Within-host diversity of SARS-CoV-2 lineages and effect of vaccination

Leo Poon (✉ llmpoon@hku.hk)
University of Hong Kong https://orcid.org/0000-0002-9101-7953
Haogao Gu
The University of Hong Kong https://orcid.org/0000-0002-7541-4262
Ahmed Abdul Quadeer
Hong Kong University of Science and Technology https://orcid.org/0000-0002-5295-9067
Pavithra Krishnan
University of Hong Kong
Lydia Chang
University of Hong Kong
Gigi Liu
University of Hong Kong
Daisy Ng
University of Hong Kong
Samuel Cheng
The University of Hong Kong
Tommy Tsan-Yuk Lam
The University of Hong Kong https://orcid.org/0000-0002-9769-1527
Malik Peiris
University of Hong Kong https://orcid.org/0000-0001-8217-5995
Matthew McKay
University of Hong Kong

Article

Keywords:

Posted Date: August 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1927944/v1

Within-host diversity of SARS-CoV-2 lineages and effect of vaccination

Haogao Gu[1], Ahmed Abdul Quadeer[2], Pavithra Krishnan[1], Daisy Y.M. Ng[1], Lydia D.J Chang[1], Gigi Y.Z. Liu[1], Samuel S.M. Cheng[1], Tommy T.Y. Lam[1,3,4], Malik Peiris[1,3,5], Matthew R. McKay[2,6,7,8], Leo L.M. Poon[1,3,5*]

[1] School of Public Health, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China.

[2] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

[3] Centre for Immunology & Infection, Hong Kong Science and Technology Park, Hong Kong, China.

[4] Laboratory of Data Discovery for Health, Hong Kong Science and Technology Park, Hong Kong, China.

[5] HKU-Pasteur Research Pole, School of Public Health, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China.

[6] Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, VIC 3010, Australia

[7] Department of Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, VIC 3000, Australia

[8] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

[*]To whom correspondence should be addressed. Email: llmpoon@hku.hk.

Word count (abstract): 148

Word count (main text excluding methods): 3990

Abstract

Viral and host factors can shape SARS-CoV-2 within-host viral diversity and virus evolution. However, little is known about lineage-specific and vaccination-specific mutations that occur within individuals. Here we analysed deep sequencing data from 2,146 SARS-CoV-2 samples with different viral lineages to describe the patterns of within-host diversity in different conditions, including vaccine-breakthrough infections. Variant of Concern (VOC) Alpha, Delta, and Omicron samples were found to have higher within-host nucleotide diversity while being under weaker purifying selection at full genome level compared to non-VOC SARS-CoV-2 viruses. Breakthrough Delta and Omicron infections in Comirnaty and CoronaVac vaccinated individuals appeared to have higher within-host purifying selection at the full-genome and/or Spike gene levels. Vaccine-induced antibody or T cell responses did not appear to have significant impact on within-host SARS-CoV-2 evolution. Our findings suggest that vaccination does not increase SARS-CoV-2 protein sequence space and may not facilitate emergence of more viral variants.

Introduction

The SARS-CoV-2 pandemic continues to spread globally. Despite vaccination of over 60% of the world population[1], the risk of SARS-CoV-2 reinfections and breakthrough infections is increasing due to the emergence of new viral variants[2,3]. Multiple variants of concern (VOC) have demonstrated the ability to evade naturally-acquired or vaccine-induced immunity[4-6]. Therefore, it is crucial to investigate the impact of vaccination on mutational and evolutionary processes of SARS-CoV-2.

Genomic surveillance has been used to trace the transmission and evolution of SARS-CoV-2 mutations at local, regional, and global scales throughout the pandemic[7-9]. However, there is still limited knowledge of how these mutations originate and accumulate within hosts. Within-host mutations can arise through replication errors or RNA damage/editing[10] and they may be subject to fixation by stochastic (genetic drift) and deterministic (natural selection) processes. We and others have previously found that the SARS-CoV-2 transmission bottleneck between hosts is narrow[8,11-14], suggesting that only few virions are transferred from the host during transmission. Most of the low-frequency mutations are not transmitted between patients, which constrains the use of intrahost single nucleotide variants (iSNVs) for effective contact tracing[12,15,16]. However, it remains important to investigate within-host diversity of SARS-CoV-2 to understand host-level evolutionary forces.

Studying SARS-CoV-2 within-host diversity under different conditions may reveal factors that control virus evolution. Host and viral factors can both contribute to within-host diversity. Host factors such as species (animals/humans)[17], viral shedding time[18], and immune status[19] were previously reported to have effects on intrahost SARS-CoV-2 diversity. It was hypothesized that prolonged infections in hosts with distinct immunological backgrounds (e.g., animals or immunocompromised patients) may hasten viral evolution and lead to emergence of novel variants[17,20]. However, there is limited knowledge about post-vaccination characteristics of within-host selection pressures, which consistently act on the virus during the entire course of breakthrough infection. Besides, viral factors such as different virus lineages may also affect SARS-CoV-2 replication properties. SARS-CoV-2 VOCs have exhibited varying capacities to evade immunity[4,6] and acquire higher transmissibility[21,22]. However, it is not clear whether different SARS-CoV-2 variants differ in within-host selection pressures.

To address these knowledge gaps, we analysed 2,146 deep sequenced SARS-CoV-2 samples collected in Hong Kong (HK) between mid-2020 and early 2022. The within-host diversity in SARS-CoV-2 infections from different lineages (VOCs: B.1.1.7 (Alpha), B.1.617.2 (Delta), B.1.1.529 (Omicron) and non-VOCs: B.1.36, B.1.36.27, and B.1.1.63) and in breakthrough (Delta or Omicron) infections after Comirnaty or CoronaVac vaccination were studied. Our results provide insights into the variation of within-host diversity, and the mutational patterns and potential selection pressures acting on viruses.

Results

Diversity of within-host mutations in SARS-CoV-2 samples

We analysed 2,146 SARS-CoV-2 samples from 2,053 different individuals, of whom 86 had multiple samples (totalling 2,053 representative samples and 93 repeated samples for technical control). The samples were collected from July-2020 to March-2022, covering the third to the fifth COVID-19 waves in HK. All samples had genome coverage ≥90% with >100 reads of sequencing depth (Supplementary Fig. 1) and with high viral loads (Ct value ≤25). The median sequencing depth ranged from 380 to 98,214 per sample. The samples belonged to major SARS-CoV-2 lineages including VOCs (B.1.1.7 (Alpha), B.1.617.2 (Delta), B.1.1.529 (Omicron, 20% are BA.1 and 80% are BA.2)) and non-VOCs (B.1.36, B.1.36.27, and B.1.1.63). The three non-VOC lineages were detected in the third (B.1.1.63) and fourth (B.1.36 and B.1.36.27) COVID-19 waves when no COVID-19 vaccine was available for use in HK[9]. For Delta and Omicron infections, samples from breakthrough infections after Comirnaty or CoronaVac (two-dose) vaccination were also included.

For reliable analysis of within-host mutations, quality filtering steps were developed and validated using technical control samples (see Methods). We identified 2,731 iSNVs, with allele frequency between 5% to 50%, at 2,058 sites from 1,117 (54.4%) samples. Of these iSNVs, 1,694 (62.0%) of them were nonsynonymous, 1,016 (37.2%) were synonymous, and 21 (0.8%) were located in untranslated regions (mutations in the heading and tailing 100bp regions were excluded in this study). We did not identify detectable iSNVs in the other 936 (45.6%) samples. Overall, the mean number of iSNVs per sample was 1.33 (Fig. 1A, dashed line). This iSNV detection rate is similar to a previously reported level[11,16], but lower than some other studies[15,23], presumably due to differences in variant filtering criteria. Of the iSNV sites, 1,801 (88%) were uniquely observed in a single patient sample. This suggests that most iSNVs are sporadic mutations occurring at distinct positions rather than recurrent mutations occurring at specific mutation hotspots (Fig. 1B).

We found a weak correlation between viral load (Ct value) and the number of iSNVs per Kb. Samples with higher viral load (lower Ct value) generally had less iSNVs (Fig. 1A and Supplementary Fig. 2A). However, while viral load decreased with detection lag (time since symptom onset) (Supplementary Fig. 2B), the correlation between detection lag and number of iSNVs was not found to be significant (Supplementary Fig. 2C). We also found that the viral load did not significantly correlate with minor allele frequency (MAF) (Supplementary Fig. 2D). Consistent with other studies, these results (Supplementary Fig. 2A-2D) suggest that enrichment of iSNVs negatively correlates with viral load[11,12,15,16], but varies less with time from symptom onset[16]. To avoid artefacts due to low viral load[16], we only included samples with a Ct value ≤25 and adjusted the number of iSNVs per Kb (referred to as incidence of iSNVs hereafter) by linear regression functions (see Methods) in the downstream comparative analysis. With these adjustments, the correlation of Ct value and number of iSNVs per Kb became insignificant (Supplementary Fig. 2E). We found the mean number of iSNVs per Kb

to be 0.045 across the full genome and the highest incidence of iSNVs were found in the ORF8 and ORF7a genes (Supplementary Table 1).

Consistent with previous reports[23,24], we found some mutation types (C→U, G→A, A→G, U→C, and G→U) occurred with higher-than-average frequencies, measured in terms of the number of synonymous/nonsynonymous iSNVs per synonymous/nonsynonymous site (i.e., d_S and d_N) (Fig. 1C; points above the dashed lines). The high frequency of C→U/G→A and A→G/U→C mutants support the hypothesis of RNA editing *in vivo* via APOlipoprotein B Editing Complex (APOBEC) and Adenosine Deaminase Acting on RNA (ADAR) enzymes[10,25], respectively. Interestingly, we observed a higher mutation frequency of G→U, but a lower mutation frequency of C→A, which suggests a strand bias of the G→U mutation. The G→U mutation may be associated to Reactive Oxygen Species (ROS)-related processes[26]. In different regions of the SARS-CoV-2 genome, we observed uneven d_S and d_N ($P<0.001$, Kruskal-Wallis rank sum test among regions with length of 1Kb) and the highest frequency was found in d_S in the Spike gene region (genomic position from 24000 to 25000 in Fig. 1D). For all regions, the number of synonymous iSNVs per Kb per synonymous site (d_S per Kb) are higher than the average number of nonsynonymous iSNVs per Kb per non-synonymous site (d_N per Kb) (Fig. 1D). There were some shared iSNVs, i.e., found in multiple samples from different patients, with five of them (labelled in Fig. 1E) observed in more than 20 samples (frequency>1%). These five high-frequent iSNVs were found in samples from more than one SARS-CoV-2 lineage and under different vaccination statuses, suggesting mutation homoplasy rather than shared mutations from direct transmissions (Supplementary Table 2).

VOC samples exhibit higher within-host diversity but weaker purifying selection than non-VOC samples

To study the within-host diversity between different groups, i.e., SARS-CoV-2 lineages for different vaccination statuses, we calculated the incidence of iSNVs (adjusted number of iSNVs per Kb), abundance of iSNVs (MAF for iSNVs), and nucleotide diversity (π, average number of nucleotide differences per site between pairwise reads)[27] for samples within each group (see Methods). Combinational use of the three complementary indices can help illustrate viral mutant spectrum dynamics[28]. Essentially, incidence of iSNVs correspond to counts of mutational sites in a sample (the breadth of the mutant spectrum), abundance of iSNVs reflects the mutational frequency of each site in the sample (the height/intensity of the mutant spectrum), and nucleotide diversity (π) is a functional index based on the total pairwise difference among observed haplotypes (the degree of polymorphism of iSNVs within a sample). Nucleotide diversity (π) can be further characterized as synonymous and nonsynonymous nucleotide diversity (π_S and π_N) in coding regions. In general, excess nonsynonymous polymorphism ($\pi_N > \pi_S$) points to diversifying/positive selection while excess synonymous polymorphism ($\pi_N < \pi_S$) indicates purifying selection. Relatively weak selection forces are observed when stochastic changes (genetic drift) dominate ($\pi_N \approx \pi_S$)[29].

Lineage-specific effects on iSNVs can be characterized by comparing unvaccinated samples between lineages. We found Delta samples without vaccination (designated "unvaccinated Delta samples") had higher incidence of iSNVs than samples from the non-VOC lineages (medians: 0.002 for Delta vs. -0.022, -0.014 and -0.023 for B.1.1.63, B.1.36 and B.1.36.27 respectively, P<0.05; Fig. 2A and Supplementary Table 3). The unvaccinated Omicron samples also had higher incidence of iSNVs than unvaccinated B.1.1.63 samples. Notably, the median incidence of iSNVs for VOC lineages (Alpha, Delta and Omicron) are all higher than non-VOC lineages (Fig. 2A), suggesting different genetic backgrounds of viruses had different within-host mutation rates. No significant difference was observed between abundance of iSNVs across unvaccinated VOC and non-VOC samples (Fig. 2B).

Similar to what was observed for the incidence of iSNVs, the nucleotide diversity in Delta samples was significantly higher than for samples from all three non-VOC lineages (Fig. 2C and Supplementary Table 3). The overall nucleotide diversity for unvaccinated Omicron and Alpha samples were statistically significantly higher than samples from the third local wave lineage B.1.1.63 samples (P<0.05, Fig. 2C and Supplementary Table 3). Overall, unvaccinated VOC samples had higher median nucleotide diversity compared to the non-VOC samples (Fig. 2C), suggesting infection with VOCs may induce greater within-host genetic variation.

We found evidence of significant purifying selection in the SARS-CoV-2 genome (top row, Full genome column in Fig. 2D) and most samples have excess synonymous polymorphisms ($\pi_N < \pi_S$, P<0.001 by two-sided Wilcoxon rank sum test). The mean value of $\pi_N - \pi_S$ is -1.18×10^{-5} ($\pi_N/\pi_S = 0.56$) across the full genome for all samples, which is consistent with previous reports[11] ($\pi_N/\pi_S = 0.55$), but differs from what was observed in other mammalian samples[17].

For the unvaccinated non-VOC (B.1.1.63, B.1.36 and B.1.36.27) samples, purifying selection was observed at the full-genome level (Full genome panel in Fig. 2D). By contrast, all three unvaccinated VOC (Alpha, Delta and Omicron) samples had overall unbiased selection ($\pi_N \approx \pi_S$) at the full genome level which is statistically indistinguishable from neutrality. At the individual gene level, evidence for positive selection was observed for the Spike gene in both unvaccinated Alpha and Delta samples ($\pi_N - \pi_S = 4.05 \times 10^{-5}$and 1.87×10^{-5}, column S in Fig. 2D and Supplementary Table 4). However, unvaccinated non-VOC samples generally showed neutral to purifying selection in the Spike gene ($\pi_N - \pi_S = -8.00 \times 10^{-5}, -5.21 \times 10^{-5}$ and -0.25×10^{-5}). This result suggests that, compared to viruses from non-VOCs lineages, those from VOC lineages are under less purifying selection pressure at the within-host level.

For other coding regions, our data suggests little evidence of lineage-specific changes in selection pressure. Neutral or purifying selection was generally observed. For example, in ORF1ab, in all cases the synonymous nucleotide diversity is higher than or similar to the non-synonymous nucleotide diversity (Fig. 2E, ORF1ab column). Possible positive selection was

observed in ORF3a in unvaccinated Delta samples (P<0.05, Fig. 2E), E in unvaccinated Alpha samples (P<0.1, Fig. 2E), and ORF7a in unvaccinated B.1.1.63 samples (P<0.1, Fig. 2E).

Vaccination appears to increase the within-host mutation rate and purifying selection pressure on VOC samples

The incidence of iSNVs and nucleotide diversity may also be affected by vaccination. By studying the samples of breakthrough infections from fully vaccinated (with two-doses of Comirnaty or CoronaVac vaccines) patients, we found that the incidence of iSNVs in Comirnaty Delta virus samples was significantly higher than that from the unvaccinated Delta samples (Fig. 3A) and the Comirnaty Omicron samples (Supplementary Fig. 3A). Within Delta samples, higher incidence of iSNVs in Comirnaty samples compared to unvaccinated samples suggests vaccine-specific effects on within-host mutation rate. However, a similar effect was not observed for Omicron samples (Fig. 3A). One possible explanation for the difference between Delta and Omicron samples could be the waning of vaccine effectiveness, as overall a longer time had passed since receiving the second dose for Omicron-infected vaccinated patients in our data (Supplementary Fig. 4). It is also possible that different levels of immune evasion between Omicron and Delta infections may play a role, since neutralizing antibody titers induced by the Comirnaty vaccine against Omicron were lower than those against Delta[30]. Unlike incidence of iSNVs (Fig. 3A), abundance of iSNVs was similar across vaccinated and unvaccinated samples (Fig. 3B), while nucleotide diversities (π) were only marginally significantly higher (P<0.1) in Comirnaty Delta samples compared to unvaccinated samples (Fig. 3C), suggesting that the overall level of genetic variation was not markedly increased by vaccination.

We found elevated purifying selection pressure at the full-genome level from Comirnaty vaccination, where significant $\pi_N < \pi_S$ were observed in vaccinated samples but not for unvaccinated samples (Fig. 3D, Full genome column). The enhanced purifying selection in Comirnaty Delta and Omicron samples was mainly contributed by increased π_S (Supplementary Table 4). At the Spike gene level, the significant positive selection on the unvaccinated Delta samples was not observed in vaccinated Delta samples ($\pi_N - \pi_S = 1.87 \times 10^{-5}$, 0.09×10^{-5} and -2.58×10^{-5} for unvaccinated Delta, Comirnaty Delta and CoronaVac Delta, respectively; column S in Fig. 3D, and Supplementary Table 4). While the selection pressure on the Spike gene in unvaccinated Omicron samples ($\pi_N - \pi_S = -0.97 \times 10^{-5}$) was not significantly different from neutrality, the purifying selection was moderately significant in those with Comirnaty or CoronaVac vaccination ($\pi_N - \pi_S = -3.94 \times 10^{-5}$ and -6.11×10^{-5} for Comirnaty and CoronaVac, respectively, P<0.1). Collectively, Comirnaty vaccination may increase synonymous nucleotide diversity and thereby purifying selection pressures on Delta and Omicron viruses at the full genome level. For the Spike gene, the observed positive/neutral selection pressures acting on Delta and Omicron samples could be shifted to neutral/purifying selection in those with CoronaVac or Comirnaty vaccination. Similar to the lineage-specific results, we did not find consistent vaccination-specific changes in selection pressure for other coding regions. Neutral or purifying selection was predominant

(Fig. 3E), with possible positive selection observed in the M gene in CoronaVac Omicron samples ($P<0.1$).

Positive selection in coding regions of VOC-specific and vaccination-specific samples (Fig. 2E and 3E) suggests diversifying mutations that can potentially lead to higher chance of phenotypic changes. To identify putative hotspot regions with excessive positive selection, we analysed sliding windows (size of 30 codons) across each protein-coding region. Consistent to the results above, we found most genomic regions were under purifying selection. Seven candidate targets of positive selection were found in ORF1ab, ORF7a and E (Supplementary Fig. 5 and Supplementary Table 5). Of these, three regions (nsp3:448-451, nsp15:278-279 and E:40-47) had partial overlap with the regions determined to be under positive selection in an independent study[24].

No significant selection on within-host mutations from immune pressure

To investigate whether the within-host mutations detected in our vaccinated samples enable immune escape, we studied the overlaps of identified within-host mutations with known neutralizing antibody (nAb) escape mutations in the Spike gene and with experimentally-determined T cell epitopes across the full genome.

Although we found mutations on the receptor-binding domain (RBD) and near the S1/S2 cleavage site (e.g., R683L), the overall mutations did not significantly cluster in any specific regions of the Spike gene (Fig. 4A, 4B). The total number of RBD mutations seem to be higher in Comirnaty Omicron samples (6 RBD mutations in 68 Comirnaty Omicron samples vs. zero mutation in 30 unvaccinated Omicron samples, Fig. 4A), however this difference was not significant ($P=0.25$, Chi-squared Test). Except for the K386E and N448K mutations found in two different Comirnaty Omicron samples (Fig. 4A), which may have mild effects on antibody escape (Supplementary Fig. 6A), the other identified mutations in the RBD region in all vaccinated Omicron and Delta samples were not on key antigenic sites (Supplementary Fig. 6A and 6B). For the NTD region, except for the A262T mutation found in one Comirnaty Delta sample, none of the other within-host mutations overlapped with the known NTD antigenic supersite[31] or with mutations that have been reported to affect neutralization of NTD-targeting nAbs[32,33].

In addition to nAbs escape mutations, T cell escape mutants have been shown to be selected under immune pressure in infections from influenza viruses[34,35]. However, the relationship between within-host mutations and T cell responses induced by SARS-CoV-2 infection or vaccination remains largely unknown. To investigate whether the variation in samples from breakthrough infections are related to host T cell responses, we studied the overlap between within-host mutations (minor allele variants) and known T cell epitopes. A total of 1324 $CD8^+$-specific and 961 $CD4^+$-specific T cell epitope-HLA (human leukocyte antigen) pairs were compiled (Methods). The distributions of these epitope-HLA pairs across SARS-CoV-2 proteins and across HLAs are shown in Supplementary Fig. 7. Considering T cell epitopes

across all proteins, the average number of overlapping $CD8^+$ and $CD4^+$ epitopes per mutation was generally similar between different groups (Supplementary Fig. 8). Focusing on vaccinated and unvaccinated samples, we observed no significant difference in the number of overlapping epitopes per mutation (Fig. 5A), which is suggestive of no T cell-based selection on within-host viral evolution. When limited to iSNVs within the Spike gene, a marginally higher number of overlapping $CD4^+$ T cell epitopes was found in Comirnaty Omicron samples compared to unvaccinated Omicron samples, but the difference was not significant (P=0.09, Supplementary Fig. 9A).

Since the samples were sequenced from HK cases, we repeated the above analysis while focusing on the epitopes associated with HLAs prevalent in the HK population (Supplementary Fig. 10, Methods). As for the above results (Fig. 5A and Supplementary Fig. 9A), we did not observe a significant difference in the number of overlapping $CD8^+$ and $CD4^+$ T cell epitopes per mutation between the vaccinated and unvaccinated samples in the full genome (Fig. 5B) or in the Spike gene (Supplementary Fig. 9B).

While the two candidate regions of positive selection mentioned in the previous section (nsp3:448-451 and E:40-47) overlapped with many $CD8^+$ T cell epitopes (N=8 and N=5), these associations did not reach statistical significance (Supplementary Table 5). Overall, we did not identify a surge of antibody escape mutations in any group, and different groups had a similar level of mutation rates in T cell epitope regions.

Discussion

In this study we have analysed Illumina amplicon data from 2,146 SARS-CoV-2 samples to estimate intra-host variation of SARS-CoV-2 under different conditions. Similar to earlier studies, we show that incidence of iSNVs in SARS-CoV-2 samples is low (0 to 2 iSNVs per sample)[11,16] and that sample viral loads negatively correlate with within-host mutation rates[11,12,15,16], which suggests low viral load specimens are prone to bias toward falsely high iSNVs rates. In agreement with reports from Tonkin-Hill et al.[15] where SARS-CoV-2 samples with lower Ct value show good concordance in allele frequencies between replicates, we also found the cut-off of Ct $\leq$25 can avoid most outliers. Evidence of RNA editing at the full genome level, e.g., the widely reported biased C→U/G→A and A→G/U→C pairs of mutations[23,24], was observed in our study. We also found strong strand asymmetry of G→U mutations in our data, suggestive of RNA damage or RNA editing (rather than replication errors) on the plus stand[15] and possible association with ROS-related processes[26]. The frequency of synonymous mutations is higher than expected ($d_N < d_S$), corresponding to overall purifying selection on within-host mutations of SARS-CoV-2. Collectively, the general within-host virus sequence diversity in the samples from HK was similar to samples from other geographical areas collected at different timepoints[11,16].

Different lineages of SARS-CoV-2 have different properties, including different levels of transmissibility[21,22], disease severity[36,37], viral load[37,38], tissue affinity[39], ability of vaccine breakthrough[4-6], etc. Here, we found SARS-CoV-2 VOC Delta, Omicron and Alpha samples had higher within-host mutation rate and/or nucleotide diversity than non-VOC lineages. Such increased mutation rate is independent of viral load, suggesting different intrinsic biological properties between variants may play a role. As the entire infected population in HK by the end of 2021 was <0.2%, our observation is unlikely affected by interference induced by prior natural infection. Various mutations have been shown to account for different viral properties, e.g., ACE2 binding (e.g., K417N, N501Y)[40], and immune escape (e.g., T478K, L452R)[41]. The increased nonsynonymous nucleotide diversity and putative diversifying selection in VOC samples (Supplementary Table 4) suggest that VOC viruses have a greater capacity to explore protein sequence space and therefore are more likely to incur a fitness change. This result is in line with VOCs' ability to spread and result in multiple sub-lineages, and warrants close monitoring of their molecular evolution in the future.

Vaccination is another factor which may affect the within-host evolution of the virus. We studied samples from Comirnaty and CoronaVac vaccine breakthrough infections and found that vaccination may be associated with increased mutation rates and increased purifying selection. We found Comirnaty vaccination may be associated with increased within-host mutation rate and nucleotide diversity in SARS-CoV-2 Delta-variant samples. Notably, the increased nucleotide diversity in specimens of Delta breakthrough infection in Comirnaty vaccinated individuals is mostly synonymous rather than non-synonymous (π_S = 3.47 and 2.90 for Comirnaty Delta and unvaccinated Delta samples respectively, Supplementary Table 4). We found Comirnaty vaccination increased synonymous nucleotide diversity and thus purifying selection pressure at both the full genome and Spike gene levels, while CoronaVac vaccination showed similar effects only at the Spike gene level. It has been reported that Comirnaty vaccine is markedly more immunogenic than CoronaVac vaccine and this may contribute to our observation[42]. It is also relevant to note that Comirnaty vaccine only has the Spike protein as an immunogen but appears to impact on purifying selection elsewhere in the genome. This may be a result of greater suppression of viral replication. Crucially, additional purifying selection pressures imposed by vaccination may limit the evolutionary protein sequence space as non-synonymous nucleotide diversity does not seem to be increasing. Overall, Comirnaty and CoronaVac vaccination seemingly amplifies the within-host purifying selection in VOCs.

We did not observe enrichment of VOC defining mutations for the non-VOC samples (data not shown), which suggests that convergent evolution of VOC mutations is infrequent. Only three of the mutations observed in our vaccinated samples overlap with known nAb escape supersites on the Spike NTD or RBD regions and the predicted RBD immune escape potential is only mildly (less than 10% immune escape) affected by these mutations. Evolution of T cell epitopes under selection by the host immune system has been reported for other viruses[43-45], and T cell responses to SARS-CoV-2 have also been reported in most COVID-19 patients[46]. However, we did not detect significant vaccination-specific T cell pressure on within-host diversity, suggesting vaccine-induced pressure may not enhance exploration of immune escape pathways.

As HK used an elimination strategy to control COVID-19, the individuals investigated in our study can be reliably categorised as immunologically naïve or vaccinated individuals, which is a significant advantage of our study. Nonetheless, our study has some limitations. The sample size for some groups in this study is small due to limited availability of samples. Although the vaccinated and unvaccinated samples in this study were collected at similar time points after symptom onset (P=0.801, two-sided Wilcoxon rank sum test), most of the studied cases have only single time point samples, and we lack serial samples data of breakthrough infections for studying the temporal changes of within-host selection pressures. In studying the effect of T cell pressure on within-host viral evolution, we could not perform an individual-based analysis since HLA typing of the patients was not performed. As most of the individuals in our study were either infection naïve or vaccinated prior to infection, the effect of hybrid immunity on SARS-CoV-2 within-host evolution could not be addressed and requires further investigation.

In conclusion, our work suggests that SARS-CoV-2 within-host evolution may exhibit different patterns in different virus lineages and in vaccinated individuals. We found that Comirnaty and CoronaVac COVID-19 vaccination increases within-host purifying selection in VOCs, providing evidence that vaccination may limit the exploration of protein sequence space and emergence of more viral variants.

Methods

Samples and sequencing

This study was conducted under ethical approval from the Institutional Review Board of the University of Hong Kong (UW 20-168). We included Illumina amplicon data from 2,053 samples from lineages B.1.1.7 (Alpha), B.1.617.2 (Delta), B.1.1.529 (Omicron) and B.1.1.63/B.1.36/B.1.36.27 (variants in the third and fourth local wave) collected from 2020-07-04 to 2022-03-01 in HK. All the samples were from patients who were either unvaccinated or fully vaccinated (received two doses of vaccines) with Comirnaty or CoronaVac vaccines. The number of samples included in the analysis are presented in Supplementary Table 6. The metadata and vaccination records of RT-PCR confirmed cases of COVID-19 were collected from the public data released by HK government since July 29, 2021 (https://gia.info.gov.hk/general/202107/29/P2021072900356_373472_1_1627542548101.pdf).

To obtain high quality sequence results, we only included samples with a cycle threshold (Ct) value ≤25 and with sufficient genome coverage and sequencing depth (sequencing depth ≥100 properly paired reads are required at >=27000 genomic sites for every sample) after Illumina sequencing. RNA samples were sent to a World Health Organization reference laboratory at the University of HK for full-genome analyses (Institutional Review Board no. UW 20–168). Virus genome was reverse transcribed with multiple gene-specific primers targeting different regions of the viral genome. The synthesized cDNA was then subjected to multiple overlapping 2-kb PCRs for full-genome amplification. PCR amplicons obtained from the same specimen were pooled and sequenced by using the Novaseq or iSeq sequencing platform (Illumina). Sequencing library was prepared by using Nextera XT (Illumina). Generated sequencing reads were quality trimmed by fastp with parameters ("-q 30 -5 -3 -c --detect_adapter_for_pe -l 50"). Potential PCR duplicates were removed by samtools markdup (v1.11). The trimmed reads were mapped to a reference virus genome by using BWA-MEM2 (v2.0pre2), and genome consensus was generated by using iVar (v1.3.1) with the PCR primer trimming protocol (minimum sequence depth of 100 and minimum Qvalue of 30).

Variant calling and quality control

The consensus-level single nucleotide polymorphisms (SNPs) and intrahost single nucleotide variants (iSNVs) were called by iVar variants (v1.3.1) with reference to the Wuhan-Hu-01 sequence. To limit the analysis to high quality SNPs and iSNVs, the following filtering criteria were applied:

1. SNVs were called from samtools mpileup files from quality-filtered reads alignment bam files using pysamstats.
2. After filtering based on MAF threshold of 0.05, we identified 24,161 iSNVs in 2,051 samples.
3. After filtering for iSNVs with strong strand bias (we kept iSNVs with strand ratio < 1/10), we identified 5,949 iSNVs in 1,643 samples.

4. After filtering for serial adjacent disjoint mutations (≥3 mutations within 30 nucleotides sliding window, likely relating to sequencing errors), we identified 5,808 iSNVs in 1,642 samples.
5. After filtering for iSNVs by minimum depth of 100 reads, we identified 5,073 iSNVs in 1,587 samples.
6. After filtering heading/tailing 100bp UTR region, binding regions of PCR primers, and previously known problematic sites[47], we identified 3,439 iSNVs in 1,129 samples.
7. Finally, removing samples with possible co-infection/contamination, we identified 2,731 iSNVs in 1,117 samples.

To further validate that the identified iSNVs are of high confidence and are reproducible, we tested another 93 technical control samples from 86 cases sequenced by different sequencing runs and platforms. Using the same filtering criteria, we found iSNVs are significantly more reproducible among technical control samples from the same patient. Specifically, for the technical control samples with at least one iSNV, 54.2% (median) of the iSNVs were reproducible between samples from the same patient, compared to 0% (median) of the iSNVs being reproducible among samples from different patients (P<0.001, two-sided Wilcoxon rank sum test).

Mutation summary statistics

Incidence of iSNVs and minor allele frequency

The incidence of iSNVs (number of iSNVs per Kb) was calculated by dividing the number of iSNVs with the number of genomic positions with sufficient coverage of reads (sequencing depth ≥ 100). The adjusted incidence of iSNVs is the residual (that is response minus fitted values) calculated by least-squares linear model ("lm" function in R 4.1.0) with the numbers of iSNVs per Kb (response variable) and Ct values (explanatory variable) from all the studied samples. The minor allele frequency (MAF), representing the abundance of iSNVs, was calculated directly from the alignment mpileup files using pysamstats (v1.1.2).

Nucleotide diversity (π)

Nucleotide diversity (π) is a summary metric of the degree of polymorphism of iSNVs within a sample and is tolerant of biases from sequencing depth[48]. We use it to measure the degree of iSNVs polymorphism within a sample. For every sample, where n_i sequences (NGS reads) of nucleotide i are observed, nucleotide diversity (π) can be calculated based on pairwise difference between sequencing reads. as

$$\pi = \frac{\sum_{i \neq j} n_i n_j}{\frac{1}{2} N(N-1)},$$

where N is the total number of sequences.

Selection analysis

The nucleotide diversity can be separately calculated for synonymous (π_S) and non-synonymous changes (π_N) in coding regions. We calculated the π_N and π_S in this study using

SNPGenie[49] with self-curated input vcf files based on the above identified iSNVs. For hypothesis testing of selection neutrality ($\pi_N = \pi_S$), Z-tests using a bootstrap method (codon unit, 10,000 replicates for genes and sliding windows) was applied. The scripts of sliding window analysis for positive selection are largely based on a previous analysis developed by the author of the software (https://github.com/krisp-kwazulu-natal/within-host-diversity-manuscript-analysis-code/blob/a276286680de3723e2b1e70f7a060750892cf8af/scripts/diversity_selection_analyses.R). The usage of this software in our study was approved by the author. Sliding windows of thirty codons and step size of one codon were used because this did not exceed the length of ORF10 (thirty-nine codons).

Neutralizing antibody escape mutations

The Spike RBD mutations found in all Omicron and Delta samples were analyzed separately with the Escape Calculator for SARS-CoV-2 RBD[50]. The calculations are based on deep mutational scanning of a large set of RBD targeting antibodies which are known to neutralize the ancestral Wuhan-Hu-1 strain. The mutation escape strength in the Escape Calculator was set to the default value of 2.

For NTD, an antigenic supersite has been defined in McCallum et al.[31] that is recognised by a large number of NTD-targeting nAbs. It includes the Spike regions: 14-20, 140-158 and 245-264. Multiple other NTD mutations have been reported to affect neutralization of NTD-targeting nAbs. These NTD mutations include[32] A67V, del69-70, T95I, G142D, del143-145, N211I, del212, and ins214 EPE. In ref.[33], NTD mutations with strong (del144, R246A), moderate (L18F, T19A, H164Y, D253G, D253Y), and mild (D80A, N149Q, S252F) effect on antibody neutralization were described. This data was collectively used in the overlap analysis of Spike NTD mutations (Fig. 4).

Acquisition of SARS-CoV-2 CD8$^+$ and CD4$^+$ T cell epitopes

We obtained SARS-CoV-2 CD8$^+$ and CD4$^+$ T cell epitope data from the dashboard reported by us[51] (https://www.mckayspcb.com/SARS2TcellEpitopes/; accessed on 15 May 2022) and the Immune Epitope Database (IEDB)[52] (https://www.iedb.org; accessed on 15 May 2022) by querying for the organism name: "SARS-CoV2" (taxonomy ID: 2697049), host: "human", and assay: "T cell positive". The compiled data was processed to only include epitopes with lengths between 9-11 residues for CD8$^+$ and 13-20 residues for CD4$^+$, which represent the typical range of HLA class I and II epitopes. Removing the epitopes with no or incomplete HLA allele information resulted in a total of 1,324 unique CD8$^+$ and 961 unique CD4$^+$ epitope-HLA pairs (Supplementary Fig. 6). The analysis in Fig. 5A is based on this complete set of known SARS-CoV-2 T cell epitopes.

For the analysis focused on epitopes targeted by T cells in the HK population (Fig. 5B), we determined class I and class II HLA alleles prevalent in HK. For class I alleles, we employed the IEDB's "Population Coverage" tool (http://tools.iedb.org/population/) to identify 12 HLA class I alleles that together cover >99% of the HK population (Supplementary Fig. 10A, left

panel). A total of 630 unique SARS-CoV-2 $CD8^+$ T cell epitopes were associated with these alleles (Supplementary Fig. 10A, right panel). For class II alleles, we employed the Allele Frequency Net Database[53] (http://www.allelefrequencies.net; accessed on 15 May 2022) and identified 13 HLA class II alleles that have an individual estimated population coverage of >5% in the HK population (Supplementary Fig. 10B, left panel). A total of 258 unique SARS-CoV-2 $CD4^+$ T cell epitopes were associated with these alleles (Supplementary Fig. 10B, right panel).

Overlapping T cell epitopes per mutation

To study whether the within-host mutations (minor allele variants) affect the T cell response generated against different SARS-CoV-2 lineages and under different vaccination status, we used the metric *Overlapping T cell epitopes per mutation*. It is computed as the number of T cell epitopes overlapping the within-host mutations observed in each group divided by the total number of within-host mutations observed in that group. The T cell epitope data used in the calculation of this metric was either from the complete set (Fig. 5A) or from the set specific to the HK population (Fig. 5B).

Statistical analysis

For bootstrapping analysis, the measurement can be taken from the same sample measured repeatedly. For the other tests (e.g., Wilcoxon tests), the measurements were taken from distinct samples. All the statistical tests in this study are two-sided and no adjustment for multiple comparisons was performed unless specified.

Data availability

The sequencing data used in this study can be access through NCBI Sequence Read Archive (SRA) with accession ID: XXX. The anonymised metadata are deposited at https://github.com/Leo-Poon-Lab/mutations-under-sarscov2-vaccination/XXX.

Code availability

Detailed scripts for the above analysis are available from https://github.com/Leo-Poon-Lab/mutations-under-sarscov2-vaccination.

Competing interests

None

Acknowledgments and funding sources

We acknowledge the technical support provided by colleagues from the Centre for PanorOmic Sciences of the University of HK. We also acknowledge the Centre for Health Protection of the Department of Health for providing epidemiological data for the study. The computations were performed using research computing facilities offered by Information Technology Services, the University of HK. The funding bodies had no role in the design of the study, the collection, analysis, and interpretation of data, or writing of the manuscript. This work is supported by the US National Institutes of Health (U01AI151810 and 75N93021C00016), the

Australian Research Council future fellowship scheme (FT200100928), Research Grants Council of HK theme-based research schemes (T11-705/21-N), InnoHK grant for the Centre for Immunology and Infection, and Health and Medical Research Fund (COVID190205).

Contributions

This study was designed by H.G. and L.L.M.P.. Data curation was performed by H.G. and A.A.Q.. The SARS-CoV-2 genome sequencing team included P.K, D.Y.M.N, L.D.J.C, G.Y.Z.L, S.S.M.C and L.L.M.P.. Data analysis was done by H.G., A.A.Q, T.T.Y.L., M.P. M.R.M, and L.L.M.P.. Data visualization was done by H.G. A.A.Q. M.R.M. and L.L.M.P.. L.L.M.P. was responsible for project supervision. The original draft of this manuscript was prepared by H.G. and A.A.Q. and was reviewed and edited by T.T.Y.L, M.P, M.R.M and L.L.M.P.

References

1. Ritchie, H., *et al.* Coronavirus pandemic (COVID-19). *Our world in data* (2020).
2. Pulliam, J.R.C., *et al.* Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. *Science* **376**, eabn4947 (2022).
3. Lipsitch, M., Krammer, F., Regev-Yochay, G., Lustig, Y. & Balicer, R.D. SARS-CoV-2 breakthrough infections in vaccinated individuals: measurement, causes and impact. *Nat Rev Immunol* **22**, 57-65 (2022).
4. McCallum, M., *et al.* Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement. *Science* **375**, 864-868 (2022).
5. Arora, P., *et al.* Comparable neutralisation evasion of SARS-CoV-2 omicron subvariants BA.1, BA.2, and BA.3. *Lancet Infect Dis* **22**, 766-767 (2022).
6. McCallum, M., *et al.* Molecular basis of immune evasion by the Delta and Kappa SARS-CoV-2 variants. *Science* **374**, 1621-1626 (2021).
7. Rochman, N.D., *et al.* Ongoing global and regional adaptive evolution of SARS-CoV-2. *Proceedings of the National Academy of Sciences* **118**, e2104241118 (2021).
8. Gu, H., *et al.* Genomic epidemiology of SARS-CoV-2 under an elimination strategy in Hong Kong. *Nature Communications* **13**, 736 (2022).
9. du Plessis, L., *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708-712 (2021).
10. Di Giorgio, S., Martignano, F., Torcia, M.G., Mattiuz, G. & Conticello, S.G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci Adv* **6**, eabb5813 (2020).
11. Lythgoe, K.A., *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* **372**, eabg0821 (2021).
12. Hannon, W.W., *et al.* Narrow transmission bottlenecks and limited within-host viral diversity during a SARS-CoV-2 outbreak on a fishing boat. *bioRxiv*, 2022.2002.2009.479546 (2022).
13. Martin, M.A. & Koelle, K. Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2". *Science Translational Medicine* **13**, eabh1803 (2021).
14. Braun, K.M., *et al.* Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLoS Pathog* **17**, e1009849 (2021).
15. Tonkin-Hill, G., *et al.* Patterns of within-host genetic diversity in SARS-CoV-2. *eLife* **10**, e66857 (2021).

16. Valesano, A.L.*, et al.* Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *PLOS Pathogens* **17**, e1009499 (2021).
17. Bashor, L.*, et al.* SARS-CoV-2 evolution in animals suggests mechanisms for rapid variant selection. *Proceedings of the National Academy of Sciences* **118**, e2105253118 (2021).
18. Voloch, C.M.*, et al.* Intra-host evolution during SARS-CoV-2 prolonged infection. *Virus Evol* **7**, veab078 (2021).
19. Gandhi, S.*, et al.* De novo emergence of a remdesivir resistance mutation during treatment of persistent SARS-CoV-2 infection in an immunocompromised patient: a case report. *Nature communications* **13**, 1-8 (2022).
20. Sonnleitner, S.T.*, et al.* Cumulative SARS-CoV-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host. *Nature Communications* **13**, 2560 (2022).
21. Zeng, C.*, et al.* SARS-CoV-2 spreads through cell-to-cell transmission. *Proceedings of the National Academy of Sciences* **119**, e2111400119 (2022).
22. Earnest, R.*, et al.* Comparative transmissibility of SARS-CoV-2 variants delta and Alpha in new England, USA. *Cell Reports Medicine* **3**, 100583 (2022).
23. Li, J.*, et al.* Two-step fitness selection for intra-host variations in SARS-CoV-2. *Cell Rep*, 110205 (2021).
24. San, J.E.*, et al.* Transmission dynamics of SARS-CoV-2 within-host diversity in two major hospital outbreaks in South Africa. *Virus Evolution* **7**(2021).
25. Pathak, A.K.*, et al.* Spatio-temporal dynamics of intra-host variability in SARS-CoV-2 genomes. *Nucleic Acids Res* **50**, 1551-1561 (2022).
26. Graudenzi, A., Maspero, D., Angaroni, F., Piazza, R. & Ramazzotti, D. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. *iScience* **24**, 102116 (2021).
27. Nei, M. & Li, W.-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences* **76**, 5269-5273 (1979).
28. Gregori, J.*, et al.* Viral quasispecies complexity measures. *Virology* **493**, 227-237 (2016).
29. Braun, K.M.*, et al.* Transmission of SARS-CoV-2 in domestic cats imposes a narrow bottleneck. *PLoS Pathog* **17**, e1009373 (2021).
30. Alidjinou, E.K.*, et al.* Immunogenicity of BNT162b2 vaccine booster against SARS-CoV-2 Delta and Omicron variants in nursing home residents: A prospective observational study in older adults aged from 68 to 98 years. *The Lancet Regional Health - Europe* **17**, 100385 (2022).
31. McCallum, M.*, et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332-2347.e2316 (2021).
32. Suryadevara, N.*, et al.* Neutralizing and protective human monoclonal antibodies recognizing the N-terminal domain of the SARS-CoV-2 spike protein. *Cell* **184**, 2316-2331.e2315 (2021).
33. Cui, Z.*, et al.* Structural and functional characterizations of infectivity and immune evasion of SARS-CoV-2 Omicron. *Cell* **185**, 860-871.e813 (2022).
34. Valkenburg, S.A.*, et al.* Acute emergence and reversion of influenza A virus quasispecies within CD8+ T cell antigenic peptides. *Nature communications* **4**, 1-10 (2013).
35. Machkovech, H.M., Bedford, T., Suchard, M.A. & Bloom, J.D. Positive selection in CD8+ T-cell epitopes of influenza virus nucleoprotein revealed by a comparative

analysis of human and swine viral lineages. *Journal of virology* **89**, 11275-11283 (2015).
36. Ulloa, A.C., Buchan, S.A., Daneman, N. & Brown, K.A. Estimates of SARS-CoV-2 Omicron Variant Severity in Ontario, Canada. *JAMA* **327**, 1286-1288 (2022).
37. Ong, S.W.X.*, et al.* Clinical and virological features of SARS-CoV-2 variants of concern: a retrospective cohort study comparing B.1.1.7 (Alpha), B.1.315 (Beta), and B.1.617.2 (Delta). *Clin Infect Dis* (2021).
38. Hoteit, R. & Yassine, H.M. Biological Properties of SARS-CoV-2 Variants: Epidemiological Impact and Clinical Consequences. *Vaccines* **10**, 919 (2022).
39. Hui, K.P.Y.*, et al.* SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo. *Nature* **603**, 715-720 (2022).
40. Khan, A.*, et al.* Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: an insight from structural data. *Journal of cellular physiology* **236**, 7045-7057 (2021).
41. Adam, D. What scientists know about new, fast-spreading coronavirus variants. *Nature* **594**, 19-20 (2021).
42. Mok, C.K.P.*, et al.* Comparison of the immunogenicity of BNT162b2 and CoronaVac COVID-19 vaccines in Hong Kong. *Respirology* **27**, 301-310 (2022).
43. O'Connor, S.L.*, et al.* Conditional CD8+ T cell escape during acute simian immunodeficiency virus infection. *Journal of virology* **86**, 605-609 (2012).
44. Bimber, B.N.*, et al.* Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing. *J Virol* **84**, 12087-12092 (2010).
45. Bull, M.B.*, et al.* Next-generation T cell activating vaccination increases influenza virus mutation prevalence. *Science Advances* **8**, eabl5209 (2022).
46. Moss, P. The T cell immune response against SARS-CoV-2. *Nature Immunology* **23**, 186-193 (2022).
47. Gu, H.*, et al.* SARS-CoV-2 under an elimination strategy in Hong Kong. (2021).
48. Zhao, L. & Illingworth, C.J.R. Measurements of intrahost viral diversity require an unbiased diversity metric. *Virus Evol* **5**, vey041 (2019).
49. Nelson, C.W., Moncla, L.H. & Hughes, A.L. SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics* **31**, 3709-3711 (2015).
50. Greaney, A.J., Starr, T.N. & Bloom, J.D. An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain. *Virus Evolution* **8**(2022).
51. Quadeer, A.A., Ahmed, S.F. & McKay, M.R. Landscape of epitopes targeted by T cells in 852 individuals recovered from COVID-19: Meta-analysis, immunoprevalence, and web platform. *Cell Reports Medicine* **2**(2021).
52. Vita, R.*, et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research* **47**, D339-D343 (2018).
53. Gonzalez-Galarza, Faviel F.*, et al.* Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research* **48**, D783-D788 (2019).

Figure Legends

Figure 1. Statistics of within-host mutations in SARS-CoV-2 samples. (A) Distribution of number of iSNV site(s) in each sample, colored by ranges of Ct values. The dashed line shows the mean value of the distribution. **(B)** Distribution of number of sample(s) sharing iSNVs (e.g., if the iSNV identified in one sample was not shared with any other sample, then the number of samples sharing that iSNV equals to one (x = 1), and so on), colored by variant types. **(C)** Distribution of the frequency of iSNVs per sample per synonymous and per non-synonymous site (d_S and d_N) for different types of mutations, colored by variant types. The dashed lines show the average frequency of synonymous and non-synonymous iSNVs among all types of mutations. The points and error bars show mean and standard deviation values based on 10,000 bootstrap replicates at mutation level. **(D)** Distribution of the frequency of iSNVs per sample per Kb for synonymous and non-synonymous site (d_S per Kb and d_N per Kb) in different genomic regions of 1Kb length, colored by variant types. The points and error bars show mean and standard deviation values based on 10,000 bootstrap replicates at mutation level. **(E)** Distribution of high-frequency mutations shared by multiple samples, colored by variant types. Coding regions of the SARS-CoV-2 genome, based on the reference genome (GenBank: MN908947.3), are shown at the bottom of the figure.

Figure 2. Comparison of within-host mutation profiles between unvaccinated VOC and unvaccinated non-VOC samples. (A-C) Full genome incidence of iSNVs (adjusted number of iSNVs per Kb), abundance of iSNVs (minor allele frequencies, MAF), and nucleotide diversity (π) of different samples. For all box plots, the bold horizontal line inside the box shows the median, the upper and lower edges of the box indicate the first and the third quartiles, and whiskers extend to span a 1.5 interquartile range from the edges. Pairwise comparisons between groups were performed by two-sided two-sample Wilcoxon tests; the pairs with P-value ≤ 0.01 and ≤ 0.05 are labelled with "**" and "*" respectively. **(D-E)** Full-genome and gene-specific within-host nonsynonymous nucleotide diversity (π_N) and synonymous nucleotide diversity (π_S) in samples from different groups. The points and error bars show the mean and standard deviation values under 10,000 bootstrap replicates at codon level. Significance was evaluated using Z-tests of the null hypothesis that $\pi_N - \pi_S = 0$ (10,000 bootstrap replicates, codon unit); P-value ≤ 0.01, ≤ 0.05 and ≤ 0.10 are labelled with "**", "*" and "^", respectively.

Figure 3. Comparison of within-host mutation profiles between vaccinated and unvaccinated Delta and Omicron samples. (A-C) Full-genome incidence of iSNVs (adjusted number of iSNVs per Kb), abundance of iSNVs (minor allele frequencies, MAF) and nucleotide diversity (π) of different samples. For all box plots, the bold horizontal line inside the box shows the median, the upper and lower edges of the box indicate the first and the third quartiles, and whiskers extend to span a 1.5 interquartile range from the edges. Pairwise comparisons between groups were performed by the two-sided two-sample Wilcoxon test; the pairs with P-value ≤ 0.01 and ≤ 0.05 are labelled with "**" and "*" respectively. **(D-E)** Full-

genome and gene-specific within-host nonsynonymous nucleotide diversity (π_N) and synonymous nucleotide diversity (π_S) in samples from different groups. The points and error bars showed the mean and standard deviation values under 10,000 bootstrap replicates at codon level. Significance was evaluated using Z-tests of the null hypothesis that $\pi_N - \pi_S = 0$ (10,000 bootstrap replicates, codon unit); P-value ≤ 0.01, ≤ 0.05 and ≤ 0.10 were labelled with "**", "*" and "^", respectively.

Figure 4. Spike mutations identified in unvaccinated and vaccinated Delta and Omicron samples. Each circle represents one mutation identified in one sample in this study. **(A)** Identified within-host mutations in Omicron samples; **(B)** Identified within-host mutations in Delta samples.

Figure 5. Overlapping known SARS-CoV-2 CD8+ and CD4+ T cell epitopes per mutation in unvaccinated and vaccinated Delta and Omicron samples. **(A)** Analysis based on all known SARS-CoV-2 T cell epitopes. **(B)** Analysis based on T cell epitopes associated with HLA alleles prevalent in the Hong Kong population. Pairwise comparisons within groups were performed by the two-sided two-sample Wilcoxon test. For all box plots, the bold horizontal line inside the box shows the median, the upper and lower edges of the box indicate the first and the third quartiles, and whiskers extend to span a 1.5 interquartile range from the edges.

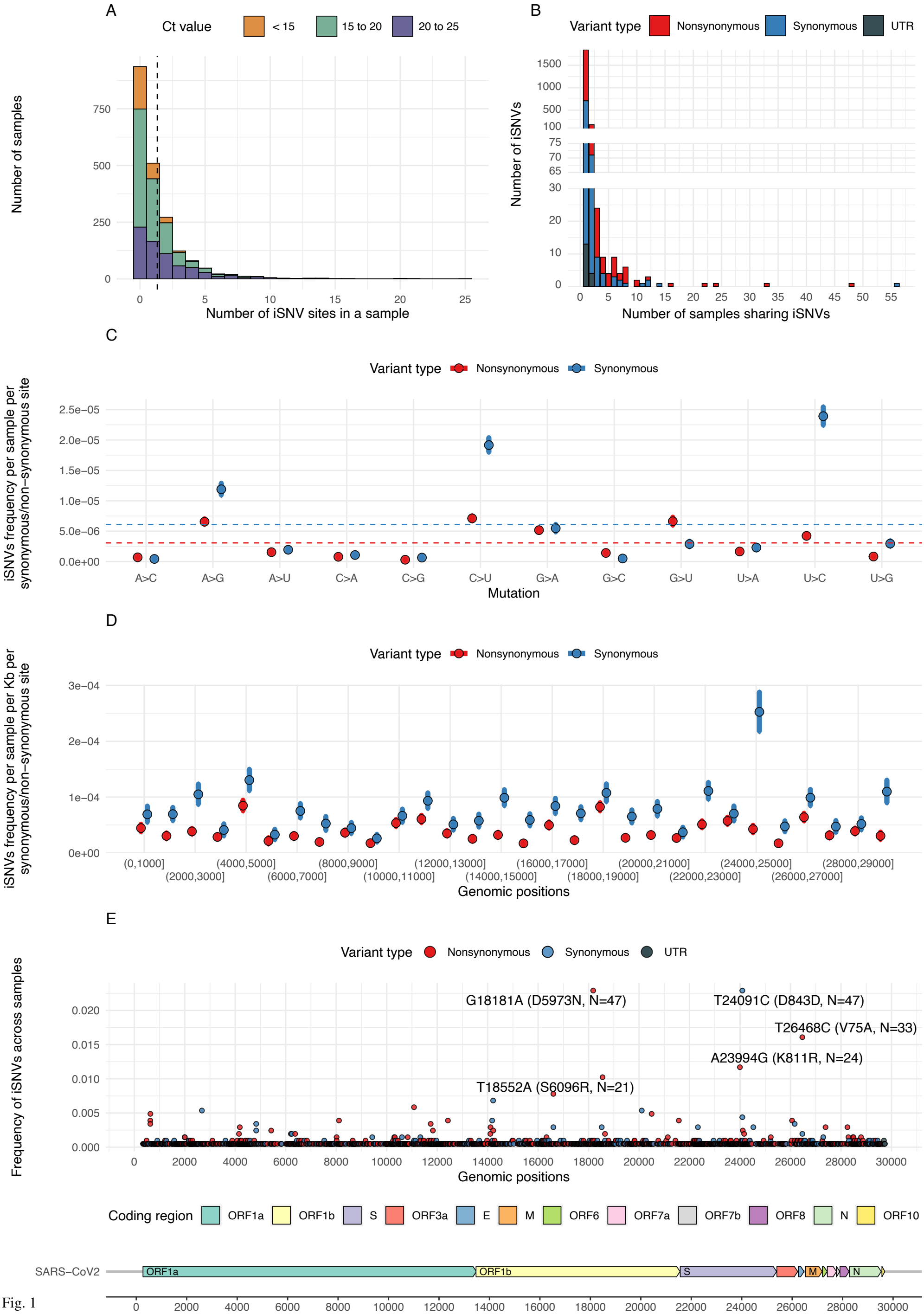

A
Ct value
< 15
15 to 20
20 to 25
Number of samples
Number of iSNV sites in a sample
B
Variant type
Nonsynonymous
Synonymous
UTR
Number of iSNVs
Number of samples sharing iSNVs
C
Variant type
Nonsynonymous
Synonymous
iSNVs frequency per sample per synonymous/non−synonymous site
A>C
A>G
A>U
C>A
C>G
C>U
G>A
G>C
G>U
U>A
U>C
U>G
Mutation
D
Variant type
Nonsynonymous
Synonymous
iSNVs frequency per sample per Kb per synonymous/non−synonymous site
Genomic positions
E
Variant type
Nonsynonymous
Synonymous
UTR
Frequency of iSNVs across samples
G18181A (D5973N, N=47)
T24091C (D843D, N=47)
T26468C (V75A, N=33)
A23994G (K811R, N=24)
T18552A (S6096R, N=21)
Genomic positions
Coding region
ORF1a
ORF1b
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
ORF10
SARS−CoV2

Fig. 1

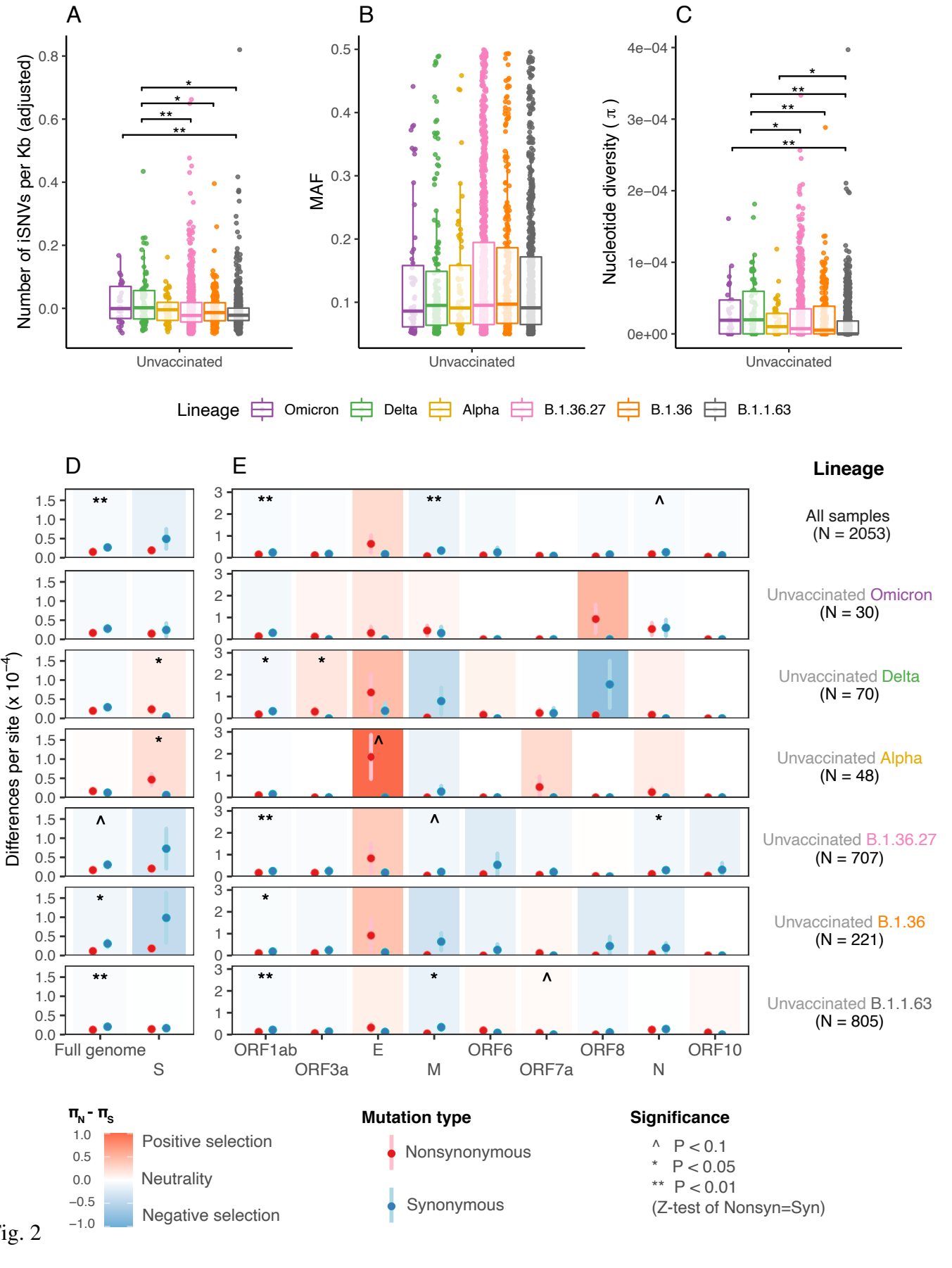

A
B
C
Number of iSNVs per Kb (adjusted)
MAF
Nucleotide diversity (π)
Unvaccinated
Lineage
Omicron
Delta
Alpha
B.1.36.27
B.1.36
B.1.1.63
D
E
Differences per site (x 10−4)
Full genome
S
ORF1ab
ORF3a
E
M
ORF6
ORF7a
ORF8
N
ORF10
Lineage
All samples (N = 2053)
Unvaccinated Omicron (N = 30)
Unvaccinated Delta (N = 70)
Unvaccinated Alpha (N = 48)
Unvaccinated B.1.36.27 (N = 707)
Unvaccinated B.1.36 (N = 221)
Unvaccinated B.1.1.63 (N = 805)
πN - πS
Positive selection
Neutrality
Negative selection
Mutation type
Nonsynonymous
Synonymous
Significance
^ P < 0.1
* P < 0.05
** P < 0.01
(Z-test of Nonsyn=Syn)

Fig. 2

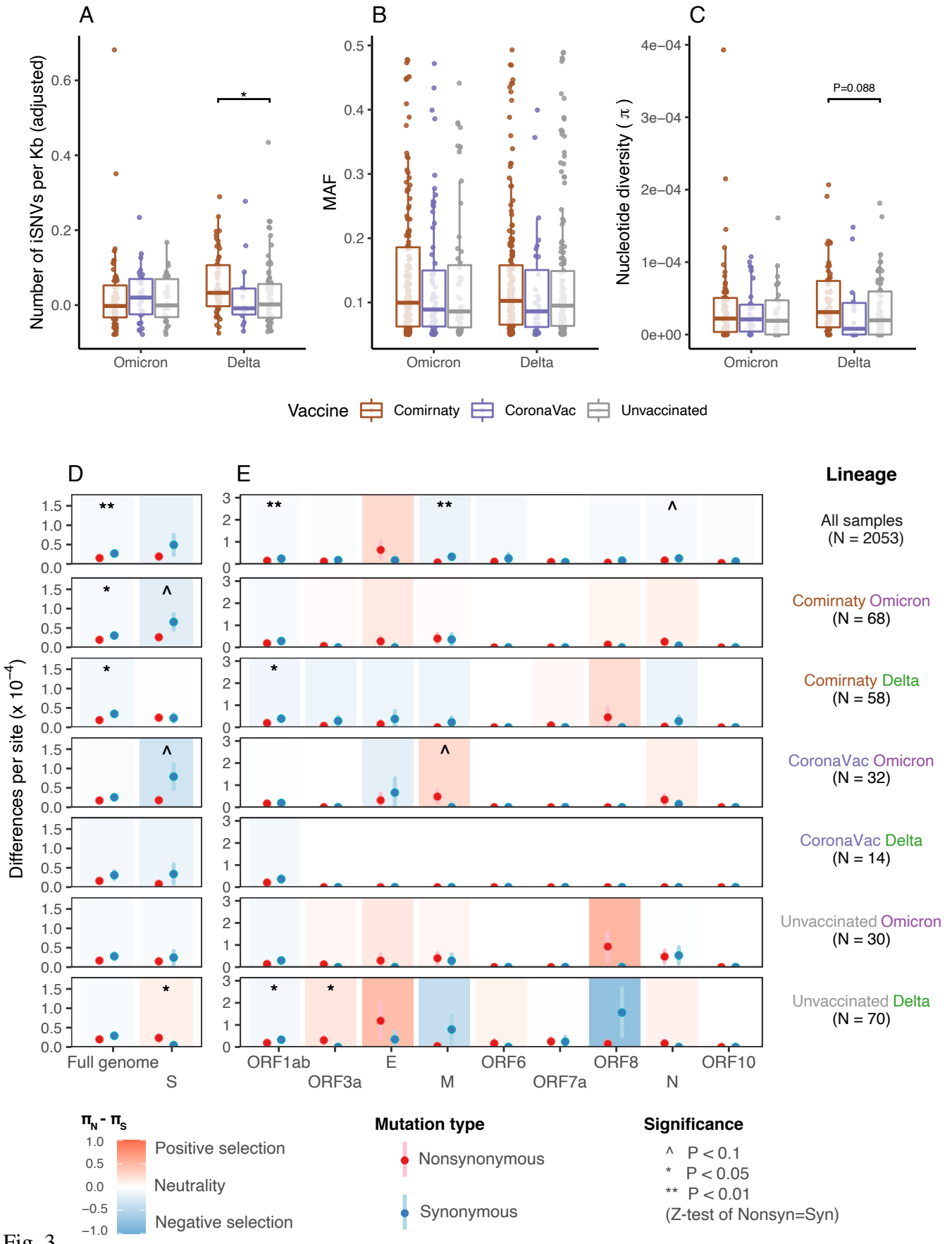

A
B
C
Number of iSNVs per Kb (adjusted)
MAF
Nucleotide diversity (π)
P=0.088
Omicron
Delta
Vaccine
Comirnaty
CoronaVac
Unvaccinated
D
E
Differences per site (x 10−4)
Full genome
S
ORF1ab
ORF3a
E
M
ORF6
ORF7a
ORF8
N
ORF10
Lineage
All samples (N = 2053)
Comirnaty Omicron (N = 68)
Comirnaty Delta (N = 58)
CoronaVac Omicron (N = 32)
CoronaVac Delta (N = 14)
Unvaccinated Omicron (N = 30)
Unvaccinated Delta (N = 70)
πN - πS
Positive selection
Neutrality
Negative selection
Mutation type
Nonsynonymous
Synonymous
Significance
^ P < 0.1
* P < 0.05
** P < 0.01
(Z-test of Nonsyn=Syn)

Fig. 3

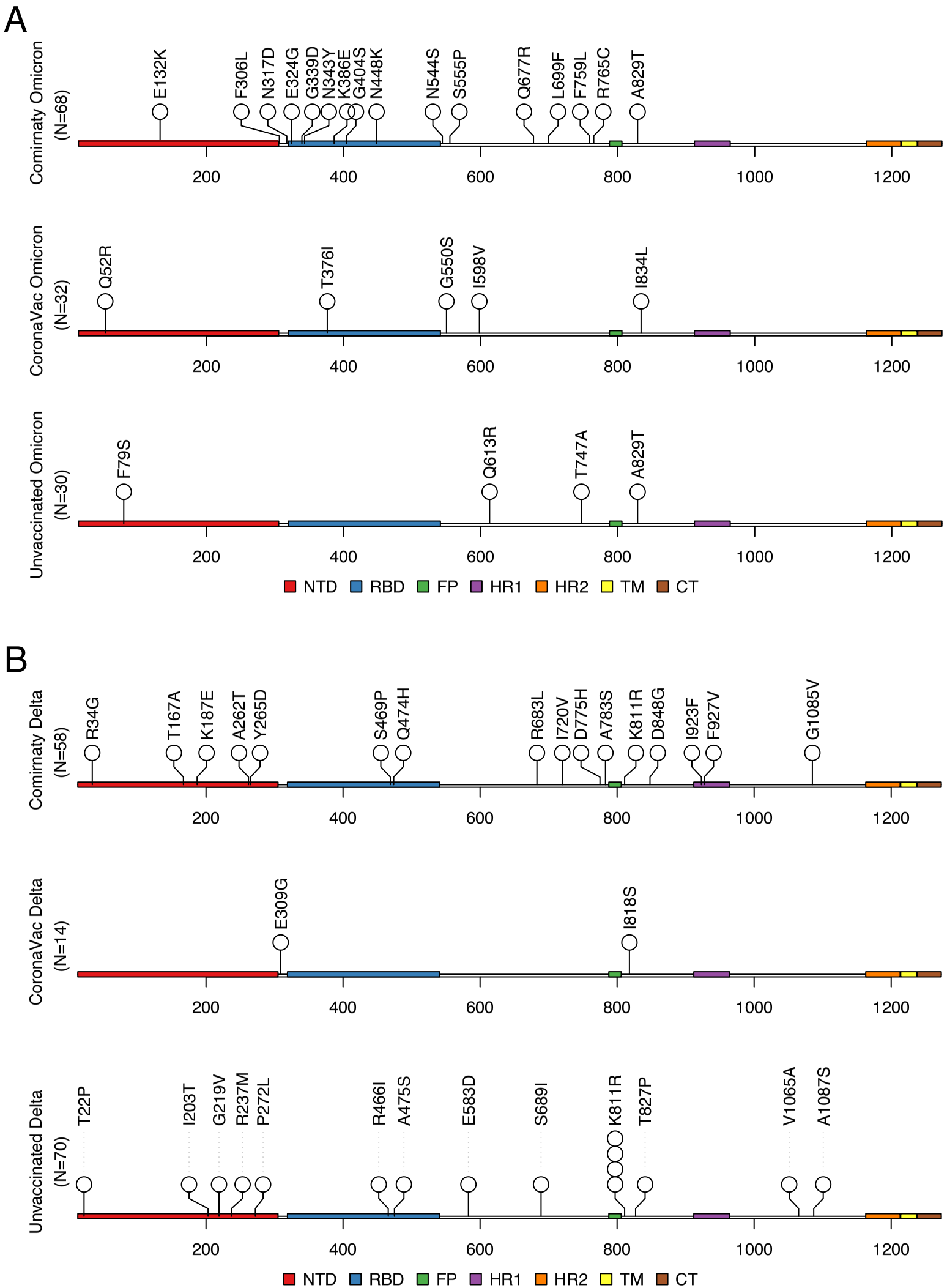

A
Comirnaty Omicron (N=68)
E132K
F306L
N317D
E324G
G339D
N343Y
K386E
G404S
N448K
N544S
S555P
Q677R
L699F
F759L
R765C
A829T
CoronaVac Omicron (N=32)
Q52R
T376I
G550S
I598V
I834L
Unvaccinated Omicron (N=30)
F79S
Q613R
T747A
A829T
200
400
600
800
1000
1200
NTD
RBD
FP
HR1
HR2
TM
CT
B
Comirnaty Delta (N=58)
R34G
T167A
K187E
A262T
Y265D
S469P
Q474H
R683L
I720V
D775H
A783S
K811R
D848G
I923F
F927V
G1085V
CoronaVac Delta (N=14)
E309G
I818S
Unvaccinated Delta (N=70)
T22P
I203T
G219V
R237M
P272L
R466I
A475S
E583D
S689I
K811R
T827P
V1065A
A1087S

Fig. 4

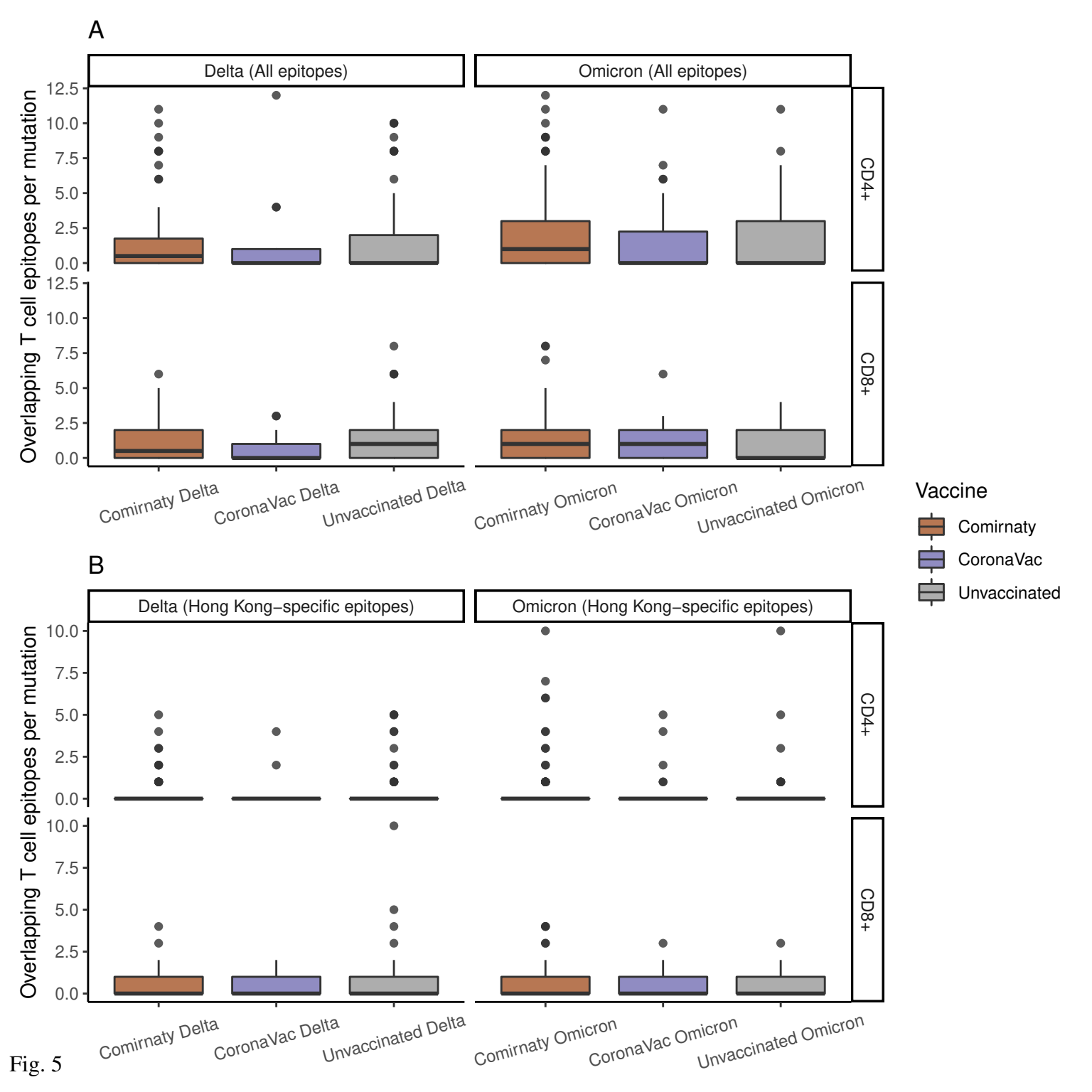

A
Delta (All epitopes)
Omicron (All epitopes)
CD4+
CD8+
Overlapping T cell epitopes per mutation
12.5
10.0
7.5
5.0
2.5
0.0
Comirnaty Delta
CoronaVac Delta
Unvaccinated Delta
Comirnaty Omicron
CoronaVac Omicron
Unvaccinated Omicron
Vaccine
Comirnaty
CoronaVac
Unvaccinated
B
Delta (Hong Kong−specific epitopes)
Omicron (Hong Kong−specific epitopes)

Fig. 5

Supplementary Tables

Table 1. iSNVs in different regions.

Gene	Mean iSNVs per Kb	Mean iSNVs per Kb (adjusted)
Full genome	0.04535019	-1.37E-18
ORF1ab	0.04392511	3.43E-18
S	0.06269981	-5.83E-18
ORF3a	0.03001364	-6.78E-18
E	0.14527307	-1.29E-17
M	0.03996223	3.60E-18
ORF6	0.05237549	-2.34E-17
ORF7a	0.0324728	1.28E-17
ORF7b	0.02319486	-1.52E-17
ORF8	0.02135847	7.58E-18
N	0.04786782	-3.81E-18
ORF10	0.03746862	-1.68E-17

Table 2. Distribution of high-frequent iSNVs in different viral lineage and vaccination groups.

Lineage	Vaccine	Gene	Mutation (amino acid)	Proportion
B.1.1.63	Unvaccinated	ORF1ab	D5973N	0.28
B.1.36.27	Unvaccinated	ORF1ab	D5973N	0.26
B.1.36	Unvaccinated	ORF1ab	D5973N	0.15
Delta	Unvaccinated	ORF1ab	D5973N	0.13
Delta	Comirnaty	ORF1ab	D5973N	0.11
Alpha	Unvaccinated	ORF1ab	D5973N	0.04
Delta	CoronaVac	ORF1ab	D5973N	0.04
B.1.36.27	Unvaccinated	S	D843D	0.73
B.1.36	Unvaccinated	S	D843D	0.27
B.1.36.27	Unvaccinated	S	K811R	0.33
B.1.1.63	Unvaccinated	S	K811R	0.25
Alpha	Unvaccinated	S	K811R	0.21
Delta	Unvaccinated	S	K811R	0.17
Delta	Comirnaty	S	K811R	0.04
B.1.36.27	Unvaccinated	ORF1ab	S6096R	0.55
B.1.1.63	Unvaccinated	ORF1ab	S6096R	0.27
B.1.36	Unvaccinated	ORF1ab	S6096R	0.09
Delta	Comirnaty	ORF1ab	S6096R	0.05
Delta	CoronaVac	ORF1ab	S6096R	0.05
B.1.36.27	Unvaccinated	E	V75A	0.61
B.1.36	Unvaccinated	E	V75A	0.27
Delta	Unvaccinated	E	V75A	0.09
Alpha	Unvaccinated	E	V75A	0.03

Table 3. Differences in iSNVs between groups at full-genome level. Only pairs with significantly large difference ($p<0.05$ in Wilcoxon Rank Sum test and difference between median values >10%) are shown.

Variable 1	Variable 2	Median of variable 1	Median of variable 2	P value
Number of iSNVs per Kb (adjusted)				
Unvaccinated_B.1.1.63	Unvaccinated_Delta	-0.0218	0.0019	0.0002
Comirnaty_Delta	Comirnaty_Omicron	0.0328	-0.0023	0.0013
Unvaccinated_B.1.36.27	Unvaccinated_Delta	-0.0225	0.0019	0.0030
Unvaccinated_B.1.36	Unvaccinated_Delta	-0.0135	0.0019	0.0138
Comirnaty_Delta	Unvaccinated_Delta	0.0328	0.0019	0.0189
Unvaccinated_B.1.1.63	Unvaccinated_Omicron	-0.0218	-0.0007	0.0243
Unvaccinated_B.1.1.63	Unvaccinated_B.1.36	-0.0218	-0.0135	0.0379
Minor allele frequency				
NA				
Nucleotide diversity (π)				
Unvaccinated_B.1.36.27	Unvaccinated_B.1.1.63	7.22E-06	0	< 0.0001
Unvaccinated_B.1.1.63	Unvaccinated_Delta	0	1.97E-05	< 0.0001
Unvaccinated_B.1.36	Unvaccinated_B.1.1.63	5.24E-06	0	0.0002
Unvaccinated_B.1.1.63	Unvaccinated_Omicron	0	1.90E-05	0.0005
Unvaccinated_B.1.36	Unvaccinated_Delta	5.24E-06	1.97E-05	0.0096
Unvaccinated_B.1.1.63	Unvaccinated_Alpha	0	1.01E-05	0.0137
Unvaccinated_B.1.36.27	Unvaccinated_Delta	7.22E-06	1.97E-05	0.0165

Table 4. Synonymous and nonsynonymous nucleotide diversity on full genome and spike gene of different groups.

Gene	Group	π_N ($\pm SD$) (10^{-5})	π_S ($\pm SD$) (10^{-5})	$\pi_N - \pi_S$ (10^{-5})	π_N/π_S
Full genome	Combined (N=2053)	1.51 (1.43 ~ 1.59)	2.69 (2.32 ~ 3.06)	-1.18	0.56
Full genome	Comirnaty Delta (N = 58)	1.85 (1.57 ~ 2.13)	3.47 (2.81 ~ 4.12)	-1.62	0.53
Full genome	Comirnaty Omicron (N = 68)	1.95 (1.7 ~ 2.21)	3.06 (2.56 ~ 3.56)	-1.10	0.64
Full genome	Unvaccinated Alpha (N = 48)	1.64 (1.36 ~ 1.93)	1.25 (0.89 ~ 1.6)	0.40	1.32
Full genome	Unvaccinated B.1.1.63 (N = 805)	1.24 (1.16 ~ 1.32)	2.04 (1.87 ~ 2.22)	-0.81	0.61
Full genome	Unvaccinated B.1.36 (N = 221)	1.16 (1.02 ~ 1.29)	3.09 (2.19 ~ 3.98)	-1.93	0.38
Full genome	Unvaccinated B.1.36.27 (N = 707)	1.66 (1.55 ~ 1.77)	3.07 (2.35 ~ 3.8)	-1.41	0.54
Full genome	Unvaccinated Delta (N = 70)	1.97 (1.7 ~ 2.23)	2.90 (2.37 ~ 3.43)	-0.93	0.68
Full genome	Unvaccinated Omicron (N = 30)	1.67 (1.32 ~ 2.02)	2.79 (2.06 ~ 3.53)	-1.12	0.60
Full genome	CoronaVac Delta (N = 14)	1.61 (1.16 ~ 2.06)	3.11 (2.04 ~ 4.18)	-1.50	0.52
Full genome	CoronaVac Omicron (N = 32)	1.74 (1.4 ~ 2.08)	2.56 (1.89 ~ 3.24)	-0.82	0.68
S	Combined (N=2053)	1.93 (1.72 ~ 2.14)	4.92 (2.34 ~ 7.51)	-2.99	0.39
S	Comirnaty Delta (N = 58)	2.49 (1.81 ~ 3.16)	2.39 (1.37 ~ 3.41)	0.09	1.04
S	Comirnaty Omicron (N = 68)	2.62 (1.84 ~ 3.4)	6.56 (4.49 ~ 8.63)	-3.94	0.40
S	Unvaccinated Alpha (N = 48)	4.66 (3.16 ~ 6.15)	0.61 (0 ~ 1.22)	4.05	7.64
S	Unvaccinated B.1.1.63 (N = 805)	1.41 (1.17 ~ 1.65)	1.66 (1.29 ~ 2.03)	-0.25	0.85
S	Unvaccinated B.1.36 (N = 221)	1.84 (1.46 ~ 2.22)	9.84 (3.33 ~ 16.3)	-8.00	0.19
S	Unvaccinated B.1.36.27 (N = 707)	2.06 (1.78 ~ 2.33)	7.26 (2.04 ~ 12.5)	-5.21	0.28
S	Unvaccinated Delta (N = 70)	2.34 (1.45 ~ 3.24)	0.48 (0.13 ~ 0.82)	1.87	4.93
S	Unvaccinated Omicron (N = 30)	1.48 (0.71 ~ 2.24)	2.45 (0.68 ~ 4.23)	-0.97	0.60
S	CoronaVac Delta (N = 14)	0.81 (0.23 ~ 1.4)	3.39 (0.87 ~ 5.91)	-2.58	0.24
S	CoronaVac Omicron (N = 32)	1.8 (0.95 ~ 2.65)	7.91 (4.73 ~ 11.1)	-6.11	0.23

Table 5. Candidate genomic regions of positive selection within hosts.

Nucleotide range	Codon range	Gene region	Codons with nonsynonymous differences	P-value	π_N	π_S	π_N/π_S	Number of overlapping CD8 epitopes	P-value	Number of overlapping CD4 epitopes	P-value	Number of overlapping CD4 and CD8 epitopes	P-value
10172 to 10201	3303 to 3312 (40 to 49)	ORF1ab (nsp5)	3304, 3307, 3308, 3310, 3311, 3312	0.00	8.06E-05	0	-	1	0.6975	0	1	1	0.7557
4061 to 4252	1266 to 1329 (448 to 511)	ORF1ab (nsp3)	1272, 1281, 1282, 1287, 1291, 1293, 1295, 1297, 1299, 1301, 1306, 1311, 1313, 1314, 1316, 1318, 1319, 1321, 1323	0.01	5.21E-05	9.9E-06	5.26	8	0.3285	0	1	8	0.4449
2246 to 2278	661 to 671 (481 to 491)	ORF1ab (nsp2)	662, 664, 669, 671	0.05	3.8E-05	0	-	1	0.7138	0	1	1	0.7704
27520 to 27564	43 to 57	ORF7a	47, 50, 57	0.06	1.75E-05	0	-	5	0.0721	0	1	5	0.2793
20452 to 20457	6730 to 6731 (278 to 279)	ORF1ab (nsp15)	6731	0.11	2.63E-05	0	-	1	0.5125	0	1	1	0.5948
26362 to 26385	40 to 47	E	41, 45	0.13	3.76E-05	1.57E-05	2.39	3	0.6761	4	0.9718	7	0.8028
3683 to 3733	1153 to 1156 (335 to 336)	ORF1ab (nsp3)	1153, 1156	0.16	7.36E-06	0	-	2	0.2708	0	1	2	0.3804

The P-value for T cell epitope overlap is defined as the probability of observing at least the same number of overlapping epitopes in the gene's window of same length as the codon range.

Table 6. Number of samples with at least one detected iSNV and number of total analysed samples, stratified by virus lineages and vaccination status.

	Comirnaty	CoronaVac	Unvaccinated
Alpha	0	0	30/48
Delta	51/56	10/12	50/70
Omicron	51/57	24/26	21/24
B.1.36	0	0	123/220
B.1.36.27	0	0	403/697
B.1.1.63	0	0	354/760

Supplementary Figures

Figure 1. The sequencing depth in sliding window of 200bp of the samples included in this study. Each grey line is representative of one individual sample, and the red line shows the average of all samples. The dashed line showed depth of 100 reads.

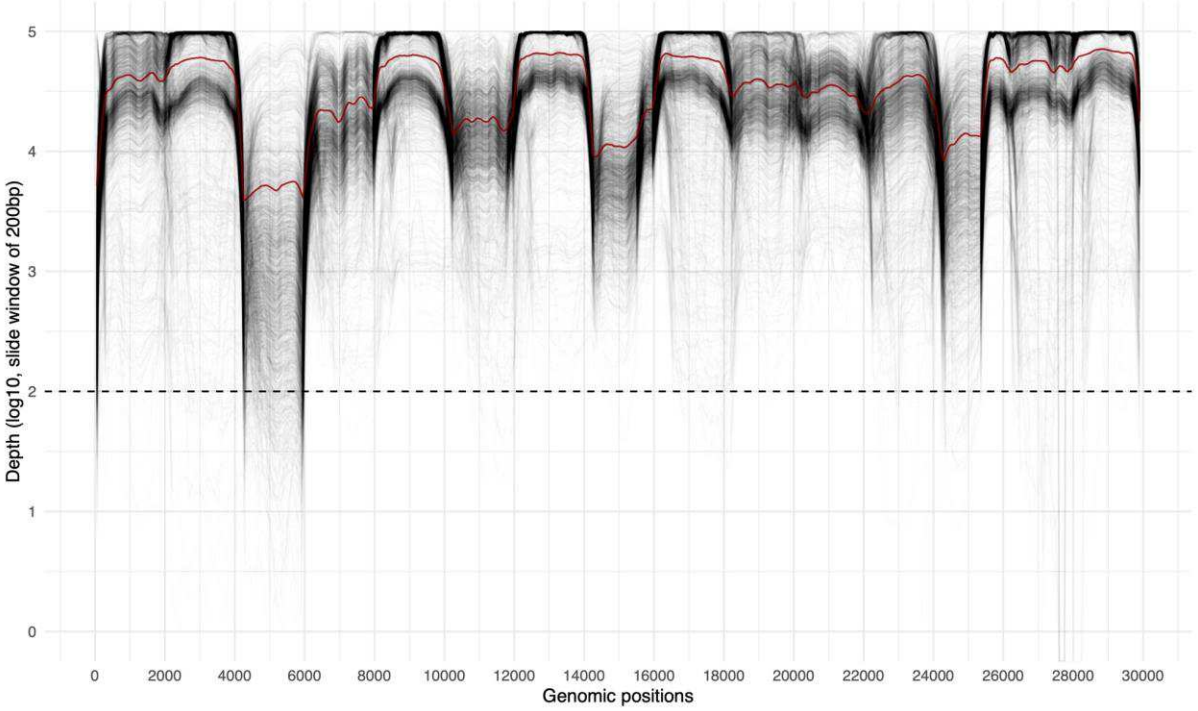


Figure 2. Correlation of Ct value between different factors. The regression lines are showed in blue. **(A)** Correlation between Ct value and number of iSNVs per Kb; **(B)** Correlation between Ct value and detection lag (time post symptom onset in days); **(C)** Correlation between detection lag and number of iSNVs per Kb; **(D)** Correlation between Ct value and minor allele frequency; **(E)** Correlation between Ct value and number of iSNVs per Kb (adjusted).

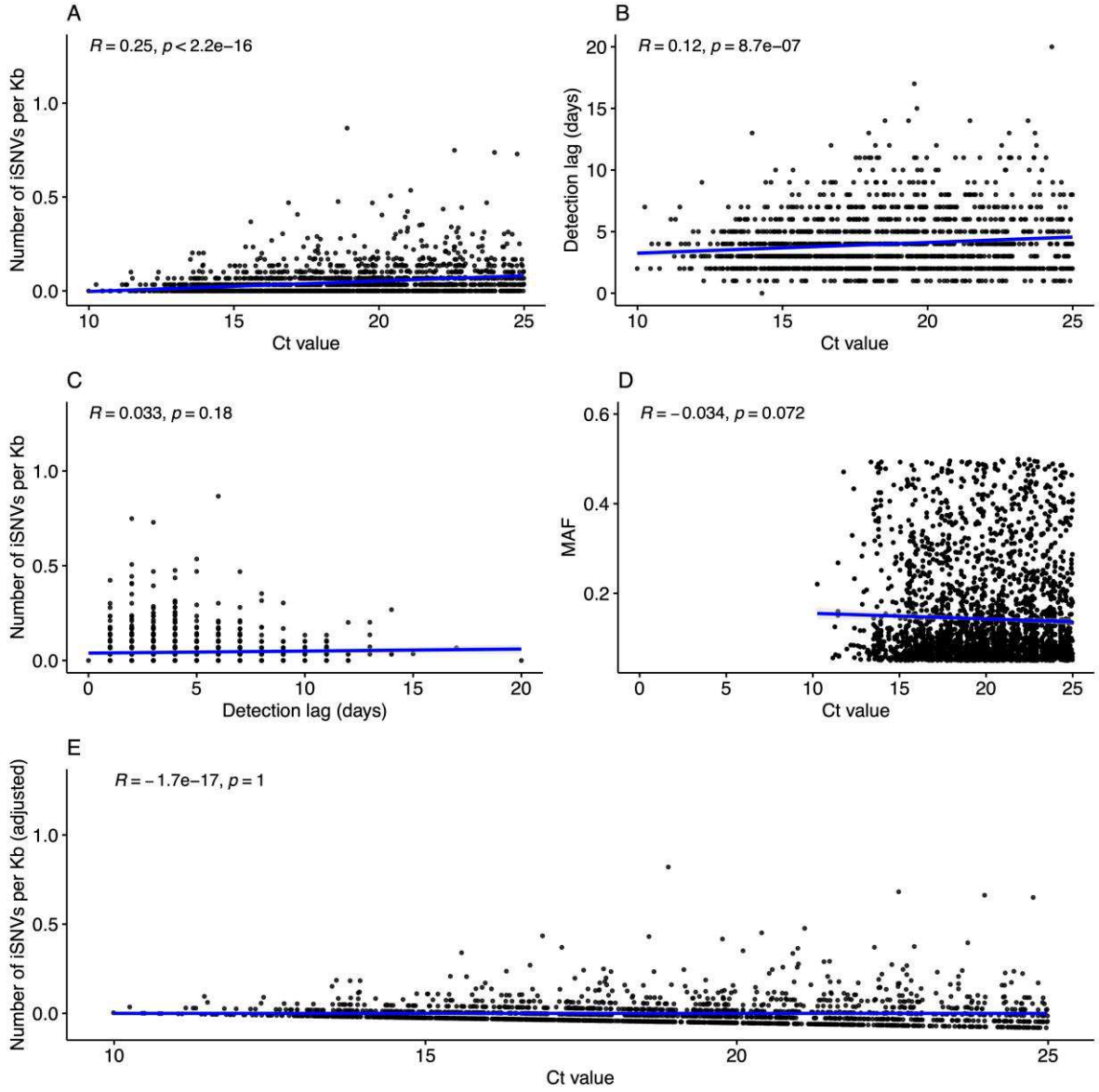


Figure 3. Within-host mutations profiles among different groups. Different vaccination statuses were compared, the data used here are the same as the data used in the Figure 2A-2C and Figure 3A-3C of the main text. Boxplots indicate median and inter-quartile ranges (IQR), and whiskers represent value ranges up to 1.5 * IQR. Pairwise comparisons within groups were tested by two-sided two-sample Wilcoxon tests, the pairs with P value ≤ 0.01 and ≤ 0.05 were labelled with "**" and "*" respectively. **(A)** Full-genome incidence of iSNVs (adjusted number of iSNVs per Kb) of different samples. **(B)** Full-genome abundance of iSNVs (minor allele frequencies) of different samples. **(C)** Full-genome nucleotide diversity (π) of different samples.

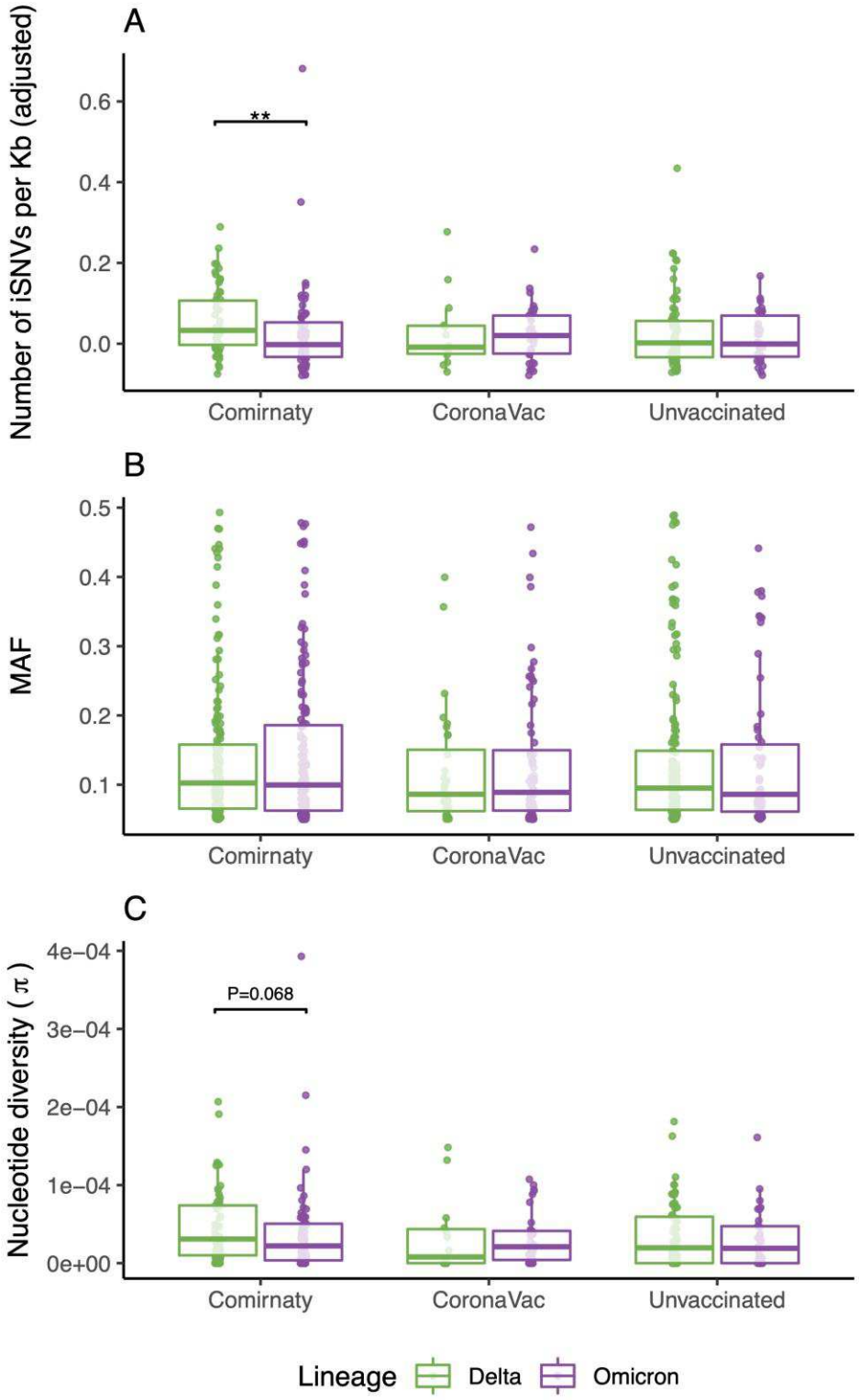


Figure 4. Distribution of time post last dose in the vaccinated samples.

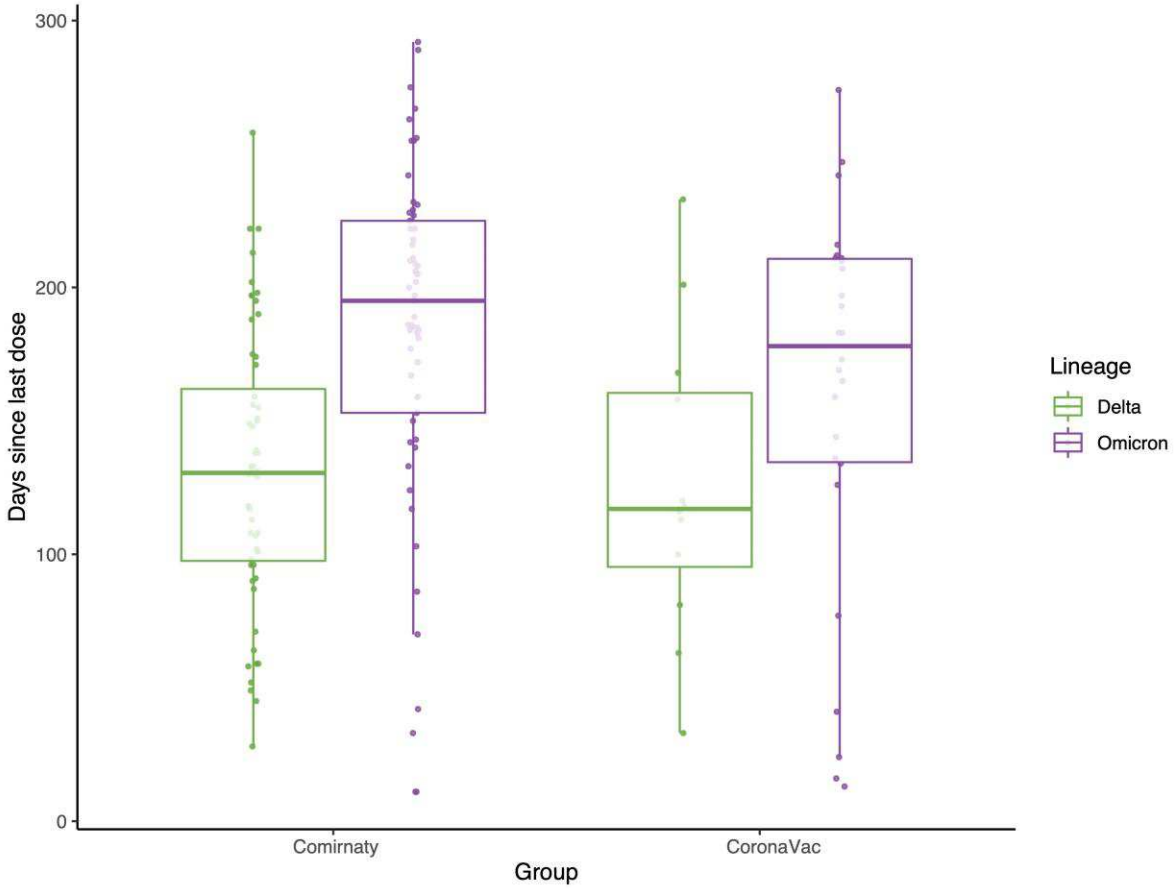


Figure 5. Sliding window analysis of Synonymous/Non-synonymous nucleotide diversity in different genes. Sliding windows size of thirty codons and step size of one codon were used because this did not exceed the length of ORF10 (thirty-nine codons).

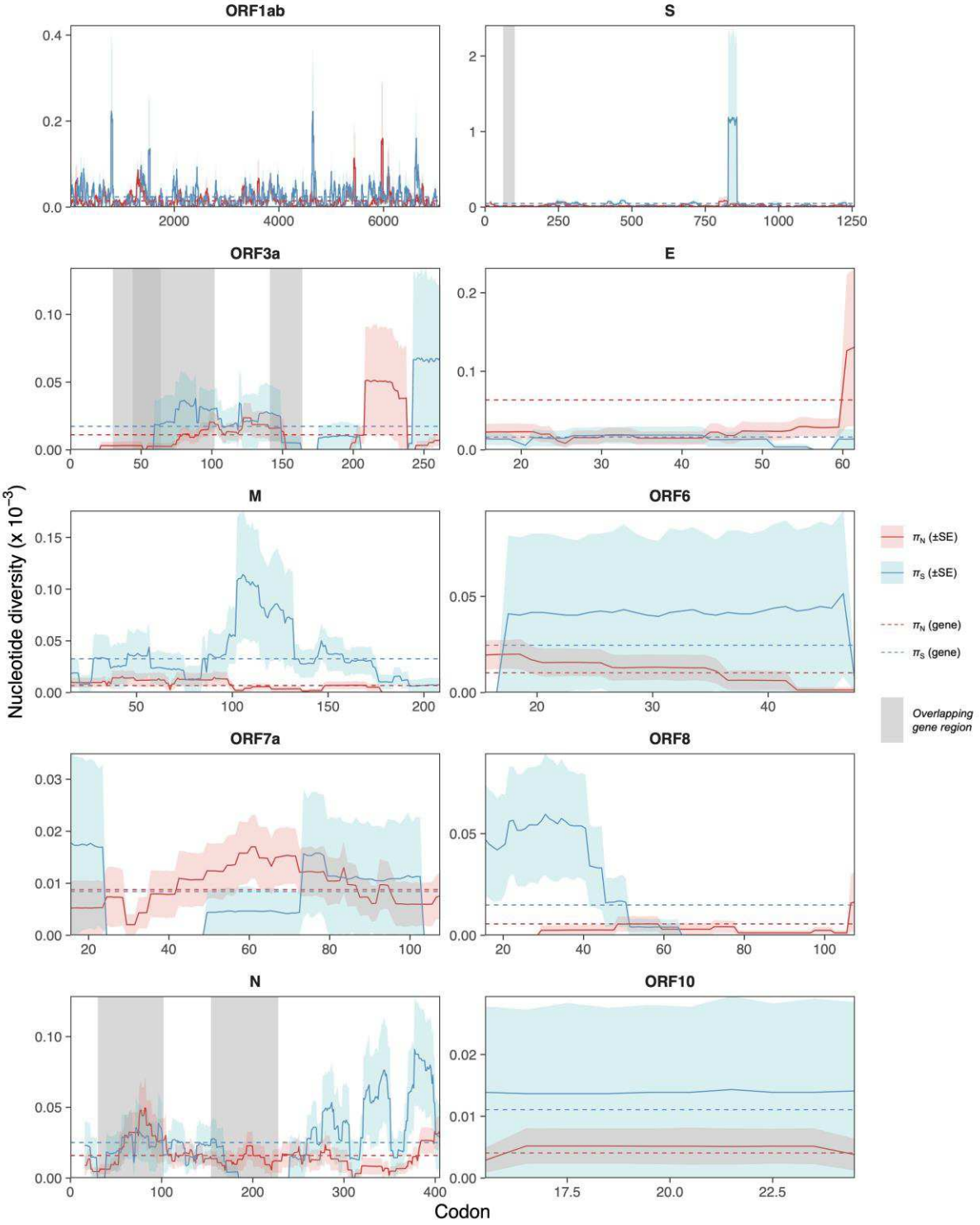


Figure 6. Antibody binding prediction of receptor-binding domain (RBD) mutations in vaccinated (A) Omicron and (B) Delta samples. The identified mutations in both Comirnaty and CoronaVac vaccinated samples were labelled in the plot (orange dots). The blue/grey lines show the total antibody binding before/after the mutations are introduced into the RBD region. The difference in y axis between orange and grey dots at the same amino acid site represents the loss of antibody binding under mutation, the differences in percentage were labelled in brackets. The calculations are based on deep mutational scanning of a large set of RBD targeting antibodies which are known to neutralize Wuhan-Hu-1. (https://jbloomlab.github.io/SARS2_RBD_Ab_escape_maps/escape-calc/).

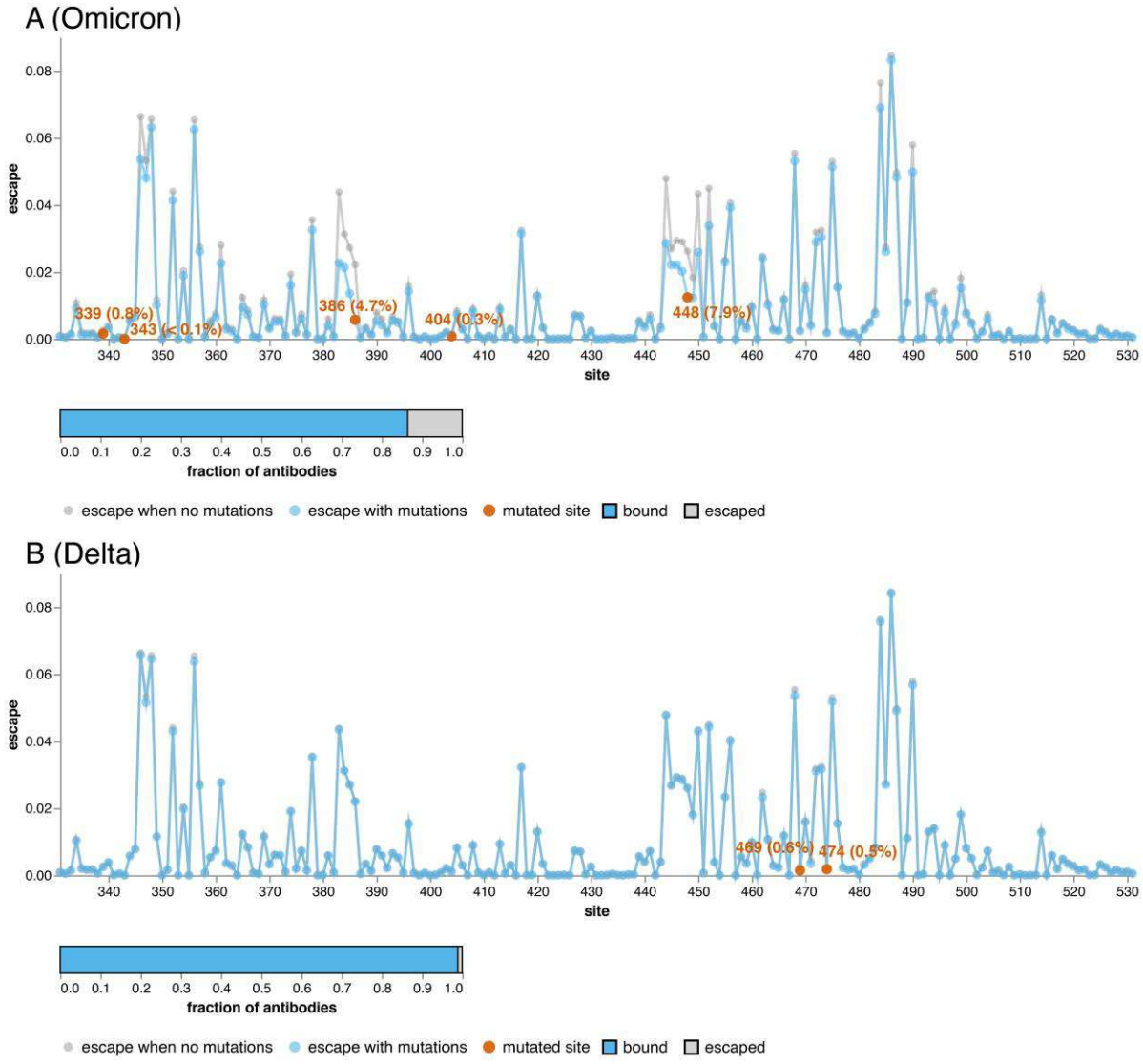


Figure 7. Distribution of CD4+/CD8+ T cell epitope-HLA pairs included in this study. **(A)** Distribution in different genomic regions; **(B)** Distribution in different HLA regions.

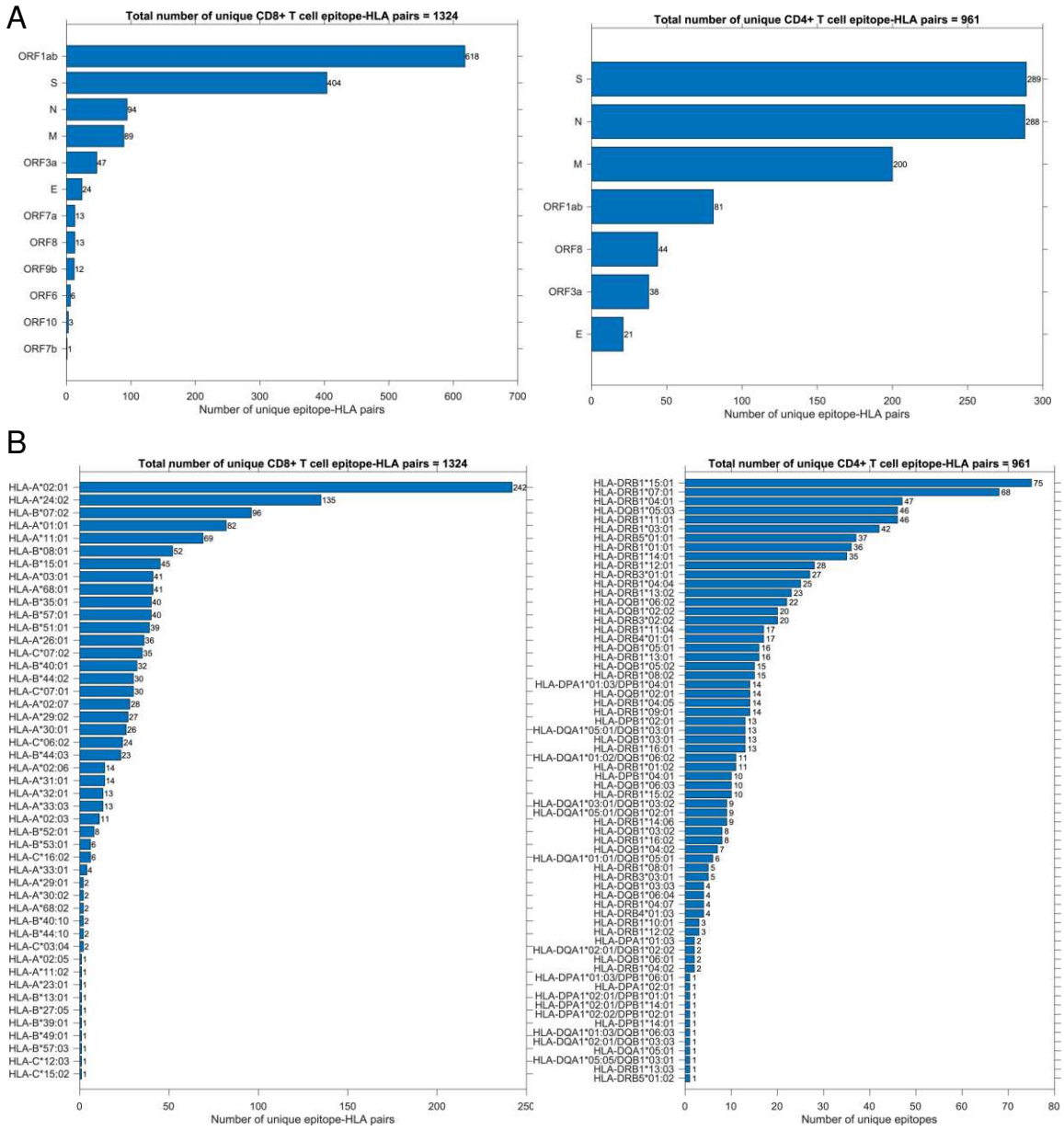


Figure 8. Average number of overlapping epitopes per mutation in different groups.

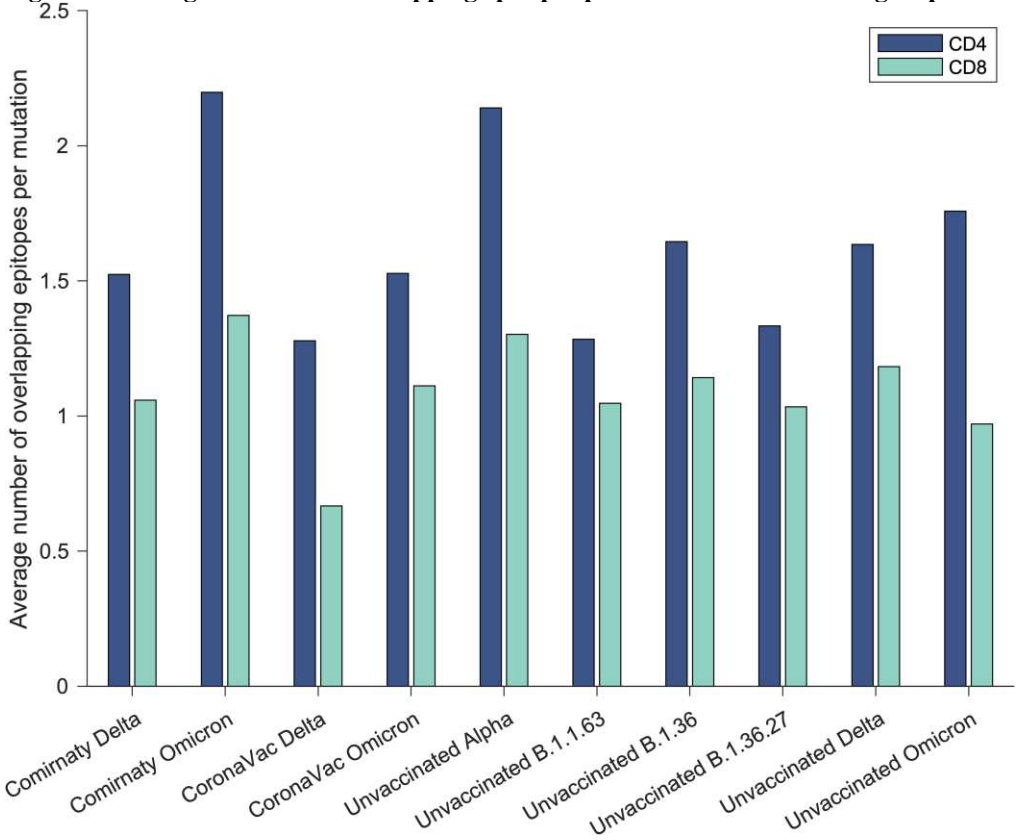


Figure 9. Overlapping of CD4+/CD8+ T cell epitopes per mutation in Spike between vaccinated and unvaccinated samples. (A) analysis based on unique T cell epitopes. **(B)** analysis based on epitope-HLA pairs specific to Hong Kong population.

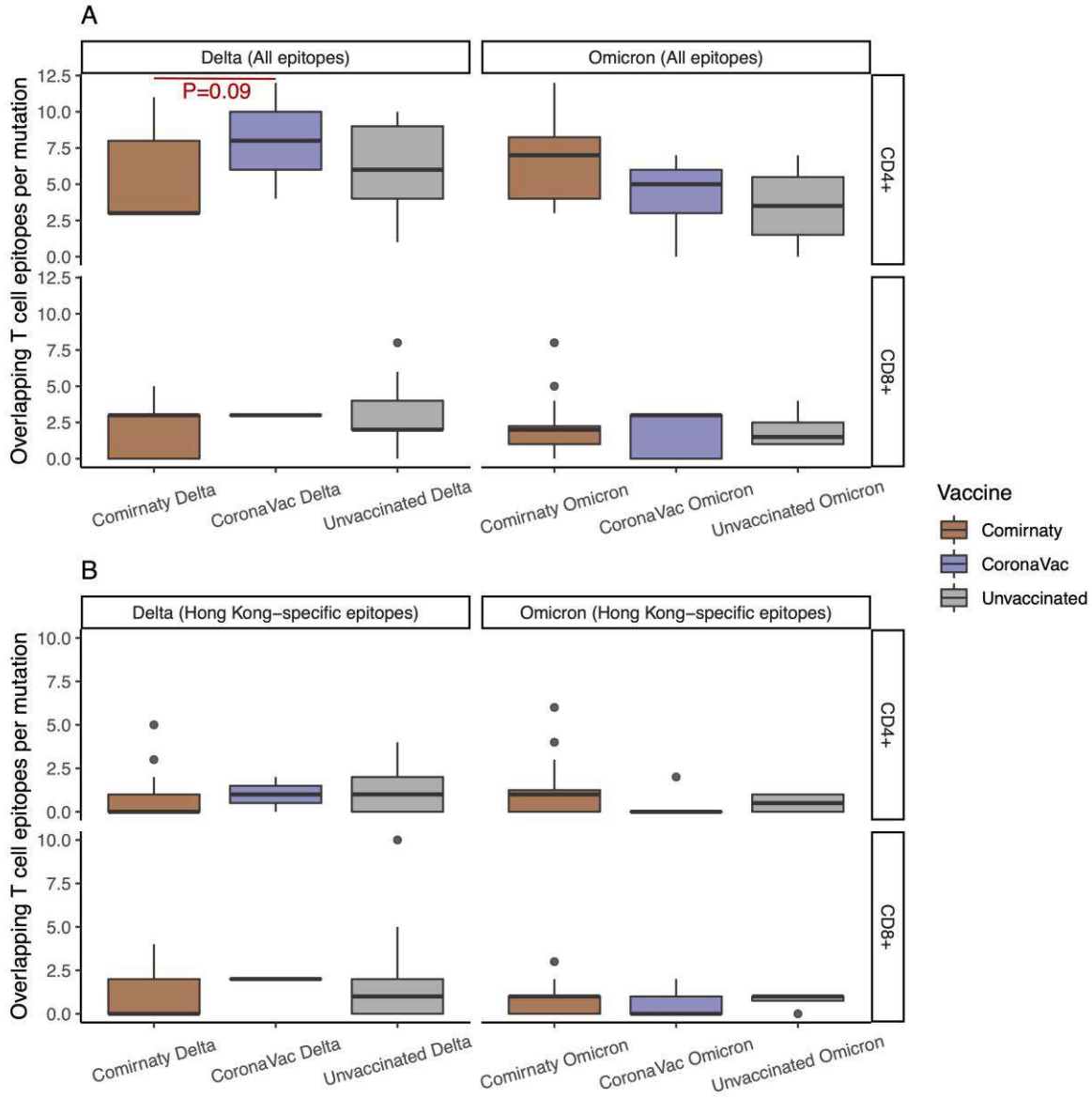
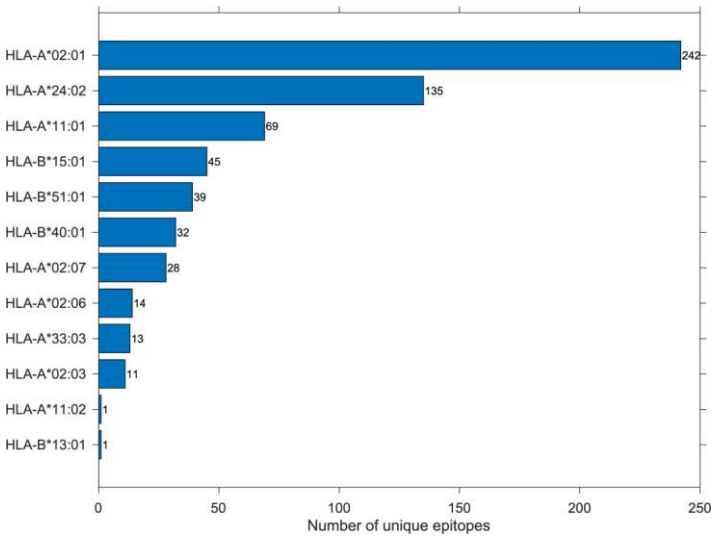


Figure 10. Distribution of CD4⁺/CD8⁺ T cell epitope-HLA pairs specific to Hong Kong population

A

No.	HLA allele	Coverage in HK population (%)#
1	HLA-A*11:01	49.12
2	HLA-A*24:02	28.4
3	HLA-B*40:01	28.01
4	HLA-A*02:07	24.46
5	HLA-A*33:03	18.98
6	HLA-A*02:03	14.98
7	HLA-B*13:01	14.95
8	HLA-A*02:01	12
9	HLA-A*02:06	9.17
10	HLA-B*51:01	8.39
11	HLA-A*11:02	7.83
12	HLA-B*15:01	4.34

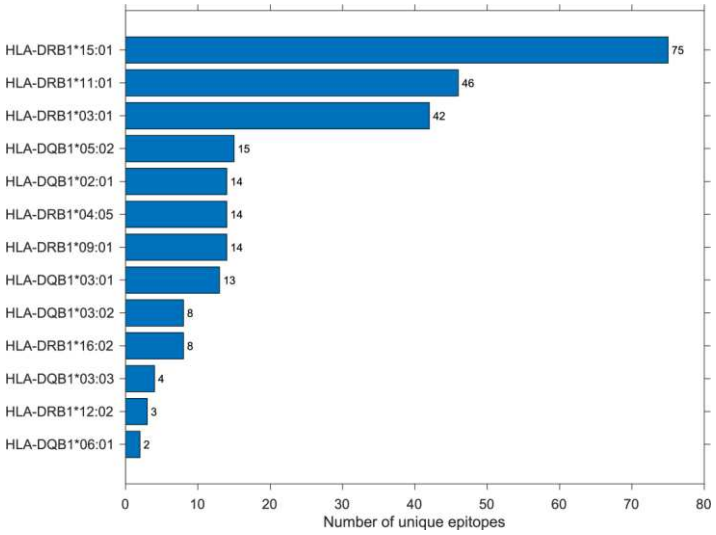
#Source: IEDB



B

No.	HLA allele	Coverage in HK population (%)#
1	HLA-DQB1*03:01	23.55
2	HLA-DQB1*03:03	15.75
3	HLA-DRB1*09:01	15.37
4	HLA-DRB1*12:02	14.39
5	HLA-DQB1*05:02	11.76
6	HLA-DQB1*06:01	11.27
7	HLA-DRB1*15:01	9.65
8	HLA-DRB1*03:01	6.79
9	HLA-DQB1*02:01	6.69
10	HLA-DQB1*03:02	6.29
11	HLA-DRB1*04:05	6.21
12	HLA-DRB1*08:03	5.94
13	HLA-DRB1*11:01	5.3

#Source: AFND (Population coverage > 5%)



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- reportingsummary.pdf
- checklist.pdf