Impacts of 203/204: RG>KR mutation in the N protein of SARS-CoV-2

Majid Vahed,[a] Tess M Calcagno,[b] Elena Quinonez,,[b], and Mehdi Mirsaeidi[d]*

[a]Division of Pulmonary and Critical Care, University of Miami, FL, USA

[b]Department of Medicine, University of Miami, Miami, FL, USA

Running Head: N protein of SARS-CoV-2 (RG>KR)

* Corresponding Author to whom requests for reprints should be addressed

Mehdi Mirsaeidi MD, 1600 NW 10th Ave # 7072B, Miami, Florida, USA 33136;

msm249@med.miami.edu

Conflicts of Interest: Authors reports not competing interest

Author contributions: Conceptualization (MM), Methodology (MV, EQ, MM), Writing (MV, TC, MM), Preparation (MV, EQ, TC), Reviewing & Editing (MV, TC, MM)

Word Count Abstract: 148

Word Count Text: 3,778

Abstract

We present a structure-based model of phosphorylation-dependent binding and sequestration of SARS-CoV-2 nucleocapsid protein and the impact of two consecutive amino acid changes R203K and G204R. Additionally, we studied how mutant strains affect HLA-specific antigen presentation and correlated these findings with HLA allelic population frequencies. We discovered RG>KR mutated SARS-CoV-2 expands the ability for differential expression of the N protein epitope on Major Histocompatibility Complexes (MHC) of varying Human Leukocyte Antigen (HLA) origin. The N protein LKR region K203, R204 of wild type (SARS-CoVs) and (SARS-CoV-2) observed HLA-A*30:01 and HLA-A*30:21, but mutant SARS-CoV-2 observed HLA-A*31:01 and HLA-A*68:01. Expression of HLA-A genotypes associated with the mutant strain occurred more frequently in all populations studied.

Importance

The novel coronavirus known as SARS-CoV-2 causes a disease renowned as 2019-nCoV (or COVID-19). HLA allele frequencies worldwide could positively correlate with the severity of coronavirus cases and a high number of deaths.

1. Introduction

A major outbreak of a novel coronavirus isolated in china called SARS-CoV2 leads to the global declaration of a worldwide pandemic (1). The novel coronavirus known as SARS-CoV2 causes a disease renowned as 2019-nCoV (or COVID-19) (1-3). Sequencing of the viral genome and characterization of immunogenic viral proteins has been crucial for understanding the virus and creating targeted vaccinations and treatments. Additionally, wide variation in the clinical presentation of COVID-19 ranging from an asymptomatic presentation to disabling multi-organ failure has led to the study of differential host-pathogen interaction specifically relating to population-based Human Leukocyte antigen (HLA) allele frequencies (4, 5).

SARS-CoV2 is an RNA virus and structurally contains 16 open reading frames (ORFs). ORF1a express 11 nonstructural proteins (Nsps) from Nsp1 to Nsp11, with the genes of ORF1b expressing proteins from Nsp12 to Nsp16 (6, 7). Major structural proteins including Spike (S), envelope (E), membrane (M) and nucleocapsid (N) proteins are encoded by other ORFs. The N proteins of SARS-CoV are highly basic structural proteins localized in the cytoplasm and the nucleolus of Trichoplusia ni BT1 Tn 5B1-4 cells (8). Previous studies have indicated the N proteins of other coronaviruses are extensively phosphorylated and bound to viral RNA to form a helical ribonucleoprotein (RNP) that comprises the viral core structure (9). Recently, mutations in N segment of SARS-CoV2 have been reported (10). Two replacements in positions R203K and G204R of N proteins have been found in several countries, but their potential effects in the protein structure have not been discussed.

We present a structure-based model of phosphorylation-dependent binding and sequestration of SARS-CoV-2 nucleocapsid protein and the impact of two consecutive amino acid changes R203K and G204R. Additionally, we studied how mutant strains affect HLA-specific antigen presentation and correlated these findings with HLA allelic population frequencies.

2. Results

2.1. Sequence Alignments and Clustering of LKR of CoVs N-protein

ClustalW multiple sequence alignment was employed to align the LKR (68 nucleotides long) of CoV N protein aligned for bat/pangolin models and human models SARS-COVs/MERS-COV. Notably, The Gly at position 204 is conserved among the closely related coronaviruses Bat coronavirus pangolin, SARS-CoV, and SARS-CoV-2, but variable for SARS-CoV2n (fig. 1(a)). The clustering trees of coronavirus are displayed in figure 1 (b). Sequence alignments suggested that other coronavirus N proteins might share the same structural organization based on intrinsic disorder predictor profiles and secondary structure predictions (Fig. 2).

2.2. Calculation of N-protein LKR Residue Energies

A plot of N-proteins residues indicates the local model quality by plotting knowledge-based energies as a function of an amino acid sequence position. A detailed energy calculation of wild and mutant types revealed residues 203/204 are located in the highest energy level area (Fig. 3). The mutant type showed slightly high free energy at residues 203/204 a.a. KR compared to the wild type; suggesting enhanced structural flexibility and increased the tendency for the formation of a coil or a bend in the secondary structure. The relative orientation of NTD and CTD, as well as the conformations of the disordered regions (N-arm, LKR, and C-tail), are drawn randomly to reflect the dynamic nature of the N protein (Fig. 3)

2.3. Calculation of Putative Phosphorylation Sites

We also analyzed predictable phosphorylation sites of N-protein by employing the NetPhos 3.1 server (http://www.cbs.dtu.dk/services/NetPhos/, accessed on 14 Sep 2020). The linker region of SARS-CoV N-protein (LKR) contains a Ser/Arg-rich region with a high number of putative phosphorylation sites (a.a. 172–206). The sites of contiguous amino acid changes of 203R>K and 204G>R are located in the SR-rich region which is known to be intrinsically disordered. We predicted a nonspecific kinase phosphorylation site at Ser202 and a specific CDK5, RSK, and GSK3 phosphorylation sites at Ser206, all of which are close to the RG>KR mutation (Fig. 4). When Ser202/206 and Thr205 are phosphorylated, charge neutralization of the nearby positively charged sidechains likely takes place. The G204R mutation decreases the conformational entropy of neutralization by increasing the positive charge in the vicinity of negatively charged P-Ser202/205 phosphate groups (Fig. 5).

2.4 Protein Localization

Subcellular localization of the N-protein was predicted using the DeepLoc 1.0 neural network algorithm. The resultant values of 0.861 and 0.913 obtained for wild and mutant N protein respectively suggest the protein is predominantly found in the nucleus (Fig. 6 a,b). The N-protein position prediction graph (Fig.6,c,d) confirmed a long peak in mutation areas K203 and R204, defined based on SARS-CoV-2 data.

2.5 Identification of N-protein B-cell epitopes

We used the Immune Epitope Database (IEDB) to determine linear B-cell hosted epitopes utilizing the incorporated Chou & Fasman Beta-Turn prediction module threshold 1.07. We supplied the FASTA sequence of the targeted protein as an input assuming all default parameters. LKR (a.a.170-206) region of N-protein shown significant antigenic epitopes with potential binding to B lymphocytes cells (Fig. 7). IEDB software also predicted epitopes based on N-protein conformation and residue exposure, and independently graphed a broad peak in mutation areas K203 and R204, the likely epitope regions defined based on SARS-CoV-2 data (Fig. S2).

2.6 Identification of T-cell epitopes

We found that MHC polymorphism typically results in differential MHC epitope recognition within the N-protein LKR K203 R204 region of wild type (SARS-CoV-2) as compared to mutant (SARS-CoV2). Wild type (SARS-CoV-2) observed HLA-A*30:01 and HLA 31:01 predictions, but mutant SARS-CoV-2 observed HLA-A*31:01 and HLA-A*68:01 predictions (Table 1,2). The frequency of HLA class I representation within South American, Japanese, and Iranian populations was recorded for both wild type and mutant strains. HLA class I molecules associated with mutant strains occurred more frequently in all three population groups, with the most significant increase seen in the South American population (18.38% in wild type versus 28.25% in the mutant).

3. Discussion

Variations in Host Human Leukocyte Antigen (HLA) gene expression can influence antigenic presentation of coronavirus epitopes. The Nucleocapsid (N) proteins of many coronaviruses are highly immunogenic epitopes expressed abundantly during infection. Using genome sequencing and prediction models, we characterized a mutation in the N-protein which affected the viral structure, function, and immunogenicity.

In the case of SARS-CoV N protein, previous studies have demonstrated that there are three intrinsically disordered regions (IDRs) (residues 1–44, 182–247, and 366–422) which modulate the RNA-binding activity of the N-terminal domain (NTD) and the C-terminal domain (CTD) (11).

The middle IDR, which we coined the linker region of SARS-CoV N protein (a.a. 182–247) LKR, and C-terminal IDR have both been implicated in the oligomerization of the N protein (12, 13). Five a.a variations at position 203 (R203K/M/S/I/G) within the SR-rich domain have been studied of the five variations, substitution R203K occurred most frequently in 68.09% of the mutated strains globally followed by the substitution G204R found in 67.94% mutated strains of the SARS-CoV-2 [19].

We identified phosphorylation sites in close proximity to the 203/204: RG>KR mutation and modeled the mutation's potential role in the alteration of protein structure and favorable electrostatic interaction between positively charged Arg and the negatively charged P-Ser202/205. A decrease in structural instability could contribute to a potential increase in immunogenicity of the mutant variant.

The N protein wild type is known to be a representative antigen for the T-cell response in a vaccine setting, inducing SARS-specific T-cell proliferation and cytotoxic activity (14, 15). We discovered RG>KR mutated SARS-CoV-2 expands the ability for differential expression of the N protein epitope on Major Histocompatibility Complexes (MHC) of varying Human Leukocyte Antigen (HLA) origin. Specifically, the N protein LKR region K203 R204 of wild type (SARS-CoVs) and (SARS-CoV-2) observed HLA-A*30:01 and HLA-A*30:21, but mutant SARS-CoV-2 observed HLA-A*31:01 and HLA-A*68:01. Expression of HLA-A genotypes associated with the mutant strain occurred more frequently in all populations studied (South American, Japanese, and Iranian) with the most significant increase in frequency seen in the South Americans.

Major MHC-I molecules play a key role in the recognition of intracellular pathogens. HLA-A*31:01, associated with mutant and wild type strains is reported to be linked with carbamazepine (CBZ)-induced severe cutaneous adverse reactions (SCAR), including medicine reaction with eosinophilia and systemic symptoms (DRESS), Stevens-Johnson syndrome (SJS), and toxic epidermal necrolysis (TEN) (16). HL-A*68:01 was only associated with the mutant strain in our analysis. Two previous independent studies linked the HLA-A*68:01 allele, which is expressed at 5.2–25% allele frequency, with severe influenza disease during the 2009 influenza pandemic (17, 18).

The frequency of the HLA-A*68:01 allomorph is high among the indigenous populations globally including Southern America (http://www.allelefrequencies.net) and Australia (19, 20). HLA-A*68:01 is at low levels in most of SE Asia, particularly the indigenous populations reflected by low levels along the West Pacific Rim including Japan (19). If HLA-A*68:01 expression correlates with the manifestation of a more severe illness course, HLA-A*68:01 allele frequencies worldwide could positively correlate with the severity of coronavirus cases and a high number of deaths seen in south American countries like Brazil (21). On the other hand, low HLA-A*68:01 expression could correlate with the low number of COVID-19-attributable deaths seen in Japan as compared to other industrialized countries (22).

4. Conclusion

Findings in this study demonstrate how variations in Host Human Leukocyte Antigen (HLA) gene expression can influence the antigenic presentation of coronavirus epitopes. We identified phosphorylation sites in close proximity to the 203/204: RG>KR mutation and modeled the mutation's potential role in the alteration of protein structure and favorable electrostatic interaction between positively charged Arg and the negatively charged P-

Ser202/205. Major MHC-I molecules play a key role in the recognition of intracellular pathogens. Low HLA-A*68:01 expression could correlate with the low number of COVID-19-attributable deaths seen in Japan as compared to other industrialized countries. Importantly, we found that RG>KR mutated SARS-CoV-2 expands the ability for differential expression of the N protein epitope on Major Histocompatibility Complexes (MHC) of varying Human Leukocyte Antigen (HLA) origin.

5. Methods

5.1. Data retrieval and sequence alignment

We used BLASTP programs from the NCBI database search (23) to find the LKR N-protein sequence (43 nucleotides long) of all SARS-CoV-2. Conserved and varied residues were identified by using the WebLogo program (24-26). Multiple alignments were performed between full-length N-protein sequences on the EMBL-EBI server. Clustal Omega is used to apply mBed algorithms for guide trees. ClustalW alignment tools executed to output alignment format (27). We analyzed all available sequences available up to September 07th, 2020.

5.2. Structure modeling

The atomic coordinates of the N-terminal domain (NTD) and C-terminal domain (CTD) were obtained from the structure that is available in a Protein Data Bank (http://www.rcsb.org/pdb) (PDB ID: 6M3M, 6WZO)(28, 29). The tertiary structure of the full 419 a.a. sequencing of N-proteins was predicted using the IntFOLD5 server (PYMOL). Sequences from residues 1-419 for N-protein native Sequence ID: YP_009724397.2 and mutant sequence ID: QIQ08827.1 were used in this study. All of the structures were visualized using

PYMOL Chimera software Version 1.7.4 (30). The calculation procedure was almost the same as that in our previous works (31-34).

5.3. Model Validation

ProSA was used to measure the energy distribution of the N-protein structure (35, 36).

5.4. Identification of B-cell epitopes

In this subsection, we used the Immune Epitope Database (IEDB) (36) to determine linear B-cell epitopes operating the incorporated Chou & Fasman Beta-Turn prediction module (37). We supplied the FASTA sequence of the targeted protein as an input considering all default parameters. We also used the Discotope 2.0 (38) method to predict epitopes based on N-protein conformation and residue exposure.

5.5. Identification of T-cell epitopes

We used the TepiTool, a T cell Epitope Tool that is used for MHC class I and II binding predictions. The IEDB team's recommendations were selected as defaults to automatically select the top peptides (39). In the MHC-I Binding Prediction feature, the default value provided is 1.0, i.e. all peptides with percentile rank ≤ 1.0 will be selected as predicted peptides. The list of representative alleles from different HLA supertypes was selected by the panel of 27 allele reference sets. The peptide selection criterion in this approach will always be predicted percentile rank. MHC-II binding prediction Results from the default value provided is 10.0, i.e. all peptides with percentile rank ≤ 10.0 will be selected as predicted peptides by using the panel of 26 most frequent alleles.

5.6. Prediction of protein subcellular localization using deep learning

The subcellular localization of wild and mutant N proteins was predicted using DeepLoc-1.0 software (40). Using the 419 sequence number at residues 203/204, a.a. of wild/ mutant strains were used to provide information about the subcellular localization of eukaryotic proteins using Neural Networks algorithm trained.

References

1. Jin Y, Yang H, Ji W, Wu W, Chen S, Zhang W, Duan G. 2020. Virology, Epidemiology, Pathogenesis, and Control of COVID-19. Viruses 12.
2. Hui DS, I Azhar E, Madani TA, Ntoumi F, Kock R, Dar O, Ippolito G, McHugh TD, Memish ZA, Drosten C, Zumla A, Petersen E. 2020. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China. International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases 91:264-266.
3. Carlos WG, Dela Cruz CS, Cao B, Pasnick S, Jamil S. 2020. Novel Wuhan (2019-nCoV) Coronavirus. Am J Respir Crit Care Med 201:P7-p8.
4. Wang W, Zhang W, Zhang J, He J, Zhu F. 2020. Distribution of HLA allele frequencies in 82 Chinese individuals with coronavirus disease-2019 (COVID-19). Hla 96:194-196.
5. Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, Nellore A, Thompson RF. 2020. Human Leukocyte Antigen Susceptibility Map for Severe Acute Respiratory Syndrome Coronavirus 2. J Virol 94.
6. Zumla A, Chan JF, Azhar EI, Hui DS, Yuen KY. 2016. Coronaviruses - drug discovery and therapeutic options. Nat Rev Drug Discov 15:327-47.
7. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, Liu W, Bi Y, Gao GF. 2016. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends Microbiol 24:490-502.
8. Ren AX, Xie YH, Kong YY, Yang GZ, Zhang YZ, Wang Y, Wu XF. 2004. Expression, purification and sublocalization of SARS-CoV nucleocapsid protein in insect cells. Acta Biochim Biophys Sin (Shanghai) 36:754-8.
9. Macneughton MR, Davies HA. 1978. Ribonucleoprotein-like structures from coronavirus particles. J Gen Virol 39:545-9.
10. Franco-Muñoz C, Álvarez-Díaz DA, Laiton-Donato K, Wiesner M, Escandón P, Usme-Ciro JA, Franco-Sierra ND, Flórez-Sánchez AC, Gómez-Rangel S, Rodríguez-Calderon LD, Barbosa-Ramirez J, Ospitia-Baez E, Walteros DM, Ospina-Martinez ML, Mercado-Reyes M. 2020. Substitutions in Spike and Nucleocapsid proteins of SARS-CoV-2 circulating in South America. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases 85:104557-104557.
11. Chang C-k, Hou M-H, Chang C-F, Hsiao C-D, Huang T-h. 2014. The SARS coronavirus nucleocapsid protein--forms and functions. Antiviral research 103:39-50.
12. Luo H, Chen J, Chen K, Shen X, Jiang H. 2006. Carboxyl terminus of severe acute respiratory syndrome coronavirus nucleocapsid protein: self-association analysis and nucleic acid binding characterization. Biochemistry 45:11827-35.
13. He R, Dobie F, Ballantine M, Leeson A, Li Y, Bastien N, Cutts T, Andonov A, Cao J, Booth TF, Plummer FA, Tyler S, Baker L, Li X. 2004. Analysis of multimerization of the SARS coronavirus nucleocapsid protein. Biochemical and biophysical research communications 316:476-483.
14. Okada M, Takemoto Y, Okuno Y, Hashimoto S, Yoshida S, Fukunaga Y, Tanaka T, Kita Y, Kuwayama S, Muraki Y, Kanamaru N, Takai H, Okada C, Sakaguchi Y, Furukawa I, Yamada K,

Matsumoto M, Kase T, Demello DE, Peiris JS, Chen PJ, Yamamoto N, Yoshinaka Y, Nomura T, Ishida I, Morikawa S, Tashiro M, Sakatani M. 2005. The development of vaccines against SARS corona virus in mice and SCID-PBL/hu mice. Vaccine 23:2269-72.

15. Gao W, Tamin A, Soloff A, D'Aiuto L, Nwanegbo E, Robbins PD, Bellini WJ, Barratt-Boyes S, Gambotto A. 2003. Effects of a SARS-associated coronavirus vaccine in monkeys. Lancet 362:1895-6.

16. Dean L. 2012. Carbamazepine Therapy and HLA Genotype. *In* Pratt VM, Scott SA, Pirmohamed M, Esquivel B, Kane MS, Kattman BL, Malheiro AJ (ed), Medical Genetics Summaries. National Center for Biotechnology Information (US), Bethesda (MD).

17. Hertz T, Oshansky CM, Roddam PL, DeVincenzo JP, Caniza MA, Jojic N, Mallal S, Phillips E, James I, Halloran ME, Thomas PG, Corey L. 2013. HLA targeting efficiency correlates with human T-cell response magnitude and with mortality from influenza A infection. Proc Natl Acad Sci U S A 110:13492-7.

18. Grant E, Wu C, Chan KF, Eckle S, Bharadwaj M, Zou QM, Kedzierska K, Chen W. 2013. Nucleoprotein of influenza A virus is a major target of immunodominant CD8+ T-cell responses. Immunol Cell Biol 91:184-94.

19. Middleton D, Menchaca L, Rood H, Komerofsky R. 2003. New allele frequency database: http://www.allelefrequencies.net. Tissue Antigens 61:403-7.

20. Clemens EB, Grant EJ, Wang Z, Gras S, Tipping P, Rossjohn J, Miller A, Tong SYC, Kedzierska K. 2016. Towards identification of immune and genetic correlates of severe influenza disease in Indigenous Australians. Immunology and cell biology 94:367-377.

21. Burki T. 2020. COVID-19 in Latin America. The Lancet Infectious diseases 20:547-548.

22. Iwasaki A, Grubaugh ND. 2020. Why does Japan have so few cases of COVID-19? EMBO molecular medicine 12:e12481-e12481.

23. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188-90.

24. Vahed M, Ishihara J-I, Takahashi H. 2019. DIpartite: A tool for detecting bipartite motifs by considering base interdependencies. PloS one 14:e0220207-e0220207.

25. Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. Nucleic acids research 18:6097-6100.

26. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47:W636-w641.

27. Kang S, Yang M, Hong Z, Zhang L, Huang Z, Chen X, He S, Zhou Z, Zhou Z, Chen Q, Yan Y, Zhang C, Shan H, Chen S. 2020. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. Acta Pharm Sin B 10:1228-1238.

28. Ye Q, West AMV, Silletti S, Corbett KD. 2020. Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. bioRxiv doi:10.1101/2020.05.17.100685:2020.05.17.100685.

29. McGuffin LJ, Adiyaman R, Maghrabi AHA, Shuid AN, Brackenridge DA, Nealon JO, Philomina LS. 2019. IntFOLD: an integrated web resource for high performance protein structure and function prediction. Nucleic Acids Res 47:W408-w413.

30. Vahed M, Vahed M, Sweeney A, Shirazi FH, Mirsaeidi M. 2020. Mutation in position of 32 (G&amp;gt;U) of S2M differentiate human SARS-CoV2 from Bat Coronavirus. bioRxiv doi:10.1101/2020.09.02.280529:2020.09.02.280529.

31. Wiederstein M, Sippl MJ. 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35:W407-10.

32. Vahed M, Sweeney A, Shirasawa H, Vahed M. 2019. The initial stage of structural transformation of Aβ(42) peptides from the human and mole rat in the presence of Fe(2+) and Fe(3+): Related to Alzheimer's disease. Comput Biol Chem 83:107128.

33. Vahed M, Neya S, Matsuzaki K, Hoshino T. 2018. Simulation Study on Complex Conformations of Aβ(42) Peptides on a GM1 Ganglioside-Containing Lipid Membrane. Chem Pharm Bull (Tokyo) 66:170-177.

34. Vahed M, Ahmadian G, Ameri N, Vahed M. 2019. G-rich VEGF aptamer as a potential inhibitor of chitin trafficking signal in emerging opportunistic yeast infection. Comput Biol Chem 80:168-176.

35. Sippl MJ. 1993. Recognition of errors in three-dimensional structures of proteins. Proteins 17:355-62.

36. Kim Y, Ponomarenko J, Zhu Z, Tamang D, Wang P, Greenbaum J, Lundegaard C, Sette A, Lund O, Bourne PE, Nielsen M, Peters B. 2012. Immune epitope database analysis resource. Nucleic Acids Research 40:W525-W530.

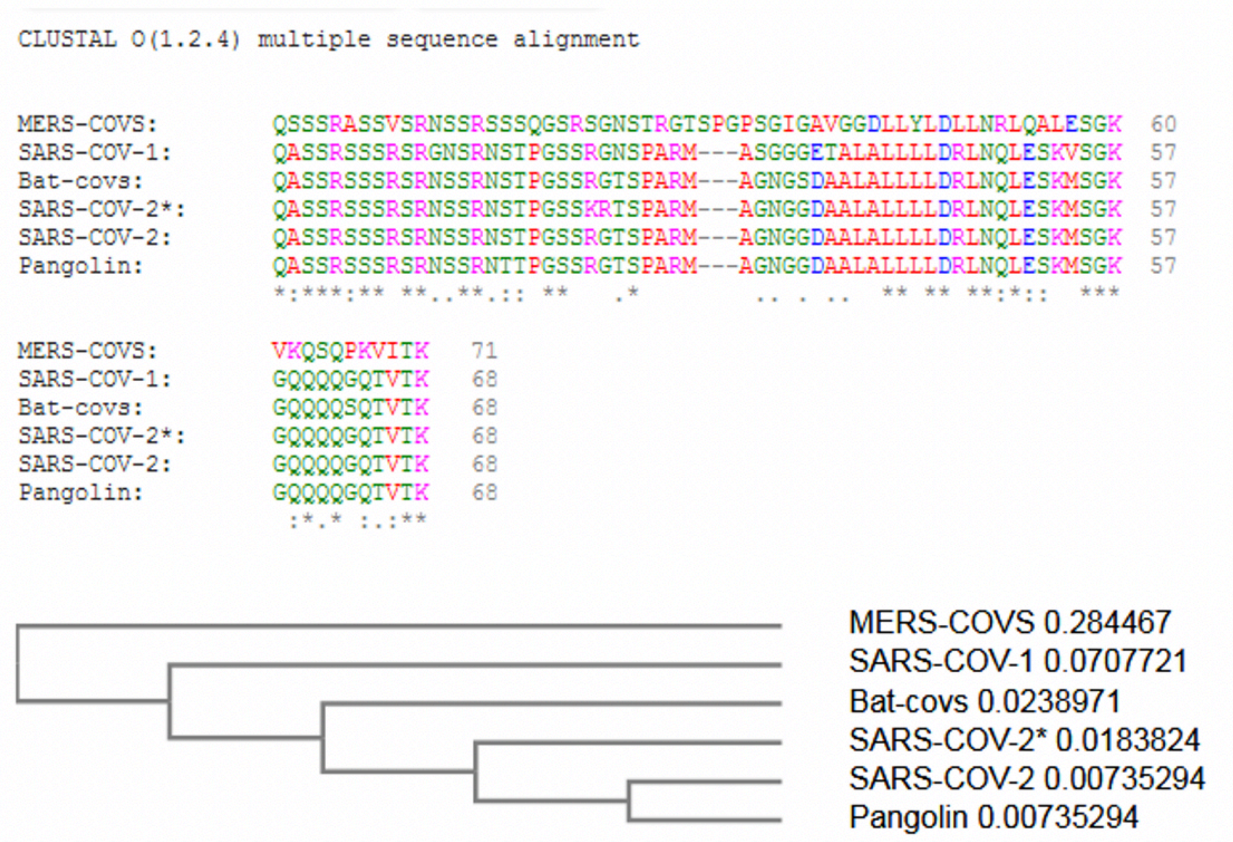
37. Chou PY, Fasman GD. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 47:45-148.

38. Kringelum JV, Lundegaard C, Lund O, Nielsen M. 2012. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. PLoS Comput Biol 8:e1002829.

39. Paul S, Sidney J, Sette A, Peters B. 2016. TepiTool: A Pipeline for Computational Prediction of T Cell Epitope Candidates. Curr Protoc Immunol 114:18.19.1-18.19.24.

40. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 33:3387-3395.

Figures and Tables



```
CLUSTAL O(1.2.4) multiple sequence alignment


MERS-COVS:      QSSSRASSVSRNSSRSSSQGSRSGNSTRGTSPGPSGIGAVGGDLLYLDLLNRLQALESGK  60
SARS-COV-1:     QASSRSSSRSRGNSRNSTPGSSRGNSPARM---ASGGGETALALLLLDRLNQLESKVSGK  57
Bat-covs:       QASSRSSSRSRNSSRNSTPGSSRGTSPARM---AGNGSDAALALLLLDRLNQLESKMSGK  57
SARS-COV-2*:    QASSRSSSRSRNSSRNSTPGSSKRTSPARM---AGNGGDAALALLLLDRLNQLESKMSGK  57
SARS-COV-2:     QASSRSSSRSRNSSRNSTPGSSRGTSPARM---AGNGGDAALALLLLDRLNQLESKMSGK  57
Pangolin:       QASSRSSSRSRNSSRNTTPGSSRGTSPARM---AGNGGDAALALLLLDRLNQLESKMSGK  57
                *:***:** **..**.:: **   .*        .. . ..  ** ** **:*::  ***

MERS-COVS:      VKQSQPKVITK   71
SARS-COV-1:     GQQQQGQTVTK   68
Bat-covs:       GQQQQSQTVTK   68
SARS-COV-2*:    GQQQQGQTVTK   68
SARS-COV-2:     GQQQQGQTVTK   68
Pangolin:       GQQQQGQTVTK   68
                 :*.* :.:**
```

Figure. 1. Coronavirus N-protein LKR sequence domains. (a) Alignment for bat coronavirus pangolin, SARS-CoV, SARS-CoV-2 and SARS-CoV2n each genotype for sequence representation. Columns with changes for nucleotide positions have been color-coded. (b) ClustalW multiple sequence alignment trees display of coronavirus. (*)The asterisk indicates the mutation type.

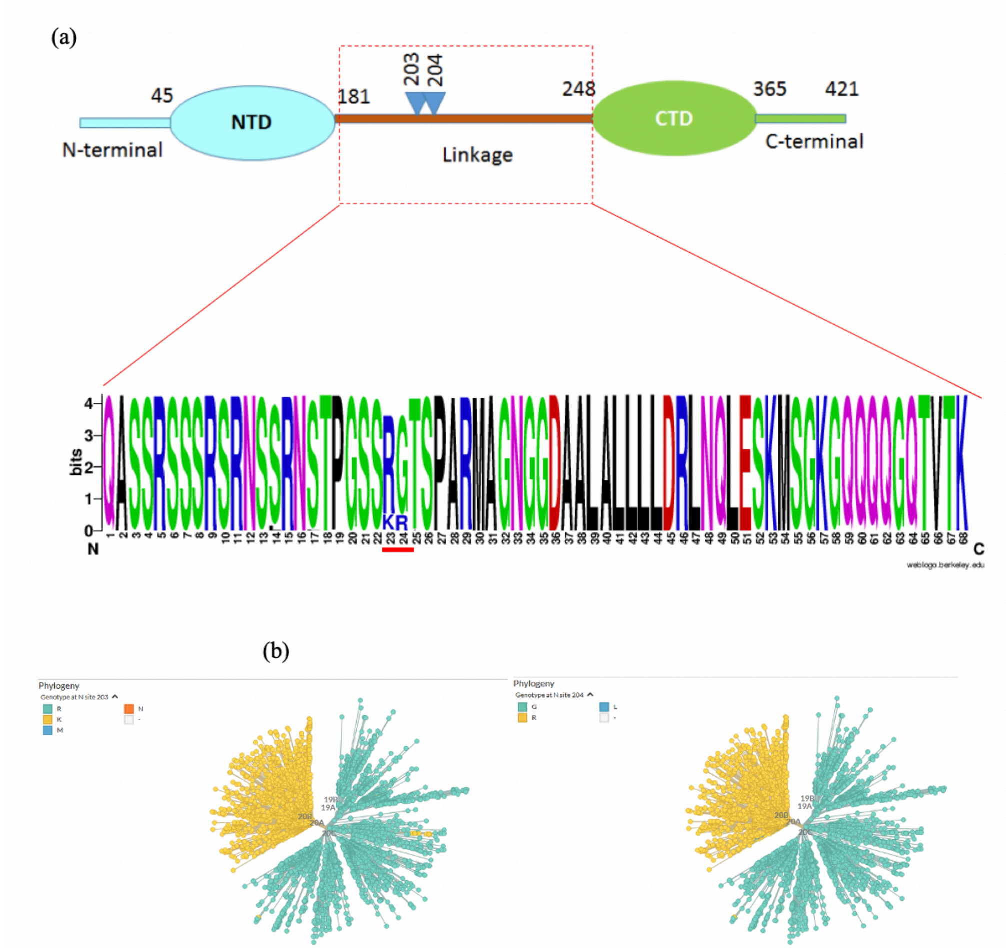


Figure. 2.. (a) Schematic representation of the N protein LKR domain, (b) Multiple sequence alignment logos of selected N protein LKR domain, a.a. letter heights indicate their frequency at the position 203/204 in SARS-COV-2. (c) The SARS-CoV-2 N protein conserved mutations at the position 203/204 in a phylogenetic graph obtained from Nextstraindatabase. Picture captured date: 9/25/2020.

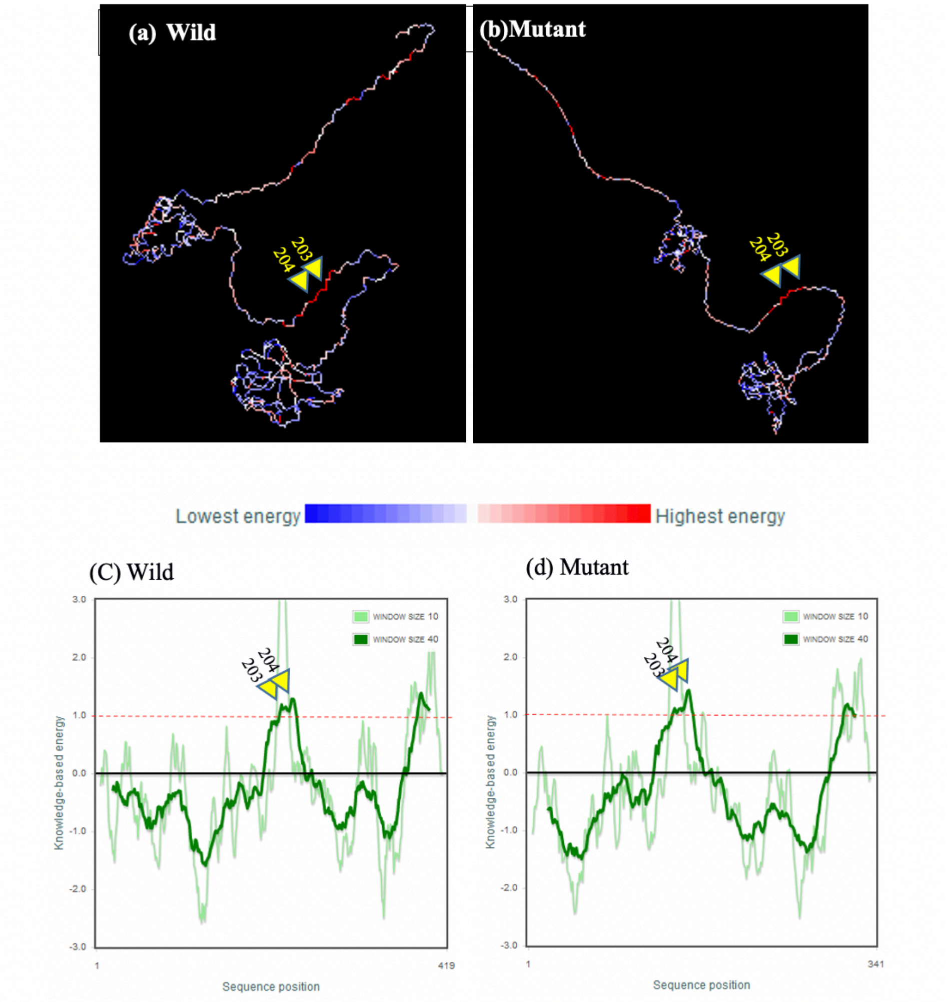
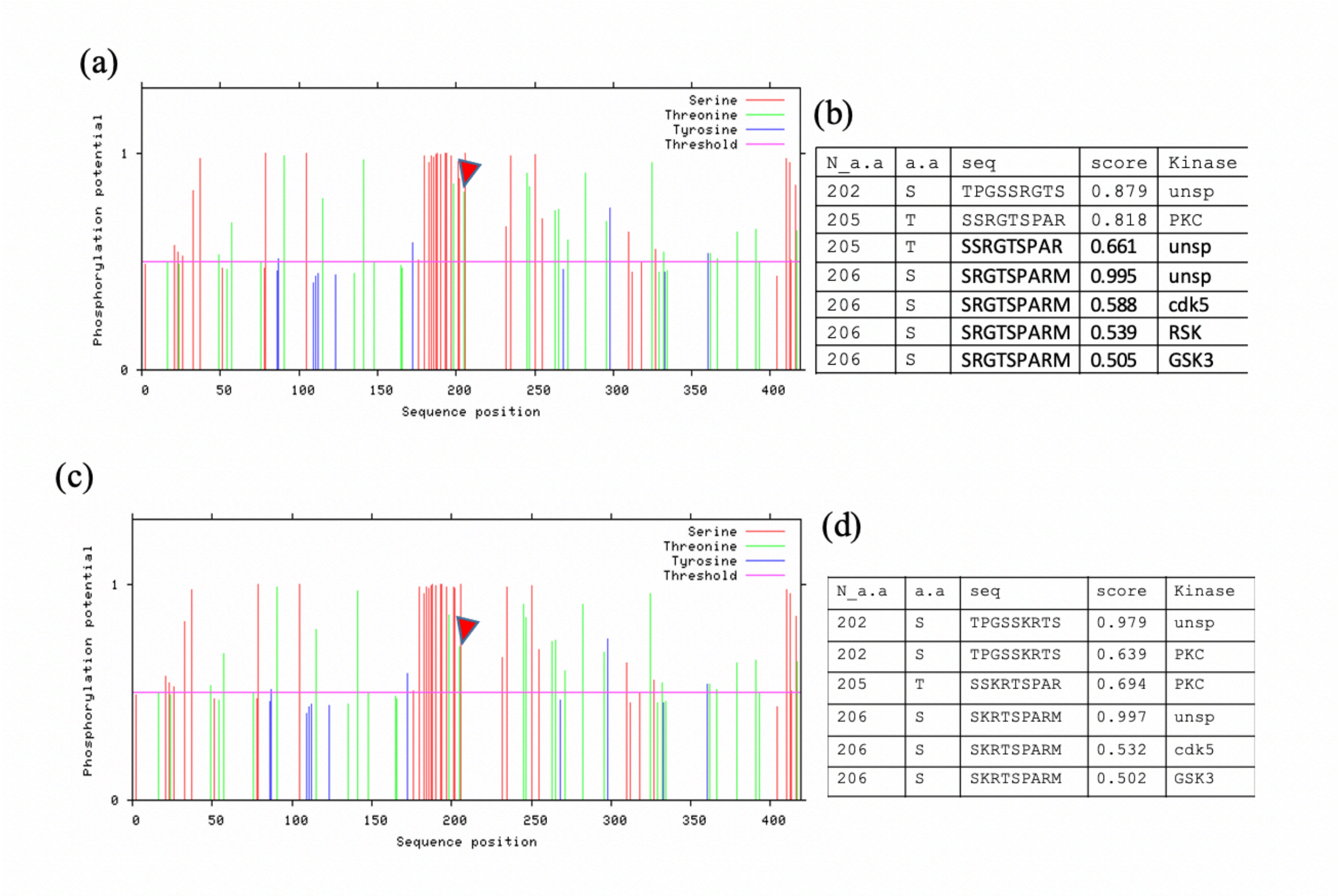


Figure. 3. Three structure dimensional depicted of the based on the energy values where "blue" represents lowest energy and "red" represents highest energy in the N-protein (a) SARS-CoV-2 wild (b) SARS-CoV-2 mutant. (c) ProSa energy plots of N-protein wild type a.a. 203/204 RG (d) mutant type a.a. 203/204 KR.



N_a.a	a.a	seq	score	Kinase
202	S	TPGSSRGTS	0.879	unsp
205	T	SSRGTSPAR	0.818	PKC
205	T	SSRGTSPAR	0.661	unsp
206	S	SRGTSPARM	0.995	unsp
206	S	SRGTSPARM	0.588	cdk5
206	S	SRGTSPARM	0.539	RSK
206	S	SRGTSPARM	0.505	GSK3

N_a.a	a.a	seq	score	Kinase
202	S	TPGSSKRTS	0.979	unsp
202	S	TPGSSKRTS	0.639	PKC
205	T	SSKRTSPAR	0.694	PKC
206	S	SKRTSPARM	0.997	unsp
206	S	SKRTSPARM	0.532	cdk5
206	S	SKRTSPARM	0.502	GSK3

382

383 **Figure. 4.** ProSa Phosphorylation plots of N-protein. (a) Phosphorylation sites predicted in the
384 local LKR SARS-CoV-2 wild (b) phosphorylation score and substrates of kinases SARS-CoV-
385 2 wild (c) Phosphorylation sites predicted in the local LKR SARS-CoV-2 mutant.(d)
386 phosphorylation score and substrates of kinases SARS-CoV-2 mutant.

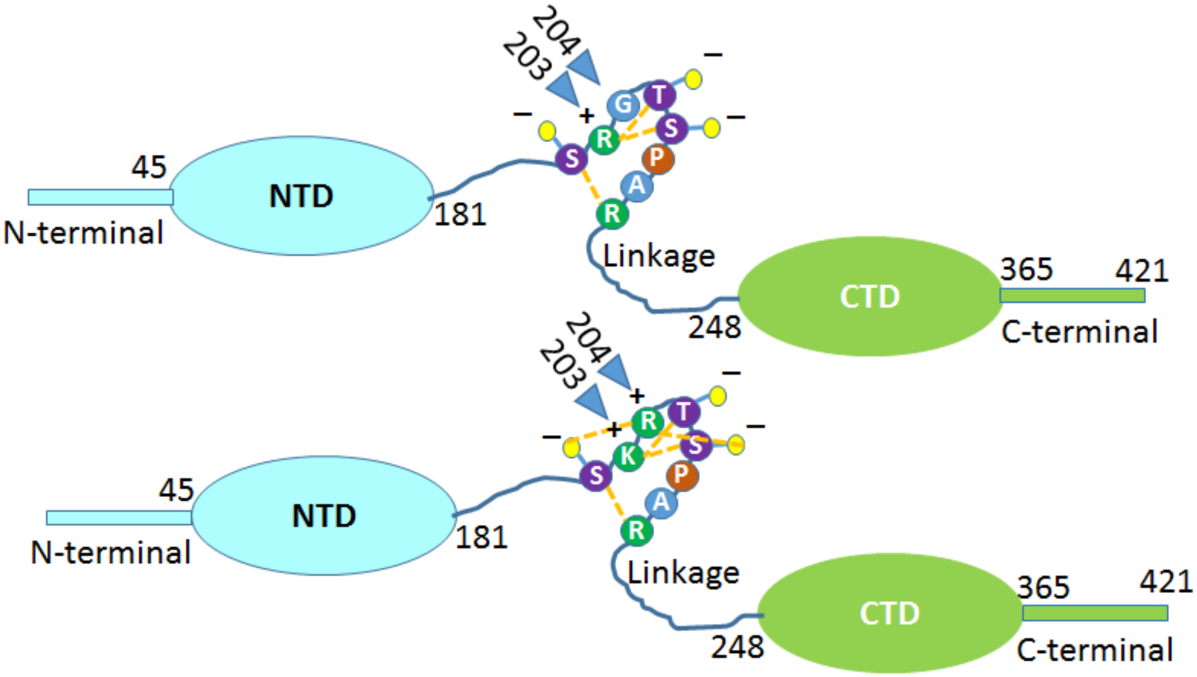


Figure. 5 Schematic representation of the mutation and possibility electrostatic interactions between positively charged Arg/Lys and the negatively charged P-Ser/Thr as shown yellow dotted lines. (a) the wild type SARS-CoV-2 N-protein with Arg and Gly at position 203 and 204, respectively (b) the variant SARS-CoV-2 N-protein with mutations at position 203 and 204 (203R>K, 204G>R). These residues are colour-coded based on their Charges (Green: positively charged, Purple: negatively charged P-Ser/Thr. The phosphate groups on Ser/Thr residues at position 202, 205 and 206 are denoted as red circles on yellow sticks

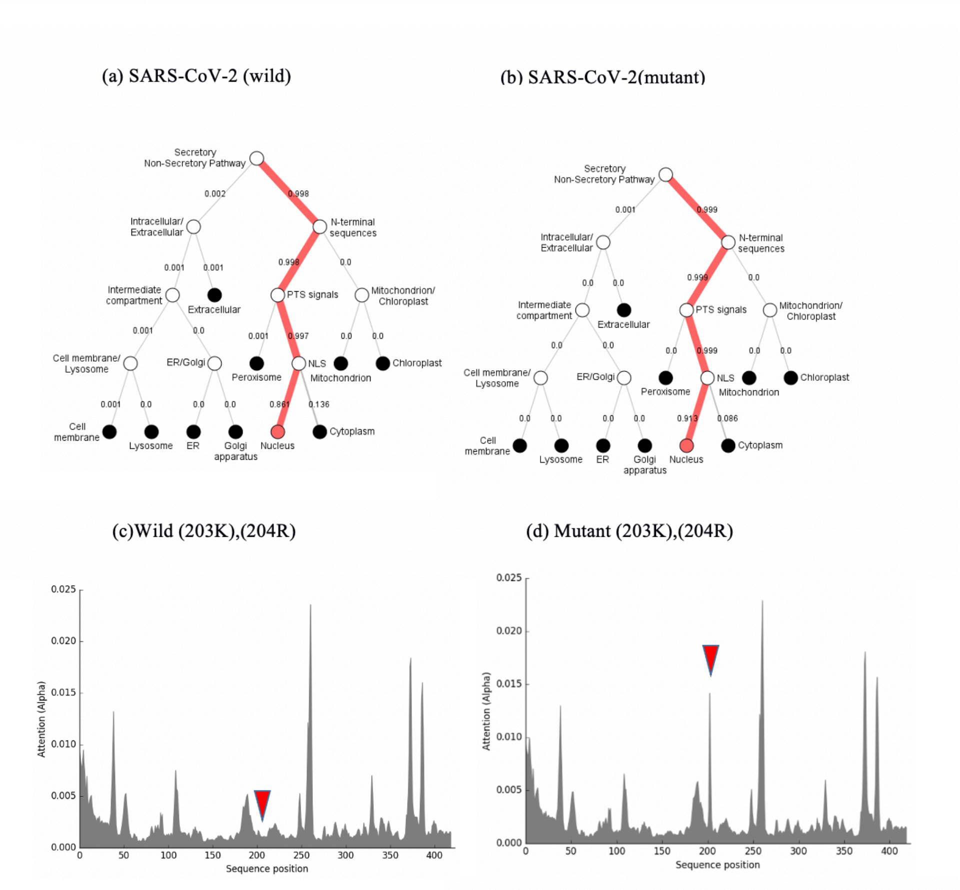
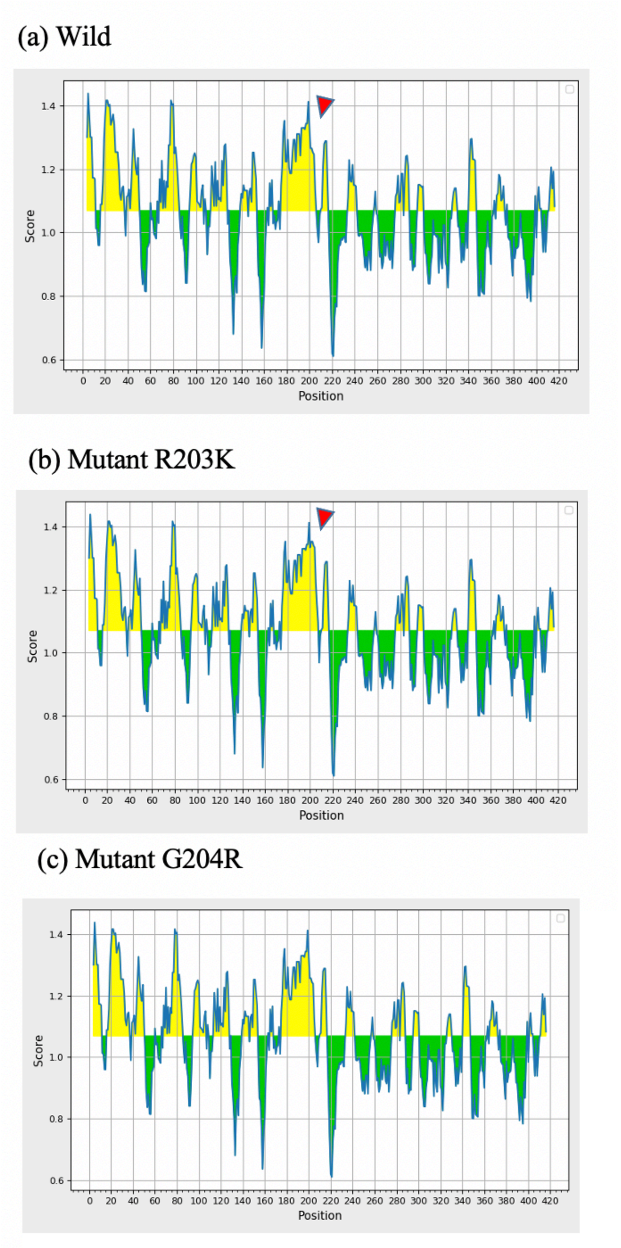


Figure 6. Hierarchical tree-predicted subcellular localizations of N-protein using neural networks algorithm. (a) Wild (203R),(204G). (b) Mutant (203K),(204R). (c) N-protein residue position prediction Wild (203K),(204R). (d) N-protein residue position prediction Mutant (203K), (204R)



(a) Wild

(b) Mutant R203K

(c) Mutant G204R

Figure. 6. Graphical representation of prediction linear B-cell epitopes within the N-protein of (a) SARS-COV-2 wild (b) SARS-COV-2 R203K mutant (c) G203R by IEDB scales. Arrow depicts the residue 204.

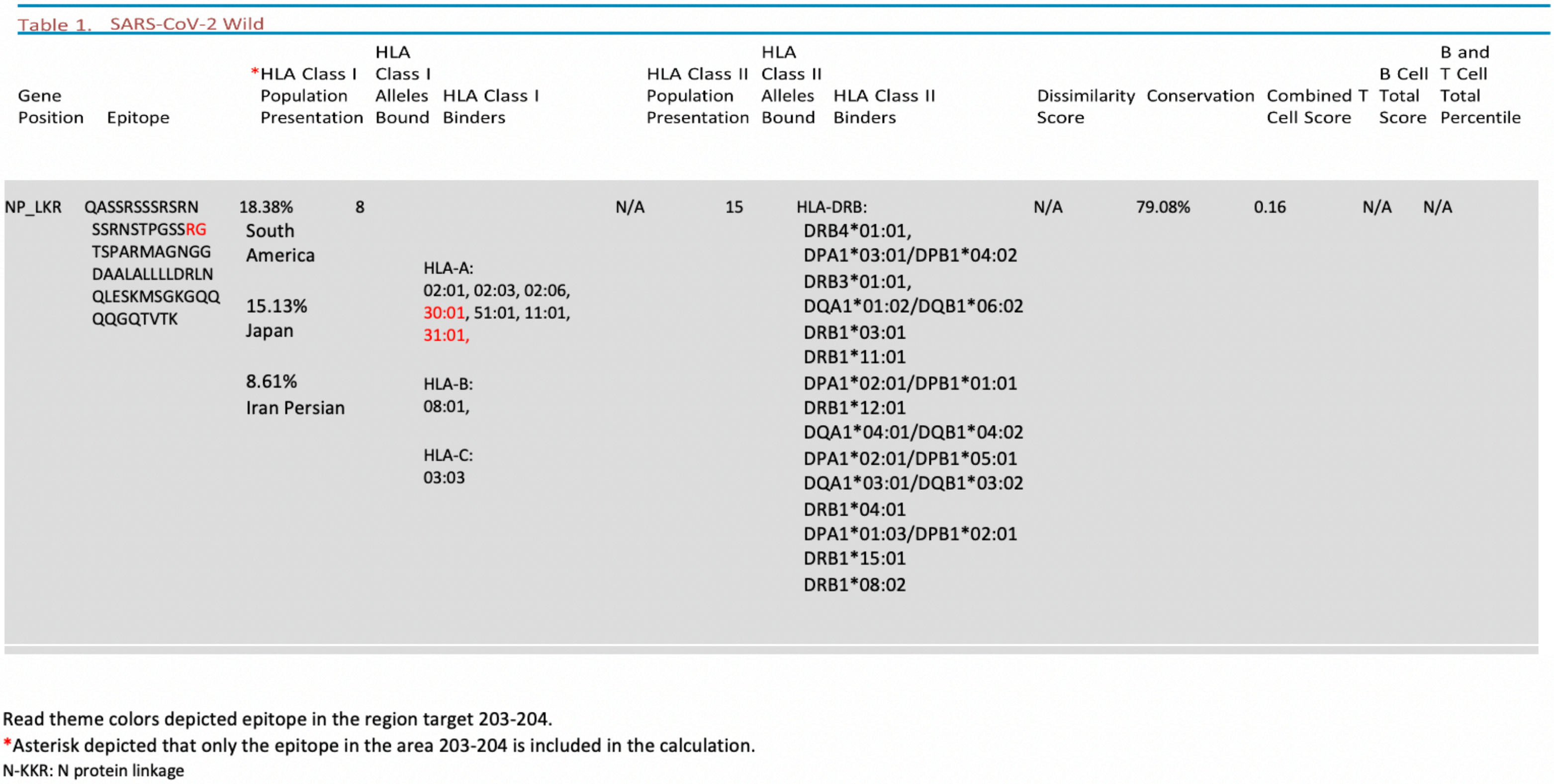


Table 1. SARS-CoV-2 Wild

Gene Position	Epitope	*HLA Class I Population Presentation	HLA Class I Alleles Bound	HLA Class I Binders	HLA Class II Population Presentation	HLA Class II Alleles Bound	HLA Class II Binders	Dissimilarity Score	Conservation	Combined T Cell Score	B Cell Total Score	B and T Cell Total Percentile
NP_LKR	QASSRSSSRSRN SSRNSTPGSSRG TSPARMAGNGG DAALALLLLDRLN QLESKMSGKGQQ QQGQTVTK	18.38% South America 15.13% Japan 8.61% Iran Persian	8	HLA-A: 02:01, 02:03, 02:06, 30:01, 51:01, 11:01, 31:01, HLA-B: 08:01, HLA-C: 03:03	N/A	15	HLA-DRB: DRB4*01:01, DPA1*03:01/DPB1*04:02 DRB3*01:01, DQA1*01:02/DQB1*06:02 DRB1*03:01 DRB1*11:01 DPA1*02:01/DPB1*01:01 DRB1*12:01 DQA1*04:01/DQB1*04:02 DPA1*02:01/DPB1*05:01 DQA1*03:01/DQB1*03:02 DRB1*04:01 DPA1*01:03/DPB1*02:01 DRB1*15:01 DRB1*08:02	N/A	79.08%	0.16	N/A	N/A

Read theme colors depicted epitope in the region target 203-204.

*Asterisk depicted that only the epitope in the area 203-204 is included in the calculation.

N-KKR: N protein linkage

Table 1: SARS-CoV-2 wild type N protein T cell HLA epitope predictions at gene position NP_LKR. HLA-class I binders depicted in red included only the epitope in the target region 203-204 in the calculation. Frequency of HLA class I representation within South American, Japanese, and Iranian populations is recorded as a frequency.

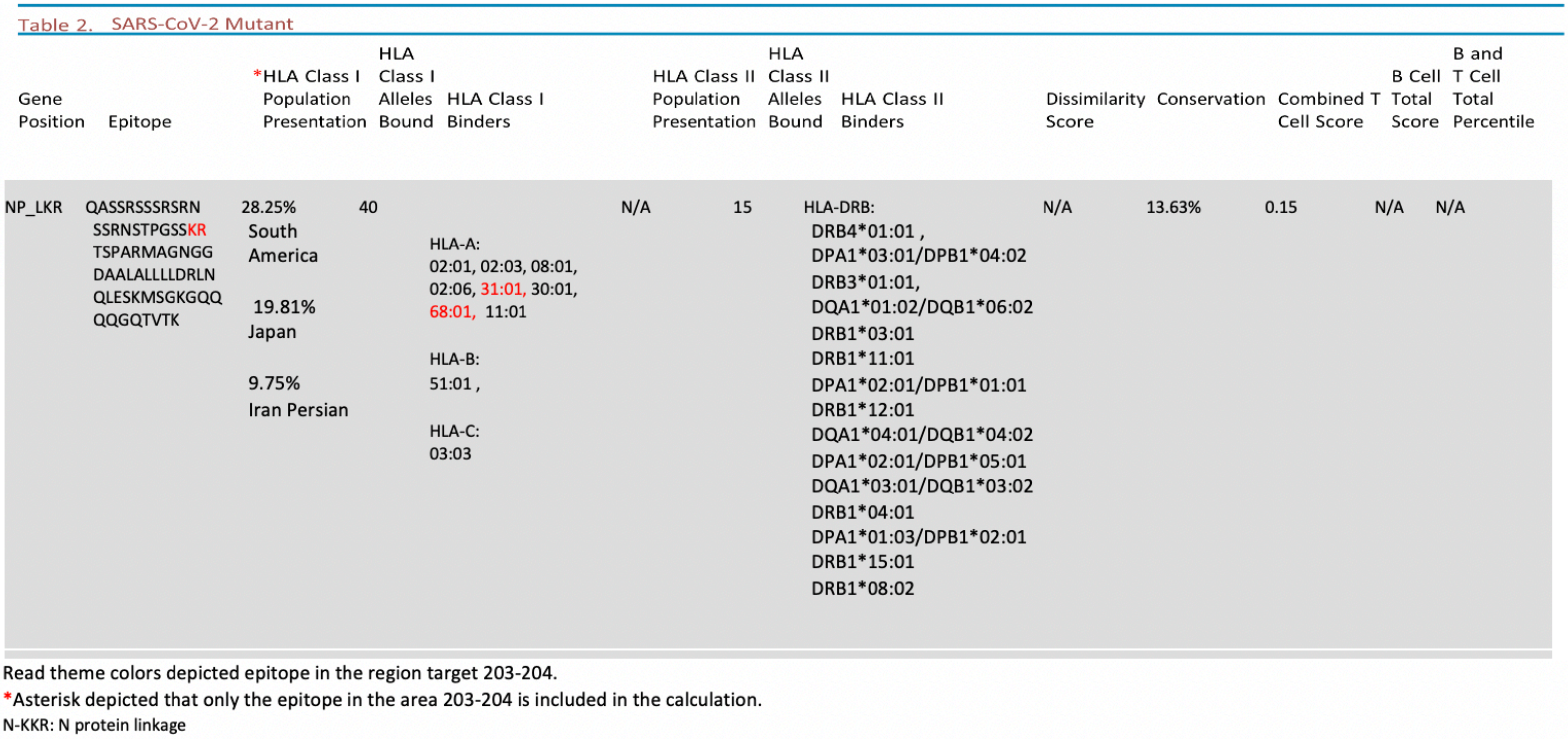


Table 2. SARS-CoV-2 Mutant

Gene Position	Epitope	*HLA Class I Population Presentation	HLA Class I Alleles Bound	HLA Class I Binders	HLA Class II Population Presentation	HLA Class II Alleles Bound	HLA Class II Binders	Dissimilarity Score	Conservation	Combined T Cell Score	B Cell Total Score	B and T Cell Total Percentile
NP_LKR	QASSRSSSRSRNSSRNSTPGSSKRTSPARMAGNGGDAALALLLLDRLNQLESKMSGKGQQQQGQTVTK	28.25% South America 19.81% Japan 9.75% Iran Persian	40	HLA-A: 02:01, 02:03, 08:01, 02:06, 31:01, 30:01, 68:01, 11:01 HLA-B: 51:01 , HLA-C: 03:03	N/A	15	HLA-DRB: DRB4*01:01 , DPA1*03:01/DPB1*04:02 DRB3*01:01, DQA1*01:02/DQB1*06:02 DRB1*03:01 DRB1*11:01 DPA1*02:01/DPB1*01:01 DRB1*12:01 DQA1*04:01/DQB1*04:02 DPA1*02:01/DPB1*05:01 DQA1*03:01/DQB1*03:02 DRB1*04:01 DPA1*01:03/DPB1*02:01 DRB1*15:01 DRB1*08:02	N/A	13.63%	0.15	N/A	N/A

Read theme colors depicted epitope in the region target 203-204.

*Asterisk depicted that only the epitope in the area 203-204 is included in the calculation.

N-KKR: N protein linkage

Table 2: SARS-CoV-2 mutant G204R type N protein T cell HLA epitope predictions at gene position NP_LKR. HLA-class I binders depicted in red included only the epitope in the target region 203-204 in the calculation. Frequency of HLA class I representation within South American, Japanese, and Iranian populations is recorded as a frequency.