Iranian Journal of Microbiology

Volume 13 Number 2 (April 2021) 145-155

ORIGINAL ARTICLE

Bioinformatic analysis of the whole genome sequences of SARS-CoV-2 from Indonesia

Maria Ulfah, Is Helianti*

Department of Biocatalyst, Centre for Bioindustrial Technology, Agency for Assessment and Application of Technology (BPPT), Laboratorium of Bioindustrial Technology, LAPTIAB BPPT, Puspiptek-Serpong, Tangerang Selatan, Indonesia

Received: November 2020**, Accepted:** February 2021

ABSTRACT

Background and Objectives: In first May 2020, Indonesia has been successfully submitted the first three full-length sequence of SARS-CoV-2 to GISAID database. Until September 10th, 2020, Indonesia had submitted 54 WGS. In this study, we have analyzed and annotated SARS-CoV-2 mutations in spike protein and main proteases.

Materials and Methods: The Whole Genome Sequence (WGS) of Indonesia were obtained from GISAID data base. The 54 data were taken from March to September 10th, 2020. The sequences corresponded to Spike Protein (SP), 3-chymotrypsin like protease (3CLpro), and papain like protease (PLpro) were selected. The Wuhan genome was used as reference.

Results: In total WGS from Indonesia, we found 5 major clades, which dominated as G clade, where the mutation of D614G was found. This D614G was identified as much as 59%, which mostly reported in late samples submitted. Beside D614G mutation, we report three unique mutations: A352S, S477I, and Q677H. Besides, some mutations were also detected in two domains that were expected to be conserved region, the main viral proteases: PLpro (P77L and V205I), 3CLpro (M49I and L50F).

Conclusion: The analysis of SARS-CoV-2 from WGS Indonesia showed a high genetic variation. The diversity in SARS-CoV-2 may epidemiologically enhance virulence and transmission of this virus. The prevalence of D614G over the time in different locations, indicating that changes in this mutation may related to host infection and the viral transmission. However, some mutations that have been reported in this study were not eligible for the most stable conformation.

Keywords: COVID-19; Genetic variation; Indonesia; Mutation; SARS-CoV-2; Spike glycoprotein; Viral proteases

INTRODUCTION

Pandemic COVID-19 which caused by the newly severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has become very serious health and social problem in the entire worldwide. Until September 13th, 2020 this pathogen was infected closed to 29 million people in the world and more than 917 thousand of mortality cases (source: World Health Organization). This pandemic become increasing more than 4 times in last three months. In Indonesia, more than 218 thousand cases were reported with more than 8 thousand cases of death (data were taken September 13th, 2020. Source: www.covid19.go.id). The transmission of this virus was massive and very fast. So, understanding and analyzing the genome diversity have been priority to design vaccine or drugs.

*Corresponding author: Is Helianti, Ph.D, Department of Biocatalyst, Centre for Bioindustrial Technology, Agency for Assessment and Application of Technology (BPPT), Laboratorium of Bioindustrial Technology, LAPTIAB BPPT, Puspiptek-Serpong, Tangerang Selatan, Indonesia.
Tel: +62-217560758 ext 7460
Fax: +62-217566922
Email: is.helianti@bppt.go.id

When the virus adapts to a new environment in new population, it could make change its genetical material/s and would bring modifications in viral proteins (1). Furthermore, the genetic material/s variation of would help the virus to survive and replicate in new host cells.

The single stranded RNA of SARS-CoV-2 was reported about >29,000 length base. The whole sequence has been identified and annotated. Foster et al, classified the three central variants differentiated by amino acid changes, named cluster A, B and C. The type A was ancestral of the group based on the bat outgroup coronavirus, recognized by mutation T29095C (2, 3). Type A and C were found significantly in Europeans and Americans, while type B belong to East Asia. Type B was derived from type A by two mutations: T8782C and C28144T, while type C was different from its parents type B, noticed by mutation G26144T, major found in Europeans (3). Meanwhile, complete viral genome sequence has been submitted and published to NCBI, Nextstrain, and GISAID databases. These platforms allowed us to access data and analyze this virus. Due to the genetic diversity of SARS-CoV-2, since July 4th, 2020 GISAID classified a nomenclature system based on marker mutation. The six clades were grouping from early split S and L. Clade L was split into V and G and later G to GH and GR. Clade O (others) was also introduced where mutation/s were not found in all classes (www.gisaid.org) (Table 1).

The SARS-CoV-2 contain at least four structural proteins: Spike glycoprotein (S), envelope (E), membrane protein (M), and nucleocapsid protein (N) (4-6). Among them, spike protein responsible for mediating to the host attachment and fusion of viral cell during infection (7, 8). The total length of SARS-CoV-2 Spike protein is 1273 amino acid residue. This consist of signal peptide (1-13 residues), the S1 subunit (14-685 residues) which separated by 4 (four) amino acid residues to S2 subunit (686-1,273 residues) (9). The alignment of S1/S2 junction with closely related bat-SL-RaTG13 found that SARS-CoV-2 contain 4 (four) amino acid residue insertion PRRA, that was not found in other sequence analysis (10). Spike protein consists of two domains: S1 and S2. In the SI subunit there is N-terminal domain (14-305 residues) and a Receptor Binding Domain (RBD, 319 -541 residues) (9). The S1 domain mediates attachment of receptor binding to host cell and S2 domain mediates in entry and fusion to host cell (1, 11). Hence, the spike protein become interesting domain to be further investigated for designing vaccine and even more as antivirus candidates.

Beside four structural proteins, SARS-CoV-2 viral genome also consist of two open reading frame (ORF) located in N-terminus, ORF1a (266-13468), and ORF1b (13, 468-21, 536) which serves to activate the intercellular pathways and drive immune response of the host (6). The main proteases are located in ORF1a, that play a role in processing the polyproteins and translated from viral RNA which include papain-like protease (PLpro) (4,955-5,900) and chymotrypsin-like protease (3CLpro) (10, 055-10, 977) (6, 12). The 3CLpro has no human homolog (13). This is one of main targets for the development of antiviral drug therapy because it plays an important role in viral replication (14). PLpro is important to proofread replication of virus. (8). Both PLpro and 3CLpro are in charge in processing polyprotein to mature protein (8). This study reported the variations in the genomes of SARS-CoV-2, especially mutation in spike protein and proteases which are distributed in Indonesia.

MATERIALS AND METHODS

Sequences accession. The fifty-four SARS-CoV-2 genomic sequences from different region of Indone-

Table 1. The list of 6 marker variants of SARS-CoV-2

Clade	Mutations
S	C8782T, T28144C NS8-L84S
L	C241, C3037, A23403, C8782, G11083, G25563, G26144, T28144, G28882
V	G11083T, G26144T NSP6-L37F + NS3-G251V
G	C241T, C3037T, A23403G, includes S-D614G
GH	C241T, C3037T, A23403G, G25563T, includes S-D614G + NS3-Q57H
GR	C241T, C3037T, A23403G, G28882A, includes S-D614G + N-G204R
O	Others

sia were obtained from GISAID data base. Data were taken from May to September 10th, 2020 arranged by the time submission (Supplementary file 1). The Wuhan genome EPI_ISL_449482 was used as reference. This sample was collected at the end of January 2020. From fifty-four WGS samples from Indonesia, 44 of full-length sequences (>29,000 nt) were analyzed, while low-quality sequences (300-900 nt) were removed. Besides, additional comparison data from other continents were also analyzed. Data from Africa/Egypt, Asia/India, Europe/ Italy and United Kingdom, North America/USA, South America/Brazil, and Oceania/Australia were also retrieved from GISAID between May to June 2020. The nucleotide sequences then translated into the protein sequences on EXPASY Translate tool. The sequences corresponded to Spike Protein (SP), 3-chymotrypsin like protease (3CLpro) and papain like protease (PLpro) were selected.

Sequence alignments and analysis. The nucleotides sequence was aligned by MAFFT version 7 (https://mafft.cbrc.jp/alignment/server/index.html). The phylogenetic tree was viewed using PhyloIO_ three, the sub menu in MAFFT version 7. All protein sequences of Spike Protein (SP), 3-chymotrypsin like protease (3CLpro), and papain like protease (PLpro) were aligned by multiple sequence alignment Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) and MAFFT version 7, while the result was viewed using MSAViewer, the sub menu in MAFFT version 7. Mutations in the protein sequence were recorded. The phylogenetic tree of Spike protein was visualized by Geneious Prime (Geneious Prime version 2021.0).

RESULTS

The first three full length sequence of SARS-COV-2 isolated from Hospital in Jakarta, was submitted on May 4th, 2020 by Eijkman Institute. The sequences were submitted on GISAID platform. Previously, the six earlier data were submitted by National Institute of Health Research and Development on March 4th, 2020, produced low quality sequences (300-900 nt), these remain the preliminary data submitted by Indonesian scientists.

From May to September 10th, we downloaded the data and performed a nucleotide alignment to recognize the variation of genomes. We choose 44 (forty-four) of full-length sequences and remove 10 (ten) of low-quality sequences. Some of data in full-length sequences showed non code nucleotide or 'N', so after translated into amino acid protein, it was appeared as 'X'. We considered the Wuhan origin genome used as reference in alignment. These nucleotides were aligned in MAFFT version 7 and analyzed the similarities and differences. From metadata database, some regions did not mention in certain area, such as Jawa Barat and Jawa Tengah. Bandung is part of Jawa Barat. Surabaya, Pasuruan and Sidoarjo are part of Jawa Timur. In this study, we mention the area based on the metadata submitted. We found that samples from different region in Indonesia clustered generally in two major clades: L and G and particularly in 5 different clades: L, G, GH, GR, and O. The G clade, including GH and GR dominated group with percentage about 59%. Two samples from Jawa Barat were classified as G clade (EPI_ISL_528748 and EPI_ISL_529719). Two samples from Jakarta and Jawa Barat were classified as GR (EPI_ISL_518820 and EPI_ISL_528747), while the rest was GH clade recorded in 22 (twenty-two) samples. One different mutation form GR and GH clade which found in alignment was G28882A. In G clade, samples spread from different region; Surabaya, Bandung, Yogyakarta, Jakarta, Jawa tengah, Jawa Barat, and Sidoarjo. Clade L was dominated from sample from Jakarta, the first outbreak was announced, then Surabaya, and even found in and Samarinda and Manado. One sample from Jakarta was identified as O clade. Uniquely, data comparison from other continents were clustered in G Clade. The North America/ USA EPI_ISL_480365 and Europe/ Italy EPI_ISL_560407 were grouped in G clade; Europe/ United Kingdom EPI_ISL_559682, Asia/ India EPI_ISL_481110, South America/Brazil EPI_ISL_492044 and Oceania/Australia EPI_ISL_521862 were recorded in GR clade; while Africa/ Egypt EPI_ISL_529031were recorded in GH clade (Fig. 1). The distribution of SARS-CoV-2 based on the first data submitted in GISAID platform was shown in map of Indonesia (Fig. 2).

From data obtained by the time of submission, the L clade was recorded in first data, which mutated less than the G clade. The theory was proven, then the L clade split into G clade, spread with some additional mutations. In this alignment, Wuhan origin genome was clustered in L clade. The sequences corresponded to Spike Protein (SP) was determined.

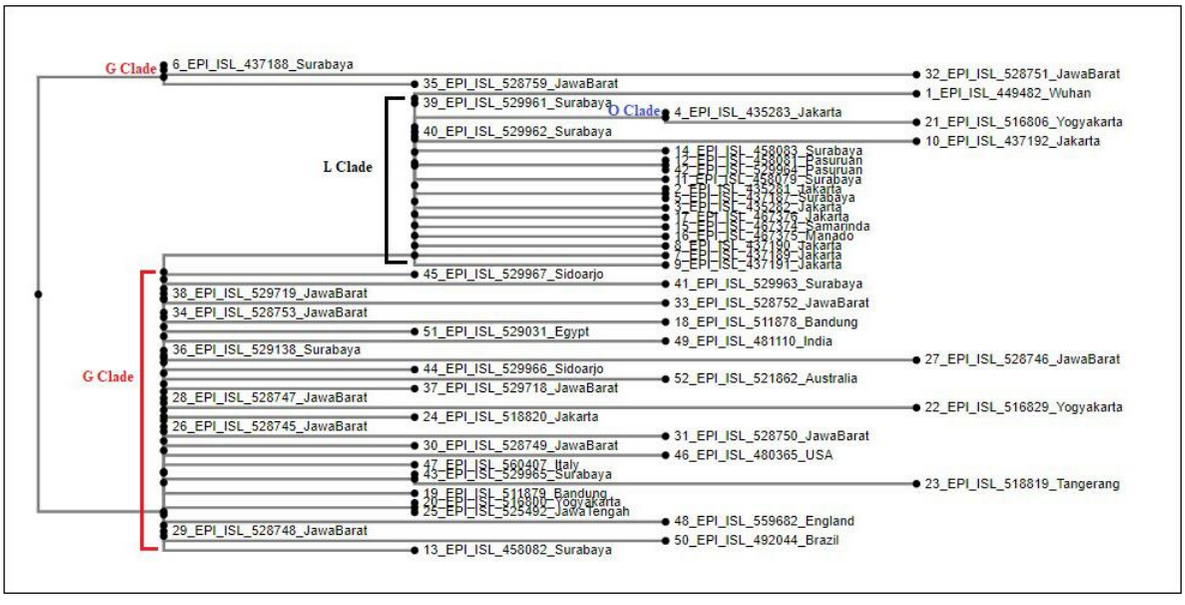

Fig. 1. Phylogenetic tree of nucleotide sequences of SARS-CoV-2 from different regions in Indonesia (44 WGS) and other continents. (PhyloIO_Tree).

Fig. 2. Distribution of SARS-CoV-2 in Indonesia based on the first data submitted to GISAID. (https://paintmaps.com/)

From alignment sequences using Clustal Omega, MAFFT version 7, and Genious prime, we recorded 18 (eighteen) kind different protein mutations from 33 (thirty-three) samples from selected region. The rest 11 (eleven) from 44 (forty-four) samples did not have mutation in Spike protein, in other words, these sample were closely like to reference. Some samples were recorded having two - three mutations (Tabel 2).

From phylogenetic tree analysis of spike protein using Genious Prime, we found the similar result as nucleotide alignment, that the G clade marked as D614G mutation, dominated the result with percentage about 59% (26/44) (Fig. 3). This mutation was early detected in Surabaya then spread to Bandung, and Jawa Barat, became the most cases found recently. The D614G mutation was also found in Yogyakarta, Tangerang, Jakarta, Jawa Tengah, and Sidoarjo. All samples represent from other continents, randomly taken from GISAID, were also recorded harbouring D614G. The alignment performed by MAFFT version 7 and visualized by MSAViewer

Table 2. Tabulation of accession ID with mutation in Spike Protein with specific region

No.	Accession ID	Spike Protein Mutation	No.	Accession ID	Spike Protein Mutation
1	EPI_ISL_425283_Jakarta	T76I	18	EPI_ISL_528747_Jawa Barat	D614G
2	EPI_ISL_437188_Surabaya	S116C; D614G; Q677H	19	EPI_ISL_528748_Jawa Barat	D614G
3	EPI_ISL_437189_Jakarta	V622F	20	EPI_ISL_528749_Jawa Barat	R214L, D614G
4	EPI_ISL_437192_Jakarta	T572I; L822F	21	EPI_ISL_528750_Jawa Barat	D614G
5	EPI_ISL_458081_Pasuruan	A352S	22	EPI_ISL_528751_Jawa Barat	Q677H; D614G
6	EPI_ISL_458082_Surabaya	D614G	23	EPI_ISL_528752_Jawa Barat	D614G
7	EPI_ISL_467374_Samarinda	A672V	24	EPI_ISL_528753_Jawa Barat	D614G
8	EPI_ISL_467376_Jakarta	C1254F	25	EPI_ISL_528759_Jawa Barat	T95I; Q677H, D614G
9	EPI_ISL_511878_Bandung	T259I, D614G	26	EPI_ISL_529138_Surabaya	D614G
10	EPI_ISL_511879_Bandung	N185Y, D614G	27	EPI_ISL_529718_Jawa Barat	D614G
11	EPI_ISL_516829_Yogyakarta	D614G	28	EPI_ISL_529719_Jawa Barat	D614G
12	EPI_ISL_516800_Yogyakarta	D614G	29	EPI_ISL_529963_Surabaya	T22I, A67V, D614G
13	EPI_ISL_518819_Tangerang	T22P, S477I, D614G	30	EPI_ISL_529964_Pasuruan	A352S
14	EPI_ISL_518820_Jakarta	D614G	31	EPI_ISL_529965_Surabaya	D614G
15	EPI_ISL_525492_Jawa Tengah	D614G	32	EPI_ISL_529966_Sidoarjo	D614G
16	EPI_ISL_528745_Jawa Barat	D614G	33	EPI_ISL_529967_Sidoarjo	D614G
17	EPI_ISL_528746_Jawa Barat	D614G			

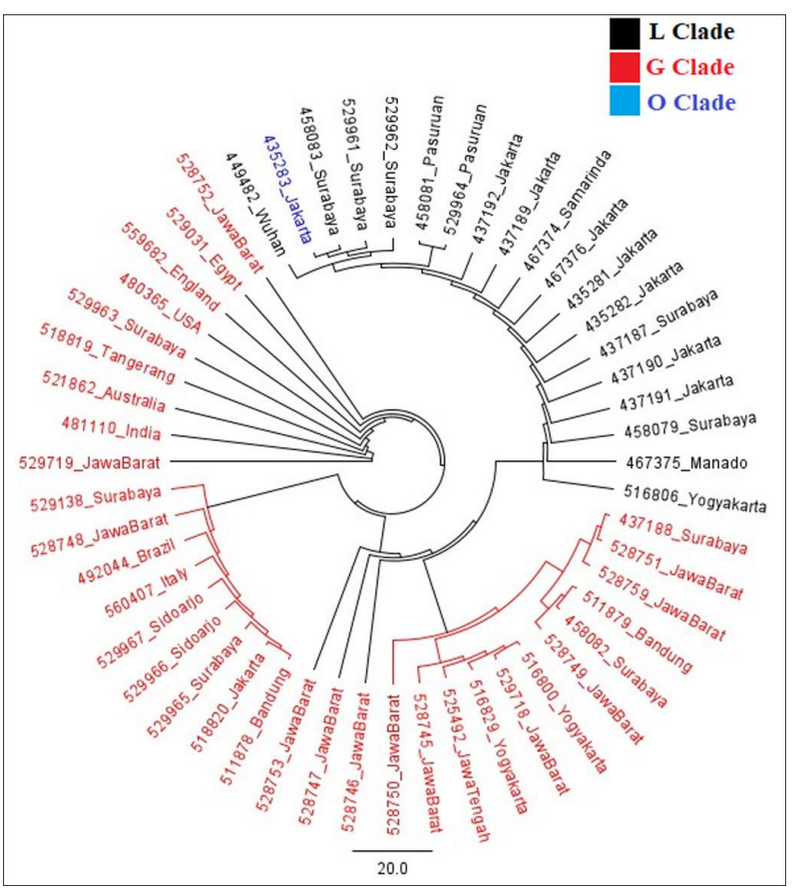

Fig. 3. Phylogenetic analysis of the spike protein sequences of SARS-CoV-2 from different regions of Indonesia and other continents. (Geneious Prime version 2021.0)

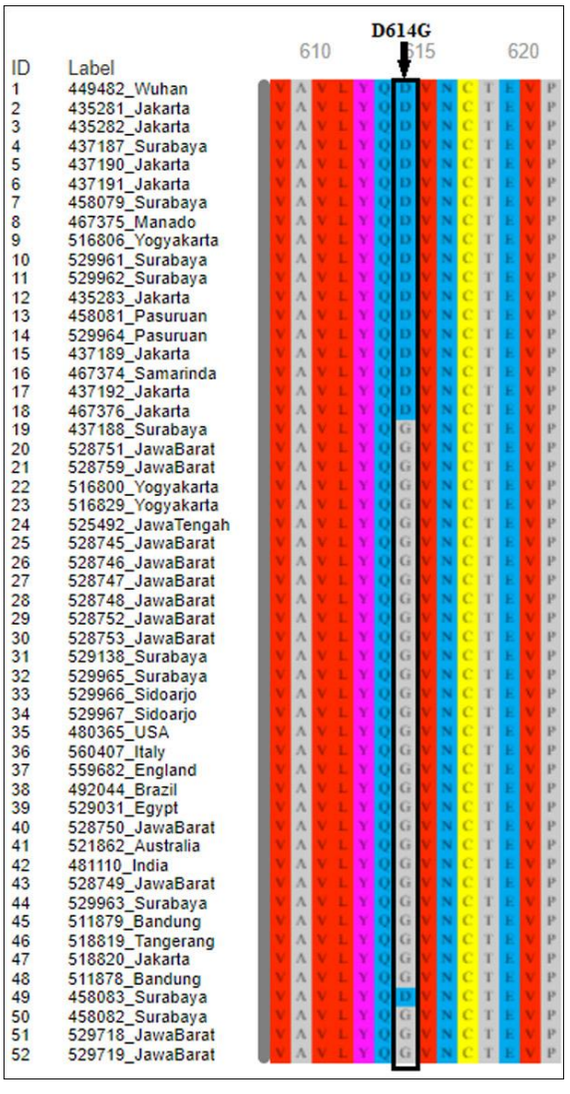

Fig. 4. Multiple sequence alignment of Spike protein from forty-four data from Indonesia and other continents. Site mutation of D614G is showed in black column.

showed a predominance of D614G mutation (Fig. 4). Recently, D614G mutation was spread massively and increased rapidly in late samples submitted. This recorded was increase since July to September 2020.

Meanwhile, we recorded three unique mutations at A352S, S477I, and Q677H. The mutation A352S was found in two samples EPI_ISL_458081 and EPI_ISL_529964. Samples were collected from the same location, Pasuruan at the same time. The alignment of this mutation was performed by MSAViewer in Fig. 5a. The position of A352S is in the spike protein S1 domain. Beside the A352S mutation, we also recorded the S477I mutation, which located in spike protein S1 domain. This mutation was found in sample from Tangerang EPI_ISL_518819. This sample was also recorded harboring D614G mutation. The alignment of S477I mutation is shown in Fig. 5b. Another unique mutation was also found in three samples, but different location. Mutation Q677H was found firstly in Surabaya on April 2020 (EPI_ISL_437188), then three months later it was found in Jawa Barat, in two different samples (EPI_ISL_518751 and EPI_ISL_518759) in July 2020. The position of Q677H is the S1 domain, relatively close to S1/S2 junction. The alignment of Q677H is showed in Fig. 5c.

In the ORF1a region of SARS-Cov-2, we detected four different mutations: P77L and V205I which found in PLpro (Fig. 6), while M49I and L50F in 3CLpro (Fig. 7). Surprisingly, substitution P77L was found in three samples from different regions, Yogyakarta, Jawa Tengah and Jawa Barat (EPI_ISL_516800; EPI_ISL_525492; EPI_ISL_528752, respectively). The mutation of V205I was observed in Bandung (EPI_ISL_511879). Interestingly, we recorded a mutation in active binding site of 3CLpro, M49I, which found in one sample from Yogyakarta (EPI_ISL_516806). The last mutation was detected L50F from Bandung (EPI_ISL_511878). In addition, the data for PL and 3CL proteases from other continents showed no mutation, or in other words the same as the reference.

DISCUSSION

As described before, regard to the phylogenetic clusters, Foster et al. 2020 distinguished three central variants groups based on amino acid changes, named A, B, and C. Type A and C were found significantly in Europeans and Americans, while type B was major common type found in East Asia. This type was derived from type A with two mutations: T8782C and C28144T. The analysis of genome variants showed that although it was widely found in East Asia, type B was also found in USA, Canada, Mexico, France, Germany, Italy, and Australia. even in a small portion (3). Analysis of genome variants from Indonesian samples demonstrated that there are 2 main clades: L and G. The variant mutations in L clade are similar to the first hypothesis of type B formation, where T8782C and C28144T are the markers. In the other words, this type B was a parent of L clade with additional variant mutations (Table 1). This complex history has one possibility, that ances-

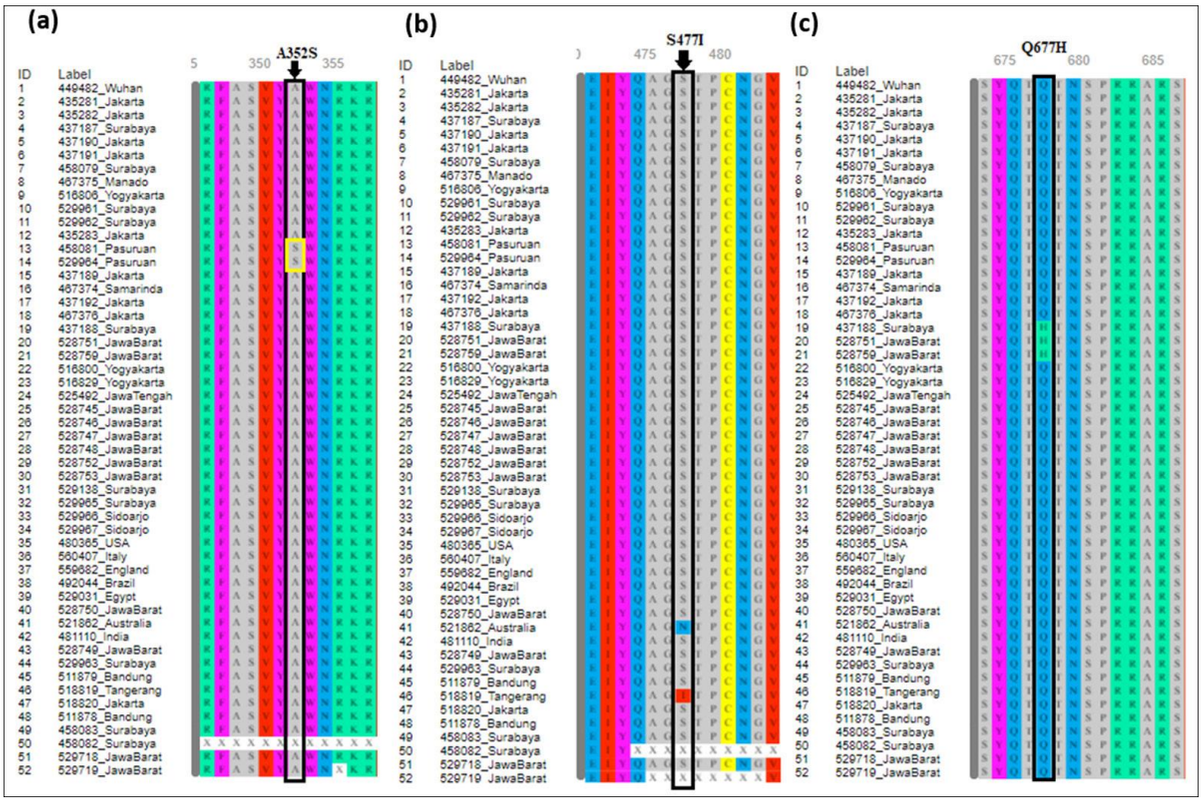

Fig. 5. Unique mutation in Spike Protein found in different region and compare to other continents. (a) Site mutation of A352S is showed in black column. (b) Site mutation of S477I is shown in black column. (c) Site mutation of Q677H is showed in black column.

tor type B was spread and environmentally adapted and mutated in Indonesian population then formed the L clade. From its origin of East Asia including Wuhan, before entering Indonesia this virus could be spread by adapting and carrying the D614G mutation (known as G clade marker), which may originated and distributed firstly in European and Americans (15, 16).

On May 2020, Ansori et al. has been reported the first-nine genetic variant of SARS-CoV-2 sample from Indonesia. The genetic analysis of Whole Genome Sequence demonstrated that there were no significant changes in SARS-CoV-2 Spike glycoprotein genes compare to Wuhan-Hu-1 origin (17). At that time, GISAID EpiCoV database had formed the nucleotide variants into 3 clades: S, G, and V, while the G clade mutation had no variation. In contrast with the analysis that we performed, in September we found 18 (eighteen) different kind of mutation in spike protein of forty-four samples. We demonstrated a lot of mutation emerged from samples in different regions. In fact, two – three mutation were found in one sample. The mutation that appeared most frequently was D614G. This mutation dominated the result with 59% and spread widely in different locations. Jawa Barat is the province where this mutation is mostly found. Even the last 12 samples submitted from Jawa Barat were detected to carry this mutation. In resume, 14 (fourteen) samples from Jawa Barat, include Bandung contribute 53% to the total result of D614G mutation. Comparing to the data taken from other continents, the D614G substitution recorded in all random data retrieved.

To the date, the mortality rate in Indonesia is the 23rd position in the world. Italy, as the first European country that infected heavily by SARS-CoV-2, including countries with high mortality rates. It has been reported in January to April, in total 79 samples were submitted and analysed, all of them harbour the D614G mutation (18). The D614G mutation may play a major role in death rates worldwide. This mutation tends to increase in many countries, as also reported by Sallam et al. that D614G mutation predominates in the Middle East and North Africa. It was noticed the escalation percentage from 63.0% in February to 98.5% in four months later (19).

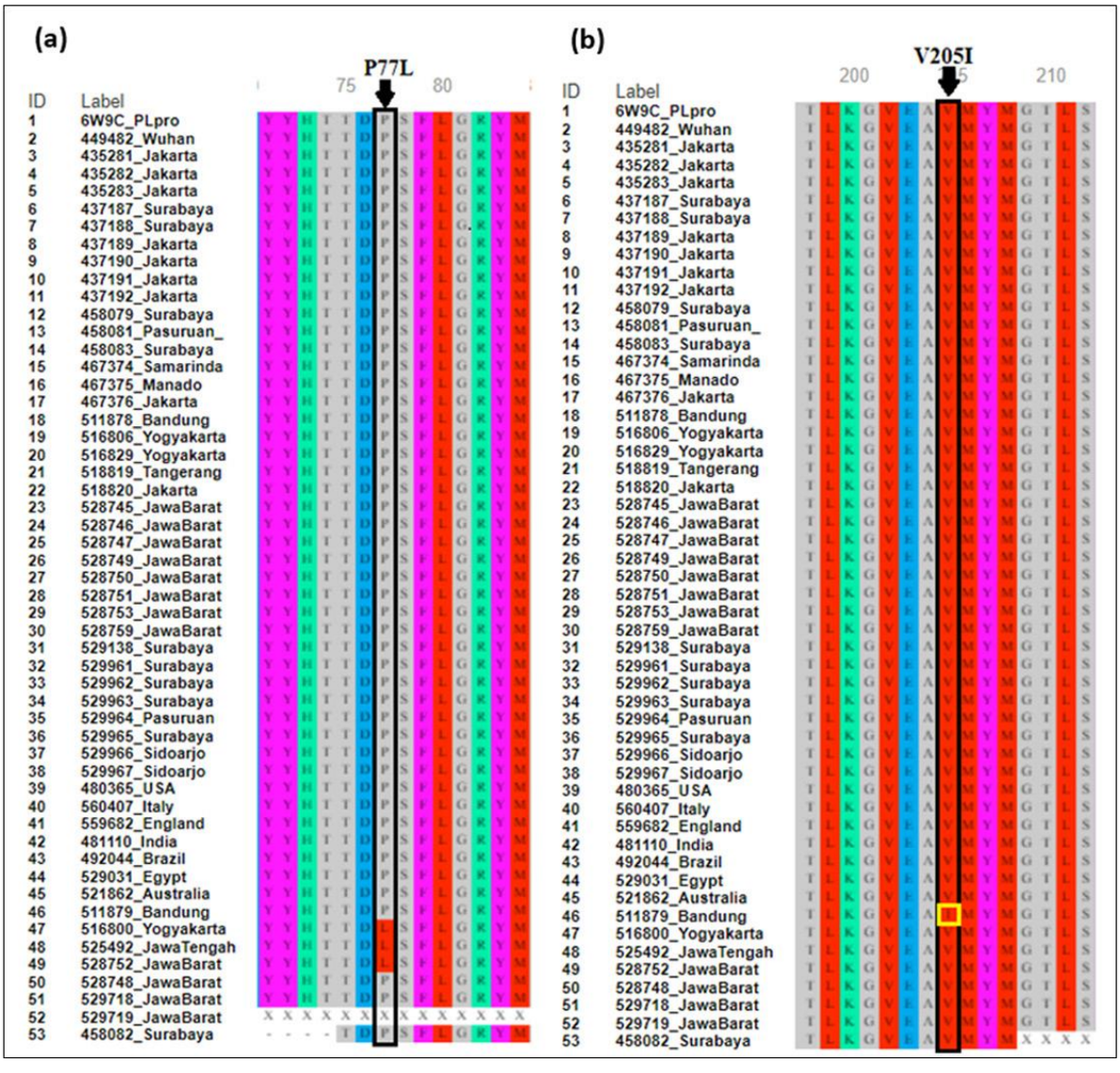

Fig. 6. The substitutions were found in PLpro. (a) The mutation of P77L is shown in black column. (b) The mutation of V205I is shown in black column.

The prevalence of D614G strain over the time in different locations, indicating that changes in this mutation may be related to host infection and the viral transmission (15, 20, 21). The DG614G mutation is located in S1 domain of spike protein. Since S1 domain mediates attachment of receptor binding to host cell, the substitution of D614G had been demonstrated to be more epidemiologically stable and increase efficiency in the host receptor binding (16, 21). Overall, worldwide increasing of G614 strains, giving speculation and indication that this mutation may be more virulent and infects more severely (18, 22). Nowadays, the D614G mutation was recognize as globally dominant variant (www.gisaid.org). Other mutations that were found uniquely in S1 subunit were A352G and S477I. These mutations were in Receptor Binding Domain (RBD, 319-541 residues). The substitution of A352G was recorded two samples from Pasuruan, Jawa Timur. This mutation was also reported in three countries with seven occurrences (23). Based on computational calculation of free energy, the mutation of A352G may contribute high energy, so that this mutation may not have meaning (23). Another mutation was S477I detected in Tangerang, western border of Jakarta. Other studies reported this S477 substitution but in different residue S477G (20) and S477R in Egypt (19).

Another unique mutation was recorded in this

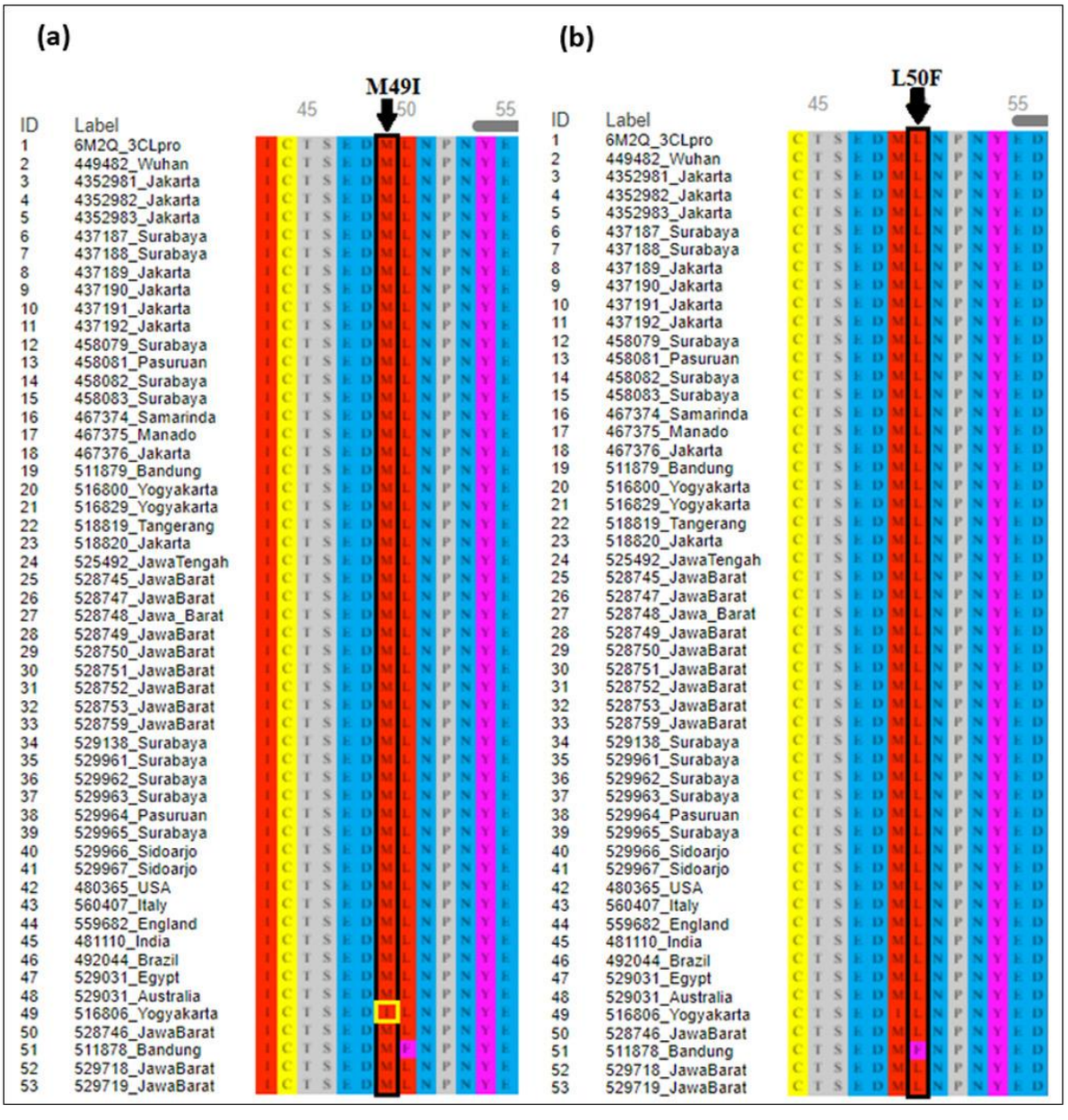

Fig. 7. The substitutions were found in 3CLpro. (a) The mutation of M49I is shown in black column. (b) The mutation of L50F is shown in black column.

study was Q677H. This mutation firstly detected one sample in Surabaya, then later two samples in Jawa Barat. In Egypt, the Q677H mutation, was detected as the most frequent mutation after D614G. It was recorded in 8 samples from total 553 samples from Middle East and North Africa (MENA) (19). Since this mutation frequently appeared in MENA region, it may indicate that this substitution spread worldwide from Middle East. The significant functional of Q677H had not been determined clearly, even though this has been described in previous study (20).

The study of variation in Spike Receptor Binding Domain protein had been reported (19, 20, 23). In MENA there are at least 8 different Spike Receptor Binding Domain variants (19). The multiple point mutations even reported in China, India, Australia, Taiwan, Malaysia, Canada, Spain, and UK (23). The mutations that have been reported were not eligible for the most stable conformation. The only possibility calculated was 2-point mutation combination of R355D and K424E that could produce strongest structural stability to the spike protein (23). Mean-

while, protein spike data from other continents did not show any substitution, but only sample from Asia/India was recorded having mutation in R1039T. In contrast to spike proteins which are identified as structural proteins, the main protease is in ORF1a. This includes PLpro and 3CLpro. In PLpro, we recorded at least two mutations; P77L and V205I. Gao et al. reported that catalytic site of PLpro includes residue C111, H272, D286, W93, W106, D108, and N109 (24). The others catalytic binding site residues that were reported, Asp164, Val165, Arg166, Glu167, Met 208, Ala246, Pro247, Pro248, Tyr 264, Gly266, Asn267, Tyr 268, Gln269, Cys217, Gly271, Tyr273, Thr301, and Asp302 (25). Since there were no report for residue P77 and V205, the substitution of Pro-77Leu and Val205Ile may not contribute to substrate stability.

In 3CLpro, we recorded two mutations, M49I and L50F. Yoshino et al. reported that residue M49 against inhibitor 2A5I ligand and Indinavir contributed with over 30% probability during simulation. However, His 41, Gly143, and Glu166 residues were showed high probability in all molecular dynamics (MD) simulations (26). Meanwhile, the catalytic residue His41 and Cys145 were reported more stable in substrate binding (12, 13). So, then the substitution M49I may not contribute high change in stability, but this assumption still needs to be proven in MD with different inhibitor or even in wet lab. The substitution L50F has not been reported to have a major contribution to the catalytic site.

In conclusion, several point mutations occurred in spike protein and proteases may influence the interaction between SARS-CoV02 in human infection. The diversity of mutation in SARS-CoV-2 may epidemiologically enhance virulence and transmission of this virus. This study showed that D614G mutation, as an issue of the worldwide today, indicated the virus more contagious. But the limitation of this study may not representative. We analysed 54 samples from 200 thousand data report infected. But this report could be taken into consideration for emerging, transmission and molecular epidemiology of SARS-CoV-2 virus, particularly would help to guide the SARS-CoV-2 vaccine development in Indonesia.

ACKNOWLEDGEMENTS

We would like to thank to the all-Indonesian institutions that submitted whole genome sequences of SARS-CoV-2 data to GISAID so that we can use for analysis, and to GISAID for making available the SARS-CoV-2 data.

This work was funded by Research and Innovation Consortium Program for Accelerated Handling of Corona Virus Disease 2019 (Covid-19) from LPDP (Indonesian Ministry of Finance-supported by Indonesian Ministry of Research and Technology granted to IH.

REFERENCES

1. Begum F, Mukherjee D, Thagriki D, Das S, Tripathi PP, Banerjee AK, et al. Analyses of spike protein from first deposited sequences of SARS-CoV2 from West Bengal, India. *bioRxiv* 2020. https://doi.org/10.1101/2020.04.28.066985.
2. Zhou P, Yang XL, Wang XG, Chen HD, Chen J, Luo Y, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;579: 270-273.
3. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* 2020;117: 9241-9243.
4. Chan JF, Kok KH, Zhu Z, Chu H, To KK, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with a typical pneumonia after visiting Wuhan. *Emerg Microbes Infect* 2020;9: 221-236.
5. Yoshimoto FK. The proteins of severe acute respiratory syndrome Coronavirus 2 (SARS CoV 2 or n COV19), the cause of COVID 19. *Protein J* 2020;39: 198-216.
6. Chitranshi N, Gupta VK, Rajput R, Godinez A, Pushpitha K, Shen T, et al. Evolving geographic diversity in SARS CoV2 and in silico analysis of replicating enzyme 3CLpro targeting repurposed drug candidates. *J Transl Med* 2020;18: 278.
7. Hoffmann M, Weber HK, Schroeder S, Kruger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* 2020;181: 271-280.e8.
8. Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm Sin B* 2020;10: 766-788.
9. Huang Y, Yang C, Xu X, Xu W, Liu S. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol Sin* 2020;41: 1141-1149.

10. 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: 3309-3325.
11. Ou X, Liu Y, Lei X, Li P, Mi Dan, Ren Lili, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat Commun* 2020;11: 1620.
12. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved a-ketoamide inhibitors. *Science* 2020;368:409-412.
13. Amamuddy OS, Verkhivker GM, Bishop ÖT. Impact of emerging mutations on the dynamic properties the SARS-CoV-2 main protease: an in silico investigation. *bioRxiv* 2020. https://doi.org/10.1101/2020.05.29.123190
14. Joshi RS, Jagdale SS, Bansode SB, Shankar SS, Tellis MB, Pandya VK, et al. Discovery of potential multi-target-directed ligands by targeting host-specific SARS-CoV-2 structurally conserved main protease. *J Biomol Struct Dyn* 2020;1-16.
15. Bette K, Will MF, Sandrasegaram G, Hyejin Y, James T, Werner A, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* 2020. https://doi.org/10.1101/2020.04.29.069054
16. Narendran PK, Prasanta S, Ashwani K. Distribution of the genetic clade "G" of Sars-cov-2 – an insight into COVID-19 virulence and spread in India. *IndiaRxiv* 2020. https://doi.org/10.35543/osf.io/5fzup
17. 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):971-978.
18. Benvenuto D, Demir AB, Giovanetti M, Bianchi M, Angeletti S, Pascarella S, et al. Evidence for mutations in SARS-CoV-2 Italian isolates potentially affecting virus transmission. *J Med Virol* 2020;92: 2232-2237.
19. Sallam M, Ababneh NA, Dababseh D, Bakri FG, Mahafzah A. Temporal increase in D614G mutation of SARS-CoV-2 in the Middle East and North Africa: Phylogenetic and mutation analysis study. *medRxiv* 2020. https://doi.org/10.1101/2020.08.24.20176792
20. Kim JS, Jang JH, Kim JM, Chung YS, Yoo CK, Han MG. Genome-wide identification and characterization of point mutations in the SARS-CoV-2 Genome. *Osong Public Health Res Perspect* 2020;11: 101-111.
21. 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. https://doi.10.1101/2020.06.12.148726
22. Eaaswarkhanth M, Al Madhoun A, Al-Mulla F. Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? *Int J Infect Dis* 2020;96: 459-460.
23. Smaoui MR, Yahyaoui H. Unraveling the stability landscape of mutations in the SARS-CoV-2 receptor-binding domain. PREPRINT (Version 1) available at Research Square 2020. https://doi.org/10.21203/rs.3.rs-59058/v1
24. Gao X, Qin B, Chen Pu, Zhu K, Hou P, Wojdyla JA, et al. Crystal structure of SARS-CoV-2 papain-like Protease. *Acta Pharm Sin B* 2021;11: 237-245.
25. Arya R, Das A, Prashar V, Kumar M. Potential inhibitors against papain-like protease of novel coronavirus (SARS-CoV-2) from FDA approved drugs. *chemRxiv* 2020. https://doi.org/10.26434/chemrxiv.11860011.v2
26. Yoshino R, Yasuo N, Sekijima M. Identification of key interactions between SARS CoV 2 main protease and inhibitor drug candidates. *Sci Rep* 2020;10: 12493.