Combinatorial approach with mass spectrometry and lectin microarray dissected glycoproteomic features of virion-derived spike protein of SARS-CoV-2

Takahiro Hiono [ORCID: 0000-0001-8897-3592]¹, Azusa Tomioka¹, Hiroyuki Kaji [ORCID: 0000-0002-0910-9999]¹, Michihito Sasaki², Yasuko Orba², Hirofumi Sawa², Atsushi Kuno [ORCID: 0000-0002-6147-6171]^{1*}

¹Molecular and Cellular Glycoproteomics Research Group, Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science & Technology, Tsukuba, Ibaraki 305-8565, Japan

²Division of Molecular Pathobiology, International Institute for Zoonosis Control, Hokkaido University, Sapporo, Hokkaido 001-0020, Japan

*To whom correspondence should be addressed:

Molecular and Cellular Glycoproteomics Research Group, Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science & Technology, Tsukuba, Ibaraki 305-8565, Japan

16 **Abstract**

17 The COVID-19 pandemic caused by the novel coronavirus, SARS-CoV-2, has a global impact on public health. Since
 18 glycosylation of the viral envelope glycoproteins is known to be deeply associated with their immunogenicity, intensive
 19 studies on the glycans of its major glycoprotein, S protein, have been conducted. Nevertheless, the detailed site-specific
 20 glycan compositions of virion-associated S protein have not yet been clarified. Here, we conducted intensive
 21 glycoproteomic analyses of SARS-CoV-2 S protein using a combinatorial approach with two different technologies: mass
 22 spectrometry (MS) and lectin microarray. Using our unique MS1-based glycoproteomic technique, Glyco-RIDGE, in
 23 addition to MS2-based Byonic search, we identified 1,759 site-specific glycan compositions. The most frequent was
 24 HexNAc:Hex:Fuc:NeuAc:NeuGc = 6:6:1:0:0, suggesting a tri-antennary *N*-glycan terminating with LacNAc and having
 25 bisecting GlcNAc and a core fucose, which was found in 20 of 22 glycosylated sites. The subsequent lectin microarray
 26 analysis emphasized intensive outer arm fucosylation of glycans, which efficiently complemented the glycoproteomic
 27 features. The present results illustrate the high-resolution glycoproteomic features of SARS-CoV-2 S protein and
 28 significantly contribute to vaccine design, as well as the understanding of viral protein synthesis.

29 **Key words:** SARS-CoV-2, COVID-19, spike protein, glycosylation, glycoproteomics, lectin microarray, mass
 30 spectrometry

31

Introduction

The COVID-19 pandemic caused by the novel coronavirus, SARS-CoV-2, has a global impact on public health¹. As a countermeasure against the COVID-19 pandemic, great efforts have been made to establish vaccines targeting the spike (S) protein of SARS-CoV-2². The coronavirus S protein is a glycoprotein required for virus attachment to host cells and is mainly targeted by the host immune system. Since glycosylation of the viral envelope glycoproteins is known to be strongly associated with their immunogenicity³, intensive studies on the glycans of SARS-CoV-2 S protein have been conducted⁴. Watanabe et al. first demonstrated a site-specific glycoform for all 22 *N*-glycosylation sites of the S protein⁵. Moreover, Shajahan et al. identified two *O*-glycosylation sites in the receptor-binding domain of the S protein⁶. Based on the revealed glycoform, the molecular dynamics of the glycosylated S proteins were simulated *in silico*⁷. The results demonstrated that two *N*-glycans at N165 and N234 modulated the conformational dynamics of the receptor-binding domain of S proteins. These results were demonstrated with recombinant S protein using a mammalian cell-protein expression system. Subsequently, Yao et al. conducted site-specific glycomic analyses using the S protein derived from intact virions, although detailed glycan structures were not provided⁸.

To date, analyses of the glycoforms of viral proteins have been mainly conducted using mass spectrometry (MS). MS-based approaches have been previously employed for glycomic and glycoproteomic analyses of viruses in many different taxonomic groups, including human immunodeficiency virus^{9,10}, influenza virus^{11–13}, and SARS coronavirus¹⁴. MS-based physical analyses are beneficial in that they provide detailed site-specific glycan compositions. We previously established a liquid chromatography (LC)/MS-based glycoproteomic approach^{15,16}. In general, analyses of site-specific glycoforms of glycopeptides require tandem MS-based approaches. Tandem MS analyses of glycopeptides often showed

low sensitivity because of the difficulty of obtaining fragment ions to identify peptide sequences, in addition to the poor ionization efficiency of glycopeptides. On the other hand, our approach enables MS1-based site-specific glycoproteomic analyses using accurate masses and LC-retention times of glycopeptides^{15,17}. This method, called Glyco-RIDGE, is MS2-independent and therefore has an advantage in sensitivity. Using this unique approach, we successfully identified lists of core peptides for poly-*N*-acetyllactosamine using human promyelocytic leukemia HL60 cells¹⁷ and those for Lewis x antigens using mouse kidneys¹⁵.

Lectin microarray (LMA)-based glycan profiling is an alternative biochemical method for analyzing glycan attached to viruses without the liberation of the glycan moiety¹⁸. Our previous studies on glycan profiling of surface glycoproteins of hepatitis B and influenza A viruses demonstrated that this method provides highly sensitive and simple platforms for the structural interpretation of viral glycans^{19,20}. As an interaction-based assay, the LMA provides information not only on the molecular structures of glycans but also on the accessibility of endogenous glycan-binding molecules to them²¹. However, this method does not provide detailed glycan composition or site-specific information on glycans, that is, glycan micro-heterogeneity in a glycoprotein. Accordingly, the combinatorial use of MS- and LMA-based approaches in the glycoproteomic analysis of the viral glycoprotein should compensate for each method's limitations to synergistically dissect the detailed glycoforms of viral glycoproteins. In this study, we analyzed the glycoform of S proteins derived from intact SARS-CoV-2 viruses by combining MS1-based glycoproteomic analysis and LMA. These two approaches cooperatively illustrated the composition, structure, and conformation of the glycan at high resolution, which will enhance our in-depth exploration of the real world, i.e., “meta-heterogeneity,” of SARS-CoV-2 glyco-particles²².

70 **Materials and methods**

71 *Viruses and cells*

72 The SARS-CoV-2, 2019-nCoV/Japan/TY/WK-521/2020 strain was provided by the National Institute of
73 Infectious Diseases. VeroE6/TMPRSS2 cells were established in our laboratory as described previously²³. The cells were
74 maintained in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum. The virus was inoculated
75 onto the cells, and the infectious cell culture supernatant was collected and cleared by low-speed centrifugation. For the
76 preparation of the sample subjected to MS analyses, virus particles in the supernatant were concentrated through a 20%
77 sucrose cushion at 133,900× g for 2 h, washed with PBS, and resuspended¹⁴. For the samples subjected to LMA analyses,
78 viruses were inactivated with 0.1% β-propiolactone at 4 °C for 1 d.

79

80 *Antibodies*

81 Rabbit monoclonal anti-SARS-CoV-2 spike S1 antibody, 007 (Cat# 40150-R007), and mouse monoclonal anti-
82 SARS-CoV-2 spike neutralizing antibody, 43 (Cat# 40591-V05H1), were purchased from Sino Biological (Beijing, China).
83 These antibodies were biotinylated using the Biotin Labeling Kit-NH2 (Dojindo, Kumamoto, Japan).

84

85 *Site-specific glycan analysis by LC-MS*

86 SARS-CoV-2 particles collected from the culture medium of Vero/TMPRSS2 cells by ultracentrifugation were
87 precipitated with acetone (final concentration of 75%) at -80 °C overnight. The precipitate was recovered and dissolved in
88 100 mM Tris-HCl (pH9.0) containing 12 mM sodium deoxycholate and 12 mM sodium lauroylsarcosinate. Proteins were

S-reduced with dithiothreitol and alkylated with iodoacetamide. After 5 x dilution, the proteins were digested with chymotrypsin, trypsin + Lys-C endopeptidase, or α -lytic protease. An aliquot of each digest was analyzed using an LC/MS system with a nanoflow LC, Ultimate 3000 (Thermo Fisher Scientific, Waltham, MA, USA), and Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific). Another aliquot of each digest was subjected to hydrophilic interaction chromatography (HILIC) on an Amide-80 column (TOSOH, Tokyo, Japan) to collect glycopeptides. An aliquot of the glycopeptide fraction was treated with peptide-*N*-glycanase F (PNGaseF, Takara, Shiga, Japan) in H₂¹⁸O to remove *N*-glycans and to label glycosylated Asn with ¹⁸O as Asp (¹⁸O). The labeled peptides were identified by LC/MS analysis followed by a database search using Mascot (Matrix Science, London, UK)^{24,25}. Another aliquot of glycopeptides was analyzed using LC/MS and the acquired MS2 spectra were analyzed by database search using Byonic. Simultaneously, the MS1 spectra were analyzed using the Glyco-RIDGE method as described previously^{15,17}.

Three-dimensional structural-based mapping of S protein

N-glycosylation sites were mapped on the S protein in a closed conformation (PDB ID: 6ZGE)²⁶ using Discovery Studio Visualizer (BIOVIA, San Diego, CA, USA). Each monomer was depicted as a soft surface model and colored in light gray, light pink, or light blue.

Immunoprecipitation

One hundred microliters of β -propiolactone-inactivated infectious cell culture supernatants were mixed with 200 ng biotinylated 007 antibody. After incubation at 37 °C for 1 h, the complexes were captured with Dynabeads MyOne

Streptavidin T1 (Thermo Fisher Scientific) at 4 °C for 1 h. After washing, virus antigens were eluted from the magnetic beads and disrupted using 20 µL of 100 mM citric acid containing 1% Triton X-100. Biotinylated antibodies incorporated into the eluted samples were depleted with Dynabeads MyOne Streptavidin T1 at 4 °C for 1 h. Antibody-depleted samples were then pH adjusted with 5 µL of Tris-HCl buffer (pH 9.0). For two-step immunoprecipitation, antibody-depleted samples were neutralized with 5 µL of HEPES-NaOH buffer (pH 8.5). The 25 µL sample was then fluorescently labeled with 10 µg Cy3-succinimidyl-ester (GE Healthcare, Buckinghamshire, UK) for 1 h at room temperature in the dark. After adjusting the volume to 125 µL with probing buffer (25 mM Tris-HCl (pH 7.5), containing 137 mM NaCl, 2.7 mM KCl, 500 mM glycine, 1 mM CaCl₂, 1 mM MgCl₂, and 1% Triton X-100), the sample solutions were incubated for 2 h at room temperature to quench the excess fluorescent label. The sample solutions were mixed with 200 ng biotinylated 43 antibody and incubated at 4 °C overnight. The complexes were captured with Dynabeads MyOne Streptavidin T1 at 4 °C for 1 h. After washing, virus antigens were eluted from the magnetic beads and disrupted using 20 µL of 100 mM citric acid containing 1% Triton X-100. Biotinylated antibodies incorporated into the eluted samples were depleted with Dynabeads MyOne Streptavidin T1 at 4 °C for 1 h. Antibody-depleted samples were then pH adjusted with 5 µL of Tris-HCl buffer (pH 9.0). PBS was used as the negative control instead of the virus solution.

123 *LMA*

124 An antibody-overlay LMA was performed as described previously, with some modifications^{20,27}. Lectins printed
125 on the array chips are listed in Supplementary Table S1. The glass slides were printed with 45 lectins with three replicates
126 (LecChip Ver. 1.0; GlycoTechnica, Yokohama, Japan) were activated with probing buffer. Immunoprecipitated samples

diluted with probing buffer were then applied to glass slides and incubated overnight at 20 °C. Subsequently, the slides were blocked with 0.3 mg/mL human IgG for 30 min at 20 °C. After removal of unbound samples and IgG by subsequent washing, 6.7 µg/mL biotinylated monoclonal antibody 43 was added to the slides and incubated for 1 h at 20 °C. Slides were washed three times and incubated with 1.7 µg/mL Streptavidin, Alexa Fluor 555 conjugate (Thermo Fisher Scientific) for 30 min at 20 °C. After washing, the slides were scanned with an evanescent-field excitation fluorescence imager (GlycoStation Reader 1200; GlycoTechnica). The obtained images were quantified using the GlycoStation™ ToolsPro Suite ver. 1.5 (GlycoTechnica). The net intensity for each lectin was calculated as the mean value of three spots minus the background. For direct labeling methods, sample solutions were added to the array glass slide and incubated overnight at 20 °C. After washing three times, the glass slides were scanned using an evanescent-field excitation fluorescence imager.

Exoglycosidase digestion

Fluorescently labeled S protein was prepared as described above using two-step immunoprecipitation. The labeled proteins were subjected to digestion with Sialidase A (Agilent Technologies, Santa Clara, CA, USA), β1-4 galactosidase (Agilent Technologies), or double digestion with Sialidase A and β-galactosidase, according to the manufacturer's instructions. Digestion was conducted at 37 °C for 12 h.

Results

Site-specific glycan composition analysis of SARS-CoV-2 S protein using the LC/MS-Glyco-RIDGE method

Protein digests obtained using chymotrypsin, trypsin+ Lys-endopeptidase, or α-Lytic protease were analyzed by

LC/MS, and the data were searched using Mascot to identify the spike protein and confirm peptides with the *N*-glycosylation potential consensus sequence, N-!P-(S|T) (!P = any amino acid except proline). In tryptic and chymotryptic digests, no consensus sequence-containing peptide was identified by Mascot search even considering semi-digestion, suggesting high glycan-occupancy for all 22 potential sites (Supplementary Tables S2–4). Conversely, *N*-glycosylated sites were identified by IGOT-LC/MS analysis of amide-bound glycopeptides. Mascot search results are shown in Supplementary Tables S5–7. In the chymotryptic digest, 20 *N*-glycosylated sites were identified from 22 potential sites (Supplementary Table S5). In addition, 16 and 11 sites were identified in the tryptic digest and α -Lytic protease digest, respectively (Supplementary Tables S6–7). Therefore, all 22 potential sites were identified as previously glycosylated sites.

Next, using LC/MS2 analysis results of amide-bound glycopeptides of each digest, a Byonic search was carried out to identify glycopeptide forms using an MS/MS spectrum-based approach. The identified glycopeptides are listed in Supplementary Table S8. A total of 919 site-specific glycans were identified for all 22 potential sites (Supplementary Table S9). Finally, using the same LC/MS analysis data, Glyco-RIDGE analyses were performed to identify glycopeptide forms (site-specific glycan compositions). All matched results are shown in Supplementary Table S10. A total of 1,449 site-specific glycans were identified (Supplementary Table S11). Combining the Byonic and Glyco-RIDGE results, 1,759 site-specific glycan compositions and 463 glycan compositions were identified (Supplementary Tables S12–13). The frequencies of the compositions that emerged at the 22 sites are shown in Supplementary Table S14. Among them, the top 21 compositions presented over 15 sites are listed in Table 1. The most frequently found composition was HexNAc:Hex:Fuc:NeuAc:NeuGc = 6:6:1:0:0, which appeared at 20 out of 22 glycosylated sites. The most plausible structure predicted from this composition is a tri-antennary glycan having a bisecting GlcNAc and a fucose, probably a

core fucose. Please note that, in the following text, Figures, and Tables, “N:N:N:N” means the actual glycan composition of HexNAc:Hex:Fuc:NeuAc:NeuGc. To present glycan compositions indicating their structure intuitively, we used the number of saccharides on the trimannosyl core (Man(3)GlcNAc(2)) and showed the number in the order of Hex-HexNAc-Fuc-NeuAc-NeuGc, e.g., glycan id 20000 for 2:5:0:0:0, or 22100 for 4:5:1:0:0. The top 21 most frequent compositions contained three high-mannose compositions, (N0000), and complex type compositions, including one or more fucoses and no sialic acid (except one, 32010). The emerging frequency of glycan compositions shows that glycans on the S protein were relatively well processed, i.e., highly branched and fucosylated, moderately sialylated, and low extension (polyLacNAc). Fig. 1 lists the number and rate of compositions at each site. Except for three C-terminal sites, high-mannose compositions were found at each site. To estimate the abundance of the high-mannose compositions, extracted ion chromatograms of the high-mannose composition (20000, 30000, 40000) and the top three compositions of each site were obtained for the common core peptide (data not shown). According to the results, the content of high-mannose composition was classified into three categories: rare (or not detected), moderate, and abundant (Fig. 1, Supplementary Fig. S1). Eight sites were found to contain abundant high-mannose glycans: Asn-61, 122, 234, 603, 709, 717, 801, and 1074. These sites also contained a relatively higher content of hybrid compositions (#HexNAc=1, over 15%). Conversely, other sites have a tendency to show higher rates of branching, fucosylation, and sialylation.

Frequency of glycan compositions

Glycan modification susceptibility at each site was estimated by counting the compositions containing the focused motif or their rate. For example, when the fucosylation rates between sites X and Y are compared, the ratio of the number

of fucosylated compositions for total number of compositions was compared. On the other hand, since high-mannose type glycans are limited in number, i.e., there are only five compositions for high-mannose glycans (20000-60000), we compared the rate of high-mannose type glycans, by the count of the high-mannose type compositions and the relative intensities of major high-mannose glycans to those of the top three glycopeptides of each site using extracted ion chromatography of the same core glycopeptides. For Asn-61 (NVT), peak intensities of 20000, 30000, and 40000 of glycopeptide positions 59–65 were compared with those of 22200, 21100, and 31100 of the same core glycopeptides. Because intensity of 30000 was significantly high compared to the others, the high-mannose rate of the site was classified as “abundant.” Conversely, for Asn-74 (NGT), the intensities of 20000, 30000, 40000, 23300, 13300, and 22300 glycopeptides (core positions 66–79) were compared. Slight peaks were observed for high-mannose compositions; thus, the rate is “rare (or not detected).” By comparison, eight sites were found to be high-mannose abundant, 61, 122, 234, 603, 709, 717, 801, and 1074 (Fig. 1).

Glycan stem distribution

Next, we moved our attention to the glycan modification pathway from high-mannose to highly modified complex-type glycans in Vero/TMPRSS2 cells. Glycan modifications can be divided into two categories: extension or excision of the glycan stem, such as the addition of GlcNAc, Gal, GalNAc, and end-capping (leaves) with fucose and sialic acid, where polysialylation is not considered. From the identified compositions, it is clear that fucosylation actively occurs in Vero/TMPRSS2 cells; therefore, the formation of a glycan stem will be illustrated on #Hex-#HexNAc matrix considering the biosynthetic pathway. The *N*-glycan precursor is attached to the Asn side-chain en bloc as Glc(3)Man(9)GlcNAc(2), so

the starting point is Hex(12)HexNAc(2) = 90000 (9-0 in Fig 2A). After successive removal of Glc3, mannoses were excised from the high-mannose glycans to form 20000 (2-0 in Fig 2A). Next, extension of glycan started with the addition of GlcNAc to form hybrid glycan 21000 (2-1 in Fig 2A), followed by Gal addition to 31000 (3-1 in Fig 2A). After removing 2 Man from glycan 3-1, successive addition of GlcNAc and Gal forms bi-antennary glycan 22000 (2-2 in Fig 2A). Further addition of 1 or 2 units of LacNAc (Gal-GlcNAc) formed tri-antennary 33000 (3-3 in Fig 2A) and tetra-antennary 44000 (4-4 in Fig 2A) glycans. Simultaneous addition of HexNAc (such as bisecting GlcNAc or GalNAc to form LacdiNAc) and removal of Hex (Gal or Man) will occur to form complex glycan stems. Fig. 2B shows the frequencies of each composition identified at 22 sites of the S protein. The most frequently observed compositions were 2-1, 2-2, 1-3, 3-3, 3-4, and 4-4, which are presumed to be bi-antennary to tetra-antennary glycans. Glycans 1-3 were presumed to have a LacdiNAc motif. From the frequency mapping, the glycan modification pathway of Vero/TMPRSS2 cells was elucidated. Mass data revealed glycan compositions, and composition mapping provided information on partial glycan motifs such as outer fucosylation, bisecting GlcNAc, LacdiNAc, and less polylactosamine formation. These motifs were suggested by LMA analysis.

215

216 *LMA analyses of SARS-CoV-2 S protein*

Glycoforms of S protein were further analyzed using LMA analysis (Fig 3). First, we tried the “antibody-overlay method,” which we employed in our previous studies (Fig 3A)^{20,27}. Immunoprecipitated S proteins were subjected to LMA glass slides and detected with a biotinylated anti-S monoclonal antibody and Alexa555-conjugated streptavidin. As a result, considerable signals in LacNAc-specific (ECA and RCA120) and mannose-specific lectins (NPA, GNA, and HHL) were obtained (Fig 4A), suggesting that both complex and high-mannose type glycans were displayed on the S protein. Signals

222 in sialic acid-specific lectins (MAL-I, SNA, SSA, TJA-I, MAH) provided only low intensity, suggesting a low abundance
 223 of sialic acid-terminated glycans. Interestingly, the results were not necessarily consistent with those obtained by MS-based
 224 analyses. We then established direct labeling methods coupled with two-step antigen purification with immunoprecipitation
 225 (Fig 3B). For this purpose, we labeled immunoprecipitated viral antigens with Cy3 and further purified S proteins from
 226 fluorescence-labeled proteins by a second immunoprecipitation with another monoclonal antibody. The purified and Cy3-
 227 labeled S protein was then directly applied to LMA glass slides and scanned without overlaying detection antibodies. The
 228 results were clearly different from those obtained with the antibody-overlay LMA analyses. Notably, sialic acid-specific
 229 lectins showed high signals in this analysis (Fig 4B). In order to further investigate the glycan structures, especially for the
 230 outer non-reducing end of glycans, the purified and Cy3-labeled S proteins were digested with exoglycosidases, such as
 231 sialidase A and/or β 1-4 galactosidase, and the digested antigens were subsequently subjected to LMA analyses (black bars
 232 in Fig 5). Signals in sialic acid-specific lectins were partially diminished by digestion with sialidase (Fig 5A). Interestingly,
 233 after digestion with β -galactosidase, signals in both sialic acid- and LacNAc-specific lectins partly decreased (Fig 5B).
 234 Moreover, double digestion with sialidase and β -galactosidase resulted in nearly complete disappearance of the signals in
 235 MAL-I, SNA, SSA, and TJA-I (Fig 5C). These results confirmed that complex-type glycans on S proteins were terminated
 236 with either sialic acid or exposed galactose. It should also be noted that signals in the AOL and AAL increased after β -
 237 galactosidase digestion. This may indicate that some of the S proteins were modified with Lewis-type fucose. Removing
 238 the terminal galactose interacted with fucose, an increase in the mobility of the fucose facilitated enhanced accessibility of
 239 AOL and AAL to the glycan. Along with low signals in UEA-I, which is an H-antigen specific lectin, most of the outer
 240 fucose structure can be assigned as the Lewis type rather than the H type. The simultaneous acquisition of the *O*-linked

glycan information with *N*-glycan suggested the presence of disialyl T-, sialyl T-, and T-antigens indicated by signals in jacalin, ACA, MPA, and MAH. This was confirmed by the enhanced signals of HPA, a Tn-antigen-specific lectin, after β -galactosidase digestion. These *O*-glycan signals derived from the interaction-based assay were relatively weak compared to those of *N*-glycan signals, reflecting the difference in density and accessibility between *N*- and *O*-glycans against the surrounding molecules.

Glycosylation sites in S proteins were evolutionarily conserved among subgenus Sarbecovirus

To explore the biological significance of *N*-linked glycan, putative *N*-glycosylation sites of S proteins were compared among five viruses belonging to the subgenus *Sarbecovirus*. Therefore, the amino acid sequences of S proteins from five SARS-CoV-2 (2019/nCoV/Japan/TY/WK-521/2020, WHU-01, hCoV-19/England/MILK-9E05B3/2020 (VOC-202012/01), hCoV-19/South Africa/KRISP-K004312/2020 (501.V2 variant), hCoV-19/Japan/IC-0561/2021 (Lineage P.1)), two bat-derived coronaviruses (RaTG13, Rc-o319), a pangolin-derived coronavirus (PCoV-GX-P4L), and SARS coronavirus Tor2 strain were aligned (Supplementary Fig. S2). Although slight variations in glycosylation sites were observed in the N-terminal region, the N-!P-(S|T) sequon motifs were well conserved among viruses belonging to the subgenus *Sarbecovirus*. On the other hand, the lineage P.1 strain, hCoV-19/Japan/IC-0561/2021, acquired two additional N-!P-(S|T) sequences during evolution. Since stepwise acquisitions of *N*-glycans lead to antigenic changes in the virus²⁸, any further acquisition of potential glycosylation in the SARS-CoV-2 S protein during evolution should be monitored.

Discussion

In the present study, we utilized two approaches, MS and LMA, for glycoproteomic analyses of the SARS-CoV-2 S protein. MS-based analyses reveal site-specific glycans with accurate glycan composition (Supplementary Table S9, 11–12). This approach is especially useful for differentiating complex or high-mannose types or for analyzing the site occupancy of glycans at each site. In contrast, the glycoforms obtained in the present study and the other study were based on cell line-derived virions⁸. Considering that glycoforms of viral glycoproteins are host cell-dependent, it is essential to dissect glycoforms of human-derived virions to utilize glycomic data to develop vaccines or diagnostic methods²⁰. For this purpose, the current MS-based approach does not have sufficient sensitivity to analyze human-derived virions. LMA is potentially sensitive enough to analyze human-derived virions. In addition, as an interaction-based approach, LMA is powerful for determining both the terminal structures and the accessibility of glycans. The present results demonstrated that the glycan profile obtained with the direct labeling method was consistent with the results obtained with the MS-based analysis (Fig. 4B). This suggests that the detailed glycoform of human-derived virions may be speculated by refining glycoproteomic data of cell line-derived virions obtained with the MS-based approach with the comparative data of glycan profiles of the cell line- and human-derived virions obtained by LMA.

For in-depth analysis, we further assigned the site-specific glycan compositions using two mass spectrometric approaches. One is the MS2-based approach using the Byonic database search engine, and the other is the MS1-based Glyco-RIDGE approach. The results identified by the two MS approaches are different from each other, and are increased in number when combined (Supplementary Table S14); however, it is therefore difficult to estimate the relative quantity between the identified members, especially when they were identified by different approaches. Similarly, when the compositions were identified from different core peptide sequences, the abundance ratio between the members was unclear.

Thus, we compared glycan compositions between sites by counting the compositions containing the focused motif or their rate (Fig. 1).

Our results with MS-based glycoproteomics are in good agreement with the results of Yao et al. for the analysis of intact viral particles and of Watanabe et al. for the recombinant protein^{5,8}. The major reason for the retention of high-mannose glycans at specific sites is thought to be the low accessibility of glycan-modifying enzymes (mannosidases, GlcNAc transferases, sialyl transferases, etc.) to the glycan. The fact that the rate of hybrid-type glycans (#HexNAc = #GlcNAc = 1) for all compositions is higher at the high-mannose-abundant sites supports this observation. In addition, these sites also tend to have low rates of branching (#HN $\geq$ 4) and sialylation, e.g., at Asn-61, 122, 234, 603, 709, and 801. This low accessibility appears to be caused by steric hindrance of the lipid bilayer (ER or Golgi membrane) around the site or by burial of the glycan in the protein cleft or the interface between subunits. In fact, sites 61, 122, and 234 were located in the cleft of the protein of the 3D model of the S protein trimer (Supplementary Fig. S1). In contrast, sites 709, 801, and 1047 were located on the surface of the protein. These sites are located on the bottom surface of the S protein, which is distant from the head domain. On the other hand, the C-terminus Asn-1158, 1173, 1194 were exclusively decorated with complex-type glycans. Since maturation of glycan and core-protein should proceed simultaneously, the higher frequency of sites 709, 801, and 1047 may reflect unknown intracellular events (e.g., host protein interactions or conformational changes) of S protein.

It is important to note that, at these sites, the high-mannose type is major but not exclusive, as indicated in soluble recombinant protein-based analyses by Watanabe et al.⁵. On the eight sites, we also found many branched, highly fucosylated, or sialylated glycan compositions, as well as the results of Yao et al., who analyzed parental Vero cell-derived

viral particles⁸. The reason for this difference is unclear; however, the sensitivity of the detection (depth/coverage of the glycome of each site) and molecular state (viral particle or over-expressed protein) may influence the distribution. In contrast, at the low high-mannose sites, the rate of HexNAc ≥ 3 (suggesting branching, including addition of bisecting GlcNAc) is relatively high, especially at Asn-1158, 1173, and 1194 (Fig. 1). Similarly, Asn-17, 149, 331, 657, 1158, and 1173 showed high fucosylation rates (> 80%). As described above, although there is a bias in the glycan composition for each site, there is no strict exclusion, and many complex-type glycans have been found at all sites. This may be partly because S proteins are not in exactly the same three-dimensional structural environment (symmetry) in the trimer of S proteins.

It is of interest that the glycoform of SARS-CoV-2 S protein was quite different from the previously reported glycoform of the SARS virus S protein¹⁴. Most of the complex-type *N*-glycans attached to the SARS virus S protein were represented in the agalactosyl form, suggesting incomplete maturation. In contrast, complex-type *N*-glycans of the SARS-CoV-2 S protein were terminated with either galactose or sialic acid. The difference in the glycoforms of S proteins of SARS virus and SARS-CoV-2 should arise from either a differential impact on the host glycosylation machinery or a differential route of the S protein synthesis pathway. It should also be noted that a murine coronavirus, mouse hepatitis virus, buds from the intermediate compartment between the endoplasmic reticulum and Golgi complex²⁹. Virion formation of coronavirus during the early to mid-stage of *N*-glycan maturation is closely related to the glycoform of coronavirus S proteins. On the other hand, the Golgi localization of mannosyl (alpha-1,3-)-glycoprotein beta-1,2-*N*-acetylglucosaminyltransferase (MGAT1)³⁰, which is the initiating enzyme for the synthesis of complex-type *N*-glycan, contradicts the fact that coronavirus S proteins are decorated with complex-type *N*-glycans. How the coronavirus S protein

encounters MGAT1 is one of the key questions in understanding their post-translational modifications. Perhaps, analyzing the factors associated with the differential glycoforms of SARS and SARS-CoV-2 S proteins may unveil the widely unknown late-stage of the coronavirus lifecycle.

The glycoprotein of the viral protein is strongly associated with its immunogenicity³. Our previous study demonstrated that virions propagated in different cell lines showed different glycan profiles²⁰. Accordingly, optimizing the glycoform of vaccine antigens by selecting the appropriate host cell line is one of the candidate strategies for developing better vaccines. Our approach, a combination of MS-based and lectin interaction-based glycoproteomics, provided a highly accurate glycan profile of SARS-CoV-2 S protein using a relatively lower amount of viral antigen, resulting in the exploration of meta-heterogeneity, to describe a higher level of glycan regulation: the variation in glycosylation across multiple sites of SARS-CoV-2 S protein. The platform also enables the quick provision of glycoform data of vaccine antigens. In summary, our new concept for glycoproteomic analyses of viral proteins should significantly contribute to establishing effective countermeasures against COVID-19, as well as future viral pandemics.

Acknowledgements

We thank Ms. Misugi Nagai, Ms. Kaori Ohki, and Ms. Masako Sukegawa for their technical help. We thank the National Institute of Infectious Diseases for providing SARS-CoV-2, 2019-nCoV/Japan/TY/WK-521/2020 strain. We would like to thank all researchers who kindly deposited and shared genomic data on GISAID.

Declarations

336 *Funding*

337 The present work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI
338 Grant Number 18K15176 to TH, and the Japan Agency for Medical Research and Development (AMED) Grant Numbers
339 16809263 and 20he0522002j0001. The funders had no role in the study design, data collection and analysis, decision to
340 publish, or preparation of the manuscript.

341

342 *Conflicts of interest/Competing interests*

343 The authors declare that there are no conflicts of interest.

344

345 *Ethics approval*

346 The present work does not contain any studies with human participants or animals performed by any of the authors.

347

348 *Authors' contributions*

349 Conceptualization: TH, HK, and AK; methodology: TH and HK; formal analysis: TH, AT, and HK; investigation:
350 TH, HK, MS, and YO; resources: MS, YO, and HS; data curation: TH, AT, HK, and AK; writing original draft: TH and
351 HK; writing review and editing: TH, HK, MS, YO, HS, and AK; visualization: TH and HK; supervision: HK, HS, and AK;
352 project administration: HS and AK; funding acquisition: TH, HK, and HS.

353

354 **References**

1. WHO Coronavirus (COVID-19) Dashboard. <https://covid19.who.int>.
2. Koff, W. C. *et al.* Development and deployment of COVID-19 vaccines for those most vulnerable. *Sci. Transl. Med.* **13**, eabd1525 (2021).
3. de Vries, R. P. *et al.* Glycan-Dependent Immunogenicity of Recombinant Soluble Trimeric Hemagglutinin. *J. Virol.* **86**, 11735–11744 (2012).
4. Lardone, R. D. *et al.* How glycobiology can help us treat and beat the COVID-19 pandemic. *J. Biol. Chem.* 100375 (2021) doi:<https://doi.org/10.1016/j.jbc.2021.100375>.
5. Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S. & Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **369**, 330–333 (2020).
6. Shajahan, A., Supekar, N. T., Gleinich, A. S. & Azadi, P. Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. *Glycobiology* **30**, 981–988 (2020).
7. Casalino, L. *et al.* Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* **6**, 1722–1734 (2020).
8. Yao, H. *et al.* Molecular Architecture of the SARS-CoV-2 Virus. *Cell* **183**, 730–738.e13 (2020).
9. Go, E. P. *et al.* Comparative Analysis of the Glycosylation Profiles of Membrane-Anchored HIV-1 Envelope Glycoprotein Trimers and Soluble gp140. *J. Virol.* **89**, 8245–8257 (2015).
10. Go, E. P. *et al.* Glycosylation Benchmark Profile for HIV-1 Envelope Glycoprotein Production Based on Eleven Env Trimers. *J. Virol.* **91**, e02428-16 (2017).
11. An, Y., McCullers, J. A., Alymova, I., Parsons, L. M. & Cipollo, J. F. Glycosylation analysis of engineered H3N2 influenza A virus hemagglutinins with sequentially added historically relevant glycosylation sites. *J. Proteome Res.* **14**, 3957–3969 (2015).
12. Khatri, K. *et al.* Integrated Omics and Computational Glycobiology Reveal Structural Basis for Influenza A Virus Glycan Microheterogeneity and Host Interactions. *Mol. Cell. Proteomics* **15**, 1895–1912 (2016).
13. She, Y.-M., Li, X. & Cyr, T. D. Remarkable Structural Diversity of N-Glycan Sulfation on Influenza Vaccines. *Anal. Chem.* **91**, 5083–5090 (2019).
14. Ritchie, G. *et al.* Identification of N-linked carbohydrates from severe acute respiratory syndrome (SARS) spike glycoprotein. *Virology* **399**, 257–269 (2010).
15. Noro, E. *et al.* Large-Scale Identification of N-Glycan Glycoproteins Carrying Lewis x and Site-Specific N-Glycan Alterations in Fut9 Knockout Mice. *J. Proteome Res.* **14**, 3823–3834 (2015).
16. Togayachi, A. *et al.* Polylactosamine on glycoproteins influences basal levels of lymphocyte and macrophage activation. *Proc. Natl. Acad. Sci.* **104**, 15829–15834 (2007).
17. Togayachi, A. *et al.* Identification of Poly-N-Acetylactosamine-Carrying Glycoproteins from HL-60 Human Promyelocytic Leukemia Cells Using a Site-Specific Glycome Analysis Method, Glyco-RIDGE. *J. Am. Soc. Mass Spectrom.* **29**, 1138–1152 (2018).
18. Kuno, A. *et al.* Evanescent-field fluorescence-assisted lectin microarray: a new strategy for glycan profiling. *Nat. Methods* **2**, 851 (2005).
19. Wagatsuma, T. *et al.* Highly Sensitive Glycan Profiling of Hepatitis B Viral Particles and a Simple Method for Dane Particle Enrichment. *Anal. Chem.* **90**, 10196–10203 (2018).

- 393 20. Hiono, T. *et al.* Lectin microarray analyses reveal host cell-specific glycan profiles of the hemagglutinins of
394 influenza A viruses. *Virology* **527**, 132–140 (2019).
- 395 21. Narimatsu, H. *et al.* A strategy for discovery of cancer glyco-biomarkers in serum using newly developed
396 technologies for glycoproteomics. *FEBS J.* **277**, 95–105 (2010).
- 397 22. Čaval, T., Heck, A. J. R. & Reiding, K. R. Meta-heterogeneity: Evaluating and Describing the Diversity in
398 Glycosylation Between Sites on the Same Glycoprotein. *Mol. Cell. Proteomics* **20**, 100010 (2021).
- 399 23. Sasaki, M. *et al.* SARS-CoV-2 variants with mutations at the S1/S2 cleavage site are generated in vitro during
400 propagation in TMPRSS2-deficient cells. *PLOS Pathog.* **17**, e1009233 (2021).
- 401 24. Kaji, H. *et al.* Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked
402 glycoproteins. *Nat. Biotechnol.* **21**, 667–672 (2003).
- 403 25. Kaji, H., Yamauchi, Y., Takahashi, N. & Isobe, T. Mass spectrometric identification of N-linked glycopeptides
404 using lectin-mediated affinity capture and glycosylation site-specific stable isotope tagging. *Nat. Protoc.* **1**, 3019–
405 3027 (2006).
- 406 26. Wrobel, A. G. *et al.* SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and
407 furin-cleavage effects. *Nat. Struct. Mol. Biol.* **27**, 763–767 (2020).
- 408 27. Hiono, T. & Kuno, A. C-Terminally tagged NA in replication-competent influenza A viruses reveals differences in
409 glycan profiles between NA and HA. *Analyst* **145**, 5845–5853 (2020).
- 410 28. Wu, N. C. & Wilson, I. A. A Perspective on the Structural and Functional Constraints for Immune Evasion: Insights
411 from Influenza Virus. *J. Mol. Biol.* **429**, 2694–2709 (2017).
- 412 29. Krijnse-Locker, J., Ericsson, M., Rottier, P. & Griffiths, G. Characterization of the budding compartment of mouse
413 hepatitis virus: evidence that transport from the RER to the Golgi complex requires only one vesicular transport
414 step. *J. Cell Biol.* **124**, 55–70 (1994).
- 415 30. Hassinen, A., Rivinoja, A., Kauppila, A. & Kellokumpu, S. Golgi N-Glycosyltransferases Form Both Homo- and
416 Heterodimeric Enzyme Complexes in Live Cells. *J. Biol. Chem.* **285**, 17771–17777 (2010).

Table 1. Top 21 frequent glycan compositions emerged in 22 potential sites of SARS CoV-2 Spike protein

Ranking	Glycan composition	No. of HexNAc	No. of Hex	No. of Fuc	No. of NeuAc	No. of NeuGc	Frequency / 22 sites	Glycan id
1	6:6:1:0:0	6	6	1	0	0	20	34100
2	4:5:1:0:0	4	5	1	0	0	19	22100
3	5:6:1:0:0	5	6	1	0	0	19	33100
4	6:7:1:0:0	6	7	1	0	0	19	44100
5	4:5:2:0:0	4	5	2	0	0	18	22200
6	4:5:3:0:0	4	5	3	0	0	18	22300
7	5:5:1:0:0	5	5	1	0	0	18	23100
8	5:6:2:0:0	5	6	2	0	0	18	33200
9	4:5:0:0:0	4	5	0	0	0	17	22000
10	5:4:1:0:0	5	4	1	0	0	17	13100
11	5:6:3:0:0	5	6	3	0	0	17	33300
12	6:7:4:0:0	6	7	4	0	0	17	44400
13	2:5:0:0:0	2	5	0	0	0	16	20000
14	3:5:1:0:0	3	5	1	0	0	16	21100
15	6:7:2:0:0	6	7	2	0	0	16	44200
16	6:7:3:0:0	6	7	3	0	0	16	44300
17	2:6:0:0:0	2	6	0	0	0	15	30000
18	2:7:0:0:0	2	7	0	0	0	15	40000
19	3:6:1:0:0	3	6	1	0	0	15	31100
20	4:6:0:1:0	4	6	0	1	0	15	32010
21	5:5:2:0:0	5	5	2	0	0	15	23200

site	No. of composition / site	No. of High-Mannose (HM) type composition (#HexNAc=2)	Rate of HM type (from XIC)	No. of Hybrid type composition (#HexNAc=3)	Rate of Hybrid type	No. of branched type (#HexNAc≥6)	Rate of branched type (HexNAc≥6)	No. of Fucosylated glycan composition (#F≥1)	Rate of Fucosylation	No. of multiply Fucosylated glycans (#Fuc≥2)	Rate of multiple fucosylation (#F≥2 / #F)	No. of Sialylated glycan composition	Rate of Sialylation
17 (NLT)	54	1	Rare	5	0.09	15	0.28	44	0.83	23	0.52	19	0.36
61 (NVT)	54	4	Abundant	8	0.15	15	0.28	36	0.72	20	0.56	11	0.22
74 (NGT)	237	1	Rare	14	0.06	141	0.59	196	0.83	120	0.61	156	0.66
122 (NAT)	102	5	Abundant	21	0.21	22	0.22	63	0.65	28	0.44	46	0.47
149 (NKS)	132	4	Rare	6	0.05	69	0.52	110	0.86	73	0.66	82	0.64
165 (NCT)	50	2	Moderate	5	0.10	20	0.40	38	0.79	20	0.53	16	0.33
234 (NIT)	46	5	Abundant	7	0.15	18	0.39	24	0.59	6	0.25	10	0.24
282 (NGT)	77	3	Rare	11	0.14	26	0.34	50	0.68	20	0.400	29	0.39
331 (NIT)	56	2	Rare	2	0.04	18	0.32	47	0.87	26	0.55	28	0.52
343 (NAT)	190	2	Rare	5	0.03	118	0.62	159	0.85	96	0.60	138	0.73
603 (NTS)	36	4	Abundant	6	0.17	6	0.17	24	0.75	11	0.46	8	0.25
616 (NCT)	57	2	Rare	8	0.14	14	0.25	33	0.60	11	0.33	23	0.42
657 (NNS)	55	3	Rare	6	0.11	23	0.42	43	0.83	25	0.58	25	0.48
709 (NNS)	28	5	Abundant	2	0.07	13	0.46	16	0.70	5	0.31	8	0.35
717 (NFT)	18	5	Abundant	5	0.28	7	0.39	8	0.62	0	0.00	8	0.62
801 (NFS)	92	7	Abundant	17	0.18	12	0.13	56	0.66	21	0.38	37	0.44
1074 (NFT)	169	9	Abundant	24	0.14	59	0.35	115	0.72	59	0.51	84	0.53
1098 (NGT)	157	6	Moderate	24	0.15	56	0.36	93	0.62	41	0.44	81	0.54
1134 (NNT)	60	4	Moderate	9	0.15	18	0.30	43	0.77	15	0.35	19	0.34
1158 (NHT)	24	0	Rare	0	0.00	22	0.92	21	0.88	11	0.52	16	0.67
1173 (NAS)	16	0	Rare	0	0.00	15	0.94	13	0.81	7	0.54	9	0.56
1194 (NES)	49	0	Rare	0	0.00	27	0.55	38	0.78	18	0.47	25	0.51
	1,759	≥4		≥0.15		≥0.5		≤0.7		≤0.4		≤0.35	

Fig. 1 Rates of glycan types and terminal modifications on each glycosylation site. Rates of glycan types such as high-mannose type, hybrid type, and branched type are estimated based on the number of glycan compositions assigned on each site. The cells in which the numbers or rate are larger than criteria presented in the lowest cells, are colored. The rate of HM type is categorized from the extracted ion chromatograms (XIC) as described in the main text. Numbers and rate of terminal modifications such as fucosylation and sialylation are obtained by the same way for type analyses. The cells less than criteria are colored.

Fig. 2 Illustration of formation of glycan stem (#Hex-#HexNAc matrix) considering the biosynthetic pathway. Extension or excision of glycan stem can be illustrated on a matrix of the numbers of Hex (Man or Gal) and HexNAc (GlcNAc or GalNAc). Frequencies of glycan stems (combination of Hex-HexNAc) in 22 potential glycosylation sites are plotted on the matrix.

Fig. 3 Schematic diagrams for LMA analyses. The samples for antibody-overlay method (A) and direct labeling method (B) were prepared from the same batch of the inactivated virus solution. In the antibody-overlay method, immunoprecipitated viral antigens were detected by overlaying an anti-S specific monoclonal antibody and fluorescent labeled streptavidin. In the direct labeling method, immunoprecipitated viral antigens were directly labeled with fluorophores and S proteins were further purified by second-round immunoprecipitation.

Fig. 4 Glycan profiles of S protein. Glycoforms of S protein were analyzed with LMA analyses using 45 lectins. S proteins were enriched from inactivated viral solution (closed bars) or PBS (open bars), and signals were detected with either antibody-overlay (A) or direct labeling (B) methods. Data are represented as mean signals of three spots $\pm$ standard deviations (SD).

Fig. 5 Effects of exoglycosidase digestion on glycan profiles of S protein. Glycoforms of S protein were analyzed with LMA analyses after the digestion with sialidase (A), β -galactosidase (B), and double digestion with sialidase and β -galactosidase (C). S protein prepared with the direct labeling method was digested with exoglycosidases and subjected to analyses. Data are represented as mean signals of three spots $\pm$ standard deviations (SD). Closed bars indicate signals in digested samples, and open bars indicate those in undigested samples.

Supplementary Fig. S1 Three-dimensional structural-based mappings of *N*-linked glycosylation sites on S proteins. *N*-linked glycosylation sites of S protein were mapped on the 3D model of S protein in the closed conformation (PDB ID: 6ZGE). Each monomer is depicted in light gray, light pink, or light blue. *N*-linked glycosylation sites on the light gray monomer are colored yellow (rare), light green (moderate), or green (abundant) based on the frequency of high-mannose type glycans. The original structure missed structural information on amino acids 1–13, 71–75, and 1147–1273. Accordingly, sites 74, 1158, 1173, and 1194, all of which were categorized as rare (yellow), are not presented on the structure.

2019-nCoV/Japan/TY/WK-521/2020	1	MFVFLVLLPLVSSQCVN--LTTRTQLPPAYT--NSFTRGVYYPDKVFRSSVLHSTQDLFL	56
SARS-CoV-2/WHU-01	1		56
hCoV-19/England/MILK-9E05B3/2020	1		56
hCoV-19/South Africa/KRISP-K004312/2020	1		56
hCoV-19/Japan/IC-0561/2021	1	F.N..S..	56
RaTG13	1	S.....L.....	56
PCoV/GX-P4L	1	FV.....GI..G..S.....I..L.....	56
Rc-o319	1	..ILVL.NV.SVYG..L.YGSHHL--V--S.R.....DA...TNVLHEGF..	54
SARS-CoV/Tor2	1	..I..LF.T.T.GSDLDRC.T.FDDVQA.N..QHT.SM.....EI...DT.YL.....	60
2019-nCoV/Japan/TY/WK-521/2020	57	PFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTSEKSNIRGWIFGTTLDSKTQS	116
SARS-CoV-2/WHU-01	57		116
hCoV-19/England/MILK-9E05B3/2020	57	S--	114
hCoV-19/South Africa/KRISP-K004312/2020	57	A.....	116
hCoV-19/Japan/IC-0561/2021	57		116
RaTG13	57	I.....	116
PCoV/GX-P4L	57	NT.NYQ.--F.K.....AR.....	114
Rc-o319	55	..D....YSFWN-----QKYSVATS..G....STID...VV...V...ND...	107
SARS-CoV/Tor2	61	..Y....G..T.N-----HT.G...I..K..I...A....VV...V...MNN.S...	113
2019-nCoV/Japan/TY/WK-521/2020	117	LLIVNNATNVVIVKVFQFCNDPFLGVYHKNKNSWMESEFRVYSSANNCTFEYVSQPFL	176
SARS-CoV-2/WHU-01	117		176
hCoV-19/England/MILK-9E05B3/2020	115		173
hCoV-19/South Africa/KRISP-K004312/2020	117		176
hCoV-19/Japan/IC-0561/2021	117	Y.....	176
RaTG13	117		176
PCoV/GX-P4L	115	T.....N...T.V.N.....I.....	174
Rc-o319	108	V.LY.DG.H.RVE..T.H..PT.VFSA-----SPHL...F...LN.TLASVR	156
SARS-CoV/Tor2	114	VI..I..S....RA.N.EL.DN..FA.----SKPMGTQHTMIFDN.F....I.DA.S	169
2019-nCoV/Japan/TY/WK-521/2020	177	MDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTP----INLVRDLPQGFSALEPLVDLPIG	232
SARS-CoV-2/WHU-01	177		232
hCoV-19/England/MILK-9E05B3/2020	174		229
hCoV-19/South Africa/KRISP-K004312/2020	177	G.....	232
hCoV-19/Japan/IC-0561/2021	177	S.....	232
RaTG13	177	P.....	232
PCoV/GX-P4L	175	V.....D.....R..A.....	230
Rc-o319	157	A.FTEVD.S..TI.....LQ..SLNV.YAS.SYVLA.GATSQ..S.VTP.V..WKI...	216
SARS-CoV/Tor2	170	L.VSE.S....H.....K..FLYV.KGYQ.----DV....S..NT.K.IFK..L.	225
2019-nCoV/Japan/TY/WK-521/2020	233	INITRFQTLALHRSYLTPGDSSSGWTAGAAAYVGYLQPRFTLLKYNENGTITDAVDCA	292
SARS-CoV-2/WHU-01	233		292
hCoV-19/England/MILK-9E05B3/2020	230		289
hCoV-19/South Africa/KRISP-K004312/2020	233	X.....	289
hCoV-19/Japan/IC-0561/2021	233		292
RaTG13	233		292
PCoV/GX-P4L	231	NLE...T.....Q.....S..Q.....S	290
Rc-o319	217	L...N.K..V-----RSDN-----PLQ...V..H.KR..MMF..D...V..I...	266
SARS-CoV/Tor2	226	...N.RAI.----TAFS.AQDI--GTS....F...K.T..M...D.....S	279
2019-nCoV/Japan/TY/WK-521/2020	293	LDPLSETKCTLKSFTEVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYA	352
SARS-CoV-2/WHU-01	293		352
hCoV-19/England/MILK-9E05B3/2020	290		349
hCoV-19/South Africa/KRISP-K004312/2020	290		349
hCoV-19/Japan/IC-0561/2021	293		352
RaTG13	293	D.....T.....	352
PCoV/GX-P4L	291	L.....I.....SK.....	350
Rc-o319	267	R..I.....QDTV.....S.....T.....	326
SARS-CoV/Tor2	280	QN..A.L..SV...EID.....V.SGDV.....K.P.....	339
2019-nCoV/Japan/TY/WK-521/2020	353	WNRKRISNCVADYSVLNYSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVQRQIAP	412
SARS-CoV-2/WHU-01	353		412
hCoV-19/England/MILK-9E05B3/2020	350		409
hCoV-19/South Africa/KRISP-K004312/2020	350		409
hCoV-19/Japan/IC-0561/2021	353		412
RaTG13	353	T.....T.....	412
PCoV/GX-P4L	351	T.....VK.....	410
Rc-o319	327	T.....Q.....S.....V.....	386
SARS-CoV/Tor2	340	.E..K.....TF.....A.....S.....VK..D.....	399

Supplementary Fig. S2 Comparison of potential N-glycosylation sites among the S protein of viruses belonging to subgenus *Sarbecovirus*. Amino acid sequences of S proteins from seven viruses belonging to subgenus *Sarbecovirus* were aligned with MUSCLE and N-IP-(S|T) sequons are highlighted. Sequons conserved across all the seven viruses are colored yellow. Sequons observed in SARS-CoV-2 but not conserved across all the members are colored green. Sequons not observed in SARS-CoV-2 are colored gray. ACE2 binding sites are indicated in red boxes. S1/S2 and S2' proteolytic sites are shown with open red arrowheads.

2019-nCoV/Japan/TY/WK-521/2020	413	GQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNLYRLFRKSNLKPFFERDISTEI	472
SARS-CoV-2/WHU-01	413		472
hCoV-19/England/MILK-9E05B3/2020	410		469
hCoV-19/South Africa/KRISP-K004312/2020	410	...N.....	469
hCoV-19/Japan/IC-0561/2021	413	...T.....	472
RaTG13	413	KHI.A.E...F...A.....	472
PCoV/GX-P4L	411	...V.....VKQ.ALT...G.....K.....	470
Rc-o319	387	...V.....L...R.Q.ASTS...F.Y...IW.SEK.R.....AHYD	446
SARS-CoV/Tor2	400	...V.....M...L...TR.I.A.TST.....K...YL...HGK.R.....NVP	459
2019-nCoV/Japan/TY/WK-521/2020	473	YQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSEFELLHAPATVCGPKKSTN	532
SARS-CoV-2/WHU-01	473		532
hCoV-19/England/MILK-9E05B3/2020	470	Y.....	529
hCoV-19/South Africa/KRISP-K004312/2020	470	K.....Y.....	529
hCoV-19/Japan/IC-0561/2021	473	K.....Y.....	532
RaTG13	473	...K...QT.L...Y...YR...Y...D...H.....N.....	532
PCoV/GX-P4L	471	QV.L...Y...ER...H...T...N...F.....NG.....L...T	530
Rc-o319	447	...V.TQ-----KS--S.KN...YSSA.DSH.....N.....Q...E	497
SARS-CoV/Tor2	460	FSPDGK..T-PPAL...W..ND...YT.T.I.....N.....L...D	518
2019-nCoV/Japan/TY/WK-521/2020	533	LVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSF	592
SARS-CoV-2/WHU-01	533		592
hCoV-19/England/MILK-9E05B3/2020	530	D.....	589
hCoV-19/South Africa/KRISP-K004312/2020	530		589
hCoV-19/Japan/IC-0561/2021	533		592
RaTG13	533		592
PCoV/GX-P4L	531	...D.....T.K.Q.....S.....	590
Rc-o319	498	.I.....D.....QS.....VS.F..S.K..K...V.....Y	557
SARS-CoV/Tor2	519	.I..Q.....P.S.R.Q.....VS.F..S...K.S...S...A	578
2019-nCoV/Japan/TY/WK-521/2020	593	GGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIG	652
SARS-CoV-2/WHU-01	593		652
hCoV-19/England/MILK-9E05B3/2020	590	G.....	649
hCoV-19/South Africa/KRISP-K004312/2020	590	G.....	649
hCoV-19/Japan/IC-0561/2021	593	G.....	652
RaTG13	593	A.....	652
PCoV/GX-P4L	591	M...E...A...A.A.....V	650
Rc-o319	558	...A.T.....D..T...E...S.....T.M...Q.....	617
SARS-CoV/Tor2	579	...A.SE.....D.ST.....A..I...N...Q.....	638
2019-nCoV/Japan/TY/WK-521/2020	653	AEHVNNSEYECDIPIGAGICASYQTQTNSPRRARSV-ASQSIIATYMSLGAENSVAYSNNNS	711
SARS-CoV-2/WHU-01	653		711
hCoV-19/England/MILK-9E05B3/2020	650	H.....	708
hCoV-19/South Africa/KRISP-K004312/2020	650	V.....	708
hCoV-19/Japan/IC-0561/2021	653	.Y.....	711
RaTG13	653	S.....	707
PCoV/GX-P4L	651	...V.....HSMSS---L...-NQR.....	705
Rc-o319	618	...D.....T.H.PSM---L..ANNKR.V..V.....T	673
SARS-CoV/Tor2	639	...DT.....H.VSL---L..T-SQK..V.....DS.I...T	693
2019-nCoV/Japan/TY/WK-521/2020	712	IAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSSTECNSLLQYGSFCTQLNRALTGIA	771
SARS-CoV-2/WHU-01	712		771
hCoV-19/England/MILK-9E05B3/2020	709	...I.....	768
hCoV-19/South Africa/KRISP-K004312/2020	709		768
hCoV-19/Japan/IC-0561/2021	712		771
RaTG13	708		767
PCoV/GX-P4L	706	I.....	765
Rc-o319	674	...VM.....V...T.....	733
SARS-CoV/Tor2	694	...S...I...VM...A...N.....A.....S...	753
2019-nCoV/Japan/TY/WK-521/2020	772	VEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADA	831
SARS-CoV-2/WHU-01	772		831
hCoV-19/England/MILK-9E05B3/2020	769		828
hCoV-19/South Africa/KRISP-K004312/2020	769		828
hCoV-19/Japan/IC-0561/2021	772		831
RaTG13	768		827
PCoV/GX-P4L	766		825
Rc-o319	734	...D.....	793
SARS-CoV/Tor2	754	A...R..R.....M...TL.Y.....L..T.....	813

Supplementary Fig. S2 (continued)

2019-nCoV/Japan/TY/WK-521/2020	832	GFIKQYGDCLGDIAARDLICAQKFNGLTVLPLLTDEMQYTSALLAGTITSGWTFGAG	891
SARS-CoV-2/WHU-01	832		891
hCoV-19/England/MILK-9E05B3/2020	829		888
hCoV-19/South Africa/KRISP-K004312/2020	829		888
hCoV-19/Japan/IC-0561/2021	832		891
RaTG13	828		887
PCoV/GX-P4L	826		885
Rc-o319	794	..V.....Q.....A..A..IS..A..A.....	853
SARS-CoV/Tor2	814	..M....E....N.....D...A..A..VS..A..A.....	873
2019-nCoV/Japan/TY/WK-521/2020	892	AALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSTASALGKLQDV	951
SARS-CoV-2/WHU-01	892		951
hCoV-19/England/MILK-9E05B3/2020	889		948
hCoV-19/South Africa/KRISP-K004312/2020	889		948
hCoV-19/Japan/IC-0561/2021	892		951
RaTG13	888		947
PCoV/GX-P4L	886		945
Rc-o319	854	E..T.....	913
SARS-CoV/Tor2	874	Q.....K..SQ..E..TT.ST.....	933
2019-nCoV/Japan/TY/WK-521/2020	952	VNQNAAQALNTLVKQLSSNFGAISSVLNDILSRDKVEAEVQIDRLITGRLQSLQTYVTQQ	1011
SARS-CoV-2/WHU-01	952		1011
hCoV-19/England/MILK-9E05B3/2020	949	A.....	1008
hCoV-19/South Africa/KRISP-K004312/2020	949		1008
hCoV-19/Japan/IC-0561/2021	952		1011
RaTG13	948		1007
PCoV/GX-P4L	946		1005
Rc-o319	914	I.....	973
SARS-CoV/Tor2	934		993
2019-nCoV/Japan/TY/WK-521/2020	1012	LIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLSFPQSAPHGTVFLHVTYVPAQ	1071
SARS-CoV-2/WHU-01	1012		1071
hCoV-19/England/MILK-9E05B3/2020	1009		1068
hCoV-19/South Africa/KRISP-K004312/2020	1009		1068
hCoV-19/Japan/IC-0561/2021	1012	I.....	1071
RaTG13	1008		1067
PCoV/GX-P4L	1006		1065
Rc-o319	974	A.....I.S.....	1033
SARS-CoV/Tor2	994	A.....S.....	1053
2019-nCoV/Japan/TY/WK-521/2020	1072	EKNFTTAPAICHGDKAHFPREGVFSVNGTHWVFVTQRNFYEPQIITDNTFVSGNCDVVIG	1131
SARS-CoV-2/WHU-01	1072		1131
hCoV-19/England/MILK-9E05B3/2020	1069	H.....	1128
hCoV-19/South Africa/KRISP-K004312/2020	1069		1128
hCoV-19/Japan/IC-0561/2021	1072		1131
RaTG13	1068	S.....	1127
PCoV/GX-P4L	1066	E.....I.....S.....	1125
Rc-o319	1034	..R.....E.....I.....T.....	1093
SARS-CoV/Tor2	1054	..R.....E....Y.....F....S..I....FS.....	1113
2019-nCoV/Japan/TY/WK-521/2020	1132	IVNNTVYDPLQPELDSFKEELDKEYFKNHTSPDVLGDISGINASVVNIQKEIDRLNEVAK	1191
SARS-CoV-2/WHU-01	1132		1191
hCoV-19/England/MILK-9E05B3/2020	1129		1188
hCoV-19/South Africa/KRISP-K004312/2020	1129		1188
hCoV-19/Japan/IC-0561/2021	1132	F.....	1191
RaTG13	1128		1187
PCoV/GX-P4L	1126		1185
Rc-o319	1094	E...D.....I.....D.....I.KD...	1153
SARS-CoV/Tor2	1114	..I.....	1173
2019-nCoV/Japan/TY/WK-521/2020	1192	NLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCMTSCCCLKGCCSCG	1251
SARS-CoV-2/WHU-01	1192		1251
hCoV-19/England/MILK-9E05B3/2020	1189		1248
hCoV-19/South Africa/KRISP-K004312/2020	1189		1248
hCoV-19/Japan/IC-0561/2021	1192		1251
RaTG13	1188	I.....	1247
PCoV/GX-P4L	1186	P.....I.....	1245
Rc-o319	1154	N.....V.....I.....	1213
SARS-CoV/Tor2	1174	V.....L.....A.....	1233
2019-nCoV/Japan/TY/WK-521/2020	1252	SCCKFDEDDSEPVLKGVKLHYT	1273
SARS-CoV-2/WHU-01	1252		1273
hCoV-19/England/MILK-9E05B3/2020	1249		1270
hCoV-19/South Africa/KRISP-K004312/2020	1249		1270
hCoV-19/Japan/IC-0561/2021	1252		1273
RaTG13	1248		1269
PCoV/GX-P4L	1246		1267
Rc-o319	1214		1235
SARS-CoV/Tor2	1234		1255

Supplementary Fig. S2 (continued)