Genomic diversity of SARS-CoV-2 in Malaysia

Noorliza Mohamad Noordin[1], Joon Liang Tan[2], Chee Kheong Chong[3], Yu Kie Chem[1], Norazimah Tajudin[1], Rehan Shuhada Abu Bakar[1], Selvanesan Sengol[1], Hannah Yik Phing Phoon[1], Nurul Aina Murni Che Azid[1], W Nur Afiza W Mohd Arifin[1], Zirwatul Adilah Aziz[1], Hani Hussin[1], Nurul Syahida Ibrahim[1], Aziyati Omar[1], Ushananthiny Ravi[1], Kamal Hisham Kamarul Zaman[1], Mohd Asri Yamin[1] and Yun Fong Ngeow[4,5]

[1] National Public Health Laboratory, Ministry of Health Malaysia, Sungai Buloh, Selangor, Malaysia
[2] Faculty of Information Science and Technology, Multimedia University, Melaka, Melaka, Malaysia
[3] Public Health Department, Ministry of Health, Putrajaya, Malaysia
[4] Faculty of Medicine and Health Sciences, Universiti Tunku Abdul Rahman, Bandar Sungai Long, Selangor, Malaysia
[5] Center for Research on Communincable Diseases, Universiti Tunku Abdul Rahman, Bandar Sungai Long, Selangor, Malaysia

ABSTRACT

Background. More than a year after its first appearance in December 2019, the COVID-19 pandemic is still on a rampage in many parts of the world. Although several vaccines have been approved for emergency use, the emergence and rapid spread of new SARS-CoV-2 variants have sparked fears of vaccine failure due to immune evasion. Massive viral genome sequencing has been recommended to track the genetic changes that could lead to adverse consequences.

Methods. We sequenced SARS-CoV-2 respiratory isolates from the National Public Health Laboratory, Malaysia and examined them together with viral genomes deposited in GISAID by other Malaysian researchers, to understand the evolutionary trend of the virus circulating in the country. We studied the distribution of virus lineages and site-wise mutations, analysed genetic clustering with the goeBURST full Minimum Spanning Tree algorithm, examined the trend of viral nucleotide diversity over time and performed nucleotide substitution association analyses.

Results. We identified 22 sub-lineages, 13 clonal complexes, 178 sequence types and seven sites of linkage disequilibrium in 277 SARS-CoV-2 genomes sequenced between January and December 2020. B.1.524 was the largest lineage group. The number of mutations per genome ranged from 0 to 19. The mean genomic diversity value over 12 months was 3.26×10^{-4}. Of 359 mutations detected, 60.5% of which were non-synonymous, the most frequent were in the *ORF1ab* (P4715L), *S* (D614G and A701V) and *N* (S194L) genes.

Conclusion. The SARS-CoV-2 virus accumulated an abundance of mutations in the first year of the COVID-19 pandemic in Malaysia. Its overall genetic diversity, however, is relatively low compared to other Asian countries with larger populations. Continuous genomic and epidemiological surveillance will help to clarify the evolutionary processes determining viral diversity and impacting on human health.

Submitted 2 June 2021
Accepted 18 October 2021
Published 3 November 2021

Corresponding author
Yun Fong Ngeow,
ngeowyf@utar.edu.my,
yunngeow@yahoo.com

Academic editor
Pedro Silva

Additional Information and Declarations can be found on page 11

DOI **10.7717/peerj.12449**

How to cite this article Mohamad Noordin N, Tan JL, Chong CK, Chem YK, Tajudin N, Abu Bakar RS, Sengol S, Phoon HYP, Che Azid NAM, W Mohd Arifin WNA, Aziz ZA, Hussin H, Ibrahim NS, Omar A, Ravi U, Kamarul Zaman KH, Yamin MA, Ngeow YF. 2021. Genomic diversity of SARS-CoV-2 in Malaysia. *PeerJ* **9**:e12449 http://doi.org/10.7717/peerj.12449

Subjects Bioinformatics, Genomics, Molecular Biology, Virology
Keywords SARS-CoV-2, Genetic diversity, Lineages, Mutations, Linkage disequilibrium

INTRODUCTION

Despite world-wide control efforts, the COVID-19 pandemic continues to ravage many populations since it was first reported in December 2019. This protraction has been attributed partially to the emergence of SARS-CoV-2 mutants that have been linked with increased infectivity and the ability to resist host immune responses. The best known variants with global spread since their introduction include the UK variant B.1.1.7 N501Y that has been shown to have higher transmissibility and risk of death than the wild-type virus (*Challen et al., 2021*), the S. African variant B.1.351 20H/501Y.V2 with spike protein mutations that apparently led to more frequent and more serious infections among young adults with no underlying illnesses and decreased neutralisation by antibodies (*Planas et al., 2021*) and the Brazilian variant P.1 GR/501Y.V3 with unique mutations in the receptor binding domain of the spike protein that has been linked with greatly increased transmissibility, higher mortality and lower susceptibility to inactivation by anti-S antibodies (*Planas et al., 2021*).

In South-East Asia, Malaysia is one of the countries to report COVID-19 infections as early as January 2020 when a few cases were imported from China. Since then, local transmission and further importation from different parts of the world increased the number of confirmed cases to 115,078 (an incidence of 352.14/100,000) with 474 (0.41%) deaths by 1 January 2021 and 514 clusters of infection reported in the 12-month period (*Fun et al., 2021*). One of the largest was associated with a Tablighi Jamaat religious mass gathering held in Kuala Lumpur between February 27 and March 3, 2020, that was attended by about 16,000 people (10% foreigners from many different countries) and 34 deaths among 3,375 infected individuals (*Fun et al., 2021*). From this event, the infection was spread all over Malaysia and to neighbouring countries as well. Other prominent outbreaks included the Pesantren cluster in April that was linked to Malaysian students returning from a religious school in East Java, Indonesia, and the Sivaganga cluster in July-August that was traced to a Malaysian who returned from India carrying the D614G variant (*Danial et al., 2020*). Non-Malaysians were predominant in a few clusters among construction workers and inmates of prisons and immigration detention centres. From October 2020 onwards, increasing cases occurred in the East Malaysian states of Sabah and Sarawak causing the incidence and the effective reproduction number (Rt) of COVID-19 there to exceed those in Peninsular Malaysia (*WHO, 2021*). Up to April 2021, five infections by B.1.1.7 N501Y and 21 by B.1.351 20H/501Y.V2 were reported by the Ministry of Health, Malaysia (*Malaymail, 2021a*; *Malaymail, 2021b*).

In this paper, we analysed 277 genome sequences from Malaysian SARS-CoV-2 isolates to study the diversity of the viral genomes over time, and to monitor the emergence of mutations that could affect the ability of the virus to spread or to cause more severe illness. We specifically looked out for rapidly spreading mutants that have emerged in other countries.

METHODOLOGY

Ethics statement

This study received waiver on informed consent by the Ministry of Health Medical Research Ethics Committee (no. KKM/NIHSEC/ P20-1094).

Source of SARS-CoV-2 genomes

The 277 SARS-CoV-2 genomes analysed comprised 30 sequenced (Table S1) by the National Public Health Laboratory Malaysia (MKAK) and 247 lodged in GISAID (*Elbe & Buckland-Merrett, 2017*) by other Malaysian research institutions, namely, the Institute for Medical Research (IMR), the Malaysian Genome Institute (MGI) and public universities (University Malaya Medical Centre and University Malaya Pahang). The genomes downloaded from the GISAID database were accessed on 5 January 2021. They represented 421 genomes deposited at different time points in Malaysia throughout the year 2020.

At the MKAK, nasopharyngeal and oropharyngeal swabs from patients seen in public hospitals were used for virus isolation in Vero E6 cells. From each patient, the isolate from the earliest available sample during illness and yielding the highest viral load was selected for whole genome sequencing using the MiSeq platform. The sequencing reads were evaluated in FastQC (*Andrews, 2010*) and the data obtained were preprocessed in PRINSEQ (*Schmieder & Edwards, 2011*), according to the quality assessments in the FastQC. In brief, the sequences were quality processed with phred probability mean score of Q20, 5′ and 3′ ends clipping, reads deduplication, removal of reads containing ambiguous nucleotide and removal of trimmed reads <67 bp. The preprocessed data were subjected to assembly in MEGAHIT (*Li et al., 2015*).

Genome analysis

All 421 sequenced and downloaded genomes were screened for bases other than IUPAC non-ambiguous nucleotides. This process resulted in 277 genomes for comparisons. All these genomes were aligned against Wuhan-Hu-1(NCBI Accession: NC_045512.2; GISAID Accession: EPI_ISL_402125) in MAFFT (*Katoh et al., 2002*). The mutations and corresponding amino acid changes were evaluated manually in MEGA X (*Kumar et al., 2018*). The clonal compositions of SARS-CoV-2 were evaluated with Minimum-Spanning Tree of Phyloviz 2.0 (*Francisco et al., 2012*). All genomes that emerged from a single founder were considered as belonging to a single cluster. Genomic and gene diversity studies were performed in DNASP 6 (*Rozas et al., 2017*). Additionally, DNASP 6 was further utilized to infer linkage disequilibrium which is the degree of non-random association of alleles (D′). Besides allele frequency, the linkage disequilibrium was calculated by considering the correlation coefficient of variant frequency (r^2) (*VanLiere & Rosenberg, 2008*). A strong association of mutations was predicted by an r^2 value > 0.8 and a Fisher one-tailed test of significance value < 0.01. All analyses were conducted with default parameters.

RESULTS

The genome analysis results based on Pangolin COVID-19 Lineage Assigner (https://pangolin.cog-uk.io/) showed 22 lineages with the lineage group B.1.524 forming the

majority (107, 38.6%), followed by B.6 (41, 14.8%), B (22, 7.9%), B.1.129 (21, 7.6%), B.6.1 (20, 7.2%) and other smaller groups (Fig. S1). All genomes were aligned at least 99% against the Wuhan-HU-1 genome except for two with only approximately 94% alignment.

Mutations in SARS-CoV-2 from Malaysia

Of the 277 genomes that underwent genome-wide alignment, nine were found to be identical to Wuhan-Hu-1. Of these, four were from Malaysian patients with no history of travel outside the country.

Viewing site-wise mutations, a total of 359 positions with variants were predicted, 63.8% of which were found to be solitary across the genomes analysed. These mutations were observed within nine protein encoding genes and four non-coding sites. Approximately 60% of mutations were non-synonymous and 67% involved a base substitution to Thymine. The highest number of mutations predicted in one genome was 19.

The most frequently seen mutations are listed in Table 1. Those shared by at least 100 genomes were at 13 genome positions (241, 3037, 6312, 8637, 10124, 14408, 17518, 21516, 21622, 23403, 23664, 28133, and 28854) with eight non-synonymous mutations in the *ORF1ab* (T2016I, T2791I, T3287A, P4715L, L5752F), *S* gene (D614G, A701V) and the *N* gene (S194L) respectively. The mutations shared by 50–99 genomes were at five positions (6312, 11083, 13730, 23929, and 28311). The non-synonymous mutations were in the *ORF1ab* (T2016K, L3606F, A4489V) and *N* (P13L). At the 6312 position there were two mutations, one with a T2016I (C- > T) change that was shared by >100 genomes and a second with a T2016K (C- > A) change that was shared by only 85 genomes. Excluding the substitution at position 241 in the 5′ UTR, the top three most frequent mutations were the C3037T (59.9%), C14408T (59.9%) and A23403G (D614G) (57.4%). These three mutations were reported to be found in SARS-CoV-2 genomes worldwide SARS-CoV-2 genomes worldwide.

Less frequent mutations (Table S2) included a unique 2 bp-deletion in *ORF8* (genomic position 28066–28067) that was found in only one genome, and the mutations (C6310A, T7621C and C19524T) reported by neighbouring countries such as Singapore, Australia, and India (*Chong et al., 2020*). There was only one genome each with the T8782C and C28144T mutations that have been used to distinguish the type B variant predominant in East Asia from the type A variant predominant in North America and Europe (*Forster et al., 2020*)

Identification of super spreader mutations

We specifically looked for mutations that have been linked to increased transmission in various countries. As shown in Table 1 above, the D614G amino acid substitution in the S spike protein was observed in 159 of 277 (57.4%) genomes. Based on the isolation dates in GISAID, this mutation was present in the viruses isolated in March 2020, suggesting that D614G was present in Malaysia before it was first reported in the Sivaganga cluster in July-August 2020. None of our D614G mutations were accompanied by the S477N substitution noted to be frequently alongside D614G in the S protein (*Singh et al., 2021*). We did not find the genomes of the UK, S. African and Brazilian super spreaders, although

Table 1 Mutations in at least 50 genomes of SARS-CoV-2.

Nucleotide substitution position	Impact of mutation	Gene	Number of strains
241	–	5′UTR	162
3037	Synonymous	*ORF1ab*	168
6312	T2016I	*ORF1ab*	106
8637	T2791I	*ORF1ab*	106
10124	T3287A	*ORF1ab*	107
14408	P4715L	*ORF1ab*	166
17518	L5752F	*ORF1ab*	106
21516	Synonymous	*ORF1ab*	107
21622	Synonymous	*S*	107
23403	D614G	*S*	159
23664	A701V	*S*	107
28133	Synonymous	*ORF8*	107
28854	S194L	*N*	112
6312	T2016K	*ORF1ab*	85
11083	L3606F	*ORF1ab*	91
13730	A4489V	*ORF1ab*	87
23929	Synonymous	*S*	81
28311	P13L	*N*	82

the E: P71L and S: A701V mutations defining the S. African variant B.1.351 501Y.V2 were found separately, in one and 107 genomes, respectively.

Clonality and diversity of SARS-CoV-2 genomes

The genomic variations of the 277 SARS-CoV-2 genomes were summarized with the goeBURST full Minimum Spanning Tree algorithm designed to detect clusters defined by genetic sequence similarities (Fig. 1). The analysis indicated 115 groups, of which 13 formed clonal complexes. There were 178 sequence types (STs) with 26 being subgroup founder's STs. The clonal complexes were not related to the month of virus isolation, or to other epidemiological features described for the outbreaks in different communities.

We further evaluated the trend of diversity in the SARS-CoV-2 genomes (Fig. 2). In 2020, the viral genomic diversity (the average number of nucleotide differences per site between sequences and its sampling variance) increased from 2.18×10^{-4} in the first quarter of the year to 3.34×10^{-4} in the second quarter and 3.86×10^{-4} in the third quarter before it fell to 3.65×10^{-4} in the final quarter. The mean genomic diversity was 3.26×10^{-4}.

Although the number of viral genomes analyzed in this study was small, the changes in diversity reflected major events in the country. Early in 2020, the diversity value was low because there were only a few viruses in circulation. A steep incline in diversity soon occurred with the spread of infections and hence opportunities for mutations to occur in the local population, and further importation from other countries which would have brought in different lineages and genetic variants of the SARS-CoV-2 virus. From August

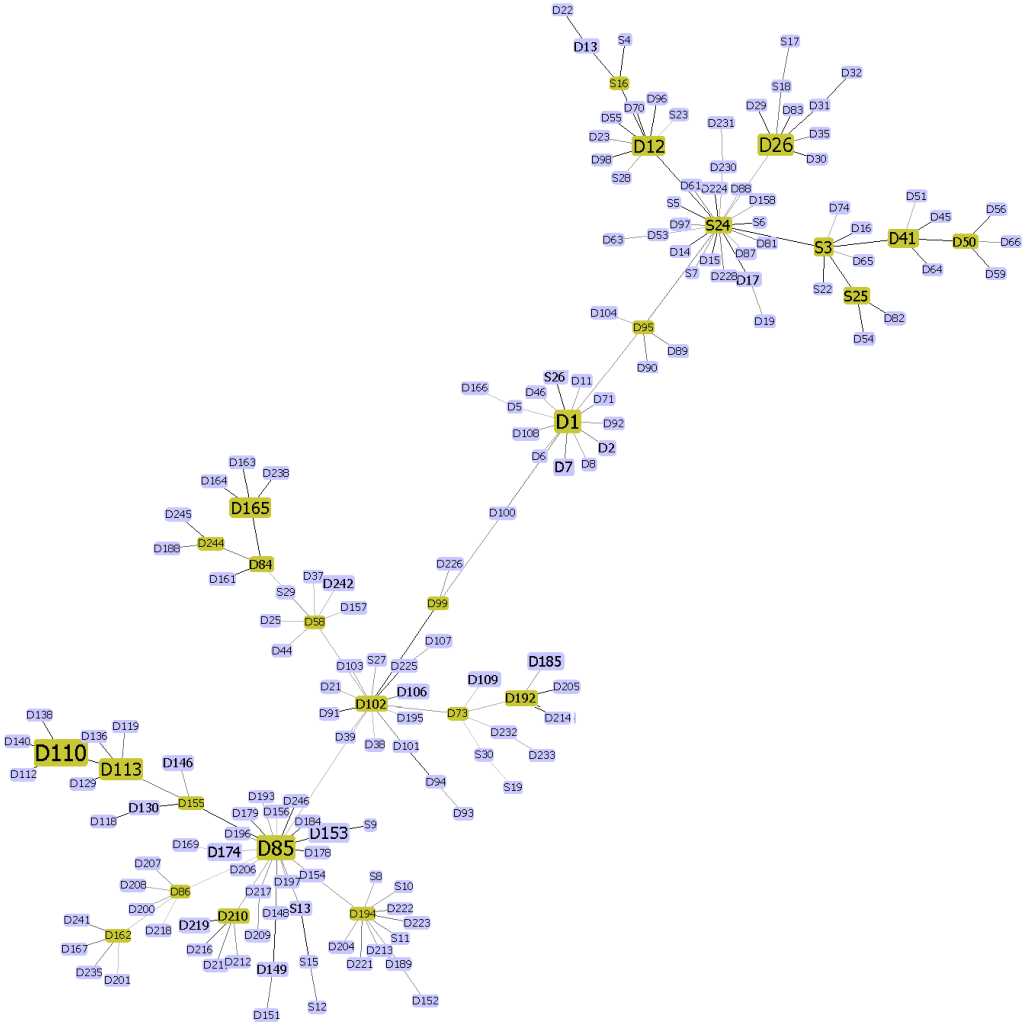

Figure 1 Diversity illustration of SARS-CoV-2 in Malaysia generated in Minimum Spanning Tree (MST). Each node represents either a strain or group of strains and linked based on MST. "D#" denotes downloaded genome and "S#" denotes genome sequenced in this study.

Full-size DOI: 10.7717/peerj.12449/fig-1

to December, although the number of reported cases increased, virus diversity showed a decline. The reason could be that the genomes we analyzed were mostly from a few large clusters and did not include new cases responsible for the spike in infections in Sabah towards the end of the year, as there were no SARS-CoV-2 genomes from Sabah in GISAID at the time we accessed the database.

To study nucleotide diversity of the SARS-CoV-2 genes affected by mutations, we looked at the genes for the four structural proteins E (envelope), M (membrane), N (nucleocapsid), S (spike), and the ORFs (open-reading frames). Figure 3 shows the diversity indices of eight genes which were, in descending order, 1.7×10^{-3} (N), 1.69×10^{-3} (*ORF8*), 0.79×10^{-3} (*ORF3a*), 0.61×10^{-3} (*S*), 0.49×10^{-3} (*M*), 0.41×10^{-3} (*ORF1ab*), 0.14×10^{-3} (*ORF7a*) and 0.13×10^{-3} (*E*). The *N, E, S* and *ORF1ab* are genes often used in RT-PCR assays for

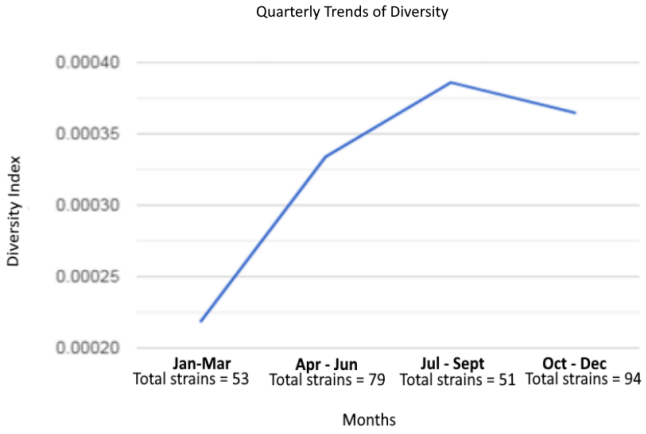

Figure 2 **Graph of SARS-CoV-2 diversity in Malaysia in 2020.**
Full-size DOI: 10.7717/peerj.12449/fig-2

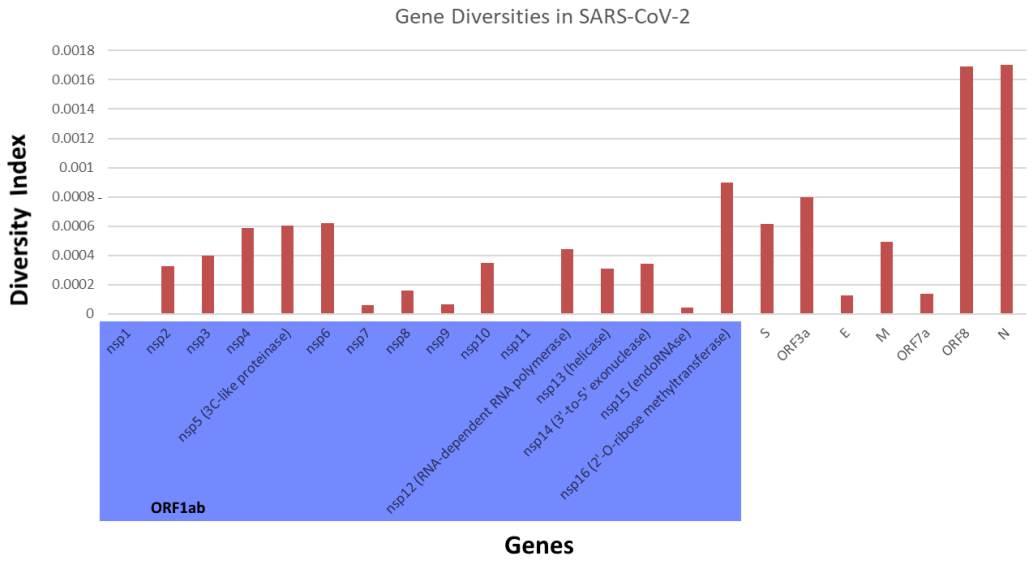

Figure 3 **Gene Diversity of SARS-CoV-2 in Malaysia.**
Full-size DOI: 10.7717/peerj.12449/fig-3

the diagnosis of COVID-19. A high diversity in the N gene is expected to decrease its sensitivity in these assays (*Hasan et al., 2021*). The high diversity in *ORF8* is consistent with reports describing it as a rapidly evolving gene (*Ceraolo & Giorgi, 2020*). Similarly, the lower diversity values in the *M* and *ORF7a* genes are not surprising as these genes have been reported to be stable genes that are suitable targets for vaccine development (*Nguyen et al., 2020*). There were 16 non-structural proteins (nsp) coded in *ORF1ab*. Both *nsp1* and *nsp11* showed a diversity index of 0, indicating high nsp conservation in the genomes. The most diverse nsp was *nsp16* (diversity of 8.99×10^{-4}). The diversity indices of the remaining nsp ranged from 4.47×10^{-5} to 6.17×10^{-4}.

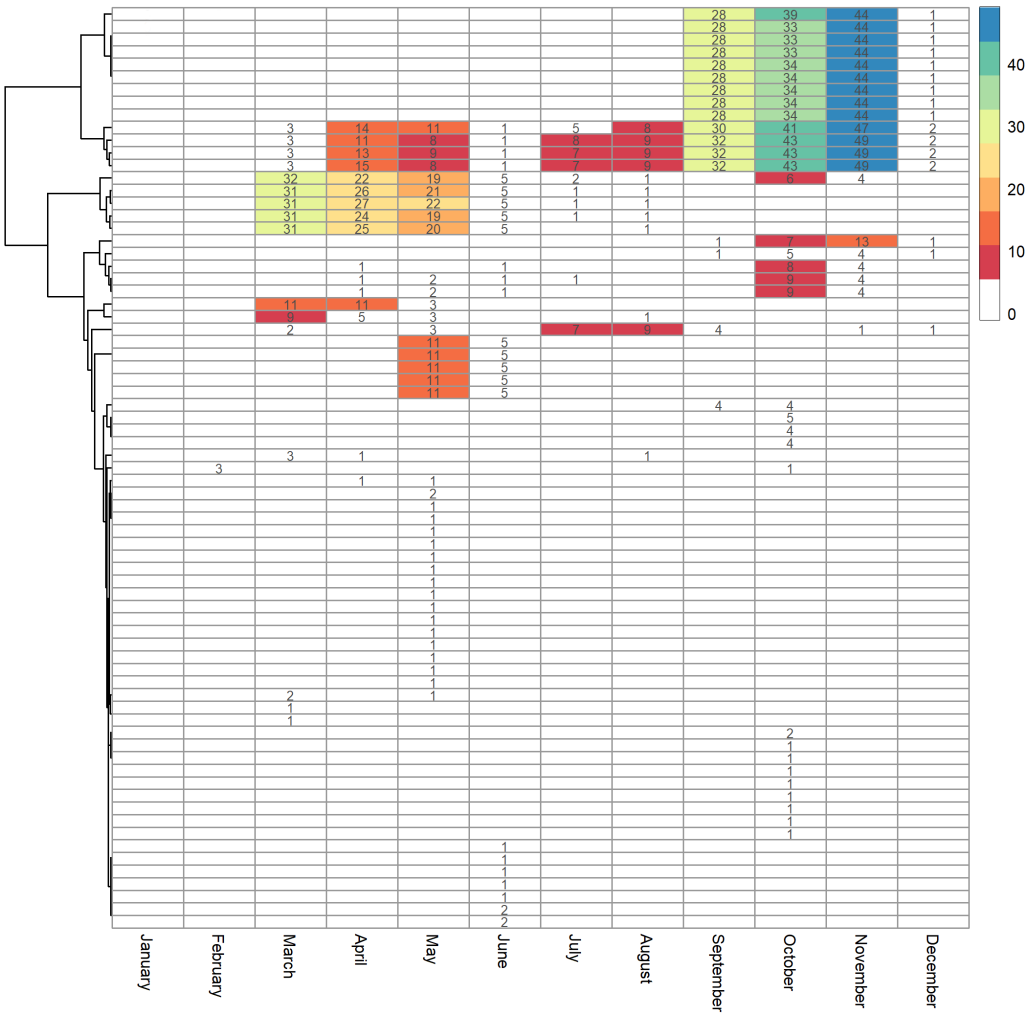

Figure 4 Heatmap of mutations in SARS-CoV-2 against months. The number in each cell indicates the number of strains having the mutation.

Full-size DOI: 10.7717/peerj.12449/fig-4

Nucleotide substitution association analyses

The temporal study on viral mutations showed that in January and February 2020, there were only a few mutations. From March to December, substitutions appeared consistently at the genomic positions C241T, C3037T, C14408T and A23403G. In addition, 10 more mutations appeared from September to December. These mutations were C5869T, C8637T, A10124G, C17518T, C21365T, C21516T, C21622A, C23664T, A28133T and C28854T. Two mutations, C13730T and C23929T were noted to be present in viruses isolated in March to August. At the genomic position of 6312, one mutation was observed in March to August (C6312A) and another appeared from September to December (C6312T). Month-specific mutations were the highest in May and not in November and December, the two months with the largest number of reported cases (Fig. 4).

When we checked for non-random mutational associations, we hypothesized that non-random association of mutations should be present in accordance with the frequency of occurrence. For this, we inferred based on r^2 and supported with Fisher one-tailed test of significance on the parsimony informative sites. The analyses returned seven sets of non-redundant site comparisons above the threshold with the r^2 ranging from 0.8 to 1.0 and with coefficient of linkage disequilibrium (LD) ranging from 0.93 to 1.0 (Table S3). All the positions identified were supported with a *p*-value of 0. The first LD set was observed in the viruses isolated in October 2020, at the genomic positions of A1904G, C12488T and G23236T. They were all non-synonymous mutations in *ORF1ab*. The second LD set comprised C3037T, C14408T and A23403G mutations in viruses isolated from March to December and involving a single synonymous and non-synonymous mutation in *ORF1ab*, together with a single non-synonymous mutation in the S protein. These are the mutations also reported by others (*Rouchka, Chariker & Chung, 2020*) to be linked. The third LD set was found in eight genomic positions, namely C8637T, A10124G, C17518T, C21516T, C21622A, C23664T, A28133T and C28854T. Except for C21516T, C21622A and A28133T, all were non-synonymous mutations, spanning *ORF1ab*, S protein, *ORF8* and N protein in viruses isolated from September to December 2020. The fourth LD set involved four mutations across three different proteins, at the genomic positions of G11083T, C13730T, C23929T and C28311T. The mutations were synonymous in the S protein, non-synonymous in N protein and non-synonymous in two positions in the ORF1ab. The fifth LD set involved genomic positions C13329T, C20823T, C26607T and A29086T. C20823T and A29086T caused non-synonymous mutations in the *ORF1ab* and *N* genes respectively, in viruses isolated in May and June. In the sixth LD set, C18877T, C26735T and G25563T were observed, the former two as synonymous mutations in *ORF1ab* and M protein, respectively, and the third as a non-synonymous mutation in *ORF3a*. Lastly, LD was observed in C24382T and G28307T causing a synonymous mutation in the S protein, and a non-synonymous mutation in the N protein, respectively. All in, the LD sets involved 21 non-synonymous and eight synonymous mutations in the *ORF1ab, ORF8, ORF3a, M, N* and *S* genes. While the LD sets were observed to be associated with the month of virus isolation, they were not clustered in the Minimum Spanning Tree.

DISCUSSION

The World Health Organisation has strongly advocated routine genomic surveillance for the SARS-CoV-2 virus (*WHO, 2021*). A principal aim is to look out for new variants that might cause adverse consequences such as increased viral infectivity and virulence, resistance to therapeutic agents, and immune evasion. As countries world-wide adopt vaccination to end the COVID-19 pandemic, many fear that under vaccine pressure, SARS-CoV-2 variants will emerge with new strategies to escape vaccine-induced immunity.

The diversity of SARS-CoV-2 is analysed to understand its pathogenicity, origin, and evolutionary implications (*Forster et al., 2020*; *Parlikar et al., 2020*). In this study, we found a mean genomic diversity index of 3.26×10^{-4} for Malaysian viruses sequenced between January and December 2020. This value is lower than those reported by *Flores-Alanis et al.*

(2021) for Asia (4.2×10^{-4}) and six other geographical regions in the world (0.44×10^{-4}). The generally low diversity in the SARS-CoV-2 virus has been noted by other researchers (*Dearlove et al., 2020*; *Rauscha et al., 2020*) and has been surmised to be due to the virus having an RNA polymerase with proofreading ability to correct errors during RNA synthesis (*Ogando et al., 2019*) as well as genetic drift and bottleneck events in the evolution of the virus (*Rausch, 2020*). Although low diversity is currently an advantage for vaccine design, it is expected that this landscape will soon be changed after the world-wide implementation of mass vaccination which would exert a tremendous immune pressure on the evolving virus.

As in most parts of Asia, the B lineage predominated among the 22 lineages identified in our Malaysian viral genomes. The largest lineage group was B.1.524 described in Pango Lineages as a Malaysian lineage that formed 65.0% and 26.0% of SARS-CoV-2 in Malaysia and Singapore, respectively, after it was first reported on 1 Sep 2020 (https://cov-lineages.org/lineages/lineage_B.1.html). Other major lineages (B.6, B.6.1, B.6.2 and B.6.6) were also those originally reported from India, Malaysia, and Singapore, possibly reflecting the frequent population movement among the three countries. There were more non-synonymous than synonymous mutations overall and among the LD sets. Among the mutation hotspots only D614G and A701V in the spike protein as well as P4715L and T2015I in *ORF1ab* are known to have a world-wide prevalence. Only one mutation, a 2 bp-deletion in *ORF8,* is possibly unique (not found among the genome sequences in GISAID).

The pattern of linkage disequilibrium (LD) in a genome is said to reflect the evolutionary history of an organism (*Illingworth & Mustonen, 2012*). We found seven sets of mutations showing LD, each with at least one non-synonymous mutation, distributed in the *ORF1ab, ORF8, ORF3a, N, M* and *S* genes. One of the LD sets included A23403G (D164G), C14408T and C3037T which were reported to be the most frequent mutations in many parts of the world (*Nguyen et al., 2020*; *Rouchka, Chariker & Chung, 2020*) and also found to be the most frequent mutations in this study. The linkage between D641G in the *S* gene and P4715L in *ORF1ab* is consistent with their presence together in >75% of global sequences (*Flores-Alanis et al., 2021*). While the LD sets were observed to be associated with the month of virus isolation, they were not clustered in the Minimum Spanning Tree, indicating independent evolution of each of the LD sets. As the linkages we observed were also reported by other researchers using different approaches and different cohorts (*Rouchka, Chariker & Chung, 2020*; *Haddad et al., 2021*), it is possible that the linked mutations might have evolutionary or survival significance in SARS-CoV-2. However, for most of the LD sets, we could not predict any association between the mutations showing LD and characteristics such as better adaptation to the host or potential for immune evasion.

This study is limited by the small number of SARS-CoV-2 genomes analyzed. Sampling bias could not be ruled out as it was not known how individual researchers selected the viral genomes they deposited in the GISAID database. As the downloaded genomes were only available in the form of assemblies in GISAID, we were unable to evaluate their quality and standardize the data preprocessing with the use of sequencing reads. There was also limited patient and epidemiological data on the genomes we analyzed. Thus, we were unable to

discuss issues such as intra-host viral heterogeneity (*Tang et al., 2020*) or the association of genetic variations with viral infectivity and pathogenicity.

CONCLUSIONS

This sample of SARS-CoV-2 genome sequences showed that within the first year of its appearance in Malaysia, the virus appeared to have accumulated an abundance of mutations including one possibly novel mutation in *ORF8*. Continuous genomic surveillance will provide information on the evolutionary trend of the virus, but adequate coverage of the circulating viruses is necessary to generate truly representative viral genetic characteristics that can be used along with epidemiological data to guide public health control strategies. This would require more extensive genome studies to enable the capture of all important developments in the evolution of the SARS-CoV-2 virus.

ACKNOWLEDGEMENTS

We would like to thank the Director General of Health Malaysia for his permission to publish this article.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Ministry of Health Malaysia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
The Ministry of Health Malaysia.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Noorliza Mohamad Noordin conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Joon Liang Tan analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Chee Kheong Chong, Yu Kie Chem, Norazimah Tajudin, Rehan Shuhada Abu Bakar, Selvanesan Sengol Hannah Yik Phing Phoon, Nurul Aina Murni Che Azid, W Nur Afiza W. Mohd Arifin, Zirwatul Adilah Aziz, Hani Hussin, Nurul Syahida Ibrahim, Aziyati Omar, Ushananthiny Ravi, Kamal Hisham Kamarul Zaman and Mohd Asri Yamin performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Yun Fong Ngeow conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

The Medical Research and Ethics Committee (MREC), Ministry of Health Malaysia (MOH) granted ethical approval to carry out the study (Ethical Application Ref: KKM/NIHSEC/ P20-1094).

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The sequenced genomes are available at GISAID: EPI_ISL_1972889 to EPI_ISL_1972895, EPI_ISL_2001114, EPI_ISL_2001115, EPI_ISL_2001213, EPI_ISL_2001429, EPI_ISL_2001430, EPI_ISL_2001481, EPI_ISL_2001482, EPI_ISL_2001643, EPI_ISL_2001644, EPI_ISL_2001645, EPI_ISL_2001654, EPI_ISL_2001655, EPI_ISL_2001656, EPI_ISL_2001659, EPI_ISL_2009116, EPI_ISL_2009193, EPI_ISL_2009232, EPI_ISL_2009233, EPI_ISL_2009234, EPI_ISL_2009305, EPI_ISL_2009529, EPI_ISL_2009530, EPI_ISL_2009595.

Data Availability

The following information was supplied regarding data availability:

The data is available at SRA: SRX11017364 to SRX11017393, SRP321799.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.12449#supplemental-information.

REFERENCES

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. *Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

Ceraolo C, Giorgi FM. 2020. Genomic variance of the 2019-nCoV coronavirus. *Journal Medical Virology* **92**:522–528 DOI 10.1002/jmv.25700.

Challen R, Brooks-Pollock E, Read JM, Dyson L, Tsaneva-Atanasova K, Danon L. 2021. Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. *British Medical Journal* **372**:n579 DOI 10.1136/bmj.n579.

Chong YM, Sam IC, Chong J, Kahar Bador M, Ponnampalavanar S, Syed Omar SF, Kamarulzaman A, Munusamy V, Wong CK, Jamaluddin FH, Chan YF. 2020. SARS-CoV-2 lineage B.6 was the major contributor to early pandemic transmission in Malaysia. *PLOS Neglected Tropical Diseases* **14**:e0008744 DOI 10.1371/journal.pntd.0008744.

Danial M, Arulappen AL, Ch'ng ASH, Looi I. 2020. Mitigation of COVID-19 clusters in Malaysia. *Journal of Global Health* **10**:0203105 DOI 10.7189/jogh.10.0203105.

Dearlove B, Lewitus E, Bai H, Li Y, Reeves DB, Joyce MG, Scott PT, Amare MF, Vasan S, Michael NL, Modjarrad K, Rolland M. 2020. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proceedings of the*

National Academy of Sciences of the United States of America **117**:23652–23662 DOI 10.1073/pnas.2008281117.

Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**:33–46 DOI 10.1002/gch2.1018.

Flores-Alanis A, Cruz-Rangel A, Rodríguez-Gómez F, González J, Torres-Guerrero CA, Delgado G, Cravioto A, Morales-Espinosa R. 2021. Molecular epidemiology surveillance of SARS-CoV-2: mutations and genetic diversity one year after emerging. *Pathogens* **10**:184 DOI 10.3390/pathogens10020184.

Forster P, Forster L, Renfrew C, Forster M. 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. *National Academy of Sciences (US)* **117**:9241–9243 DOI 10.1073/pnas.2004999117.

Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carrico JA. 2012. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics* **13**:87 DOI 10.1186/1471-2105-13-87.

Fun WH, Ang ZY, Shakirah MS, Anis-Syakira J, Cheah KY, Kong YL, Selvarajah S, Balqis-Ali NZ, Sararaks S. 2021. The COVID-19 Chronicles of Malaysia | Navigating the Storm. 1 2020-31 2020. The 3rd Chronicle. Shah Alam, Selangor: National Institutes of Health. *Available at http://maera.nih.gov.my/index.php/component/advportfoliopro/project/41-the-covid-19-chronicles-of-malaysia-edition-3*.

Haddad D, John SE, Mohammad A, Hammad MM, Hebbar P, Channanath A, Nizam R, Al-Qabandi S, Al Madhoun A, Alshukry A, Ali H, Thanaraj TA, Al-Mulla F. 2021. SARS-CoV-2: possible recombination and emergence of potentially more virulent strains. *PLOS ONE* **16**:e0251368 DOI 10.1371/journal.pone.0251368.

Hasan MR, Sundararaju S, Manickam C, Mirza F, Al-Hail H, Lorenz S, Tang P. 2021. A novel point mutation in the N gene of SARSCoV-2 may affect the detection of the virus by reverse transcription-quantitative PCR. *Journal of Clinical Microbiology* **59**:e03278–20 DOI 10.1128/JCM.03278-20.

Illingworth CJ, Mustonen V. 2012. Components of selection in the evolution of the influenza virus: linkage effects beat inherent selection. *PLOS Pathogens* **8**:e1003091 DOI 10.1371/journal.ppat.1003091.

Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**:3059–3066 DOI 10.1093/nar/gkf436.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* **35**:1547–1549 DOI 10.1093/molbev/msy096.

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct De Bruijn graph. *Bioinformatics* **31**:1674–1676 DOI 10.1093/bioinformatics/btv033.

Malaymail. 2021a. *Available at https://www.malaymail.com/news/malaysia/2021/04/15/covid-19-five-uk-b.1.1.7-variant-cases-detected-in-malaysia-says-dr-noor-hi/1966772* (accessed on 27 July 2021).

Malaymail. 2021b. *Available at https://www.malaymail.com/news/malaysia/2021/04/24/health-d-g-says-four-more-covid-19-cases-of-south-african-variant-found-in/1969176* (accessed on 27 July 2021).

Nguyen TT, Pham TN, Van T D, Nguyen TT, Nguyen DTN, Le HNM, Eden JS, Rockett RJ, Nguyen TTH, Vu BTN, Tran GV, Le TV, Dwyer DE, vanDoorn HR, OUCRU COVID-19 Research Group. 2020. Genetic diversity of SARS-CoV-2 and clinical, epidemiological characteristics of COVID-19 patients in Hanoi, Vietnam. *PLOS ONE* **15**:e0242537 DOI 10.1371/journal.pone.0242537.

Ogando NS, Ferron F, Decroly E, Canard B, Posthuma CC, Snijder EJ. 2019. The curious case of the nidovirus exoribonuclease: its role in RNA synthesis and replication fidelity. *Frontiers in Microbiology* **10**:1813 DOI 10.3389/fmicb.2019.01813.

Parlikar A, Kalia K, Sinha S, Patnaik S, Sharma N, Vemuri SG, Sharma G. 2020. Understanding genomic diversity, pan-genome, and evolution of SARS-CoV-2. *PeerJ* **8**:e9576 DOI 10.7717/peerj.9576.

Planas D, Bruel T, Grzelak L, Guivel-Benhassine F, Staropoli I, Porrot F, Planchais C, Buchrieser J, Rajah MM, Bishop E, Albert M, Donati F, Prot M, Behillil S, Enouf V, Maquart M, Smati-Lafarge M, Varon E, Schortgen F, Yahyaoui L, Gonzalez M, De Sèze J, Péré H, Veyer D, Sève A, Simon-Lorière E, Fafi-Kremer S, Stefic K, Mouquet H, Hocqueloux L, Van der Werf S, Prazuck T, Schwartz O. 2021. Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to neutralizing antibodies. *Nature Medicine* **27**:917–924 DOI 10.1038/s41591-021-01318-5.

Rausch JW, Capoferri AA, Katusiime MG, Patro SC, Kearney MF. 2020. Low genetic diversity may be an Achilles heel of SARS-CoV-2. *Proceedings of the National Academy of Sciences of the United States of America* **117**:24614–24616 DOI 10.1073/pnas.2017726117.

Rauscha JW, Capoferria AA, Katusiimea MG, Patro SC, Kearneya MF. 2020. Low genetic diversity may be an Achilles heel of SARS-CoV-2. *Proceedings of the National Academy of Sciences of the United States of America* **117**:24614–24616 DOI 10.1073/pnas.2017726117.

Rouchka EC, Chariker JH, Chung D. 2020. Variant analysis of 1,040 SARS-CoV-2 genomes. *PLOS ONE* **15**:e0241535 DOI 10.1371/journal.pone.0241535.

Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sanchez-Gracia A. 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution* **34**:3299–3302 DOI 10.1093/molbev/msx248.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**:863–864 DOI 10.1093/bioinformatics/btr026.

Singh S, Steinkellner G, Köchl K, KarlGruber , Gruber CC. 2021. Serine 477 plays a crucial role in the interaction of the SARS CoV 2 spike protein with the human receptor ACE2. *Scientifc Reports* **11**:4320 DOI 10.1038/s41598-021-83761-5.

Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J. 2020. On the origin and continuing evolution of SARS-CoV-2. *National Science Review* **7**:1012–1023 DOI 10.1093/nsr/nwaa036.

VanLiere JM, Rosenberg NA. 2008. Mathematical properties of the r^2 measure of linkage disequilibrium. *Theoretical Population Biology* **74**:130–137 DOI 10.1016/j.tpb.2008.05.006.

WHO. 2020. WHO urges greater surveillance as new COVID-19 variants emerge. *Available at https://www.afro.who.int/news/who-urges-greater-surveillance-new-covid-19-variants-emerge* (accessed on 20 April 2021).

WHO. 2021. Malaysia. Coronavirus Disease 2019 (COVID-19) Situation Report. Weekly report for the week ending 25 Oct 2020. *Available at https://www.who.int/docs/default-source/wpro---documents/countries/malaysia/coronavirus-disease-(covid-19)-situation-reports-in-malaysia/covid19-sitrep-mys-20201025.pdf?sfvrsn=f0cb9408_4&download=true* (accessed on July 2021).