virus research, 2019. **105**: p. 93-116.
5. Wrapp, D., et al., *Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.* Science, 2020. **367**(6483): p. 1260-1263.
6. Yang, T.-J., et al., *Cryo-EM analysis of a feline coronavirus spike protein reveals a unique structure and camouflaging glycans.* Proceedings of the National Academy of Sciences, 2020.
7. Shajahan, A., et al., *Deducing the N-and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2*. bioRxiv, 2020.
8. Watanabe, Y., et al., *Site-specific analysis of the SARS-CoV-2 glycan shield.* BioRxiv, 2020.
9. Yao, H., et al., *Molecular architecture of the SARS-CoV-2 virus.* bioRxiv, 2020: p. 2020.07.08.192104.
10. Aricescu, A.R., W. Lu, and E.Y. Jones, *A time-and cost-efficient system for high-level protein production in mammalian cells.* Acta Crystallographica Section D: Biological Crystallography, 2006. **62**(10): p. 1243-1250.
11. Nettleship, J.E., et al., *Transient expression in HEK 293 cells: an alternative to E. coli for the production of secreted and intracellular mammalian proteins*, in *Insoluble proteins*. 2015, Springer. p. 209-222.
12. Chang, V.T., et al., *Glycoprotein structural genomics: solving the glycosylation problem.* Structure, 2007. **15**(3): p. 267-273.

Acknowledgements

The authors wish to thank Professor David Harvey for critical reading of the manuscript and for helpful comments. We thank Professor Ray Owens for kindly providing the RBD-6H construct and Professor Gavin Screaton and Dr. Juthathip Mongkolsapaya for kindly providing the Spike construct. AdvancedBio Peptide HPLC column was a gift from Agilent Technologies. This project has received

funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 875510. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and Ontario Institute for Cancer Research, Royal Institution for the Advancement of Learning McGill University, Kungliga Tekniska Hoegskolan, Diamond Light Source Limited. The SGC is a registered charity (number 1097737) that receives funds from AbbVie, Bayer Pharma AG, Boehringer Ingelheim, Canada Foundation for Innovation, Eshelman Institute for Innovation, Genentech, Janssen, Merck KGaA, Darmstadt, Germany, MSD, Ontario Ministry of Research, Innovation and Science (MRIS), Pfizer, São Paulo Research Foundation-FAPESP, Takeda, and Wellcome [106169/ZZ14/Z].