Evolutionary insights into the envelope protein of SARS-CoV-2

M. Shaminur Rahman[1*], M. Nazmul Hoque[1*,2], M. Rafiul Islam[1*], Israt Islam[1], Israt Dilruba Mishu[1], Md. Mizanur Rahaman[1], Munawar Sultana[1], M. Anwar Hossain[1,3*]

[1]Department of Microbiology, University of Dhaka, Dhaka-1000, Bangladesh

[2]Department of Gynecology, Obstetrics and Reproductive Health, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur-1706, Bangladesh

[3]Present address: Vice-Chancellor, Jessore University of Science and Technology, Jessore 7408, Bangladesh

*Equal contribution

**Corresponding to:

M. Anwar Hossain, PhD
Professor
Department of Microbiology
University of Dhaka, Dhaka, Bangladesh
E-mail: hossaina@du.ac.bd

Running Head: Envelope protein of SARS-CoV-2

 doi:10.20944/preprints202008.0665.v1

Abstract

The ongoing mutations in the structural proteins of SARS-CoV-2 is the major impediment for prevention and control of the COVID-19 disease. The envelope (E) protein of SARS-CoV-2 is a structural protein existing in both monomeric and homopentameric forms, associated with a multitude of functions including virus assembly, replication, dissemination, release of virions, infection, pathogenesis, and immune response stimulation. In the present study, 81,818 high quality E protein sequences retrieving from the GISAID were subjected to mutational analyses. Our analysis revealed that only 0.012 % (982/81818) stains possessed amino acid (aa) substitutions in 63 sites of the genome while 58.77% mutations in the primary structure of nucleotides in 134 sites. We found the V25A mutation in the transmembrane domain which is a key factor for the homopentameric conformation of E protein. We also observed a triple cysteine motif harboring mutations (L39M, A41S, A41V, C43F, C43R, C43S, C44Y, N45R) which may hinder the binding of E protein with spike glycoprotein. These results therefore suggest the continuous monitoring of each structural protein of SARS-CoV-2 since the number of genome sequences from across the world are continuously increasing.

Keywords: SARS-CoV-2, envelop protein, mutations, transmembrane domain, triple cysteine motif.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0665.v1

The Study

SARS-CoV-2, the etiologic agent of COVID-19 disease has impacted the entire world, and created a public health emergency since December 2019[1,2]. The inherently higher mutations in the genome of SARS-CoV-2 have already produced many descendants from the original Wuhan strain, thereby escaping the host immune responses[3-6]. The genome of SARS-CoV-2 virus encodes for four major structural proteins such as the spike (S) protein, nucleocapsid (N) protein, membrane (M) protein, and the small envelope (E) protein, all of which are required to complete a successful infectious event/replication cycle of virus including entry, assembly, packaging and release of new virus particles within the human cells[8-11]. The E protein is the smallest of the major structural proteins, and associated with viral assembly, budding, envelope formation, and pathogenesis[7]. During the replication cycle, the virus expresses the E protein in high abundance inside the host cell, however, only a small portion is incorporated into the virion envelope. This protein carries out its functions by interacting membrane (M) and other accessory proteins viz. ORF3a, ORF7a, and host cell proteins[9,10]. The ongoing rapid transmission, and global spread of COVID-19 have raised intriguing questions whether the evolution and adaptation of SARS-CoV-2 is driven by synonymous mutations, deletions and/or replacements[5,12,13]. Although, the mutational spectra of different structural proteins (S, M, and N) of SARS-CoV-2 has been reported by several research groups[5,6,13-15] over a short period of time, however, available literature on the nucleotide and aa-level mutations of E protein is till limited.

To comprehensively analyze the mutational spectra of E protein of SARS-CoV-2 as a continuous part of the coronavirus genomic mutational research[5,6,15,16], we retrieved 83,607 complete or near-complete genome sequences of SARS-CoV-2 (human host) from the global initiative on sharing all influenza data (GISAID) (https://www.gisaid.org/) belonging to 159

 doi:10.20944/preprints202008.0665.v1

countries or territories till 20 August 2020 (Supplementary Data 1). We obtained 81,818 cleaned sequences (97.86%) after removing the low quality sequences. Multiple sequence alignment was performed in MAFFT by using Wuhan strain as a reference (NCBI accession no. NC_045512)[17], and nonsynonymous mutations were retrieved using the previously reported methods[5,6,15].

The mutational analysis of the present study revealed that only 0.012 % (982/81818) stains possessed amino acid (aa) substitutions in 63 sites of the E protein. Highest aa mutations (n=4) (C43R, C44Y, N45R, V47G) were found in a Moroccan strain of SARS-CoV-2 (EPI_ISL_467299) followed by two aa mutations in nine strains from England, Israel, Netherlands, Northern Ireland, Scotland, and Sierra Leone at different positions. Remarkably, rest of the strains (n= 972) possessed only one aa mutation in different positions of the E protein of the SARS-CoV-2 genome (Fig. 1, Supplementary Data 1).

The nucleotide (nt) level analysis of mutational spectra identified 58.77% mutations in 134 sites of the primary structure of the E protein. We found 7 nt variations in a Moroccan strain (EPI_ISL_467299), and 2 nt variations in 54 strains of SARS-CoV-2 from Australia, Austria, Canada, England, Guangdong, India, Iran, Israel, Netherlands, Northern_Ireland, Scotland, Sichuan, Sierra_Leone, Sweden, and Wales. The rest of the strains (n=1514) showed only 1 nt mutation (Supplementary Data 1).

In this study, we also observed worldwide mutational variations within the primer-probes binding sites of SARS-CoV-2 *E* gene (Table 1, Supplementary Data 1). We found a total of 74 nucleotide mutations that occupied the binding sites of primer-probes recommended by several research groups[18-20] for *E* gene targeted PCR-based detection of SARS-CoV-2 (Table 1). The forward primer of Charité, German contained 15 mismatches within primer in the viral strains of USA, England, India, Scotland and Wales, whereas a USA strain showed the 3´ end mismatch.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0665.v1

The reverse primer and the probe of the same set exhibited 14 and 17 mismatches, respectively. No strain showed 3´ end mismatch with the reverse primer while a strain of Netherlands showed 5´ end mismatch in the probe of the primer set (Table 1). Moreover, Park et al.[18] recommended primer set possessing 28 mismatches (forward=18 and reverse=10) in SARS-CoV-2 strains of many countries including Taiwan, Canada, Scotland, USA, Sierra_Leone, England, Austria, Guangdong and Spain. Noteworthy, the 3´ end of the forward primer of this set mismatched with SARS-CoV-2 strains of USA and England. However, taking only 180 strains into consideration, Nalla et al.[21] found one mismatch within reverse and one mismatch within probe binding sites of the primer set recommended by Charité, German. Thus, more mismatches in the primer-probe binding sites are found in our study that warrants the ongoing evolution of SARS-CoV-2 *E* gene[21].

However, mutations in the primer binding sites, importantly at 3´ end of primers, may affect the RT-PCR-based COVID-19 detection resulting in false negative results[21,22]. Besides primer mismatches, sequence variations in probe recognition sites might also affect the efficiency of RT-PCR-based detection of COVID-19 providing false negative (unable to bind) or false positive (non-specific binding) results[21,23]. Our study provides a global insight of *E* gene evolution, and its likely consequences on primer-based detection of SARS-CoV-2 by RT-PCR method. Overall this study warrants continuous monitoring and to update the primer-probe sequences based on the regional viral genomic sequences for efficient and accurate detection of COVID-19.

The primary structure of the E protein contains 75 aa of which 84.0% (63/75) sites underwent to 115 unique aa mutations (Fig. 1). Our analysis showed that 35 sites in the E protein structure underwent to more than one aa mutations, and of them, aa position 5 and 72 had aa variation numbers of 4 and 6, respectively (Table 2). Comparing the individual strain level mutations, we found that the S68F mutation in 250 strains (highest frequency) followed by L73F,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0665.v1

R69I, and P71L mutations noticed in 100, 88, and 59 strains, respectively. The N-terminal, transmembrane domain (TMD), and C-terminal domain of the E protein had 7, 25, and 31sites for aa substitutions, respectively (Fig. 1). Several earlier studies[5,8] also reported aa mutations is 10 sites (aa positions: 26, 36, 37, 39, 46, 58, 68, 71, 72, 73) of the E protein corroborating our current findings.

Identity, similarity, and the gap between SARS-CoV-2 (NC_045512.2) and SARS-CoV (NC_004718.3) E protein were 94.8%, 96.1%, and 1.3%, respectively. Two aa mutations, N15A and V25F were found in the TMD which may abolish the ion channeling capability of SARS-CoV E viroporin structure, a key factor of its homopentameric conformation[9,24-26]. We observed the V25A mutation in six strains from Spain, Canada, and England that may hamper the oligomerisation of the E protein of SARS-CoV-2, at least to some extent. Moreover, a triple cysteine motif (38-NH_2-L**C**AY**CC**N-COOH-44), and similar motif located in the C-terminus of S protein of SARS-CoV were predicted to interact with each other. This interaction can serve as a structural basis between E and S proteins which would be enhanced by the disulphide bonding to the corresponding cysteine residues[9,27]. Mutations (L39M, A41S, A41V, C43F, C43R, C43S, C44Y, N45R) in this interacting motif of the E protein were also evident in different strains (Supplementary Data 1). The C43F substitution was observed in six strains from England, Saudi Arabia, and the USA whereas C43R and C44Y mutations were noticed in two different strains of SARS-CoV-2 deposited to the GISAID from Morocco. We also found C43S mutation in one of the Australian strains (Supplementary Data 1). The mutations found in the E protein may hamper the genomic structure of SARS-CoV-2, and the mutated E protein might affect the viral assembly, replication, propagation, and pathogenesis as also previously observed in SARS-CoV and MERS-CoV[27,28]. Therefore, mutated E protein can be a potential target for SARS-CoV-2 viral

 doi:10.20944/preprints202008.0665.v1

inactivation, and reduction of pathogenicity. The identification of the nucleotides and amino acids which are involved in virulence reduction should be investigated by further studies. The results of the present study should be interpreted cautiously given the existing uncertainty of SARS-CoV-2 genomic data to develop potential prophylaxis and mitigation for tackling the pandemic COVID-19 crisis.

 doi:10.20944/preprints202008.0665.v1

Acknowledgements

The authors would like to acknowledge the frontliner's who are working restlessly against this COVID-19 pandemic situations. The authors also appreciate the researchers worldwide who were kind enough to deposit and share the complete genomes of SARS-CoV-2 and other coronaviruses to the GISAID.

Competing interest

The authors declare no competing interests.

Data availability

We used the SARS-CoV-2 genome sequences available in the open shared database (GISAID).

Author contributions

MSR conducted the overall study. MSR, MNH, MRI, II, and IDM interpreted the results and drafted the manuscript. MNH finally compiled and edited the manuscript. MMR, MS and MAH contributed intellectually to the interpretation and presentation of the results.

Ethical statements

We confirm that the ethical policies of the journal, as noted on the journal's authors guideline page, have been adhered to. No ethical approval was required since the study didn't include any animal or human sample.

Supplementary Information

Supplementary information supporting the findings of this study are available in this article as Supplementary Data.

 doi:10.20944/preprints202008.0665.v1

References

1. Zhang Y-Z, Holmes EC. A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell.* 2020.

2. Tang X, Wu C, Li X, et al. On the origin and continuing evolution of SARS-CoV-2. *National Science Review.* 2020.

3. Li Y, Yang X, Wang N, et al. The divergence between SARS-CoV-2 and RaTG13 might be overestimated due to the extensive RNA modification. *Future Virology.* 2020(0).

4. DeDiego ML, Pewe L, Alvarez E, Rejas MT, Perlman S, Enjuanes L. Pathogenicity of severe acute respiratory coronavirus deletion mutants in hACE-2 transgenic mice. *Virology.* 2008;376(2):379-389.

5. Islam MR, Hoque MN, Rahman MS, et al. Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. *Scientific Reports.* 2020;10(1):1-9.

6. Rahman MS, Islam MR, Hoque MN, et al. Comprehensive annotations of the mutational spectra of SARS-CoV-2 spike protein: a fast and accurate pipeline. *bioRxiv.* 2020.

7. Hoque MN, Chaudhury A, Akanda MAM, Hossain MA, Islam MT. Genomic Diversity and Evolution, Diagnosis, Prevention, and Therapeutics of the Pandemic COVID-19 Disease. 2020. PeerJ, 8:e9689 http://doi.org/10.7717/peerj.9689.

8. Hassan SS, Choudhury PP, Roy B. SARS-CoV2 envelope protein: non-synonymous mutations and its consequences. 2020.

9. Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. *Virology journal.* 2019;16(1):1-22.

10. McBride R, Van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. *Viruses.* 2014;6(8):2991-3018.

11. Rahman MS, Hoque MN, Islam MR, et al. Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2, the etiologic agent of COVID-19 pandemic: an in silico approach. *PeerJ.* 2020;8:e9572.

12. Bal A, Destras G, Gaymard A, et al. Molecular characterization of SARS-CoV-2 in the first COVID-19 cluster in France reveals an amino acid deletion in nsp2 (Asp268del). *Clinical Microbiology and Infection.* 2020.

13. Pachetti M, Marini B, Benedetti F, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine.* 2020;18:1-9.

14. Phan T. Genetic diversity and evolution of SARS-CoV-2. *Infection, genetics and evolution.* 2020;81:104260.

15. Rahman MS, Islam MR, Alam ARU, et al. Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein (N protein) and its consequences. *BioRxiv.* 2020.

16. Ul Alam AR, Rafiul Islam M, Shaminur Rahman M, Islam OK, Anwar Hossain M. Understanding the possible origin and genotyping of first Bangladeshi SARS-CoV-2 strain. *Journal of Medical Virology.* 2020.

17. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics.* 2019;20(4):1160-1166.

18. Park M, Won J, Choi BY, Lee CJ. Optimization of primer sets and detection protocols for SARS-CoV-2 of coronavirus disease 2019 (COVID-19) using PCR and real-time PCR. *Experimental & molecular medicine.* 2020;52(6):963-977.

19. Chu DK, Pan Y, Cheng SM, et al. Molecular diagnosis of a novel coronavirus (2019-nCoV) causing an outbreak of pneumonia. *Clinical chemistry.* 2020;66(4):549-555.

20. D'Cruz RJ, Currier AW, Sampson VB. Laboratory testing methods for novel severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2). *Frontiers in Cell and Developmental Biology.* 2020;8.

21. Nalla AK, Casto AM, Huang M-LW, et al. Comparative performance of SARS-CoV-2 detection assays using seven different primer-probe sets and one assay kit. *Journal of clinical microbiology.* 2020;58(6).

22. Rana DR, Pokhrel N. Sequence mismatch in PCR probes may mask the COVID-19 detection in Nepal. *Molecular and Cellular Probes.* 2020;53:101599.

23. Kamau E, Agoti CN, Lewa CS, et al. Recent sequence variation in probe binding site affected detection of respiratory syncytial virus group B by real-time RT-PCR. *Journal of Clinical Virology.* 2017;88:21-25.

24. Torres J, Maheswari U, Parthasarathy K, Ng L, Liu DX, Gong X. Conductance and amantadine binding of a pore formed by a lysine-flanked transmembrane domain of SARS coronavirus envelope protein. *Protein science.* 2007;16(9):2065-2071.

25. Verdiá-Báguena C, Nieto-Torres JL, Alcaraz A, et al. Coronavirus E protein forms ion channels with functionally and structurally-involved membrane lipids. *Virology.* 2012;432(2):485-494.

26. Torres J, Parthasarathy K, Lin X, Saravanan R, Kukol A, Liu DX. Model of a putative pore: the pentameric α-helical bundle of SARS coronavirus E protein in lipid bilayers. *Biophysical journal.* 2006;91(3):938-947.

27. Wu Q, Zhang Y, Lü H, et al. The E protein is a multifunctional membrane protein of SARS-CoV. *Genomics, proteomics & bioinformatics.* 2003;1(2):131-144.

28. DeDiego ML, Nieto-Torres JL, Jimenez-Guardeño JM, et al. Coronavirus virulence genes with main focus on SARS-CoV envelope gene. *Virus research.* 2014;194:124-137.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0665.v1

Figure

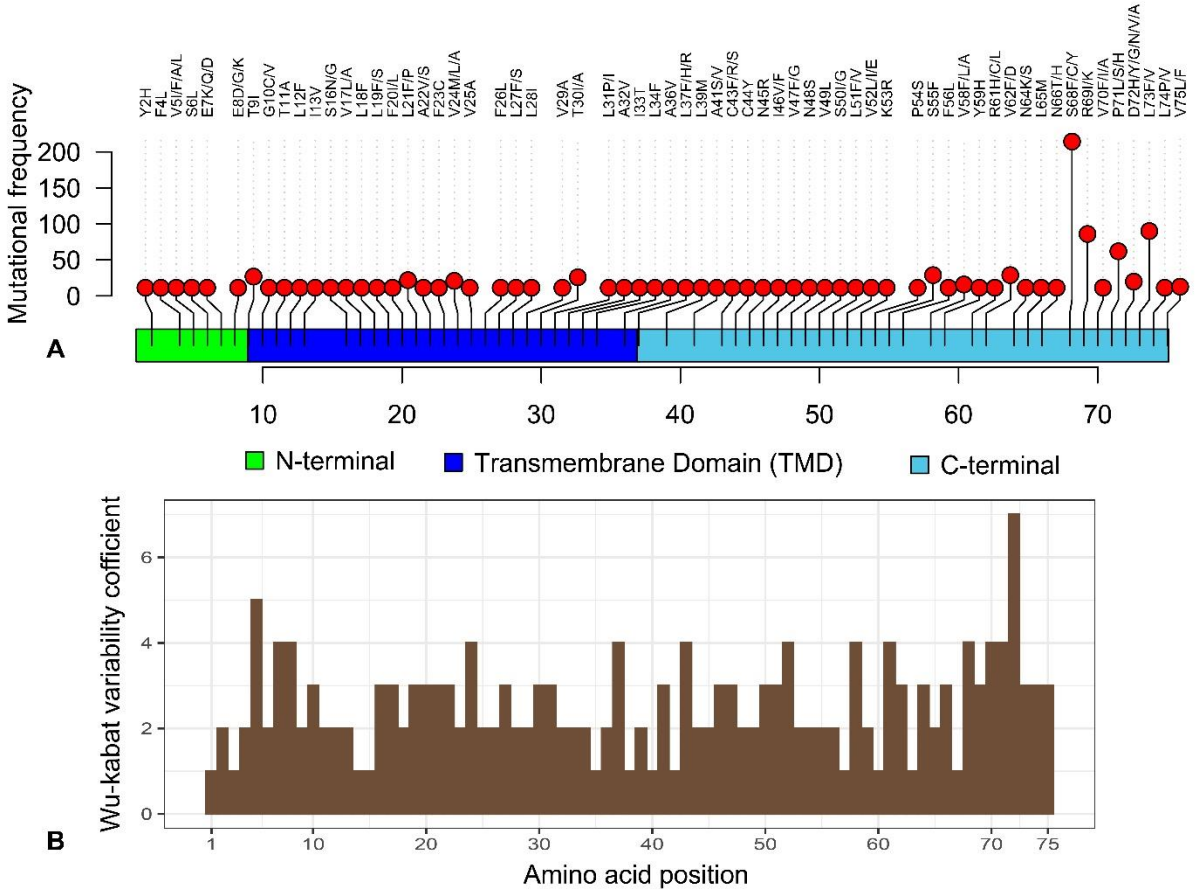

Fig. 1: Overview and variability coefficient of the envelop (E) protein of SARS-CoV-2. (A) Mapping and frequency distribution of mutations in the E protein of SARS-CoV-2 strains through Loliplot visualization. (B) Wu-Kabat variability coefficient of E protein of SARS-CoV-2. Here, variability coefficient 1 indicates the conservancy, whereas coefficients > 1 indicate relative variability of the respective positions. The more the coefficient value the more the variability or diversity (Supplementary Data 1).

 doi:10.20944/preprints202008.0665.v1

Tables

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0665.v1

Table 1: Mutation in primer probe binding sites of SARS-CoV-2 E gene

Reference	F/R/P*	Sequence	Position	Mutation within primer region	No. of mismatch	Countries went mutation	3' end mismatch country
Chu et al., 2020; D'Cruz et al., 2020; Charité, Germany	F	ACAGGTACGTTAATAGTTAATAGCGT	25-50	C26T, G28T, G29T, T30C, A31G, G33T, A36G, A36T, A37G, T42C, A46G, G47A, C48T, G49T, T50C	15	USA England India Scotland Wales	USA
	R	ATATTGCAGCAGTACGCACACA	116-137	G121T, C122T, G123A, C126T, T127C, G128T, G128C, C129T, G131A, C132T, A134G, T135G, A136G, A136T	14	Austria Scotland USA Guangdong Spain England	-
	P	FAM-ACACTAGCCATCCTTACTGCGCTTCG-BBQ	88-113	A88G, C89T, A90G, C91A, T92C, C95T, C96T, T98C, C99T, C100T, C107T, G108T, C109T, T110A, T110G, T111G, T111A	17	Netherlands Canada USA England Wales India Turkey	-
Park et al. 2020	F	TTCGGAAGAGACAGGTACGTTA	15-36	C17T, G18A, G18T, G19A, G19C, A21T, G22A, A23G, G24T, G24C, C26T, G28T, G29T, T30C, A31G, G33T, A36G, A36T	18	Taiwan Canada Scotland USA Sierra_Leone England	USA, England
	R	AGCAGTACGCACACAATCG	112-130	A114G, T115A, G121T, C122T, G123A, C126T, T127C, G128T, G128C, C129T	10	Scotland USA Austria Guangdong Spain England	-

*F=Forward primer, R=Reverse primer, P=Probe

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2020 doi:10.20944/preprints202008.0665.v1

Table 2: Amino acid (aa) variation in envelop (E) protein of SARS-CoV-2.

Position	Number of aa variation	aa (Ref:position:strain)	Reference aa characteristics	Strains aa characteristics
72	6	D72Y,D72G,D72H,D72N,D72V,D72A	NC	P,NP,PC,P,NP,NP
5	4	V5I,V5F,V5A,V5L	NP	NP,NP,NP,NP
7	3	E7K,E7Q,E7D	NC	PC,P,NC
8	3	E8G,E8D,E8K	NC	NP,NC,PC
24	3	V24M,V24L,V24A	NP	NP.NP.NP
37	3	L37H,L37F,L37R	NP	PC,NP,PC
43	3	C43F,C43R,C43S	NP	NP,PC,P
52	3	V52I,V52L,V52E	NP	NP,NP,NC
58	3	V58F,V58L,V58A	NP	NP,NP,NP
61	3	R61H,R61C,R61L	PC	PC, P,NP
68	3	S68F,S68C,S68Y	P	NP,NP,P
70	3	V70F,V70I,V70A	NP	NP,NP,NP
71	3	P71L,P71S,P71H	NP	NP,P,PC
10	2	G10C,G10V	NP	NP,NP
16	2	S16N,S16G	P	P,NP
17	2	V17L,V17A	NP	NP,NP
19	2	L19S,L19F	NP	P,NP
20	2	F20I,F20L	NP	NP,NP
21	2	L21F,L21P	NP	NP,NP
22	2	A22V,A22S	NP	NP,P
27	2	L27F,L27S	NP	NP,P
30	2	T30I,T30A	P	NP,NP
31	2	L31P,L31I	NP	NP,NP
41	2	A41S,A41V	NP	P,NP
46	2	I46V,I46F	NP	NP,NP
47	2	V47F,V47G	NP	NP,NP
50	2	S50I,S50G	P	NP,NP
51	2	L51F,L51V	NP	NP,NP
62	2	V62F,V62D	NP	NP,NC
64	2	N64K,N64S	P	PC,P
66	2	N66T,N66H	P	P,PC
69	2	R69I,R69K	PC	NP,PC
73	2	L73F,L73V	NP	NP,NP
74	2	L74P,L74V	NP	NP,NP
75	2	V75L,V75F	NP	NP,NP

*NP=Non-Polar, P=Polar, PC=Positive Charge, NC=Negative Charge