RESEARCH ARTICLE

Mutation and Phylogenetic Analysis of Spike Glycoprotein of Indonesian Isolates of *Severe-Acute-Respiratory-Syndrome-Coronavirus-2* (SARS-CoV-2)

Shabarni Gaffar, Syifa Al Fauziah Rahmani, Ari Hardianto
Chemistry Department, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Indonesia

Abstract

Coronavirus disease-2019 (COVID-19) is an infectious acute respiratory disease caused by SARS-CoV-2. The protein that plays a role in the entry of SARS-CoV-2 into human cells is the surface protein, or the Spike, which is thought to be the effective vaccine target to prevent SARS-CoV-2 infection. Until December 2020, Indonesia has reported 106 SARS-CoV-2 genome sequences identified from COVID-19 positive patients. The purpose of this study was to analyze the phylogenetic relationship of the Spike protein of the Indonesian isolates of SARS-CoV-2 Indonesian, as well as the virus mutations and their effects on changes in the amino acid. The 106 Indonesian SARS-CoV-2 genomes were downloaded from GISAID and the Spike nucleotide and amino acid sequences were analyzed by multiple sequence alignment (MSA) and mutation analysis using the ClustalW method. Phylogenetic trees were created using the Neighbor-Joining method in MEGA-X software. The results showed that 30 of the 106 Indonesian isolate SARS-CoV-2 Spike were 100% identical to the Wuhan-Hu-1, while the remaining 76 had experienced mutations at 1-4 sites. There were 43-point mutations in the Spike gene, 27 of which led to amino acid changes and four had not been reported in other countries. The global mutation D614G was found in 60 Indonesian isolates , of which West Java was the province with the most reports. The phylogenetic of Spike showed that the Indonesian samples have been divided into several branches that are far from Wuhan-Hu-1. This study indicates the possibility of differences in the protein structure of Indonesian isolate SARS-CoV-2 Spike that need to be further studied to manufacture a vaccine against the Indonesian strain of SARS-CoV-2.

Keywords: Mutation, phylogenetic analysis, SARS-CoV-2, spike

Analisis Mutasi dan Filogenetik Glikoprotein *Spike Severe-Acute-Respiratory-Syndrome-Coronavirus-2* (SARS-CoV-2) Isolat Indonesia

Abstrak

Penyakit Coronavirus-2019 (COVID-19) adalah penyakit saluran pernapasan akut menular yang disebabkan oleh SARS-CoV-2. Protein yang berperan pada masuknya SARS-CoV-2 ke dalam sel manusia adalah protein permukaan Spike yang dianggap sebagai target vaksin yang efektif untuk mencegah infeksi SARS-CoV-2. Sampai Desember 2020, Indonesia telah melaporkan 106 urutan genom SARS-CoV-2 dari pasien positif COVID-19. Tujuan penelitian ini adalah menganalisis hubungan filogenetik dari *Spike* SARS-CoV-2 Indonesia, serta menganalisis mutasi dan pengaruhnya terhadap perubahan asam amino. Sebanyak 106 genom SARS-CoV-2 Indonesia diunduh dari GISAID, urutan nukleotida dan asam amino *spike* dianalisis dengan *Multiple-Sequence-Alignment*/MSA dan analisis mutasi menggunakan metode Clustal W. Pohon filogenetik dibuat menggunakan metode *Neighbor-Joining* dalam perangkat lunak MEGA-X. Hasil penelitian menunjukkan 30 dari 106 *spike* SARS-CoV-2 Indonesia 100% identik dengan Wuhan-Hu-1, sedangkan 76 sisanya mengalami mutasi di 1-4 lokasi. Ada 43 titik mutasi pada gen *Spike*, 27 di antaranya menyebabkan perubahan asam amino dan empat di antaranya belum dilaporkan di negara lain. Mutasi global D614G ditemukan pada 60 orang Indonesia, Jawa Barat merupakan provinsi dengan laporan terbanyak. Pohon filogenetik *spike* menunjukkan sampel Indonesia telah terbagi atas beberapa cabang yang jauh dari Wuhan-Hu-1. Simpulan, penelitian ini menunjukkan kemungkinan perbedaan struktur protein SARS-CoV-2 Indonesia yang perlu dikaji lebih lanjut guna pembuatan vaksin melawan SARS-CoV-2 *strain* Indonesia.

Kata kunci: Analisis filogenetik, mutasi, SARS-CoV-2, *spike*

Corresponding Author: Shabarni Gaffar, Chemistry Department, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Jalan Raya Bandung-Sumedang KM 21, West Java, Indonesia, Email: shabarni.gaffar@unpad.ac.id

Introduction

Coronavirus disease 2019 (COVID-19) pandemic caused by Severe Acute Respiratory Syndrome-Coronavirus 2 (SARS-CoV-2) has recently become a global health threat. Coronaviruses are a large family of viruses that infect the upper respiratory tract in humans and other mammals. Three coronavirus pandemics have been reported in the 21st century, i.e., SARS in 2002, MERS in 2012, and COVID-19 in 2019.

The current coronavirus pandemic is caused by a new type of coronavirus found in December 2019 that causes an outbreak of the respiratory inflammatory disease in the city of Wuhan, which is the capital city of China's Hubei province.[1,2] This coronavirus is then named Severe Acute Respiratory Syndrome-Coronavirus 2 (SARS-CoV-2), owing to its RNA genome that is 79.5% identical to the SARS-CoV.[1,3] Both viruses belong to the B clade of the genus *Beta coronavirus*. COVID-19 was initially associated with the Huanan seafood and animal market in Wuhan; however, this is still debatable. The fact that this coronavirus has gained the ability for human-to-human transmission making it a highly infectious zoonotic virus that grows exponentially, leading to the emergence of millions of cases globally and creating a worldwide pandemic.[1] By 3 April 2021, the number of COVID-19 cases has reached more than 130 millions worldwide with 2.8 million deaths. In Indonesia, more than 1.5 million people have been infected by this virus and 41,151 of them died because of this disease.

The SARS-CoV-2 Wuhan-Hu-1 (NC_045512.2) virus is known to have 96.2% homology with SARS-CoV RaTG13 (MN996532), which is a virus that attacks the respiratory tract in bats. This has led to the assumption that SARS-CoV-2 is a virus that undergoes zoonosis from bats to humans after several mutations in the genome, causing this virus to infect and be transmitted between humans.[1]

The SARS-CoV-2 virus is a positive-strand RNA virus with a genome size of ~29,900 nucleotides.[2] The genomic RNA (gRNA) of coronavirus contains 5'cap and 3'poli A and can act as an mRNA that can be directly translated into poly-proteins. At the 5 'end, there are two open reading frames (ORF) that encode very long replicase polyproteins, ORF1a and ORF1b, which cover 20 kb or two-thirds of the genome. Each of these ORFs are translated into large polyproteins, namely 1a and 1ab, where the pp1ab is expressed through the translation shift coded by the 3'ORF1a end. Polyproteins are co-translationally digested into 16 proteins, which generally have enzymatic activities including two proteases that modify RNAs, such as polymerase and helicase. After ORF1b, there is ORF2-ORF10 which encodes structural proteins consisting of Spike (S), Membrane protein (M), an Envelope protein (E), and Nucleo-capsid protein (N). Several nonstructural genes that encode small accessory proteins that do not play a role in replication but positioned among structural genes that play a role in interactions with the host are also observed.[4]

The phylogenetic and bioinformatics analysis of SARS-CoV-2 against other virus species has been performed by several researchers. One of these researchers is Lu et al.[2] who demonstrated that of the 5 representative coronavirus sub-genus, SARS-CoV-2 belongs to the *Sarbe covirus* subgenus. This subgenus can be further classified into 3 clades. Clade 1 consists of SARS-related CoV that is found in Kenya and Bulgaria; clade 2 consists of 10 sequences of SARS-CoV-2 from Wuhan and SARS that are derived from bats from Zhoushan, China; and clade 3 that consists of Bat SARS-like and several SARS strains that attack humans. SARS-CoV-2 has a very distant relationship with SARS-CoV. The results of this phylogenetic analysis prove that SARS-CoV-2 is a new type of *Betacoronavirus* under the B lineage of the subgenus *Sarbecovirus*.[2]

Jaimes et al.[5] have conducted phylogenetic and bioinformatic analysis, as well as structural modeling of the SARS-CoV-2 Spike protein based on its homology with SARS-CoV. Spike protein plays a very important role in viral infection as it mediates the attachment, fusion, and entry of the virus into the target cells. The Receptor Binding Domain (RBD) on the Spike protein plays a role in recognizing and interacting with the Angiotensin-Converting Enzyme 2 (ACE-2) receptor on the target cell surface. This interaction occurs between several amino acids on RBD, which is referred to as the Receptor Binding Motif (RBM) with ACE-2 receptors. Furthermore, fusion occurs to the host membrane, involving the fusion of peptides (FP) and HR1 and 2 (HR =Heptad region). The process used by the virus to enter the host requires the role of a trans-membrane host protease, namely TMPRSS2 (Trans-membrane serine protease), which will bypass the ACE-2 receptor and activate Spike, enabling the SARS-CoV-2 to enter the host cell through endocytosis.[6]

Spike is a large glycoprotein with a size of about 180 kDa and consists of two domains: the S1 domain (receptor binding domain) and

the S2 domain (fusion domain). The N-terminal and C-terminal regions of the S protein form a distinctive crown-like trimer when observed under an electron microscope.[7] The S1 domain contains the RBD (amino acids333-527), in which the RBM is the amino acids438-506, that interacts directly with the ACE-2 receptor. Furthermore, there are Fusion peptide (FP), Heptad Region (HR) 1 and 2, and Trans-membrane TM area required for fusion to the membrane and for introducing the virus into the target cells.[8]

Spike, RBD, and RBM are the targets for vaccines against SARS-CoV-2. Several studies have reported the expression of whole Spike, S1 domain, or RBD in heterologous protein expression systems.[9,10] Viral genomes, including that of the coronavirus, are reported to be rapidly evolving due to the accumulation of mutations caused by replication errors in the host cells. Although SARS-CoV-2 is known to have a proofreading mechanism, several different strains have been found, which indicate the occurrence of mutations.[3]

In this study, an analysis of the phylogenetic relationship of Spike protein from the Indonesian SARS-CoV-2 strain was carried out based on the comparison of the nucleotide and amino acids sequence with the Wuhan-Hu-1 strain and strains from other countries. Analysis on the relationship between mutations and changes in the amino acid sequence of the Spike protein of the SARS-CoV-2 strain was performed. The results of this study are expected to provide information about phylogenetic relationships and mutations in the Spike protein of SARS-CoV-2 to be used as a reference for vaccine candidates in Indonesia.

Methods

This was a bioinformatic study without any wet lab experiment involving human participants or animals. It employed an ASUS X200MA-KX636D laptop with a turbo boost speed specification of ~2.58 GHz and 2GB RAM. The gene encoded SARS-CoV-2 Spike glycoprotein (3822 bp) was obtained from the GenBank (National Center for Biotechnology Information, USA) and the Global Initiative on Sharing All Influenza Data (GISAID) EpiCoV database (Germany). The software used in this study was Lasergene version 7.1.0 (DNASTAR, USA), Snapgene Viewer 5.1.5 (USA), and MEGA-X (Molecular Evolutionary Genetic Analyzes). Data were collected from March to December 2020.

Indonesian and nine global SARS-CoV-2 genome sequences were retrieved from the GISAID database. The sequences (supplemental information file) were annotated according to the sample ID on GISAID with additional serial numbers to facilitate analysis. All genome sequences were analyzed by Seqman on Lasergene to determine the position of the *S* gene by comparing it with the reference sequence (GenBank: NC_045512.2). Then, the *S* gene sequence from each sample was cut according to the length of the reference sequence, which removes any truncated sequences. The nucleotide sequences of the *S* genes were translated into amino acid sequences using the EditSeq on Lasergene based on standard genetic code. All nucleotide and amino acid sequence files were then converted into FASTA format.

The SARS-CoV-2 Spike gene reference sequence used was the Wuhan-Hu-1 available in the GenBank (NC_045512.2) with a length of 29,903 bases. For the Spike, the Spike glycoprotein reference sequence (GenBank: YP_009724390.1) with a length of 1273 amino acids was used as reference. In this study, 106 sequences of the SARS-CoV-2 genome from Indonesia were included and 9 genome sequences from other countries were randomly selected and genome sequences from India, Italy, Malaysia, Kazakhstan, Saudi Arabia, China, Singapore, Taiwan, and the United States were used in the study (The ID of samples are presented in the supplement file).

Both nucleotide and amino acid sequences of Spike were analyzed using the MSA with the default Clustal W method in MEGA-X using Wuhan-Hu-1 as the reference. Gap opening and extension penalties were set to 10.0 and 0.20, respectively. The negative matrix was off and delay divergent cutoff was 30%. The resulting alignment was stored in the FASTA format for further analysis using the Snap gene program to interpret the mutations and changes in amino acids in the Spike genes/protein.

MSA with the Clustal W method was used to analyze the nucleotide and amino acid sequences of Spike in FASTA format in this study.[11] This program aligns the nucleotide/amino acid sequences so that mutations that occur in the Indonesian SARS-CoV-2 Spike protein sequence can be analyzed and compared to the Wuhan-Hu-1 sequence. Clustal W uses a progressive alignment method where the most similar sequences are aligned first, followed by a more different sequence to obtain a global alignment. The algorithm begins by calculating the distance
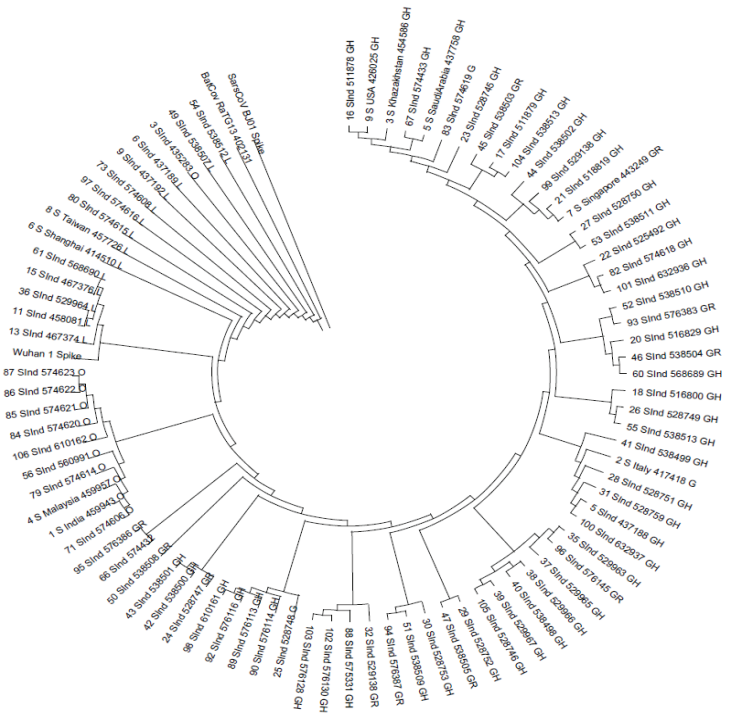
Tabel Point Mutation in 76 Indonesian Samples

No	Wuhan-1	1 (14) C	2 (64) A	3 (65) C	4 (76) C	5 (90) T	6 (180) C	7 (186) T	8 (200) C	9 (208) G	10 (225) T	11 (227) C	12 (247) G	13 (284) C	14 (347) C	15 (414) T	16 (438) C	17 (474) A	18 (553) A	19 (638) T	20 (641) G	21 (772) T	22 (776) C	23 (826) C	24 (882) C	25 (888) C	26 (1054) G	27 (1430) G	28 (1715) C	29 (1743) A	30 (1841) A	31 (1864) G	32 (1960) G	33 (1971) C	34 (2015) C	35 (2031) G	36 (2367) C	37 (2432) A	38 (2464) C	39 (2556) A	40 (2631) G	41 (3371) G	42 (3444) C	43 (3761) G	JML
1	3_435283											T																																	1
2	5_437188														G																G							T							3
3	6_437189																C																T												2
4	9_437192																												T												T				2
5	11_458081																										T																		1
6	13_467374																																				T								1
7	15_467376																																										T		1
8	16_511878																									T					G														2
9	17_511879																		T												G														2
10	18_516800																														G														1
11	20_516829																														G														1
12	21_518819		C																								T				G														3
13	22_525492																														G														1
14	23_528745																														G														1
15	24_528747																														G														1
16	25_528748																														G														1
17	26_528749																	T													G														2
18	27_528750																														G														1
19	28_528751																														G						T								2
20	29_528752																														G														1
21	30_528753																														G														1
22	31_528759											T																			G						T								3
23	32_529138																														G														1
24	35_529963			T					T																						G														3
25	36_529964																										T																		1
26	37_529965																														G														1
27	38_529966																														G														1
28	39_529967																														G														1
29	40_538498														T																G														2
30	41_538499																														G														1
31	42_538500																					C									G														2
32	43_538501																					C									G														2
33	44_538502																														G														1
34	45_538503																														G														1
35	46_538504																														G			C											2
36	47_538505									C																					G														2
37	49_538507																														G									G					1
38	50_538508																														G														1
39	51_538509			T																											G														2
40	52_538510																														G														1
41	53_538511																														G														1
42	54_538512																					C										G								G					3
43	55_538513																														G														1
44	56_560991																																					T							1
45	60_568689																														G														1
46	61_568690																																										T		1
47	66_574432																						C								G														2
48	67_574433																														G														1
49	71_574606																																			T					T				2
50	73_574608																																T												1
51	79_574614																																			T									1
52	80_574615																								C																				1
53	82_574618																														G														1
54	83_574619																														G														1
55	84_574620																														G							T							2
56	85_574621																														G							T							2
57	86_574622																														G							T							2
58	87_574623																														G							T							2
59	88_575331																														G														1
60	89_576113				T																	C									G														3
61	90_576114																														G														1
62	92_576116				T																	C									G														3
63	93_576383																														G										T				2
64	94_576387				T																										G														2
65	95_576386																										T				G														2
66	96_576145																														G														1
67	97_574616									T																																			1
68	98_610161				T																	C									G														3
69	99_529138																														G														1
70	100_632937											C																			G							T							3
71	101_632936																														G								T						2
72	102_576130	T																													G														2
73	103_576128	T																													G														2
74	104_538513																														G														1
75	105_528746																														G												T		2
76	106_610162				C		C																			C					G														4
	Jumlah	2	1	1	2	1	3	1	1	1	1	1	1	1	1	1	1	1	3	1	1	1	1	1	2	3	2	1	1	1	60	1	1	1	3	7	1	1	2	1	1	1	2	123
	Asam Amino Wuhan-1	L	T	T	P				A	V		T	V	T	S			R	N	V	R		T				A	S	T		D	V	E		A	Q		K	L			G		C	27
	Hasil mutasi	F	P	I	S				V	F		I	L	I	C			S	Y	A	L		I				S	I	I		G	F	Q		V	H		I	F			V		F	
	Posisi asam amino	5	22	22	26				67	70		76	83	95	116			158	185	213	214		259				352	477	572		614	622	654		672	677		811	822			1124		1254	
	Jenis mutasi	C	NC	NC	SC				SC	SC		NC	C	NC	SC			NC	NC	SC	NC		NC				C	NC	NC		SC	SC	C		SC	C		NC	C			NC		NC	
		U	B	U	U				U	U		U	U	U	B			U	B	U	U		U				U	U	U		U	U	U		U	U		B	U			U		U	

Table is constructed based on MSA of the nucleotide and amino acid sequences of the Spike protein. From 76 Indonesian samples that were not 100% identical to the Wuhan-1 sequence, 43 mutation points were found, with 27 mutations causing amino acid changes. The type of amino acid change varied between C (conservative), SC (semi-conservative), and NC (non-conservative). U=was also found in other countries' samples; B=was only seen in the Indonesian sample.

Figure 1 Phylogenetic tree of 76 Indonesia's Spike Genes Compared to Nine Overseas Samples, Wuhan-Hu-1, BatCoVRatG, and SARS-CoV

The Phylogenetic Tree was constructed based on the Spike Gene's 3822 bases using the Neighbor-Joining Method in MEGA X Software
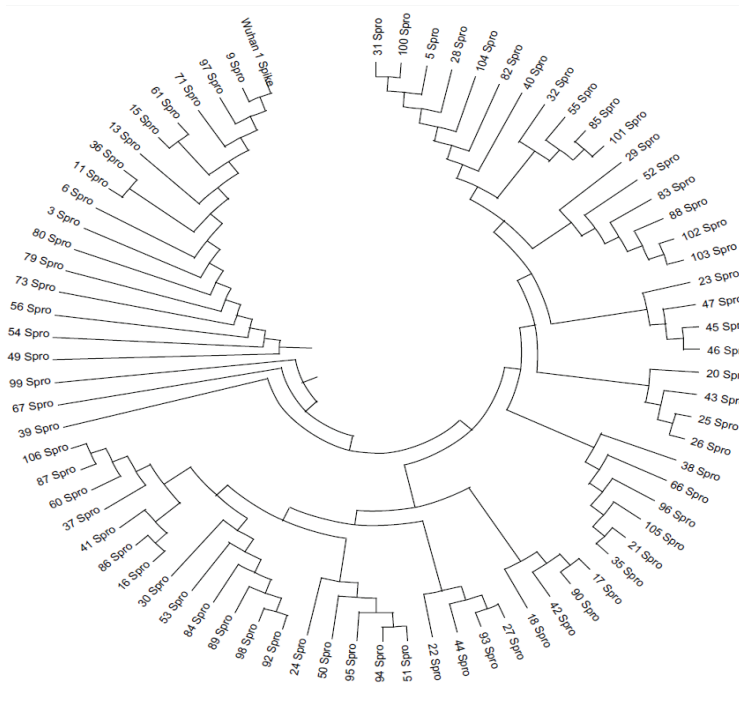

Figure 2 Phylogenetic Trees of 1273 Amino Acids of S Protein from 76 Indonesian Samples Compared to the Wuhan-1 Sequence
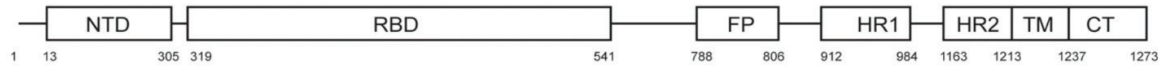
The tree was created using MEGA-X software

Figure 3 SARS-CoV-2 Spike Protein Structure
Protein Spike consists of S1 domain containing NTD=N terminal domain, RBD=Receptor Binding domain, SD=Sub Domain 1 and 2; domain S2 contains FP=fusion peptide, HR=Heptad Repeat 1 and 2; TM=transmembrane Region and IC=Intracellular domain.[29]

matrix between each pair of sequences based on the value in pairwise sequence alignment, after which the neighbor-joining method algorithm with midpoint rooting is used to build a guide tree used to make the global alignment.[11]

The MSA result was then used to construct the phylogenetic tree to determine the relationship between the Indonesian samples, reference sequences, and samples from other countries. A Neighbor-Joining method in the MEGA-X was selected to construct the phylogenetic tree since the data was free from any truncated sequence, which would be vulnerable for such a method.[12,13]

Results

The MSA analysis of Indonesian sequences, when compared to the Wuhan-Hu-1 sequence, demonstrated that 30 of 106 Indonesian sequences were 100% identical to Wuhan-Hu-1.

The comparison between the remaining 76 mutated Indonesian sequences and Wuhan-Hu-1 is shown in Table 1. Based on the results, it was revealed that there were 43 points mutations in the Indonesian sequences, with each sequence mutated at 1–4 points. These indicate a high mutation rate with varying positions in the Spike gene.

Based on the MSA of the Spike amino acid sequence of 76 mutated samples, 27 non-synonymous mutations were identified to cause amino acid changes. In addition, there were 16 mutations that were synonymous mutations. From the perspective of the amino acid changes, it was demonstrated that 6 were conservative mutations (size and hydropathy conserved), 8 were semi-conservative mutations (size or hydropathy conserved) and 13 were non-conservative mutations (size and hydropathy non-conserved). Non-conservative mutations are mutations that produce changes in amino
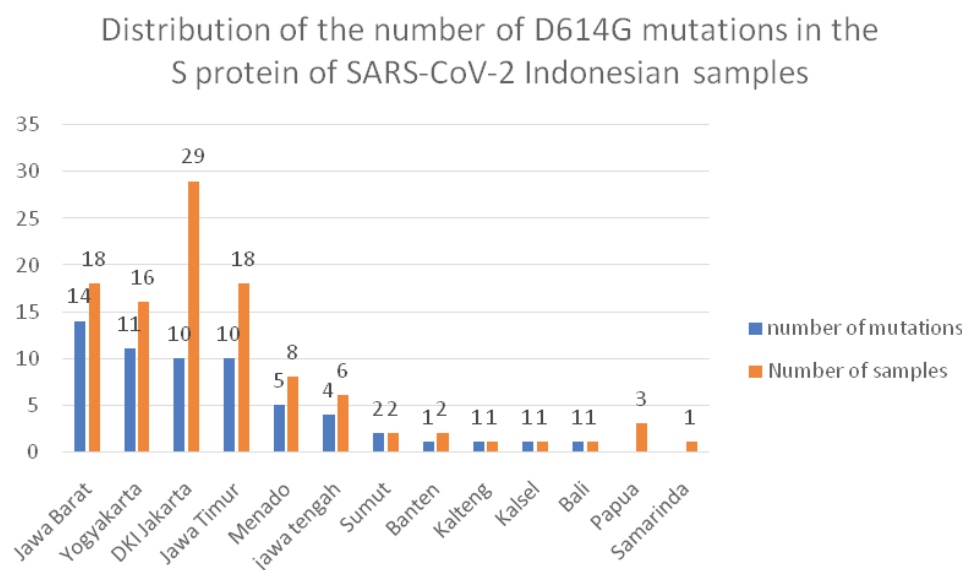
Figure 4 D614G Mutation Distribution of in Spike Protein from 106 Indonesian Samples

acids of different sizes and properties. These mutations are thought to result in changes in the three-dimensional structure of the protein. When compared to the data in the GISAID database, the following four mutations were identified as only occuring in the Indonesian samples: T22P, S116C, N185Y, and K811I. Other mutations identified in the remaining Indonesian samples were also seen in samples from other countries.

Thirty Indonesian samples were100% identical to the Wuhan-Hu-1. These sequences were not included in the phylogenic analysis, as this analysis was only performed on samples that was different from Wuhan-Hu-1 The S gene from 30 Indonesian samples that were identical to Wuhan-Hu-1 based on the MSA results are located in the same branch as Wuhan-Hu-1 and different from BatCoV-RatG13 with a distance of 3.6 x 100 nucleotide substitutions (data not shown). Most of the samples belonged to clade L and only two samples belonged to clade O (sample no. 63 and 81) and one belonged to GH clade (sample 91). Based on the information in the GISAID database, most of these samples were sent to GISAID in May-October 2020.

The phylogenetic tree of 76 samples that are different from Wuhan-Hu-1 is shown in Figure 1. Based on this tree, the Indonesian samples were divided into several branches of small groups based on the mutation that occurred. This group may be large if the number of samples was added. Furthermore, it seems that the similarity of some samples to samples from abroad might link to the patient's travel history; however, it is also possible that it happened by chance because this virus mutates rapidly. There are even some sample sequences that are presented earlier than the Wuhan-Hu-1 reference sequence. This may happen because the virus mutates repeatedly. The phylogenetic tree uses a specific algorithm to be able to count back mutation events. The phylogenetic tree also reflects the the fact that rapid mutation of changes do not occur with time; thus, the phylogenetic tree of the SARS-CoV-2 virus cannot be used to describe the common ancestor of the mutations. The SARS-CoV-2 virus mutates less quickly than other viruses because this virus has a proofreading mechanism to correct errors[3] but because the virus spreads very rapidly, there is not enough time to correct the reading errors, which leads to mutations.

The phylogenetic tree based on the amino acid sequence of the Spike protein is shown in Figure 2. It can be seen that the relationship of S gene is slightly different from the amino acid sequence. This is because not all mutations cause amino acid to change and in some mutations, even when there is a change in the amino acid, the properties are still the same. Phylogenetic analysis based on the amino acid sequence of the Indonesian Spike protein shows that the Indonesian Spike strain is divided into several major branches based on the mutations that change amino acids.

Discussion

The SARS-CoV-2 Spike protein monomer consists of NTD, N-terminal domain; RBD, Receptor binding domain; SD1, subdomain 1; SD2, subdomain 2; FP, fusion peptide; HR1, heptad repeat 1; HR2, heptad repeat 2; TM, transmembrane region; IC, intracellular domain (Figure 3). The position of RBD is at the 319-541 residues. This RBD contains the RBM (Receptor Binding Motive), which is the key to binding Spike with ACE-2, with N439, L455, F486, Q493, Q498, and N501 as the amino acids that play a role in the binding.[14]

Similar studies have been reported from different areas.[15] One of these study uses full-length genome sequences to yield a phylogenetic analysis of SARS-CoV-2 circulating in Yogyakarta and Central Java, Indonesia. This is different from this present study as this study focuses on the phylogeny of amino acid sequences translated from, spike protein genome of SARS-CoV-2 from various regions in Indonesia, not only from Yogyakarta and Central Java

The results of this study demonstrated that 15 mutations that cause amino acid changes occur in the NTD domain, but each of these mutations only occurs in 1-2 samples, with two mutations in RBD causes amino acid changes (A352S and S477I). The A352S mutation is a conservative mutation that changes the hydrophobic amino acid, Alanine, to a polar Serine while the S477I mutation is non-conservative where the amino acid Serine (polar) changes to Isoleucine (hydrophobic). Yuan et al.[16] reported that the A352S mutation is located in the opposite position to the binding site of RBD in ACE-2 receptor and in the same position as the epitope binding of the CR3022 antibody (neutralizing antibody) isolated from convalescent SARS-CoV patients. However, the A352S mutation in SARS-CoV-2 has not been previously reported, so its effect on binding to ACE-2 receptors and antibodies is not known. Meanwhile, the S477I mutation has been reported to strengthen the bond between Spike and the ACE receptor.[17]

Twelve other mutations have been identified to cause amino acid changes in post-RBD domains. These mutations have also generally been reported in other countries. The analysis in this study indicates that the D614G mutation occurs in 60 Indonesian samples (Table 1). This mutation is semi-conservative in that there is a change in the amino acid Aspartic acid (negatively charged) to glycine (neutral).

Several studies have reported that the D614G mutation in the SARS-CoV-2 Spike protein is a global mutation that occurs in many countries.[18] This mutation first appears in Europe, then spreads to the USA, Canada, and Australia, and worldwide. The MSA results of the SARS-CoV-2 Spike protein in some samples from different countries, when compared against the Spike reference sequence (YP_009724390.1), present a D614G mutation of the SARS-CoV-2 Spike protein in patients: USA 426025, Singapore 443249, Saudi Arabia 437758, Kazakhstan 454586, and Italy 417418. Based on the analysis of the SARS-CoV-2 genome in Europe from late December 2019 to mid-March 2020, the D614G mutation in S protein is the most dominant mutation in Europe, estimated to appear in mid to late January 2020.[19]

The D614G mutation is reported to change the conformation of S protein to be more dynamic, which affects the ability of Spike to be infected, facilitates fusion, and increases viral titer in the host. D614 is located on the surface of the Spike protomer that interacts with neighboring protomers. Hydrogen bonds are formed between the D614 side chains of protomer-1 and T859 of protomer-2. The change in G614 removes this hydrogen bond, thereby increasing the flexibility and changing the interactions between protomers. This substitution can also modulate the glycosylation of N616, leading to an increase in the dynamics of the fusion peptide and probably also causing other effects. High viral titer is reported due to G614, causing virus entry no longer requires TMPRSS2 activities.[18]

D614G mutation is also known to be the cause of individuals to be susceptible to infection for the second time.[18] One of the reasons for the increase in the global spread of the SARS-CoV-2 virus is the mutation at position D614G.[21] The increase in the spread of the SARS-CoV-2 virus due to the D614G mutation is also supported by other studies.[22,23]

The D614G mutation is thought to reduce the interaction between S1 and S2 units and facilitate the release of S1 from S2 bound to the viral membrane. The alternative structure-based hypothesis suggested that this mutation could affect the binding of RBD-ACE-2, although not directly because this site is not proximal to the binding interface. The RBD needs to be in an "up" position to engage with the ACE-2 receptor, and this site may allosterically alter the "up" and "down" transitions of RBD.[15] The slightest change allows a conformational change of the dynamic S protein trimer. Therefore, experimental studies and more detailed modeling are needed to explain the effects of these mutations on the RBD transition.

Figure 4 depicts the distribution of D614G mutations in Indonesian patients. It can be seen from the 106 genome samples sent to GISAID that although DKI Jakarta province is the province that sends the highest number of of the SARS-CoV-2 genome sequences compared to other provinces in Indonesia, most cases of mutations are actually found in West Java Province, where 14 of the 18 samples sent show have undergone mutations. The D614G mutation is reported to make the virus more abundant in the upper respiratory tract, i.e., nose and mouth, making it easier for the virus to spread. However, the clinical symptoms of the D614G mutant are weaker than the wild type SARS-CoV-2 virus.[24] A study by Ansori et al.[25] in May 2020 that analyzed nine Indonesian samples in the GISAID database only found one D614G mutation in samples from Surabaya with ID: 437188. In November 2020, however, 60 D614G mutants are detected in 106 Indonesian samples, showing how fast this SARS-CoV-2 mutants are spreading.

The phylogenetic analysis of S protein using the neighbor-joining method showed a variation in the distribution of the SARS-CoV-2 virus clade in Indonesia, and also a rapid mutation. Furthermore, it is also known that Indonesia's SARS-CoV-2 may not have originated directly from China, but originated from different countries as reflected in the travel history of patients to various countries. There is also a possibility of local transmission that leads to mutations of the virus.

The results of phylogenetic analysis that is based on the S protein of SARS-CoV-2 Indonesian in this study are similar to the results of phylogenetic analysis on Indonesian SARS-CoV-2 genome by the Next strain team based on data from GISAID from March to April 2020, showing the distribution of Indonesian clades that are categorized into clade B (L), B (O), and B. 1.36 (GH). This shows a close relationship with several countries such as the United States, Bangladesh, Nigeria, China, Taiwan, and Sweden.

The findings of mutations that occur in the Spike gene of Indonesia's SARS-CoV-2 and the results of the phylogenetic analysis of Indonesia's S protein SARS-CoV-2 prove that Indonesia's coronavirus strains are not 100% similar to viruses from other countries. There are several mutations in the S gene that need to be studied further to see their effect on the ability of SARS-CoV-2 virus to infect and its interactions with antibodies considering that S protein is the main target of the subunit vaccine being developed in Indonesia.

In conclusion, this study shows the presence of highly mutated S gene in Indonesian SARS-CoV-2 samples. Although the mutations identified are not located on the binding site of RBD or the antibody binding site, the effect of these mutations on changes in the structure and function of Spike needs to be studied in more in-deph manner, especially for the four specific mutations that only onccur in Indonesian samples. The D614G mutation, which is a global mutation, is seen in 56% of the samples, indicating a higher spread of this mutant virus. The phylogenetic analysis of the nucleotide and amino acid sequences of Indonesian Spike samples shows that the Indonesian patients' Spike has been divided into several small branches based on the mutation.

References

1. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; 579(7798): 270–73.
2. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. 2020;395(10224): 565–74.
3. Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R. A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping. Cells. 2020; 9(1267):1-22.
4. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, et al. Genome composition and divergence of the novel Coronavirus (2019-nCoV) originating in China. Cell Host Microbe. 2020;27:325–28.
5. Jaimes JA, André NM, Chappie JS, Millet JK, Whittaker GR. Phylogenetic analysis and structural modeling of SARS-CoV-2 Spike protein reveals an evolutionary distinct and proteolytically sensitive activation Loop. J Mol Biol. 2020; 432(10): 3309–25.
6. Rabi FA, Al Zoubi MS, Kasasbeh GA, Salameh DM, Al-Nasser AD. SARS-CoV-2 and Coronavirus Disease 2019: What We Know So Far. Pathogens. 2020; 9(3):231.
7. Belouzard S, Millet JK, Licitra BN, Whittaker GR. Mechanisms of coronavirus cell entry mediated by the viral Spike protein. Viruses. 2012; 4(6): 1011–33.
8. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020;581(7807):215–20.
9. Wang M, Fu T, Hao J, Li L, Tian M, Jin N, Ren L, Li C. A recombinant Lactobacillus plantarum strain expressing the spike protein of SARS-CoV-2. International J Biol Macromol. 2020; 160:736–40.
10. Zha L, Zhao H, Mohsen MO, Hong L, Zhou Y, Li Z, et al. Development of a COVID-19 vaccine based on the receptor binding domain displayed on virus-like particles. bioRxiv. 2020;1–6.
11. Gunadi, Wibawa H, Marcellus, Hakim MS, Daniwijaya EW, Rizki LP, et al. Full-length genome characterization and phylogenetic analysis of SARS-CoV-2 virus strains from Yogyakarta and Central Java, Indonesia. PeerJ. 2020;8:e10575.
12. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4): 406–25.
13. Kumar S, Stecher G, Li M, Knyaz C,Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms F. U. Battistuzzi, ed. Mol Biol Evol. 2018;35(6): 1547–49.
14. Higgins D, Thompson J, Gibson D, Tompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence aligment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acid Res. 1994;22:4673–80.
15. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 Spike glycoprotein. Cell. 2020;181(2):281–92.
16. Yuan M, Wu NC, Zhu X, Lee CCD, So RTY, Lv H, et al. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science. 2020;368(6491):630–3.
17. Chaturvedi P, Han Y, Král P, Vuković L. Adaptive evolution of peptide inhibitors for mutating SARS-CoV-2. [published online

ahead of print, 2020 Oct 8]. Adv Theory Simul. 2020;2000156.

18. Korber B, Fischer W, Gnanakaran SG, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020;069054.

19. Isabel S, Graña-Miraglia L, Gutierrez JM, Bundalovic-Torma C, Groves HE, Isabel MR, et al. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. Sci Rep. 2020;10(1):14031.

20. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 Spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020;581(7807):215–20.

21. Hu J, He CL, Gao QZ, Zhang GJ, Cao XX, Long QX, et al. The D614G mutation of SARS-CoV-2 spike protein enhances viral infectivity. bioRxiv. 2020;161323.

22. Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izard T, et al The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity.

bioRxiv. 2020;2020.06.12.148726.

23. Hou YJ, Chiba S, Halfmann P, Ehre C, Kuroda M, Dinnon KH, et al. SARS-CoV-2 D614G variant exhibits enhanced replication ex vivo and earlier transmission in vivo. Preprint. bioRxiv. 2020;2020.09.28.317685.

24. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness and neutralization susceptibility. Preprint. bioRxiv. 2020;2020.09.01.278689. Published 2020 Sep 2.

25. Ansori ANM, Kharisma VD, Muttaqin SS, Antonius Y, Parikesit AA. Genetic Variant of SARS-CoV-2 Isolates in Indonesia: Spike Glycoprotein Gene. J Pure Appl Microbiol. 2020;14(suppl 1):971–8.

26. Planas D, Bruel T, Grzelak L, Guivel-Benhassine F, Staropoli I, Porrot F, et al. Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to neutralizing antibodies. [published online ahead of print, 2021 Mar 26]. Nat Med. 2021;10.1038/s41591-021-01318-5.