RESEARCH ARTICLE
doi: 10.2306/scienceasia1513-1874.2022.138

ScienceAsia 49 (2023): **77–84**

Computational analysis of the novel Thailand-specific mutations in SARS-CoV-2 spike glycoprotein sequences

Pongpat Chaisawasd, Sirawit Ittisoponpisan*

[a] Center for Genomics and Bioinformatics Research, Division of Biological Science, Faculty of Science, Prince of Songkhla University, Songkhla 90110 Thailand

*Corresponding author, e-mail: sirawit.i@psu.ac.th

Received 19 May 2022, Accepted 4 Sep 2022
Available online 10 Oct 2022

ABSTRACT: As of 14 January 2022, 2.3 million people in Thailand had been reportedly infected with SARS-CoV-2, and 21,883 people had died. Spike glycoprotein, on the SARS-CoV-2 membrane, is a key factor for viral infection. Some scientists have demonstrated that some amino acid mutations in the protein increase infectivity and transmissibility of the virus. However, many studies concerning mutations in the spike glycoprotein, particularly in Thailand, were not comprehensive enough to illustrate the impacts of the mutations on the spike glycoprotein. To narrow this gap, we examined approximately 8.3 million spike glycoprotein sequences retrieved from GISAID Initiative and NCBI Virus databases to identify novel mutations. Limiting our scope to the Thai samples, we demonstrated how local SARS-CoV-2 strains changed over 2021. In addition, we found that T95I had emerged and become one of the main characteristics of delta strains in Thailand. We further detected 28 Thailand-specific novel mutations, which were then analyzed with amino acid-based analysis tools to gain insights into their impacts on the spike glycoprotein. Upon closer examination, we found that 2 novel mutations, L249E and R457W, were likely to diminish the interactions between the spike glycoprotein and neutralizing antibodies *in silico*. This finding suggests that both mutations may hinder the neutralization, allowing the virus to escape the antibodies. Additionally, our study highlights the significant effects of some novel mutations on the stability and functionality of the spike glycoprotein, which may be useful for COVID-19 diagnosis and vaccine development.

KEYWORDS: SARS-CoV-2, spike glycoprotein, Corona variant, mutagenesis, Thailand-specific mutation

INTRODUCTION

As of early January 2022, the COVID-19 pandemic has claimed approximately 5.5 million people's lives, leaving over 346 million people infected [1]. In Thailand, people have been encouraged to follow technical guidance for COVID-19 prevention since the first country-wide outbreak of the SARS-CoV-2 [2]. In addition to local restrictions, disease surveillance should also be comprehensively conducted.

COVID-19 is caused by the coronavirus named SARS-CoV-2. The virus uses its protruding membrane proteins called spike glycoproteins to enter the host cell cytoplasm. Initially, the spike glycoprotein in trimer form (3 monomers hinged symmetrically) binds to a region of the host receptor angiotensin-converting enzyme 2 (ACE2). As a result, the trimer form is reorganized into post-fusion conformation. At this stage, the protein subsequently binds to a region of the enzyme transmembrane serine protease 2 (TMPRSS2), triggering membrane fusion and viral entry [3, 4]. To prevent such a phenomenon, some scientists used neutralizing antibodies (Nabs), which primarily bind to N-terminal domain (NTD) and receptor-binding domain (RBD) of the spike glycoprotein, to block interactions between the spike glycoprotein and the host cell enzymes [5, 6]. Several studies concerning these 2 domains have reported that some mutations in the domains help the virus invade the host cell more effectively. Dicken et al [7] suggested that the deletions, Δ69, Δ70, and Δ144 (all located in the N-terminal domain) in the spike glycoprotein of the Alpha variant, could help increase transmissibility of the virus. In addition, some mutations in the receptor-binding domain of the protein have been reported to increase RBD-ACE2 interaction and diminish binding to neutralizing antibodies [8–12]. Another concern is that mutations, which cause changes in the stability and functionality of the spike glycoproteins in pre- and post-fusion conformations, influence infectivity and transmissibility of the virus [13]. These findings suggest that a thorough study of novel mutations in the spike glycoprotein should be conducted.

Despite several studies looking into the effects of mutations on the spike glycoprotein and its interactions with antibodies or human ACE2 [9–12, 14–18], they only focused on well-known mutations and did not implement comprehensive instruments to measure the stability and functionality of the protein. In Thailand, such studies were absent, and the concerns of most SARS-CoV-2 studies were on genomic surveillance, the pattern of transmission of COVID-19, and partially structural analysis of the spike glycoprotein [19–23]. To fill this gap, we retrieved over 8.3 million spike glycoprotein sequences from GISAID Initiate and NCBI Virus databases, deposited globally and analyzed them to detect novel mutations. Consequently, the novel mutations were examined in terms of hydrogen

Table 1 Thai sequences containing novel mutations.

Accession no.	Novel spike mutation(s)	Collection date
EPI_ISL_3419394	N824E	1 Jul 2021
EPI_ISL_4106140	R457W	8 Aug 2021
EPI_ISL_4106141	Q1208P, P1213K	22 Aug 2021
EPI_ISL_4106144	E918N	1 Aug 2021
EPI_ISL_4254660	Y1272G	1 Sep 2021
EPI_ISL_4348473	L241W	7 Sep 2021
EPI_ISL_4348474	D290P	6 Sep 2021
EPI_ISL_4635439	H66K	7 Sep 2021
EPI_ISL_4636457	L249E	4 Sep 2021
EPI_ISL_5030519	ins19RQKR	24 Aug 2021
EPI_ISL_5052380	S735C, M740H, I742K	1 Sep 2021
EPI_ISL_5446194	V70H	7 Oct 2021
EPI_ISL_5881197	P330M	30 Sep 2021
EPI_ISL_5881514	ins68KRTLLFWN**	9 Oct 2021
EPI_ISL_5884497	ins63KKLKK	4 Oct 2021
EPI_ISL_5887156	H519E	27 Sep 2021
EPI_ISL_5916960	L1224E, I1225E, I1227R	15 Oct 2021
EPI_ISL_5916980	Q992C, V1176R	15 Oct 2021
EPI_ISL_6136054	I1130E, I1132R	9 Oct 2021
EPI_ISL_6136120	R102C	18 Oct 2021
EPI_ISL_6334729	Ins99K*	24 Oct 2021
EPI_ISL_6695346	ins62NLRK	28 Oct 2021

There are 23 substitution mutations and 5 insertion mutations. * The smallest insertion; ** the largest insertion.

bond formation across nearby amino acid residues, protein stability change upon mutation and clash among surrounding residues. These terms were used to describe the impacts of the mutations on the spike glycoprotein and interactions between the protein and neutralizing antibodies. These findings may be useful for scientists who study the spike glycoprotein in the fields of biochemistry, physiochemistry, and bioinformatics. Furthermore, our results could benefit disease surveillance, help researchers track circulating SARS-CoV-2 strains as well as analyze the severity of the disease to keep the outbreak of COVID-19 under control.

MATERIALS AND METHODS

SARS-CoV-2 spike glycoprotein sequence retrieval and mutation identifications

We retrieved 8,330,646 SARS-CoV-2 spike glycoprotein sequences from NCBI Virus (www.ncbi.nlm.nih.gov/sars-cov-2) and GISAID Initiative (www.gisaid.org) databases on 6 January 2022. Then, we used Python programming to obtain sequences whose lengths were between 1,209 and 1,337. This range was based on the length of the spike glycoprotein reference sequence (NCBI Protein Reference Sequence: YP_009724390.1) (1,273 ± 5%). Accordingly, 8,214,373 sequences (98.6%) remained in our final dataset — 6,576,102 were from GISAID Initiative sequences (10,355 were from Thailand) and 1,638,271 from NCBI Virus sequences (302 were from Thailand). In total, 10,657 were from Thailand, and 8,203,716 were from other countries.

Next, we identified mutations in each sequence by performing pairwise alignment between each sequence of interest and the Wuhan reference sequence (YP_009724390.1) using PairwiseAligner method available in Bio.Align sub-package from Biopython [24]. The BLOSUM or PAM scoring matrices were excluded from the amino acid sequence alignments due to the time-consuming process. For quick analysis, simple scoring criteria for pairwise alignment were set as follows: match_score = 1.0, mismatch_score = −2.0, open_gap_score = −3.0, and extend_gap_score = −2.5. For more details, see the Python scripts in Supplementary File 1.

After identifying mutations in each sequence in our dataset, we then categorized them into “Thai” and “non-Thai” groups. As a result, 1,388 spike glycoprotein mutations were detected in the Thai samples, whereas 15,175 mutations were detected in the non-Thai samples. Accordingly, 28 mutations in Thai spike glycoprotein sequences were novel — they were not found in any non-Thai sequences (Table 1).

Structural investigation of the mutations on the SARS-CoV-2 spike glycoprotein

The 28 Thailand-specific mutations were set to be analyzed with Missense3D (http://missense3d.bc.ic.ac.uk/missense3d) for the structural changes upon amino acid substitutions [25] and with mCSM (http://biosig.unimelb.edu.au/mcsm) for the protein stability changes upon the mutations [26]. PDB structures used in these analyses are the spike glycoproteins in 2 conformations — the close and the post-fusion — and the complexes of partial spike/antibody (see Table 2). Our PDB structure selection criteria were based on the quality of the structure (< 3.5 Å resolution). Due

Table 2 List of PDB structures in our analysis.

ID	Type	Resolution	Purpose	Ref.
6XR8	EM	2.90 Å	To represent the distinct conformational state of the spike glycoprotein trimer for the analyses of the impacts of the mutations on the protein structure.	[30]
6XRA	EM	3.00 Å	To represent the post-fusion conformational state of the spike glycoprotein for the analyses of the impacts of the mutations on the protein structure.	[30]
7B3O**	X-Ray	2.00 Å	To analyze how mutations affect the interaction between the neutralizing antibodies and the spike glycoprotein.	[31, 32]
7E7X*	X-Ray	2.78 Å		
7E86**	X-Ray	2.90 Å		
7E88**	X-Ray	3.14 Å		
7E8F*	EM	3.18 Å		

* PDB structures used in NTD-mAbs interaction analysis.
** PDB structures used in RBD-mAbs interaction analysis.

to limitations, single amino acid substitution analyses are possible on the web servers. Indel mutations and some amino acid substitutions, which were not detected in the PDB structures, were all excluded from the analyses.

RESULTS AND DISCUSSION

Statistical analysis of the spike mutations and the variants of concern in 2021

We have determined the presence of the top 10 most common mutations, detected in 9,940 Thai sequences throughout year 2021 (Fig. 1a). Notably, the D614G mutation had been detected in nearly all Thai sequences throughout the year. Other 8 mutations: P681R, L452R, T478K, T19R, D950N, E156-, R158G, and F157-, had been detected since June 2021. Interestingly, these mutations indicated the presence of the Delta strains which had emerged in Thailand since June 2021. Another mutation, T95I, had also been detected in July 2021. Although the T95I mutation was not listed as one of the characteristics of the Delta strain reported by WHO [27], this mutation was detected only in a portion of Delta-strain sequences from the Thai samples. The result indicates that the T95I mutation could have newly emerged among the Delta strains.

Furthermore, the Thai sequences from the year 2021 were analyzed to determine the presence of the 5 Variants of Concern (VOCs): Alpha, Beta, Gamma, Delta, and Omicron, all classified according to their constituent mutations reported by WHO [27] (Fig. 1b). The Alpha variant had been found to be dominant from April to June. The Delta variant had been found to be dominant from July to December. One Gamma variant sequence was detected in April. According to GISAID Initiative sequence information, which is freely available for registered users at https://gisaid.org, the sequence (EPI_ISL_1708575) was from a patient with a travel history from France. Two Omicron variant sequences (EPI_ISL_7398578 and EPI_ISL_774767) were detected in late December 2021, indicating the emergence of the Omicron variant in Thailand. Although the Omicron variant became dominant in Thailand after early January 2022, the data retrieved as of 6 January 2022 did not well cover the outbreak of the variant. Additionally, 1,718 sequences remained unclassified, as their mutations did not meet the criteria for any VOC.

In addition to the common mutations which have been well-documented, 28 Thailand-specific mutations have also been detected. Notably, each was found in only one strain (see Table 1). These mutations were located as follows:
11 in the N-terminal domain, 3 in the receptor-binding domain, 3 in the subdomain 2, 1 in the central helix domain, 2 in the connector domain, 1 in the heptad repeat 1 domain, 5 in the transmembrane domain, 5 in the heptad repeat 2 domains, and 1 in the cytoplasmic tail domain (Fig. 1c). According to Table 1, there are 23 substitutions and 5 insertions. The largest amino acid insertion was ins68KRTLLFWN. All the insertions were detected in the N-terminal domain. In addition, the highest density of mutations was measured in the N-terminal domain (AVG: 1.67 mutations/residue) (see Supplementary Table S1). These findings suggest that the N-terminal domain is the most susceptible to mutation.

Hydrogen bond analysis

Fifteen of 28 Thailand-specific mutations (including H66K, R102C, L241W, D290P, P330M, R457W, H519E, S735C, M740H, I742K, N824E, E918N, Q992N, I1130E, and I1132R) were analyzed on the PDB structure 6XR8 (chain A). However, the other 13 mutations were excluded, as their residues were not located in the PDB structure.

When analyzed with Missense3D, all the 15 mutations were found to form at least one hydrogen bond with nearby amino acids. Six novel mutations were predicted to form extra hydrogen bonds, whereas the other 5 mutations were predicted to reduce hydrogen bonds. The remaining 4 mutations did not alter the number of hydrogen bonds. Interestingly, the D290P mutation reduced most hydrogen bonds from four to one (Fig. 2a). D290 formed 4 hydrogen bonds with PHE58 (3.19 Å), ARG273 (3.29 Å), ALA292 (3.30 Å), and SER297 (3.79 Å), whereas P290 formed only one hydrogen bond with SER297 (3.79 Å). The result suggests that the D290P mutation may reduce the structural stability of the N-terminal domain (see Supplementary File 2).

In the post-fusion spike glycoprotein structure, 7 mutations including S735C, M740H, I742K, E918N, Q992N, I1130E, and I1132R could be analyzed on the PDB structure 6XRA representing the post-fusion spike glycoprotein, as their residues were covered within the PDB structure. None of these mutations was found to alter the total number of hydrogen bonds. However, some changes can be observed in the number of intra- and interchain hydrogen bonds. Fig. 2b demonstrates the change in hydrogen bonds caused by the Q992C mutation. The Q992 (chain A) formed hydrogen bonds with GLU988 in the same chain (2.69 Å) and with GLN755 in chain B (2.67 Å) and LYS1157 in chain C (3.89 Å). The C992 formed hydrogen bonds in the same chain with GLU988 (3.74 Å) and ARG995 (3.37 Å) and with GLN755 in chain B (2.92 Å). The depletion of interchain hydrogen bonds caused by the Q992C mutation suggests that the mutation may diminish the interaction between chain A and chain C, leading to the separation of the chains.

The L249E mutation was predicted to form an additional intrachain hydrogen bond at the interface of the spike glycoprotein/N11 complex (Fig. 2c). Orig-

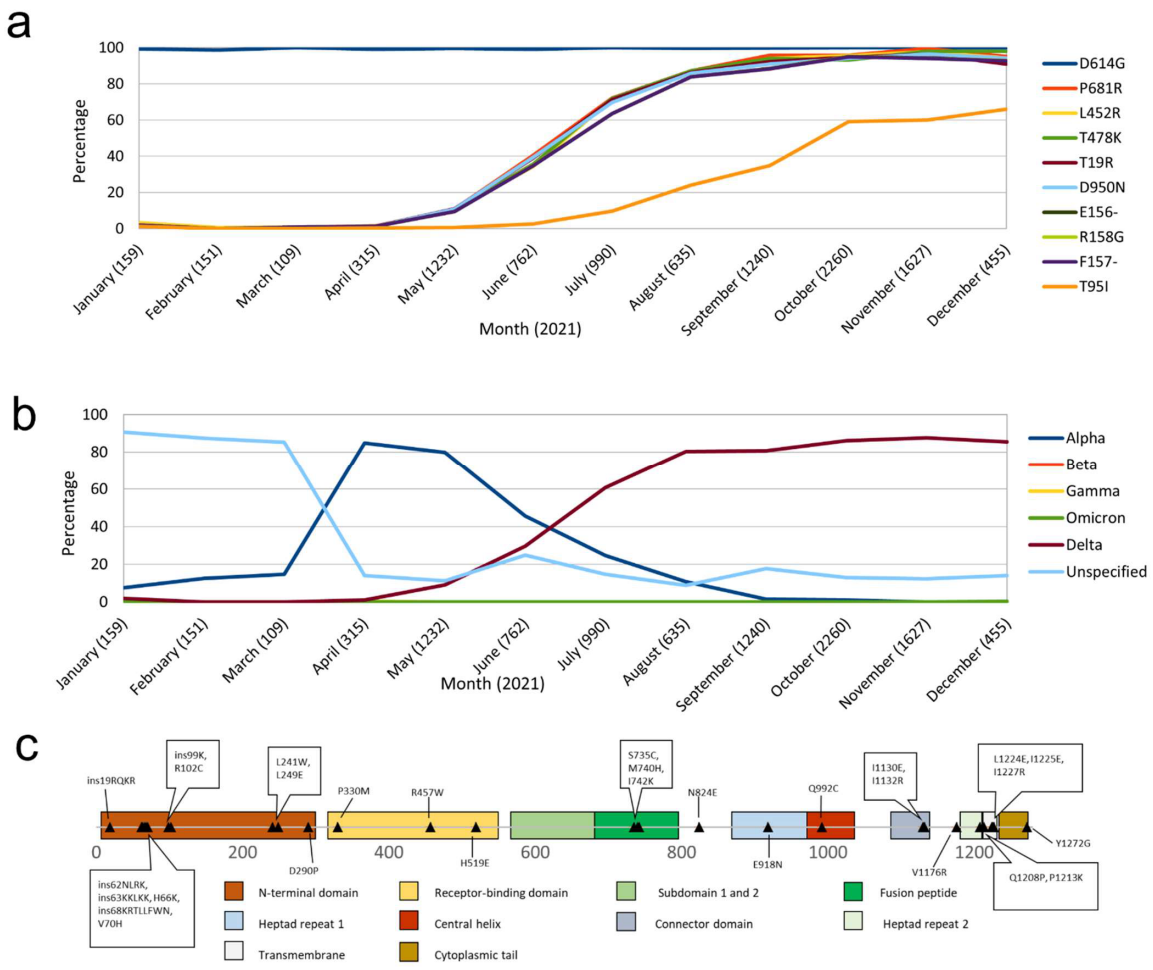

Fig. 1 Analysis of the mutations and variants of concern found in Thailand in 2021. (a) Monthly distribution of Top 10 mostly found spike mutations in 2021. (b) Monthly distribution of the 5 variants of concern in 2021. (c) Mapping of 28 Thailand-specific novel mutations (represented by black triangles) onto SARS-CoV-2 spike glycoprotein sequence.

inally, the L249 (chain A) formed hydrogen bonds with ASP253 (3.00 Å) from the same chain and with THR107 (3.55 Å) and TRP109 (3.73 Å) from the heavy chain of N11, whereas the E249 (chain A) formed hydrogen bonds with SER247 (2.74 Å) and ASP253 (3.00 Å) from the same chain and with THR107 (3.52 Å) and TRP109 (3.73 Å) from the heavy chain of N11. Despite the additional hydrogen bond, the number of interchain hydrogen bonds between the spike glycoprotein and N11 remained the same, so it is unlikely that the L249E mutation would increase the interaction between the spike glycoprotein and N11. In contrast, the R457W mutation greatly reduced hydrogen bonds in the spike/STE90-C11 complex (Fig. 2d). The R457 (chain A) originally formed hydrogen bonds with SER459 (2.90 Å), ASP460 (3.59 Å), GLU465 (3.81 Å), and ASP467 (2.76 and 3.17 Å) from the same chain and with SER53 (2.55 Å) from the heavy chain of STE90-C11, whereas the W457 (chain A) formed only a hydrogen bond with only SER53 (2.55 Å). Therefore, the R457W mutation may diminish the interaction between the spike protein structure and the STE90-C11 antibody (see Supplementary File 2).

Protein stability changes upon mutations

The stability changes upon mutations (defined by the difference in folding free energy (ΔG) between the wild type and the mutant structures: $\Delta\Delta G = \Delta G_m - \Delta G_w$) were calculated with mCSM. Table 3 shows the $\Delta\Delta G$ (in kcal/mol) of the 15 Thailand-specific spike mutations. According to mCSM's terminology [26], positive $\Delta\Delta G$ indicate stabilizing mutations, while negative ones mean they are destabilizing mutations. We found that the N824E mutation was the only one

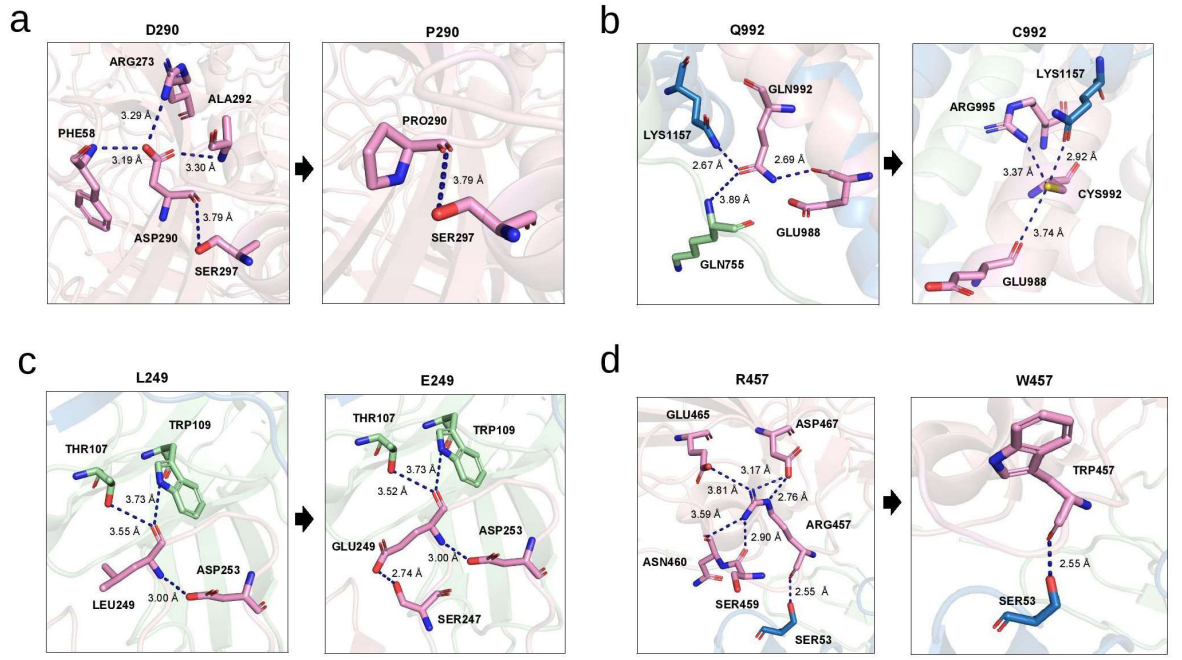

Fig. 2 Analysis of hydrogen bond formation caused by Thailand-specific spike glycoprotein mutations. (a) D290P mutation on the spike glycoprotein trimer (PDB: 6XR8; chain A). (b) Q992C mutation on the post-fusion spike glycoprotein (PDB: 6XRA; pink for chain A, light green for chain B, and sky blue for chain C). (c) L249E mutation on the N-terminal domain (pink) interacting with the heavy chain of antibody N11 (light green) (PDB: 7E7X; pink for chain A and light green for chain H). (d) R457W mutation on the receptor-binding domain (pink) interacting with the heavy chain of antibody STE90-C11 (sky blue) (PDB: 7B3O; pink for chain A and sky blue for chain H). The sky-blue amino acid residues represent amino acids from the antibody. All figures were generated with PyMOL.

that had a slightly stabilizing effect on the spike glycoprotein trimer (0.185 kcal/mol), whereas the other 14 mutations had destabilizing effects (Table 3). Notably, the greatest stability changes, considered to have highly destabilizing effects, were caused by L241W (−2.175 kcal/mol) and I742K (−2.275 kcal/mol). The result suggests that both mutations could reduce the structural stability in the domains of the spike glycoprotein trimer.

Seven mutations including S735C, M740H, I742K, E918N, Q992N, I1130E, and I1132R were analyzed on the full PDB structure 6XRA representing the post-fusion spike glycoprotein. The 2 highest energy changes were 0.233 kcal/mol (stabilizing effect) caused by E918N and −1.802 kcal/mol (destabilizing effect) caused by I742K. The 2 lowest energy changes were 0.184 kcal/mol (stabilizing effect) caused by I1132R and −0.217 kcal/mol (destabilizing effect) caused by the S735C mutation.

The spike glycoprotein in the post-fusion conformation plays an important role in the host cell invasion. After binding to TMPRSS2, the protein refolds and embeds its fusion peptide-proximal region (FPPR) into the host cell membrane, triggering membrane fusion [28–31]. Therefore, amino acid mutations that have stabilizing effects on the post-fusion spike glycoprotein (in our result: M740H, I1130E, and I1132R) may facilitate the membrane fusion by preventing the separation of the chains of the protein (Table 3).

Additionally, we determined the protein stability changes in the domains of the spike glycoprotein/-monoclonal antibody (mAb) complex. The domains of the protein are the N-terminal domain (NTD) and the receptor-binding domain (RBD). The mutations occurred at the interface between the spike residues and the mAbs are L249E (located in NTD) in contact with N9 (PDB: 7E8F) and N11 (PDB: 7E7X) [32] and R457W (located in RDB) in contact with BD-508 (PDB: 7E86), BD-515 (PDB: 7E88) [32], and STE90-C11 (PDB: 7B3O) [33]. The stability change energies upon L249E are −2.23 (with N9) and −1.30 kcal/mol (with N11) (Table 3). Remarkably, L249E had a highly destabilizing effect on N9. This suggests that the L249E mutation may hinder the neutralization by the N9 antibody, allowing the virus to escape the antibody. The stability changes upon the R457W mutation were −0.384 (with BD-508), −0.489 (with BD-515), and −0.235 kcal/mol (with STE90-C11) (Table 3). These findings suggest that both mutations, L249E and R457W, could diminish the interactions between the

Table 3 Protein stability change and local clash score of the mutations on each domain.

Mutation	PDB	Protein stability change	Local clash score	
		(kcal/mol)	Wild type	Mutant
H66K	6XR8	−0.814	33.11	33.58
R102C	6XR8	−1.375	22.49	22.61
L241W	6XR8	−2.175	27.56	46.80
L249E	7E7X	−2.230	17.08	17.13
	7E8F	−1.299	28.65	28.72
D290P	6XR8	−0.222	14.89	28.72
P330M	6XR8	−0.548	17.79	22.80
R457W	6XR8	−0.158	21.05	25.48
	7B3O	−0.235	9.60	11.86
	7E86	−0.384	30.55	32.73
	7E88	−0.489	37.95	40.70
H519E	6XR8	−0.102	18.5	21.74
S735C	6XR8	−0.532	14.57	15.34
	6XRA	−0.217	21.63	25.23
M740H	6XR8	−0.232	12.79	15.11
	6XRA	0.236	20.16	20.16
I742K	6XR8	−2.275	10.52	15.76
	6XRA	−1.802	11.90	27.50
N824E	6XR8	0.185	7.37	7.37
E918N	6XR8	−0.952	14.68	15.86
	6XRA	−1.085	19.98	18.89
Q992C	6XR8	−0.446	13.25	17.87
	6XRA	−0.399	22.35	25.34
I1130E	6XR8	−0.404	16.84	16.88
	6XRA	0.233	31.38	32.34
I1132R	6XR8	−0.779	11.26	13.21
	6XRA	0.184	24.70	25.40

In the Protein Stability Change column, positive values (shown in bold) denote stabilizing effects, whereas the negative values denote destabilizing effects, according to mCSM web server.

spike glycoprotein and the monoclonal antibodies.

Steric clash analysis

The clash scores of mutations indicate the number of atomic collisions among the surrounding areas where the mutations occur. The 15 mutations, on the spike glycoprotein trimer, were analyzed with Missense3D on PDB 6XR8 chain A for the differences in clash scores between wild types and the mutants (Table 3). Notably, the highest clash score difference is 19.24, caused by L241W (wild type = 27.56 and mutant = 46.8). This result suggests that the L241W mutation may slightly alter the conformation of the N-terminal domain, preventing the binding of mAbs to the domain.

In the post-fusion spike glycoprotein, the 7 mutations were analyzed with Missense3D on PDB structure 6XRA chain A (Table 3). Interestingly, the I1130E mutation had the highest clash score (32.34). In addition, the I742K mutation had the highest clash score difference of 15.6 (wild type = 11.9 and mutant = 27.5). These results suggest that both mutations, I742K and I1130E, may cause the helix breakage, preventing the embedding of the post-fusion spike glycoprotein into the host cell membrane [28–30].

The clash scores before and after the L249E mutation are similar in the PDBs: 7E8F (wild type = 28.65 and mutant = 28.72) and 7E7X (wild type = 17.08 and mutant = 17.13), which represent the N-terminal domain bound to N9 and N11, respectively (Table 3). R457W had slightly increased clash scores in the PDBs: 7E86 (wild type = 37.95 and mutant = 40.70), 7E88 (wild type = 30.55 and mutant = 32.73), and 7B3O (wild type = 9.6 and mutant = 11.86), which represent the receptor-binding domain bound with BD-508, BD-515, and STE90-C11, respectively (Table 3). These minor changes in the clash scores of both mutations suggest that they are unlikely to be the key factors that influence the interaction between the spike glycoprotein and the antibodies. For more details of the Missense3D analysis results, see Supplementary File 3.

Broad analysis of Thailand-specific variants and limitations of this study

Recently, reports from WHO and Thailand's Department of Public Health showed that Omicron strains have become dominant throughout the country since late January 2022 [34, 35]. However, the data were all retrieved on 6 January 2022. Therefore, only a few Omicron sequences were detected in our results. In the previous Thailand-specific SARS-CoV-2 variant analysis covering the first outbreak, we reported 6 Thailand-specific novel spike glycoprotein mutations. However, the study covered the data of the outbreak up to early 2021. When the analysis was repeated in this study, we found that the previously reported 6 mutations were no longer Thailand-specific [23]. It is also worth noting that although we were able to analyze over 10,000+ Thai sequences in this study, the incident rate of COVID-19 may vary among different countries. One study found that the incidence rates were significantly positively correlated with Human Development Index (HDI) and Inequality-adjusted Human Development Index (IHDI) from 177 countries analyzed [36]. Thus, keeping the SARS-CoV-2 strains well monitored is paramountly important, and appropriate measures should be undertaken to keep the SARS-CoV-2 outbreak under control.

Due to computational resource limitations, we had to avoid using BLOSUM and PAM scoring matrices, which are very time-consuming, for the pairwise sequence alignment of over 8.2 million closely related pairs. The use of such scoring matrices normally has

a benefit in terms of alignment accuracy — especially when aligning distantly related proteins. However, most of the SARS-CoV-2 spike glycoprotein sequences are highly similar, with over 98% sequence identity. In that case, using simple match/mismatch scores would be sufficient for rapid and accurate sequence alignment to identify mutations. Another limitation is that the structure-based web servers that we used were designed to analyze only single amino acid substitutions. In fact, novel SARS-CoV-2 strains normally contain multiple mutations. The effect of multiple mutations on a single strain could potentially have diverse effects on the functionality of the spike glycoprotein, the infectivity of the virus, and the severity of the disease. To best address multiple mutations, these mutations should be analyzed together since combining the prediction results of individual mutations present on the same viral strain could possibly yield a different result. Thus, new tools that are more comprehensive and capable of assessing multiple mutations are required. Additionally, as the SARS-CoV-2 strains mutate over time, those with more insertion and deletion mutations may predominate the outbreak. Therefore, to best annotate these new strains, more tools or web servers are needed to analyze not only the substitution mutations but also deletion and insertion mutations.

CONCLUSION

We performed large-scale *in silico* analyses to study novel mutations detected in Thai samples by retrieving over 8.3 million sequences from NCBI Virus and GISAID Initiative databases and using Python programming to examine the data. We identified 28 Thailand-specific novel mutations in almost every domain of the spike glycoprotein. Some of these specific mutations are likely to be deleterious to the virus: I742K had the strongest destabilizing effect, and L241W had the highest local clash score. Nevertheless, we found that the mutation L249E and R457W, which resulted in increased clash scores in the spike/antibody complexes, could potentially hinder the neutralization, helping the virus escape the mAbs. However, additional laboratory studies of these mutations are needed to confirm the effects of the mutations on the spike glycoprotein. Finally, our study provides useful information for disease surveillance in terms of virus detection and COVID-19 severity analysis. Furthermore, the results could benefit the vaccine development to protect people from the outbreak of COVID-19.

Appendix A. Supplementary data

Supplementary data associated with this article can be found at http://dx.doi.org/10.2306/scienceasia1513-1874.2022.138.

Supplementary File 1: is available at https://drive.google.com/file/d/1hEyJng3-jN-GSSNhjyfQCIQqYH7FKjJj/view?usp=sharing,

Supplementary File 2: is available at https://drive.google.com/file/d/1GSLLjzjvNpFS7kuuUTikvpQ4EUgdJ0zk/view?usp=sharing, and

Supplementary File 3: is available at https://drive.google.com/file/d/1Gfcg_4r2yTIVQYBGZlSynxQW5ohHM3lj/view?usp=sharing.

Acknowledgements: This work was supported by the Division of Biological Science, Faculty of Science, Prince of Songkhla University.

REFERENCES

1. WHO (2022) Weekly epidemiological update on COVID-19 – 25 January 2022. Available at: www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---25-january-2022.
2. Sirilak S (2020) Thailand's experience in the Covid-19 response. Available at: https://ddc.moph.go.th/viralpneumonia/eng/file/pub_doc/LDoc9.pdf.
3. Duan L, Zheng Q, Zhang H, Niu Y, Lou Y, Wang H (2020) The SARS-CoV-2 spike glycoprotein biosynthesis, structure, function, and antigenicity: Implications for the design of spike-based vaccine immunogens. *Front Immunol* **11**, 576622.
4. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, Schiergens TS, Herrler G, et al (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* **181**, 271–280.
5. Cerutti G, Guo Y, Zhou T, Gorman J, Lee M, Rapp M, Reddem ER, Yu J, et al (2021) Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. *Cell Host Microbe* **29**, 819–833.
6. Suryadevara N, Shrihari S, Gilchuk P, VanBlargan LA, Binshtein E, Zost SJ, Nargi RS, Sutton RE, et al (2021) Neutralizing and protective human monoclonal antibodies recognizing the N-terminal domain of the SARS-CoV-2 spike protein. *Cell* **184**, 2316–2331.
7. Dicken SJ, Murray MJ, Thorne LG, Reuschl A-K, Forrest C, Ganeshalingham M, Muir L, Kalemera MD, et al (2021) Characterisation of B.1.1.7 and Pangolin coronavirus spike provides insights on the evolutionary trajectory of SARS-CoV-2. bioRxiv, 2021.03.22.436468.
8. Shah M, Ahmad B, Choi S, Woo HG (2020) Mutations in the SARS-CoV-2 spike RBD are responsible for stronger ACE2 binding and poor anti-SARS-CoV mAbs cross-neutralization. *Comput Struct Biotechnol J* **18**, 3402–3414.
9. Barton MI, Macgowan S, Kutuzov M, Dushek O, Barton GJ, Anton van der Merwe P (2021) Effects of common mutations in the SARS-CoV-2 spike RBD and its ligand, the human ACE2 receptor on binding affinity and kinetics. *eLife* **10**, e70658.
10. Greaney AJ, Starr TN, Barnes CO, Weisblum Y, Schmidt F, Caskey M, Gaebler C, Cho A, et al (2021) Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat Commun* **12**, 4196.
11. Lupala CS, Ye Y, Chen H, Su X-D, Liu H (2021) Mutations on RBD of SARS-CoV-2 Omicron variant result in stronger binding to human ACE2 receptor. bioRxiv, 2021.12.10.472102.
12. Ou J, Zhou Z, Dai R, Zhang J, Zhao S, Wu X, Lan W, Ren Y, et al (2021) V367F Mutation in SARS-CoV-2 spike RBD

emerging during the early transmission phase enhances viral infectivity through increased human ACE2 receptor binding affinity. *J Virol* **95**, 617–638.
13. Omotuyi IO, Nash O, Ajiboye OB, Iwegbulam CG, Oyinloye BE, Oyedeji OA, Kashim ZA, Okaiyeto K (2020) Atomistic simulation reveals structural mechanisms underlying D614G spike glycoprotein-enhanced fitness in SARS-COV-2. *J Comput Chem* **41**, 2158–2161.
14. Ashwaq O, Manickavasagam P, Haque SM (2021) V483A: An emerging mutation hotspot of SARS-CoV-2. *Future Virol* **16**, 419–429.
15. Jangra S, Ye C, Rathnasinghe R, Stadlbauer D, Alshammary H, Amoako AA, Awawda MH, Beach KF, et al (2021) SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. *Lancet Microbe* **2**, e283–284.
16. Nonaka CKV, Franco MM, Gräf T, De Lorenzo Barcia CA, De Ávila Mendonça RN, De Sousa KAF, Costa Neiva LM, Fosenca V, et al (2021) Genomic evidence of SARS-CoV-2 reinfection involving E484K spike mutation, Brazil. *Emerg Infect Dis* **27**, 1522.
17. Ortega JT, Pujol FH, Jastrzebska B, Rangel HR (2021) Mutations in the SARS-CoV-2 spike protein modulate the virus affinity to the human ACE2 receptor, an *in silico* analysis. *EXCLI J* **20**, 585.
18. Weissman D, Alameh MG, de Silva T, Collini P, Hornsby H, Brown R, LaBranche CC, Edwards RJ, et al (2021) D614G Spike mutation increases SARS CoV-2 susceptibility to neutralization. *Cell Host Microbe* **29**, 23–31.
19. Okada P, Buathong R, Phuygun S, Thanadachakul T, Parnmen S, Wongboot W, Waicharoen S, Wacharapluesadee S, et al (2020) Early transmission patterns of coronavirus disease 2019 (COVID-19) in travellers from Wuhan to Thailand, January 2020. *Euro Surveill* **25**, 2000097.
20. Puenpa J, Suwannakarn K, Chansaenroj J, Nilyanimit P, Yorsaeng R, Auphimai C, Kitphati R, Mungaomklang A, et al (2020) Molecular epidemiology of the first wave of severe acute respiratory syndrome coronavirus 2 infection in Thailand in 2020. *Sci Rep* **10**, 16602.
21. Buathong R, Chaifoo W, Iamsirithaworn S, Wacharapluesadee S, Joyjinda Y, Rodpan A, Ampoot W, Putcharoen O, et al (2021) Multiple clades of SARS-CoV-2 were introduced to Thailand during the first quarter of 2020. *Microbiol Immunol* **65**, 405409.
22. Joonlasak K, Batty EM, Kochakarn T, Panthan B, KÃijmpornsin K, Jiaranai P, Wangwiwatsin A, Huang A, et al (2021) Genomic surveillance of SARS-CoV-2 in Thailand reveals mixed imported populations, a local lineage expansion and a virus with truncated ORF7a. *Virus Res* **292**, 198233.
23. Ittisoponpisan S, Yahangkiakan S, Sternberg MJE, David A (2022) The SARS-CoV-2 infection in Thailand: analysis of spike variants complemented by protein structure insights. bioRxiv, 2022.01.01.474713.
24. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, et al (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423.
25. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE (2019) Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated? *J Mol Biol* **431**, 2197–2212.
26. Pires DEV, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **30**, 335–342.
27. WHO (2022) VOC profiles of spike amino acid changes. Available at: www.who.int/docs/default-source/coronaviruse/s.pdf?sfvrsn=990a05c2_14.
28. Tamm LK, Han X (2000) Viral fusion peptides: A tool set to disrupt and connect biological membranes. *Biosci Rep* **20**, 501–518.
29. Earp LJ, Delos SE, Park HE, White JM (2004) The many mechanisms of viral membrane fusion proteins. *Curr Top Microbiol Immunol* **285**, 25–66.
30. White JM, Delos SE, Brecher M, Schornberg K (2008) Structures and mechanisms of viral membrane fusion proteins: Multiple variations on a common theme. *Crit Rev Biochem Mol Biol* **43**, 189.
31. Cai Y, Zhang J, Xiao T, Peng H, Sterling SM, Walsh RM, Rawson S, Rits-Volloch S, Chen B (2020) Distinct conformational states of SARS-CoV-2 spike protein. *Science* **369**, 1586–1592.
32. Cao Y, Yisimayi A, Bai Y, Huang W, Li X, Zhang Z, Yuan T, An R, et al (2021) Humoral immune response to circulating SARS-CoV-2 variants elicited by inactivated and RBD-subunit vaccines. *Cell Res* **31**, 732–741.
33. Bertoglio F, Fühner V, Ruschig M, Heine PA, Abassi L, Klünemann T, Rand U, Meier D, et al (2021) A SARS-CoV-2 neutralizing antibody selected from COVID-19 patients binds to the ACE2-RBD interface and is tolerant to most known RBD mutations. *Cell Rep* **36**, 109433.
34. Department of Disease Control (2022) Thailand weekly situation update on 7 January 2022. Available at: https://ddc.moph.go.th/viralpneumonia/eng/file/situation/situation-no723-070165.pdf.
35. WHO Thailand (2022) COVID-19 Situation, Thailand 26 January 2022. https://cdn.who.int/media/docs/default-source/searo/thailand/2022_01_26_tha-sitrep-220-covid-19.pdf?sfvrsn=bf4f76d6_5.
36. Chunbao M, Xiaoting M, Tingyu M, Jiansheng C, Xia X, Chuntao N, Dechan T, Shuzhen L, et al (2021) Association between COVID-19 incidence and outcome and national development levels: An ecologic analysis. *ScienceAsia* **47**, 211–219.

Appendix A. Supplementary data

Table S1 Distribution of Thai mutations in each domain.

Domain	Residue	No. of mutation (Thailand-specific)	Density (mutation/residue)
signal peptide	1–13	18 (0)	1.38
N-terminal domain	14–306	202 (11)	0.69
receptor-binding domain	319–540	371 (3)	1.67
subdomain 1 and 2	541–686	116 (3)	0.79
usion peptide	789–808	38 (0)	1.90
heptad repeat 1	912–983	48 (1)	0.67
central helix	984–1034	18 (1)	0.36
connector domain	1079–1140	48 (2)	0.77
heptad repeat 2	1163–1212	63 (1)	1.26
transmembrane	1213–1236	35 (5)	1.46
cytoplasmic tail	1237–1273	44 (1)	1.19
Total	1–1273	1294 (28)	1.02

Numbers in the parentheses represent the number of Thailand-specific mutations found in each domain. Density is calculated from the number of mutations divided by the number of residues, in the domain.