ELSEVIER

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Journal Pre-proof

Comparative analysis of within-host diversity among vaccinated COVID-19 patients infected with different SARS-CoV-2 variants

Hebah A. Al-Khatib, Maria K. Smatti, Fatma H. Ali, Hadeel T. Zedan, Swapna Thomas, Muna N. Ahmed, Reham A. El kahlout, Mashael A. Al Bader, Dina Elgakhlab, Peter V. Coyle, Laith J. Abu-Raddad, Asma A. Al Thani, Hadi M. Yassine

PII: S2589-0042(22)01710-2

DOI: https://doi.org/10.1016/j.isci.2022.105438

Reference: ISCI 105438

To appear in: *ISCIENCE*

Received Date: 24 April 2022

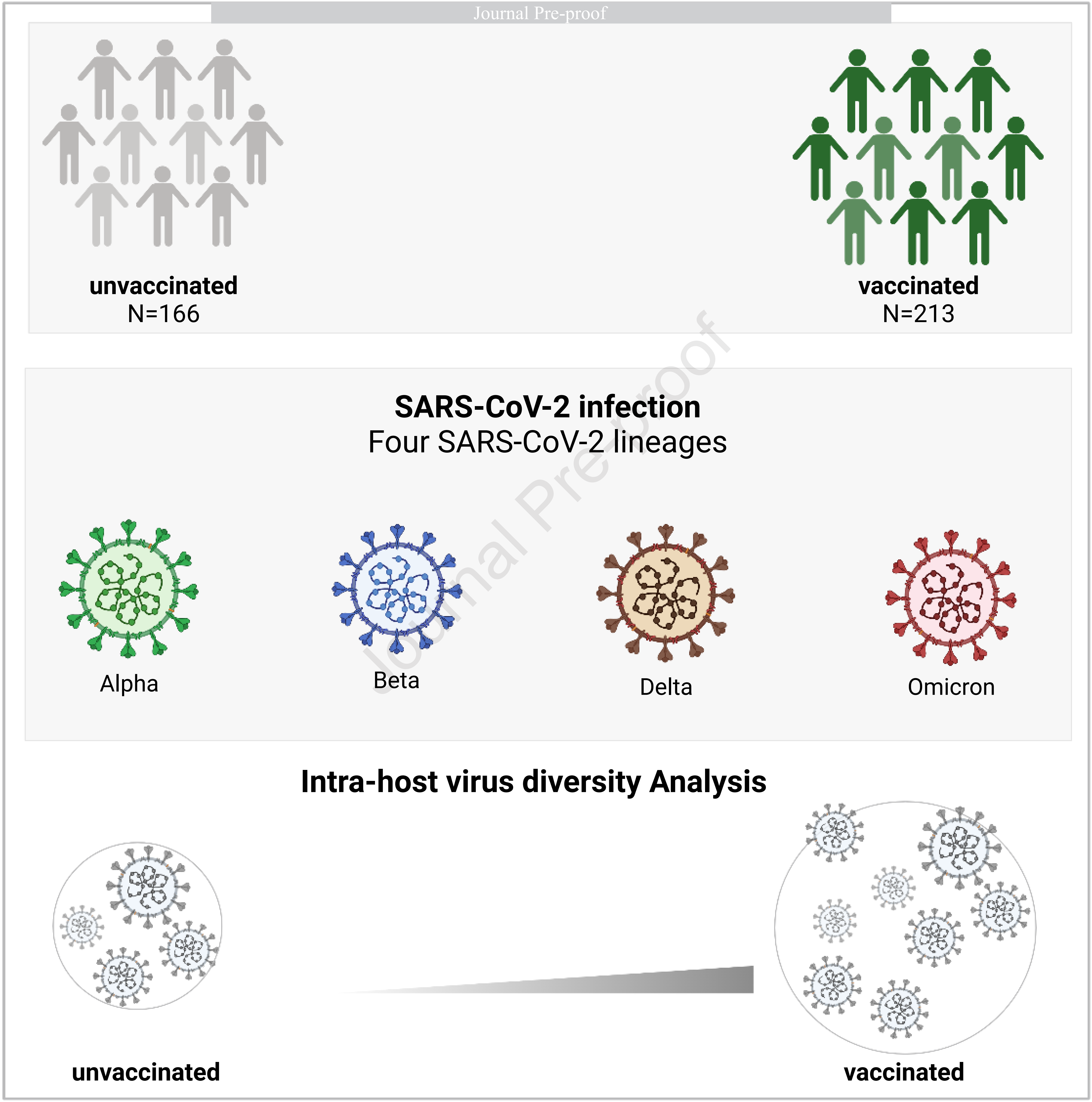
Revised Date: 18 September 2022

Accepted Date: 19 October 2022

Please cite this article as: Al-Khatib, H.A., Smatti, M.K., Ali, F.H., Zedan, H.T., Thomas, S., Ahmed, M.N., El kahlout, R.A., Al Bader, M.A., Elgakhlab, D., Coyle, P.V., Abu-Raddad, L.J., Al Thani, A.A., Yassine, H.M., Comparative analysis of within-host diversity among vaccinated COVID-19 patients infected with different SARS-CoV-2 variants, *ISCIENCE* (2022), doi: https://doi.org/10.1016/j.isci.2022.105438.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 The Author(s).



unvaccinated
N=166

vaccinated
N=213

SARS-CoV-2 infection
Four SARS-CoV-2 lineages

Alpha

Beta

Delta

Omicron

Intra-host virus diversity Analysis

unvaccinated

vaccinated

Journal Pre-proof

Journal Pre-proof

Comparative analysis of within-host diversity among vaccinated COVID-19 patients infected with different SARS-CoV-2 variants

Authors

Hebah A. Al-Khatib[1#], Maria K. Smatti[1], Fatma H. Ali[1], Hadeel T. Zedan[1], Swapna Thomas [1], Muna N. Ahmed[1], Reham A. El kahlout[2], Mashael A. Al Bader[3], Dina Elgakhlab[4], Peter V. Coyle[2], Laith J. Abu-Raddad[5], Asma A. Al Thani[1,6], Hadi M. Yassine[1,6]*

1 Biomedical Research Center, Qatar University, Doha 2713, Qatar.
2 Virology Laboratory, Hamad Medical Corporation, Doha, Qatar.
3 National Reference Laboratory, Ministry of Public Health, Doha 42, Qatar.
4 Qatar Biobank, Qatar Foundation, Doha, Qatar.
5 Weill Cornell Medicine–Qatar, Qatar Foundation, Doha, Qatar
6 Department of Biomedical Sciences, College of Health Sciences, Qatar University, Doha, Qatar.

*** Corresponding author**

Hadi M. Yassine
email: hyassine@qu.edu.qa
Phone: +974-4403-6819

#Lead contact

Hebah A. Al-Khatib
email h.alkhatib@qu.edu.qa
phone +974-44037709

Journal Pre-proof

Summary

SARS-CoV-2 is a rapidly evolving RNA virus that mutates within hosts and exists as viral quasispecies. Here, we evaluated the within-host diversity among vaccinated and unvaccinated individuals (n=379) infected with different SARS-CoV-2 variants of concern. The majority of samples harbored less than 14 iSNVs. Deep analysis revealed a significantly higher intra-host diversity in Omicron samples compared to other variants (p-value < 0.05). Vaccination status and type had a limited impact on intra-host diversity except for Beta-B.1.315 and Delta-B.1.617.2 vaccinees, who exhibited higher diversity compared to unvaccinated individuals (p-values: < 0.0001 and < 0.0021; respectively). Three immune-escape mutations were identified: S255F in Delta; and R346K and T376A in Omicron-B.1.1.529. The latter two mutations were fixed in BA.1 and BA.2 genomes, respectively. Overall, the relatively higher intra-host diversity among vaccinated individuals, and the detection of immune-escape mutations, despite rare, suggest a potential vaccine-induced immune pressure in vaccinated individuals.

Keywords: SARS-CoV-2 variants, within-host diversity, immune-escape mutations, mutation-selection

Journal Pre-proof

Introduction

Since its emergence in November 2019, SARS-CoV-2 has evolved rapidly, accumulating mutations, and generating new variants [1-4]. Emerging variants show variable characteristics of transmissibility, virulence, and immune evasion [5-7]. Based on these characteristics, the CDC has classified the new variants into “variants of concern”, “variants of interest” and “variants under monitoring” [8,9]. Recent variants of concern include the Delta and Omicron variants, which have caused the third and fourth waves of infection in many countries, respectively. The two variants are characterized by increased transmissibility and reduced neutralization by post-vaccination sera [6,10-12]. As SARS-CoV-2 continues to circulate globally, several genetic mutations have accumulated and will continue to accumulate, possibly at a faster rate as greater immunity develops in the population. The origin of SARS-CoV-2 variants remains unclear and there is no clear evidence explaining the mechanism(s) that led to their emergence. Several hypotheses have been proposed including (i) virus evolution in animals (zoonotic origin), (ii) virus evolution in long-term infected immunocompromised individuals, and (iii) virus evolution in immunocompetent individuals with pre-existing immunity (vaccination, infection, treatment with convalescent sera and monoclonal antibodies).

The evolution of coronaviruses, like other RNA viruses, begins with the accumulation of mutations as the virus replicates within hosts. Therefore, coronaviruses exist within hosts as a cloud of genomes referred to as within-host diversity (quasispecies). Among detected mutations, only a few may rise in frequency, transmit to other hosts, or even fix in the virus population [13,14]. Factors that determine within-host evolution of RNA viruses are not well understood. Multiple factors may affect within host evolution of RNA viruses including antigenic selection, antiviral treatment, tissue specificity, spatial structure, and multiplicity of infection [15-17]. Potentially advantageous mutations that confer enhanced receptor binding affinity, increased transmissibility, and immune escape properties might be selected and become dominant [18,19]. Recently, concerns have been raised that expanding massive vaccination could increase within-host selection for vaccine-escape mutations, ultimately undermining vaccine effectiveness [20,21]. Within-host SARS-CoV-2 diversity was commonly reported in COVID-19 patients, particularly among immunocompromised patients with persistent infection [22-25]. Investigating the within-host evolution of SARS-CoV-2 in immunocompromised patients revealed a dynamic within-host diversity that continues to change throughout the course of the infection [23,24]. More importantly, the lack of effective immune response in those patients allowed for a relaxed within host virus evolution which resulted in the emergence of immune escape mutations, many of which were found in other variants of concern (Alpha and Beta) [23].

Published data in immunocompetent patients reported variable levels of within-host diversity among COVID-19 patients [22,25,26]. These differences could be attributed to host and viral related factors such as age, underlying comorbidities, and SARS-CoV-2 lineage. We have previously shown that higher within-host diversity is commonly seen among elderly patients (>60 years old) and patients with severe respiratory symptoms [22]. Here, we evaluated within-host diversity among non-hospitalized

Journal Pre-proof

symptomatic COVID-19 infected with different SARS-CoV-2 variants of concern. We sequenced 340 SARS-CoV-2 genomes from samples collected during the four waves of infection in Qatar. The four waves were caused by Alpha, Beta, Delta, and Omicron variants, respectively. We further subdivided samples based on vaccination status and vaccination type to compare within-host diversity between vaccinated and unvaccinated individuals and to investigate the possible emergence of immune escape mutations among vaccinated individuals.

Results

Evaluating within-host diversity of SARS-CoV-2

To evaluate the within-host diversity of SARS-CoV-2, we called all intra-host single nucleotide variants (iSNVs) occurring above MAF of 0.05 in each of the analyzed samples. Overall, low levels of within-host diversity (less than 14 iSNVs) were reported among the majority of samples regardless of SARS-CoV-2 lineage (**figure 1a**). As expected, within-host diversity was significantly higher in Omicron positive samples compared to other lineages (**Table 2**). On average, Omicron positive samples exhibited the highest number of iSNVs (mean=14, SD=11.3), followed by Delta-B.1.1617.2 (mean=6, SD=7.5) and Beta (mean=6, SD=3.4), while the lowest diversity was reported among Alpha and Delta-AY.4 positive samples (mean=4). In all lineages, the total number of mutations in the virus genome was proportional to the number of mutations in the Spike (S) gene (**figure 1b**). All samples harbored at least one iSNV (**figure 2a**). The majority of samples had less than 14 iSNVs regardless of lineage. Eight samples had a higher number of iSNVs ranging from 30 to 70 iSNVs, however, the higher diversity within those samples was not associated with a particular lineage (**figure 2b**). Those samples belonged to Alpha, Delta-B.1.617.2, BA.1 and BA.2 lineages.

Then, we evaluated the impact of the vaccine on within-host diversity. While vaccination status did not seem to affect the within-host diversity in Omicron positive samples, significant differences were seen between vaccinated and unvaccinated samples collected from Beta (p-value <0.001) and Delta-B.1.617.2 (p-value <0.001) positive samples. Intriguingly, this significance was driven by Pfizer-vaccinated individuals in Beta positive samples (p-value <0.001) and by Moderna-vaccinated individuals in Delta-B.1.617.2 positive samples (**figure 3**). Lower within-host diversity was reported among BA.1 and BA.2 individuals who received three doses of the vaccine compared to those who received two doses. Moderna vaccinated individuals who received their third dose have generally exhibited lower diversity compared to those who received two doses. However, only a few samples were collected from individuals who received three doses of the vaccine so no confirmative conclusions could be drawn from this finding. We have also investigated the correlation between within-host diversity and the duration between vaccination and infection, however, no correlation was seen regardless of lineage, vaccination status, or vaccination type. A linear model was also performed to study the interaction effect of lineage and vaccine type on the prevalence of iSNVs. Analysis of the

Journal Pre-proof

interaction effect showed no significance effect of vaccine type on iSNVs regardless of SARS-CoV-2 lineage (**Supp. Figure 1)**.

Distribution of iSNVs across the genome

We next looked at the distribution of the identified iSNV sites across the genome. In all groups, the majority of iSNVs were found in the 3’ and 5’ untranslated (UTRs) and intragenic regions of SARS-CoV-2 sequences, and those were excluded from subsequent analysis. Overall, lineages exhibited variable numbers of iSNV sites ranging: 40 iSNVs in Alpha sequences, 59 iSNVs in Omicron-B.1.1.529, 57 iSNVs Omicron-BA.1, 61 iSNVs in Omicron-BA.2, 77 in Delta-AY.4 sequences, and 79 iSNVs in Beta sequences. The largest number of iSNV sites, though, were reported in Delta-B.1.617.2 sequences (n=188 iSNVs). The majority of iSNV sites in Omicron sequences are lineage-specific mutations that are mainly found in the S gene. On the other hand, the majority of mutations in Beta and Delta sequences are non-lineage specific mutations that are distributed across the genome. In all lineages, the distribution of iSNVs across the genes was considerably variable, with open-reading frames (ORFs) ORF1ab, 3a, nucleocapsid (N), and spike (S) genes showing the highest densities (**figure 4**). The majority of mutations in ORF1ab were localized in nsp3 and nsp12 (RNA Dependent RNA Polymerase, RdRp). The higher number of iSNVs in these two regions was associated with higher iSNV numbers in the receptor-binding domain (RBD) of S1, ORF3a, ORF8, and N genes. This was particularly seen in sequences from mRNA-1273-vaccinated individuals infected with Beta or Delta-AY.4 and from BNT162b2-vaccinated individuals infected with Delta-B.1.617.2 (**figure 4**).

In-depth analysis of non-synonymous, low-frequency mutations

Among the identified within-host mutations, only a few may rise in frequency and possibly transmit to other hosts [15]. Here, we focused the analysis on non-synonymous mutations with MAF 0.05-0.5 and evaluated their emergence, frequency, and prevalence among vaccinated and unvaccinated individuals. Only iSNVs sites not detected in the control and found in more than 2% of samples were included in the subsequent analysis.

In Alpha and Beta sequences, the vast majority of identified iSNVs were high-frequency, lineage-specific mutations (**figure 5**). In Beta sequences, high frequency, non-lineage mutations were found in ORF1ab (n=9), ORF3a (n=2) and N (n=1). Two mutations, L3829F in ORF1ab and A23V in ORF3a, were found to be under positive selection. Only one low-frequency, a non-lineage mutation was found, S: V1264L, and was found in 5% of Beta samples vaccinated with BNT162b2 (**figure 6**). Some mutations showed variable frequencies among the groups. The V202L mutation in ORF3a, for example, was found at low frequency (<0.5) in BNT162b2 vaccinated individuals and at higher frequencies (> 0.5) in unvaccinated individuals. Notably, none of the mutations in Beta sequences were associated with immune escape (**figure 6**).

Unlike Alpha and Beta sequences, Delta-B.1.617.2 exhibited a high number (n=42) of both high- and low-frequency non-lineage mutations, particularly in the S gene (n=10 iSNVs) (**figure 5**). Four mutations (out of 10 in S) were under positive selection: T29A, A67S, S255F, and T859N (**figure 6**). The S255F mutation in the N-terminal domain (NTD) of the S gene is also an escape mutant that demonstrated reduced neutralization by the potent NTD monoclonal antibody, mAb_S2L28 [27]. Another immune escape mutation that was found in the RBD of S protein is R346K. This mutation showed resistance to monoclonal antibodies such as C135. Of note, this mutation was detected later on at a higher prevalence in Omicron variant B.1.1.529 and was fixed in all BA.1 sequences (**figure 6**).

Delta-B.1.617.2 sequences have also exhibited a high number of ORF1ab mutations (n=19), seven of which are in the RdRp coding region, the highest compared to other lineages. Interestingly, six (out of seven) of RdRp mutations were found only in vaccinated individuals. This could partially explain the higher number of iSNVs in Delta-B.1.617.2 samples collected from mRNA-1273 vaccinated individuals.

Analysis of low-frequency mutations in Delta-B.1.617.2, in particular, revealed a large number of low-frequency mutations (n=33) compared to other lineages. Low-frequency mutations were found in ORF1ab (n=18), S (n=8), ORF3a (n=3), ORF8 (n=2) and N (n=2). Notably, though, the majority (24 out of 33) of identified low-frequency mutations were detected in vaccinated samples, particularly in mRNA-1273 vaccinated individuals. The prevalence of low-frequency mutations among mRNA-1273 vaccinated individuals was variable ranging from 9% to 18% (**figure 6**). While this may suggest a possible transmission of these low-frequency mutations, their emergence in mRNA-vaccinated individuals exclusively may favor the *de novo* emergence assumption rather than transmission. Seven of these mutations were found to be under positive selection pressure, but not associated with immune escape: four in the S gene, two in ORF 3a, and 2 in the N gene (**figure 6**).

The other Delta variant, AY.4, has also exhibited a relatively high number of non-lineage specific mutations (n=28). Unlike Delta-B.1.617.2, though, low-frequency mutations were less common among Delta-AY.4 samples. Mutations were located in ORF1ab (n=12), S (n=7), ORF3a (n=3) and N (n=4). All non-lineage mutations in S were found at high frequency except for K1073N which was detected at low frequency (<0.1) in 33% of mRNA-1273 vaccinated individuals. The rest of the non-lineage mutations, on the other hand, did not show any specific association with vaccination status and/or type. Moreover, six mutations (out of seven) in the S gene were found to be under positive selection pressure. Two S mutations, S255F and T29A were also found in Delta-B.1.617.2 sequences, however, S255F prevalence was higher in Delta-AY.4 samples regardless of vaccination status (**figure 6**).

In ORF1ab, mutations were found in nsp3 (n=1), nsp4 (n=1), nsp13 (n=9) and 3'-to-5' exonuclease (nsp14A, n=1) coding regions. Three of these mutations M2683I, E2993Q, and K6498T were found exclusively in mRNA-1273 vaccinated individuals (**figure 6**). Only four (out of the 12 non-lineage mutations) were found at low frequency. The rest of the ORF1ab mutations were detected at the consensus sequence level of Delta-AY.4 sequences. The prevalence of three low-frequency mutations in ORF1b: A942V (RdRp), A1779V (exonuclease), and K2097T (exonuclease) were higher in vaccinated samples.

Four non-lineage mutations were in the N gene, two were detected exclusively at low frequencies (<0.15) among vaccinated individuals: R40P and G71R. The other two were detected at consensus sequences of vaccinated (S232G), and unvaccinated (T135I) samples.

Omicron sequences generally exhibited a smaller number of non-lineage mutations. Each of the Omicron variants, B.1.1.529, BA.1, and BA.2, carried 22 non-lineage mutations. The majority of these mutations were located in ORF1ab and fewer mutations were found in S, ORF6, and N genes (**figure 6**). All Omicron variants shared five non-lineage mutations: ORF1ab: H236Q, ORF1b: L1639V, ORF6: D61H, N: D63G, N: D343G. Interestingly, ORF1b: L1639V which is located in the exonuclease coding region (nsp14A2) was found at low frequency exclusively in 33% of Omicron-B.1.1.529 and BA.1 individuals vaccinated with the mRNA-1273 vaccine. The prevalence of this mutation was found later on- at low frequency- in more than 50% of BA.2 samples regardless of vaccination status. A similar pattern was also seen for D61H mutation in ORF6. It appeared first as a low-frequency mutation (<0.3) in 30% of Omicron-B.1.1.529 and BA.1 individuals who received mRNA-1273, then was seen in the consensus sequences of all BA.2 samples regardless of vaccination status, suggesting a possible selection of this mutation.

BA.1 and BA.2 shared additional six non-lineage mutations; all were low-frequency mutations. The only exceptions were ORF1ab mutation, T1543I, and ORF1b mutation, T591I, which appeared at high frequencies in BA.2 and BA.1 omicron sub-lineages, respectively.

Within the host non-lineage, spike mutations were limited among all the three Omicron variants (**figure 6**). BA.1 and BA.2 shared one low-prevalent S mutation, S643L, which was found at the consensus sequence level in BA.1 samples and low frequency in BA.2. Two S mutations were identified as immune escape mutations: R346K (in Omicron-B.1.1.529 and BA.1) and T376A (in Omicron-B.1.1.529). All these mutations were found at low frequency in vaccinated individuals. Of these, R346K and T376A were found later on at high frequency and prevalence in the consensus sequences of BA.1 and BA.2, respectively.

Discussion

SARS-CoV-2 has evolved rapidly since its emergence in 2019 and generated hundreds of variants that caused multiple waves of infection worldwide [28]. Most analyses report on variants' mutations observed in virus consensus genomes and neglect mutations that appear at sub-consensus level which may affect virus characteristics [15,19]. Here, we looked beneath the consensus to analyze genetic variation within SARS-CoV-2 viral populations in individuals infected with four of the SARS-CoV-2 variants of concern. Overall, we reported low levels of within-host diversity among all samples regardless of causative SARS-CoV-2 variants. The limited number of within-host mutations can be attributed to the low mutation rate of coronaviruses (1.1 × 10^{-3} substitutions/site/year) [29]. Unlike other RNA viruses, coronaviruses exhibit a unique proofreading activity of its 3'-to-5' exoribonuclease which may correct some of the errors that occur during replication [30]. The limited number of within-host mutations in SARS-CoV-2 samples we reported here are consistent with other reported levels [24,31,32] but lower than in some other studies [26,33], likely reflecting differences in immune status of participants, sample selection criteria and variant calling methods. Higher within-host diversity is commonly reported among immunocompromised patients [23,25]. The absence of immune pressure in immunocompromised patients allows the virus to replicate and accumulate mutations at a faster rate compared to viruses replicating in immunocompetent patients. Conversely, the infection in immunocompetent patients as the case in our study is usually a self-limiting infection with limited diversity within-host diversity [22,24]. The limited within-host diversity can also be attributed to the dynamics of SARS-CoV-2 infection. Studies have demonstrated a dynamic within-host diversity throughout infection in both immunocompetent and immunocompromised patients [23,26,32,34]. A longitudinal study in an immunocompromised patient showed that virus diversity tends to increase during infection (after 14 days) [23]. In their study, Weigang and colleagues were able to identify several mutations, however, at later stages of infection in an immunocompromised patient. In this study, we analyzed patients' samples collected within 1 to 3 days following symptoms onset which may also explain the limited diversity and may not be reflecting the actual dynamicity of within-host diversity.

The emergence and diversity of within-host mutations in RNA viruses, including SARS-CoV-2, are driven by many factors including tissue specificity, antiviral treatment, and antigenic selection [15,17]. Antigenic selection is one of the major contributors to within-host virus evolution. Immune pressure resulting from vaccination and/or infection could in theory maximize the within-host diversity and potentially speed up the virus evolution rate. Therefore, we investigated the impact of the vaccine on the within-host diversity of different lineages. As expected, vaccinated individuals exhibited an overall higher number of within-host mutations, particularly in Delta-B.1.617.2 and Beta infected individuals. In Delta-B.1.617.2, higher diversity was particularly seen in mRNA-1273 vaccinees. The reasons for the higher within-host diversity among mRNA-1273 vaccinees are not clear. However, several studies have reported higher antibody response and more adverse side effects among mRNA-1273 vaccinees compared to BNT162b2 vaccinees [35-37]. The higher immune response following mRNA-1273 vaccination could exert immune pressure on B.1.617.2 to change and hence may explain the higher within-host diversity in those patients. Interestingly, the effectiveness of the mRNA-1273 vaccine was found to be higher in preventing B.1.617.2 infection compared to the BNT162b2 vaccine [38-40]. Unlike other lineages, the higher diversity seen among vaccinated Beta individuals was derived from BNT162b2 vaccinated individuals. This could be related to the higher number of samples compared to mRNA-1273 vaccinated individuals. This was due to the limited use of Moderna vaccine during Beta variant outbreak in the country. Altogether, this may suggest that the higher immune response elicited following mRNA-1273 vaccination has resulted in higher within-host mutations and hence a broader immune response which offered better protection against Delta-B.1.617.2 infections. Overall, the

Journal Pre-proof

mRNA-1273 vaccine has also demonstrated better effectiveness against other variants compared to BNT162b2. Yet, no significant difference in within-host diversity was found between mRNA-1273 and BNT162b2 vaccinated individuals in all other variants.

In addition to overall diversity, we examined the emergence, frequency, and spread of immune-escape mutations, especially among vaccinated individuals. Current data estimated that 65% of the world population has received at least one dose of vaccine [41]. In a highly seropositive population, the emergence of immune escape mutations is inevitable. Reports on vaccine breakthrough infections in vaccinated individuals are also accumulating, raising concern of escape mutations emergence as a result of immune selection [42,43]. The emergence of immune-escape mutations was clearly demonstrated at later stages in long-infected immunocompromised patients [23]. Here, we reported, despite rarely, the emergence of immune escape mutations in different variants. Of these mutations, R346K in S-RBD was of particular interest. In our data, this mutation appeared first at low prevalence in Delta-B.1.617.2 and Omicron-B.1.1.529, then at a higher prevalence in BA.1. Later on, BA.1 sequences carrying R346K mutation were assigned to a new Omicron sub-lineage, BA.1.1. However, the first appearance of R346K mutation was reported in Mu (B.1.621) variant which appeared in early 2021 in South America and spread later on to Europe [44]. Prediction and experimental methods showed that this mutation can escape recognition by more than 10 monoclonal antibodies [45,46]. Another escape mutation of interest is S255F in NTD of S. In our study, this mutation was seen in Delta lineages: B.1.617.2 and AY.4. In B.1.617.2 positive samples, S255F was detected in 10 samples (out of 81) and particularly among vaccinated individuals (>90%). Its prevalence among unvaccinated samples was higher (50%) in AY.4 samples. Unlike R346K, this mutation is not associated with any specific variant of concern. S255F is located within multiple T- and B cell epitopes and was found to be associated with reduced neutralization by a monoclonal antibody, mAb_S2L28 [27]. Despite its immune escape properties, this mutation was not fixed in Delta lineages and its prevalence remained limited. Taken together, putative within-host mutations may emerge in antigenic sites, particularly in vaccinated samples. However, few may rise in frequency and prevalence.

Conclusion

Since the first identification of SARS-CoV-2, hundreds of variants have emerged and spread globally causing multiple waves of infection. As the virus circulate in seropositive populations, more variants are expected to rise due to the immune pressure of previous infection and/or vaccination. Pre-existing immunity is expected to maximize the number of within host mutations and result in higher within host diversity. Here, we reported an relatively higher within-host mutations among vaccinated individuals, particularly among Beta and Delta-B.1.617.2 infected individuals. We have also investigated the emergence of immunity-evading mutations and reported, despite rare, mutations in Delta and Omicron lineages. Within-host mutations with resistance against natural or vaccine-induced immunity would probably be selected and replace previously circulating strains. Therefore, the continuous tracking of novel and potentially clinically important mutations is of great importance in light of public health, disease control, and the design of new preventive immunization strategies.

Limitation of Study

This study has some limitations that should be addressed. We could not sequence the effective sample size of some groups (Methods section). We did not study other “variants of concern” as those were not detected in Qatar. It is noteworthy to mention that more than 26 Delta sub-lineages were circulating in Qatar during the period between May and November 2022, however, those were not included in this analysis. Only the most prevalent Delta sub-lineages: B.1.1617.2 and AY.4 were included. Future work should focus on studying changes in within-host diversity in fully vaccinated individuals who received three doses of the vaccine. It should also investigate the impact of other vaccine types (non-RNA based vaccines) on within host diversity.

Acknowledgments

The authors are grateful for the leadership and assistance provided by the Ministry of Public Health in Qatar, the virology laboratory staff at Hamad Medical Corporation, and Qatar Biobank (QBB) team. This project was funded by Qatar National Research Fund (QNRF; Project number UREP28-164-3-048) and Qatar University (Project number QUCG-BRC-22/23-547).

Authors’ contribution

H.A. and H.Y. designed the concept; M.S., F.A., H.Z., S.T., M.N. performed sequencing and analyzed the data; M.A., R.E., P.C. helped in samples collection and other logistics; and D.E. helped in collecting demographic and clinical data; H.A. and A.A. provided the funding; H.A. wrote the first manuscript draft; All authors read and approved the final draft of the manuscript.

Declaration of Interests

The authors declare no conflict of interest.

Journal Pre-proof

Main Figure

Figure 1: Number of iSNVs (MAF > 0.05) observed in each sample of the SARS-CoV-2 lineages/sub-lineages. (**a**) Number of iSNVs seen in the whole genome of each sample. (**b**) Number of iSNVs seen in the spike gene of each sample. All mutation were called with respect to Wuhan Hu-1 reference sequence (RefSeq ID NC_045512). Data represent the mean of the mean number of iSNVs ± SD reported among samples that belong to the same lineage/sub-lineage. All samples (n= 379) in our dataset are included in this figure. Number of samples within each group is listed in **Table 1**.

Figure 2: Histograms showing the number of samples exhibiting N number of iSNVs (MAF > 0.05). (a) Histogram showing the total number of samples with N number of iSNVs. **(b)** Stacked histogram showing the number of samples that had N number of iSNV sub-categorized based on lineage: Alpha, Beta, Delta and Omicron lineages and sub-lineages. All identified sites were included in this figure except for those located in intragenic regions and in the upper and lower untranslated regions of the genome.

Figure 3: Average number of iSNVs observed in SARS-CoV-2 positive samples. Number of iSNVs seen in each sample of the SARS-CoV-2 lineages sub-divided based on **(a)** vaccination status: vaccinated and unvaccinated; and **(b)** mRNA vaccine type: mRNA-1273 (Moderna) and BNT162b2 (Pfizer). All mutations were included except for those located in intragenic regions and in the upper and lower untranslated regions of the genome. Significance is indicated as follows: * for <0.033, ** for <0.0021, *** for < 0.0002 and **** for < 0.0001. A detailed linear model analysis that incorporates the interaction between the virus lineage and vaccine type is demonstrated in **supp. Figure 1 and supp file 1**.

Figure 4: Heatmap demonstrating the distribution of iSNVs throughout SARS-CoV-2 genome of each lineage. All identified intra-host variation sites (MAF>0.05) were included in the heatmap analysis except for those located in intragenic regions and the upper and lower untranslated regions of the genome. The range on the right demonstrates the number of intra-host variations sites in each region of the genome where 0 indicates no mutations while 100 indicates that 100 mutations were found in this region.

Figure 5: Distribution and frequency of the non-synonymous mutations (MAF> 0.05) across SARS-CoV-2 genomes of each lineage.

Figure 6: Prevalence of non-lineage mutations identified in vaccinated and unvaccinated individuals of each lineage. This figure displays non-synonymous mutations that showed significant differences in their prevalence among the three groups: unvaccinated, Moderna-vaccinated and Pfizer-vaccinated individuals. The squares and triangles above the bars indicate immune escape mutations and positively selected site, respectively.

Journal Pre-proof

Main Tables

Table 1: Numbers of SARS-CoV-2 sequences within each group/subgroup included in this study.

Lineage	**Sub-lineage**	**vaccinated**		**unvaccinated**	**Total**	**Sample collection time period**
		mRNA-1273	**BNT162b2**			
Alpha		-	-	33	33	Dec-Jan, 2021
Beta		12	29	33	74	Mar-May, 2021
Delta	**B.1.617.2**	22	17	30	69	July-August, 2021
	AY.4	12	15	22	49	Sep-Oct, 2021
Omicron	**B.1.1.529**	8	9	4	21	Dec, 2021
	BA.1	12	10	10	32	Dec 2021-Jan 2022
	BA.2	35	32	34	101	Dec 2021-Jan 2022

Table 2: Comparison of within-host diversity between Omicrons and other SARS-CoV-2 lineages.

Multiple comparisons test	**Adjusted p-value**	**Summary**
Alpha vs. Omicron-B.1.1.529	0.0024	*
Alpha vs. Omicron-BA.1	<0.0001	****
Alpha vs. Omicron-BA.2	<0.0001	****
Beta vs. Omicron-B.1.1.529	0.4078	ns
Beta vs. Omicron-BA.1	0.0059	*
Beta vs. Omicron-BA.2	0.0095	*
Delta-B.1.617.2 vs. Omicron-B.1.1.529	0.0039	*
Delta-B.1.617.2 vs. Omicron-BA.1	<0.0001	****
Delta-B.1.617.2 vs. Omicron-BA.2	<0.0001	****
Delta-AY.4 vs. Omicron-B.1.1.529	0.0010	***
Delta-AY.4 vs. Omicron-BA.1	<0.0001	****
Delta-AY.4 vs. Omicron-BA.2	<0.0001	****

Significance is indicated as follows: * for <0.03, ** for <0.0021, *** for < 0.0002 and **** for < 0.0001

Journal Pre-proof

Star Methods

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Hebah A. Al-Khatib (h.alkhatib@qu.edu.qa)

Materials Availability

All sequences generated in this study are publicly available in NCBI website (NCBI BioProject ID PRJNA863945)

Data and Code availability

This paper does not report original code. Data reported in this paper and any additional information required to reanalyze the data will be provided from the lead contact upon request.

Experimental models and subject details

Sample selection criteria

Nasopharyngeal swabs were collected and tested in the virology laboratory at Hamad Medical Corporation, Qatar. Aliquots of viral transport medium of positive samples (Ct value < 25) were transported to be sequenced in the Biomedical Research Center at Qatar University. A representative number of samples were selected from the four SARS-CoV-2 pandemic waves: Alpha (December 2020-March 2021), Beta (February 2021-April 2021), Delta (April 2021-November 2021), and Omicron (December 2021-now) (**Table 1**). To avoid contamination and spillover across variants, samples were selected from the peaks of each wave during which only one lineage was dominating in the country. Samples were selected from patients with no history of SARS-CoV-2 infection. Further, samples were selected from patients aged between 12- and 60-years old with mild to moderate symptoms. These selections were made based on our previous results that displayed higher within-host mutations in elderly (older than 60 years old) and severely ill patients [22]. Also, only samples collected within 1 to 3 days following the onset of symptoms were selected to minimize variations in within-host diversity reported over the course of infection as previously reported [24]. Finally, only samples with Ct values ranging from 18 to 22 were selected. Lythgoe et al (2021) reported that calling within-host mutations at a minimum frequency of 3% is highly reproducible for samples having 50,000 uniquely mapped reads which correspond to a cycle threshold of ~22. Selection of samples based on age range, infection severity and Ct values was done to minimize their known effect on intra-host diversity and focus on the impact of lineage, vaccination status and vaccination type. A total number of 379 samples were selected for deep sequencing analysis (**Table 1**).

Journal Pre-proof

Sample size calculation was performed to determine the number of samples required within each group. However, this number could not be achieved for some study groups due to the following issues:

- Vaccinated individuals infected with Alpha variant. None of Alpha positive cases had received SARS-CoV-2 vaccine. During the Alpha peak that lasted during January and March 2021, the use of vaccine to restricted to high-risk groups and hence it was difficult to find vaccinated individuals who fit the above-mentioned criteria.
- Moderna-vaccinated individuals infected with Beta variant. Pfizer vaccine was the main vaccine used during the circulation of Beta variant (February 2021-May 2021). The majority of Moderna vaccinated individuals had received their first dose of vaccine only and hence were not included in the analysis.
- Omicron-B.1.1.529 samples. This Omicron variant circulated for only 2 weeks in Qatar and hence we could not find enough number of samples.

Journal Pre-proof

Ethical approval

This study was approved by the IRB committees of Qatar Biobank (QF-QBB-RES-ACC-0184).

Extraction and quantification of viral RNA

Viral RNA was extracted from 150 uL of viral transport medium of nasopharyngeal and/or oropharyngeal samples using the MGISP-960 sample preparation system (MGI, China). Viral load was quantified using quantitative polymerase chain reaction (RT-qPCR) using TaqPath COVID-19 Combo Kits (Thermo Fisher Scientific, USA) on an ABI 7500 FAST (Thermo Fisher Scientific, USA) that targets the viral S, N, and ORF1ab gene regions. Viral load was then estimated from Ct values against a standard curve that has been generated using a serially diluted viral RNA control (0 – 200,000 copies/reaction).

Sequencing full-length SARS-CoV-2 genome

Libraries were generated using the CleanPlex SARS-CoV-2 Research and Surveillance panel as described by the manufacturer (Paragon Genomics, China). Briefly, extracted RNA was quantified using a Qubit RNA HS assay kit and 100 ng of viral RNA was used for the reverse transcription step. This was followed by a multiplex PCR reaction using two sets of primer pools to ensure full coverage of the viral genome. All primers sequences used to sequence the variants can be requested as bed files from Paragon Genomics company. A second PCR was then performed to add specific indexes to each sample. PCR products were purified using CleanMag Magnetic Beads (Paragon Genomics, China). Indexed libraries were quantified, normalized, and pooled to generate a final yield of 155 ng of DNA. Pooled libraries were then converted into single-stranded DNA and circularized using the MGIEasy circularization kit (MGI, China) as instructed in the protocol. Circularized DNA was bead-purified and used for DNA nanoball (DNB) generation. DNBs were quantified and at least 800 ng were loaded in the MGI-G50 sequencer. Each sequencing run included 94 samples, negative buffer control, and an RNA extracted from non-COVID-19 patients. All sequencing runs were performed using DNBSEQ-G50RS High-throughput Sequencing Set that includes the large, paired-end flow cell (FCL PE100, MGI).

Analysis of next-generation sequencing data

Analysis of sequence reads was performed using the SARS-CoV-2_Multi-PCR_V1.0 pipeline available at GitHub (https://github.com/MGI-tech-bioinformatics/SARS-CoV-2_Multi-PCR_v1.0). In short, demultiplexed sequence read pairs were trimmed to remove the adaptors and primer sequences using the SOAPnuke filtration toolkit [47]. Trimmed reads were then mapped to the SARS-CoV-2 Reference genome Wuhan-Hu-1 (NC_045512.2) using bwa mem version 1.5.7 as the mapper, and samtools were used for the final analysis [48-50]. Only properly paired reads with insert size <500 bp and with at least 90% sequence identity to the reference were retained. Primer sequences were then masked from mapped reads (BAM files) using fgbio software package [51]. Clean mapped reads were then used for consensus sequence construction and variant calling. Variants were called using Freebayes variant detector tools and were restricted only to positions with a minimum depth of 100, frequency of 60%, and quality of 30 [52]. For analysis of consensus genomes, consensus calls required a minimum of ten uniquely mapped reads per position. Lineages were assigned by the Pangolin web server using the determined consensus genome for each sequenced sample [53,54].

Intra host single nucleotide variants (iSNVs) calling

Full coverage sequences (> 95% coverage) were considered for subsequent within-host diversity analysis. Previous studies have estimated within-host diversity by evaluating the number of within-host single nucleotide variants (iSNVs) occurring above a specific minor allele frequency (MAF) threshold. Here, we evaluated the within-host diversity by counting the number of single-nucleotide variants (iSNVs) in each sample including (i) mutations occurring above the minor allele frequency (MAF) threshold of 5%; (ii) mutations with a minimum sequencing depth threshold of 500 reads; (iii) mutations not occurring in RNA control, and (iv) mutations occurring in coding regions. MAFs were computed at every position using low-frequency variants calling tools: LoFreq and Freebayes, with the default parameters of no indel calling and a maximum pileup depth of 1000000 [55,56]. The ESC and IEDB (https://www.iedb.org) resources were used to investigate the immune escape properties of within-host mutations [45]. Positive selection in each site was estimated using the Bayesian approach, FUBAR (Fast, Unconstrained Bayesian Approximation), which infers nonsynonymous (dN) and synonymous (dS) substitution rates in each position for a given coding alignment assuming a constant selection pressure for each position for all sequences in the alignment.

Statistical analysis

Comparison of within-host mutations between samples of each group and among different groups was determined using the one-way ANOVA followed by a Kruskal–Wallis test (within-group) and post-hoc Dunn's multiple comparisons test (between groups) using GraphPad Prism 9. Linear model analysis was performed to study the relationship between lineage and vaccine type using the R. The R package emmeans was used to calculate the estimated marginal means, the broom package was used to format the linear model results in a readable manner and ggblot for visualization. The full linear model is described in **supp. file 1**. Significance was considered for p-values < 0.05.

Journal Pre-proof

Supplemental Data

Supp. Figure 1: Linear model displaying the interaction effect of lineage and vaccine type on iSNVs prevalence. The linear model was performed using R packages: emmeans for calculating the estimated marginal means, broom package for format the linear model results in a readable manner and ggblot for visualization. Statistical significance was considered for p-values < 0.05. This figure is related to **Figure 3**. The full linear model is described in **supp. file 1.**

Supp. file 1: Full model analysis of lineage and vaccine type interaction. This file is related to **supp. Figure 1** and the "Statistical analysis" section in Star Methods.

Journal Pre-proof

Journal Pre-proof

References

1. Safari, I., InanlooRahatloo, K., and Elahi, E. (2021). Evolution of SARS-CoV-2 genome from December 2019 to late March 2020: Emerged haplotypes and informative Tag nucleotide variations. J Med Virol *93*, 2010-2020. 10.1002/jmv.26553.
2. Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., et al. (2020). On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev *7*, 1012-1023. 10.1093/nsr/nwaa036.
3. Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E.J., Msomi, N., et al. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. Nature *592*, 438-443. 10.1038/s41586-021-03402-9.
4. Singh, J., Rahman, S.A., Ehtesham, N.Z., Hira, S., and Hasnain, S.E. (2021). SARS-CoV-2 variants of concern are emerging in India. Nat Med *27*, 1131-1133. 10.1038/s41591-021-01397-4.
5. Volz, E., Mishra, S., Chand, M., Barrett, J.C., Johnson, R., Geidelberg, L., Hinsley, W.R., Laydon, D.J., Dabrera, G., O'Toole, A., et al. (2021). Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature *593*, 266-269. 10.1038/s41586-021-03470-x.
6. Planas, D., Bruel, T., Grzelak, L., Guivel-Benhassine, F., Staropoli, I., Porrot, F., Planchais, C., Buchrieser, J., Rajah, M.M., Bishop, E., et al. (2021). Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to neutralizing antibodies. Nat Med *27*, 917-924. 10.1038/s41591-021-01318-5.
7. Yadav, P.D., Sapkal, G.N., Ella, R., Sahay, R.R., Nyayanit, D.A., Patil, D.Y., Deshpande, G., Shete, A.M., Gupta, N., Mohan, V.K., et al. (2021). Neutralization of Beta and Delta variant with sera of COVID-19 recovered cases and vaccinees of inactivated COVID-19 vaccine BBV152/Covaxin. J Travel Med *28*. 10.1093/jtm/taab104.
8. Garcia-Beltran, W.F., Lam, E.C., St Denis, K., Nitido, A.D., Garcia, Z.H., Hauser, B.M., Feldman, J., Pavlovic, M.N., Gregory, D.J., Poznansky, M.C., et al. (2021). Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. Cell *184*, 2372-2383 e2379. 10.1016/j.cell.2021.03.013.
9. CDC (2020). SARS-CoV-2 Variant Classifications and Definitions. https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html.
10. Riediker, M., Briceno-Ayala, L., Ichihara, G., Albani, D., Poffet, D., Tsai, D.H., Iff, S., and Monn, C. (2022). Higher viral load and infectivity increase risk of aerosol transmission for Delta and Omicron variants of SARS-CoV-2. Swiss Med Wkly *152*, w30133. 10.4414/smw.2022.w30133.
11. Saxena, S.K., Kumar, S., Ansari, S., Paweska, J.T., Maurya, V.K., Tripathi, A.K., and Abdel-Moneim, A.S. (2022). Transmission dynamics and mutational prevalence of the novel Severe acute respiratory syndrome coronavirus-2 Omicron Variant of Concern. J Med Virol *94*, 2160-2166. 10.1002/jmv.27611.
12. Edara, V.V., Pinsky, B.A., Suthar, M.S., Lai, L., Davis-Gardner, M.E., Floyd, K., Flowers, M.W., Wrammert, J., Hussaini, L., Ciric, C.R., et al. (2021). Infection and Vaccine-Induced Neutralizing-Antibody Responses to the SARS-CoV-2 B.1.617 Variants. N Engl J Med *385*, 664-666. 10.1056/NEJMc2107799.
13. Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., and VandePol, S. (1982). Rapid evolution of RNA genomes. Science (New York, N.Y.) *215*, 1577-1585.

14. Lauring, A.S., and Andino, R. (2010). Quasispecies theory and the behavior of RNA viruses. PLoS pathogens *6*, e1001005. 10.1371/journal.ppat.1001005.

15. Xue, K.S., Moncla, L.H., Bedford, T., and Bloom, J.D. (2018). Within-Host Evolution of Human Influenza Virus. Trends in microbiology *26*, 781-793. 10.1016/j.tim.2018.02.007.

16. Wang, Y., Wang, D., Zhang, L., Sun, W., Zhang, Z., Chen, W., Zhu, A., Huang, Y., Xiao, F., Yao, J., et al. (2021). Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. Genome Med *13*, 30. 10.1186/s13073-021-00847-5.

17. Gaiarsa, S., Giardina, F., Batisti Biffignandi, G., Ferrari, G., Piazza, A., Tallarita, M., Novazzi, F., Bandi, C., Paolucci, S., Rovida, F., et al. (2022). Comparative analysis of SARS-CoV-2 quasispecies in the upper and lower respiratory tract shows an ongoing evolution in the spike cleavage site. Virus Res *315*, 198786. 10.1016/j.virusres.2022.198786.

18. Martin, D.P., Weaver, S., Tegally, H., San, J.E., Shank, S.D., Wilkinson, E., Lucaci, A.G., Giandhari, J., Naidoo, S., Pillay, Y., et al. (2021). The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. Cell *184*, 5189-5200 e5187. 10.1016/j.cell.2021.09.003.

19. Gelbart, M., Harari, S., Ben-Ari, Y., Kustin, T., Wolf, D., Mandelboim, M., Mor, O., Pennings, P.S., and Stern, A. (2020). Drivers of within-host genetic diversity in acute infections of viruses. PLoS pathogens *16*, e1009029. 10.1371/journal.ppat.1009029.

20. Fontanet, A., and Cauchemez, S. (2020). COVID-19 herd immunity: where are we? Nat Rev Immunol *20*, 583-584. 10.1038/s41577-020-00451-5.

21. Zhou, D., Dejnirattisai, W., Supasa, P., Liu, C., Mentzer, A.J., Ginn, H.M., Zhao, Y., Duyvesteyn, H.M.E., Tuekprakhon, A., Nutalai, R., et al. (2021). Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. Cell *184*, 2348-2361 e2346. 10.1016/j.cell.2021.02.037.

22. Al Khatib, H.A., Benslimane, F.M., Elbashir, I.E., Coyle, P.V., Al Maslamani, M.A., Al-Khal, A., Al Thani, A.A., and Yassine, H.M. (2020). Within-Host Diversity of SARS-CoV-2 in COVID-19 Patients With Variable Disease Severities. Front Cell Infect Microbiol *10*, 575613. 10.3389/fcimb.2020.575613.

23. Weigang, S., Fuchs, J., Zimmer, G., Schnepf, D., Kern, L., Beer, J., Luxenburger, H., Ankerhold, J., Falcone, V., Kemming, J., et al. (2021). Within-host evolution of SARS-CoV-2 in an immunosuppressed COVID-19 patient as a source of immune escape variants. Nature communications *12*, 6405. 10.1038/s41467-021-26602-3.

24. Lythgoe, K.A., Hall, M., Ferretti, L., de Cesare, M., MacIntyre-Cockett, G., Trebes, A., Andersson, M., Otecko, N., Wise, E.L., Moore, N., et al. (2021). SARS-CoV-2 within-host diversity and transmission. Science (New York, N.Y.) *372*. 10.1126/science.abg0821.

25. Siqueira, J.D., Goes, L.R., Alves, B.M., de Carvalho, P.S., Cicala, C., Arthos, J., Viola, J.P.B., de Melo, A.C., and Soares, M.A. (2021). SARS-CoV-2 genomic analyses in cancer patients reveal elevated intrahost genetic diversity. Virus Evol *7*, veab013. 10.1093/ve/veab013.

26. Tonkin-Hill, G., Martincorena, I., Amato, R., Lawson, A.R.J., Gerstung, M., Johnston, I., Jackson, D.K., Park, N., Lensing, S.V., Quail, M.A., et al. (2021). Patterns of within-host genetic diversity in SARS-CoV-2. eLife *10*. 10.7554/eLife.66857.

27. McCallum, M., De Marco, A., Lempp, F.A., Tortorici, M.A., Pinto, D., Walls, A.C., Beltramello, M., Chen, A., Liu, Z., Zatta, F., et al. (2021). N-terminal domain

Journal Pre-proof

antigenic mapping reveals a site of vulnerability for SARS-CoV-2. Cell *184*, 2332-2347.e2316. 10.1016/j.cell.2021.03.028.

28. Safari, I., and Elahi, E. (2022). Evolution of the SARS-CoV-2 genome and emergence of variants of concern. Archives of virology *167*, 293-305. 10.1007/s00705-021-05295-5.
29. Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., and Baele, G. (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evol *6*, veaa061. 10.1093/ve/veaa061.
30. Yan, L., Yang, Y., Li, M., Zhang, Y., Zheng, L., Ge, J., Huang, Y.C., Liu, Z., Wang, T., Gao, S., et al. (2021). Coupling of N7-methyltransferase and 3'-5' exoribonuclease with SARS-CoV-2 polymerase reveals mechanisms for capping and proofreading. Cell *184*, 3474-3485 e3411. 10.1016/j.cell.2021.05.033.
31. Shen, Z., Xiao, Y., Kang, L., Ma, W., Shi, L., Zhang, L., Zhou, Z., Yang, J., Zhong, J., Yang, D., et al. (2020). Genomic Diversity of Severe Acute Respiratory Syndrome-Coronavirus 2 in Patients With Coronavirus Disease 2019. Clin Infect Dis *71*, 713-720. 10.1093/cid/ciaa203.
32. Valesano, A.L., Rumfelt, K.E., Dimcheff, D.E., Blair, C.N., Fitzsimmons, W.J., Petrie, J.G., Martin, E.T., and Lauring, A.S. (2021). Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. PLoS pathogens *17*, e1009499. 10.1371/journal.ppat.1009499.
33. Popa, A., Genger, J.W., Nicholson, M.D., Penz, T., Schmid, D., Aberle, S.W., Agerer, B., Lercher, A., Endler, L., Colaco, H., et al. (2020). Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Science translational medicine *12*. 10.1126/scitranslmed.abe2555.
34. Morris, S.K., Parkin, P., Science, M., Subbarao, P., Yau, Y., O'Riordan, S., Barton, M., Allen, U.D., and Tran, D. (2012). A retrospective cross-sectional study of risk factors and clinical spectrum of children admitted to hospital with pandemic H1N1 influenza as compared to influenza A. BMJ Open *2*, e000310. 10.1136/bmjopen-2011-000310.
35. Kelliher, M.T., Levy, J.J., Nerenz, R.D., Poore, B., Johnston, A.A., Rogers, A.R., Stella, M.E.O., Snow, S.E., Cervinski, M.A., and Hubbard, J.A. (2022). Comparison of Symptoms and Antibody Response Following Administration of Moderna or Pfizer SARS-CoV-2 Vaccines. Arch Pathol Lab Med. 10.5858/arpa.2021-0607-SA.
36. Bajema, K.L., Dahl, R.M., Evener, S.L., Prill, M.M., Rodriguez-Barradas, M.C., Marconi, V.C., Beenhouwer, D.O., Holodniy, M., Lucero-Obusan, C., Brown, S.T., et al. (2021). Comparative Effectiveness and Antibody Responses to Moderna and Pfizer-BioNTech COVID-19 Vaccines among Hospitalized Veterans - Five Veterans Affairs Medical Centers, United States, February 1-September 30, 2021. MMWR Morb Mortal Wkly Rep *70*, 1700-1705. 10.15585/mmwr.mm7049a2.
37. Twohig, K.A., Nyberg, T., Zaidi, A., Thelwall, S., Sinnathamby, M.A., Aliabadi, S., Seaman, S.R., Harris, R.J., Hope, R., Lopez-Bernal, J., et al. (2022). Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: a cohort study. The Lancet. Infectious diseases *22*, 35-42. 10.1016/S1473-3099(21)00475-8.
38. Tang, P., Hasan, M.R., Chemaitelly, H., Yassine, H.M., Benslimane, F.M., Al Khatib, H.A., AlMukdad, S., Coyle, P., Ayoub, H.H., Al Kanaani, Z., et al. (2021). BNT162b2 and mRNA-1273 COVID-19 vaccine effectiveness against the SARS-CoV-2 Delta variant in Qatar. Nat Med *27*, 2136-2143. 10.1038/s41591-021-01583-4.

Journal Pre-proof

39. Nanduri, S., Pilishvili, T., Derado, G., Soe, M.M., Dollard, P., Wu, H., Li, Q., Bagchi, S., Dubendris, H., Link-Gelles, R., et al. (2021). Effectiveness of Pfizer-BioNTech and Moderna Vaccines in Preventing SARS-CoV-2 Infection Among Nursing Home Residents Before and During Widespread Circulation of the SARS-CoV-2 B.1.617.2 (Delta) Variant - National Healthcare Safety Network, March 1-August 1, 2021. MMWR Morb Mortal Wkly Rep *70*, 1163-1166. 10.15585/mmwr.mm7034e3.
40. Harder, T., Kulper-Schiek, W., Reda, S., Treskova-Schwarzbach, M., Koch, J., Vygen-Bonnet, S., and Wichmann, O. (2021). Effectiveness of COVID-19 vaccines against SARS-CoV-2 infection with the Delta (B.1.617.2) variant: second interim results of a living systematic review and meta-analysis, 1 January to 25 August 2021. Euro Surveill *26*. 10.2807/1560-7917.ES.2021.26.41.2100920.
41. ourworldindata. Coronavirus (COVID-19) Vaccinations. https://ourworldindata.org.
42. Butt, A.A., Khan, T., Yan, P., Shaikh, O.S., Omer, S.B., and Mayr, F. (2021). Rate and risk factors for breakthrough SARS-CoV-2 infection after vaccination. J Infect *83*, 237-279. 10.1016/j.jinf.2021.05.021.
43. Teran, R.A., Walblay, K.A., Shane, E.L., Xydis, S., Gretsch, S., Gagner, A., Samala, U., Choi, H., Zelinski, C., and Black, S.R. (2021). Postvaccination SARS-CoV-2 Infections Among Skilled Nursing Facility Residents and Staff Members - Chicago, Illinois, December 2020-March 2021. MMWR Morb Mortal Wkly Rep *70*, 632-638. 10.15585/mmwr.mm7017e1.
44. Hernandez-Ortiz, J., Cardona, A., Ciuoderis, K., Averhoff, F., Maya, M.A., Cloherty, G., and Osorio, J.E. (2022). Assessment of SARS-CoV-2 Mu Variant Emergence and Spread in Colombia. JAMA Netw Open *5*, e224754. 10.1001/jamanetworkopen.2022.4754.
45. Rophina, M., Pandhare, K., Shamnath, A., Imran, M., Jolly, B., and Scaria, V. (2022). ESC: a comprehensive resource for SARS-CoV-2 immune escape variants. Nucleic Acids Res *50*, D771-D776. 10.1093/nar/gkab895.
46. Weisblum, Y., Schmidt, F., Zhang, F., DaSilva, J., Poston, D., Lorenzi, J.C., Muecksch, F., Rutkowska, M., Hoffmann, H.H., Michailidis, E., et al. (2020). Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. Elife *9*. 10.7554/eLife.61312.
47. Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., Zhang, X., et al. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience *7*, 1-6. 10.1093/gigascience/gix120.
48. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) *25*, 1754-1760. 10.1093/bioinformatics/btp324.
49. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics (Oxford, England) *27*, 2156-2158. 10.1093/bioinformatics/btr330.
50. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*. 10.1093/gigascience/giab008.
51. fgbio software package. https://github.com/fulcrumgenomics/fgbio.
52. Garrison, E.a.M., G. (2012). Haplotype-based variant detection from short-read sequencing.
53. O'Toole, A., Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J.T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B., et al. (2021). Assignment of

Journal Pre-proof

epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol *7*, veab064. 10.1093/ve/veab064.

54. Rambaut, A., Holmes, E.C., O'Toole, A., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., and Pybus, O.G. (2021). Addendum: A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol *6*, 415. 10.1038/s41564-021-00872-5.

55. Wilm, A., Aw, P.P., Bertrand, D., Yeo, G.H., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., and Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res *40*, 11189-11201. 10.1093/nar/gks918.

56. Garrison, E.a.M., G. (2012). Haplotype-based variant detection from short-read sequencing.

Journal Pre-proof

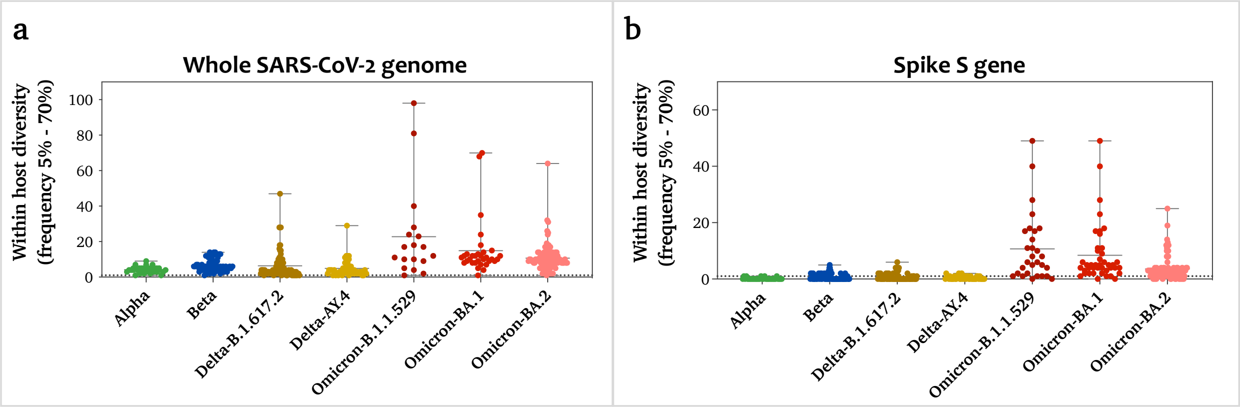


Figure 1: Number of iSNVs (MAF > 0.05) observed in each sample of the SARS-CoV-2 lineages/sub-lineages. (**a**) Number of iSNVs seen in the whole genome of each sample. (**b**) Number of iSNVs seen in the spike gene of each sample. All mutation were called with respect to Wuhan Hu-1 reference sequence (RefSeq ID NC_045512). Data represent the mean of the mean number of iSNVs ± SD reported among samples that belong to the same lineage/sub-lineage. All samples (n= 379) in our dataset are included in this figure. Number of samples within each group is listed in **Table 1**.

Journal Pre-proof

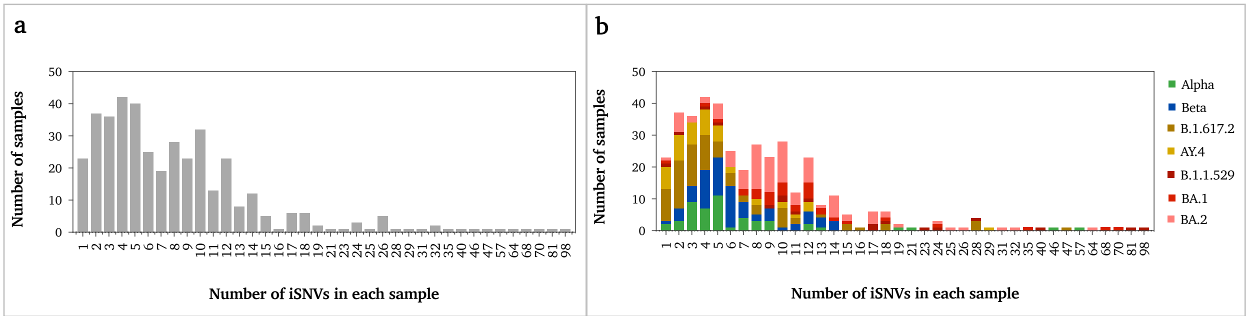


Figure 2: Histograms showing the number of samples exhibiting N number of iSNVs (MAF > 0.05). (a) Histogram showing the total number of samples with N number of iSNVs. **(b)** Stacked histogram showing the number of samples that had N number of iSNV sub-categorized based on lineage: Alpha, Beta, Delta and Omicron lineages and sub-lineages. All identified sites were included in this figure except for those located in intragenic regions and in the upper and lower untranslated regions of the genome.

Journal Pre-proof

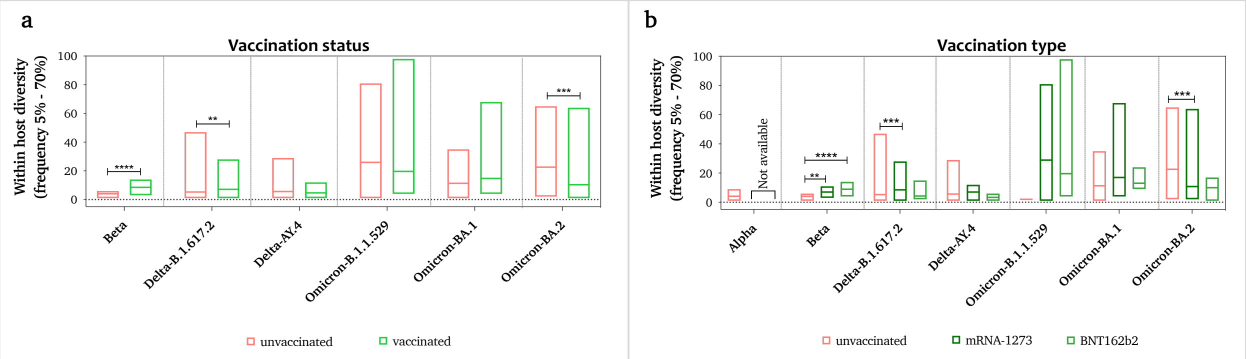
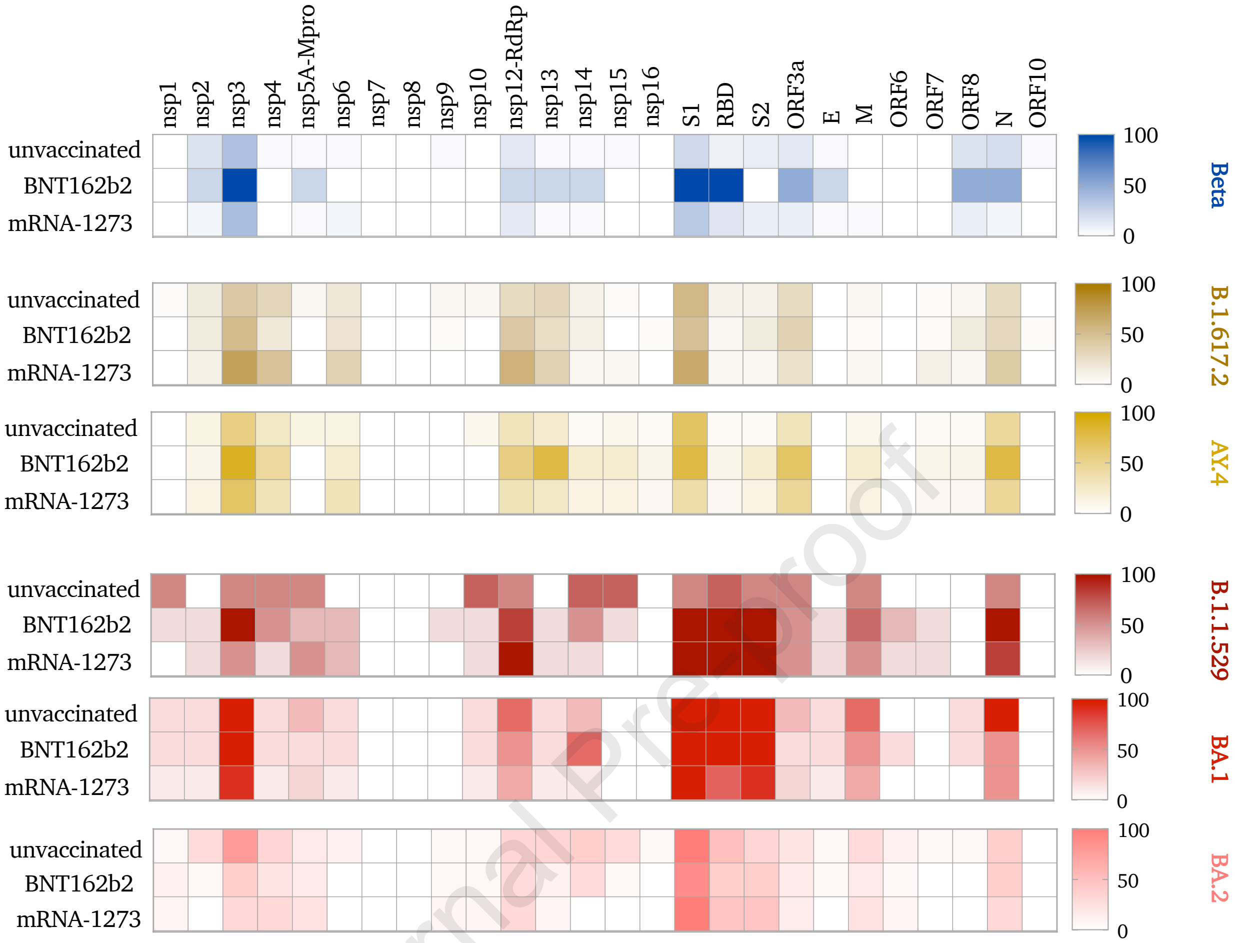


Figure 3: Average number of iSNVs observed in SARS-CoV-2 positive samples. Number of iSNVs seen in each sample of the SARS-CoV-2 lineages sub-divided based on **(a)** vaccination status: vaccinated and unvaccinated; and **(b)** mRNA vaccine type: mRNA-1273 (Moderna) and BNT162b2 (Pfizer). All mutations were included except for those located in intragenic regions and in the upper and lower untranslated regions of the genome. Significance is indicated as follows: * for <0.033, ** for <0.0021, *** for < 0.0002 and **** for < 0.0001. A detailed linear model analysis that incorporates the interaction between the virus lineage and vaccine type is demonstrated in **supp. Figure 1 and supp file 1**.

Journal Pre-proof



Journal Pre-proof

Figure 4: **Heatmap demonstrating the distribution of iSNVs throughout SARS-CoV-2 genome of each lineage.** All identified intra-host variation sites (MAF>0.05) were included in the heatmap analysis except for those located in intragenic regions and the upper and lower untranslated regions of the genome. The range on the right demonstrates the number of intra-host variations sites in each region of the genome where 0 indicates no mutations while 100 indicates that 100 mutations were found in this region.

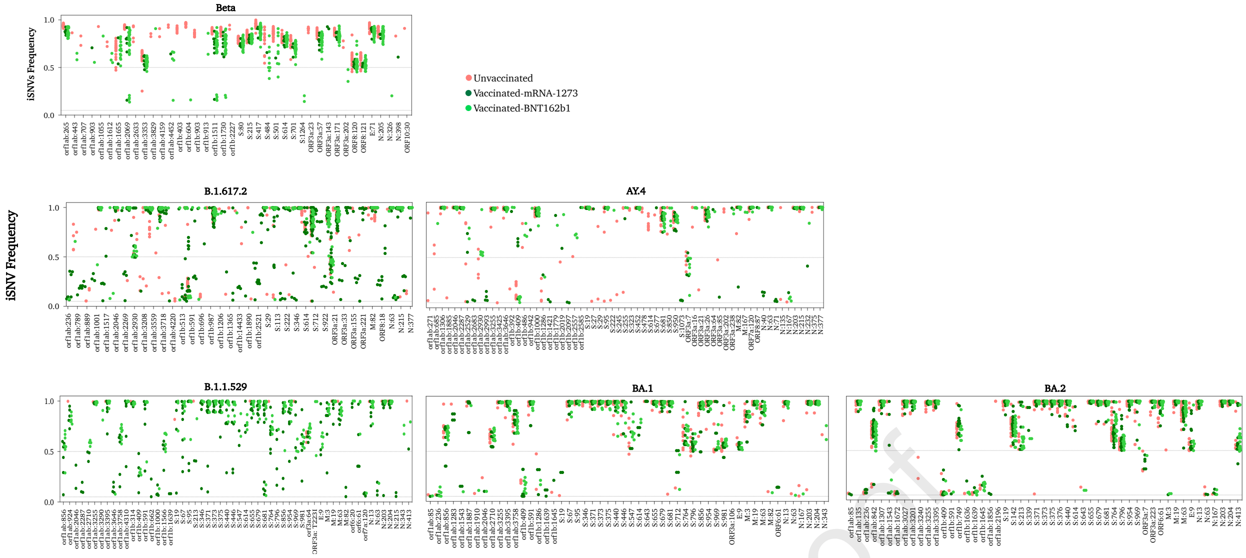


Figure 5: Distribution and frequency of the non-synonymous mutations (MAF> 0.05) across SARS-CoV-2 genomes of each lineage.

Journal Pre-proof

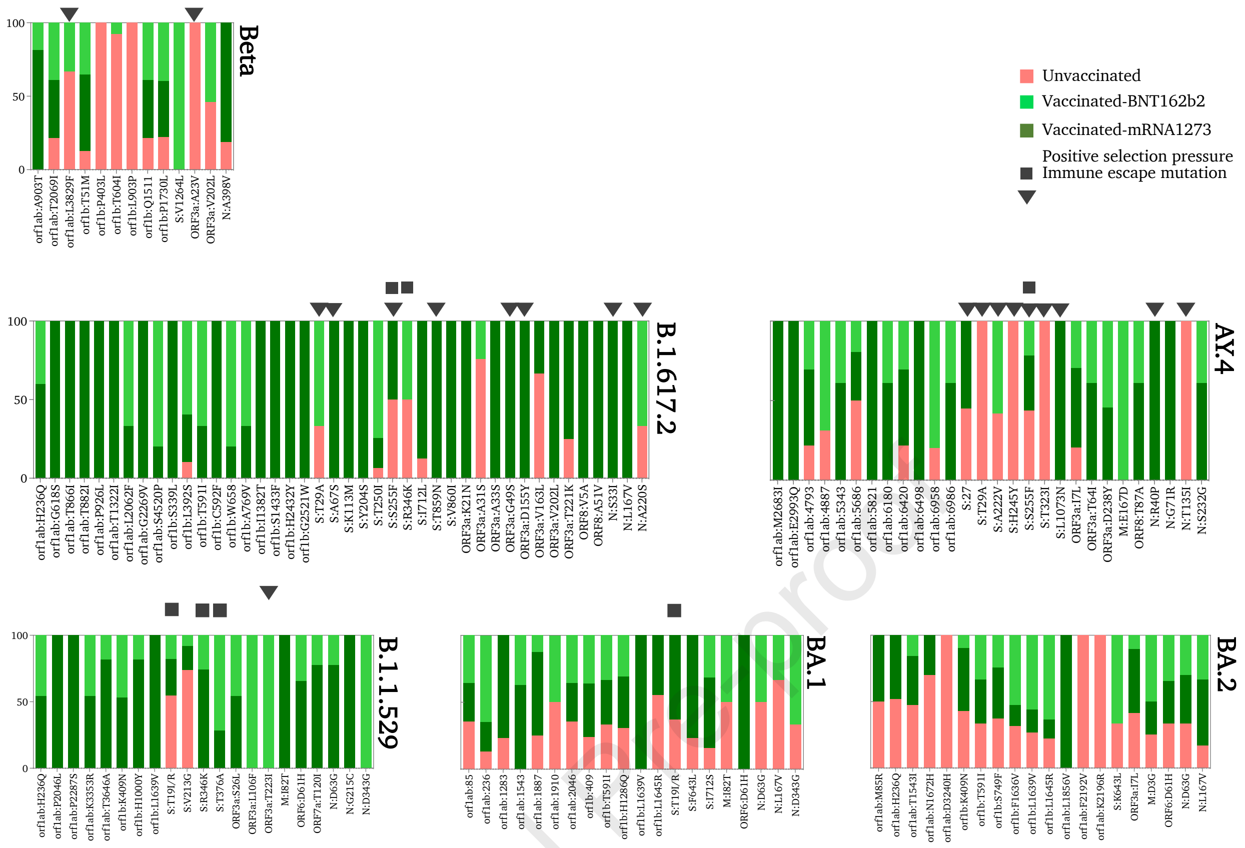


Figure 6: Prevalence of non-lineage mutations identified in vaccinated and unvaccinated individuals of each lineage. This figure displays non-synonymous mutations that showed significant differences in their prevalence among the three groups: unvaccinated, Moderna-vaccinated and Pfizer-vaccinated individuals. The squares and triangles above the bars indicate immune escape mutations and positively selected site, respectively.

Journal Pre-proof

Comparative analysis of within-host diversity among vaccinated COVID-19 patients infected with different SARS-CoV-2 variants

Highlights

- Higher within-host diversity among omicron positive samples.
- Higher within-host diversity among vaccinated individuals regardless of virus lineage.
- Limited impact of vaccine types on within-host diversity of SARS-CoV-2.

Journal Pre-proof

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
CleanPlex SARS-CoV-2 Research and Surveillance panel	Paragon Genomics	Cat#918002
CleanPlex for MGI Single-Indexed PCR Primers	Paragon Genomics	Cat#318007
CleanMag Magnetic Beads	Paragon Genomics	Cat# 718003
MGIEasy circularization kit	MGI	Cat#1000005259
DNBSEQ-G50RS High-throughput Sequencing Set	MGI	Cat# 1000019859
Qubit RNA HS assay kit	Invitrogen	Cat# Q32852
Deposited data		
NCBI BioProject ID PRJNA863945	This paper	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA863945
Software and algorithms		
SARS-CoV-2_Multi-PCR_V1.0 pipeline	MGI Tech bioinformatics	https://github.com/MGI-tech-bioinformatics/SARS-CoV-2_Multi-PCR_v1.0
SOAPnuke filtration toolkit (version 2.0.6)	Chen at al 2018[1]	https://github.com/BGI-flexlab/SOAPnuke
BWA (version 0.7.17)	Li 2013[2]	https://github.com/lh3/bwa
Samtools (version 1.7)	Danecek et al 2021[3]	https://github.com/samtools/samtools
Fgbio software package		https://github.com/fulcrumgenomics/fgbio
LoFreq (version 2)	Wilm et al 2012[4]	https://csb5.github.io/lofreq/
GraphPad Prism 9.0		www.graphpad.com
R (version 4.2.1)		www.R-project.org

Journal Pre-proof

References

1. Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., Zhang, X., et al. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. Gigascience *7*, 1-6. 10.1093/gigascience/gix120.
2. Li (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
3. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*. 10.1093/gigascience/giab008.
4. Wilm, A., Aw, P.P., Bertrand, D., Yeo, G.H., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., and Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res *40*, 11189-11201. 10.1093/nar/gks918.

Journal Pre-proof