2022

IDENTIFICATION OF NOVEL PHOSPHORYLATION SITES IN SARS-CoV-2 NUCLEOCAPSID PROTEIN

Josie Daldegan Rezende
Saint John's University, Jamaica New York

IDENTIFICATION OF NOVEL PHOSPHORYLATION SITES IN SARS-CoV-2 NUCLEOCAPSID PROTEIN

A thesis submitted in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

to the faculty of the

DEPARTMENT OF CHEMISTRY

of

ST. JOHN'S COLLEGE OF LIBERAL ARTS AND SCIENCES

at

ST. JOHN'S UNIVERSITY
New York
by
Josie Daldegan Rezende

Date Submitted 4/21/2022 Date Approved 5/17/2022

_____	_____
Josie Daldegan Rezende	Dr. Erica Jacobs

ABSTRACT

IDENTIFICATION OF NOVEL PHOSPHORYLATION SITES IN SARS-CoV-2 NUCLEOCAPSID PROTEIN

Josie Daldegan Rezende

COVID-19, the 2019 disease outbreak, continues to cause severe mortality worldwide. Fast and reliable tests that identify the SARS-CoV-2 virus, which is responsible for causing the COVID-19 disease, are crucial for controlling the spread of the virus. Antigen tests are ideal for the rapid diagnosis of COVID-19. Current antigen tests for COVID-19 detect the SARS-CoV-2 nucleocapsid (N) protein. The nucleocapsid is the most abundant protein produced by the virus upon infection, which makes it a superior target for detection by an antigen test. Nucleocapsid proteins from several coronaviruses, including SARS, have previously been shown to be modified by posttranslational modification (PTM), including phosphorylation. PTMs are important mechanisms to increase the functional diversity of the virus within its host cell and modulate the antiviral response. Antigen tests require an antibody to bind to nucleocapsid protein, however, phosphorylation may differ between the nucleocapsid used to elicit the diagnostic antibody and phosphorylation on the viral protein in an infected person. In this research, his-tagged SARS-CoV-2 nucleocapsid protein was expressed in insect cells using a baculovirus expression system and analysis of purified samples using bottom-up proteomics approach followed by ESI-MS/MS along with MALDI-TOF analysis was used to localize phosphorylation sites in the SARS-CoV-2 nucleocapsid protein, which

may benefit the development of better antigen tests. Results from MALDI shows that N protein is phosphorylated by an average of 13 phosphate groups. A total of 12 phosphopeptides were identified by ESI-MS/MS, of which 3 are novel sites not previously reported (T135, T265 and T329).

ACKNOWLEDGEMENTS

I would like to acknowledge and give a special thanks to my supervisor and mentor Dr. Erica Jacobs who made this work possible and provided expertise and guidance throughout this research. Many thanks to the rest of my thesis committee members, Dr. Francisco Vazquez, and Dr. Kona Arachchila Senevirathne, for their insightful comments and feedback. I would also like to thank the Department of Chemistry for the support to the success of this study.

I gratefully acknowledge Paul Dominic B. Olinares and Brain T. Chait from the Laboratory of Mass Spectrometry and Gaseous Ion Chemistry at Rockefeller University, also Jennifer Aguilan and Simone Sidoli from Albert Einstein College of Medicine for advice and use of their instrumentations much needed to accomplish this research.

Finally, to my caring and loving family and friends my deepest gratitude. The support, encouragement, and sacrifices when times get rough are much appreciated and noted.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ---------- ii

LIST OF TABLES ---------- v

LIST OF FIGURES ---------- vi

CHAPTER 1: INTRODUCTION ---------- 1

1.1 COVID-19 disease ---------- 1

1.2 COVID-19 Testing ---------- 1

1.3 Antigen Test ---------- 3

1.4 SARS-CoV-2 Virology ---------- 3

1.5 Posttranslational modification ---------- 5

CHAPTER 2: EXPERIMENTAL METHODS ---------- 7

2.1 Recombinant N protein purification using IMAC column ---------- 7

2.2 Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) ---------- 8

2.3 In-gel digestion and peptide purification ---------- 8

2.4 ESI Mass spectrometry ---------- 9

2.5 Raw data processing ---------- 10

2.6 In-solution phosphatase treatment and Lys-C digestion ---------- 10

2.7 Ultra-thin layer ---------- 11

2.8 MALDI Mass spectrometry ---------- 11

CHAPTER 3: RESULTS AND DISCUSSION ---------- 13

3.1 SARS-CoV-2 SR-rich domain is predicted to be highly phosphorylated ----------13

3.2 Overall phosphorylation state of recombinant SARS-CoV-2 N protein via MALDI-MS ----------16

3.3 Identification of phosphorylated sites on recombinant SARS-CoV-2 N protein via ESI-MS and bottom-up analysis ----------22

3.4 Autoproteolysis occurrence on SARS-CoV-2 N protein----------29

CHAPTER 4: CONCLUSION AND FUTURE WORK ----------31

4.1 Conclusion----------31

4.2 Future Work ----------32

REFERENCES ----------34

LIST OF TABLES

Table 1: Enzyme concentration and amount used in different protein digestion procedures. ---------- 9

LIST OF FIGURES

Figure 1: Coronavirus structure and SARS-CoV-2 Nucleocapsid protein ---------- 5

Figure 2: Protein Multiple Sequence Alignment using Clustal Omega ----------14

Figure 3: NetPhos 3.1 phosphorylation site prediction ----------16

Figure 4: MALDI results of intact nucleocapsid protein ----------19

Figure 5: MALDI spectra of peptides from recombinant SARS-CoV-2 N protein digestion with Lys-C----------21

Figure 6: Expasy peptide sequence prediction from Lys-C digestion ----------22

Figure 7: ESI-MS/MS spectra showing the phosphorylation amino acid along with their peptide sequence ----------27

Figure 8: Comparison of HCD and ETD fragmentation of phosphopeptides from SARS-CoV-2 N protein SR-rich domain----------28

Figure 9: Phosphorylated sites and peptide sequences identified from each enzymatic digestion procedure ----------28

Figure 10: Phosphorylation sites found in SARS-CoV-2 N protein within the protein domains and amino acid sequence----------29

CHAPTER 1: INTRODUCTION

1.1 COVID-19 disease

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) virus causes Coronavirus disease 2019 (COVID-19). It was first reported in China by the World Health Organization (WHO) [1]. SARS-CoV-2 is highly infectious, causing severe mortality worldwide and rapidly developing into a pandemic. By January 1st, 2022, SARS-CoV-2 had infected over 36 million people worldwide and killed more than 691,000 individuals [2]. Although effective vaccines against SARS-CoV-2 have recently been developed, a better understanding of the epidemiology and the viral pathogenesis of SARS-CoV-2 infections in humans is still urgently needed as a basis for the development of effective products for diagnosing, treating, preventing, and controlling the spread of COVID-19 infectious disease [3].

1.2 COVID-19 Testing

The data provided from testing the population and the number of tests that are been performed provides us with a perspective into the pandemic and how the virus is spreading. Without testing the true number of infections may be far higher than the number of confirmed cases, according to source from February 2022 [2]. Fast, sensitive, and accurate early diagnostic testing allows us to identify infected individuals and provide a quick response for prevention and treatment of the disease. To this extent, the spread of the virus can be better controlled, and the acute phase of the pandemic could end [3]. Early availability of diagnostic testing plays an important role in patient management and public health [3].

There are currently three main types of tests to detect the SARS-CoV-2 virus. The molecular test reverse transcription-polymerase chain reaction (RT-PCR) test is extremely specific, and it is the most accurate test, with the lowest limit of detection. In order to test the genetic information of the virus in the laboratory, a respiratory or saliva sample from the person being tested would be taken and analyzed for potential infection of SARS-CoV 2 [4]. However, this test has some limitations. First, the results can take a few hours to more than a week to be ready depending on the location and resources, which is a drawback for early diagnostic testing. Second, it has a high cost due to the complexity of the technique requiring highly trained professionals and high-cost equipment. Also, false-negative amplification results might occur from cross-react with the non-specific nucleic acid. An alternative diagnostic approach, the antibody test, also known as the serology test, is a fast diagnosis test. It detects antibodies that the immune system develops in response to the viral infection, but it does not necessarily detect the presence of the virus itself. Our bodies can take several weeks to develop antibodies against the virus, and it can remain in the body for an undetermined amount of time. So, although the antibody test is a quick test, it does not diagnose an active infection.

The antigen test, by contrast, detects a specific viral antigen providing an indication of active coronavirus infection. In general, the test is cheaper and much faster than the molecular RT-PCR test. The accessibility of the various home antigen tests makes it easier for individuals to test themselves more often and without presenting to medical care or a testing center [2]. The COVID-19 antigen test is therefore ideal for early diagnostic testing of SARS-CoV-2.

1.3 Antigen Test

Currently, most COVID-19 rapid antigen tests on the market target SARS-CoV-2 nucleocapsid (N) protein. SARS-CoV-2 N protein is the most abundant viral protein produced within the infected cells [5]. The presence of IgA, IgM, and IgG antibodies against nucleocapsid antigen have been confirmed by immunoblot assay on patients infected with COVID-19 [6]. Research results show that SARS-CoV-2 nucleocapsid protein has a high specific antigen detection of 100%, and a sensitivity of 75% compared to the nucleic acid test, in the early phase of infection [7]. Although the antigen test is highly specific, it is usually less sensitive than RT-PCR test, which can lead to potential false negative results or less accurate results overall [8][9]. Because of its abundance and antigenicity, the SARS-CoV-2 nucleocapsid protein is an ideal target for early antigen diagnostic detection for COVID-19 infection. However, improvements need to be made to increase antigen test sensitivity for better accuracy of results. Epitope binding of antibodies and therefore, antigen test sensitivity may be influenced by posttranslational modification (PTM) of the antigen. Identification, quantification, and characterization of PTMs on SARS-CoV-2 nucleocapsid protein can help to understand antigen binding, molecular recognition, and the development of new types of antibodies that are more specific, have higher affinities and are not affected by differences in posttranslational modifications between the viral antigen and the recombinant antigen typically used to elicit diagnostic antibodies for the antigen test.

1.4 SARS-CoV-2 Virology

SARS-CoV-2, which has a long RNA genome size of about 29.9 kb [10], belongs to the betacoronavirus genus, one of the four genera of coronavirus. The virus consists of an enveloped, single-stranded positive-sense RNA coated with nucleocapsid protein and

is packed with three structural proteins including membrane (M) protein, envelope (E) protein and spike (S) protein. The genome also encodes 16 non-structural proteins (nsp1-16), which contribute to survival, signaling, and replication of the virus as well as 7 accessory proteins responsible for attenuating host response against SARS-CoV-2 [10][11].

The spike protein is responsible for fusion with the host cell membrane, allowing the virus to penetrate the host cell causing infection. Coronavirus nucleocapsid protein, as previously reported from other viruses, plays a key role in viral replication, transcription, and translation [5][12]. N protein is essential for the SARS-CoV-2 life cycle and plays critical roles in viral RNA synthesis as a multifunctional protein [13]. Spike and nucleocapsid protein antigens from SARS-CoV-2 are highly immunogenic [14][15], and they are the target for diagnosis and treatment of COVID-19. However, the spike protein is more prone to mutations [16] that affect antibody binding, with several cases already reported [17][18], which is a drawback for the development of diagnosis and treatment of COVID-19.

The SARS-CoV-2 N protein is 419 amino acids in length and predicted to be 46 KDa in mass, and it binds to the viral RNA genome via its N-terminal domain [6]. The N protein contains two domains, an N-terminal RNA binding domain (NTD), and a C-terminal oligomerization domain (CTD), which are connected by linker containing a serine-arginine rich region (SR-rich; 176–204 aa), **Figure 1**B [13]. Furthermore, there is also an interaction between N protein and coronavirus M (membrane) protein mediated by the serine-arginine rich domain [19].

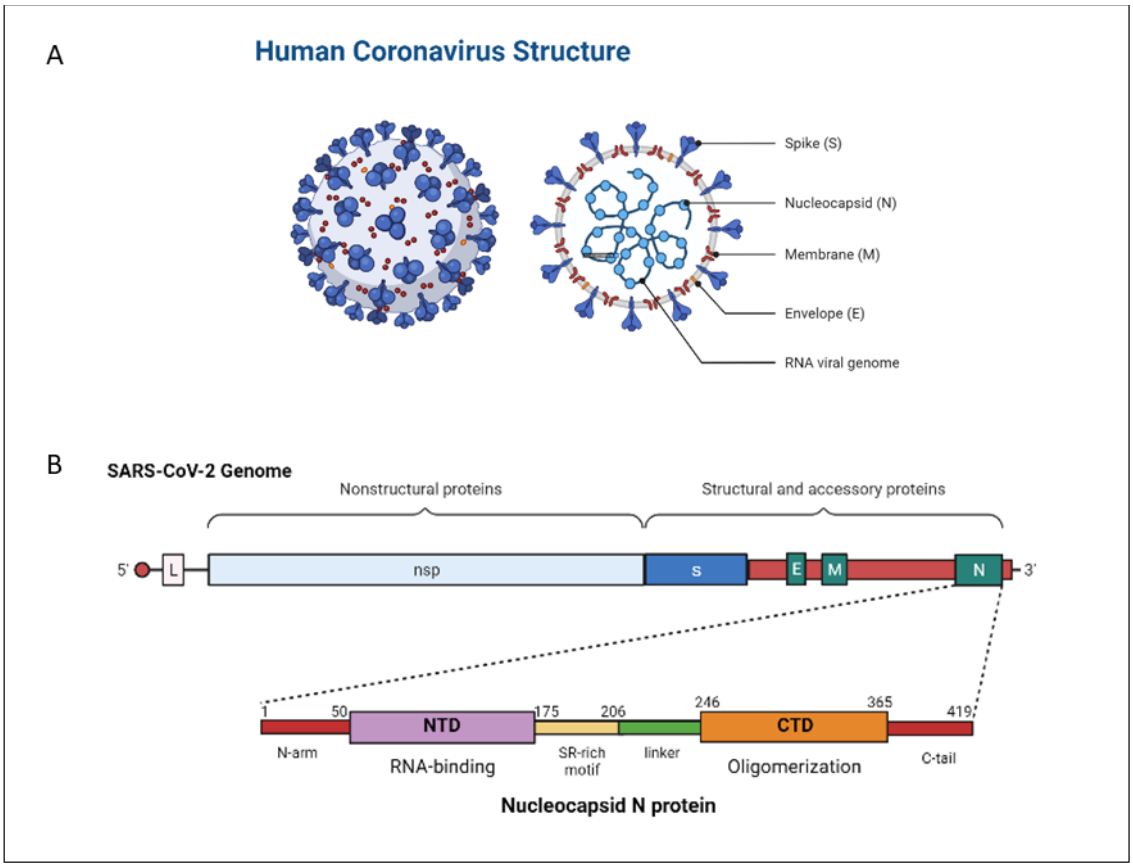


Figure 1: Coronavirus structure and SARS-CoV-2 Nucleocapsid protein. *(A) Structure of human SARS-CoV-2 coronavirus virion. (Reprinted from "Human Coronavirus Structure", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates). (B) Schematic diagram of SARS-CoV-2 genome structure and domain structure of the 419 amino acid N protein (not to scale) with NTD domain in purple which represent the RNA-binding region, the SR-rich motif in yellow, and the CTD domain in orange which represents the oligomerization region. (Adapted from "Genome Organization of SARS-CoV", by BioRender.com (2022). Retrieved from https://app.biorender.com/biorender-templates).*

1.5 Posttranslational modification

Posttranslational modifications (PTM) of proteins involve the covalent addition of functional groups, usually via enzymatic catalysis. Phosphorylation, which is the addition of a phosphate group, usually to a serine, threonine, or tyrosine residue, is an essential posttranslational modification. Antigenicity of the recombinant nucleocapsid protein has been shown to be closely linked to nucleocapsid phosphorylation [20]. Dephosphorylation of coronavirus nucleocapsid protein increases its binding affinity to viral RNA over cellular RNA. Phosphorylation of coronavirus nucleocapsid protein,

regulated by kinases, is required to release RNA for viral infection [21]. Phosphorylation and dephosphorylation of coronavirus nucleocapsid protein control the viral life cycle in the host cell [22]. SARS-CoV-2 infection promotes activation and production of a kinase that arrests the cell cycle and rewires the phosphorylation status of both host and viral proteins [23].

The coronavirus N protein is mainly modified by phosphorylation, but modifications by arginine methylation [24], proteolytic cleavage, and ADP-ribosylation [25] were previously detected as well. Nucleocapsid protein from other viruses are often also phosphorylated. Phosphorylation of coronavirus N protein was identified previously in the Mouse hepatitis virus (MHV) [22][26], the Transmissible gastroenteritis virus (TGEV) [27], the Infectious bronchitis virus (IBV) [21], and the severe acute respiratory syndrome coronavirus (SARS-CoV) [28][29].

This research aims to map phosphorylation sites of SARS-CoV-2 Nucleocapsid protein, using tagged recombinant protein expressed in insect cells. The previous literature along with computational predictions suggest that N protein is highly phosphorylated. This study utilizes mass spectrometric analysis of intact N protein and peptides using matrix-assisted laser desorption/ionization (MALDI-MS). Additionally, a bottom-up proteomic approach was used to identify phosphorylated sites using electrospray ionization mass spectrometry (ESI-MS). MALDI-MS results show an average of 13 phosphorylated sites in recombinant SARS-CoV-2 N protein and ESI-MS results confirmed 12 phosphorylated serines and threonines including 3 novel phosphorylated sites.

CHAPTER 2: EXPERIMENTAL METHODS

2.1 Recombinant N protein purification using IMAC column

Recombinant SARS-CoV-2 Nucleocapsid protein expressed in baculovirus using Bac-to-Bac™ Baculovirus Expression System (Thermo Fisher) was purified from insect cells. The Nucleocapsid protein was cloned into two versions, one version with a 7x histidine-tag and green fluorescent protein (GFP) tag on the C-terminus and the second version with only a 7x histidine-tag on the C-terminus. The recombinant Nucleocapsid protein was then affinity purified using Co-NTA (cobalt-nitrilotriacetic acid) resin as follows; the 10ml cell lysate was harvested by centrifugation at 12,000 rpm for 25 min at 4°C after thawing from −80°C on ice and adding 100 µl of protease inhibitor solution (18 mg/ml phenylmethylsulphonyl fluoride 0.3 mg/ml pepstatin A in ethanol). 100 µl of 1 M HEPES 7.4 was added in order to raise the pH to about 7. The supernatant was loaded onto a 750 µl HisPur™ Cobalt Resin column, equilibrated with 5000ul equilibration buffer (20 mM HEPES [pH 7.4], 150 mM NaCl). After being incubated for 45 min with tilting on ice, extract was allowed to flow through, and the column with resin was washed two times with 10 ml of 20 mM HEPES pH 7.4, 1.25 M NaCl buffer, and two times with 10 ml of buffer 20 mM HEPES 7.4, 150 mM NaCl, 10 mM imidazole buffer. Purified N protein was sequentially eluted from the column two or three times with 1 ml of 20 mM HEPES 7.4, 150 mM NaCl, 250 mM imidazole buffer. The samples were collected, and fractions were analyzed by SDS-PAGE. A buffer exchange step (Nanosep® Centrifugal Devices with Omega™ Membrane 10K) was used to desalt and to increase the concentration of the N protein fractions.

2.2 Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)

Buffer-exchanged affinity chromatography purified N protein preparations were separated by SDS-PAGE. 4x gel loading buffer was added to samples to a final concentration of 1x. Samples were then heated to 70°C for 5 min and 1/20 volume of 0.5 M iodoacetamide was added and samples were incubated at room temperature for 10 min in the dark. The purpose of this procedure was to denature the N protein. Samples were loaded onto precast polyacrylamide gels (Invitrogen NuPAGE™ 4 to 12%, Bis-Tris, 1.0 mm, Mini Protein Gel, 15-well) and electrophoresis was performed using 1x MOPS running buffer and 500 µl Invitrogen NuPAGE™ Antioxidant at 190 volts for 45 min. Precision Plus Protein Standards™ Dual Color (BioRad) were used as a marker. Gels were stained using colloidal Coomassie dye G-250 (GelCode™ Blue Stain Reagent, Thermo Fisher) on a rocking platform overnight and distained in water. N protein band was compared to BSA standards to estimate the protein concentration on the gel using a public domain Java image processing program, Image J [30].

2.3 In-gel digestion and peptide purification

The N protein bands were cut out and diced into small pieces, with each tube receiving ~3 µg of N protein. Gel pieces were de-stained until clear using 50 mM ammonium bicarbonate (ABC) and 50% acetonitrile (ACN) and dehydrated with 100% ACN. Aspiration was used to remove ACN, and gel pieces were air dried. Dried gel pieces were rehydrated with 10 µl of the indicated concentration of Lys-C enzyme or a combination of Lys-C and trypsin according to Table 1. After 10 min on ice, 50 µl of Lys-C buffer (25 mM Tris-HCl, pH 8.0, 1 mM EDTA) was added and incubated at 37°C according to time indicated in Table 1. After removal of supernatant, additional

extraction was performed by shaking the gel pieces with 175 µl of peptide extraction buffer (1.67% FA, 67% ACN, 0.05% TFA), followed by 100 µl of 0.1% FA. Then, extracted peptides were purified over reverse-phase C18 column packed into a microcapillary tube, washed three times with 50 µl of 0.1% TFA, 90% ACN with 0.1% TFA and 0.1% acetic acid respectively [31]. Samples were sequentially eluted with 50 µl 40% ACN, 0.1% acetic acid and then with 90% ACN, 0.1% acetic acid.

Procedure	Enzyme concentration	Enzyme amount	incubation time	incubation temperature
1	Trp 0.05 µg/µl	10 µl	3 hours	37°C
2	LysC 0.05 µg/µl	10 µl	3 hours	
	LysC 0.05 µg/µl + **Trp 0.00005 µg/µl**			
	LysC 0.05 µg/µl + **Trp 0.00005 µg/µl**		5 hours	
3	LysC 0.05 µg/µl	10 µl	3 hours	
	LysC 0.05 µg/µl + **Trp 0.00005 µg/µl**			
	LysC 0.05 µg/µl + **Trp 0.0005 µg/µl**			
	LysC 0.05 µg/µl + **Trp 0.005 uµg/µl**			

Table 1: Enzyme concentration and amount used in different protein digestion procedures.

2.4 ESI Mass spectrometry

Samples were analyzed with a nano-LC 1200 (Thermo Fisher) using a EASYspray PepMap RSLC C18 3 micron, 100 Å, 75 µm by 15 cm column using a Orbitrap Lumos instrument (Thermo Fisher) and separated using a 120 minute gradient.

For HCD fragmentation experiments, the instrument was operated in data-dependent mode, and the most abundant ions were selected in each full scan for the maximum parallelizable time and sequentially fragmented by high-energy collisional dissociation using a normalized collision energy of 28%; dynamic exclusion was enabled. Target resolution was 120,000 for MS. The quadrupole isolation window was 1.4 m/z,

and the MS/MS used a maximum injection time of 250 ms with 1 microscan and a minimum intensity threshold of 5E3. Charge states were limited to between 2-6.

For ETD experiments, the instrument was also operated in data-dependent mode, and the most abundant ions were selected in each full scan for the maximum parallelizable time and sequentially fragmented by electron transfer dissociation with a reaction time of 50 ms, a reagent target of 1E6, and a maximum reagent injection time of 200 ms; dynamic exclusion was enabled. Target resolution was 120,000 for MS. The quadrupole isolation window was 1.4 m/z, and the MS/MS used a maximum injection time of 250 ms with 1 microscan and a minimum intensity threshold of 5E3. Charge states were limited to between 4-8.

2.5 Raw data processing

MaxQuant (v 2.0.3.0) was used to process all ESI-MS raw data, with the recombinant N protein sequence and a *Trichoplusia ni* (*T. ni*) protein sequence database from the insect host cells. MaxQuant default parameters were used with the exception that label-free quantification was turned on. Numbers of modifications allowed per peptide was set to 8.

2.6 In-solution phosphatase treatment and Lys-C digestion

Purified recombinant SARS-CoV-2 N protein was dephosphorylated using CIAP (Calf Intestinal Alkaline Phosphatase) (Invitrogen™) and recommended manufacture's protocol was performed with 1 hour and 30 min incubation at 37°C. Purified recombinant SARS-CoV-2 N protein was also dephosphorylated using lambda protein phosphatase (400,000 U/ml, New England Bio Labs Inc.) with 1 hour and 30min incubation at 30°C following manufacture's protocol. Protein digestion with Lys-C was performed on

sample treated with lambda protein phosphatase and buffer exchanged into Lys-C digestion buffer. 10ul of 0.1µg/µl Lys-C was added to 30ul (~ 2 µg) of Nc protein concentrated from buffer exchange.

2.7 Ultra-thin layer

For the analysis and comparison of N protein and peptides, MALDI-TOF/TOF-MS (Bruker Ultraflextreme) was used. The preparation for the matrix-assisted laser desorption was done following the ultra-thin layer method [32]. Intact and dephosphorylated protein and peptides had matrix crystals incorporated into them prepared with α-Cyano-4-hydroxycinnamic acid (4-HCCA). The thin layer substrate solution (1 part of saturated 4-HCCA and 3 parts of isopropanol) was added to a cleaned MALDI plate. The matrix solution was prepared with 1:3:2 FWI (formic acid, water, and isopropanol) and saturated 4-HCCA. The analyte solutions were diluted in matrix solution to a final concentration of ~ 2.5 µM. 1µl of each sample/matrix solution was spotted on the plate to crystallize of matrix and analyte and the excess liquid was aspirated with a vacuum line. 0.5µl of myoglobin/matrix mixture was used as a calibrant and applied in a similar manner with a final concentration of 0.1µM. The spots were washed twice with 2 µl of ice cold 0.1% TFA solution aspirating the excess liquid with a vacuum line.

2.8 MALDI Mass spectrometry

Samples spotted on an MTP 384 well polished steel target plate were analyzed using the Ultraflextreme MALDI-TOF/TOF mass spectrometer (Bruker Daltonics, Billerica, MA, USA) equipped with a smart beam-II laser set at 1,000 Hz repetition rate. The MS spectra were externally calibrated with Protein Calibration Standard I. The Lys-

C digested Nc peptides were analyzed from 5,000-10,000 m/z range while the intact Nc proteins were analyzed from 40,000-60,000 m/z range in the linear positive mode with matrix suppression set at 4,500 and 55,000 Da respectively. A total of 3,000 accumulated shots were acquired for each sample. The following ion source parameters were used: ion source 1 at 25.00 kV; ion source 2: 23.30 kV; lens at 6.56 kV and pulsed ion extraction at 400 ns. MALDI-TOF spectra were acquired using the FlexControl 3.4 acquisition software while the raw data were processed using the FlexAnalysis 4.0 software.

CHAPTER 3: RESULTS AND DISCUSSION

3.1 SARS-CoV-2 SR-rich domain is predicted to be highly phosphorylated

A protein multiple sequence alignment performed on N protein from SARS-CoV-2, SARS-CoV, MHV, TGEV, and IBV using UniProt [33] is shown in Figure 2. Previously determined serine and threonine phosphorylation sites are highlighted in green. All the nucleocapsid proteins are phosphorylated in the serine-arginine rich domain, highlighted in yellow, as shown in previous studies [21][22][26][27][28][29].

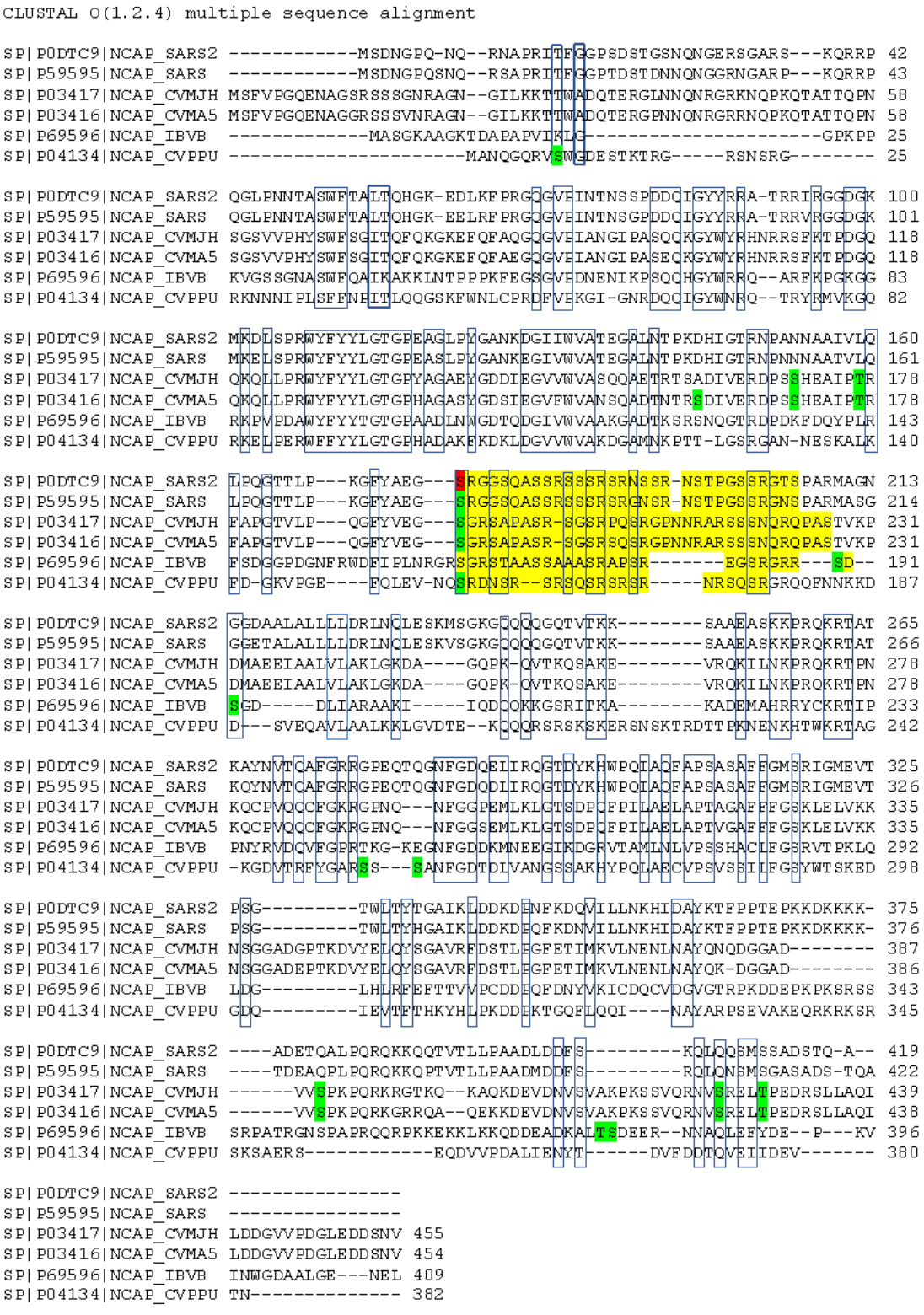


Figure 2: Protein Multiple Sequence Alignment using Clustal Omega [34]. *NCAP_SARS2: SARS-CoV-2 Nucleocapsid protein. NCAP_SARS: SARS-CoV Nucleocapsid protein* [28][29]. *NCAP_CVMJH: MHV-*

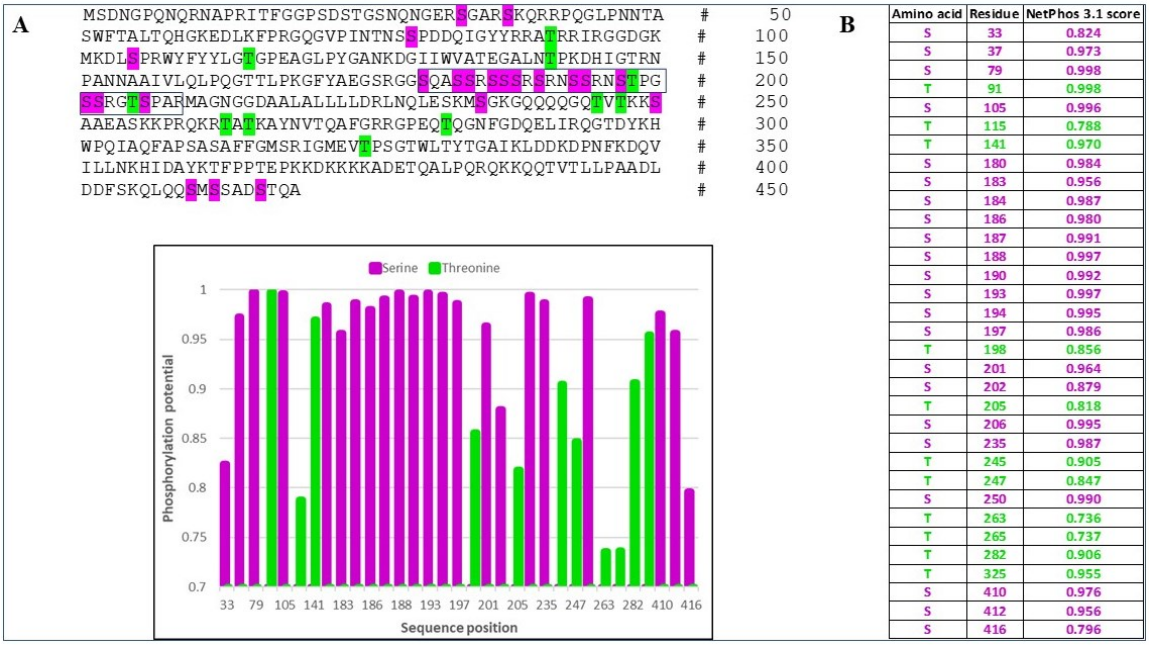
JHM Nucleocapsid protein [26]. *NCAP_CVMA5: MHV-A59 Nucleocapsid protein* [22]. *NCAP_IBVB: IBV Nucleocapsid protein* [21]. *NCAP_CVPPU: TGEV Nucleocapsid protein* [27]. *Identical and similar residues are boxed in blue lines. Previously experimentally identified phosphorylated amino acids are highlighted in green background. S176 residue (highlighted in red) was computationally predicted to be phosphorylated in UniProt* [33] *by homology to SARS. The serine-arginine rich domains are highlighted in yellow background.*

Molecular evolutionary research has shown that SARS-CoV-2 belongs to the beta coronavirus B lineage, in the same branch as SARS-CoV. As they belong to the same protein family, nucleocapsid protein from SARS-CoV and SARS-CoV-2 share 79% sequence identity [35] and 90% homology [10][36]. A posttranslationally modified residue on SARS-CoV nucleocapsid protein on serine position 176 (P0DTC9 · NCAP_SARS2) was previously reported in UniProt [33].

Bioinformatic approaches can be an important tool to predict phosphorylation of proteins. Computational analysis was therefore used to identify possible phosphorylation sites on N protein. A phosphorylation prediction program, NetPhos 3.1, analysis shows 22 possible serine phosphorylation sites and 11 possible threonine phosphorylation sites shown in Figure 3 [37][38]. Figure 3 shows a dense region of possible serine phosphorylation between amino acids 141 and 197, which is within the serine-arginine rich domain. In conclusion, the SR-rich domain is predicted to be highly phosphorylated, and some other sites outside the SR-rich domain are also predicted to be phosphorylated.

A comparison of previous experimental data from UniProt [33] with computational predictions from NetPhos 3.1 [38], both shows phosphorylation on serine 176. While NetPhos 3.1 [38] is based on computational analysis and predicted results, UniProt [33] results are based on actual data. The data from UniProt [33] and the

predictions from NetPhos 3.1 [38] provide positive indications for discovering SARS-CoV-2 nucleocapsid protein phosphorylation sites.



Amino acid	Residue	NetPhos 3.1 score
S	33	0.824
S	37	0.973
S	79	0.998
T	91	0.998
S	105	0.996
T	115	0.788
T	141	0.970
S	180	0.984
S	183	0.956
S	184	0.987
S	186	0.980
S	187	0.991
S	188	0.997
S	190	0.992
S	193	0.997
S	194	0.995
S	197	0.986
T	198	0.856
S	201	0.964
S	202	0.879
T	205	0.818
S	206	0.995
S	235	0.987
T	245	0.905
T	247	0.847
S	250	0.990
T	263	0.736
T	265	0.737
T	282	0.906
T	325	0.955
S	410	0.976
S	412	0.956
S	416	0.796

Figure 3: NetPhos 3.1 phosphorylation site prediction. *Predicted phosphorylation sites of SARS-CoV-2 Nucleocapsid protein on serine (in purple) and threonine (in green) amino acids. (A top) The SR-rich region is boxed in blue lines within the amino acid sequence. (A bottom) NetPhos 3.1 threshold score is 0.7. (B) Location of predicted phosphorylation sites of SARS-CoV-2 N protein with their corresponding NetPhos 3.1 scores are presented in a table* [38].

3.2 Overall phosphorylation state of recombinant SARS-CoV-2 N protein via MALDI-MS

Accurate molecular mass measurements via MALDI-TOF-MS can indicate the presence and identity of posttranslational modifications of proteins and peptides. The m/z values generated from MALDI-TOF/TOF-MS usually represent the true molecular mass as most of the ions have +1 charge. Intact recombinant SARS-CoV-2 N protein was compared before and after treatment with lambda phosphatase (Figure 4BC) and CIP phosphatase (Figure 4DE). Intact protein represents the full length N protein not proteolytically digested into peptides. Our his-tagged recombinant SARS-CoV-2 N protein has a predicted molecular weight of 47,467 Da [39], including the his-tag, which

matches well with our MALDI data of 47,578 Da from Figure 4E, and 47,354 Da from Figure 4B representing the dephosphorylated intact recombinant SARS-CoV-2 N protein.

Comparing the MALDI peaks between the samples that were treated with protein phosphatase and samples that were not treated with protein phosphatase (Figure 4) we can observe that the control samples, which were not treated with protein phosphatase, have a higher m/z than the samples treated with protein phosphatase. The mass difference between the samples can be related to the phosphorylation of N protein.

The m/z mass difference between the phosphorylated (Figure 4B) and dephosphorylated (Figure 4C) intact recombinant SARS-CoV-2 N protein using lambda phosphatase is 764 Da and the m/z mass difference between the phosphorylated (Figure 4D) and dephosphorylated (Figure 4E) intact recombinant SARS-CoV-2 N protein using calf intestinal phosphatase (CIP, Invitrogen) is 1040 Da. Assuming that the only difference between those two groups of samples is that in Figure 4BD the samples were not treated with protein phosphatase and in Figure 4CE the samples were treated with protein phosphatase in order to remove the phosphoryl groups, then we can conclude that the mass difference implies that the recombinant SARS-CoV-2 N protein is indeed phosphorylated. The loss of a phosphoryl group from the reaction of a phosphorylated serine and/or threonine with a protein phosphatase represent a shift in molecular mass of minus 80 Da [40][41]. The m/z discrepancy of 1040 Da thus reveals that the recombinant SARS-CoV-2 N protein treated with CIP phosphatase were modified by thirteen phosphoryl groups. The m/z discrepancy of 764 Da reveals that the recombinant SARS-CoV-2 N protein treated with lambda phosphatase were modified by at least nine phosphoryl groups.

Some considerations on those peaks should be taken into account. First, Figure 4E shows smooth peaks while the others MALDI shows peaks more jagged and uneven due to the difference in scale intensity (figure 4E has maximum intensity of 2.5x 10^4 and the other have maximum intensity of ~1.25 x 10^3). Furthermore, it is notable that Figure 4C shows a long shoulder on the right and on the left of the peak. Shoulder peaks are frequently observed as a result of the existence of two closely unresolved compounds. However, the possibility that this shoulder might represent heavier and lighter versions of the same compound cannot be disregarded. If this is the case, the shoulder peak will indicate versions of SARS-CoV-2 N protein with higher phosphorylation states as well with lower phosphorylation state. Even without considering the shoulder, this peak is broader than the peak in Figure 4E which also might represent different phosphorylated versions of SARS-CoV-2 N protein. Taken together, this evidence suggests that the lambda phosphatase-treated sample (Figure 4C) may not have been completely dephosphorylated. Another corroboration that supports this idea is the fact that the MALDI peak from the sample treated with lambda has a lower intensity compared to the sample treated with CIP. Phosphorylation suppresses the peak intensity in MALDI, and ESI [42] and the signal is further diluted by the presence of multiple phosphorylation sites.

Additionally, confirmation of efficacy of the dephosphorylation reaction using lambda can be interpreted from SDS-PAGE (Figure 4A) as the band from a sample treated with lambda phosphatase is lower than the band from a control sample not treated with lambda. The lower band represent samples with lower molecular weight due to the removal of phosphate groups from the sample treated with lambda phosphatase.

However, there is no proof that the reaction with lambda went to completion by removing all phosphate group from the N protein, based on SDS-PAGE results. The SDS-PAGE bands were observed to be slightly slanted perhaps due to the presence of high salt concentration in the phosphatase buffer.

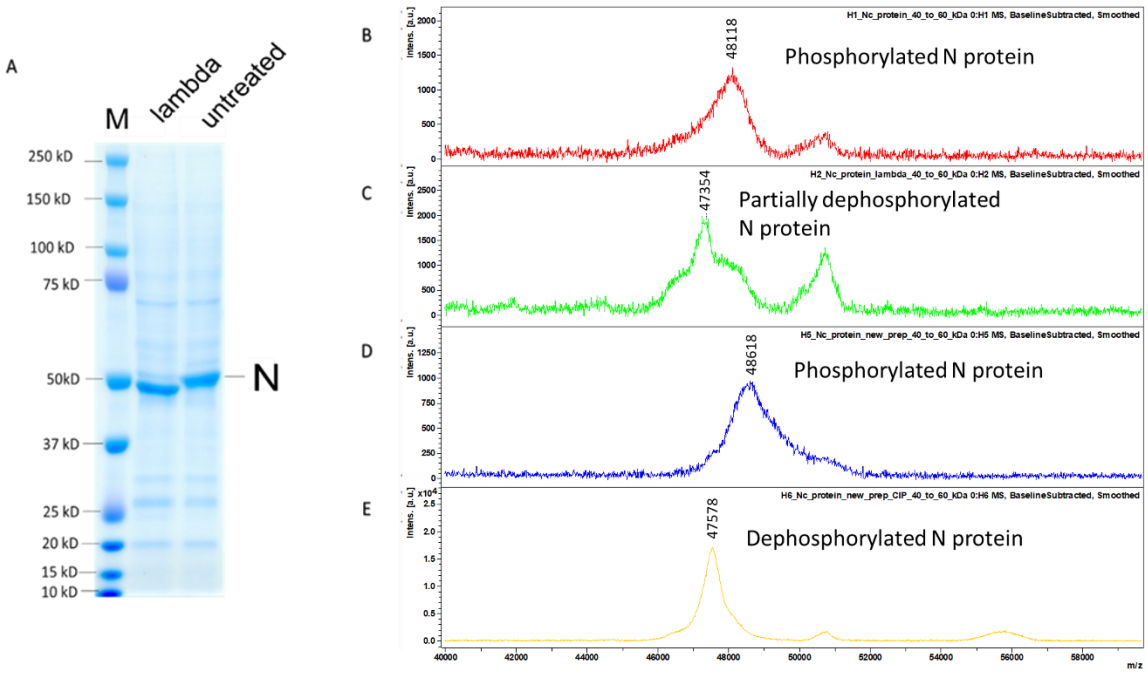


Figure 4: MALDI results of intact nucleocapsid protein. *(A) SDS-PAGE of recombinant SARS-CoV-2 N protein treated and untreated (control) with lambda protein phosphatase. (B) Intact recombinant SARS-CoV-2 N protein MALDI analysis before lambda phosphatase treatment. (C) Intact recombinant SARS-CoV-2 N protein MALDI analysis after lambda phosphatase treatment. (D) Intact recombinant SARS-CoV-2 N protein MALDI analysis before CIP phosphatase treatment. (E) Intact recombinant SARS-CoV-2 N protein MALDI analysis after CIP phosphatase treatment.*

Expasy results from in silico digestion of our his-tagged recombinant SAR-CoV-2 N protein with the protease Lys-C predicts the peptide GFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGDAALALLLLDRLNQLESK containing 64 amino acids, which includes the SR-rich domain, to have a peptide mass of 6544.07 Da (

Figure 6A), without posttranslational modifications. The Figure 5B represents the MALDI spectra of peptides of recombinant SAR-CoV-2 N protein from digestion with Lys-C and the partial lambda phosphatase treatment. This means that this sample will have peptides that are phosphorylated and peptides that are dephosphorylated as well. The dephosphorylated spectrum shows a peak with m/z of 6544.2 Da, which we can

conclude represents the SR-domain containing peptide, due its high accuracy match with the predicted mass from Expasy.

However, considerations about the peptide's formation from lambda protein digestion with Lys-C must also be addressed. The peak with m/z 7695.1 Da is shown in the sample treated with lambda phosphatase (Figure 5B), but the same peak is not shown in the intact sample, which was not treated with lambda protein phosphatase (Figure 5A). Additionally, this 7695.1 Da peak matches with one of the predicted peptide masses from lambda protein digestion with Lys-C equal to 7694.657 Da, highlighted in yellow on Figure **6**B. This led us to conclude that the peak is associated with a peptide from the lambda phosphatase digested with Lys-C. Phosphorylated peptide can not be observed in MALDI analysis of the Lys-C digested protein and this might be due the signal suppression from phosphorylation [42].

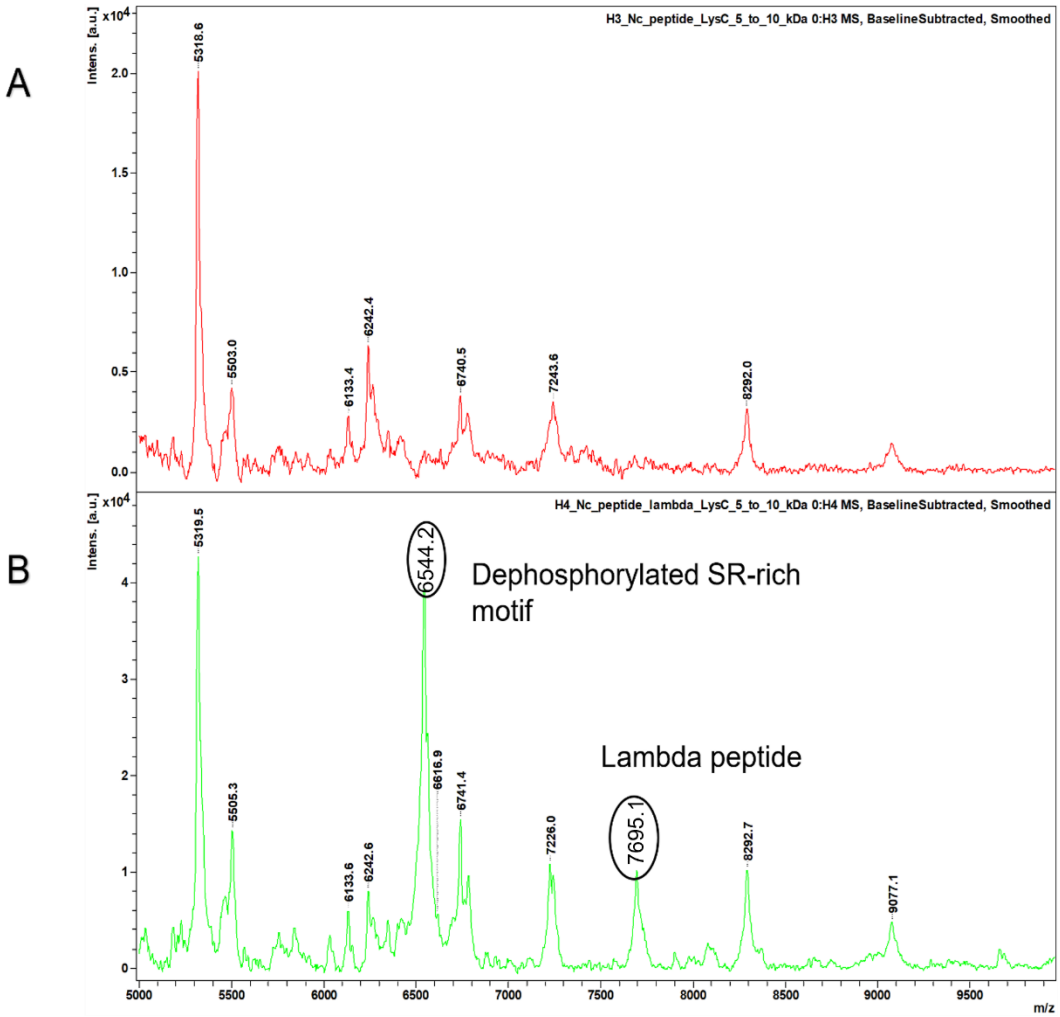


Figure 5: MALDI spectra of peptides from recombinant SARS-CoV-2 N protein digested with Lys-C. *(A) N protein not treated with lambda phosphatase. (B) N protein partially treated with lambda phosphatase.*

A

Position of cleavage site	Name of cleaving enzyme(s)	Resulting peptide sequence	Peptide length [aa]	Peptide mass [Da]
38	LysC	MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSK	38	3977.17
61	LysC	QRRPQGLPNNTASWFTALTQHGK	23	2608.902
65	LysC	EDLK	4	503.553
100	LysC	FPRGQGVPINTNSSPDDQIGYYRRATRRIRGGDGK	35	3906.295
102	LysC	MK	2	277.382
127	LysC	DLSPRWYFYYLGTGPEAGLPYGANK	25	2836.155
143	LysC	DGIIWVATEGALNTPK	16	1684.91
169	LysC	DHIGTRNPANNAAIVLQLPQGTTLPK	26	2740.115
233	LysC	GFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGDAALALLLLDRLNQLESK	64	6544.07
237	LysC	MSGK	4	421.512
248	LysC	GQQQQGQTVTK	11	1202.289
249	LysC	K	1	146.189
256	LysC	SAAEASK	7	662.698
257	LysC	K	1	146.189
261	LysC	PRQK	4	527.624
266	LysC	RTATK	5	575.666
299	LysC	AYNVTQAFGRRGPEQTQGNFGDQELIRQGTDYK	33	3746.027
338	LysC	HWPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAIK	39	4243.867
342	LysC	LDDK	4	489.526
347	LysC	DPNFK	5	619.675
355	LysC	DQVILLNK	8	942.123
361	LysC	HIDAYK	6	745.833
369	LysC	TFPPTEPK	8	916.042
370	LysC	K	1	146.189
372	LysC	DK	2	261.278
373	LysC	K	1	146.189
374	LysC	K	1	146.189
375	LysC	K	1	146.189
387	LysC	ADETQALPQRQK	12	1384.512
388	LysC	K	1	146.189
405	LysC	QQTVTLLPAADLDDFSK	17	1862.067
433	end of sequence	QLQQSMSSADSTQAENLYFQSHHHHHHH	28	3323.482

B

Position of cleavage site	Name of cleaving enzyme(s)	Resulting peptide sequence	Peptide length [aa]	Peptide mass [Da]
6	LysC	MRYYEK	6	889.037
11	LysC	IDGSK	5	518.567
31	LysC	YRNIWVVGDLHGCYTNLMNK	20	2396.767
40	LysC	LDTIGFDNK	9	1022.123
41	LysC	K	1	146.189
108	LysC	DLLISVGDLVDRGAENVECLELITFPWFRAVRGNHEQMMIDGLSERGNVNHWLLNGGGWFFNLDYDK	67	7694.66
113	LysC	EILAK	5	572.702
118	LysC	ALAHK	5	538.647
131	LysC	ADELPLIIELVSK	13	1439.712
133	LysC	DK	2	261.278
134	LysC	K	1	146.189
151	LysC	YVICHADYPFDEYEFGK	17	2096.298
174	LysC	PVDHQQVIWNRERISNSQNGIVK	23	2718.027
177	LysC	EIK	3	388.464
191	LysC	GADTFIFGHTPAVK	14	1460.653
194	LysC	PLK	3	356.465
221	end of sequence	FANQMYIDTGAVFCGNLTLIQVQGEGA	27	2861.237

Figure 6: Expasy peptide sequence prediction from Lys-C digestion. *(A) Expasy prediction of sequence and mass (Da) of peptides resulting from digestion of recombinant SARS-CoV-2 N protein with Lys-C protease. Highlighted in yellow is the largest peptide from the Lys-C digestion, containing 64 amino acids and a peptide mass of 6544.070 Da. (B) Expasy prediction of sequence and mass (Da) of peptides resulting from digestion of lambda protein with Lys-C. Highlighted in yellow is the peptide with mass 7694.657 Da.*

3.3 Identification of phosphorylated sites on recombinant SARS-CoV-2 N protein via ESI-MS and bottom-up analysis

In this study a recombinant SARS-CoV-2 N protein was expressed in eukaryotic cells, which allows for the possibility for it to be phosphorylated by endogenous eukaryotic kinases. Phosphorylation is a major posttranslational modification (PTM) where a phosphoryl group is added to a serine, threonine and /or tyrosine. In this reaction, the hydroxyl group from these residues is deprotonated, in basic conditions, which allows the negatively charged oxygen to act as a nucleophile and attack the electrophilic phosphorus atom and covalently add a phosphate group to the amino acid.

Qualitative and quantitative analysis of phosphorylation can be done using a bottom-up approach in which the protein is enzymatically digested into peptides followed by electrospray ionization tandem mass spectrometry (ESI-MS/MS). Three different procedures were used to proteolytically digest the purified recombinant N protein. In the first procedure, the samples were digested with trypsin only and in the second and third procedures the samples were digested with different amounts of Lys-C and trypsin combinations, according to Table 1. Trypsin cuts proteins after both lysine and arginine, while Lys-C only cuts after lysine. The SR domain is rich in arginine, making it unlikely that we could detect phosphorylation sites within this domain using only trypsin, since the peptides generated would be too small to be detected or for phosphorylation sites to be confidently assigned within them. In procedures 1 and 2, high-energy C-trap dissociation (HCD) was used to further fragment the peptides in an Orbitrap analyzer, which gives b ions from the amino terminus and y ions from the carboxylic terminus. In procedure 3, the peptides were fragmented using electron transfer dissociation (ETD), generating multiply charged c and z ions mostly. ESI-MS/MS is helpful for localizing phosphorylation sites since the generation of those ions from HCD and ETD fragmentation can reveal the modified residues. The identification of mass-to-charge ratios (m/z) from those ions allows the determination of the peptide sequence along with, in many cases, the specific site of PTM occurrence. The fragmentation spectra of the phosphorylated peptides identified in this study are shown in Figure 7 with their respective sequences.

ETD fragmentation allows for more highly charged peptides and longer amino acid sequences to be detected compared to HCD fragmentation, which can be

advantageous for identifying phosphorylation sites. ETD has the additional benefit of relatively unbiased peptide backbone fragmentation, compared to HCD. These features have the potential to overcome some of the challenges to analyzing SR-rich domains [43].

In this study, phosphorylation of S176 was identified using HCD fragmentation from a sample digested with Lys-C and a low amount of trypsin, conditions which would be expected to result in of complete digestion at lysine residues but only partial digestion at arginine residues. Phosphorylation of S176 was also identified using ETD fragmentation along with phosphorylated S180 in a peptide digested with Lys-C and higher amount of trypsin (

Figure **8**). The two digestion conditions generated the same peptide sequence, however the peptide from HCD was found with a +2 charge and one phosphorylation site the peptide from ETD was found with a +3 charge and 2 phosphorylation sites. In Figure 7A the fragmentation spectrum was rich in y and b ions and accurately localized the phosphorylation site [44][45].

MaxQuant was used to analyze the raw ESI-MS/MS data files. Analysis of our recombinant SARS-CoV-2 N protein expressed in insect cells revealed 12 phosphorylation sites including S2, S21, T24, T135, located in/near the RNA binding region, S176, S180, S190, S193, S194, T198, located within the SR-rich region, T265, and T329 located at the CTD domain where oligomerization occurs (Figure 9, Figure 10). The 6 phosphorylation sites from the SR-rich domain as well as the first 3 phosphorylation sites from the N terminus were found in previous studies of cell lines infected with actual SARS-CoV-2 virus, in which phosphopeptides were enriched using

affinity techniques [46][47][48]. Collectively these studies found a total of 37 phosphorylation sites in SARS-CoV-2 N protein from the virus, 16 of them within the SR-rich domain. Since the SR-rich domain contains 16 serines and threonines total, all the possible phosphorylation sites were found to be in fact modified, according to those studies although they couldn't estimate occupancy. However, our MALDI data provides occupancy of about 13 phosphorylation sites on average. Our research detected 3 novel phosphorylation sites, T135, T265 and T329 located at the NTD and the CTD domains (Figure 7). These experiments' inability to identify the three unique phosphorylation sites might be attributed to differences in the technical approaches, or alternatively, the three novel sites could be phosphorylated in the recombinant SARS-CoV-2 from insect cells but not in the genuine virus.

Protein coverages, defined as the percentage of the amino acid residues identified by ESI-MS/MS analysis, were 68.3%, 75.8%, and 60.7% from procedure 1, 2 and 3 respectively. These numbers reflect a successful protein analysis with high coverage since total amino acid coverage is seldom accomplished with ESI-MS, unless different combinations of proteases are used. This is because the digested protein may generate too large or too small peptides that are not detectable within the mass range from the instrument which cause the highly hydrophilic peptides to pass through the reverse phase column and the highly hydrophobic peptides not to be eluted from the. In SARS-CoV-2 N protein, the SR-rich region also can influence a lower protein coverage because this domain has a long 31 amino acid sequence with mass of 3129.1 Da and it is highly phosphorylated with a dense negatively charge region. Protein digestion of SR-rich

region is challenging as in the absence of missed cleavages extremely small peptides from trypsin digestion occur which are not detectable or assignable in ESI-MS.

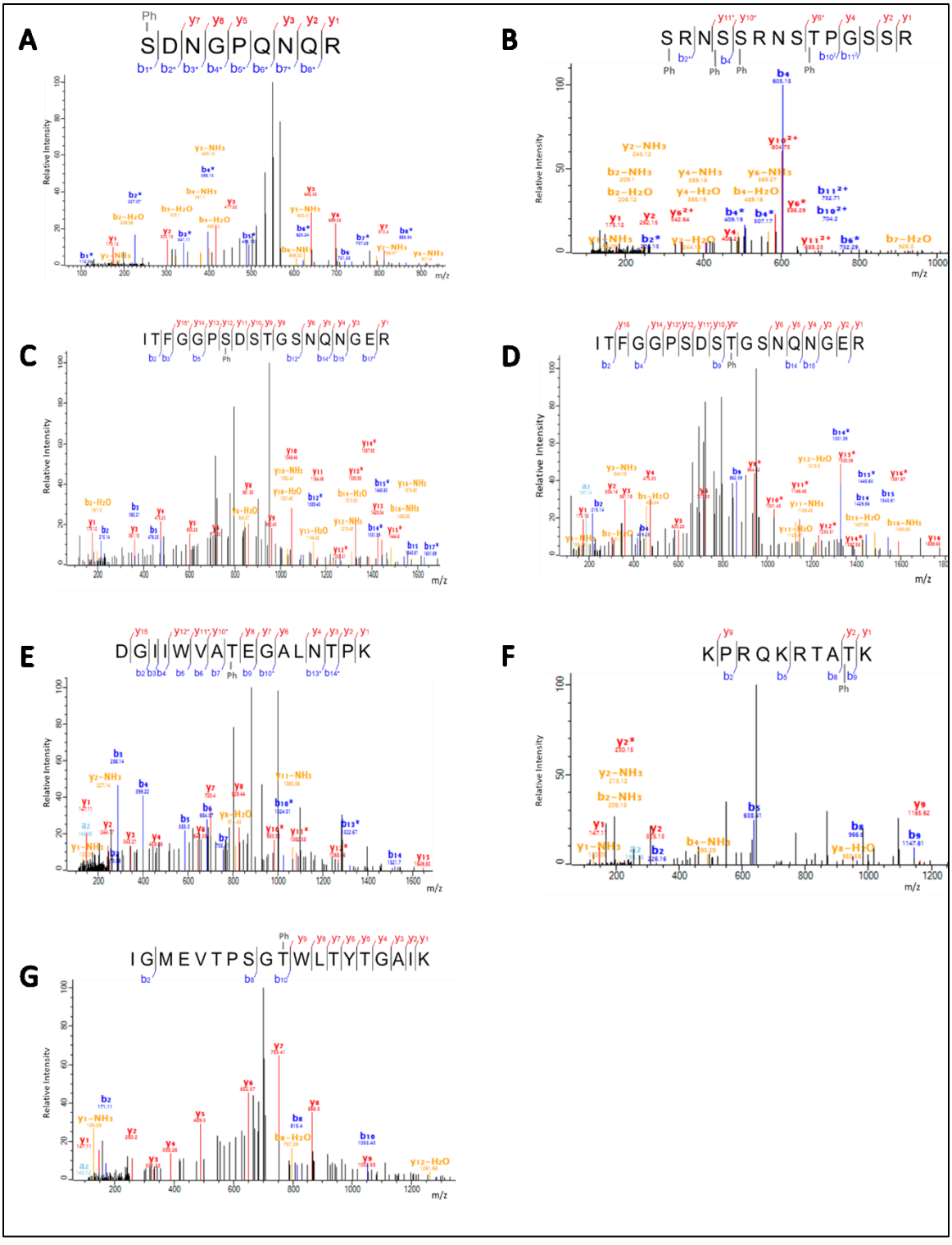


Figure 7: ESI-MS/MS spectra showing the phosphorylation amino acid along with their peptide sequence. *(A) S2 (B) S190, S193, S194, T198 (C) S21 (D) T24 (E) T135 (F) T265 (G) T329.*

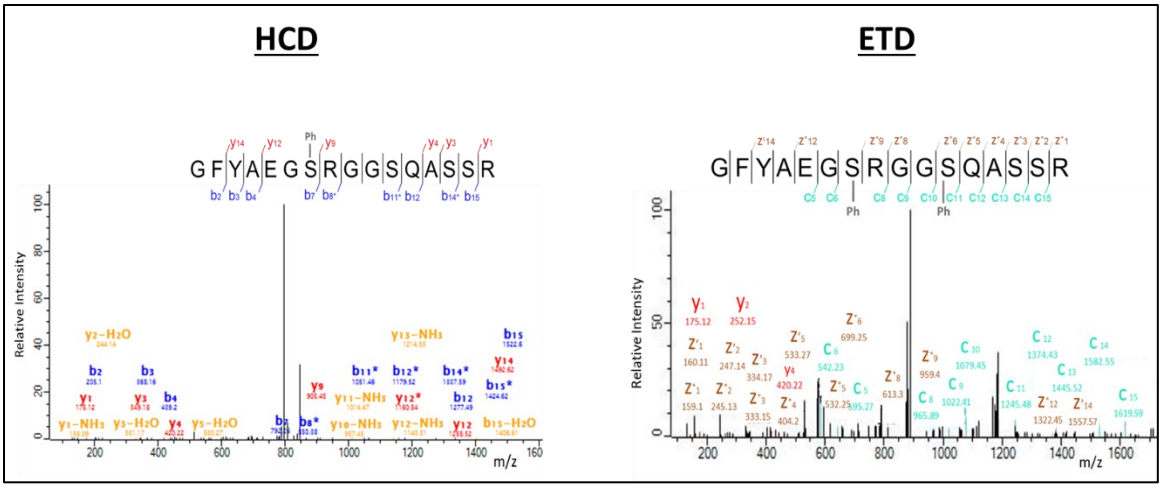


Figure 8: Comparison of HCD and ETD fragmentation of phosphopeptides from SARS-CoV-2 N protein SR-rich domain. *HCD fragmentation identified a phosphoserine at 176 while ETD fragmentation identified two phosphorylation sites (S176 and S180).*

Digestion procedure	Enzyme digestion amount	Incubation time	ESI-MS fragmentation	Phosphorylated sites	Peptide sequence	Domain	Protein % coverage
1	Trp 0.5 µg	3 hours	HCD	S2	**S**DNGPQNQR	RNA binding	68.30%
				S21	ITFGGP**S**DSTGSNQNGER		
				T24	ITFGGPSDS**T**GSNQNGER		
				T135	DGIIWVA**T**EGALNTPK	NTD/ RNA binding	
				S176	GFYAEG**S**R	SR-rich domain	
				T329	IGMEVTPSG**T**WLTYTGAIK	CTD/ Oligomerization	
2	LysC 0.5 µg	3 hours	HCD	T135	DGIIWVA**T**EGALNTPK	NTD/ RNA binding	75.80%
				T265	KPRQKRTA**T**K	CTD/ Oligomerization	
	LysC 0.5 µg + Trp 0.0005 µg			S190	**S**RNSSRNSTPGSSR	SR-rich domain	
				S193	SRN**S**SRNSTPGSSR		
				S194	SRNS**S**RNSTPGSSR		
				T198	SRNSSRNS**T**PGSSR		
	LysC 0.5 µg + Trp 0.0005 µg	5 hours		S176	GFYAEG**S**RGGSQASSR		
				S2	**S**DNGPQNQRNAPR	RNA binding	
3	LysC 0.5 µg + Trp 0.005 µg	3 hours	ETD	S2	**S**DNGPQNQRNAPR	RNA binding	60.70%
				S176	GFYAEG**S**RGGSQASSR	SR-rich domain	
				S180	GFYAEG**S**RGG**S**QASSR		

Figure 9: Phosphorylated sites and peptide sequences identified from each enzymatic digestion procedure.

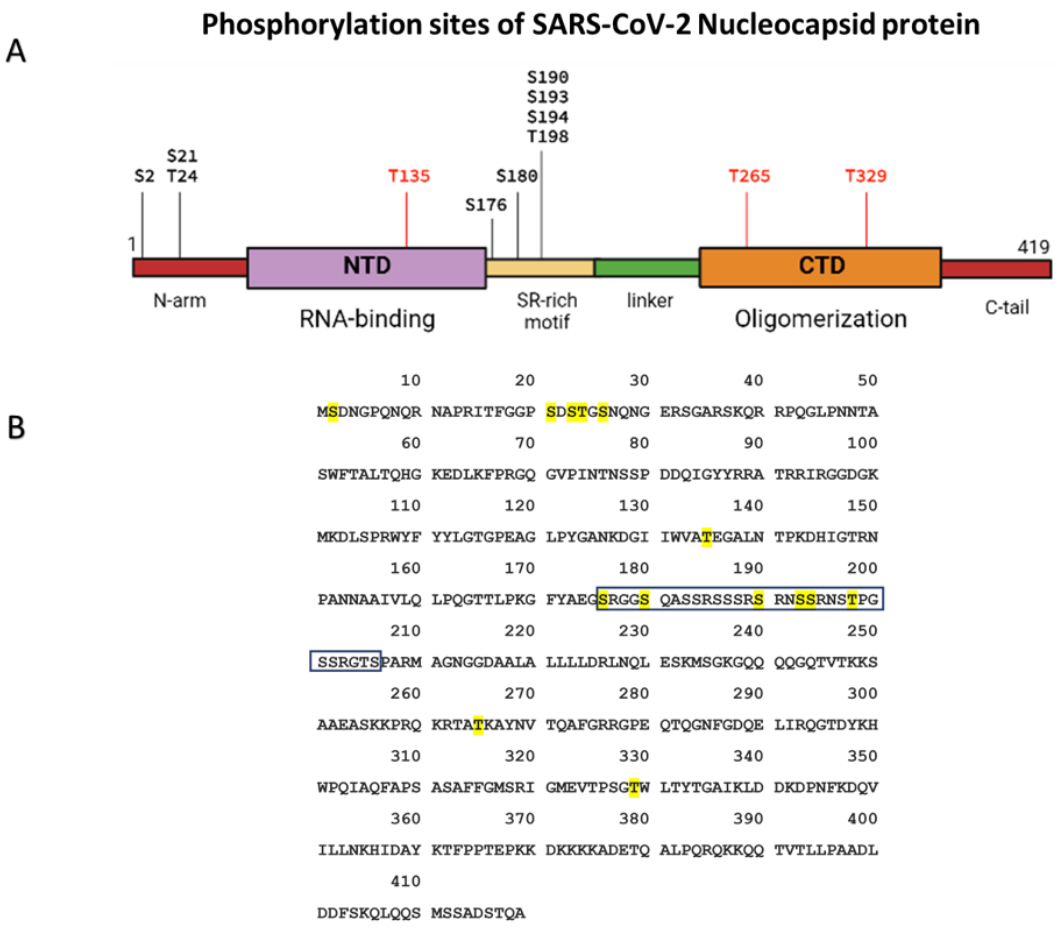


Figure 10: Phosphorylation sites found in SARS-CoV-2 N protein within the protein domains and amino acid sequence. *(A) Phosphorylation sites of SARS-CoV-2 Nucleocapsid protein. Novel sites in red. (B) Amino acid sequence of SARS-CoV-2 Nucleocapsid protein. Amino acids highlighted in yellow represent the phosphorylated residues and the SR-rich region is boxed in black. (Created with BioRender.com).*

3.4 Autoproteolysis occurrence on SARS-CoV-2 N protein

Throughout the progress of this experiment, a consistent decrease over time in the amount of purified recombinant SARS-CoV-2 N protein was observed via SDS-PAGE. The SDS-PAGE band representing the recombinant SARS-CoV-2 N protein became smaller and faded. For the GFP-tagged version of nucleocapsid the SDS-PAGE band representing the GFP-tag became thicker with a more intense color. Therefore, the

protein's concentration, estimated using web application, Image J, decreased over time. This observation led us to conclude that the recombinant SARS-CoV-2 N protein is easily degraded. For this reason, extra precautions were taken in handling the protein to avoid further degradation. Some of these precautions included storing the protein in cold temperatures, such as -80° C degrees, whenever possible, and minimizing unnecessary freezing/thawing cycles in an attempt to maintain the protein's integrity. However, further examination of SDS-PAGE results indicated that the extra care did not completely prevent protein degradation. One possibility is that the observed protein degradation may be autoproteolytic.

A recent publication has also indicated that SARS-CoV-2 N protein may be autoproteolytically degraded [49]. According to this publication, autoproteolysis led to the formation of five proteoforms with potential functions associated with the SARS-CoV-2 N protein domains they contain. In 2020, the authors of this study proposed that the proteoforms containing the CTD domain of SARS-CoV-2 N protein locate antigenic regions and the ability to form oligomers as a function of pH, aiding in antibody neutralization. Moreover, the proteoforms containing the NTD domain along with SR-rich domain binds to virus RNA promoting virus proliferation [49].

CHAPTER 4: CONCLUSION AND FUTURE WORK

4.1 Conclusion

The MALDI mass difference between the N protein samples that were treated and non-treated with CIP and lambda clearly represents phosphorylation occurrence on recombinant SARS-CoV-2 N protein. The recombinant SARS-CoV-2 sample treated with CIP revealed that the untreated sample was phosphorylated by 13 phosphoryl groups and the recombinant SARS-CoV-2 sample partially treated with lambda revealed that the intact sample was phosphorylated by at least 9 phosphoryl groups. Correlation of MALDI results along with SDS-PAGE gels stained with Coomassie blue revealed the phosphatase reaction with lambda protein was successful however it didn't remove all the phosphorylation from the recombinant SARS-CoV-2 N protein. This study encounters analytical challenges to find phosphorylation on peptides generated from Lys-C digestion and analyzed by MALDI-TOF. The difficulty detecting phosphopeptides by MALDI might be associated with the fact that the signal is suppressed by the low sensitivity of MALDI for phosphorylated peptides, and also the low abundance of these peptides [41].

Twelve phosphorylation sites were detected using ESI-MS/MS, of 3 are novel sites. The majority of those phosphorylated sites are located within the RNA binding region and SR-rich region. According to previous studies, SR-rich domain is detected to be fully phosphorylated on serine and threonine [46]. Phosphorylation on SR-rich domain changes the shape and charge of the domain compared to unphosphorylated SR-rich domain. We can hypothesize that phosphorylation of SR-rich domain modulates RNA binding. SARS-CoV-2 N protein is predicted to be slightly positively charged and have a relatively high isoelectric point (pI) ~ 10, due to higher number of basic amino acids over

the number of acidic amino acids. The addition of negatively charged phosphate group will cause a decrease in the pI making the N protein less positive. This will repel the interaction with the negatively charged RNA from SARS-CoV-2. Thus, we hypothesized that phosphorylation of SARS-CoV-2 N protein decreases the RNA binding. The disentangled interaction between the protein and the RNA allows RNA to be free for translation and duplication. By contrast, the unphosphorylated N protein will likely binds strongly to RNA which might serve as a protection to the genetic material.

4.2 Future Work

Further work can be done to find more phosphorylation sites in SARS-CoV-2 nucleocapsid protein, especially within the SR domain, which is predicted to be highly modified. Improvements in methods procedures can be done to maximize results. For example, increasing washings steps, trying different columns such as Ni-NTA during purification procedure can help optimize percent yield and increase protein % coverage analysis by mass spectrometry.

Peptide maps of proteins separated by 2 dimensional gel electrophoresis followed by an enzymatic and chemically (such as cyanogen bromide) degradation in-gel reaction is an alternative for detecting the presence of modifications. The peptides fragments can be extracted from the gel with compatible solution for accurate MALDI-MS measurements analysis without fractionation [32]. Another powerful approach to investigate posttranslational modification on protein is the combination of manipulated wet chemistry degradation in conjunction with MALDI-MS measurements of the amino acid using protein ladder sequencing technique [32].

Additionally, a simple and rapid technique for protein phosphorylation analysis can be done by measuring the protein isoelectric point using gel-based isoelectric focusing. Phosphorylated protein will cause a shift from predicted unphosphorylated protein as the addition of negatively charged phosphate group (PO_4^{-3}) will lower the isoelectric point and therefore effect protein migration on the 2-dimentional gel [50]. Highly phosphorylated protein will cause a higher shift on its isoelectric point.

Another significant technique to find phosphorylation on proteins in combination with mass spectrometry is phosphopeptides enrichment using strategies such as affinity chromatography, chemical modification or immunoprecipitants, which will increase the number of phosphopeptides isolated from samples. This technique can be advantageous to overcome the low abundance of phosphorylated proteins and peptides and signal suppression caused by the phosphorylation.

REFERENCES

[1] "Promed Post - ProMED-mail." https://promedmail.org/promed-post/?id=6864153 #COVID19.

[2] "Coronavirus Pandemic (COVID-19) - Our World in Data." https://ourworldindata.org/coronavirus#coronavirus-country-profiles (accessed Feb. 18, 2022).

[3] P. J. Rosenthal, "Editorial The Importance of Diagnostic Testing during a Viral Pandemic: Early Lessons from Novel Coronavirus Disease (COVID-19)," *Am. J. Trop. Med. Hyg*, vol. 102, no. 5, pp. 915–916, 2020, doi: 10.4269/ajtmh.20-0216.

[4] L. Caulley *et al.*, "Salivary Detection of COVID-19," *Ann. Intern. Med.*, Aug. 2020, doi: 10.7326/M20-4738.

[5] R. McBride, M. van Zyl, and B. C. Fielding, "The Coronavirus Nucleocapsid Is a Multifunctional Protein," *Viruses*, vol. 6, no. 8, p. 2991, Aug. 2014, doi: 10.3390/V6082991.

[6] W. Zeng *et al.*, "Biochemical characterization of SARS-CoV-2 nucleocapsid protein," *Biochem. Biophys. Res. Commun.*, vol. 527, no. 3, Jun. 2020, doi: 10.1016/j.bbrc.2020.04.136.

[7] B. Diao *et al.*, "Accuracy of a nucleocapsid protein antigen rapid test in the diagnosis of SARS-CoV-2 infection," *Clin. Microbiol. Infect.*, Oct. 2020, doi: 10.1016/j.cmi.2020.09.057.

[8] J. Dinnes *et al.*, “Rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection,” *Cochrane Database Syst. Rev.*, vol. 2020, no. 8, Aug. 2020, doi: 10.1002/14651858.CD013705.

[9] FDA, “Coronavirus Disease 2019 Testing Basics,” *fda.gov*, 2020. https://www.fda.gov/consumers/consumer-updates/coronavirus-disease-2019-testing-basics.

[10] M.-Y. Wang, R. Zhao, L.-J. Gao, X.-F. Gao, D.-P. Wang, and J.-M. Cao, “SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development,” doi: 10.3389/fcimb.2020.587269.

[11] A. Shah *et al.*, “Comparative mutational analysis of SARS-CoV-2 isolates from Pakistan and structural-functional implications using computational modelling and simulation approaches,” *Comput. Biol. Med.*, vol. 141, p. 105170, Feb. 2022, doi: 10.1016/J.COMPBIOMED.2021.105170.

[12] Y. Cong *et al.*, “Nucleocapsid Protein Recruitment to Replication-Transcription Complexes Plays a Crucial Role in Coronaviral Life Cycle,” *J. Virol.*, vol. 94, no. 4, Nov. 2019, doi: 10.1128/JVI.01925-19.

[13] C. A. Lutomski, T. J. El-Baba, J. R. Bolla, and C. V Robinson, “Proteoforms of the SARS-CoV-2 nucleocapsid protein are primed to proliferate the virus and attenuate the antibody response,” doi: 10.1101/2020.10.06.328112.

[14] S. Jiang, C. Hillyer, and L. Du, “Neutralizing Antibodies against SARS-CoV-2 and Other Human Coronaviruses,” *Trends Immunol.*, vol. 41, no. 5, p. 355, May

2020, doi: 10.1016/J.IT.2020.03.007.

[15] L. Premkumar *et al.*, “The receptor binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in SARS-CoV-2 patients,” *Sci. Immunol.*, vol. 5, no. 48, Jun. 2020, doi: 10.1126/SCIIMMUNOL.ABC8413.

[16] G. Wu and S. Yan, “Reasoning of spike glycoproteins being more vulnerable to mutations among 158 coronavirus proteins from different species,” *J. Mol. Model.*, vol. 11, no. 1, p. 8, Feb. 2005, doi: 10.1007/S00894-004-0210-0.

[17] T. Mohammad *et al.*, “Genomic Variations in the Structural Proteins of SARS-CoV-2 and Their Deleterious Impact on Pathogenesis: A Comparative Genomics Approach,” *Front. Cell. Infect. Microbiol.*, vol. 11, p. 951, Oct. 2021, doi: 10.3389/FCIMB.2021.765039/BIBTEX.

[18] F. D. Mast *et al.*, “Highly synergistic combinations of nanobodies that target SARS-CoV-2 and are resistant to escape,” *Elife*, vol. 10, Dec. 2021, doi: 10.7554/ELIFE.73027.

[19] K. R. Hurst, L. Kuo, C. A. Koetzner, R. Ye, B. Hsue, and P. S. Masters, “A Major Determinant for Membrane Protein Interaction Localizes to the Carboxy-Terminal Domain of the Mouse Coronavirus Nucleocapsid Protein,” *J. Virol.*, vol. 79, no. 21, pp. 13285–13297, Nov. 2005, doi: 10.1128/jvi.79.21.13285-13297.2005.

[20] C. K. Gu-Choul Shin 1, Yoon-Seok Chung1, In-Soo Kim, Hae-Wol Cho, “Antigenic characterization of severe acute respiratory syndrome-coronavirus

nucleocapsid protein expressed in insect cells: The effect of phosphorylation on immunoreactivity and specificit," *Virus Res.*, vol. 127, no. 2007, pp. 71–80, 2007, doi: https://doi.org/10.1016/j.virusres.2007.03.019.

[21] H. Chen *et al.*, "Mass Spectroscopic Characterization of the Coronavirus Infectious Bronchitis Virus Nucleoprotein and Elucidation of the Role of Phosphorylation in RNA Binding by Using Surface Plasmon Resonance," *J. Virol.*, vol. 79, no. 2, pp. 1164–1179, Jan. 2005, doi: 10.1128/jvi.79.2.1164-1179.2005.

[22] T. C. White, Z. Yi, and B. G. Hogue, "Identification of mouse hepatitis coronavirus A59 nucleocapsid protein phosphorylation sites," *Virus Res.*, vol. 126, no. 1–2, pp. 139–148, Jun. 2007, doi: 10.1016/j.virusres.2007.02.008.

[23] M. Bouhaddou *et al.*, "The Global Phosphorylation Landscape of SARS-CoV-2 Infection," *Cell*, vol. 182, no. 3, pp. 685-712.e19, Aug. 2020, doi: 10.1016/j.cell.2020.06.034.

[24] C. F. Invernizzi *et al.*, "Arginine methylation of the HIV-1 nucleocapsid protein results in its diminished function," *AIDS*, vol. 21, no. 7, pp. 795–805, Apr. 2007, doi: 10.1097/QAD.0b013e32803277ae.

[25] T. S. Fung and D. X. Liu, "Post-translational modifications of coronavirus proteins: roles and function," doi: 10.2217/fvl-2018-0008.

[26] S. G. Siddell, A. Barthel, and V. Ter Meulen, "Coronavirus JHM: a Virion-assoeiated Protein Kinase," 1981.

[27] E. Calvo *et al.*, “Phosphorylation and subcellular localization of transmissible gastroenteritis virus nucleocapsid protein in infected cells,” *J. Gen. Virol.*, vol. 86, no. 8, pp. 2255–2267, Aug. 2005, doi: 10.1099/vir.0.80975-0.

[28] M. Surjit, R. Kumar, R. N. Mishra, M. K. Reddy, V. T. K. Chow, and S. K. Lal, “The Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein Is Phosphorylated and Localizes in the Cytoplasm by 14-3-3-Mediated Translocation,” *J. Virol.*, vol. 79, no. 17, pp. 11476–11486, Sep. 2005, doi: 10.1128/jvi.79.17.11476-11486.2005.

[29] C.-H. Wu *et al.*, “Glycogen Synthase Kinase-3 Regulates the Phosphorylation of Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein and Viral Replication *,” *J. Biol. Chem.*, vol. 284, pp. 5229–5239, 2008, doi: 10.1074/jbc.M805747200.

[30] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, “NIH Image to ImageJ: 25 years of Image Analysis HHS Public Access,” *Nat Methods*, vol. 9, no. 7, pp. 671–675, 2012.

[31] J. Rappsilber, Y. Ishihama, and M. Mann, “Stop and Go Extraction Tips for Matrix-Assisted Laser Desorption/Ionization, Nanoelectrospray, and LC/MS Sample Pretreatment in Proteomics,” *Anal. Chem.*, vol. 75, no. 3, pp. 663–670, Feb. 2002, doi: 10.1021/AC026117I.

[32] R. C. Beavis and B. T. Chait, “Matrix-assisted laser desorption ionization mass-spectrometry of proteins,” *Methods Enzymol.*, vol. 270, no. 1978, pp. 519–551, 1996, doi: 10.1016/s0076-6879(96)70024-1.

[33] D480-D489, “UniProt: the universal protein knowledgebase in 2021 The UniProt Consortium,” *Nucleic Acids Res.*, vol. 49, 2021, doi: 10.1093/nar/gkaa1100.

[34] M. Goujon *et al.*, “A new bioinformatics analysis tools framework at EMBL-EBI,” doi: 10.1093/nar/gkq313.

[35] C. C. Lai, T. P. Shih, W. C. Ko, H. J. Tang, and P. R. Hsueh, “Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges,” *International Journal of Antimicrobial Agents*, vol. 55, no. 3. Elsevier B.V., p. 105924, Mar. 01, 2020, doi: 10.1016/j.ijantimicag.2020.105924.

[36] A. Grifoni, J. Sidney, Y. Zhang, R. H. Scheuermann, B. Peters, and A. Sette, “A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2,” *Cell Host Microbe*, vol. 27, no. 4, pp. 671-680.e2, Apr. 2020, doi: 10.1016/j.chom.2020.03.002.

[37] N. Blom, T. Sicheritz-Pontén, R. Gupta, S. Gammeltoft, and S. Brunak, “Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence,” *Proteomics*, vol. 4, no. 6, pp. 1633–1649, Jun. 2004, doi: 10.1002/pmic.200300771.

[38] N. Blom, S. Gammeltoft, and S. Brunak, “Sequence and structure-based prediction of eukaryotic protein phosphorylation sites,” *J. Mol. Biol.*, vol. 294, no. 5, pp. 1351–1362, Dec. 1999, doi: 10.1006/JMBI.1999.3310.

[39] B. A. Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., "ExPASy: the proteomics server for in-depth protein knowledge and analysis Nucleic Acids," *Res. 31:3784-3788(2003).* https://web.expasy.org/docs/expasy_ref.html.

[40] S. B. Breitkopf and J. M. Asara, "Determining in vivo Phosphorylation Sites using Mass Spectrometry," doi: 10.1002/0471142727.mb1819s98.

[41] C. E. Parker, V. Mocanu, M. Mocanu, N. Dicheva, and M. R. Warren, "Mass Spectrometry for Post-Translational Modifications," *Neuroproteomics*, pp. 93–113, Jan. 2010, doi: 10.1201/9781420076264.ch6.

[42] H. Steen, J. A. Jebanathirajah, J. Rush, N. Morrice, and M. W. Kirschner, "Phosphorylation Analysis by Mass Spectrometry MYTHS, FACTS, AND THE CONSEQUENCES FOR QUALITATIVE AND QUANTITATIVE MEASUREMENTS*," 2006, doi: 10.1074/mcp.M500135-MCP200.

[43] S. R. Kundinger, I. Bishof, E. B. Dammer, D. M. Duong, and N. T. Seyfried, "Middle-Down Proteomics Reveals Dense Sites of Methylation and Phosphorylation in Arginine-Rich RNA-Binding Proteins," *J. Proteome Res.*, vol. 19, no. 4, pp. 1574–1591, Apr. 2020, doi: 10.1021/ACS.JPROTEOME.9B00633/SUPPL_FILE/PR9B00633_SI_002.XLSX.

[44] D. L. Tabb, D. B. Friedman, and A.-J. L. Ham, "Verification of automated peptide identifications from proteomic tandem mass spectra," doi: 10.1038/nprot.2006.330.

[45] S. Banerjee and S. Mazumdar, “Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte,” *Int. J. Anal. Chem.*, vol. 2012, p. 40, 2012, doi: 10.1155/2012/282574.

[46] N. S. Tomer M. Yaron, Brook E. Heaton, Tyler M. Levy, Jared L. Johnson, Tristan X. Jordan, Benjamin M. Cohen, Alexander Kerelsky, Ting-Yu Lin, Katarina M. Liberatore, Danielle K. Bulaon, Edward R. Kastenhuber, Marisa N. Mercadante, Kripa Shobana-Ganesh, Long He and Heaton., “The FDA-approved drug Alectinib compromises SARS-CoV-2 nucleocapsid phosphorylation and inhibits viral infection in vitro,” pp. 1–46, 2020, doi: https://doi.org/10.1101/2020.08.14.251207.

[47] K. Klann, D. Bojkova, G. Tascher, S. Ciesek, C. Münch, and J. Cinatl, “Growth Factor Receptor Signaling Inhibition Prevents SARS-CoV-2 Replication,” *Mol. Cell*, vol. 80, no. 1, pp. 164-174.e4, Oct. 2020, doi: 10.1016/J.MOLCEL.2020.08.006/ATTACHMENT/9DEC5616-6956-443B-BC07-318C94865C1B/MMC6.XLSX.

[48] A. D. Davidson *et al.*, “Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein,” *Genome Med.*, vol. 12, no. 1, pp. 1–15, Jul. 2020, doi: 10.1186/S13073-020-00763-0/FIGURES/5.

[49] C. A. Lutomski, T. J. El-Baba, J. R. Bolla, and C. V Robinson, “Proteoforms of the SARS-CoV-2 nucleocapsid protein are primed to proliferate the virus and attenuate the antibody response,” *bioRxiv*, p. 2020.10.06.328112, 2020, [Online].

Available: https://doi.org/10.1101/2020.10.06.328112.

[50] J. C. Anderson and S. C. Peck, “A simple and rapid technique for detecting protein phosphorylation using one-dimensional isoelectric focusing gels and immunoblot analysis,” *Plant J.*, vol. 55, no. 5, pp. 881–885, 2008, doi: 10.1111/J.1365-313X.2008.03550.X.

Vita

Name	*Josie Daldegan Rezende*
Baccalaureate Degree	*Bachelor of Science*, *Bloomfield College, Bloomfield* *Major: Chemistry*
Date Graduated	*May, 2015*