Site-specific N-glycosylation Characterization of Recombinant SARS-CoV-2 Spike Proteins using High-Resolution Mass Spectrometry

Yong Zhang[a, #], Wanjun Zhao[b, #], Yonghong Mao[c], Shisheng Wang[a], Yi Zhong[a], Tao Su[a], Meng Gong[a], Xiaofeng Lu[a], Jingqiu Cheng[a], Hao Yang[a,*]

[a]Key Laboratory of Transplant Engineering and Immunology, MOH, West China-Washington Mitochondria and Metabolism Research Center, West China Hospital, Sichuan University, Chengdu 610041, China

[b]Department of Thyroid Surgery, West China Hospital, Sichuan University, Chengdu 610041, China

[c]Thoracic Surgery Research Laboratory, West China Hospital, Sichuan University, Chengdu 610041, China

*Corresponding author:

E-mail address: yanghao@scu.edu.cn (H. Yang)

[#]These authors contributed equally to this work.

ABSTRACT

The global pandemic of severe acute pneumonia syndrome (COVID-19) caused by SARS-CoV-2 urgently calls for prevention and intervention strategies. The densely glycosylated spike (S) protein highly exposed on the viral surface is a determinant for virus binding and invasion into host cells as well as elicitation of a protective host immune response. Herein, we characterized the site-specific N-glycosylation of SARS-CoV-2 S protein using stepped collision energy (SCE) mass spectrometry (MS). Following digestion with two complementary proteases to cover all potential N-glycosylation sequons and integrated N-glycoproteomics analysis, we revealed the N-glycosylation profile of SARS-CoV-2 S proteins at the levels of intact N-glycopeptides and glycosites, along with the glycan composition and site-specific number of glycans. All 22 potential canonical N-glycosites were identified in S protein protomer. Of those, 18 N-glycosites were conserved between SARS-CoV and SARS-CoV-2 S proteins. Nearly all glycosites were preserved among the 753 SARS-CoV-2 genome sequences available in the public influenza database Global Initiative on Sharing All Influenza Data. By comparison, insect cell-expressed SARS-CoV-2 S protein contained 38 N-glycans, which were primarily assigned to the high-mannose type N-glycans, whereas the human cell-produced protein possessed up to 140 N-glycans largely belonging to the complex type. In particular, two N-glycosites located in the structurally exposed receptor-binding domain of S protein exhibited a relatively conserved N-glycan composition in human cells. This N-glycosylation profiling and determination of differences between distinct expression systems could shed light on the infection mechanism and promote development of vaccines and targeted drugs.

Keywords:

SARS-CoV-2; Spike protein; N-glycosylation; Mass spectrometry

Introduction

In the last month of 2019, a novel coronavirus (SARS-CoV-2) emerged and rapidly spread, developing into an epidemic of severe acute pneumonia syndrome (COVID-19), which engulfed the city of Wuhan and Hubei Province of China. The virus quickly swept through the entirety of China and subsequently the outbreak came out in Asia and the entire world within a couple of months. Similar to SARS-CoV and MERS-CoV that emerged in 2002 and 2013, respectively, SARS-CoV-2 is highly transmissible from infected individuals, even without symptoms, to healthy humans and can cause lethal respiratory symptoms[1-3]. The World Health Organization has declared the spread of SARS-CoV-2 a Public Health Emergency of International Concern as over 160 countries have reported confirmed cases. From SARS-CoV to SARS-CoV-2, the periodic outbreak of coronavirus infection in humans urgently calls for prevention and intervention measures. However, there are no approved vaccines or effective antiviral drugs for either SARS-CoV or SARS-CoV-2.

Decoding the critical component and molecular characteristics of the virus is the key to developing a cure strategy. SARS-CoV-2 is a single-stranded RNA virus. RNA sequencing revealed that SARS-CoV-2 belongs to the beta-coronavirus genus and is most closely related to SARS-CoV, with a genome size of approximately 30 kb encoding 15 non-structural proteins, 4 structural proteins, and 8 auxiliary proteins[4,5]. The structural proteins of mature SARS-CoV-2 include spike (S) protein, envelope (E) protein, membrane (M) protein, and nucleocapsid (N) protein[2]. Theoretically, all structural proteins can serve as antigens for vaccine development or targets for anti-viral treatment. Of these proteins, the transmembrane S protein protruding from the virus surface is highly exposed and responsible for invasion into host cells, which has attracted the special attention of researchers. S protein is homotrimeric and is highly glycosylated on the virus surface, allowing for binding to the angiotensin converting enzyme II (ACE2) receptor on host cells to promote the fusion of viral and host cellular membranes[6,7]. Given its indispensable role in virus entry and infectivity, S protein is a promising target for vaccine design and drug discovery to block the interaction between the virus and host cells. S protein has been revealed as a crucial antigen for raising neutralizing antibodies and eliciting protective humoral as well as cellular immunity upon infection or vaccination with S protein-based vaccines[8-11].

Typically, S protein has an ectodomain linked to a single-pass transmembrane anchor and a short C-terminal intracellular tail[12]. The ectodomain comprises a receptor-binding S1 subunit and a membrane-fusion S2 subunit. Following attachment to the host cell surface via S1, SARS-CoV-2 S protein is cleaved at the S1/S2 boundary sites together with the junction site and the S2′ site by host cellular proteases for S protein priming, consequently mediating membrane fusion driven by S2 and making way for the viral genetic materials to enter the host cell[10]. Based on cryoelectron microscopy (Cryo-EM) observations, recognition of S protein to the ACE2 receptor primarily involves extensive polar residue interactions between the SARS-CoV-2 receptor binding domain (RBD) and the peptidase domain of ACE2[7,13]. The S protein RBD is located in the S1 subunit and undergoes a hinge-like dynamic movement to capture the receptor through the interaction of three group residue clusters between the RBD and ACE2. Compared to S protein of SARS-CoV, that of SARS-CoV-2 displays up to

10–20-fold higher affinity for the human ACE2 receptor, which partially explains the higher transmissibility of this new virus[7,13].

In addition to the structural information and core amino acid residues for receptor binding, SARS-CoV-2 S protein possesses 22 potential N-linked glycosylation motifs (N-X-S/T, X≠P) in each monomer. The N-glycans on S protein play a pivotal role in proper protein folding and protein priming by host proteases. Importantly, glycosylation is an underlying mechanism for coronavirus to evade both the innate and adaptive immune responses, as the glycans might shield the amino acid residues from cell and antibody recognition[10,11,14]. Cryo-EM observations provided evidence of the existence of glycans on 14–16 of 22 potential sites in SARS-CoV-2 S protein[7,10]. However, these glycosites and glycans need to be experimentally identified in detail. Glycosylation analysis via glycopeptides can provide insight into the N-glycan microheterogeneity of a specific site, as variation in site-specific glycosylation levels can be greater than that at the protein level[15]. Therefore, further identification of site-specific N-glycosylation information of SARS-CoV-2 S protein, including intact N-glycopeptides, glycosites, glycan composition, and the site-specific number of glycans, could be meaningful to obtain a deeper understanding of the mechanism of virus invasion, providing guidance for vaccine design and antiviral therapeutics development[10,16]

Herein, we characterized the site-specific N-glycosylation of recombinant SARS-CoV-2 S proteins by combined analysis of intact and deglycosylated N-glycopeptides using tandem mass spectrometry (MS/MS). Based on this integrated method, we identified 22 potential canonical N-glycosites and their corresponding N-glycans from the recombinant ectodomain (residues 16–1213) expressed in insect cells. For comparison, glycosylation of the recombinant S1 subunit (residues 16–685) expressed in human cells was resolved in parallel. All of these glycosites were found to be highly conserved among 753 SARS-CoV-2 genome sequences from the Global Initiative on Sharing All Influenza Data (GISAID) database. These detailed glycosylation profiles decoded from MS/MS analysis are complementary to those observed from Cryo-EM and might help in the development of vaccines and therapeutic drugs. The raw MS data are publicly accessible at ProteomeXchange (ProteomeXchange.com) under accession number PXD018068.

Materials and Methods

Materials and chemicals

Dithiothreitol (DTT), iodoacetamide (IAA), formic acid (FA), trifluoroacetic acid (TFA), Tris base, and urea were purchased from Sigma (St. Louis, MO, USA). Acetonitrile (ACN) was purchased from Merck (Darmstadt, Germany). The zwitterionic hydrophilic interaction liquid chromatography (Zic-HILIC) materials were purchased from Fresh Bioscience (Shanghai, China). Commercially available recombinant SARS-CoV-2 S protein (S1+S2 ECD, His tag) expressed in insect cells via baculovirus and S protein (S1, His tag) expressed in human embryonic kidney (HEK293) cells were purchased from Sino Biological (Beijing, China). Sequencing-grade trypsin and Glu-C were obtained from Enzyme & Spectrum (Beijing, China). The quantitative colorimetric peptide assay kit was purchased from Thermo Fisher Scientific (Waltham, MA, USA). Deionized water was prepared using a Milli-Q system (Millipore, Bedford, MA, USA). All other chemicals and reagents of the best available grade were purchased from Sigma-Aldrich or Thermo Fisher Scientific.

Protein digestion

The recombinant S proteins were proteolyzed using an in-solution protease digestion protocol. In brief, 50 μg of protein in a tube was denatured for 10 min at 95 °C. After reduction by DTT (20 mM) for 45 min at 56 °C and alkylating with IAA (50 mM) for 1 h at 25 °C in the dark, 2 μg of protease (trypsin or/and Glu-C) was added to the tube and incubated for 16 h at 37 °C. After desalting using a pipette tip packed with a C18 membrane, the peptide concentration was determined using a peptide assay kit based on the absorbance measured at 480 nm. The peptide mixtures (intact N-glycopeptides before enrichment) were freeze-dried for further analysis.

Selective enrichment of intact N-glycopeptides

Intact N-glycopeptides were enriched with Zic-HILIC materials (Fresh Bioscience, Shanghai, China). Specifically, 20 μg of peptides was resuspended in 100 μL of 80% ACN/0.2% TFA solution, and 2 mg of processed Zic-HILIC was added to the peptide solution and rotated for 2 h at 37 °C. Finally, the mixture was transferred to a 200-μL pipette tip packed with the C8 membrane and washed twice with 80% ACN/0.2% TFA. After enrichment, intact N-glycopeptides were eluted three times with 70 μL of 0.1% TFA and dried using SpeedVac for further analysis.

Deglycosylation

Enriched intact N-glycopeptides were digested using 1 U PNGase F dissolved in 50 μL of 50 mM NH_4HCO_3 for 2 h at 37 °C. The reaction was terminated by the addition of 0.1% FA. The deglycosylated peptides were dried using SpeedVac for further analysis.

Liquid chromatography-MS/MS analysis

All samples were analyzed by SCE-higher-energy collisional dissociation (HCD)-MS/MS using an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific). In brief, intact N-glycopeptides before or after enrichment and deglycosylated peptides were dissolved in 0.1% FA and separated on a column (ReproSil-Pur C18-AQ, 1.9 μm, 75 μm inner diameter, length 20 cm; Dr Maisch) over a 78-min gradient (buffer A, 0.1% FA in water; buffer B, 0.1% FA in 80% ACN) at a flow rate of 300 nL/min. MS1 was analyzed with a scan range (m/z) of 800–2000 (intact N-glycopeptides before or after enrichment) or 350–1550 (deglycosylated peptides) at an Orbitrap resolution of 120,000. The RF lens, AGC target, maximum injection time, and exclusion duration were 30%, 2.0 e^4, 100 ms, and 15 s, respectively. MS2 was analyzed with an isolation window (m/z) of 2 at an Orbitrap resolution of 15,000. The AGC target, maximum injection time, and HCD type were standard, 250 ms, and 30%, respectively. The stepped collision mode was turned on with an energy difference of ±10%.

Data analysis

The raw data files were searched against the SARS-CoV-2 S protein sequence using Byonic software (version 3.6.0, Protein Metrics, Inc.)[17] with the mass tolerance for precursors and fragment ions set at ±10 ppm and ±20 ppm, respectively. Two missed cleavage sites were allowed for trypsin or/and Glu-C digestion. The fixed modification was carbamidomethyl (C), and variable modifications included oxidation (M), acetyl (protein N-term), and deamidation (N). In addition, 38 insect N-glycans or 182 human N-glycans were specified as N-glycan modifications for intact N-glycopeptides before or after enrichment. We then checked the protein database options, including the decoy database. All other parameters were set at the default values, and protein groups were filtered to a 1% false discovery rate based on the number of hits obtained for searches against these databases. Stricter quality control methods for intact N-glycopeptides and peptide identification were implemented, requiring a score of no less than 200 and at least 7 amino acids to be identified. Furthermore, all of these peptide spectrum matches (PSMs) and glycopeptide-spectrum matches (GPSMs) were examined manually and filtered using the following standard criteria: PSMs were accepted if there were at least 3 b/y ions in the peptide backbone, and GPSMs were accepted if there were at least two glycan oxonium ions and at least 3 b/y ions in the peptide backbone. N-glycosite conservation analysis was performed using R software packages. Model building based on the Cryo-EM structure (PDB: 6VSB) of SARS-CoV-2 S protein was performed using PyMOL.

Results and Discussion

Strategy for site-specific N-glycosylation characterization

Previous studies have revealed that glycosylated coronavirus S protein plays a critical role in the induction of neutralizing antibodies and protective immunity. However, the glycans decorated on S protein might also shield the protein surface and lead to virus evasion from the immune system[8,11,14]. Herein, we aimed to decode the detailed site-specific N-glycosylation profile of SARS-CoV-2 S protein. A commercial S protein ectodomain expressed by the baculovirus expression vector in insect cells was first used to analyze the glycosylation patterns, since the baculovirus vector can express a large protein without resulting in splicing of the native S protein via host proteases[10,18]. The recombinant SARS-CoV-2 S ectodomain contains 1209 amino acids (residues 16–1213) that were translated from the complete genome

(GenBank: MN908947.3)[19] and 22 putative N-glycosites (motif N-X-S/T, X≠P). Analysis of the theoretical enzymatic peptides showed that trypsin (hydrolyzing proteins at K and R) alone did not produce a sufficient amount of appropriate peptides to cover all potential N-glycosites. The missing potential N-glycosites could be found by introducing the endoproteinase Glu-C (hydrolyzing proteins at D in ammonium bicarbonate solution)[20]. Hence, we took advantage of this complementary trypsin and Glu-C digestion approach (Figure 1). Since the N-glycan compositions of S protein expressed in insect cells would be different from those of the native S protein expressed in human host cells, despite insect cells mimicking the process of mammalian cell glycosylation[18], the recombinant SARS-CoV-2 S protein S1 subunit expressed in human cells was obtained for comparison. S2 subunit N-glycosylation sequons are conserved among SARS-CoV-2 and SARS-CoV and have been confirmed in previous studies[10,11]. The S1 subunit contains 681 amino acids (residues 16–685) and 13 potential N-glycosites. Analysis of the theoretical enzymatic peptides showed that trypsin digestion alone would produce a sufficient amount of appropriate peptides to cover all potential N-glycosites on the S1 subunit.

In general, the relative content of N-glycosylated peptides in a glycoprotein is low; hence, enrichment of intact N-glycopeptides is necessary[21]. For this purpose, we used Zic-HILIC materials to enrich intact glycopeptides through hydrophilic interactions owing to their high selectivity and reproducibility[22]. However, due to the microheterogeneity (different glycans attached to the same glycosite) and macroheterogeneity (glycosite occupancy) of glycosylation[23], there are no materials available that can capture all glycopeptides without preference[24,25]. For these reasons, site-specific glycosylation was determined based on a combined analysis of the intact N-glycopeptides before and after enrichment and the deglycosylated peptides following enrichment using SCE-HCD-MS/MS[26,27] (Figure 1). Analysis of intact N-glycopeptides before enrichment can retrieve the missing intact N-glycopeptides from Zic-HILIC materials, while detection of deglycosylated peptides can simultaneously confirm the N-glycosites. Therefore, integration of complementary digestion and N-glycoproteomics analysis from three levels is a promising approach to comprehensively and confidently profile the site-specific N-glycosylation of recombinant SARS-CoV-2 S proteins.

N-glycosite profiling of recombinant SARS-CoV-2 S proteins

S protein produced by the baculovirus insect cell expression system contains 22 potential N-glycosites. Using our integrated analysis method described above, 20 N-glycosites were assigned unambiguously with high-quality (score ≥ 200) spectral evidence (Figure 2A, Table S1 and Table S2). Two N-glycosites (N17 and N1134) were ambiguously assigned with relatively lower spectral scores (score < 200) (Figure S1). However, the N-glycosite N1134 has been reported in the Cryo-EM structure of SARS-CoV-2 S protein[10]. In addition, three non-canonical motifs of N-glycosites (N164, N334, and N536) involving N-X-C sequons were not N-glycosylated. Before enrichment, 11 N-glycosites from trypsin-digested peptides and 9 N-glycosites from Glu-C-digested peptides were assigned unambiguously, whereas hydrophilic enrichment resulted in an increase of these glycosites to 14 and 11, respectively (Table S1).

To further assess the necessity for enrichment, we compared the spectra of two intact

N-glycopeptides (N61 and N74) before and after enrichment. Without interference from the non-glycosylated peptides, the intact N-glycopeptide had more fragmented ions assigned to N-glycosites after enrichment (Figure S2). Exceptionally, the intact N-glycopeptide containing an N-glycosite (N17) was missed after enrichment, presumably because of the selectivity of Zic-HILIC (Table S1).

Complementary digestion with trypsin and Glu-C promoted the confident identification of two N-glycosites (N709 and N717) on an intact N-glycopeptide (Figure S3). The introduction of Glu-C digestion resulted in the production of a short intact N-glycopeptide containing 23 amino acids, which is more suitable for achieving good fragmentation than the long peptide of 48 amino acids obtained from trypsin digestion (Figure S3). Similarly, the N-glycosylation analysis strategy combining intact N-glycopeptides with deglycosylated peptides improved the identification of N-glycosite N234, which was ambiguously assigned in the spectrum of the intact N-glycopeptides alone (Figure S4).

For the recombinant protein S1 subunit expressed in human cells, 12 out of 13 N-glycosites were assigned unambiguously with high-quality spectral evidence. One N-glycosite (N17) was assigned ambiguously with a relatively low spectral score (Figure 2B, Table S2 and Table S3). The relatively low spectral evidence of two N-glycosites (N17 and N1134) indicate the existence of low-frequency glycosylation on the these ambiguous glycosites since deglycosylation failed to improve the identification of all the two sites. Finally, using this strategy, we profiled all 22 potential N-glycosites of S protein. These sites were preferentially distributed in the S1 subunit of the N-terminus and in the S2 subunit of the C-terminus, including two sites in the RBD (Figure 2A). To visualize N-glycosylation on the protein structure, all of the experimentally determined N-glycosites were hand-marked on the surface of trimeric S protein following refinement of the recently reported SARS-CoV-2 S protein Cryo-EM structure (PDB: 6VSB) (Figure 2C)[7].

Based on these findings, we further analyzed the conservation of the glycosites among 753 SARS-CoV-2 genome sequences from the GISAID database. After removal of redundant sequences of S protein at the amino acid residue level, we found a very low frequency of alterations in 38 residue sites uniformly spanning over the full length of S protein among 145 protein variants, with the exception of the substitution G614D, which was found at relatively high frequency in 47 variants (Table S4). However, nearly all of the 22 N-glycosylated sequons were conserved in S protein, except for loss of the N717 glycosite due to the T719A substitution in only one S protein variant. Compared to SARS-CoV S protein, 18 of the 22 N-glycosites were found to be conserved in SARS-CoV-2 S protein, indicating the importance of glycosylation for the virus. Four newly arised N-glycosites (N17, N74, N149, and N657) are located in the SARS-CoV-2 S protein S1 subunit away from the RBD. Moreover, four confirmed N-glycosites (N29, N73, N109, and N357) in SARS-CoV S protein were missing in SARS-CoV-2 S, one of which (N357) lies in the RBD[11,28].

Intact N-glycopeptides of recombinant SARS-CoV-2 S proteins

Precise characterization of intact N-glycopeptides is critical for understanding biological functions[29]. Although intact N-glycopeptide analysis is more challenging than analysis of separate N-glycosites or N-glycans, it can provide more comprehensive information, including N-glycosites, N-glycan compositions, and the number of N-glycans[30-32]. The potential N-glycopeptides in the S protein sequence are shown in Figure S5. Comparison of the intact N-glycopeptides spectra to the total spectra showed that the average enrichment efficiency of the Zic-HILIC materials reached up to 97%. Ultimately, 646 non-redundant intact N-glycopeptides were identified from SARS-CoV-2 S proteins (Table S1), and 410 non-redundant intact N-glycopeptides were identified from the recombinant S1 subunit (Table S3). Representative and high-quality spectra of intact N-glycopeptides are shown in Figure S6. The number of intact N-glycopeptides and N-glycans significantly increased after glycopeptide enrichment (Figure 3A and Figure 3B).

Regarding the N-glycan composition, N-glycopeptides of S protein expressed in insect cells had smaller and fewer complex N-glycans compared with those of the S1 subunit produced in human cells. Both recombinant products contained the common N-acetylglucosamine as a canonical N-glycan characteristic (Figure 3C and 3D). N-glycopeptides of S protein expressed in insect cells were decorated with 38 N-glycans, with the majority preferentially comprising paucimannose- and fucose-type oligosaccharides (Figure 3C and Table S1). By contrast, N-glycopeptides of the S1 subunit expressed in human cells were attached with up to 140 N-glycans, mainly containing fucose-type and unique sialic acid-type oligosaccharides (Figure 3D and Table S3). Returning to the glycosite level, most of the N-glycosites in S protein were modified with 17–35 types of N-glycans, with a high proportion of high-mannose N-glycans and a lower proportion of hybrid N-glycans (Figure 3E). For the S1 subunit, three N-glycosites (N122, N282, and N657) were surprisingly decorated with markedly heterogeneous N glycans of up to 113 types, including a high proportion of complex N-glycans and a small proportion of hybrid or high-mannose

N-glycans (Figure 3F). These results showed that the two S proteins expressed in different cells displayed different N-glycosylation patterns with a distinctive N-glycan composition (microheterogeneity) and different numbers of N-glycans at the same site, along with distinct site occupancies in intact glycopeptides (macroheterogeneity) (Figure S7).

The glycosylation patterns of proteins expressed in insect cells have been found to be more immunogenic than those produced in human cells[33,34], and antigen production from mammalian cells does not always induce a strong humoral immune response[35]. The complex and highly heterogeneous N-glycans modified on SARS-CoV-2 S protein expressed in human cells may be related to the differential immune response, which could possibly be caused by epitope masking by the glycan shield, although this hypothesis requires further testing and clarification. Therefore, the less complex N-glycans covering S protein expressed in insect cells might favor the development of vaccines to elicit neutralizing antibodies against SARS-CoV-2 virus.

Two potential N-glycosites (N331 and N343) in RBD were confirmed in the N-glycopeptides from both insect cell-expressed S protein and human cell-produced S1 subunit (Figure 3E and 3F). Both sites closely located in the same glycopeptide (Figure 4). Intriguingly, the composition of the N-glycans in human cells exhibited relatively high conservation compared to most of other N-glycosites (Figure 3F). However, the N-glycan composition were more variable in RBD expressed in insect cells (Figure 3E). These results imply that N-glycosylation modification might be associated with receptor binding since the recognition of RBD to ACE2 mainly depends on polar residue interaction[13,14]. In addition, complex N-glycans are also ligands for galectins, which can engage different glycoproteins and regulate immune cell infiltration and activation upon virus infection[36-38]. The RBDs of SARS-CoV-2 decorated with distinct N-glycans in different expression systems could be candidates for SARS-CoV-2 vaccine design as an alternative to full-length S protein which can lead to undesired immunopotentiation with respect to increased infectivity and eosinophilic infiltration[9]. Based on these results, a large-scale intact N-glycopeptides database of recombinant SARS-CoV-2 S proteins was developed. Nevertheless, the implication of S protein site-specific N-glycosylation (including N-glycosites and N-glycans) on receptor binding, viral infectivity, and immunogenicity should be further investigated.

Conclusions

A comprehensive analysis of site-specific N-glycosylation of SARS-CoV-2 S protein was performed at the levels of intact N-glycopeptides, glycosites, glycan composition, and site-specific numbers of N-glycans. By taking advantage of two complementary protease digestion systems and N-glycoproteomics analysis through enrichment and deglycosylation, we provided a global and site-specific profile of N-glycosylation on SARS-CoV-2 S proteins, revealing extensive heterogeneity in N-glycan composition and site occupancy. Almost all of these glycosites were conserved among the 753 published SARS-CoV-2 genome sequences. In particular, two N-glycosites in the S protein RBD produced in human cells showed relative conservation compared to the N-glycan composition of insect cells, suggesting the potential impact of N-glycosylation on receptor binding. Overall, our data indicate that N-glycosylation profiling and identifying differences among distinct expression systems might help to elucidate the infection mechanism toward development of an effective vaccine and targeted drugs.

ASSOCIATED CONTENT

Supporting Information

Figure S1, Spectra of intact N-glycopeptides with ambiguously assigned N-glycosites; Figure S2, Comparison of the spectra of intact N-glycopeptides before (A) and after (B) enrichment; Figure S3, Comparison of the spectra of intact N-glycopeptides from trypsin digestion (A) and Glu-C digestion (B); Figure S4, Comparison of the spectra of intact N-glycopeptides (A) and deglycopeptides (B); Figure S5, Amino acid sequence alignment of recombinant SARS-CoV-2 S proteins expressed in insect cells (A) and human cells (B). Yellow background: putative sequence containing a signal sequence; Red: potential N-glycosites; Red bold: identified N-glycosites; Green: theoretical cleavage sites of trypsin; Blue: theoretical cleavage sites of Glu-C; Figure S6, Representative and high-quality spectra of intact N-glycopeptides and deglycosylated peptides. Figure S7, Microheterogeneity and macroheterogeneity of the N-linked glycopeptides of S protein; Table S1, Site-specific N-glycosylation characterization of recombinant SARS-CoV-2 S protein expressed in insect cells; Table S2, Glycoproteomic identification results of recombinant SARS-CoV-2 spike protein using the combination of trypsin and Glu-C digestion; Table S3, Site-specific N-glycosylation characterization of recombinant SARS-CoV-2 S protein expressed in human cells; Table S4, Mutation frequency of SARS-CoV-2 spike protein.

AUTHOR INFORMATION

Corresponding Author

* E-mail address: yanghao@scu.edu.cn (H. Yang)

Author Contributions

#Y. Zhang and W. Zhao contributed equally to this work.

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENT

This work was funded by grants from the National Natural Science Foundation of China (grant number 31901038), the 1.3.5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (ZYGD18014, CJQ), and the Chengdu Science and Technology Department Foundation (grant number 2020-YF05-00240-SN).

REFERENCES

(1) Lu, R.; Zhao, X.; Li, J.; Niu, P.; Yang, B.; Wu, H.; Wang, W.; Song, H.; Huang, B.; Zhu, N.; Bi, Y.; Ma, X.; Zhan, F.; Wang, L.; Hu, T.; Zhou, H.; Hu, Z.; Zhou, W.; Zhao, L.; Chen, J., et al. *Lancet* **2020**, *395*, 565-574.
(2) Zhou, P.; Yang, X.-L.; Wang, X.-G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.-R.; Zhu, Y.; Li, B.; Huang, C.-L.; Chen, H.-D.; Chen, J.; Luo, Y.; Guo, H.; Jiang, R.-D.; Liu, M.-Q.; Chen, Y.; Shen, X.-R.; Wang, X.; Zheng, X.-S., et al. *Nature* **2020**, 10.1038/s41586-41020-42012-41587.
(3) Drosten, C.; Günther, S.; Preiser, W.; van der Werf, S.; Brodt, H.-R.; Becker, S.; Rabenau, H.; Panning, M.; Kolesnikova, L.; Fouchier, R. A. M.; Berger, A.; Burguière, A.-M.; Cinatl, J.; Eickmann, M.; Escriou, N.; Grywna, K.; Kramme, S.; Manuguerra, J.-C.; Müller, S.; Rickerts, V., et al. *N Engl J Med* **2003**, *348*, 1967-1976.
(4) Wu, F.; Zhao, S.; Yu, B.; Chen, Y. M.; Wang, W.; Song, Z. G.; Hu, Y.; Tao, Z. W.; Tian, J. H.; Pei, Y. Y.; Yuan, M. L.; Zhang, Y. L.; Dai, F. H.; Liu, Y.; Wang, Q. M.; Zheng, J. J.; Xu, L.; Holmes, E. C.; Zhang, Y. Z. *Nature* **2020**, *579*, 265-269.
(5) Wu, A.; Peng, Y.; Huang, B.; Ding, X.; Wang, X.; Niu, P.; Meng, J.; Zhu, Z.; Zhang, Z.; Wang, J.; Sheng, J.; Quan, L.; Xia, Z.; Tan, W.; Cheng, G.; Jiang, T. *Cell Host Microbe* **2020**, S1931-3128(1920)30072-X.
(6) Li, W.; Moore, M. J.; Vasilieva, N.; Sui, J.; Wong, S. K.; Berne, M. A.; Somasundaran, M.; Sullivan, J. L.; Luzuriaga, K.; Greenough, T. C.; Choe, H.; Farzan, M. *Nature* **2003**, *426*, 450-454.
(7) Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C. L.; Abiona, O.; Graham, B. S.; McLellan, J. S. *Science* **2020**, *367*, 1260-1263.
(8) Du, L.; He, Y.; Zhou, Y.; Liu, S.; Zheng, B. J.; Jiang, S. *Nat Rev Microbiol* **2009**, *7*, 226-236.
(9) Chen, W.-H.; Strych, U.; Hotez, P. J.; Bottazzi, M. E. *Current Tropical Medicine Reports* **2020**.
(10) Walls, A. C.; Park, Y. J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D. *Cell* **2020**.
(11) Walls, A. C.; Xiong, X.; Park, Y. J.; Tortorici, M. A.; Snijder, J.; Quispe, J.; Cameroni, E.; Gopal, R.; Dai, M.; Lanzavecchia, A.; Zambon, M.; Rey, F. A.; Corti, D.; Veesler, D. *Cell* **2019**, *176*, 1026-1039 e1015.
(12) Li, F. *Annual Review of Virology* **2016**, *3*, 237-261.
(13) Yan, R.; Zhang, Y.; Li, Y.; Xia, L.; Guo, Y.; Zhou, Q. *Science* **2020**.
(14) Walls, A. C.; Tortorici, M. A.; Frenz, B.; Snijder, J.; Li, W.; Rey, F. A.; DiMaio, F.; Bosch, B. J.; Veesler, D. *Nat Struct Mol Biol* **2016**, *23*, 899-905.

(15) Clerc, F.; Reiding, K. R.; Jansen, B. C.; Kammeijer, G. S.; Bondt, A.; Wuhrer, M. *Glycoconj J* **2016**, *33*, 309-343.

(16) Chang, D.; Zaia, J. *Mol Cell Proteomics* **2019**, *18*, 2348-2358.

(17) Medzihradszky, K. F.; Maynard, J.; Kaasik, K.; Bern, M. *Molecular & Cellular Proteomics* **2014**, *13*, S36-S36.

(18) Kost, T. A.; Condreay, J. P.; Jarvis, D. L. *Nat Biotechnol* **2005**, *23*, 567-575.

(19) Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; Yuan, M.-L.; Zhang, Y.-L.; Dai, F.-H.; Liu, Y.; Wang, Q.-M.; Zheng, J.-J.; Xu, L.; Holmes, E. C.; Zhang, Y.-Z. *Nature* **2020**.

(20) Gimenez, E.; Ramos-Hernan, R.; Benavente, F.; Barbosa, J.; Sanz-Nebot, V. *Analytica chimica acta* **2012**, *709*, 81-90.

(21) Suttapitugsakul, S.; Sun, F.; Wu, R. *Analytical Chemistry* **2020**, *92*, 267-291.

(22) Pohlentz, G.; Marx, K.; Mormann, M. *Methods in molecular biology* **2016**, *1394*, 163-179.

(23) Zhang, Y.; Xie, X.; Zhao, X.; Tian, F.; Lv, J.; Ying, W.; Qian, X. *J Proteomics* **2018**, *170*, 14-27.

(24) Sun, N.; Wu, H.; Chen, H.; Shen, X.; Deng, C. *Chem Commun (Camb)* **2019**, *55*, 10359-10375.

(25) Ruhaak, L. R.; Xu, G.; Li, Q.; Goonatilleke, E.; Lebrilla, C. B. *Chem Rev* **2018**, *118*, 7886-7930.

(26) Zhang, Y.; Mao, Y.; Zhao, W.; Su, T.; Zhong, Y.; Fu, L.; Zhu, J.; Cheng, J.; Yang, H. *Journal of Proteome Research* **2020**, *19*, 655-666.

(27) Zhang, Y.; Zhao, W.; Zhao, Y.; Mao, Y.; Su, T.; Zhong, Y.; Wang, S.; Zhai, R.; Cheng, J.; Fang, X.; Zhu, J.; Yang, H. *J Proteome Res* **2019**.

(28) Yuan, Y.; Cao, D.; Zhang, Y.; Ma, J.; Qi, J.; Wang, Q.; Lu, G.; Wu, Y.; Yan, J.; Shi, Y.; Zhang, X.; Gao, G. F. *Nat Commun* **2017**, *8*, 15092.

(29) Zhang, S.; Cao, X.; Liu, C.; Li, W.; Zeng, W.; Li, B.; Chi, H.; Liu, M.; Qin, X.; Tang, L.; Yan, G.; Ge, Z.; Liu, Y.; Gao, Q.; Lu, H. *Molecular & cellular proteomics : MCP* **2019**, *18*, 2262-2272.

(30) Wang, Y.; Xu, F.; Xiao, K.; Chen, Y.; Tian, Z. *Chem Commun (Camb)* **2019**, *55*, 7934-7937.

(31) Riley, N. M.; Hebert, A. S.; Westphall, M. S.; Coon, J. J. *Nat Commun* **2019**, *10*, 1311.

(32) Sun, S.; Hu, Y.; Jia, L.; Eshghi, S. T.; Liu, Y.; Shah, P.; Zhang, H. *Anal Chem* **2018**, *90*, 6292-6299.

(33) Li, D.; von Schaewen, M.; Wang, X.; Tao, W.; Zhang, Y.; Li, L.; Heller, B.; Hrebikova, G.; Deng, Q.; Ploss, A.; Zhong, J.; Huang, Z. *J Virol* **2016**, *90*, 10486-10498.

(34) Urbanowicz, R. A.; Wang, R.; Schiel, J. E.; Keck, Z. Y.; Kerzic, M. C.; Lau, P.; Rangarajan, S.; Garagusi, K. J.; Tan, L.; Guest, J. D.; Ball, J. K.; Pierce, B. G.; Mariuzza, R. A.; Foung, S. K. H.; Fuerst, T. R. *J Virol* **2019**, *93*.

(35) Ozdilek, A.; Paschall, A. V.; Dookwah, M.; Tiemeyer, M.; Avci, F. Y. *Proc Natl Acad Sci U S A* **2020**, *117*, 1280-1282.

(36) Patnaik, S. K.; Potvin, B.; Carlsson, S.; Sturm, D.; Leffler, H.; Stanley, P. *Glycobiology* **2006**, *16*, 305-317.

(37) Robinson, B. S.; Arthur, C. M.; Evavold, B.; Roback, E.; Kamili, N. A.; Stowell, C. S.; Vallecillo-Zuniga, M. L.; Van Ry, P. M.; Dias-Baruffi, M.; Cummings, R. D.; Stowell, S. R. *Front Immunol* **2019**, *10*, 1762.

(38) Wang, W. H.; Lin, C. Y.; Chang, M. R.; Urbina, A. N.; Assavalapsakul, W.; Thitithanyanont, A.; Chen, Y. H.; Liu, F. T.; Wang, S. F. *J Microbiol Immunol Infect* **2019**.

Figures and Legends

Figure 1. Workflow for site-specific N-glycosylation characterization of recombinant SARS-CoV-2 S proteins using two complementary proteases for digestion and simultaneous N-glycoproteomics analysis.

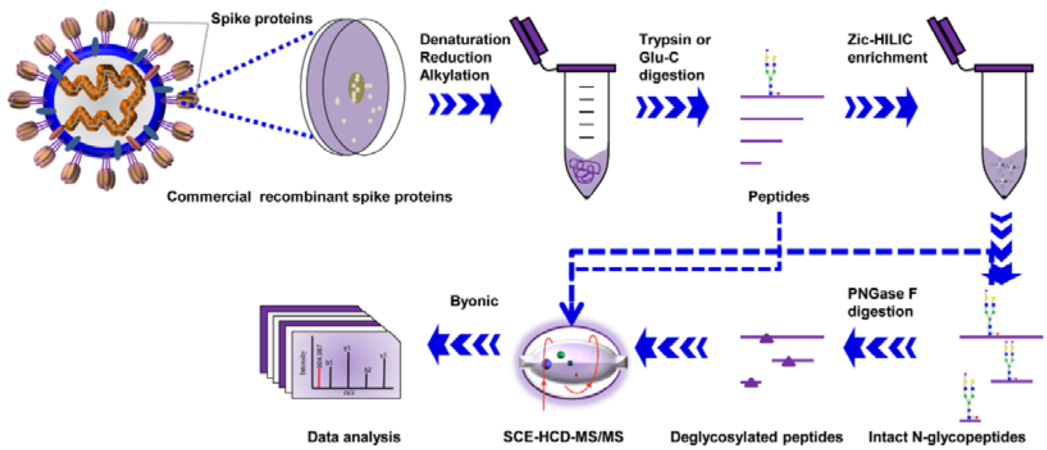


Figure 2. N-glycosites characterization of SARS-CoV-2 S proteins

(A and B) N-glycosites of recombinant SARS-CoV-2 S proteins expressed in insect cells (A) and human cells (B). PSS: putative signal sequence; RBD: receptor-binding domain; S1/S2: S1/S2 protease cleavage site; Oval: potential N-glycosite; Yellow oval: ambiguously assigned N-glycosite; Red oval: unambiguously assigned N-glycosite; Blue arrow: unambiguously assigned N-glycosite using trypsin digestion; Green arrow: unambiguously assigned N-glycosite using Glu-C digestion; Black arrow: retrieved N-glycosite using the combination of trypsin and Glu-C digestion. (C) Surface demonstration of the SARS-CoV-2 S protein (PDB code: 6VSB) ectodomain trimers with RBDs (yellow) and N-glycosites (blue)

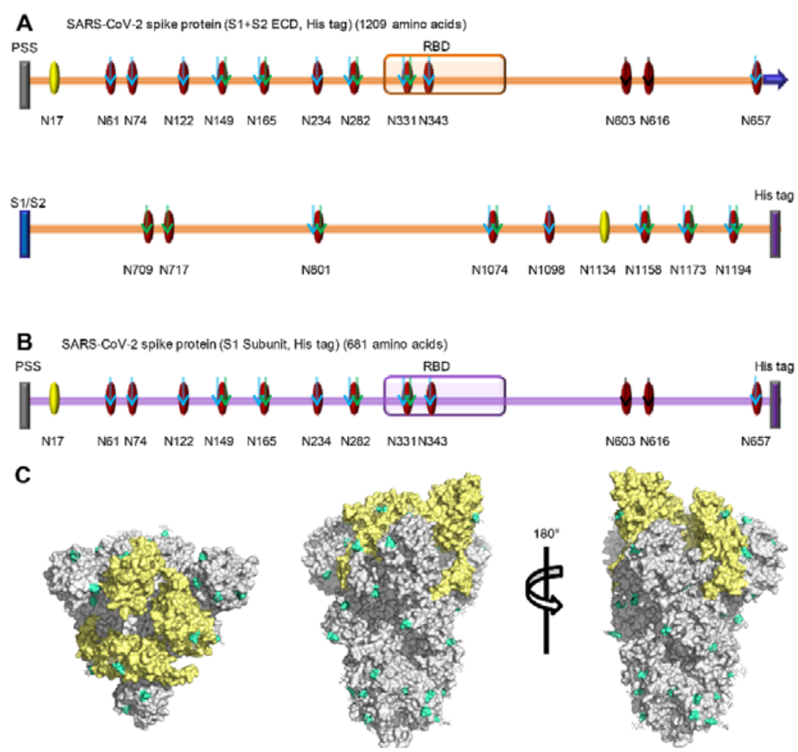


Figure 3. Site-specific N-glycosylation of recombinant SARS-CoV-2 S proteins.

The number of intact N-glycopeptides and N-glycans in recombinant SARS-CoV-2 S proteins expressed in insect cells (A) or human cells (B). Top 10 N-glycans and their putative structures on the intact N-glycopeptides from recombinant SARS-CoV-2 S proteins expressed in insect cells (C) and human cells (D). Different types and numbers of N-glycans on each N-glycosite of recombinant SARS-CoV-2 S protein or subunit expressed in insect cells (E) or human cells (F).

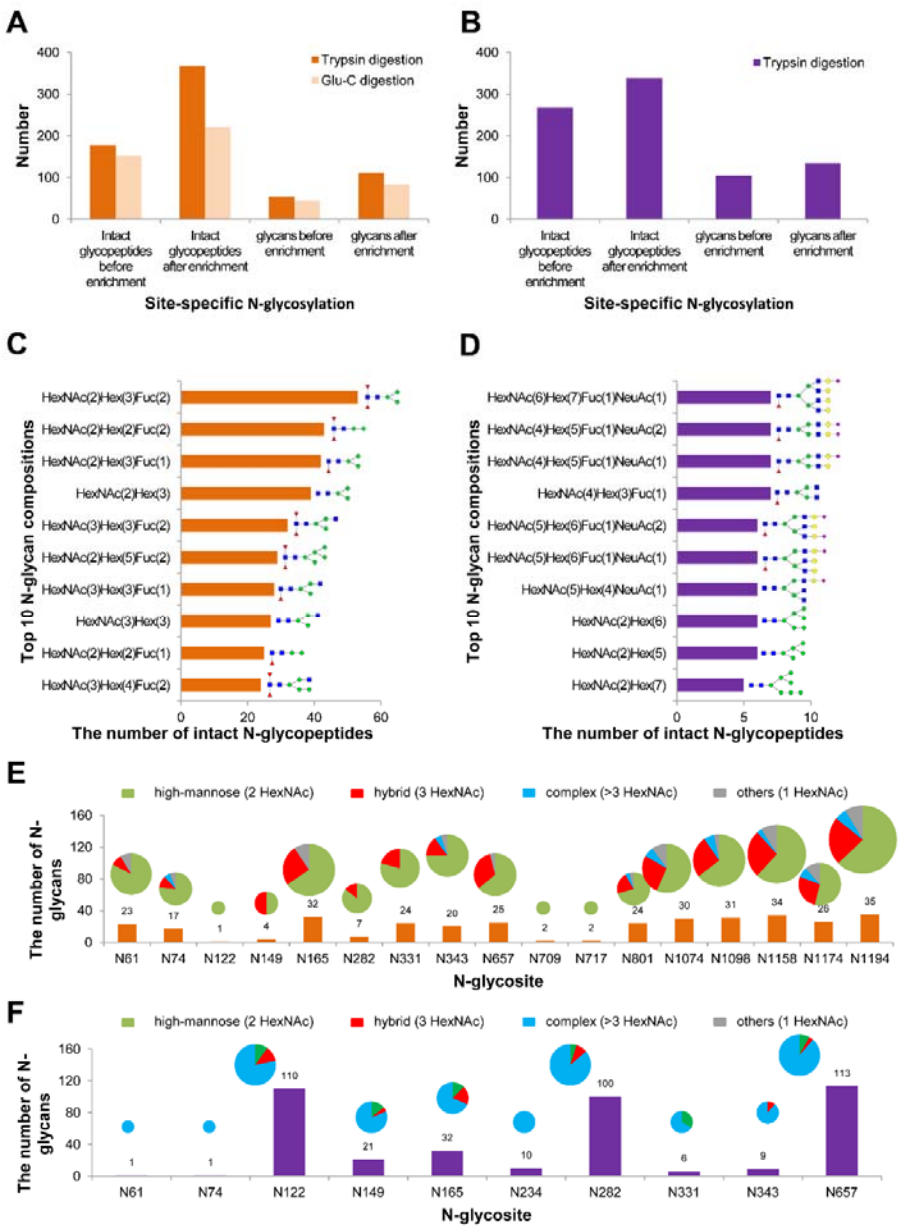
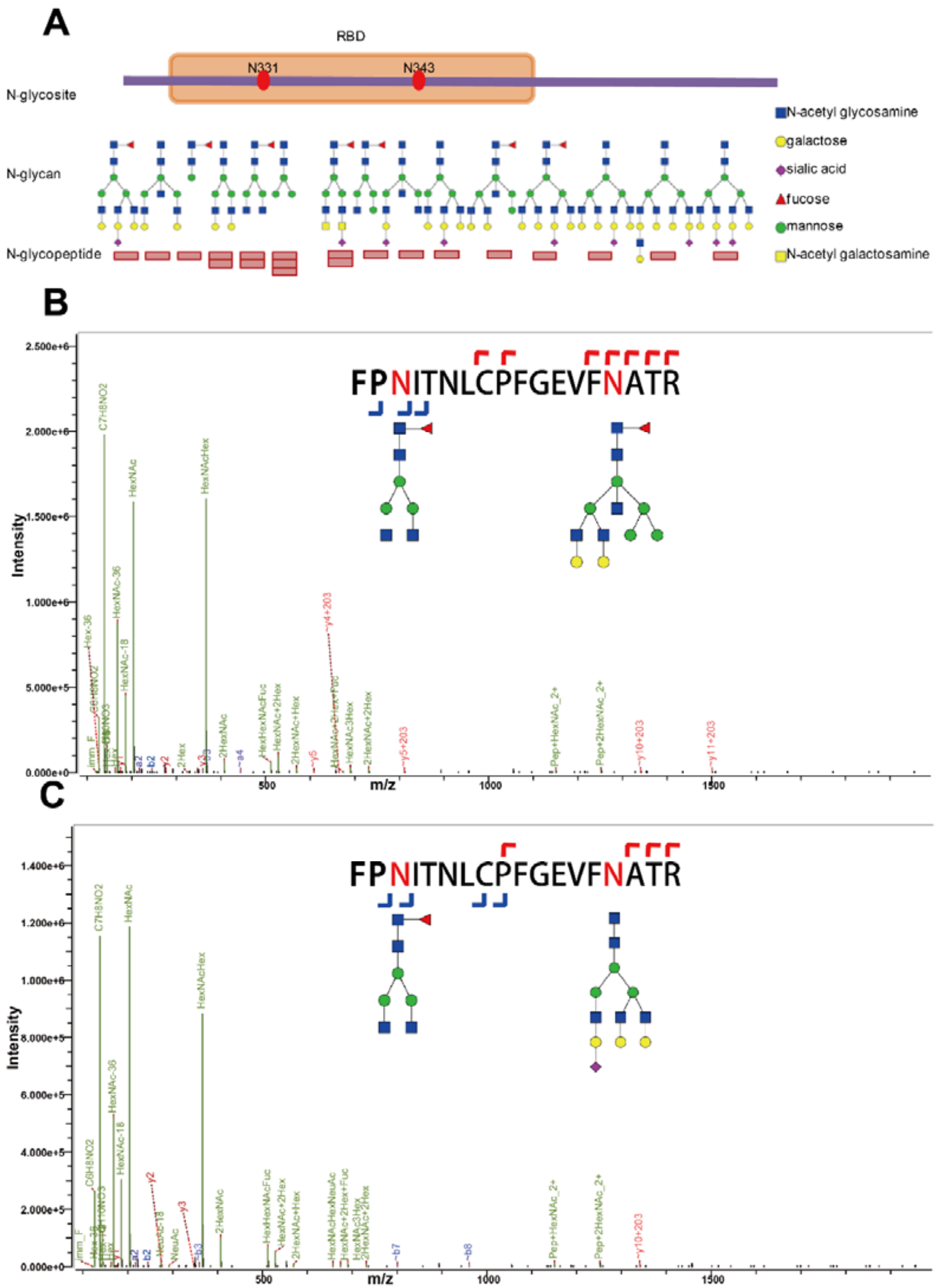


Figure 4. Site-specific N-glycosylation profile of RBD. (A) The putative N-glycans on N331 and N343 of RBD. (B and C) Representative spectra of the intact N-glycopeptides from recombinant SARS-CoV-2 S protein expressed in human cells.



Supplementary Figures:

Supplementary Figure S1. Spectra of intact N-glycopeptides with ambiguously assigned N-glycosites

Supplementary Figure S2. Comparison of the spectra of intact N-glycopeptides before (A) and after (B) enrichment

Supplementary Figure S3. Comparison of the spectra of intact N-glycopeptides from trypsin digestion (A) and Glu-C digestion (B)

Supplementary Figure S4. Comparison of the spectra of intact N-glycopeptides (A) and deglycopeptides (B)

Supplementary Figure S5. Amino acid sequence alignment of recombinant SARS-CoV-2 spike proteins expressed in insect cells (A) and human cells (B). Yellow background: putative sequence containing signal sequence; Red: potential N-glycosites; Red bold: identified N-glycosites; Green: theoretical cleavage sites of trypsin; Blue: theoretical cleavage sites of Glu-C

Supplementary Figure S6. Representative and high-quality spectra of intact N-glycopeptides and deglycosylated peptides

Supplementary Figure S7. Microheterogeneity and macroheterogeneity of the N-linked glycopeptides of S protein