



# In-solution buffer-free digestion allows full-sequence coverage and complete characterization of post-translational modifications of the receptor-binding domain of SARS-CoV-2 in a single ESI–MS spectrum

Luis Ariel Espinosa<sup>1</sup> · Yassel Ramos<sup>1</sup> · Ivan Andújar<sup>1</sup> · Enso Onill Torres<sup>1</sup> · Gleysin Cabrera<sup>1</sup> · Alejandro Martín<sup>1</sup> · Diamilé Roche<sup>1</sup> · Glay China<sup>1</sup> · Mónica Becquet<sup>1</sup> · Isabel González<sup>1</sup> · Camila Canaán-Haden<sup>1</sup> · Elías Nelson<sup>1</sup> · Gertrudis Rojas<sup>2</sup> · Beatriz Pérez-Massón<sup>2</sup> · Dayana Pérez-Martínez<sup>2</sup> · Tamy Boggiano<sup>2</sup> · Julio Palacio<sup>2</sup> · Sum Lai Lozada Chang<sup>2</sup> · Lourdes Hernández<sup>2</sup> · Kathya Rashida de la Luz Hernández<sup>2</sup> · Saloheimo Markku<sup>3</sup> · Marika Vitikainen<sup>3</sup> · Yury Valdés-Balbín<sup>4</sup> · Darielys Santana-Medero<sup>4</sup> · Daniel G. Rivera<sup>5</sup> · Vicente Vérez-Bencomo<sup>4</sup> · Mark Emalfarb<sup>6</sup> · Ronen Tchelet<sup>6</sup> · Gerardo Guillén<sup>1</sup> · Miladys Limonta<sup>1</sup> · Eulogio Pimentel<sup>1</sup> · Marta Ayala<sup>1</sup> · Vladimir Besada<sup>1</sup> · Luis Javier González<sup>1</sup>

Received: 15 June 2021 / Revised: 16 September 2021 / Accepted: 5 October 2021 / Published online: 5 November 2021  
© Springer-Verlag GmbH Germany, part of Springer Nature 2021

## Abstract

Subunit vaccines based on the receptor-binding domain (RBD) of the spike protein of SARS-CoV-2 provide one of the most promising strategies to fight the COVID-19 pandemic. The detailed characterization of the protein primary structure by mass spectrometry (MS) is mandatory, as described in ICHQ6B guidelines. In this work, several recombinant RBD proteins produced in five expression systems were characterized using a non-conventional protocol known as in-solution buffer-free digestion (BFD). In a single ESI–MS spectrum, BFD allowed very high sequence coverage ( $\geq 99\%$ ) and the detection of highly hydrophilic regions, including very short and hydrophilic peptides (2–8 amino acids), and the His<sub>6</sub>-tagged C-terminal peptide carrying several post-translational modifications at Cys<sup>538</sup> such as cysteinylolation, homocysteinylolation, glutathionylation, truncated glutathionylation, and cyanylation, among others. The analysis using the conventional digestion protocol allowed lower sequence coverage (80–90%) and did not detect peptides carrying most of the above-mentioned PTMs. The two C-terminal peptides of a dimer [RBD<sub>(319–541)</sub>-(His)<sub>6</sub>]<sub>2</sub> linked by an intermolecular disulfide bond (Cys<sub>538</sub>-Cys<sub>538</sub>) with twelve histidine residues were only detected by BFD. This protocol allows the detection of the four disulfide bonds present in the native RBD, low-abundance scrambling variants, free cysteine residues, O-glycoforms, and incomplete processing of the N-terminal end, if present. Artifacts generated by the in-solution BFD protocol were also characterized. BFD can be easily implemented; it has been applied to the characterization of the active pharmaceutical ingredient of two RBD-based vaccines, and we foresee that it can be also helpful to the characterization of mutated RBDs.

**Keywords** Buffer-free digestion · RBD · SARS-CoV-2 · Modified cysteine · Hydrophilic peptides

✉ Luis Javier González  
luis.javier@cigb.edu.cu

<sup>1</sup> Center for Genetic Engineering and Biotechnology, Ave 31, e/ 158 y 190, Cubanacán, Playa, Havana, Cuba

<sup>2</sup> Center of Molecular Immunology, 216 St., P.O. Box 16040, Havana, Cuba

<sup>3</sup> VTT Technical Research Centre of Finland Ltd, P.O. Box 1000, 02044 VTT Espoo, Finland

<sup>4</sup> Finlay Vaccine Institute, 200 and 21 Street, 11600 Havana, Cuba

<sup>5</sup> Laboratory of Synthetic and Biomolecular Chemistry, Faculty of Chemistry, University of Havana, Zapata & G, 10400 Havana, Cuba

<sup>6</sup> Dyadic International, Inc, 140 Intercoastal Pointe Drive, Suite #404, Jupiter, FL 33477, USA

## Introduction

The development of effective vaccines as well as the universal access for their massive introduction is urgently needed to control the COVID-19 pandemic [1]. Nowadays, there are several vaccine platforms being evaluated according to the draft landscape published by the World Health Organization (WHO), including inactivated and live attenuated virus; non-replicating and replicating viral vectors; DNA-, mRNA-, and virus-like particles; and protein subunit vaccines [2]. Some of them have already been approved by the WHO and regulatory authorities and introduced with favorable results in the clinic [3].

SARS-CoV-2 uses the receptor-binding domain (RBD) of the spike (S) protein for entry into the host cells [4, 5]. The RBD has been proposed for the rational development of protective vaccines against SARS-CoV-2 [6, 7] and nowadays, subunit vaccines are well-represented among the candidates investigated in preclinical studies and clinical trials [2]. For a successful introduction of vaccines, the immunogens need to be produced at scale and prices affordable for all, including middle- and low-income countries [1, 8].

Probably this is one of the reasons why RBD of SARS-CoV-2, besides its production in mammalian cells [9], has also been produced in several systems [10–14], including bacteria [15], despite the challenges of expressing a non-globular protein with four disulfide bonds and the requirement of the *N*-glycosylation for its proper expression and folding [11].

According to the test procedures and acceptance criteria for Biotechnological/Biological products (ICHQ6B guidelines [16]), mass spectrometry (MS) is the analytical tool of choice for the verification of the amino acid sequence, to demonstrate the integrity of the *N*- and *C*-terminal ends, and to detect post-translational modifications (PTMs) in natural and recombinant proteins. The PTMs may modify the physico-chemical and immunological properties of the proteins. In particular, a disulfide bond arrangement identical to the one present in the native protein is mandatory for biotherapeutics as well as for vaccine development in cases where the antigen should be well folded to raise conformational and topological neutralizing antibodies [3, 17].

Sample processing prior to MS analysis also plays a determinant role in the quality of the results. An efficient proteolytic digestion and the recovery of the proteolytic peptides are mandatory to obtain the highest sequence coverage and mapping all PTMs present in the analyzed molecule. In particular, if electrospray ionization mass spectrometry (ESI-MS) is used, a desalting step is

needed to ionize properly the proteolytic peptides. This step, although necessary, often comprises the recovery of highly hydrophilic and hydrophobic peptides when micro-columns based on reverse phase chromatography are used.

Arbeitman et al. [11] analyzed by MALDI-MS the in-solution tryptic digests of two reduced and S-alkylated recombinant RBD of SARS-CoV-2. The tryptic peptides, desalted by C18-ZipTips prior to MALDI-MS analysis, allowed the unambiguous identification of RBD expressed in *P. pastoris* and HEK-293 T cells, but with a sequence coverage of only 40 and 60%, respectively.

The hydrophilic *C*-terminal peptide (LPETGHHHHHH) tagged with a repeat of six histidine residues (His<sub>6</sub> tag) was only detected for the RBD expressed in *P. pastoris*, suggesting variable results in the desalting step. Other PTMs such as *N*-, and *O*-glycosylation were not detected in this study [11]. In this manuscript, the arrangement of disulfide bonds and the presence of free cysteine residues were not verified. Free cysteine residues, even present as low-abundance species, may promote disulfide exchange and generate scrambling variants of proteins [18].

In our laboratory, we initially demonstrated that proteins separated by SDS-PAGE can be efficiently in-gel desalted and digested in water with trypsin in absence of traditional saline buffers [19]. This procedure avoids a desalting step of the proteolytic peptides and allows their direct analysis by ESI-MS, and the sequence coverage [19] was higher than what is achieved by the traditional in-gel digestion protocol.

Recently, the principles of the in-gel buffer-free digestion protocol [19] were extended to in-solution buffer-free digestion (BFD) of other proteins [20]. In-solution BFD protocol improved the sequence coverage of certain regions of proteins represented by short and hydrophilic peptides including some *N*-glycopeptides, short peptides linked by disulfide bonds, and hydrophilic His<sub>6</sub> tag *C*-terminal peptides [20].

In this work, we adapted the in-solution BFD protocol [20] to the analysis of the products of six SARS-CoV-2 RBD expression constructs from five different expression systems. The implemented in-solution BFD method avoids buffers and desalting is carried out by protein precipitation, allowing very high sequence coverage ( $\geq 99\%$ ) and the detection of PTMs including those located at the *N*- and the *C*-terminal end. The in-solution BFD protocol allowed the identification, in a single mass spectrum, of the four native disulfide bonds as well as scrambled disulfide bonds, the presence of free cysteine residues, *N*- and *O*-glycosylation, and other PTMs of known and unknown nature linked to an unpaired cysteine residue located at the *C*-terminal peptide in some of the analyzed RBD molecules. A non-peer-reviewed preprint version of this article was posted in bioRxiv [21].

**Table 1** Sequences of the recombinant receptor-binding domain of SARS-CoV-2 characterized in this work

Code <sup>a)</sup>	Expression system	Amino acid sequence <sup>b)</sup>
<i>RBD</i> <sub>(319-541)</sub> - <i>HEK_A3</i>	<i>HEK-293T</i>	<sup>319</sup> <b>RVQPTESIVRFPNITNL</b> <i>CPFGEVFNATRFASVYAWNRKRISN</i> <b>CVADYSVLYN</b> <b>SASFSTFKCYGVSPTKLNDL</b> <i>CF</i> <b>TNVYADSFVIRGDEVRQIAPGQTGKIADYNYK</b> <b>LPDDFTG</b> <i>CVIAWNSNNLDSKVGGNYNLYRLFRKSNLKP</i> <b>FERDISTEIQAGST</b> <b>PCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVLSFELLHAPATVCGPKKSTN</b> <b>LVKNKCVNF</b> <sup>541</sup> - <i>AAAHHHHHH</i>
<i>RBD</i> <sub>(319-541)</sub> - <i>HEK</i>	<i>HEK-293T</i>	<sup>319</sup> <b>RVQPTESIVRFPNITNL</b> <i>CPFGEVFNATRFASVYAWNRKRISN</i> <b>CVADYSVLYN</b> <b>SASFSTFKCYGVSPTKLNDL</b> <i>CF</i> <b>TNVYADSFVIRGDEVRQIAPGQTGKIADYNYK</b> <b>LPDDFTG</b> <i>CVIAWNSNNLDSKVGGNYNLYRLFRKSNLKP</i> <b>FERDISTEIQAGST</b> <b>PCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVLSFELLHAPATVCGPKKSTN</b> <b>LVKNKCVNF</b> <sup>541</sup> - <i>HHHHHH</i>
<i>(RBD</i> <sub>(319-541)</sub> - <i>CHO</i> ) <sub>2</sub> <sup>c)</sup>	<i>CHO-K1</i>	( <sup>319</sup> <b>RVQPTESIVRFPNITNL</b> <i>CPFGEVFNATRFASVYAWNRKRISN</i> <b>CVADYSVLY</b> <b>NSASFSTFKCYGVSPTKLNDL</b> <i>CF</i> <b>TNVYADSFVIRGDEVRQIAPGQTGKIADYNY</b> <b>KL</b> <i>PDDFTG</i> <b>CVIAWNSNNLDSKVGGNYNLYRLFRKSNLKP</b> <b>FERDISTEIQAGS</b> <b>TP</b> <i>CNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVLSFELLHAPATVCGPKKST</i> <b>NLVKNKCVNF</b> <sup>541</sup> - <i>HHHHHH</i> ) <sub>2</sub>
<i>RBD</i> <sub>(331-529)</sub> - <i>Ec</i>	<i>E. coli</i>	<i>GSSHSHHHHSSGLVPRGSHMAS</i> - <sup>331</sup> <b>NITNL</b> <i>CPFGEVFNATRFASVYAWNRKRI</i> <b>SNCVADYSVLYNSASFSTFKCYGVSPTKLNDL</b> <i>CF</i> <b>TNVYADSFVIRGDEVRQIAP</b> <b>GQTGKIADYNYKL</b> <i>PDDFTG</i> <b>CVIAWNSNNLDSKVGGNYNLYRLFRKSNLKP</b> <b>FER</b> <b>DISTEIQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVLSFELLHA</b> <b>PATVCGPKK</b> <sup>529</sup>
<i>RBD</i> <sub>(333-527)</sub> - <i>C1</i>	<i>T. heterothallica</i> <i>C1</i>	<sup>333</sup> <b>TNL</b> <i>CPFGEVFNATRFASVYAWNRKRI</i> <b>SNCVADYSVLYNSASFSTFKCYGVS</b> <b>PTKLNDL</b> <i>CF</i> <b>TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKL</b> <i>PDDFTG</i> <b>CVIAWNS</b> <b>NNLDSKVGGNYNLYRLFRKSNLKP</b> <b>FERDISTEIQAGSTPCNGVEGFNCYFPL</b> <b>QSYGFQPTNGVGYQPYRVVLSFELLHAPATVCGP</b> <sup>527</sup> - <i>GGGGSEPEA</i>
<i>RBD</i> <sub>(331-530)</sub> - <i>cmcy-Pp</i>	<i>P. pastoris</i>	<i>EFS</i> - <sup>331</sup> <b>NITNL</b> <i>CPFGEVFNATRFASVYAWNRKRISN</i> <b>CVADYSVLYNSASFSTFK</b> <b>CYGVSPTKLNDL</b> <i>CF</i> <b>TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKL</b> <i>PDDFTG</i> <b>CV</b> <b>VIAWNSNNLDSKVGGNYNLYRLFRKSNLKP</b> <b>FERDISTEIQAGSTPCNGVEGF</b> <b>NCYFPLQSYGFQPTNGVGYQPYRVVLSFELLHAPATVCGPKKS</b> <sup>530</sup> - <i>REQQLISEEDLNSAVDHHHHHH</i>

<sup>a)</sup>The numbers between parentheses correspond to the amino acid positions of the RBD of SARS-CoV-2 (UniprotKB access: P0DTC2)

<sup>b)</sup>The sequences written in bold correspond to the cloned regions of the RBD of SARS-CoV-2. Sequences written in italics indicate other sequence segments not related to RBD but added to the N- and/or C-terminal end of the protein during the cloning strategy. Cysteines are highlighted in red

<sup>c)</sup>Two molecules of *RBD*<sub>(319-541)</sub>-*CHO* are linked by an intermolecular disulfide bond between Cys<sup>538</sup>-Cys<sup>538</sup>

## Materials and methods

### Cloning expression and purification of RBD variants

Six RBD recombinant proteins, produced at laboratory scale in a wide range of host cells, were used as model antigens to develop and refine suitable analytical methods for RBD characterization. Table 1 summarizes their sequences. A more detailed description of the procedures for cloning, expression, and purification of these proteins is provided in the [Electronic Supplemental File](#) (see Experimental Section in ESM).

### In-solution buffer-free digestion protocol

Fifty micrograms of the glycoproteins, dissolved in PBS (pH 7.4) containing 0.5 M guanidine hydrochloride, was reacted with 5 mM *N*-ethylmaleimide (NEM) for 30 min at room temperature (22 °C). Then, 1 µL of PNGase F (New England Biolabs) was added and the deglycosylation reaction was allowed to proceed for 2 h at 37 °C. In the case of *N*-glycosylated *RBD*<sub>(333–527)</sub>-*CI*, the protein was not deglycosylated but reduced with 10 mM dithiothreitol and 0.2 M Tris–HCl buffer pH 8.0 for 1 h at 37 °C, and then *S*-alkylated with 25 mM iodoacetamide under exclusion of light for 20 min at 22 °C. All samples were cooled at room temperature and proteins were precipitated with ten volumes of cold acetone (–20 °C) or 80% ethanol (v/v) and the solution was kept at –80 ± 5 °C for 1 h. The sample was centrifuged at 9000×*G* during 5 min and the supernatant was discarded. The precipitate was washed by vortexing with 75% cold acetone or ethanol (–20 °C), centrifuged at 10,000 rpm during 5 min and the supernatant was discarded. This procedure was repeated twice and the final precipitate was dried up in a vacuum centrifuge during 15 min. The precipitate was dissolved in 50 µL of 20% (v/v) acetonitrile in water solution with 1 min vortexing and 10 min sonication in a water bath. One microgram of sequencing grade trypsin (Promega) dissolved in water was added to the protein solution and the specific proteolytic digestion proceeded for 16 h at 37 °C in a thermomixer (Thermo Fisher Scientific). Digestion was centrifuged at 9000×*G* for 1 min and 4 µL of the resultant mixture of tryptic peptides was mixed with 0.3 µL of 90% formic acid and it was loaded into a metal-coated borosilicate nanocapillary for MS analysis.

### Standard digestion (SD) protocol

Fifty micrograms of the protein dissolved in PBS (pH 7.2) containing 0.5 M guanidine hydrochloride reacted with 5 mM NEM during 30 min at room temperature (22 °C).

One microliter of PNGase F (New England Biolabs) was added and the deglycosylation reaction proceeded for 2 h at 37 °C. The sample was fourfold diluted and the protein digested in the presence of 0.2 M Tris–HCl buffer pH 8.0 and 1 µg of sequencing grade trypsin (Promega) previously dissolved in 20 mM acetic acid. Tryptic digestion proceeded for 16 h at 37 °C and digestion was stopped by adding formic acid to final concentration of 5% (v/v). The resulting peptides were desalted with ZipTip C18 (Millipore, USA), washed with 0.2% (v/v) formic acid solution, and eluted in 4 µL of 60% acetonitrile in water containing 0.2% formic acid (v/v).

### Electrospray ionization mass spectrometry analysis

For measuring the molecular masses of the deglycosylated RBDs and the *N*-glycosylated *RBD*<sub>(333–527)</sub>-*CI*, 7 µg of the total protein was mixed with equal volume of 6 M guanidine hydrochloride solution and desalted by using ZipTip C18 (Millipore, USA). The proteins were extensively washed with 0.2% (v/v) formic acid solution and finally eluted in 3 µL of 60% acetonitrile in water containing 0.2% formic acid (v/v). The elution was loaded into the metal-coated nanocapillary for ESI–MS analysis.

The mixture of tryptic peptides contained in 4 µL of the 20% acetonitrile hydrolysis solution was acidified by adding 0.5 µL of formic acid (90% v/v) and directly analyzed in a hybrid orthogonal QToF-2 tandem mass spectrometer (Micromass, Manchester, UK) by spraying the sample into the ion source using 1200 and 35 V for the capillary and the entrance cone, respectively. The ESI–MS were acquired from *m/z* 200–2000 and the multiply-charged ions were manually fragmented by collision-induced dissociation using appropriated collision energies (20–50 eV) to obtain sufficient structural information in the MS/MS spectra. Argon was used as a collision gas and the mass spectra were processed by using MassLynx v4.1 (Micromass, UK). The ESI–MS/MS of tryptic peptides with *z* ≥ 3+ were deconvoluted by MaxEnt 3.0. The ESI–MS spectrum (*m/z* 400–3000) of the protein deglycosylated with PNGase F was deconvoluted (mass 5000–70,000) by using the MaxEnt1.0 tool (Micromass, UK). The theoretical *m/z* for tryptic peptides as well as for the intact protein was calculated by using the peptide and protein editor available in the MassLynx v4.1 software (Micromass, UK).

### SDS-PAGE analysis

*RBD*<sub>(319–541)</sub>-*HEK\_A3*, (*RBD*<sub>(319–541)</sub>-*CHO*)<sub>2</sub>, and *RBD*<sub>(331–529)</sub>-*Ec* proteins were separated by SDS-PAGE as described by Laemmli [22], under reducing and non-reducing conditions. Two micrograms of *N*-glycosylated and deglycosylated proteins were applied in a 12.5%T, 3%C

acrylamide-bisacrylamide separating gel at 30 mA/gel until the tracking dye left the gel. Proteins were detected by silver staining [23] or Coomassie Brilliant Blue G-250; gel images were analyzed with a GS-900 calibrated imaging densitometer (Bio-Rad) and processed with Image Lab v6.0 software (Bio-Rad).

## NP-HPLC analysis

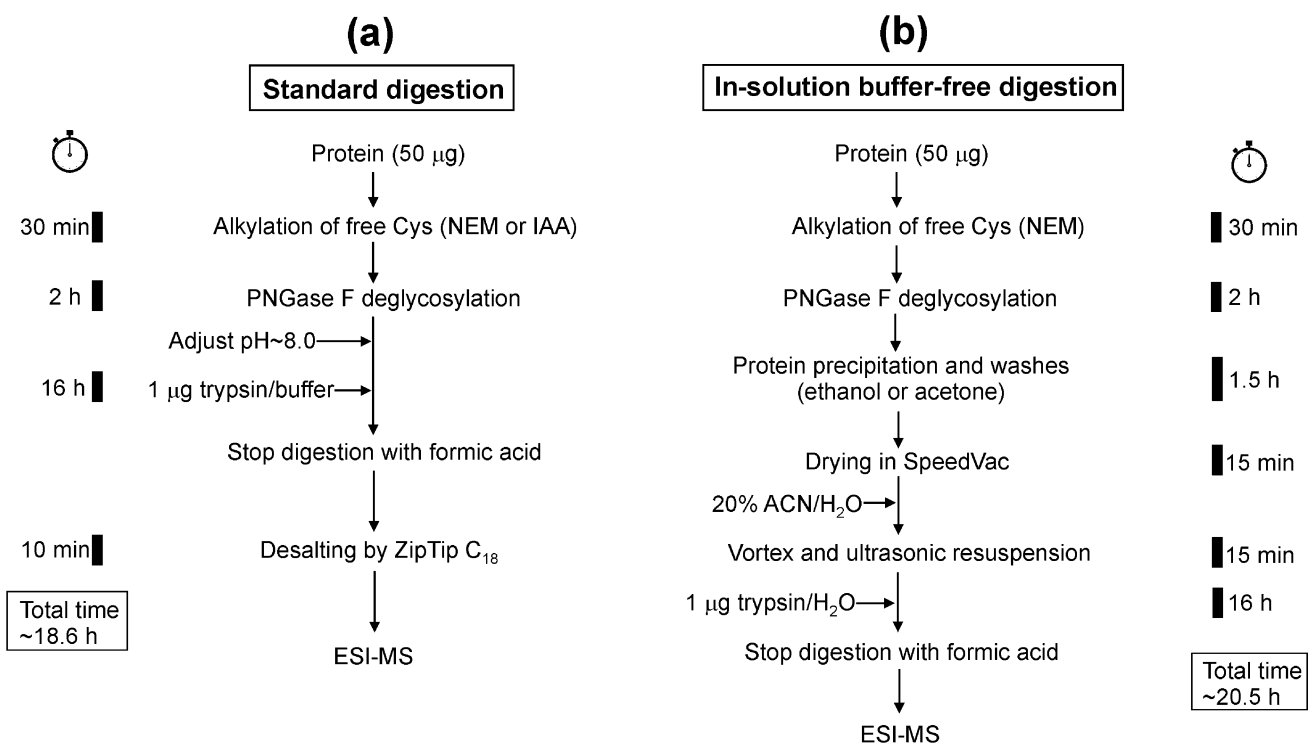
*N*-glycosylation profile was determined by using the procedure described by Guile et al. [24]. Briefly, the *N*-glycans released by PNGase F treatment were derivatized with 2-amino benzamide (2AB) by reductive amination. The chromatographic separation was carried out in an HPLC Prominence-Shimadzu (Japan) using a linear gradient from 20 to 53% of 50 mM, pH 4.4 ammonium formate (solution A), and pure acetonitrile (solution B). 2AB *N*-glycan separation was performed on an Amide-80 column (TSK-gel 250 × 46 mm, 5 μm, Tosohaas, Japan) and the derivatized oligosaccharides were detected on-line by fluorescence using an excitation and detection wavelengths of 330 nm and 420 nm, respectively. The structural assignment was performed by comparing the experimental GU values with the GlycoStore database (<https://glycostore.org/>). GU values were calculated from the retention time of each peak using

as a reference an HPLC separation ran under similar conditions for the 2AB derivatives of a dextran ladder generated by acid partial hydrolysis. Glycans structures were represented according to GlycoStore nomenclature.

## Results and discussion

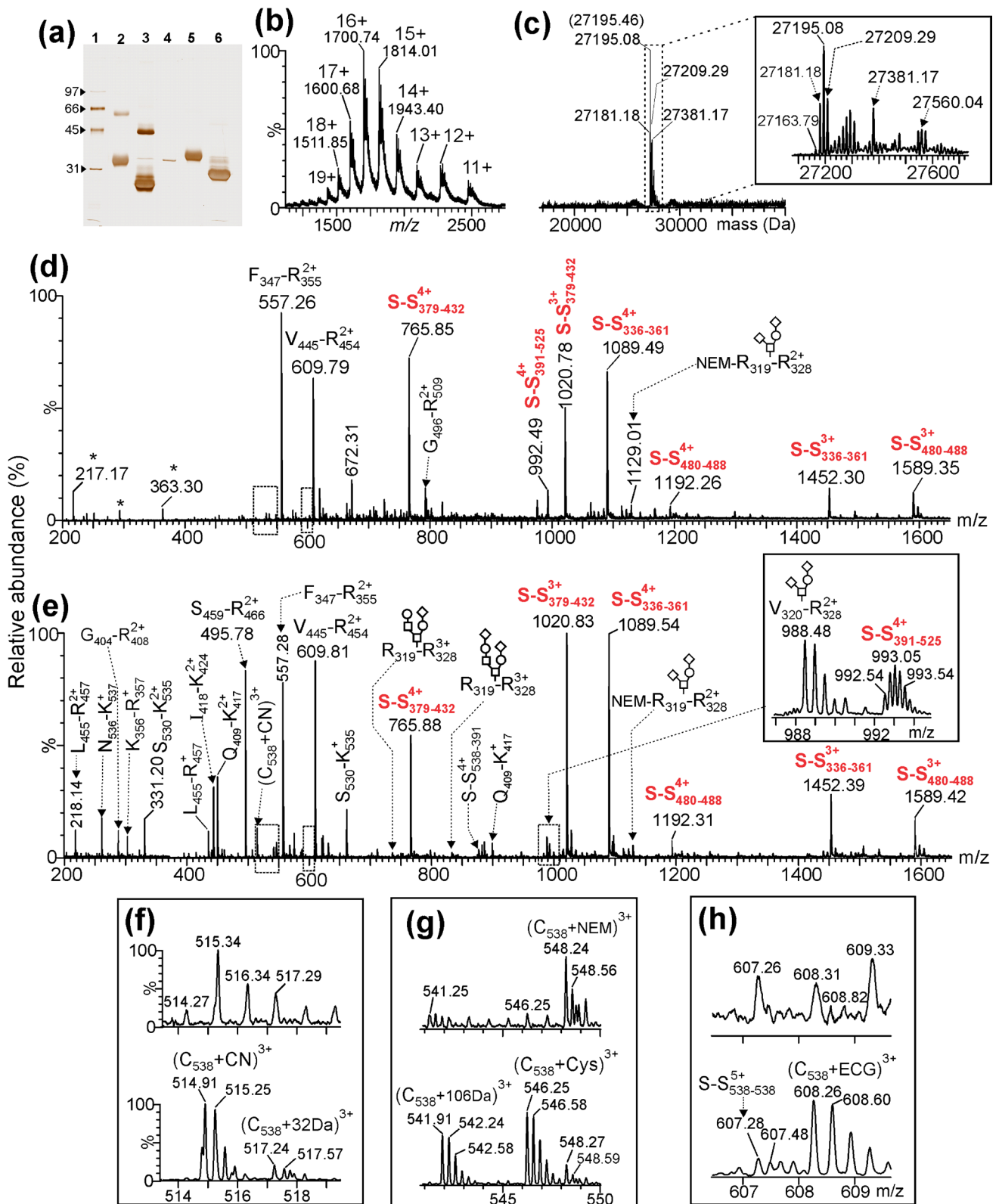
### Comparison between the standard digestion and the in-solution buffer-free digestion protocols

Both the SD (Fig. 1a) and the in-solution BFD [20] (Fig. 1b) protocols start with the S-alkylation of free cysteine residues by adding an excess of *N*-ethylmaleimide (NEM) or iodoacetamide (IAA). This step blocks the free thiol groups that can be present either because the RBD contains an odd number of cysteine residues or these groups were not quantitatively linked and thus remain partially free by a non-correct folding. At the same time, the alkylating agent added at the beginning of the protocol avoids artifacts due to the disulfide bond exchange during the subsequent steps [18]. This could be more critical in the conventional protocol using a basic pH during tryptic digestion [25, 26]. The use of a slightly acidic pH for trypsin digestion (pH 5.5–6.0) with BFD minimizes



**Fig. 1** A comparison between the in-solution standard digestion (a) and buffer-free digestion [20] (b) protocols for the ESI-MS analysis of the tryptic digests. Black rectangles at the left and right sides of the figure indicate the time required for the individual steps in each

protocol. Square boxes at the bottom-left and bottom-right in the figure indicate the total time consumed for each protocol. NEM and IAA mean *N*-ethylmaleimide and iodoacetamide, respectively



artificial modifications introduced during sample preparation such as scrambling due to the presence of free Cys in the analyzed protein. The S-alkylating agent introduces

an artificial mass tag that facilitates the assignment when any Cys is partially free and differentiates them from species modified with natural thiol-blocking groups due to

**Fig. 2 a** SDS-PAGE analysis under reducing and non-reducing conditions of *N*-glycosylated and deglycosylated *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub> and detected with silver staining. Lane 1: Molecular weight markers of low-range from 31 to 97 kDa (Bio-Rad). Lanes 2–3: *N*-glycosylated and deglycosylated protein in non-reducing conditions detecting the monomer and a low-abundance (13%) dimer species of *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub>. Lane 4: Control of PNGase F used in the *N*-deglycosylation. Lanes 5–6: *N*-glycosylated and deglycosylated protein under reducing conditions. **b** ESI–MS analysis of the *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub> deglycosylated with PNGase F. **c** Resultant ESI–MS spectrum after deconvolution with MaxEnt v 1.0 software. The inset shown in (c) corresponds to the expanded ESI–MS spectrum in the range shown by a broken line rectangle. The masses between parentheses indicate the expected molecular masses of the detected species. A detailed assignment of this ESI–MS spectrum is shown in Table 2. The ESI–MS spectra shown in (d) and (e) correspond to the ESI–MS analysis of the resultant tryptic peptides of *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub> digested with trypsin following the SD and in-solution BFD (with ethanol precipitation) protocols shown in Fig. 1(a) and (b). Asterisks in (d) correspond to background signals, not assigned to tryptic peptides. The inset shown in (e) corresponds to an expanded region where the *O*-glycosylated *N*-terminal end peptide (Val<sup>320</sup>-Arg<sup>328</sup> + [HexNAc:Hex:NeuAc<sub>2</sub>]<sup>2+</sup> and two disulfide bonded peptides (assigned as S-S<sub>391–525</sub><sup>4+</sup>) were detected. Monosaccharide symbols follow the SNFG system [60] and the *O*-glycan structures as previously reported [33]. The upper and lower mass spectra shown in (f), (g), and (h) correspond to expanded regions of the ESI–MS spectra shown in (d) and (e), respectively. A detailed assignment for all tryptic peptides in this figure is summarized in Table 3

alkylating or thiol-containing species present in the culture media [27].

As a second step, both protocols comprise the deglycosylation with PNGase F of the recombinant RBDs and convert the fully glycosylated asparagines (Asn<sub>331</sub>/Asn<sub>343</sub>) into aspartic acids. This step also facilitates the detection and sequencing of two peptides (Phe<sub>329</sub>-Arg<sub>346</sub>) and (Ile<sub>358</sub>-Lys<sub>378</sub>) linked by an intermolecular disulfide bond between Cys<sub>336</sub> and Cys<sub>361</sub>. For the particular cases of *RBD*<sub>(333–527)</sub>-*CI* and *RBD*<sub>(331–530)</sub>-*cmv-c-Pp* (Table 1), the peptide with the disulfide bond Cys<sub>336</sub>-Cys<sub>361</sub> at the same time contains the *N*-terminal end of the protein. The identification of the disulfide bonds and the *N*-terminal sequencing of the protein are aspects inquired by regulatory agencies to develop well-characterized products according to the ICHQ6B guidelines [16].

For the in-solution SD protocol (Fig. 1a), the pH of the solution is adjusted at basic pH and the deglycosylated RBD is digested with trypsin during 16 h due to our interest to guarantee a complete digestion. Also note that even after disulfide reduction, this protein has been digested overnight by other authors [11, 28]. Finally, the digestion is quenched by adding formic acid and the resultant tryptic peptides are desalted by using C18-ZipTips and eluted in a solution compatible with ESI–MS analysis.

For the in-solution BFD protocol (Fig. 1b), a desalting step is achieved at the protein level by conventional precipitation protocols using either cold acetone [29] or ethanol

[30]. Here, washing steps are included to minimize inorganic ions that may provoke adduct signals in the mass spectra. Protein resuspension is guaranteed by vigorous vortex and ultrasonic bath in 20% acetonitrile, before adding trypsin previously dissolved in water. There is no appreciable difference in the two workflows (Fig. 1a and 1b) with respect to the processing time before MS analysis.

Characterization of *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub> and *RBD*<sub>(319–541)</sub>-*HEK* proteins.

*RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub> (Table 1) expressed in HEK-293 T mammalian cell line has four disulfide bonds and a free cysteine residue (Cys<sub>538</sub>) located towards the *C*-terminal region of the protein. The high reactivity of Cys<sub>538</sub> can be used for site-directed chemical conjugation to highly immunogenic carrier proteins such as tetanus toxoid [31].

*RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub> was analyzed by SDS-PAGE in non-reducing conditions (Fig. 2a, lane 2) showing an intense and diffuse band at 33.3 kDa corresponding to the monomer with the heterogeneity of *N*-glycosylation. Also, a band detected at 59.7 kDa representing approximately ~13% was assigned to the dimer. After treatment with PNGase F and analyzed under non-reducing conditions, these bands migrated at 29.3 and 43.9 kDa (Fig. 2a, lane 3) confirming that *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub> is *N*-glycosylated. The presence of *O*-glycosylation was not excluded because PNGase F does not hydrolyze *O*-glycans covalently linked to serine or threonine. When the same samples were analyzed by SDS-PAGE under reducing conditions, only protein bands corresponding to the glycosylated monomer (Fig. 2a, lane 5) and the deglycosylated monomer (Fig. 2a, lane 6) were detected. No evidence of the dimer was observed suggesting that dimerization of the molecule was mediated by disulfide bonds and was not due to an aggregation artifact.

To confirm the integrity, the *N*-deglycosylated protein was analyzed by ESI–MS (Fig. 2b) and it showed intense multiply-charged ions of the protein. The deconvoluted ESI–MS spectrum (Fig. 2c) shows the most intense signal with molecular mass of 27,195.08 Da that agreed with the expected mass (27,195.46 Da) considering the *N*-deglycosylated monomer of *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub>, cysteinylated and *O*-glycosylated with HexNAc:Hex:NeuAc<sub>2</sub> (Table 2). Other groups that also expressed RBD molecules in HEK-293 with an odd number of cysteine residues reported cysteinylation [28, 31]. *O*-glycosylation has been reported for the native RBD of SARS-Cov-2 [32, 33] as well as for several RBD versions expressed in mammalian cells [28, 31].

Also, other signals observed in Fig. 2c (see inset) and summarized in Table 2 suggest the presence of other modified species of the *N*-deglycosylated *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub>. Separately, the *N*-deglycosylated protein was digested in-solution with trypsin by using the SD (Fig. 1a) and BFD (Fig. 1b) protocols and the resultant ESI–MS spectra are shown in Fig. 2d and 2e, respectively. The sequence

**Table 2** Summary of the ESI-MS analysis for the SD and the in-solution BFD protocols and sequence coverage of RBD proteins characterized in this work

Protein	Molecular mass			Sequence assignment <sup>a)</sup>	Sequence coverage <sup>b)</sup>	
	Exp. (Da)	Theor. (Da)	Error (ppm)		SD	BFD
<i>RBD</i> <sub>(319–541)</sub> - <i>HEK</i> <sub>A3</sub>	27,163.79	-	-	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + 87 Da	82	100
	27,181.18	-	-	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + 106 Da		
	27,195.08	27,195.46	- 13.97	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + Cys		
	27,209.29	-	-	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + hCys		
	27,308.95	-	-	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + 232 Da		
	27,381.17	27,381.62	- 16.43	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + ECG		
	27,560.04	27,560.69	- 23.58	RBD + HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc <sub>2</sub> + Cys		
	27,746.39	27,746.96	- 20.54	RBD + HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc <sub>2</sub> + ECG		
<i>RBD</i> <sub>(319–541)</sub> - <i>HEK</i>	26,982.06	26,982.22	- 5.93	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + Cys	85	100
	26,995.61	-	-	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + hCys		
	27,009.65	-	-	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + 147 Da		
	27,053.73	-	-	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + 191 Da		
	27,095.12	-	-	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + 232 Da		
	27,166.19	27,168.38	- 80.6	RBD + HexNAc:Hex:NeuAc <sub>2</sub> + ECG		
	27,181.66	-	-	RBD + HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc <sub>2</sub> - 47 Da		
	27,196.27	-	-	RBD + HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc <sub>2</sub> - 32 Da		
	27,347.31	27,347.55	- 8.78	RBD + HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc <sub>2</sub> + Cys		
	27,476.17	27,476.67	- 18.19	RBD + HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc <sub>2</sub> + EC		
<i>(RBD)</i> <sub>(319–541)</sub> - <i>CHO</i> <sub>2</sub>	53,141.06	53,141.62	- 10.54	<i>(RBD + HexNAc:Hex:NeuAc)</i> <sub>2</sub>	80.6	100
	53,433.40	53,432.87	- 9.92	<i>(RBD)</i> <sub>2</sub> + HexNAc:Hex:NeuAc + HexNAc:Hex:NeuAc <sub>2</sub>		
	53,724.72	53,724.14	- 10.79	<i>(RBD + HexNAc:Hex:NeuAc)</i> <sub>2</sub>		
<i>RBD</i> <sub>(331–529)</sub> - <i>Ec</i>	25,117.44	25,117.14	- 11.94	RBD reduced and carbamidomethylated <sup>c)</sup>	-	99
<i>RBD</i> <sub>(333–527)</sub> - <i>CI</i>	22,590.33	22,590.26	- 3.09	RBD <i>N</i> -deglycosylated	-	100
	23,481.58	23,482.09	- 21.92	RBD + M3	-	100
	23,683.30	23,685.29	- 84.91	RBD + M3A1		
	23,644.03	23,644.23	- 8.45	RBD + M4		
	23,847.28	23,847.43	- 6.29	RBD + M4A1		
	24,009.12	24,009.57	- 18.74	RBD + M5A1		
	23,969.29	23,968.52	+ 32.12	RBD + M6		
	24,172.71	24,171.71	+ 41.37	RBD + M6A1		
	24,130.26	24,130.66	- 16.57	RBD + M7		
	24,333.80	24,333.86	- 2.46	RBD + M7A1		
	24,292.36	24,292.81	- 18.52	RBD + M8		
	24,495.52	24,496.00	- 19.59	RBD + M8A1		
	24,455.07	24,454.95	+ 4.90	RBD + M9		
	24,658.32	24,658.14	+ 7.29	RBD + M9A1		
	24,819.64	24,820.29	- 26.19	RBD + M10A1		
<i>RBD</i> <sub>(331–530)</sub> - <i>Pp</i>	25,835.29	25,434.41	-	RBD + 400 Da	-	99

<sup>a)</sup>HexNAc: *N*-acetyl hexosamine, Hex: hexose, SA: sialic acid, M: mannose, GlcNAc: *N*-acetylglucosamine, ECG, glutathione; Cys, cysteine; hCys, homocysteine. Glycans structures were represented according to GlycoStore nomenclature

<sup>b)</sup>Expressed in % of the sequences provided in Table 1. SD and BFD mean that the RBD molecule was characterized by in-solution SD and BFD protocols, respectively

<sup>c)</sup>Non-reduced molecular mass of *RBD*<sub>(331–529)</sub>-*Ec* was estimated by SDS-PAGE analysis and observed between the stacking and separating gel (> 97,000 Da) in Fig. 5a

assignments based on the agreement between the expected and experimental  $m/z$  of tryptic peptides are summarized in Table 3. The four disulfide bonds present in the native RBD of S protein of SARS-CoV-2 were identified by both

protocols (Fig. 2d and 2e) and confirmed by MS/MS analysis (Fig. S1a–S1d).

In the SD protocol, only the *N*-terminal peptide R<sub>319</sub>-R<sub>328</sub> containing HexNAc:Hex:NeuAc<sub>2</sub> was detected ( $m/z_{Exp}$  1066.52 and  $m/z_{Exp}$  711.36; Fig. 2d, Table 3),



**Table 3** Summary of the 100% sequence coverage assignment by ESI-MS of the tryptic digestion using the in-solution buffer-free (BFD) and 82% by the standard digestion (SD) protocol of *RBD*<sub>(319-541)</sub>-*HEK*<sub>A3</sub> expressed in HEK293T

Code <sup>b)</sup>	<i>m/z</i> <sub>Theor</sub>	<i>z</i>	<i>m/z</i> <sub>Exp</sub>		Assignment <sup>a)</sup>
			BFD	SD	
V <sub>320</sub> -R <sub>328</sub>	514.79	2	514.81	-	<sup>320</sup> VQPTESIVR <sup>328</sup>
F <sub>347</sub> -R <sub>355</sub>	557.28	2	557.28	557.26	<sup>347</sup> FASVYAWNR <sup>355</sup>
	1113.55		1113.57	1113.50	
K <sub>356</sub> -R <sub>357</sub>	303.21	1	303.22	-	<sup>356</sup> KR <sup>357</sup>
G <sub>404</sub> -R <sub>408</sub>	575.28	1	575.29	-	<sup>404</sup> GDEV <sup>R</sup> <sup>408</sup>
	288.14	2	288.15	-	
Q <sub>409</sub> -K <sub>417</sub>	899.50	1	899.51	-	<sup>409</sup> QIAPGQTGK <sup>417</sup>
	450.25	2	450.26	-	
I <sub>418</sub> -K <sub>424</sub>	886.43	1	886.44	-	<sup>418</sup> IADYNYK <sup>424</sup>
	443.72	2	443.73	-	
V <sub>445</sub> -R <sub>454</sub>	1218.59	1	1218.59	1218.55	<sup>445</sup> VGGNYNYLYR <sup>454</sup>
	609.80	2	609.81	609.79	
L <sub>455</sub> -R <sub>457</sub>	435.27	1	435.28	-	<sup>455</sup> LFR <sup>457</sup>
	218.14	2	218.14	-	
K <sub>458</sub> -R <sub>466</sub>	559.82	2	559.81	559.78	<sup>458</sup> KSNLKPFR <sup>466</sup>
	373.55	3	373.54	373.53	
G <sub>496</sub> -R <sub>509</sub>	792.38	2	792.39	792.36	<sup>496</sup> GFQPTNGVGYQP <sup>YR</sup> <sup>509</sup>
S <sub>459</sub> -R <sub>466</sub>	495.77	2	495.78	495.76	<sup>459</sup> SNLKPFR <sup>466</sup>
K <sub>529</sub> -K <sub>535</sub>	395.25	2	395.25	-	<sup>529</sup> KSTNLVK <sup>535</sup>
S <sub>530</sub> -K <sub>535</sub>	661.39	1	661.40	-	<sup>530</sup> STNLVK <sup>535</sup>
	331.20	2	331.20	-	
N <sub>536</sub> -K <sub>537</sub>	261.16	1	261.16	-	<sup>536</sup> NK <sup>537</sup>
<b>O-glycopeptides</b>					

presumably linked to either at Thr<sub>323</sub> or Ser<sub>325</sub> according to previous reports [28, 33]. On the contrary, by in-solution BFD protocol, two peptides (R<sub>319</sub>-R<sub>328</sub> and V<sub>320</sub>-R<sub>328</sub>) linked to HexNAc; HexNAc:Hex; HexNAc:Hex:NeuAc; HexNAc:Hex:NeuAc<sub>2</sub>; HexNAc<sub>2</sub>:Hex<sub>2</sub>:NeuAc and HexNAc<sub>2</sub>:Hex<sub>2</sub>:NeuAc<sub>2</sub> were detected (Fig. 2e, Table 3). Five out of the six *O*-glycans structures were detected exclusively by the in-solution BFD protocol and these six *O*-glycans structures agree very well with the previous reports of *O*-glycosylation of Thr<sub>323</sub>/Ser<sub>325</sub> in the SARS-CoV-2 spike protein [32, 33]. MS/MS spectra of these *O*-glycopeptides confirmed this assignment (Fig. S2) by showing intense neutral losses of *O*-glycans from the precursor ions fragmented by CID according to previous reports [34].

Full-sequence coverage of *RBD*<sub>(319-541)</sub>-*HEK*<sub>A3</sub> was verified by using in-solution BFD protocol, while using the SD protocol 82% of sequence coverage was achieved (Table 2).

Several signals in the low-mass region (*m/z* 200–700) were exclusively detected when *RBD*<sub>(319-541)</sub>-*HEK*<sub>A3</sub> was analyzed by the BFD protocol and they were assigned

to short and hydrophilic internal peptides (<sup>356</sup>KR<sup>357</sup>, <sup>536</sup>NK<sup>537</sup>, <sup>455</sup>LFR<sup>457</sup>, <sup>404</sup>GDEV<sup>R</sup><sup>408</sup>, <sup>409</sup>QIAPGQTGK<sup>417</sup>, <sup>418</sup>IADYNYK<sup>424</sup>, <sup>529</sup>KSTNLVK<sup>535</sup>, and <sup>530</sup>STNLVK<sup>535</sup>; Table 3). These peptides represent the 18% of the *RBD*<sub>(319-541)</sub>-*HEK*<sub>A3</sub> sequence. Most of them (<sup>356</sup>KR<sup>357</sup>, <sup>455</sup>LFR<sup>457</sup>, <sup>409</sup>QIAPGQTGK<sup>417</sup>, <sup>418</sup>IADYNYK<sup>424</sup>, <sup>529</sup>KSTNLVK<sup>535</sup>, and <sup>530</sup>STNLVK<sup>535</sup>) were not detected by Arbeitman et al. [11] when the same RBD protein expressed in *P. pastoris* and HEK-293 T cell line was digested with a protocol similar to the in-solution SD protocol and analyzed by MALDI-MS.

The C-terminal peptide with the C<sub>538</sub> alkylated with NEM (<sup>538</sup>CVNF<sup>541</sup>-AAHHHHHH, *m/z*<sub>Exp</sub> 548.24, 3+; Fig. 2g and Table 3) was detected by both protocols (Fig. 1a–b). It confirmed that a fraction of this RBD contains an unpaired free C<sub>538</sub> residue. However, the low intensity of the signal assigned to the C-terminal peptide with a C<sub>538</sub> alkylated with NEM (*m/z*<sub>Exp</sub> 548.27, 3+; Fig. 2g and Table 3) when BFD protocol was applied suggested us that Cys<sub>538</sub> should be modified with other chemical groups.

Table.3 (continued)

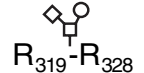


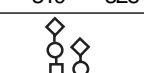
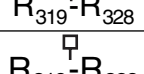
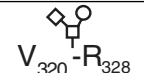
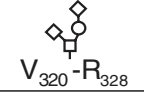
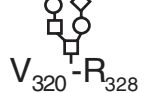

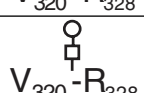
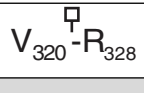





	920.96	2	920.97	-	<sup>319</sup> RVQPTEIVR <sup>328</sup> +HexNAc:Hex:NeuAc (Nt-free + O-glycosylation)
	1066.50 711.34	2 3	1066.52 711.36	1066.50 711.32	<sup>319</sup> RVQPTEIVR <sup>328</sup> +HexNAc:Hex:NeuAc <sub>2</sub> (Nt-free + O-glycosylation)
	736.02	3	736.04	-	<sup>319</sup> RVQPTEIVR <sup>328</sup> +HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc (Nt-free + O-glycosylation)
	1249.07 833.05	2 3	1249.12 833.06	-	<sup>319</sup> RVQPTEIVR <sup>328</sup> +HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc <sub>2</sub> (Nt-free + O-glycosylation)
	694.38	2	694.38	-	<sup>319</sup> RVQPTEIVR <sup>328</sup> +HexNAc (O-glycosylation)
	842.90	2	842.93	-	<sup>320</sup> VQPTEIVR <sup>328</sup> +HexNAc:Hex:NeuAc (O-glycosylation)
	988.45	2	988.48	-	<sup>320</sup> VQPTEIVR <sup>328</sup> +HexNAc:Hex:NeuAc <sub>2</sub> (O-glycosylation)
	1025.47	2	1025.50	-	<sup>320</sup> VQPTEIVR <sup>328</sup> +HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc (O-glycosylation)
	1171.02	2	1171.05	-	<sup>320</sup> VQPTEIVR <sup>328</sup> +HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc <sub>2</sub> (O-glycosylation)
	697.36	2	697.37	-	<sup>320</sup> VQPTEIVR <sup>328</sup> +HexNAc:Hex (O-glycosylation)
	616.33	2	616.33	-	<sup>320</sup> VQPTEIVR <sup>328</sup> +HexNAc (O-glycosylation)
<b>Native disulfide bonds</b>					
<b>S-S<sub>336-361</sub></b>	1452.35 1089.51	3 4	1452.39 1089.54	1452.30 1089.49	<sup>329</sup> FPDITNLCPFGEVFDATR <sup>346</sup>     <sup>358</sup> ISNCVADYSVLVNSASFSTFK <sup>378</sup> <b>(Native C336-C361)</b>
<b>S-S<sub>379-432</sub></b>	1530.71 1020.81 765.86	2 3 4	1530.73 1020.83 765.88	1530.65 1020.78 765.85	<sup>379</sup> CYGVSPTK <sup>386</sup>   <sup>425</sup> LPDDFTGCVIAWNSNNLDSK <sup>444</sup> <b>(Native C379-C432)</b>
<b>S-S<sub>391-525</sub></b>	1323.02 992.52 794.22	3 4 5	1323.07 992.54 794.25	1323.00 992.49 794.20	<sup>387</sup> LNDLCFTNVYADSFVIR <sup>403</sup>   <sup>510</sup> VVVLSEFLLHAPATVCGPK <sup>528</sup> <b>(Native C391-C525)</b>

Table 3 (continued)

<b>S-S<sub>480-488</sub></b>	1589.38 1192.29	3 4	1589.42 1192.31	1589.35 1192.26	<sup>467</sup> DISTEIYQAGSTPC <sup>541</sup> NGVEGFNC <sup>541</sup> YF PLQSYGFQPTNGVGYQP <sup>509</sup> YR <sup>509</sup> (Native C480-C488)
<b>Scrambled disulfide bonds</b>					
S-S <sub>538-379</sub>	790.36 593.02	3 4	790.38 593.03	- -	<sup>538</sup> CVNF <sup>541</sup> -AAAH <sup>541</sup> HHHHH   <sup>379</sup> CYGVSP <sup>386</sup> TK <sup>386</sup> (Scrambling C538-C379)
S-S <sub>538-432</sub>	931.67	4	931.71	-	<sup>538</sup> CVNF <sup>541</sup> -AAAH <sup>541</sup> HHHHH   <sup>425</sup> LPDDFTG <sup>541</sup> CVIAWNSNNLDSK <sup>444</sup> (Scrambling C538-C432)
S-S <sub>538-538</sub>	607.27 506.23	5 6	607.28 506.24	- -	<sup>538</sup> CVNF <sup>541</sup> -AAAH <sup>541</sup> HHHHH   <sup>538</sup> CVNF <sup>541</sup> -AAAH <sup>541</sup> HHHHH (Ct homodimer, C538-C538)
<b>Free and modified cysteines</b>					
C <sub>336</sub> +NEM	1084.01	2	1084.02	1083.99	<sup>329</sup> FPDITNLC <sup>541</sup> NEMPFGEVFD <sup>541</sup> ATR <sup>346</sup> (C336+NEM)
C <sub>432</sub> +NEM	1167.54	2	1167.56	1167.51	<sup>425</sup> LPDDFTG <sup>541</sup> C <sup>541</sup> NEMVIAWNSNNLDSK <sup>444</sup> (C432+NEM)
C <sub>391</sub> +NEM	1058.02	2	1058.02	1057.99	<sup>387</sup> LNDLC <sup>541</sup> NEMFTNVYADSFVIR <sup>403</sup> (C391+NEM)
C <sub>538</sub> +NEM	548.25	3	548.27	548.24	<sup>538</sup> C <sup>541</sup> NEMVNF <sup>541</sup> -AAAH <sup>541</sup> HHHHH (C538+NEM, +125 Da)
C <sub>538</sub> +Cys	818.84 546.23	2 3	818.86 546.25	- -	<sup>538</sup> CVNF <sup>541</sup> -AAAH <sup>541</sup> HHHHH   NH <sub>2</sub> -C-COOH (Cys+C538, +119 Da)
C <sub>538</sub> +ECG	608.25 456.44	3 4	608.26 456.45	- -	<sup>538</sup> CVNF <sup>541</sup> -AAAH <sup>541</sup> HHHHH   NH <sub>2</sub> -ECG-COOH (C538+ECG, +305 Da)
C <sub>538</sub> +CG	565.24	3	565.26	-	<sup>538</sup> CVNF <sup>541</sup> -AAAH <sup>541</sup> HHHHH   NH <sub>2</sub> -CG-COOH (C538+CG, +176 Da)
C <sub>538</sub> -34 Da	495.23	3	495.25	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAH <sup>541</sup> HHHHH (C538 - 34 Da, -SH <sub>2</sub> , dehydroalanine)
C <sub>538</sub> +CN	771.84 514.89	2 3	771.86 514.91	- -	<sup>538</sup> CVNF <sup>541</sup> -AAAH <sup>541</sup> HHHHH   C≡N (C538+CN, +25 Da)
C <sub>538</sub> +64Da	527.88	3	527.90	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAH <sup>541</sup> HHHHH (C538 + 64 Da, +SO <sub>2</sub> )

Table.3 (continued)

C538+87 Da	-	3	535.58	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAHHHHHH (C538 + 87 Da, unknown)
C538+90 Da	-	3	536.57	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAHHHHHH (C538 + 90 Da, unknown)
C538+106 Da	-	3	541.91	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAHHHHHH (C538 + 106 Da, unknown)
C538+134 Da	-	3	551.24	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAHHHHHH (C538 + hCys, homocysteine)
C538+147 Da	-	3	555.58	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAHHHHHH (C538 + 147 Da, unknown)
C538+150 Da	-	3	556.59	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAHHHHHH (C538 + 150 Da, unknown)
C538+168 Da	-	3	562.59	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAHHHHHH (C538 + 168 Da, unknown)
C538+191 Da	-	3	570.25	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAHHHHHH (C538 + 191 Da, unknown)
C538+230 Da	-	3	583.25	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAHHHHHH (C538 + 230 Da, unknown)
C538+249 Da	-	3	589.60	-	<sup>538</sup> C*VNF <sup>541</sup> -AAAHHHHHH (C538 + 249 Da, unknown)
<b>Artifacts of the protocol</b>					
 NEM-R <sub>319</sub> -R <sub>328</sub>	983.48	2	983.50	-	NEM- <sup>319</sup> RVQPTEIVR <sup>328</sup> +HexNAc:Hex:NeuAc (Nt-NEM + O-glycosylation)
 NEM-R <sub>319</sub> -R <sub>328</sub>	1129.03	2	1129.05	1129.01	NEM- <sup>319</sup> RVQPTEIVR <sup>328</sup> +HexNAc:Hex:NeuAc <sub>2</sub> (Nt-NEM + O-glycosylation)
 NEM-R <sub>319</sub> -R <sub>328</sub>	1166.05	2	1166.06	-	NEM- <sup>319</sup> RVQPTEIVR <sup>328</sup> +HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc (Nt-NEM + O-glycosylation)
 NEM-R <sub>319</sub> -R <sub>328</sub>	1311.59	2	1311.63	-	NEM- <sup>319</sup> RVQPTEIVR <sup>328</sup> +HexNAc <sub>2</sub> :Hex <sub>2</sub> :NeuAc <sub>2</sub> (Nt-NEM + O-glycosylation)
 NEM-R <sub>319</sub> -R <sub>328</sub>	837.93	2	837.95	-	NEM- <sup>319</sup> RVQPTEIVR <sup>328</sup> + HexNAc:Hex (Nt-NEM + O-glycosylation)
K <sub>356</sub> <sup>NEM</sup> -R <sub>357</sub>	428.26	1	428.27	-	<sup>356</sup> KR <sup>357</sup> +NEM at Lys <sub>356</sub>
K <sub>458</sub> <sup>NEM</sup> -K <sub>466</sub>	457.77	2	457.78	457.76	<sup>529</sup> KSTNLVK <sup>535</sup> +NEM at Lys <sub>529</sub>
K <sub>458</sub> <sup>NEM</sup> -R <sub>466</sub>	622.34	2	457.78	457.76	<sup>458</sup> KSNLKPFER <sup>466</sup> +NEM at Lys <sub>458</sub>
NEM-Cys+C <sub>538</sub>	587.91	3	587.93	-	<sup>538</sup> CVNF <sup>541</sup> -AAAHHHHHH   NEM-C-COOH (NEM at Cys+C <sup>538</sup> , +244 Da)
OH-NEM-	593.91	3	593.93	-	<sup>538</sup> CVNF <sup>541</sup> -AAAHHHHHH 

**Table 3** (continued)

Cys+C <sub>538</sub>					OH-NEM- <b>C</b> -COOH (hydrolyzed NEM at Cys+C <sup>538</sup> , +262 Da)
NEM- ECG+C <sub>538</sub>	649.93	3	649.95	-	<sup>538</sup> <b>C</b> VNF <sup>541</sup> -AAAHHHHHH   NEM-EC <b>G</b> -COOH (NEM at ECG, +430 Da)
OH-NEM- ECG+C <sub>538</sub>	655.93	3	655.96	-	<sup>538</sup> <b>C</b> VNF <sup>541</sup> -AAAHHHHHH   OH-NEM-EC <b>G</b> -COOH (hydrolyzed NEM at ECG, +448 Da)

<sup>a)</sup>The three alanine and six histidine residues located at the C-terminal end (residues 542–550) of the protein do not correspond to the RBD and were inserted in the cloning stage to facilitate the purification process of the recombinant protein by using IMAC. The superscript numbers indicate the location of the tryptic peptides within the analyzed protein. A brief description of the PTMs linked to the corresponding peptides is included. NEM, N-ethylmaleimide; C<sub>NEM</sub> cysteine alkylated with N-ethylmaleimide at the thiol group. Nt and Ct indicate an N- and C-terminal end. The residues indicated as D correspond to potential N-glycosylation sites located at Asn<sup>331</sup> and Asn<sup>343</sup> that were transformed into Asp by PNGase F

<sup>b)</sup>Monosaccharide symbols follow the SNFG system [60] and the O-glycans structures as previously reported [33]

Cyanylation ( $m/z_{Exp}$  541.91, 3+; (C<sub>538</sub>+CN)<sup>3+</sup>; Fig. 2f), cysteinylolation ( $m/z_{Exp}$  546.25, 3+; (C<sub>538</sub>+Cys)<sup>3+</sup>; Fig. 2g), and glutathionylation ( $m/z_{Exp}$  608.26, 3+; (C<sub>538</sub>+ECG)<sup>3+</sup>; Fig. 2h) of the unpaired Cys<sub>538</sub> in the C-terminal peptide of RBD<sub>(319-541)</sub>-HEK\_A<sub>3</sub> were detected exclusively when using the BFD protocol. The assignment of these modified peptides was confirmed by MS/MS analysis (Fig. 3a–c). Signals detected at  $m/z_{Exp}$  565.26, 3+ and  $m/z_{Exp}$  551.24, 3+ were also only observed when RBD<sub>(319-541)</sub>-HEK\_A<sub>3</sub> was analyzed by BFD (Table 3). MS/MS analyses demonstrated that they corresponded to the same C-terminal peptide (C<sub>538</sub>+CG)<sup>3+</sup> with the C<sub>538</sub> linked to a truncated variant of glutathione (+176 Da, +CG; Fig. S3) and homocysteine (Fig. 3d), respectively.

Signals detected at  $m/z_{Exp}$  517.24, 3+ (Fig. 2f) and  $m/z_{Exp}$  541.91, 3+ (Fig. 2g) were assigned as (C<sub>538</sub>+32 Da)<sup>3+</sup> and (C<sub>538</sub>+106 Da)<sup>3+</sup>, corresponding to the C-terminal peptide with C<sub>538</sub> linked to modifying groups of unknown chemical nature. These signals were only detected when in-solution BFD was applied to the characterization of RBD<sub>(319-541)</sub>-HEK\_A<sub>3</sub>. We also found thirteen other different variants of the C-terminal peptide (confirmed by MS/MS; see Fig. S3) that were not assigned to a known chemical structure of Cys<sup>538</sup> (see Table 3).

The alkylation with NEM, inserted in our protocols (Fig. 1a, b), transformed the hydrophilic C-terminal peptide (containing the unpaired C<sub>538</sub>) in a more hydrophobic species and in consequence, it was detected even using the SD protocol. On the contrary, the remaining Cys-capping modifications [27] mentioned above (see Fig. S3 and summarized in Table 3) did not increase the hydrophobicity of the C-terminal peptide sufficiently to be retained by

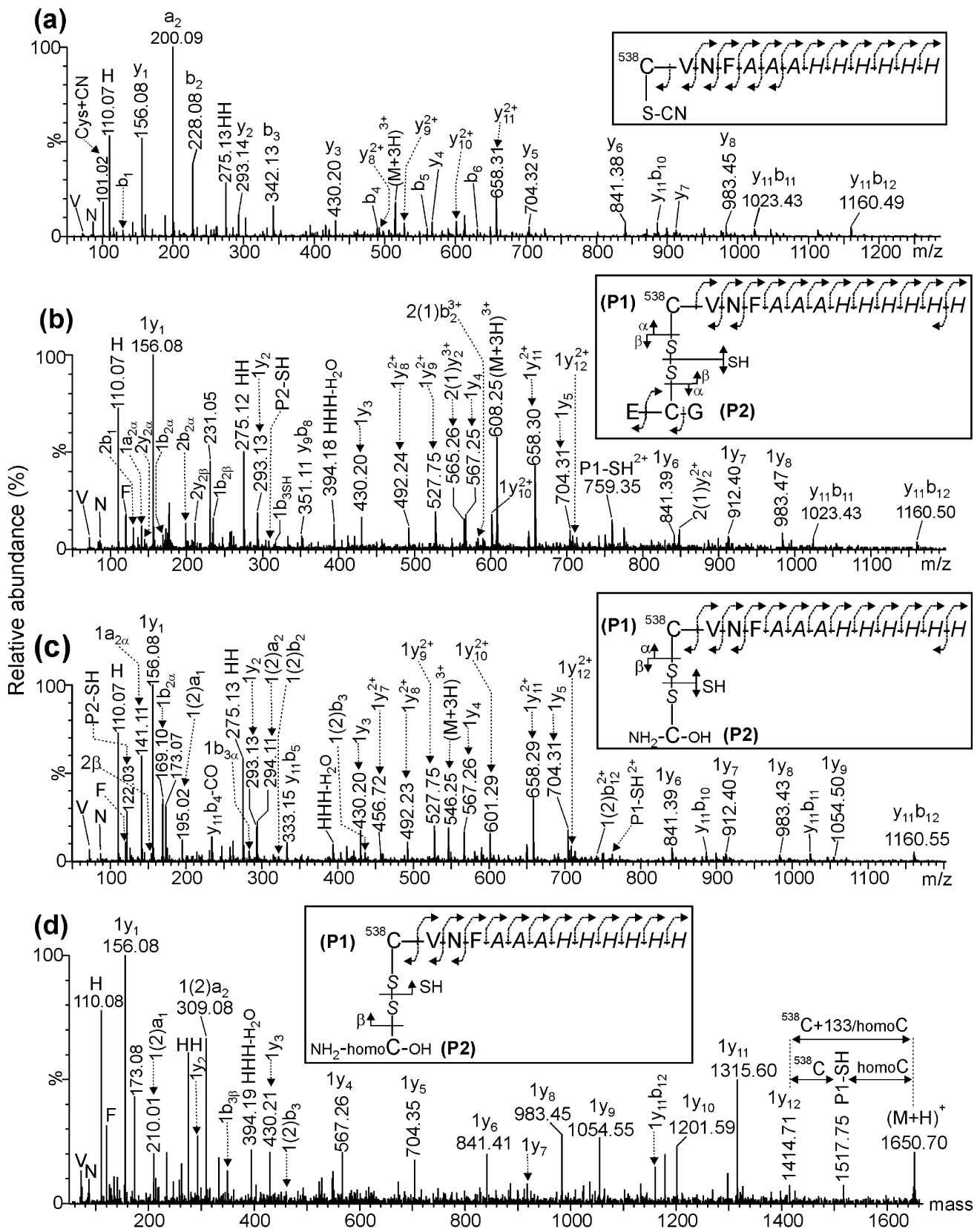
ZipTip-C<sub>18</sub> and they were detected exclusively when in-solution BFD was applied.

In contrast to the hypothesis proposing that oxidoreductase-mediated protein disulfide bonding with free cysteine or glutathione in the lumen of endoplasmic reticulum [35–37] as the source of these modifications, Zhong et al. have demonstrated that these capping modifications are generated outside mammalian cells and are sensitive to the culture medium composition [27].

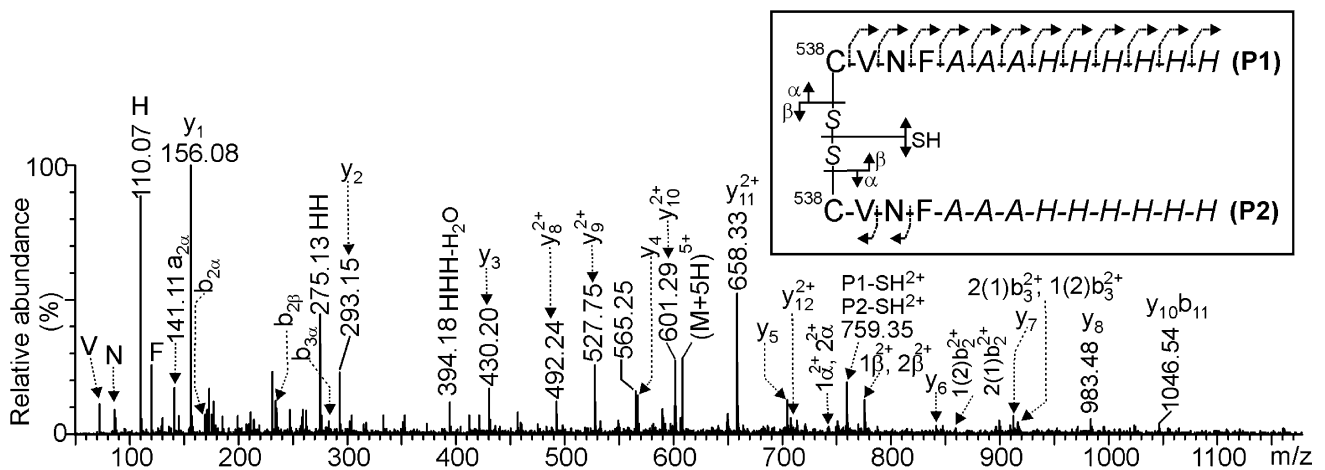
Cysteinylolation at Cys<sub>538</sub> has been reported by other authors [28, 31], but to our knowledge, the other Cys-modifying groups (Table 3) have not previously been reported for recombinant RBDs. The species with Cys<sub>538</sub> modifications and O-glycoforms detected at protein level (Table 2) were further confirmed at tryptic peptide level by the in-solution BFD (Table 3).

The use of culture media with defined composition and a well-characterized downstream process would avoid unexpected modifications of free cysteine residues [38–40], although endogenous cell metabolites may also contribute to increase protein heterogeneity at unpaired Cys.

Although Cys-capping modifications protect the molecule from aggregation and scrambling mediated by inter- and intra-molecular disulfide bonds, respectively, it needs to be addressed if the final outcome is to use the unpaired Cys for further modification, for example, in a drug conjugation process [41, 42]. Another issue also to be addressed is the potential protein heterogeneity if the final intention is the use of the dimer molecule through disulfide bonds [35, 36, 43, 44].



**Fig. 3** ESI-MS/MS spectra of C-terminal peptides (<sup>538</sup>CVNF<sup>541</sup>-AAAH<sub>11</sub>HHH) of RBD<sub>(319-541)</sub>-HEK<sub>A3</sub> containing C<sup>538</sup> modified by **a** cyanylation, **b** glutathionylation, **c** cysteinylation, and **d** homocysteinylation



**Fig. 4** MS/MS spectrum of two copies of the C-terminal peptide ( $^{538}$ CVNF $^{541}$ -AAHHHHHHH) of  $RBD_{(319-541)}$ -HEK $_A3$  linked by an intermolecular disulfide bond between two Cys $_{538}$ . The nomenclature of fragment ions is in agreement with that proposed by Mormann et al. [61]

A low-intensity signal at  $m/z_{Exp}$  607.28, 5+ and assigned to (S-S $^{5+}_{538-538}$ ) in Fig. 2h was exclusively detected when in-solution BFD protocol was applied. It suggests that a fraction of this molecule (~13% estimated by SDS-PAGE; Fig. 2a) is a dimer mediated by an intermolecular disulfide bond between two Cys $_{538}$  residues (Fig. 2a, lane 2 and lane 3). MS/MS of this signal confirmed this assignment (Fig. 4). This result matches with SDS-PAGE of  $RBD_{(319-541)}$ -HEK $_A3$  ran under reducing and non-reducing conditions (Fig. 2a).

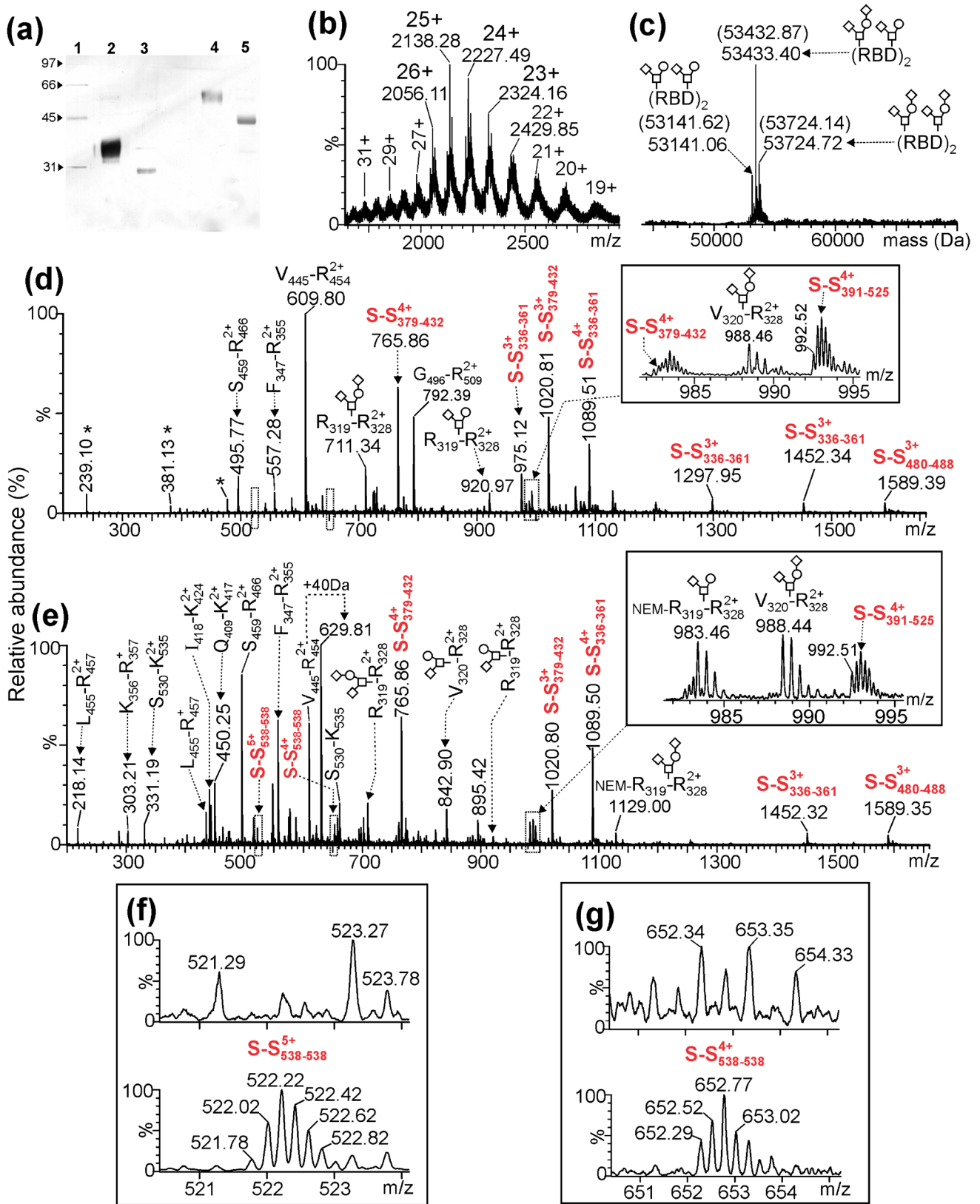
The presence of two low-abundance scrambling variants (C $_{538}$ -C $_{379}$ , C $_{538}$ -C $_{432}$ ) and the homodimer (C $_{538}$ -C $_{538}$ ) of this molecule agrees with the presence of a free Cys $_{538}$  detected in this preparation (Table 3). These two scrambled species were exclusively detected by using the in-solution BFD protocol. Also, a low-abundance population of the protein with free C $_{336}$ , C $_{391}$ , C $_{432}$ , and C $_{538}$  was detected by both protocols. All the above-mentioned assignments of scrambled and free Cys variants were confirmed by the MS/MS spectra (Figs. S4 and S5). The presence of an unpaired Cys residue may also promote disulfide exchange [18] and in consequence generates low-abundance scrambling variants of the desired molecule.

Our results indicate that Cys reduction and S-alkylation of the RBD protein before MS analysis are not convenient as important information is lost. The most striking results obtained with the BFD protocol are the detection of the disulfide-containing peptides (including low-abundance scrambled variants) and the finding of several modifications linked to free cysteines that probably most of them would be missed if reduction of disulfides takes place during sample preparation.

The analysis of the same gene construct ( $RBD_{(319-541)}$ -HEK) for the expression in HEK-293 T of the same protein without the C-terminal spacer arm of three alanines

(Table 1) by in-solution SD and BFD protocol (Fig. S6, Tables 2 and S1) yields similar results to that described here for  $RBD_{(319-541)}$ -HEK $_A3$ , at protein and peptide level (Fig. 2, Tables 2 and 3). Full-sequence coverage was achieved in the analysis of  $RBD_{(319-541)}$ -HEK by using in solution BFD protocol while using the SD protocol 85% was achieved (Table 2). C-terminal peptide containing C $_{538}$  modified with NEM was detected in both protocols (Fig. S6e,  $m/z_{Exp}$  = 477.22, 3+). However, the same C-terminal peptide containing the His $_6$  tag and other PTMs assigned to (C $_{538}$  + 106 Da) $^{3+}$  (Fig. S6e,  $m/z_{Exp}$  = 470.85, 3+), cysteinylated (Fig. S6e,  $m/z_{Exp}$  = 475.18, 3+), truncated glutathionylation (see in Fig. S6f, (C $_{538}$  + CG) $^{3+}$ ,  $m/z_{Exp}$  = 494.18, 3+), and glutathionylation (see in Fig. S6g, (C $_{538}$  + ECG) $^{3+}$ ,  $m/z_{Exp}$  = 537.20, 3+), among other PTMs previously described for  $RBD_{(319-541)}$ -HEK $_A3$  were only detected by using BFD protocol (Table S1).

In the characterization of these RBDs, short 2–9 amino acids long tryptic peptides can be detected by our in-solution BFD method; however, they are useful only to verify the sequence of already known proteins. When characterizing unknown protein species, it is preferable to resort to Lys-C and chymotrypsin, which are compatible with in-solution BFD conditions and can provide information on overlapping sequence stretches. Direct proteolysis of the RBD with Glu-C did not yield an efficient digestion with our in-solution BFD protocol except when used in tandem after Lys-C (Table S2). Shorter trypsin digestion times (15 min–4 h) of  $RBD_{(319-541)}$ -HEK $_A3$  did not yield larger peptides containing missed cleavage sites (see Fig. S7), and it provided the same information as overnight digestion, although measurement time had to be increased considerably to obtain ESI-MS spectra with a similar S/N ratio. Increasing the acetonitrile content in the spraying solution up to 50–60%





**Fig. 5** **a** SDS-PAGE analysis under reducing and non-reducing conditions of *N*-glycosylated and deglycosylated ( $RBD_{(319-541)}-CHO$ )<sub>2</sub> and detected with silver staining. Lane 1: Molecular weight markers of low-range from 31 to 97 kDa (Bio-Rad). Lanes 2–3: *N*-glycosylated and deglycosylated protein under reducing conditions detecting the reduced monomer. Lanes 4–5: *N*-glycosylated and deglycosylated protein in non-reducing conditions detecting the dimer species [( $RBD_{(319-541)}-CHO$ )<sub>2</sub>]. **b** ESI–MS spectrum of a dimeric RBD deglycosylated with PNGase F. **c** Deconvolution of the ESI–MS spectrum shown in **(b)** reveals the presence of the three major *O*-glycoforms of ( $RBD_{(319-541)}-CHO$ )<sub>2</sub>. Between parentheses the expected molecular masses of the different *O*-glycoforms are shown. ( $RBD$ )<sub>2</sub> represents an abbreviated form for referring to the ( $RBD_{(319-541)}-CHO$ )<sub>2</sub> molecule. Monosaccharide symbols follow the SNFG system [60] and the *O*-glycan structures are as previously reported [33]. The ESI–MS spectra shown in **(d)** and **(e)** correspond to the ( $RBD_{(319-541)}-CHO$ )<sub>2</sub> digested with trypsin following the SD and in-solution BFD (precipitated with acetone) protocol, respectively. Asterisks in **(d)** correspond to background signals, not assigned to tryptic peptides and (S–S)<sup>2+</sup> to peptides containing a disulfide bond between the described cysteines. The insets shown in **(d)** and **(e)** correspond to the expanded regions of the mass spectra ( $m/z$  981.5–995.5) shown by rectangles with broken lines showing the *O*-glycosylated peptides and two disulfide bond peptides (assigned as S–S<sub>391-525</sub><sup>4+</sup> and S–S<sub>379-432</sub><sup>4+</sup>). The upper- and lower-mass spectra shown in **(f)** and **(g)** correspond to two expanded regions ( $m/z$  520.4–524.1 and  $m/z$  650.5–655.2) of the ESI–MS spectra shown in **(d)** and **(e)**, respectively. A detailed assignment for all tryptic peptides in this figure is summarized in Table S4

avored the detection in the ESI–MS spectrum of three large and hydrophobic disulfide-bonded peptides containing one to three missed cleavage sites and some of the short 2–9 amino acids long tryptic peptides previously mentioned (see Fig. S8 and Table S3).

Ammonium bicarbonate is probably the most frequently used buffer for trypsin digestion of proteins. The removal of this salt by successive evaporation/dilution steps enables the direct analysis of the sample by ESI–MS without a desalting reverse-phase chromatography step and the consequent loss of valuable hydrophilic peptides. However, ammonium bicarbonate digestions do not yield ESI–MS spectra with high S/N ratio typical of the in-solution BFD protocol. It could hinder the detection of those low-abundance peptides carrying PTMs such as the detected here by applying the in-solution BFD protocol.

While the in-solution BFD protocol (Fig. 1b) was implemented with a considerable amount of recombinant RBD (50 μg, 1.5 nmol), it must be pointed out that ESI–MS analysis requires 1–3 μL out of a 100 μL sample volume. Processing lower amounts of the starting material is also possible if a more efficient protocol for protein precipitation is used (for instance with acetone at room temperature and in the presence of sodium chloride [29, 45]), and in fact, we obtained results similar to those depicted in Fig. S9 from starting amounts of 5 μg (see Experimental section in ESM). Using even lower starting amount of sample is challenging, due to

the difficulties in handling small protein pellets and the risk of sample loss during the two subsequent washing steps.

## Characterization of ( $RBD_{(319-541)}-CHO$ )<sub>2</sub>

The RBD dimer ( $RBD_{(319-541)}-CHO$ )<sub>2</sub> resulting from an intermolecular disulfide bond Cys<sub>538</sub>–Cys<sub>538</sub> was originally obtained as a by-product during the attempt to obtain  $RBD_{(319-541)}-CHO$ . The increased immunogenicity of RBD-dimer promoted its use in at least two vaccines currently in clinical trials [7, 46].

In the ( $RBD_{(319-541)}-CHO$ )<sub>2</sub> protein non-treated (lane 2, Fig. 5a) and treated (lane 3, Fig. 5a) with PNGase F and analyzed by SDS-PAGE under reducing conditions, only the presence of a glycosylated and deglycosylated monomer, respectively, was observed. When the same samples were analyzed by non-reducing conditions, the glycosylated (lane 4, Fig. 5a) and the deglycosylated (lane 5, Fig. 5a) dimers were observed. This result confirmed the covalent dimer ( $RBD_{(319-541)}-CHO$ )<sub>2</sub> and its *N*-glycosylated nature.

The ESI–MS spectrum (Fig. 5b) of the PNGase F deglycosylated dimer after the deconvolution (Fig. 5c) showed three major signals corresponding to the three combinations of two short *O*-glycan chains linked to the dimer as indicated in Fig. 5c [32]. The assignment of these *O*-glycoforms is summarized in Table 2.

The *N*-deglycosylated protein was digested with trypsin by using the in-solution SD and BFD protocols and the resultant ESI–MS spectra are shown in Fig. 5d and e, respectively. Full-sequence coverage was achieved for the in-solution BFD protocol while using the SD protocol, only 80.6% of the sequence was verified (Table 2 and Table S4).

The four disulfide bonds present in the native RBD of SARS-CoV-2 were detected by applying both protocols (Fig. 5d and 5e). *O*-glycosylated *N*-terminal peptides (R<sub>319</sub>–R<sub>328</sub> and V<sub>320</sub>–R<sub>328</sub>) with *O*-glycosylation sites located at Thr<sub>323</sub>/Ser<sub>325</sub> residues [32] were detected with appreciable intensities ( $m/z_{Exp}$  711.34, 3+; 842.90, 2+ and 988.44, 2+ in Fig. 5d and e). The mass shift provoked by these *O*-glycans observed for the *N*-deglycosylated protein (Fig. 5c) agreed with the one observed at the peptide level (Fig. 5d and e). Additionally, two low-intensity signals at  $m/z_{Exp}$  616.33, 2+ and 697.36, 2+ assigned to peptide V<sub>320</sub>–R<sub>328</sub> linked to HexNAc and HexNAc:Hex were detected only by in-solution BFD protocol (Table S4).

The most striking differences between both ESI–MS spectra (Fig. 5d and e) were observed in the low-mass region where short and hydrophilic peptides [L<sub>455</sub>–R<sub>457</sub> ( $m/z_{Exp}$

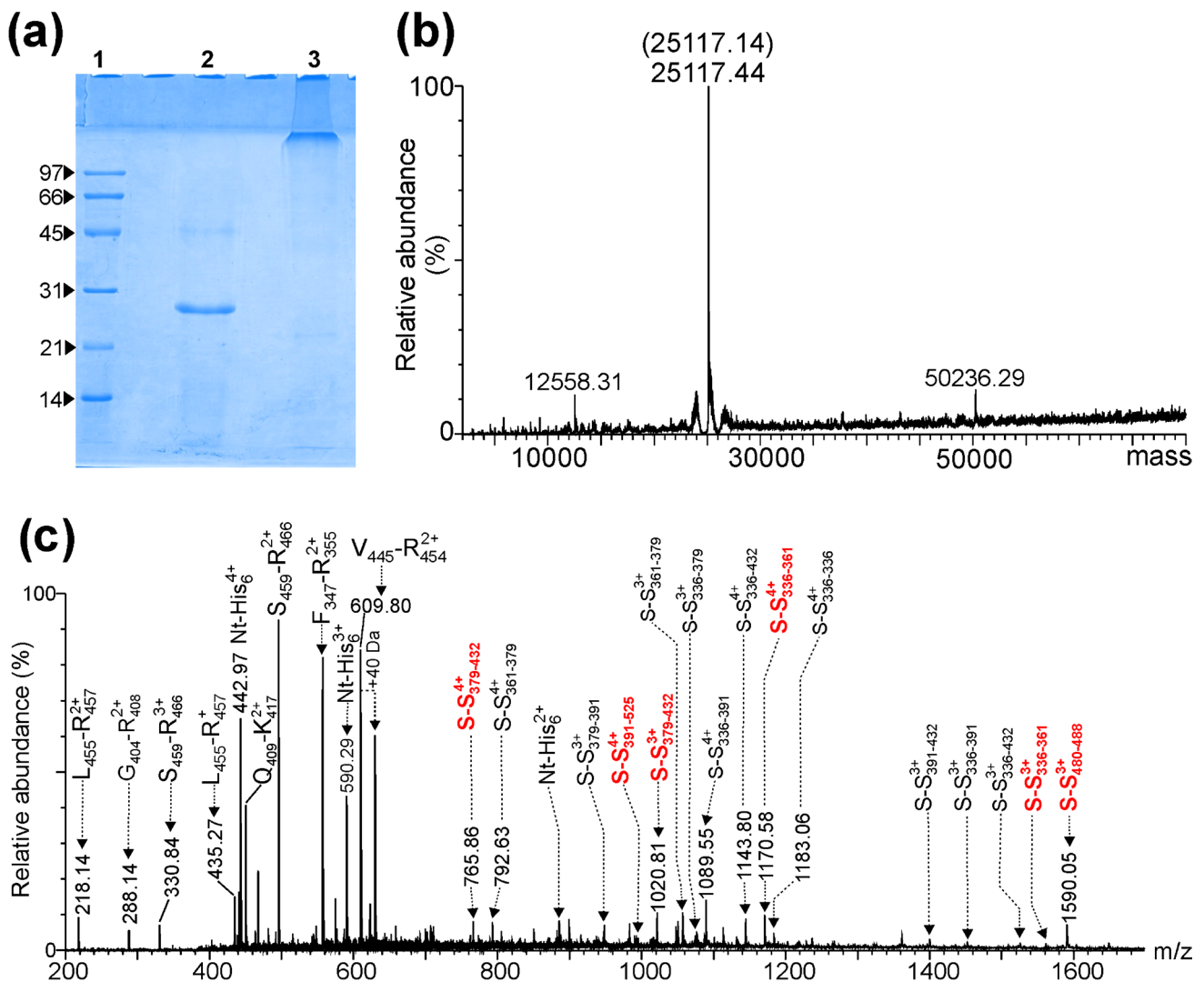
218.14, 2+), G<sub>404</sub>-R<sub>408</sub> ( $m/z_{Exp}$  288.14, 2+), K<sub>356</sub>-R<sub>357</sub> ( $m/z_{Exp}$  303.21, 1+), and S<sub>530</sub>-K<sub>535</sub> ( $m/z_{Exp}$  331.19, 2+) were only detected by applying the in-solution BFD protocol.

The ESI-MS signals that confirm the dimer nature of (RBD<sub>(319-541)</sub>-CHO)<sub>2</sub> are corresponding to the peptide [C<sub>538</sub>-H<sub>547</sub>]-S-S-[C<sub>538</sub>-H<sub>547</sub>] containing Cys<sub>538</sub> and Cys<sub>538</sub> linked by intermolecular disulfide bond ( $m/z_{Exp}$  522.02, 5+ in Fig. 5f and  $m/z_{Exp}$  652.29, 4+ in Fig. 5g). These signals that also enabled the verification of the C-terminal end of this molecule were exclusively detected by applying the in-solution BFD protocol. Probably, the presence of two His<sub>6</sub> tags (in total twelve histidine residues) in the

structure of [C<sub>538</sub>-H<sub>547</sub>]-S-S-[C<sub>538</sub>-H<sub>547</sub>] makes its retention difficult by the C<sub>18</sub>-ZipTip during the desalting step. The verification of the C-terminal end of proteins is a very important aspect included in the ICHQ6B guidelines [16].

### Characterization of RBD<sub>(331-529)</sub>-Ec

The non-correctly folded RBD is not useful for a vaccine against SARS-CoV-2 because a tridimensional structure identical to the native protein is required to generate neutralizing antibodies recognizing conformational epitopes [17].



**Fig. 6** **a** SDS-PAGE analysis of the recombinant RBD<sub>(331-529)</sub>-Ec analyzed under reducing (Lane 2) and non-reducing (Lane 3) conditions and detected with Coomassie staining. Lane 1 corresponds to the molecular weight markers of low-range from 14 to 97 kDa (Bio-Rad). **b** Deconvoluted ESI-MS spectrum of the reduced and S-carbamidomethylated protein. The expected molecular mass is indicated in parentheses. **c** ESI-MS analysis of the recombinant protein expressed

in *E. coli* and digested with trypsin by using in-solution BFD protocol. Signals assigned as (S-S)<sup>n+</sup> correspond to the peptides containing disulfide bonds between the cysteines that are described. The signals labeled with (Nt-His<sub>6</sub>)<sup>n+</sup> correspond to the N-terminal peptide containing a His<sub>6</sub> tag in its amino acid sequence. A detailed assignment for all tryptic peptides in this figure is summarized in Table S5

For this reason, the detection of non-native disulfide bonds, if present, is of tremendous importance [16].

SDS-PAGE analysis under reducing conditions (Fig. 6a, lane 2) of  $RBD_{(331-529)}-Ec$  shows a band that migrates with an estimated molecular mass of 27.3 kDa. The good agreement between the expected (25,117.14 Da) and the experimental (25,117.44 Da) molecular masses for the reduced and S-alkylated protein determined by ESI-MS analysis confirmed this result (Fig. 6b and Table 2). However, when  $RBD_{(331-529)}-Ec$  was analyzed by SDS-PAGE under non-reducing conditions (Fig. 6a, lane 3), aggregates with molecular masses higher than expected were observed. Probably, these aggregates are formed by multiple and random intermolecular disulfide bonds.

ESI-MS analysis of  $RBD_{(331-529)}-Ec$  digested with trypsin by using the in-solution BFD protocol showed several multiply-charged ion signals assigned to peptides containing Cys corresponding to the four native disulfide bonds (signals written in red and assigned as S-S<sub>##</sub><sup>n+</sup>; Fig. 6c). The good agreement between the expected and experimental molecular masses of other signals written in black and assigned as S-S<sub>##</sub><sup>n+</sup> (Fig. 6c) were assigned to tryptic peptides containing scrambled disulfide bonds in  $RBD_{(331-529)}-Ec$  (Table S5). The MS/MS spectra that confirmed these assignments are shown in Fig. S10.

The results shown here demonstrated that in-solution BFD protocol [20] in combination with ESI-MS analysis of RBD enabled in a single mass spectrum the detection of the four native disulfide bonds, the scrambled variants, and free cysteine residues that might be responsible for promoting disulfide exchange and protein aggregation [18]. Ninety-nine percent of sequence coverage for  $RBD_{(331-529)}-Ec$  was achieved when used the in-solution BFD protocol.

## Characterization of $RBD_{(333-527)}-C1$

*Thermothelomyces heterothallica* was engineered to develop an industrialized protein production host expression system with high yields (> 10 g/L) and a very significant reduction of the protease load thus minimizing unwanted degradation during fermentation [13]. Unlike other proteins characterized in this work,  $RBD_{(333-527)}-C1$  has only one *N*-glycosylation site located at Asn<sub>343</sub>.

NP-HPLC profile showed the structural assignment based on the GU indexes for the individual *N*-glycans released with PNGase F and labeled with 2AB (Fig. 7a). Deconvoluted ESI-MS spectrum of the intact  $RBD_{(333-527)}-C1$  confirmed the presence of several non-fucosylated glycoforms being M4, M5A1, and M4A1 the predominant ones (Fig. 7b). The

experimental and expected molecular masses agreed very well (Fig. 7b, Table 2).

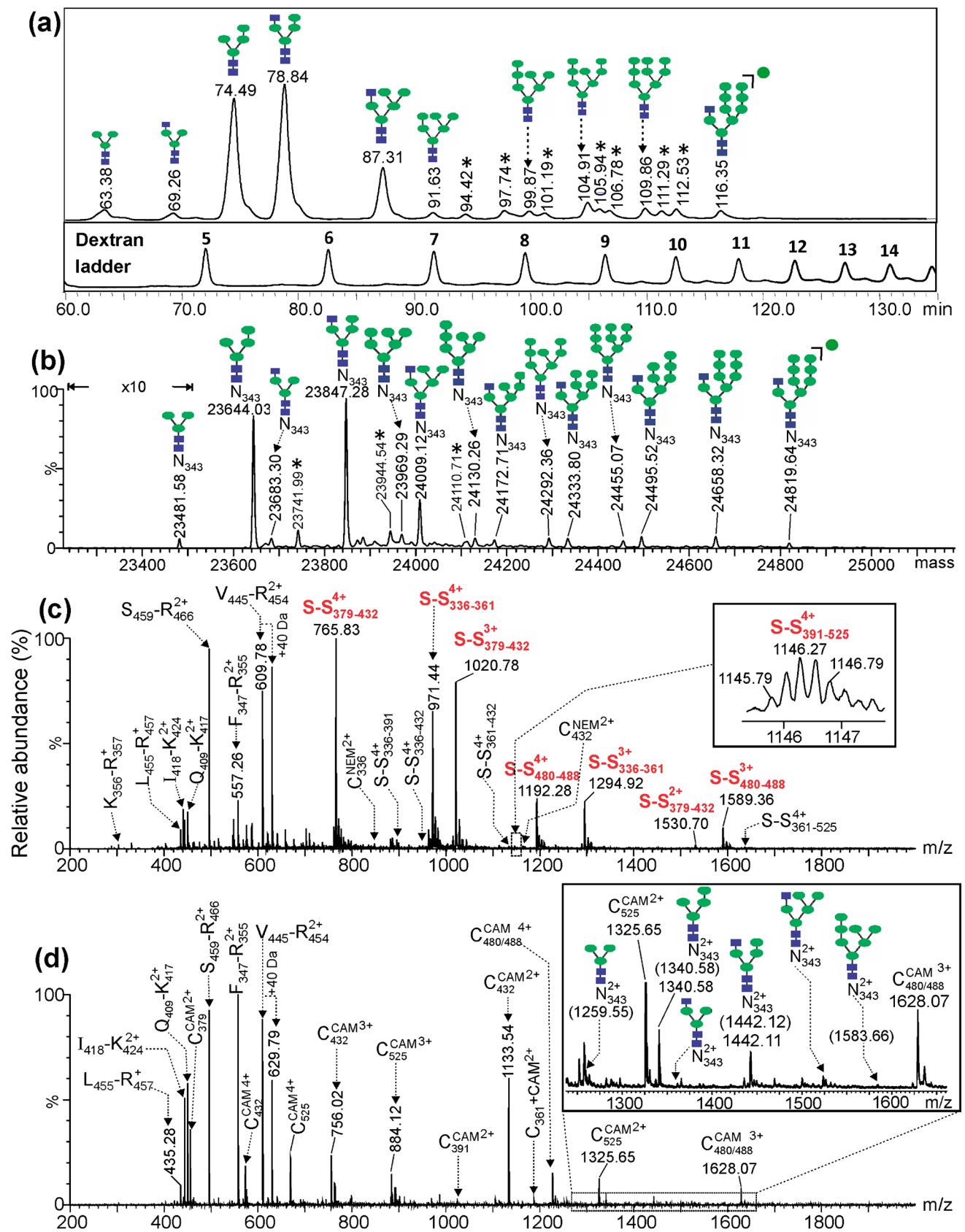
The ESI-MS spectrum of this protein (Fig. S11a and S11b) after treatment with PNGase F showed an intense signal with a mass of 22,590.33 Da (Table 2). This result agrees very well with the expected (22,590.26 Da) assuming the  $RBD_{(333-527)}-C1$  *N*-deglycosylated monomer with four disulfide bonds.

The *N*-deglycosylated protein digested with trypsin by in-solution BFD protocol (Fig. 7c) and analyzed by the ESI-MS allowed a full-sequence coverage (Table 2) and allowed the identification of the four native disulfide bonds (S-S<sub>379-432</sub>, S-S<sub>336-361</sub>, S-S<sub>480-488</sub>, and S-S<sub>391-525</sub>; Table S6). Very low-abundance signals (Fig. 7c) were detected at  $m/z_{Exp}$  847.87, 2+ and 1167.52, 2+ and assigned to the peptides T<sub>333</sub>-R<sub>346</sub> and L<sub>425</sub>-K<sub>444</sub> containing the Cys<sub>336</sub> and Cys<sub>432</sub> modified with NEM (Fig. S12a and S12b). It indicates that a minor fraction of  $RBD_{(333-527)}-C1$  contains Cys<sub>336</sub> and Cys<sub>432</sub> with free thiols in the original molecule. In addition, the same Cys<sub>336</sub> and Cys<sub>432</sub> were also detected in three low-intensity signals detected at  $m/z_{Exp}$  889.72, 4+;  $m/z_{Exp}$  944.43, 4+; and  $m/z_{Exp}$  1131.26, 4+ (see Table S6) that were assigned to (T<sub>333</sub>-R<sub>346</sub>)-S-S-(L<sub>387</sub>-R<sub>403</sub>), (T<sub>333</sub>-R<sub>346</sub>)-S-S-(L<sub>425</sub>-K<sub>444</sub>), and (I<sub>358</sub>-K<sub>378</sub>)-S-S-(L<sub>425</sub>-K<sub>444</sub>) linked by the scrambled disulfide bonds between Cys<sub>336</sub>-Cys<sub>391</sub>, Cys<sub>336</sub>-Cys<sub>432</sub>, and Cys<sub>361</sub>-Cys<sub>432</sub>, respectively (Fig. S12c-S12e). Scrambled Cys<sub>361</sub>-Cys<sub>525</sub> was also detected and the MS/MS spectrum supporting this assignment was identical to the shown in Fig. S10h. The presence of free cysteine in the molecule probably is responsible for the generation of these two low-abundance scrambling variants according to the proposed mechanisms [18].

The size heterogeneity of *N*-glycans linked to Asn<sub>343</sub> in  $RBD_{(333-527)}-C1$  was not revealed by ESI-MS analysis of the tryptic digestion (Fig. 7c) due to the removal of *N*-glycans by a PNGase F treatment. A variant of the in-solution BFD protocol without the PNGase F treatment did not provide this information because the *N*-terminal peptide (T<sub>333</sub>-R<sub>346</sub>) of the  $RBD_{(333-527)}-C1$  containing the glycosylated Asn<sub>343</sub> is linked to the peptide (I<sub>358</sub>-K<sub>378</sub>) by a disulfide bond (Cys<sub>336</sub>-Cys<sub>361</sub>).

Probably, the microheterogeneity of *N*-glycosylation gives rise to low-abundance *N*-glycopeptides that combined with their high molecular masses (over 4 kDa) have an ionization suppressed by the presence of shorter tryptic peptides in the sample. The combination of all these aspects made it difficult for the ESI-MS analysis of these *N*-glycopeptides.

However, when the *N*-glycosylated  $RBD_{(333-527)}-C1$  was reduced and S-alkylated with iodoacetamide and digested using the in-solution BFD, all cysteine-containing peptides



**Fig. 7 a** NP-HPLC profile (upper chromatogram) of the 2AB-*N*-glycans released by PNGase F treatment of the recombinant *RBD*<sub>(333–527)</sub>-*CI* and corresponding dextran ladder (lower chromatogram) used to calculate the GU indexes for all 2AB-*N*-glycans and to perform for the structural assignment. The asterisks correspond to non-assigned glycoforms. The numbers above peaks in the dextran ladder indicate the corresponding glucose units. The nomenclature used in the structural assignment of the 2-AB *N*-glycans agrees with the ones proposed by the SNFG system [60]. The deconvoluted ESI–MS spectrum shown in **(b)** corresponds to the intact protein with potential *N*-glycosylation site located at the Asn<sup>343</sup> occupied to several glycoforms. A magnification of 10× is shown in the low molecular mass region of **(b)**. The ESI–MS spectrum shown in **(c)** corresponds to the *RBD*<sub>(333–527)</sub>-*CI* treated with PNGase F and digested following the in-solution BFD protocol shown in Fig. 1b. The ESI–MS spectrum shown in **(d)** corresponds to the reduced and S-alkylated glycosylated *RBD*<sub>(333–527)</sub>-*CI*. Signals assigned as (C# + cam)<sup>n+</sup> correspond to tryptic peptides containing carbamidomethyl cysteine residues at position #. The inset shown in **(d)** corresponds to an expanded region (*m/z* 1237–1662) showing the presence of several signals assigned to the *N*-terminal end glycopeptides (T<sub>333</sub>-R<sub>346</sub>) with several *N*-glycans linked to the glycosylated Asn<sup>343</sup>. Signal assigned as (C<sub>480/488</sub> + cam)<sup>3+</sup> corresponds to the peptide D<sub>467</sub>-R<sub>509</sub> containing the Cys<sub>480</sub> and Cys<sub>488</sub> S-alkylated with iodoacetamide. A detailed assignment for all tryptic peptides in this figure is summarized in Table S6

were detected (Fig. 7d, Fig. S13a–d, Table S6) including the *N*-terminal peptide T<sub>333</sub>-R<sub>346</sub> containing Cys<sub>336</sub> and several glycoforms as shown in the inset of Fig. 7d. MS/MS spectra supporting these assignments are shown in Fig. S13e–f.

### Characterization of *RBD*<sub>(331–530)</sub>-*Cmyc*-*Pp*

RBD of SARS-CoV-2 was also expressed in *P. pastoris* with a His<sub>6</sub> tag and the *Cmyc* tag fused at the *C*-terminal end (*RBD*<sub>(331–530)</sub>-*Cmyc*-*Pp*; see Table 1) to be used for analytical purposes. The ESI–MS spectrum of *RBD*<sub>(331–530)</sub>-*Cmyc*-*Pp* deglycosylated with PNGase F (Fig. 8a) after deconvolution (Fig. 8b) yields an intense signal with a molecular mass of 25,835.29 Da that is 400.88 Da higher than expected (25,434.41 Da; Table 2).

The *N*-deglycosylated protein was digested with trypsin by the in-solution BFD protocol (Fig. 1b) and the resultant ESI–MS spectrum (Fig. 8c) showed an unexpected signal of appreciable intensity at *m/z*<sub>Exp</sub> 1219.32, 4+. The MS/MS spectrum of this signal (Fig. 8d) confirmed that two peptides [EAEAEFS-(D<sup>331</sup>-R<sup>346</sup>)-S-S-(I<sup>358</sup>-R<sup>378</sup>)] were linked by an intermolecular disulfide bond between Cys<sub>336</sub> and Cys<sub>361</sub>. One of these peptides [EAEAEFS-(D<sub>331</sub>-R<sub>346</sub>)] contains an incomplete processed fragment of the alpha mating factor signal peptide (EAEA-) [47] linked to the expected *N*-terminal end EFS-(D<sub>331</sub>-R<sub>346</sub>) of the mature

*RBD*<sub>(331–530)</sub>-*Cmyc*-*Pp*. The expected molecular mass of the residues (EAEA-) linked to the *N*-terminal end (400.39 Da) agrees with the mass difference observed between the experimental and calculated molecular mass for the *N*-deglycosylated protein (400.88 Da; Fig. 8b).

Table S7 shows a summary for the assignment of all signals observed in the ESI–MS spectrum of Fig. 8c. In-solution BFD protocol in combination with ESI–MS analysis achieved a sequence coverage of 99% (Table 2).

The α-mating factor prepro peptide secretion signal is the most commonly used signal sequence for recombinant proteins expressed in *P. pastoris* [48]. Processing of the alpha mating factor should occur in three steps; in particular, the last step involves the Ste13 protein that cleaves the Glu-Ala repeats in Golgi [49]. All the purified protein was detected exclusively with the EAEA-linked to the *N*-terminal end. Probably the high expression level of this protein (40 g/L) impaired the complete processing of the propeptide. The characterization of the *N*-terminal end is also one of the aspects requested by the ICHQ6B guidelines [16].

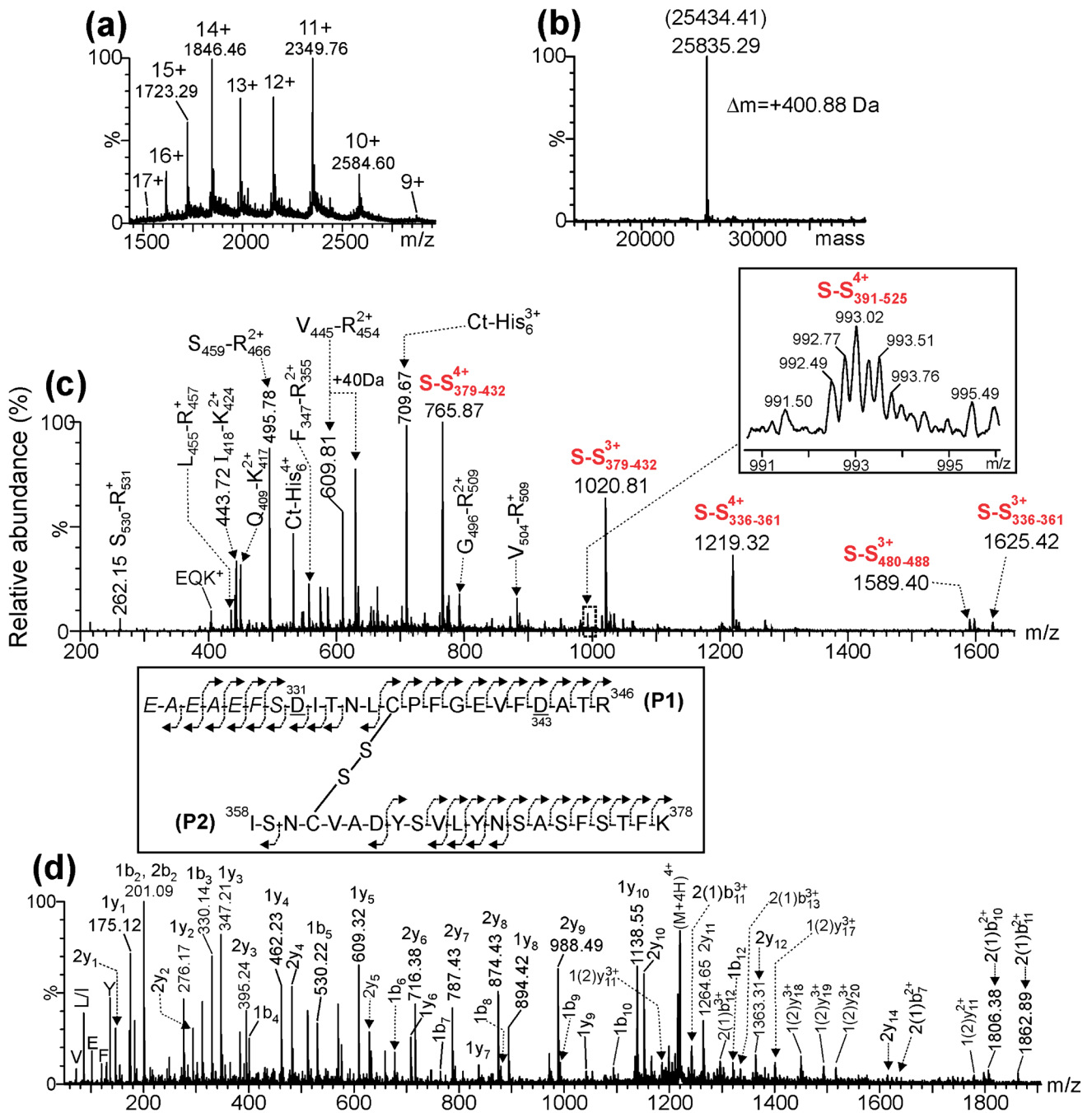
### Artificial modifications introduced during sample processing by the in-solution BFD protocol

In the characterization of all RBDs by using in-solution BFD protocol, we initially used acetone for protein precipitation (Fig. 1b). We noticed in the ESI–MS spectra an unexpected doubly-charged signal at *m/z*<sub>Exp</sub> 629.81 (Fig. 9a) having a variable intensity. This signal was not detected when RBDs were processed by using in-solution SD protocol (Fig. 2d, 5d, and S6c) and when the protein precipitation step (Fig. 1b) was carried out with cold ethanol (Fig. 9b) instead of acetone (Fig. 9a).

Comparison between the MS/MS spectra of the unmodified peptide (<sup>445</sup>VGGNYNYLYR<sup>454</sup>, *m/z*<sub>Exp</sub> 609.80, 2+; Fig. 9c) and the signal detected at *m/z*<sub>Exp</sub> 629.81, 2+ (Fig. 9d) revealed that it corresponds to the same internal peptide (Val<sup>445</sup>-Arg<sup>454</sup>) modified by adding 40 Da alternatively at Gly<sup>446</sup> (<sup>445</sup>V[G+40]GNYNYLYR<sup>454</sup>) and at the *N*-terminal end (<sup>445</sup>[V+40]GGNYNYLYR<sup>454</sup>).

Although in literature a structure for this modification has not been proposed yet, a previous work indicated that it is specific only for those peptides having Gly at position *n*+2 that were derived from tryptic digests of proteins previously precipitated with acetone [50]. All RBDs characterized here have only one internal tryptic peptide <sup>445</sup>VG\*GNYNYLYR<sup>454</sup> with this characteristic.

The acetone traces that remain adhered in the pellet, during trypsin digestion at 37 °C for 16 h, are responsible for this modification [50]. The intensity of this modified peptide can be reduced considerably if a 15 min vacuum drying step



**Fig. 8** **a** ESI-MS analysis of the deglycosylated *RBD*<sub>(331-530)</sub>-*cmyc-Pp* expressed in *P. pastoris*. **b** Deconvoluted ESI-MS spectrum. The expected mass of the *N*-deglycosylated protein is shown in parentheses. **c** ESI-MS analysis of the in-solution BFD trypsin digestion of the *N*-deglycosylated *RBD*<sub>(331-530)</sub>-*cmyc-Pp*. The inset shows the isotopic ion distribution of a 4+ ion corresponding to peptides [Leu<sub>387</sub>-Arg<sub>403</sub>]-S-S-[Val<sub>510</sub>-Lys<sub>528</sub>] linked by a disulfide bond between C<sub>391</sub>-C<sub>525</sub>. A summary of the above results is shown in Tables 2–3 and the detailed assignment for all signals in (c) is shown

in Table S7. **d** ESI-MS/MS spectrum of peptides [EAEAEFS-Asn<sub>331</sub>-Arg<sub>346</sub>]-S-S-[Ile<sub>358</sub>-Lys<sub>378</sub>] linked by a disulfide bond between C<sub>336</sub> and C<sub>361</sub>. This species contains an extension of seven amino acids (EAEAEFS-) added to the expected *N*-terminal end [Asn<sub>331</sub>-Arg<sub>346</sub>] due to an incomplete processing of the propeptide (alpha mating factor) during protein expression. Asn<sub>331</sub> and Asn<sub>343</sub> are transformed into Asp residues due to the action of PNGase F. The nomenclature for the fragment ions observed in the MS/MS spectrum agrees with the proposed by Mormann et al. [61]

is inserted in the protocol after acetone protein precipitation. However, care should be taken because an extensive drying makes dissolving the protein pellet in water/acetonitrile more difficult.

In-solution BFD of proteins precipitated with ethanol and acetone yield very similar results and they can be used indistinctively. However, during the analysis of *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub> after acetone precipitation, the isotopic ion distributions of the modified <sup>445</sup>Val-Arg<sup>454</sup> + 40 Da peptide ( $m/z_{Exp}$  629.81, 2+; Fig. 9a) and the C-terminal peptide (<sup>538</sup>CVNF<sup>541</sup>-AAAHHHHHH) carrying a + 374 Da modification at Cys<sup>538</sup> (Fig. 9b) were partially overlapped and thus, it impaired its detection. This modification at Cys<sup>538</sup> was only detected when the protein *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub> was precipitated with ethanol and analyzed by in-solution BFD (Fig. 9b).

Another artifact originated by the sample processing was the partial addition of NEM to the N-terminal end of the RBD proteins despite the fact that maleimide has 1000-fold selectivity for thiols over amine groups at neutral pH [51].

The addition of NEM was verified by ESI-MS analyses of the RBD deglycosylated with PNGase F (Table 3) and confirmed by the ESI-MS/MS analysis of the N-terminal tryptic O-glycopeptides (Fig. S14). Despite the abundant fragmentation of glycans in the MS/MS of Fig. S14, three b<sub>n</sub> ions (b<sub>1</sub>, b<sub>3</sub>, and b<sub>4</sub>) were detected containing the N-terminal end of the peptide R<sub>319</sub>-R<sub>328</sub> and increased their masses by 125 Da due to the addition of NEM.

In addition, the cysteinylated *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub> also partially added two molecules of NEM, one at the N-terminal end of Arg<sup>319</sup> (Fig. S14a–S14b) and a second one to the N-terminal end of Cys linked to Cys<sup>538</sup> (Fig. 9e). The ESI-MS/MS spectrum of the cysteinylated C-terminal peptide (<sup>538</sup>CVNF<sup>541</sup>-AAAHHHHHH,  $m/z_{Exp}$  587.92, 3+) of *RBD*<sub>(319–541)</sub>-*HEK*<sub>A3</sub> (Fig. 9e) confirms this finding. This result is in agreement with a publication that reports the alkylation (+ 125 Da) at the N-terminal end of proteins treated with NEM [52].

Using the in-solution BFD protocol (Fig. 1b), the remaining internal tryptic peptides were not modified with NEM at their N-terminal ends because this S-alkylating reagent was eliminated during sample precipitation and the subsequent washing steps before proceeding to the proteolytic digestion. However, three low-intensity signals in the ESI-MS analysis of tryptic digestion corresponding to peptides <sup>356</sup>KR<sup>357</sup> ( $m/z_{Exp}$  428.26, 1+), <sup>458</sup>KSNLKPFR<sup>466</sup> ( $m/z_{Exp}$  622.36, 2+), and <sup>529</sup>KSTNLVK<sup>535</sup> ( $m/z_{Exp}$  457.78, 2+) with the epsilon amino group of Lys<sub>356</sub>, Lys<sub>458</sub>, and Lys<sub>529</sub> modified with NEM (+ 125 Da) were detected using the in-solution BFD protocol and confirmed by MS/MS (Fig. S15).

On the contrary, when the RBD is digested in-solution by using the SD protocol and NEM is present even at a very low concentration ( $\leq 5$  mM) during all sample processing, it will be added to the N-terminal end of most of the internal tryptic peptides (data not shown).

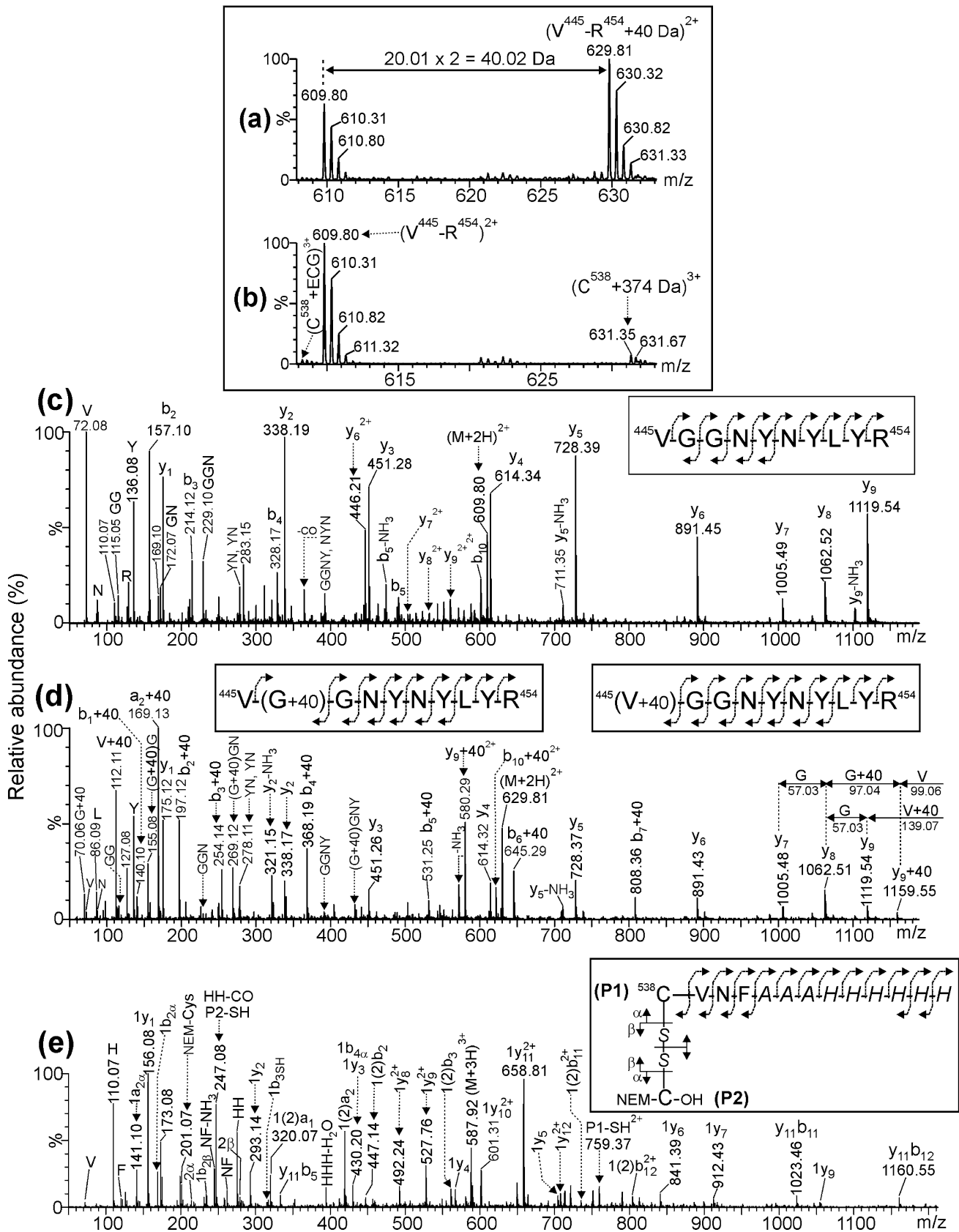
NEM is added in excess at a concentration of 5 mM and it remains during the N-deglycosylation step (2 h at 37 °C) at a pH slightly over neutral (7.2–7.4). It seems that these conditions make this side reaction favorable at the N-terminal end of the deglycosylated RBDs as well as for the cysteine linked by disulfide bond to Cys<sup>538</sup>. In a minor extension, few epsilon amino groups of Lys residues were partially modified. Therefore, the partial addition of NEM at the N-terminal end of the protein is a side reaction to be considered when in-solution BFD is used.

We also observed hydrolysis of the thiosuccinimide ring after derivatization of free Cys residues by NEM, especially when digesting the resulting RBD preparation according to the SD protocol at basic pH [53] (Fig. 1a).

Side reactions associated with the addition of NEM [52] will be present in both protocols. Using other alkylating agents (e.g., iodoacetamide, iodoacetic acid, 4-vinylpyridine, and acrylamide) to block free cysteine residues at the initial steps of the protocol was not evaluated here, but they could also be useful. Potential side reactions related to the presence of Cys-blocking groups should definitively be explored in depth to develop a well-characterized protocol [52–56].

## Conclusions

In-solution BFD in a single ESI-MS spectrum enabled the full-sequence coverage for most recombinant RBD sequences characterized in this work and outperformed the in-solution SD protocol in this aspect. The in-solution BFD protocol in combination with ESI-MS analysis has been demonstrated to be sensitive for the detection of PTMs present in the recombinant RBDs produced in different expression systems. Most of these PTMs were only detected when in-solution BFD was applied. The identification of the highly hydrophilic C-terminal peptides of these RBD proteins containing a His<sub>6</sub> tag and twelve histidine residues, an important aspect requested in the ICHQ6B guidelines, was always possible by applying the in-solution BFD while with the SD sample processing, the identification was achieved only in few cases. The results shown here support that in-solution BFD protocol in combination with ESI-MS analysis can be implemented successfully for the characterization of RBDs used as active pharmaceutical ingredients of SARS-CoV-2 subunit-based vaccines [31, 57] including those derived from mutated variants of the virus [4, 58, 59].





**Fig. 9** The ESI-MS spectra shown in (a) and (b) correspond to expanded regions of the tryptic peptides derived from *RBD*<sub>(319–541)-HEK<sub>A</sub>3 digested by in-solution BFD protocol after precipitation with acetone and ethanol, respectively. The signals assigned in (b) as (C<sup>538</sup> + ECG)<sup>3+</sup> and (C<sup>538</sup> + 374 Da)<sup>3+</sup> correspond to the C-terminal peptide <sup>538</sup>CVNF<sup>541</sup>-AAAAHHHHHH with the C<sup>538</sup> modified with glutathione and a chemical modification of unknown chemical nature that increased its molecular mass by 374 Da, respectively. The MS/MS spectra shown in (c) and (d) correspond to the internal non-modified Val<sup>445</sup>-Arg<sup>454</sup> peptide (*m/z*<sub>Exp</sub> 609.80, 2+) and the same peptide with a modification that increased its molecular mass by 40.02 Da (*m/z*<sub>Exp</sub> 629.81, 2+), respectively. This chemical modification introduced in the precipitation step with acetone is located alternatively at the N-terminal end (V+40) or at the second position glycine (G+40). The MS/MS spectra shown in (e) correspond to the cysteinylated peptide C-terminal end peptide (<sup>538</sup>CVNF<sup>541</sup>-AAAAHHHHHH) with the C<sup>538</sup> linked by a disulfide bond (-S-S-) to a Cys residue (C-OH) modified at the N-terminal end with an N-ethylmaleimide group (NEM-) introduced during the sample processing. Peptide and C-OH have been assigned as P1 and P2, respectively. The nomenclature of fragment ions is in agreement with the proposed by Mormann et al. [61]</sub>

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00216-021-03721-w>.

**Acknowledgements** The authors acknowledge their institutional support for the COVID-19 vaccine project.

**Author contribution** Luis Ariel Espinosa: Conceptualization, investigation, data curation, writing-original draft; Yassel Ramos: Conceptualization/investigation; Ivan Andújar: Investigation; Enso Onill Torres: Investigation; Gleysin Cabrera: Investigation; Alejandro Martín: Investigation; Diamilé Roche: Investigation; Glay China: Investigation; Mónica Becquet: Investigation; Isabel González: Investigation; Camila Canaán-Haden: Investigation; Elías Nelson: Investigation; Gertrudis Rojas: Investigation/writing review and editing; Beatriz Pérez-Massón: Investigation; Dayana Pérez-Martínez: Investigation; Tamy Boggiano: Investigation/supervision; Julio Palacio: Investigation; Sum Lai Lozada: Investigation; Lourdes Hernández: Investigation; Kathya Rashida de la Luz Hernández: Investigation/supervision; Saloheimo Markku: Investigation; Marika Vitikainen: Investigation; Yury Valdés-Balbín: Investigation, review and editing; Darielys Santana-Medero: Investigation, review and editing; Daniel G. Rivera: Investigation, review and editing; Vicente Vérez-Bencomo: Funding acquisition, review and editing; Mark Emalfarb: Investigation; Ronen Tchelet: Review and editing; Gerardo Guillén: Funding acquisition/supervision; Miladys Limonta: Supervision; Eulogio Pimentel: Funding acquisition; Marta Ayala: Funding acquisition/supervision; Vladimir Besada: Writing review and editing; Luis Javier González: Conceptualization, data curation, writing—original draft, writing review and editing.

**Funding** This research was supported by the Grant awarded to the COVID-19 vaccine project by the National Science and Technology Program of the Cuban Ministry of Science and Technology.

## Declarations

**Consent to participate** Consent to submit this manuscript has been received from all co-authors.

**Conflict of interest** The authors declare no competing interests.

## References

1. Wouters OJ, Shadlen KC, Salcher-Konrad M, Pollard AJ, Larson HJ, Teerawattananon Y, et al. Challenges in ensuring global access to COVID-19 vaccines: production, affordability, allocation, and deployment. *Lancet*. 2021;397:1023–34. [https://doi.org/10.1016/S0140-6736\(21\)00306-8](https://doi.org/10.1016/S0140-6736(21)00306-8).
2. WHO. Target product profiles for COVID-19 vaccines <https://www.who.int/publications/m/item/whotarget-product-profiles-for-covid-19-vaccines> [updated version 3. April 29, 2020].
3. Li Y, Lai D-y, Zhang H-n, Jiang H-w, Tian X, Ma M-l, et al. Linear epitopes of SARS-CoV-2 spike protein elicit neutralizing antibodies in COVID-19 patients. *Cell Mol Immunol*. 2020;17(10):1095–7. <https://doi.org/10.1038/s41423-020-00523-5>.
4. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270–3. <https://doi.org/10.1038/s41586-020-2012-7>.
5. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*. 2020;367(6483):1260–3. <https://doi.org/10.1126/science.abb2507>.
6. Yang J, Wang W, Chen Z, Lu S, Yang F, Bi Z, et al. A vaccine targeting the RBD of the S protein of SARS-CoV-2 induces protective immunity. *Nature*. 2020;586(7830):572–7. <https://doi.org/10.1038/s41586-020-2599-8>.
7. Valdes-Balbin Y, Santana-Mederos D, Paquet F, Fernandez S, Climent Y, Chiodo F, et al. Molecular aspects concerning the use of the SARS-CoV-2 receptor binding domain as a target for preventive vaccines. *ACS Cent Sci*. 2021. <https://doi.org/10.1021/acscentsci.1c00216>.
8. Hotez PJ, Bottazzi ME. Developing a low-cost and accessible COVID-19 vaccine for global health. *PLoS Negl Trop Dis*. 2020;14(7):e0008548. <https://doi.org/10.1371/journal.pntd.0008548>.
9. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature*. 2020;581(7807):221–4. <https://doi.org/10.1038/s41586-020-2179-y>.
10. Li T, Zheng Q, Yu H, Wu D, Xue W, Xiong H, et al. SARS-CoV-2 spike produced in insect cells elicits high neutralization titres in non-human primates. *Emerg Microbes & Infect*. 2020;9(1):2076–90. <https://doi.org/10.1080/22221751.2020.1821583>.
11. Arbeitman CR, Auge G, Blaustein M, Bredeston L, Corapi ES, Craig PO, et al. Structural and functional comparison of SARS-CoV-2-spike receptor binding domain produced in *Pichia pastoris* and mammalian cells. *Sci Rep*. 2020;10(1):21779–97. <https://doi.org/10.1038/s41598-020-78711-6>.
12. Pollet J, Chen W-H, Versteeg L, Keegan B, Zhan B, Wei J, et al. SARS-CoV-2 RBD219-N1C1: a yeast-expressed SARS-CoV-2 recombinant receptor-binding domain candidate vaccine stimulates virus neutralizing antibodies and T-cell immunity in mice. *Hum Vaccin Immunother*. 2021:1–11. <https://doi.org/10.1080/21645515.2021.1901545>.
13. Visser H, Joosten V, Punt PJ, Gusakov AV, Olson PT, Joosten R, et al. Development of a mature fungal technology and production platform for industrial enzymes based on a Myceliophthora thermophila isolate, previously known as Chrysosporium luc-knowense C1. *Ind Biotechnol*. 2011;7(3):214–23. <https://doi.org/10.1089/ind.2011.7.214>.
14. Fujita R, Hino M, Ebihara T, Nagasato T, Masuda A, Lee JM, et al. Efficient production of recombinant SARS-CoV-2 spike protein using the baculovirus-silkworm system. *Biochem Biophys*

- Res Commun. 2020;529(2):257–62. <https://doi.org/10.1016/j.bbr.2020.06.020>.
15. Prahlad J, Struble L, Lutz WE, Wallin SA, Khurana S, Schnaubelt A, et al. Bacterial expression and purification of functional recombinant SARS-CoV-2 spike receptor binding domain. bioRxiv. 2021. <https://doi.org/10.1101/2021.02.03.429601>.
  16. Rudge SR, Nims RW. ICH Q6B specifications: test procedures and acceptance criteria for biotechnological/biological products. ICH Quality Guidelines: An Implementation Guide. 2017:467. <https://doi.org/10.1002/9781118971147.ch17>.
  17. Poh CM, Carissimo G, Wang B, Amrun SN, Lee CY-P, Chee RS-L, et al. Two linear epitopes on the SARS-CoV-2 spike protein that elicit neutralising antibodies in COVID-19 patients. Nat Commun. 2020;11(1):1–7. <https://doi.org/10.1038/s41467-020-16638-2>.
  18. Lakhub JC, Shipman JT, Desaire H. Recent mass spectrometry-based techniques and considerations for disulfide bond characterization in proteins. Anal Bioanal Chem. 2018;410(10):2467–84. <https://doi.org/10.1007/s00216-017-0772-1>.
  19. Castellanos-Serra L, Ramos Y, Huerta V. An in-gel digestion procedure that facilitates the identification of highly hydrophobic proteins by electrospray ionization-mass spectrometry analysis. Proteomics. 2005;5(11):2729–38. <https://doi.org/10.1002/pmic.2004011644>.
  20. Betancourt LH, Espinosa LA, Ramos Y, Bequet-Romero M, Rodríguez EN, Sánchez A, et al. Targeting the hydrophilic regions of recombinant proteins by MS via in-solution buffer-free trypsin digestion. Eur J Mass Spectrom. 2020;26(3):230–7. <https://doi.org/10.1177/1469066719893492>.
  21. Espinosa LA, Ramos Y, Andujar I, Torres EO, Cabrera G, Martin A, et al. In-solution buffer-free digestion for the analysis of SARS-CoV-2 RBD proteins allows a full sequence coverage and detection of post-translational modifications in a single ESI-MS spectrum. bioRxiv. 2021. <https://doi.org/10.1011/2021.05.10.443404>.
  22. Laemmli UK. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature. 1970;227(5259):680–5. <https://doi.org/10.1038/227680a0>.
  23. Heukeshoven J, Dernick R. Characterization of a solvent system for separation of water-insoluble poliovirus proteins by reversed-phase high-performance liquid chromatography. J Chromatogr A. 1985;326:91–101. [https://doi.org/10.1016/S0021-9673\(01\)87434-3](https://doi.org/10.1016/S0021-9673(01)87434-3).
  24. Guile GR, Rudd PM, Wing DR, Prime SB, Dwek RA. A rapid high-resolution high-performance liquid chromatographic method for separating glycan mixtures and analyzing oligosaccharide profiles. Anal Biochem. 1996;240(2):210–26. <https://doi.org/10.1006/abio.1996.0351>.
  25. Kerr J, Schlosser JL, Griffin DR, Wong DY, Kasko AM. Steric effects in peptide and protein exchange with activated disulfides. Biomacromol. 2013;14(8):2822–9. <https://doi.org/10.1021/bm400643p>.
  26. Monahan FJ, German JB, Kinsella JE. Effect of pH and temperature on protein unfolding and thiol/disulfide interchange reactions during heat-induced gelation of whey proteins. J Agric Food Chem. 1995;43(1):46–52. <https://doi.org/10.1021/jf00049a010>.
  27. Zhong X, He T, Prashad AS, Wang W, Cohen J, Ferguson D, et al. Mechanistic understanding of the cysteine capping modifications of antibodies enables selective chemical engineering in live mammalian cells. J Biotechnol. 2017;248:48–58. <https://doi.org/10.1016/j.jbiotec.2017.03.006>.
  28. Gstöttner C, Zhang T, Resemann A, Ruben S, Pengelley S, Suckau D, et al. Structural and functional characterization of SARS-CoV-2 RBD domains produced in mammalian cells. Anal Chem. 2021;93(17):6839–47. <https://doi.org/10.1021/acs.analchem.1c00893>.
  29. Crowell AM, Wall MJ, Doucette AA. Maximizing recovery of water-soluble proteins through acetone precipitation. Anal Chim Acta. 2013;796:48–54. <https://doi.org/10.1016/j.aca.2013.08.005>.
  30. Ma J, Stoter G, Verweij J, Schellens JH. Comparison of ethanol plasma-protein precipitation with plasma ultrafiltration and trichloroacetic acid protein precipitation for the measurement of unbound platinum concentrations. Cancer Chemother Pharmacol. 1996;38(4):391–4. <https://doi.org/10.1007/s002800050501>.
  31. Valdes-Balbin Y, Santana-Mederos D, Quintero L, Fernández S, Rodriguez L, Sanchez Ramirez B, et al. SARS-CoV-2 RBD-tetanus toxoid conjugate vaccine induces a strong neutralizing immunity in preclinical studies. ACS Chem Biol. 2021;16(7):1223–33. <https://doi.org/10.1021/acscchembio.1c00272>.
  32. Sanda M, Morrison L, Goldman R. N- and O-glycosylation of the SARS-CoV-2 spike protein. Anal Chem. 2021;93(4):2003–9. <https://doi.org/10.1021/acs.analchem.0c03173>.
  33. Shajahan A, Supekar NT, Gleinich AS, Azadi P. Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. Glycobiology. 2020;30(12):981–8. <https://doi.org/10.1093/glycob/cwaa042>.
  34. Mechref Y. Use of CID/ETD mass spectrometry to analyze glycopeptides. Curr Protoc Protein Sci. 2012;68(1):12.1. 1–1. 1. <https://doi.org/10.1002/0471140864.ps1211s68>.
  35. Gadgil HS, Bondarenko PV, Pipes GD, Dillon TM, Banks D, Abel J, et al. Identification of cysteinylolation of a free cysteine in the Fab region of a recombinant monoclonal IgG1 antibody using Lys-C limited proteolysis coupled with LC/MS analysis. Anal Biochem. 2006;355(2):165–74. <https://doi.org/10.1016/j.ab.2006.05.037>.
  36. Buchanan A, Clementel V, Woods R, Harn N, Bowen MA, Mo W, et al. Engineering a therapeutic IgG molecule to address cysteinylolation, aggregation and enhance thermal stability and expression. MAbs. 2013;5(2):255–62. <https://doi.org/10.4161/mabs.23392>.
  37. Banks DD, Gadgil HS, Pipes GD, Bondarenko PV, Hobbs V, Scavazza JL, et al. Removal of cysteinylolation from an unpaired sulfhydryl in the variable region of a recombinant monoclonal IgG1 antibody improves homogeneity, stability, and biological activity. J Pharm Sci. 2008;97(2):775–90. <https://doi.org/10.1002/jps.21014>.
  38. Bayer M, König S. Abundant cysteine side reactions in traditional buffers interfere with the analysis of posttranslational modifications and protein quantification—how to compromise. Rapid Commun Mass Spectrom. 2016;30(15):1823–8. <https://doi.org/10.1002/rcm.7613>.
  39. Kim HJ, Ha S, Lee HY, Lee KJ. ROSics: chemistry and proteomics of cysteine modifications in redox biology. Mass Spectrom Rev. 2015;34(2):184–208. <https://doi.org/10.1002/mas.21430>.
  40. Moya G, Gonzalez LJ, Huerta V, Garcia Y, Morera V, Perez D, et al. Isolation and characterization of modified species of a mutated (Cys125–Ala) recombinant human interleukin-2. J Chromatogr A. 2002;971(1–2):129–42. [https://doi.org/10.1016/S0021-9673\(02\)00845-2](https://doi.org/10.1016/S0021-9673(02)00845-2).
  41. Junutula JR, Bhakta S, Raab H, Ervin KE, Eigenbrot C, Vandlen R, et al. Rapid identification of reactive cysteine residues for site-specific labeling of antibody-Fabs. J Immunol Methods. 2008;332(1–2):41–52. <https://doi.org/10.1016/j.jim.2007.12.011>.
  42. Stimmel JB, Merrill BM, Kuyper LF, Moxham CP, Hutchins JT, Fling ME, et al. Site-specific conjugation on serine→cysteine variant monoclonal antibodies. J Biol Chem. 2000;275(39):30445–50. <https://doi.org/10.1074/jbc.M001672200>.
  43. Chen X, Nguyen M, Jacobson F, Ouyang J, editors. Charge-based analysis of antibodies with engineered cysteines: from multiple peaks to a single main peak. MAbs; 2009: Taylor & Francis.
  44. Tang HY, Speicher DW. Experimental assignment of disulfide-bonds in purified proteins. Curr Protoc Protein Sci. 2019;96(1):e86. <https://doi.org/10.1002/cpps.86>.
  45. Nickerson JL, Doucette AA. Rapid and quantitative protein precipitation for proteome analysis by mass spectrometry. J Proteome

- Res. 2020;19(5):2035–42. <https://doi.org/10.1021/acs.jproteome.9b00867>.
46. Dai L, Zheng T, Xu K, Han Y, Xu L, Huang E, et al. A universal design of betacoronavirus vaccines against COVID-19, MERS, and SARS. *Cell*. 2020;182(3):722–33. e11. <https://doi.org/10.1016/j.cell.2020.06.035>.
47. Kurjan J, Herskowitz I. Structure of a yeast pheromone gene (MF $\alpha$ ): a putative  $\alpha$ -factor precursor contains four tandem copies of mature  $\alpha$ -factor. *Cell*. 1982;30(3):933–43. [https://doi.org/10.1016/0092-8674\(82\)90298-7](https://doi.org/10.1016/0092-8674(82)90298-7).
48. Lin-Cereghino GP, Stark CM, Kim D, Chang J, Shaheen N, Poerwanto H, et al. The effect of  $\alpha$ -mating factor secretion signal mutations on recombinant protein expression in *Pichia pastoris*. *Gene*. 2013;519(2):311–7. <https://doi.org/10.1016/j.gene.2013.01.062>.
49. Brake AJ, Merryweather JP, Coit DG, Heberlein UA, Masiarz FR, Mullenbach GT, et al. Alpha-factor-directed synthesis and secretion of mature foreign proteins in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA*. 1984;81(15):4642–6. <https://doi.org/10.1073/pnas.81.15.4642>.
50. Simpson DM, Beynon RJ. Acetone precipitation of proteins and the modification of peptides. *J Proteome Res*. 2010;9(1):444–50. <https://doi.org/10.1021/pr900806x>.
51. Kratz H, Haeckel A, Michel R, Schönzart L, Hanisch U, Hamm B, et al. Straightforward thiol-mediated protein labelling with DTPA: synthesis of a highly active 111 In-annexin A5-DTPA tracer. *EJN-MMI Res*. 2012;2(1):17. <https://doi.org/10.1186/2191-219X-2-17>.
52. Suttapitugsakul S, Xiao H, Smeekens J, Wu R. Evaluation and optimization of reduction and alkylation methods to maximize peptide identification with MS-based proteomics. *Mol Biosyst*. 2017;13(12):2574–82. <https://doi.org/10.1039/C7MB00393E>.
53. Boyatzis AE, Bringans SD, Piggott MJ, Duong MN, Lipscombe RJ, Arthur PG. Limiting the hydrolysis and oxidation of maleimide–peptide adducts improves detection of protein thiol oxidation. *J Proteome Res*. 2017;16(5):2004–15. <https://doi.org/10.1021/acs.jproteome.6b01060>.
54. Guo M, Weng G, Yin D, Hu X, Han J, Du Y, et al. Identification of the over alkylation sites of a protein by IAM in MALDI-TOF/TOF tandem mass spectrometry. *RSC Adv*. 2015;5(125):103662–8. <https://doi.org/10.1039/C5RA18595E>.
55. Kuznetsova KG, Levitsky LI, Pyatnitskiy MA, Ilina IY, Bubis JA, Solovyeva EM, et al. Cysteine alkylation methods in shotgun proteomics and their possible effects on methionine residues. *J Proteomics*. 2021;231:104022. <https://doi.org/10.1016/j.jprot.2020.104022>.
56. Müller T, Winter D. Systematic evaluation of protein reduction and alkylation reveals massive unspecific side effects by iodine-containing reagents. *Mol Cell Proteomics*. 2017;16(7):1173–87. <https://doi.org/10.1074/mcp.M116.064048>.
57. Limonta-Fernández M, Chinea-Santiago G, Martín-Dunn AM, Gonzalez-Roche D, Bequet-Romero M, Marquez-Perera G, et al. The SARS-CoV-2 receptor-binding domain expressed in *Pichia pastoris* as a candidate vaccine antigen. medRxiv. 2021. <https://doi.org/10.1101/2021.06.29.21259605>.
58. Zhu Z, Meng K, Meng G. Genomic recombination events may reveal the evolution of coronavirus and the origin of SARS-CoV-2. *Sci Rep*. 2020;10(1):1–10. <https://doi.org/10.1038/s41598-020-78703-6>.
59. Cele S, Gazy I, Jackson L, Hwa S-H, Tegally H, Lustig G, et al. Escape of SARS-CoV-2 501Y. V2 from neutralization by convalescent plasma. *Nature*. 2021;593(7857):142–6. <https://doi.org/10.1038/s41586-021-03471-w>.
60. Varki A, Cummings RD, Aebi M, Packer NH, Seeberger PH, Esko JD, et al. Symbol nomenclature for graphical representations of glycans. *Glycobiology*. 2015;25(12):1323–4. <https://doi.org/10.1093/glycob/cwv091>.
61. Mormann M, Eble J, Schwöppe C, Mesters RM, Berdel WE, Peter-Katalinić J, et al. Fragmentation of intra-peptide and inter-peptide disulfide bonds of proteolytic peptides by nanoESI collision-induced dissociation. *Anal Bioanal Chem*. 2008;392(5):831–8. <https://doi.org/10.1007/s00216-008-2258-7>.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.