nature ecology & evolution

Article

https://doi.org/10.1038/s41559-023-02123-8

Immune evasion and ACE2 binding affinity contribute to SARS-CoV-2 evolution

Received: 6 February 2023

Accepted: 13 June 2023

Published online: 13 July 2023

Check for updates

Wentai Ma[1,2], Haoyi Fu[1,2], Fanchong Jian[3], Yunlong Cao[3,4] ✉ & Mingkun Li[1,2] ✉

Mutations in the SARS-CoV-2 genome could confer resistance to pre-existing antibodies and/or increased transmissibility. The recently emerged Omicron subvariants exhibit a strong tendency for immune evasion, suggesting adaptive evolution. However, because previous studies have been limited to specific lineages or subsets of mutations, the overall evolutionary trajectory of SARS-CoV-2 and the underlying driving forces are still not fully understood. Here we analysed all open-access SARS-CoV-2 genomes (up to November 2022) and correlated the mutation incidence and fitness changes with the impacts of mutations on immune evasion and ACE2 binding affinity. Our results show that the Omicron lineage had an accelerated mutation rate in the RBD region, while the mutation incidence in other genomic regions did not change dramatically over time. Mutations in the RBD region exhibited a lineage-specific pattern and tended to become more aggregated over time, and the mutation incidence was positively correlated with the strength of antibody pressure. Additionally, mutation incidence was positively correlated with changes in ACE2 binding affinity, but with a lower correlation coefficient than with immune evasion. In contrast, the effect of mutations on fitness was more closely correlated with changes in ACE2 binding affinity than with immune evasion. Our findings suggest that immune evasion and ACE2 binding affinity play significant and diverse roles in the evolution of SARS-CoV-2.

Recent SARS-CoV-2 lineages (for example, BA.2 sublineages and BA.5 sublineages) have an average of over 80 mutations relative to the earliest genome (www.nextstrain.org). Some variants exhibit significant alterations in their transmissibility, antigenicity and pathogenicity compared with their predecessors[1–3]. The most notable are those referred to as variants of concern, including Alpha, Beta, Delta, Gamma and Omicron, which are more transmissible or able to escape pre-existing immune pressures and thus led to multiple surges of infection peaks on local or global scales.

Intrinsic transmissibility and immune pressure have been proposed as the two primary forces driving the evolution of SARS-CoV-2 as well as other viruses[4]. For example, the rapidly spreading D614G and N501Y mutations could enhance viral transmission by increasing ACE2 binding affinity[5,6], while E484K reduces susceptibility to neutralizing antibodies[7,8]. Moreover, with the development and optimization of the high-throughput deep mutational scanning (DMS) method, it became more feasible to assess the effect of RBD mutations in the binding affinity to antibodies[9,10] and the human ACE2 receptor[11,12], which facilitated the identification of a number of RBD mutations in Omicron and its sublineages that conferred significant immune evasion against antibodies induced by prior infections or vaccinations, while maintaining sufficient binding affinity to human ACE2 (refs. 11,13). Recent studies

[1]Beijing Institute of Genomics, Chinese Academy of Sciences, and China National Center for Bioinformation, Beijing, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing, China. [4]Changping Laboratory, Beijing, China. ✉e-mail: yunlongcao@pku.edu.cn; limk@big.ac.cn

have proposed significant enrichment of mutations in the RBD region in the Omicron lineages and unprecedented convergent evolution of BA.2 and BA.4/5 subvariants (for example, BQ.1, XBB and BM.1), which enable a near-complete evasion against most known antibodies, emphasizing the significant role of immune pressure in SARS-CoV-2 evolution[14–17]. However, previous studies mainly focused on a few fast-growing variants and mutations that were known to have a great impact on immune evasion. These studies thus may suffer from survival bias—that is, the findings may not be applicable to other variants that account for a large fraction of the data. A thorough investigation of all mutations, including those that do not define any virus lineage and are found in a small number of samples, is therefore needed to disentangle the evolutionary trajectory and underlying evolutionary driving forces of SARS-CoV-2.

There are other unanswered scientific questions concerning the evolution of SARS-CoV-2. For example, did adaptive evolution occur only in the RBD region, which is the primary target of neutralizing antibodies and to which human ACE2 binds? Did the SARS-CoV-2 evolutionary pattern vary over time and between different lineages? Finally, do the two factors—immune pressure and intrinsic transmissibility—contribute equally to viral evolution, and do their effects change over time? The evolution of SARS-CoV-2 would be better understood if these questions could be answered.

In this study, we have investigated the mutations inferred from more than six million open-access SARS-CoV-2 sequences as of 23 November 2022 and correlated the mutation spectrum and incidence with their immune-evasive and ACE2 binding potentials estimated from the DMS and neutralization data. We found that different SARS-CoV-2 macro-lineages exhibited distinct mutation patterns in the RBD region that are likely to be associated with continually changing humoral immune pressures, and immune-pressure-driven mutations became more evident in the recent BA.2 and BA.4/5 sublineages. ACE2 binding affinity also played a significant role in the evolution of the virus, especially at the early stage after the emergence of new variants when the humoral immune pressure was relatively low. Although immune pressure was more correlated with the occurrence of mutations than was ACE2 binding affinity, it is interesting to note that enhanced ACE2 binding affinity was more closely correlated with increased virus fitness than immune evasion.

Results

The accelerated mutation rate in the SARS-CoV-2 RBD region

Since mid-2022, a large number of immune-evasive BA.2 and BA.4/5 subvariants have emerged in the population[15,18], implying that the evolution of SARS-CoV-2 might be accelerating. To verify this hypothesis, we retrieved 200 sequences per month between January 2020 and November 2022 to estimate the mutation rate of SARS-CoV-2 over time. The automatic piecewise linear regression analysis found two turning points associated with a dramatic rise in the mutation number (Fig. 1a), corresponding to the emergence of the Alpha and Omicron variants, which had far more mutations than their possible predecessors. The slopes of the regression lines before and after the turning points were similar, indicating that the mutation rate did not significantly increase over time. Meanwhile, the analysis performed on different variants of concern confirmed that the mutation rate did not increase in the latest Omicron sublineages (Fig. 1b). Interestingly, we noted that the synonymous mutation rate tended to increase while the non-synonymous mutation rate tended to decrease in the latest Omicron variants (Extended Data Fig. 1). However, we found that the Omicron variants showed an accelerated rate of amino acid changes compared with previous viral lineages in the RBD region (3- to 83-fold higher) (Fig. 1c).

To further verify the accelerated mutation rate in the RBD region and test whether other genes were also subject to a higher mutation rate in recent Omicron sublineages, we retrieved all mutations inferred from the UShER mutation-annotated tree constructed from 6,484,070 complete high-quality SARS-CoV-2 genomes[19]. We found that the proportion of mutations in the RBD region increased over time, which was mainly due to the excess number of non-synonymous mutations (Fig. 1d)—that is, the proportion of non-synonymous mutations in the RBD region increased from 52.8% in the early stage of the pandemic to 62.1% in the recent Omicron sublineages. However, the same tendency was not observed in other genes or other regions in the *S* gene.

RBD mutations showed a lineage-specific pattern

To further investigate the evolution of the sequences in the RBD region, we analysed the distribution and incidence of all amino acid mutations identified in this region, which included 855 mutations that accounted for 6,328 mutation events. The mutations were inferred by comparing the sequence with its putative direct ancestor sequence identified from the phylogenetic tree (see the Methods for more details). First, we found that the incidence differed among different mutations. The overall proportion of mutations that accounted for 50% of all mutation events (P50) was 7.5%, and the proportion decreased over time (Fig. 2a), suggesting that mutations tended to become more aggregated in recent lineages. Second, the high-frequency mutations tended to be shared among variants belonging to the same macro-lineages. The viral lineages (with ≥6,000 sequences) could be classified into four clusters according to the similarity of the incidence of 59 mutations that showed the highest incidences (top five) in at least one lineage (Fig. 2b,c). These four clusters correspond to four viral macro-lineages (B.1, Delta, BA.1/2 and BA.4/5), suggesting that variants belonging to the same lineages, including those circulating at the same time, tended to have the same high-incidence mutations, which is a signature of convergent evolution. Third, mutations occurring at the same position could be of different types. Forty high-frequency mutations were found at 16 positions (Extended Data Fig. 2), while the other 19 high-frequency mutations were found at 19 distinct positions. Different high-frequency mutations at the same position differed in amino acid polarity, acidity and charge, which may result in varying effects on the affinity between the virus and the host cell and antibodies. We also noted that distinct mutations at the same position could arise in variants belonging to the same lineage, implying that the virus may react differently to the same pressures. Additionally, we noted that the convergence among the most frequent mutations was unlikely to be explained by RNA editing because A-to-I and C-to-U mutations were not enriched among those mutations (Extended Data Fig. 3).

The high-incidence mutations in the B.1 and Delta clusters were similar, as indicated by the high incidence of the S373L, P384L, A522V, E484Q, S477I and G446V mutations compared with the Omicron clusters; the first three were more frequently observed in the B.1 cluster, and the latter three were more frequently observed in the Delta cluster. In particular, the mutation G446V was 2.1 times more frequent in the Delta cluster than in the B.1 cluster. The difference between the BA.1/2 and BA.4/5 clusters was more remarkable than that between the B.1 and Delta clusters, even though some of the BA.2 and BA.4/5 subvariants were circulating at the same time in the population. BA.1/2 had a higher incidence of E484V and Q493L, while BA.4/5 had a higher incidence of K444R/N and R346I/T.

To test whether convergent mutation also occurred in other genomic regions, the same analysis was conducted on all genes in the SARS-CoV-2 genome. We found that the same trend was not observed in other genes, as indicated by an average silhouette value less than 0, which indicates that the similarity of the mutation pattern within the same lineages is lower than that between different lineages (Fig. 2d). The NTD region and other regions in the *S* gene showed signs of convergent mutations in some lineages, but the magnitude was much lower than that in the RBD region.

Immune pressure is correlated with SARS-CoV-2 evolution

To investigate whether the convergent evolution in the RBD region can be explained by a similar humoral immune pressure from antibodies

https://doi.org/10.1038/s41559-023-02123-8

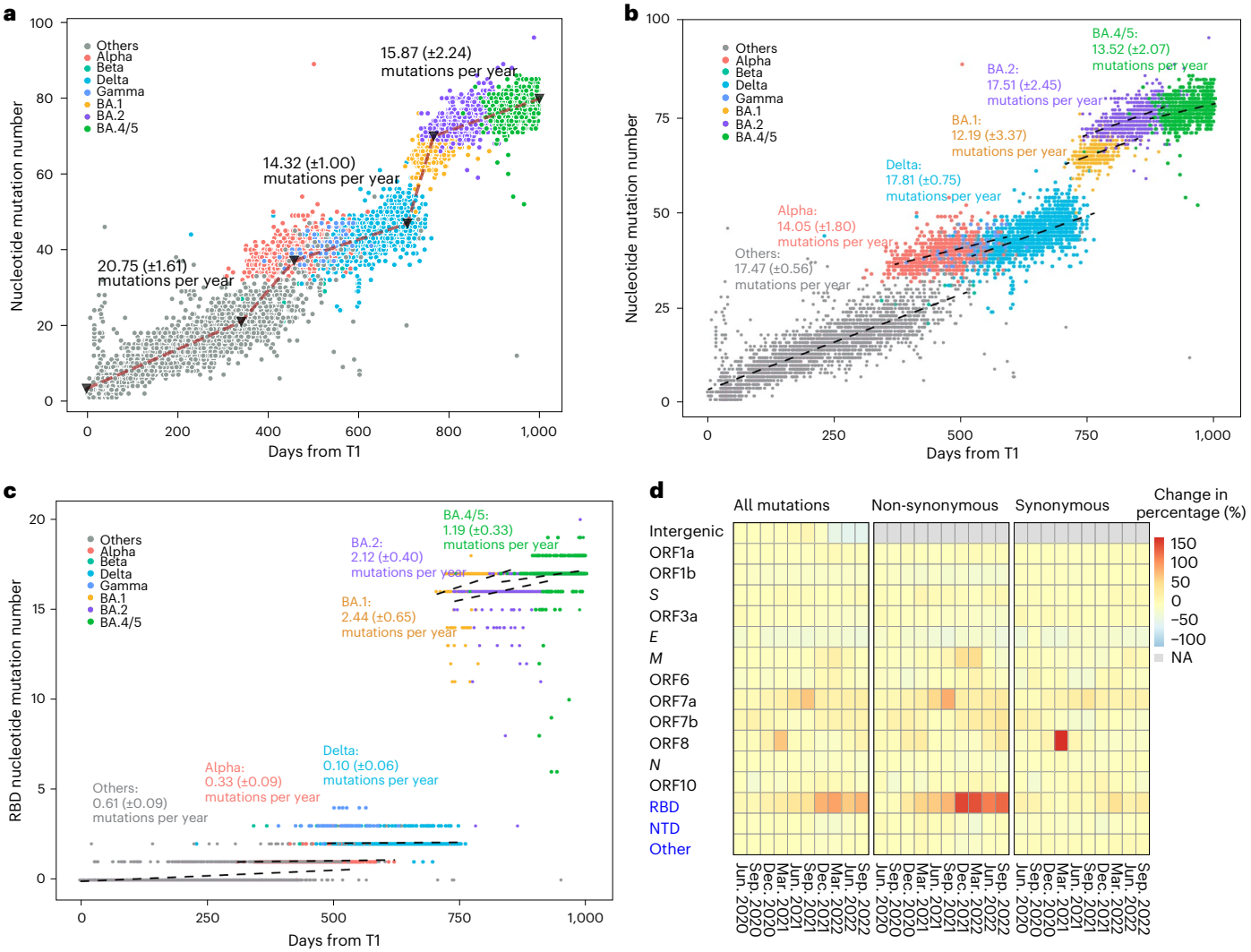


Fig. 1 | The mutations in the SARS-CoV-2 genome. a, The correlation between mutation count and collection date. Two hundred sequences were randomly selected from each month on the basis of the collection date. The segmented regression line was fitted using automatic piecewise linear regression, and the mutation rate was estimated as the slope of the regression line. We repeated the sampling 100 times and added the resulting median and standard deviation of the estimated mutation rate above the regression line. T1 represents 24 December 2019, which is the collection date of the first open-access SARS-CoV-2 sequence. **b**, The correlation between mutation count and collection date for major lineages. The estimated mutation rates for six major lineages are shown above the linear regression lines. **c**, The mutation rate of the RBD region (residues 331–531 of the *S* gene) in major lineages. **d**, The distribution of mutations across different genomic regions over time. The three regions (RBD, NTD and other regions) in the *S* gene are shown separately (marked in blue). The time window size is three months, with the first three months (March, April and May 2020) used as a reference (samples collected prior to March are limited and were not included in the analysis). The cell colours indicate the degree of change relative to the reference. NA, not applicable.

acting on the same macro-lineage, we classified the mutations into immune escape mutations and non-immune-escape mutations according to the DMS scores against over 2,000 antibodies that belong to 12 epitope groups (hereafter referred to as antibody types). A mutation that significantly reduced the affinity to any of the 12 antibody types was defined as an immune escape mutation. First, we noted that the proportion of immune escape mutations was higher in the Omicron lineages than in other earlier lineages (median proportion of immune escape mutations in two macro-lineages, 67.0% versus 47.3%; $P < 0.001$, Wilcoxon rank sum test). Meanwhile, the spectrum of immune evasion caused by mutations changed markedly over time and tended to be more concentrated on specific antibody types in recent lineages (Fig. 3a). Specifically, mutations escaping D1 and D2 antibody types became more enriched in the Delta lineages than in the earlier lineages; Omicron BA.1/2 sublineages exhibited an increased proportion of mutations escaping A, C, E2.1 and E2.2 antibodies; and Omicron BA.4/5 sublineages showed the strongest immune escape against D1 and D2 antibody types.

We noted that the variants with similar background RBD sequences (belonging to the same macro-lineage) tended to have mutations that evaded the same antibody types (Fig. 3b), suggesting that the immune pressure was similar among variants in the same macro-lineage. We also found that immune pressure was altered with the accumulation of immune escape mutations. For example, the variant BA.2.12.1 acquired an extra L452Q mutation compared with its predecessor BA.2 variant; this mutation could facilitate escape from the E2 and D1 antibody types according to the DMS data, making its antigenicity more similar to that of BA.4/5, which had a L452R mutation that could evade the same antibody types. The mutation pattern of BA.2.12.1 is thus more similar to BA.4/5 subvariants than other BA.2 subvariants. Additionally, the variant BA.4.6 possessed an extra RBD mutation (R346T) compared with other BA.4 lineages, which is capable of compromising the efficacy

https://doi.org/10.1038/s41559-023-02123-8

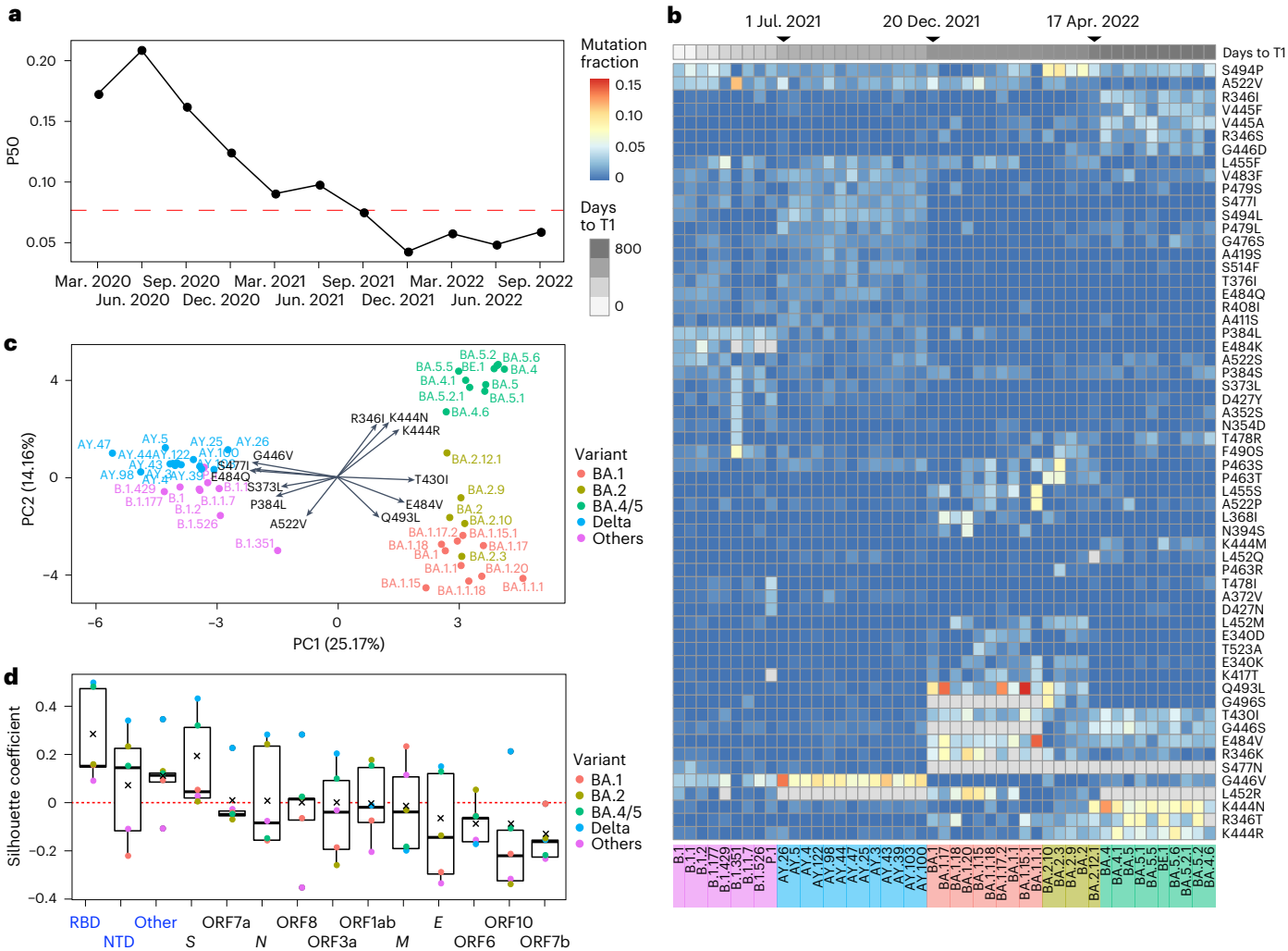


Fig. 2 | Convergent evolution of the RBD sequences in the SARS-CoV-2 genome. a, The P50 at different time points. The red dashed line indicates the P50 for all mutations. All available sequences were included in this and subsequent analyses. **b**, The mutation incidence in various SARS-CoV-2 lineages. The colours denote the ratio of each mutation's frequency to the frequency of all mutations in the lineage. The top five most frequent mutations in each lineage are shown. Mutations that have been fixed in the lineage are labelled in grey. **c**, Principal component analysis plot of mutation incidence across different lineages. The input data were taken from **b**. The top three mutations that explain the highest variance in each quadrant are labelled in the figure. **d**, The clustering significance of different lineages based on mutation incidence. The silhouette coefficient of the clustering of lineages in various genomic regions is shown. The analysis included 40 lineages with over 6,000 sequences. The lineages are sorted by the collection date of the earliest 5% of sequences belonging to each lineage. To prevent errors caused by misplaced sequences in the phylogenetic tree, back/reverse mutations were excluded from the analysis. The centre line denotes the median value, the black cross represents the mean value, the box represents the interquartile range, the whiskers extend to the furthest data point in each direction that is within 1.5 times the interquartile range values, and the points represent different variants.

of D1, E1 and E2.1 antibody types; this variant thus displayed a distinctive pattern of immune-evasive mutations compared with other BA.4/5 lineages (Fig. 3a,b).

To further quantify the impact of humoral immune pressure on the immune-evasive mutations in the RBD region, we estimated the immune pressure on four major variants (wild type (WT), BA.1, BA.2 and BA.4/5) from each antibody type using pseudovirus-neutralization data (antibodies induced by different variants were retrieved from previous studies)[13,14]. A marginal correlation was found between the incidence of the immune escape mutation and the immune pressure on the WT variants (Fig. 3c). The correlation became more significant in the Omicron lineages, and the correlation coefficients increased over time, suggesting a stronger immune pressure on recent lineages, which may be related to the increased antibody prevalence in the population as a result of mass vaccination campaigns and infections. However, the correlation between immune pressure and mutation incidence was less significant at the individual mutation level (Extended Data Fig. 4), which was consistent with the results of a previous study[9]. This suggests that other factors, such as codon preference, epistatic effects and RNA editing, may also influence the occurrence of the mutation. Nonetheless, gene set enrichment analysis (GSEA) indicated that the high-incidence mutations were more likely to confer a stronger resistance to the highly potent antibodies against the variant (Fig. 3d), and the tendency was more remarkable in the latest Omicron lineages (as indicated by lower rank and higher ES values), implying that the humoral immune pressure played a stronger role in the recent evolution of SARS-CoV-2.

ACE2 binding affinity contributes to SARS-CoV-2 evolution

Besides immune escape, increased transmissibility is another major direction of viral evolution[4,20]. One of the variables that correlates with viral transmissibility is ACE2 binding affinity, which determines how

https://doi.org/10.1038/s41559-023-02123-8

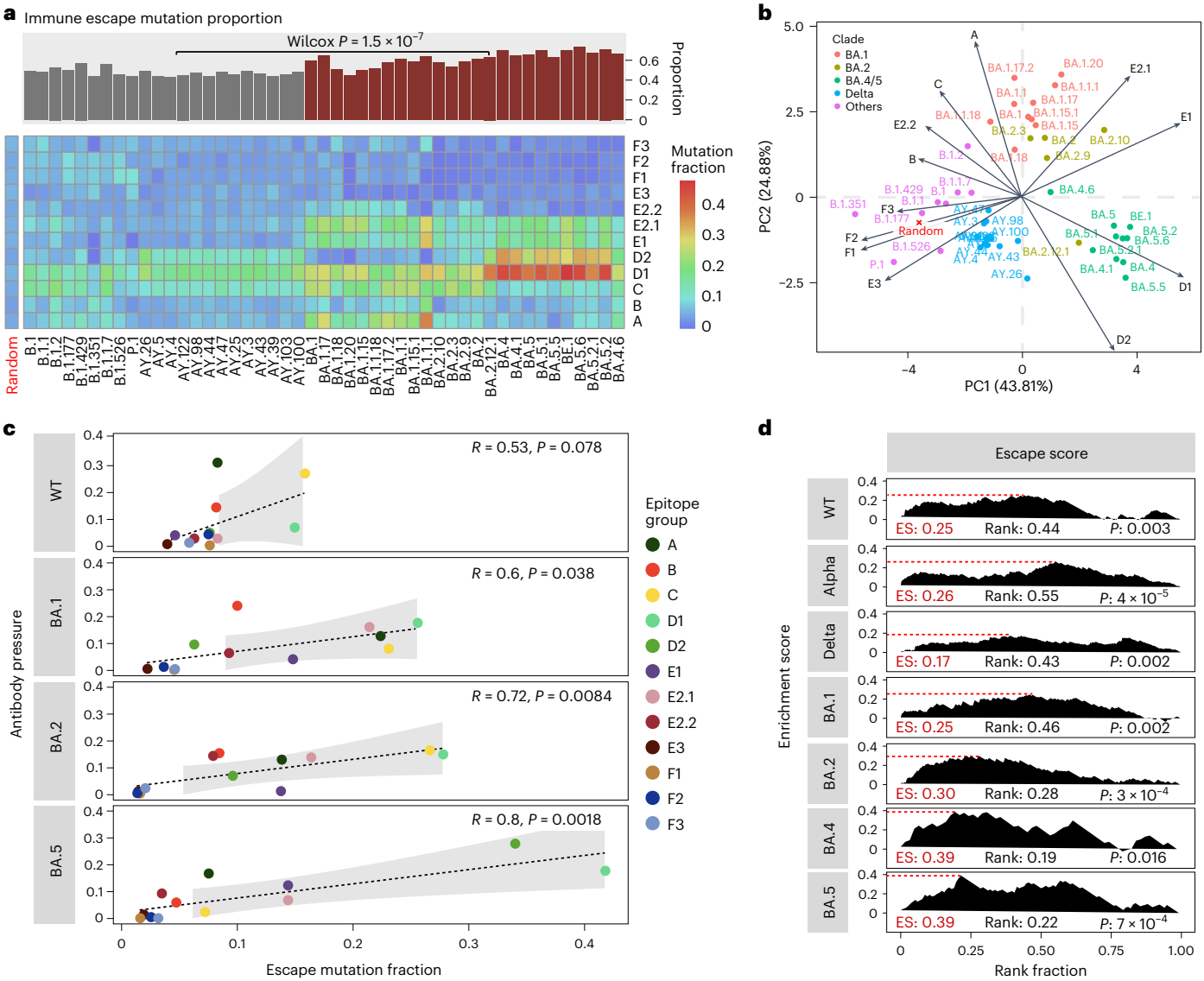


Fig. 3 | The correlation between the incidence of RBD mutations and humoral immune pressure. a, The distribution of escape mutations in 12 antibody epitope groups. The bar plot on top shows the proportion of immune escape mutations among all mutation events in different lineages. The null distribution of mutations, assuming no epitope group preference, is shown on the left side of the heat map. **b**, Principal component analysis plot of the mutation distribution in different lineages. The input data were taken from **a**. **c**, The correlation between the prevalence of escape mutations and antibody pressure in 12 epitope regions. The Pearson correlation coefficient (*R*) and two-sided unadjusted *P* value are shown in each plot. The *x* axis represents the proportion of immune escape mutations, which was calculated as the ratio of the incidence of escape mutations for that particular epitope group over the total escape mutations for all epitope groups. The *y* axis represents the proportion of immune pressure exerted on a particular epitope region, calculated by summing the neutralizing activities of all antibodies that belong to this epitope group. The shading represents the 99% confidence interval. **d**, The correlation between immune evasion capacity and mutation incidence. All mutations in each major lineage were reverse-sorted by their incidences, and GSEA was conducted to examine whether the high-weight escape mutations were enriched among the high-incidence mutations (the details are provided in the Methods). The ES value represents the highest cumulative score (the peak), the rank value represents the *x*-axis position of the peak, and the unadjusted *P* value was calculated through 1,000 runs of random reordering (how often a dataset with an ES value greater than the tested value was observed). WT has no mutations compared with the NC_045512.2 in the RBD region.

easily an infection/transmission is established. By analysing the effect of the mutation on ACE2 binding affinity under different genetic backgrounds (WT, Alpha, Delta, BA.1, BA.2 and BA.4/5), which was measured by high-throughput DMS screening[12], we found that 34% of the mutations that occurred in the WT had the potential to enhance ACE2 binding affinity, while the proportion decreased to 19% and 23% in the Alpha and Delta lineages, respectively (Fig. 4a). In the Omicron lineages, the proportion was 32% in the BA.1 sublineages and 33% in the BA.2 sublineages, and it dropped to 18% in the BA.4/5 sublineages. Meanwhile, the proportion of immune escape mutations also varied among different macro-lineages, with Omicron lineages having a higher proportion of immune escape mutations than earlier variants, such as R346K/T, K444N/R and E484V (Extended Data Fig. 2). The dynamic of the mutation pattern may reflect a shift in the force that drove the evolution of the virus.

When tracing the evolution of the two major variants B (excluding Omicron lineages) and BA.2 that spread over a long period of time, we found that mutations that occurred at different stages of transmission had varying impacts on immune evasion and ACE2 binding affinity. Specifically, the immune evasion capacity of the mutations increased

https://doi.org/10.1038/s41559-023-02123-8

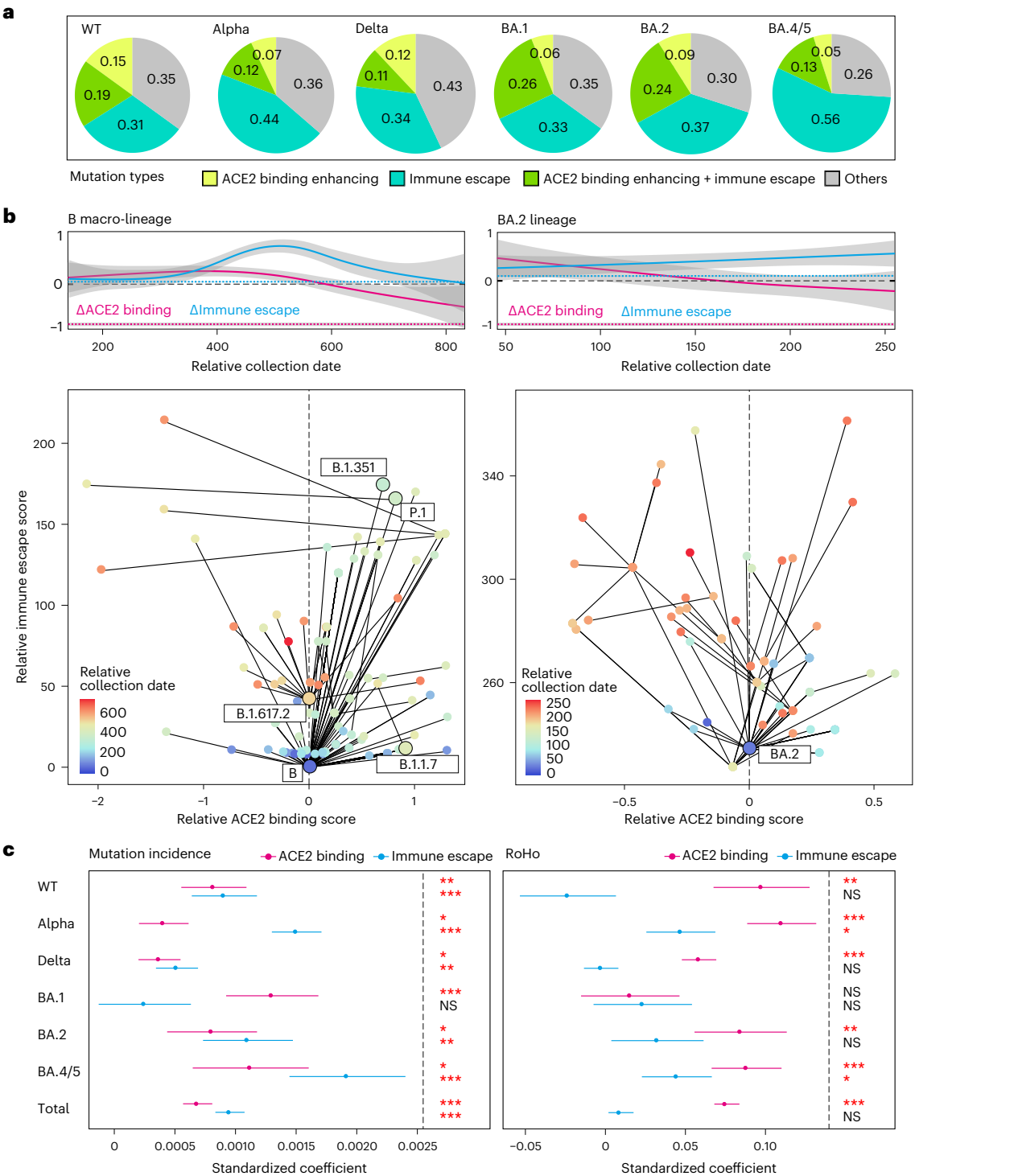


Fig. 4 | Immune evasion and ACE2 binding affinity drove the evolution of SARS-CoV-2. a, The composition of different mutation types in six major lineages. The ACE2-binding-enhancing mutations are those having a positive ACE2 binding score (the sum of the ACE2 binding score and RBD expression score). The immune escape mutations are those having an escape score that is greater than three times the average escape score of all mutations in at least one antibody epitope group. **b**, The mutation trajectories of two major variants, B (excluding the Omicron lineage) and BA.2. The relative ACE2 score and immune escape score of each variant represent the change in comparison with the ancestor variant. The regression lines depicting the relationship between time and the relative change (normalized to −1 to +1) in ACE2 binding affinity and immune evasion due to the additional mutations in the new lineage were generated using a generalized additive model and are superimposed on the top of the figures. The dashed lines in the plot indicate the expected values that were estimated by randomly selecting mutations. The grey shading represents the 99% confidence intervals. The relative collection date is the number of days after the B/BA.2 prototype was first sampled (collection date for B, 24 December 2019; for BA.2, 1 February 2022). **c**, The correlation between mutation incidence (left) or fitness change (right) and the ACE2 binding and immune escape scores of the mutation. The correlation coefficient and significance were obtained from a multivariate linear regression. The numbers of mutations used in the analysis are 147, 208, 431, 99, 100, 112 and 1,097 for the WT, Alpha, Delta, BA.1, BA.2, BA.4/5 and the total, respectively. The data are presented as mean values ± standard errors. Unadjusted *P* values are marked on the basis of their significance: $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.0001$. NS, not significant. The exact *P* values are given in Supplementary Table 2.

over time, while the ACE2 binding capacity decreased over time (Fig. 4b). When the mutations were classified into early-stage mutations (those identified in sequences collected at the first quartile of the time distribution; threshold for the B macro-lineages, 31 October 2020; threshold for the BA.2 lineage, 12 May 2022) and later-stage mutations (those that occurred later), we found that for both the B and BA.2 lineages, later-stage mutations had higher immune escape scores and a greater proportion of immune escape mutations than early-stage mutations (Extended Data Fig. 5). Furthermore, the proportion of ACE2-binding-enhancing mutations was greater in early-stage mutations, and the ACE2 binding score was higher for early-stage mutations than for later-stage mutations in the BA.2 lineage. Our results suggest that ACE2-binding-enhancing mutations are more advantageous for viral transmission in the early stage of transmission, while immune evasion is more advantageous in the later stage, probably when a large population has been infected by the variant or herd immunity has been established.

Through multiple linear regression analysis, we found that both immune evasion and ACE2 binding affinity independently correlated with the mutation incidence in all macro-lineages (Fig. 4c), with the former having a greater impact on the mutation incidence in all lineages except BA.1. Notably, immune evasion showed the highest standardized correlation coefficients in the BA.4/5 sublineages, which is consistent with the observation that the recent Omicron subvariants showed an accelerated rate of evading the antibodies induced by infections by earlier variants[14]. However, the preponderance of specific mutation types during particular pandemic stages does not necessarily mean that these mutations were advantageous to the virus's fitness. To evaluate the effect of mutations on the fitness of the new variant, we calculated the RoHo score for each pair of a new variant and its predecessor[21], which differed by one or two mutations, to represent the fitness advantage of the mutation. We found that the ACE2 binding affinity of the mutations showed a more significant correlation with the fitness advantage than immune evasion, while the contribution of immune evasion to the variant's fitness was greater in the recent BA.4/5 subvariants (Fig. 4c).

Discussion

In this study, we have provided a comprehensive evolutionary analysis of SARS-CoV-2 from the perspective of immune evasion and ACE2 binding affinity properties, which were the only two functional features available for a large number of mutations. Although other factors (such as virus particle stability, replication efficiency, incubation time and cell tropism) may also be involved in the evolution of SARS-CoV-2, the lack of relevant data prevented us from considering them in our study[22].

We found that the mutation rate of SARS-CoV-2 was similar in different macro-lineages. Estimations of the mutation rate in other studies were much higher than the segmented mutation rates estimated in our study[23,24], because the emergence of variants that had far more mutations (for example, Alpha and Omicron) than the other circulating variants at the time was not considered our estimation. As the underlying mechanism for the emergence of these new variants of concern is still mysterious[25,26], our study only focused on how the variants evolved in the population after their emergence. The distribution of mutations in the viral genome changed over time, with an increase in the proportion of mutations in the RBD region since early 2021. Notably, the rise is primarily attributable to an increase in non-synonymous mutations, suggesting that the RBD region is under growing positive selection. Given that the RBD region is where the most neutralizing antibodies and host cells bind to[27], mutations in this region could have a significant impact on the virus's ability to evade the immune system and spread to other cells; thus, it is not surprising that this region would be subject to stronger natural selection. The increased selection pressure on the RBD region over time might be caused by the rising vaccination and infection rates, which have resulted in highly concentrated humoral immune pressure.

Convergent evolution has been observed in recent Omicron subvariants, presumably caused by the concentrated humoral immunity pressures[14]. Our study indicates that convergent evolution was present even at the beginning of the pandemic, albeit to a lesser extent. The low-intensity convergent evolution may reflect mutation bias or low levels of immunity pressure. Besides the immune escape mutations (G446V, E484Q, R346I, K444N/R, T430I, S494P and E484V) and ACE2-binding-enhancing mutations (P384L), other convergent mutations whose functions are not known (S477I, S373L and A522V) repeatedly occurred in particular lineages; the underlying mechanism requires further investigation. In this study, only the RBD region showed significant convergent evolution among high-frequency mutations; however, this does not rule out the possibility that convergent mutation will take place in other regions in the future when the predominant immune pressure switches to other regions. For example, Tyr144 deletion has been frequently observed in the Omicron and early variants[28], which confers resistance to the neutralizing antibody targeting the NTD region; other mutations may arise in this region when most RBD-targeting antibodies have been escaped.

We noted that the most prevalent convergent mutations included distinct mutation types occurring at the same position, suggesting that diverse mutation types may be able to offset the same selection pressure. However, the mutation type could also be lineage-specific at some positions, suggesting that different mutation types at the same location may have distinct functional effects. For example, R346K is primarily observed in the BA.1 subvariants, while R346T/S/I is more abundant in the BA.4/5 subvariants[2]. We hypothesize that this phenomenon may reflect the shifting pressures on the virus at different phases. When Omicron first emerged in the population, it had an overwhelming growth advantage over Delta variants due to the numerous immune-evasive mutations in the Omicron RBD region. The high prevalence of R346K in BA.1 might be explained by the higher mutation rate from A to G (resulting in R346K) than the mutations from A to C/T (R346S) and mutation from G to C (R346T) in the SARS-CoV-2 genome[29]. The immune pressure against the virus became much stronger after Omicron infected a large proportion of the population; hence, immune escape mutations offered a higher transmission advantage than the ACE2-binding-enhancing mutations at the later stage. Then, mutation R346T, which offers the highest immune evasion in the BA.4/5 genetic background (Supplementary Table 1), became the most predominant mutation in the BA.4/5 subvariants. This hypothesis coincides with our finding that the immune escape mutations were more common once the variant infected a large proportion of the population, while the ACE2-binding-enhancing mutations were more prevalent when a new variant (with a significant antigenicity change) first appeared (Fig. 4).

We found that the occurrence of new mutations was significantly correlated with immune evasion as well as ACE2 binding affinity, agreeing with previous studies based on limited lineages and mutations[3], and thus confirmed function-driven virus evolution. Moreover, we have quantified the effects of the two factors on mutation occurrence and mutation fitness under different genomic backgrounds, whose effects are difficult to distinguish using incomplete data, as the same mutation occurred multiple times in different lineages, where their impact may vary. Overall, immune evasion showed a stronger correlation with mutation incidence than increased ACE2 binding affinity. Given that RNA viruses have high mutation rates, the large number of viruses in the human body could produce a population of viruses with high genetic diversity[30–32]. The viral population would then be subject to selection by antibodies induced by infections, and the variants with mutations conferring higher resistance to antibodies would have a replication advantage over other variants, making them more likely to dominate the viral population and be observed as a mutation at the individual level. Meanwhile, we found that the correlation coefficient between immune evasion and mutation occurrence increased remarkably in the recent Omicron subvariants of BA.5, which is consistent with the recent

observation that a large number of immune-evasive mutations were found in Omicron subvariants[15]. This tendency is probably attributed to the enhanced immune pressure in the population caused by mass vaccination campaigns and infections. Imprinted humoral immunity, resulting in reduced diversity of neutralizing antibodies, may also contribute to the observed trend by exerting more concentrated immunological pressure on recent variants[14,33].

Compared with immune evasion, ACE2 binding affinity exhibited a less significant correlation with mutation occurrence but a more significant correlation with viral fitness. Although previous studies have reported that ACE2-binding-enhancing mutations can promote viral transmission[1,34] (including the recent acquisition of S486P in XBB.1.5, which could significantly increase ACE2 binding affinity[35]), the importance of ACE2 binding affinity has been overlooked in previous studies of SARS-CoV-2 evolution. We hypothesize that the impact of increased ACE2 binding affinity on viral fitness is more universal than that of immune evasion, while the effect of the latter is more context-dependent. Namely, all mutations that increase the ACE2 binding effect confer a transmission advantage, as it is much easier to establish an infection[36], whereas the immune-evasive mutations provide an additional transmission advantage only in populations that have been infected by the prototype (this probably also needs to be in the recent past when the neutralizing antibody titre is high). Therefore, as the rate of reinfection rises in the Omicron era, we would expect that immune evasion will contribute more to the viral transmission advantage in the future.

Our study has several limitations. First, the effects of mutations on immune evasion and ACE2 binding affinity were estimated on the basis of the data generated in limited RBD backgrounds; thus, the interactions between mutations, as demonstrated in previous studies[12,37], were not considered when there are multiple mutations relative to the background sequences. Second, while both the humoral and cellular immune systems exert pressure on the virus, only the former was taken into account in this study, as data on the latter are not yet available. Third, the antibody compositions were estimated from a small number of samples infected by a specific variant; thus, they may not accurately reflect the humoral immune pressure on the virus due to the complex history of vaccination and infection across different populations. Additionally, the quantity of the antibodies in the human body is unknown due to technical limitations, and therefore, different antibodies were quantitatively equally weighted when estimating the immune pressure on the virus, potentially leading to biased estimates.

The unprecedented number of SARS-CoV-2 viral genomes enables us to track the trajectory of SARS-CoV-2 evolution. Meanwhile, advancements in technologies and accumulated data have greatly advanced our understanding of mutation functions. Our study showed a significant correlation between the immune evasion and ACE2 binding affinity of the mutation and mutation incidence and fitness change, which improved our understanding of the underlying forces driving the evolution of SARS-CoV-2. However, accurately determining the driving force behind a specific mutation remains a challenging task, and precisely predicting the direction of viral evolution is still elusive due to a lack of comprehensive functional data on mutations and their interactions.

Methods

Estimation of the mutation rate of SARS-CoV-2

A total of 6,484,070 high-quality open-access SARS-CoV-2 sequences and corresponding metadata were downloaded from the UShER website on 23 November 2022[19]. To calculate the mutation rate, 200 sequences from each month were randomly selected on the basis of their collection date. The number of mutations (relative to the NC_045512.2) was analysed using the optimal piecewise linear regression analysis method, a mathematical programming technique that divides the data into optimal segments and fits a linear regression function to each segment to minimize the overall absolute error[38]. The Akaike and Bayesian metrics were employed to balance predictive accuracy and model complexity to determine the optimal number of segments[39]. The mutation rate was estimated as the slope of the regression line. Meanwhile, the mutation rate was also estimated for six macro-lineages (WT, Alpha, Delta, and Omicron BA.1, BA.2 and BA.4/5) using a linear regression model. The process was repeated 100 times, and the median and standard deviation of the mutation rate were calculated. In addition, the RoHo value of the mutation (the ratio of the number of descendants in sister clades with and without a specific mutation[21]) was obtained using the matUtils tool from the UShER toolkit to represent its fitness.

Estimation of mutation incidence from the UShER phylogenetic tree

The mutation events were retrieved from the masked UShER mutation-annotated tree. First, the matUtils tool from the UShER toolkit was used to convert the protocol buffer format to the JSON format. Then, to reduce the number of false-positive events caused by the incorrect placement of the sequence in the phylogenetic tree (which included an unprecedented number of sequences), a mutation event was called only from leaf nodes (that is, real sequences) or internal nodes with at least one offspring that was a leaf node. No more than two mutations were allowed between those nodes and their parental nodes. The number of mutation events identified on the phylogenetic tree was used to represent the incidence of mutation. Only viral lineages with more than 6,000 sequences were considered for comparing the mutation incidence between different lineages to minimize the bias caused by small sample size. Notably, mutations fixed in a particular lineage did not appear on the lineage tree and were consequently not included in our analyses.

Estimation of the mutation escape score and humoral immune pressure

The antibody spectrum, neutralizing activity, antibody epitope group and mutation escape score were obtained from a previous study[14]. Briefly, 2,170 antibodies were identified from the sera of vaccinated individuals and convalescent patients of the WT, BA.1, BA.2 and BA.5 variants using single-cell V(D)J sequencing. The neutralizing activities of these antibodies against the WT, BA.1, BA.2 and BA.5 variants were determined using a pseudovirus neutralization assay. The impact of single-amino-acid mutation in the RBD region on the neutralization efficacy of antibodies was assessed using a high-throughput DMS strategy. For each mutation, an escape score was computed by fitting an epistasis model to reflect the degree of the change in antibody neutralization capacity caused by the mutation[40,41]. The raw escape score for each antibody was normalized by the maximum score among all mutations.

The antibodies were classified into 12 epitope groups according to their mutational escape profiles (the distribution of escape scores across different RBD sites for a particular antibody) based on the WT using multidimensional scaling followed by *k*-means clustering[13,14]. The antibodies in the same epitope group thus tended to be escaped by the same set of mutations. For each epitope group, mutations with an escape score (the average of the scores against all antibodies belonging to the epitope group) greater than three times the average escape score of all mutations were defined as immune escape mutations to eliminate the background noise generated during the DMS experiment (Supplementary Fig. 1). The immune pressure induced by each antibody epitope group was calculated by summing the neutralizing activity of all antibodies belonging to the group.

Calculation of the correlation between immune evasion capacity and mutation incidence

Mutations that can evade neutralizing antibodies can give the virus a transmission advantage. GSEA was used to test the significance of the correlation between mutation incidence and the immune evasion

capacity of the mutation[42]. The immune evasion capacity of the escape mutation for each antibody was calculated as the product of the escape score and the neutralizing activity of the antibody, and the values for all antibodies were summed up to represent the overall immune evasion capacity of this mutation. The weight of the escape mutation (reward) was then set to be the proportion of immune evasion capacity contributed by the mutation relative to all mutations, while the weight of non-escape mutations (penalty) was set to be the reciprocal of the number of non-immune-escape mutations, so the sum of the absolute weight of both mutation types was 1. All mutations in each lineage were reverse-sorted by incidence, and GSEA was conducted to determine whether high-weight escape mutations were enriched among high-incidence mutations. The *P* value indicating the significance of enrichment was calculated after 1,000 runs of random reordering. The escape score and the neutralizing activity for the WT variant were used for the Alpha and Delta variants due to the lack of data for these variants.

Estimation of ACE2 binding affinity

The DMS data for ACE2 binding and RBD expression were obtained from previous studies[12,43], which included measurements on the WT, Alpha, Beta, Delta, and Omicron BA.1 and BA.2 variants. Since ACE2 binding affinity and RBD expression are both critical for viral transmission, their effects were merged by summing their values (the two values are both log fold changes); a similar method was used in a previous study[14]. The data for Omicron BA.2 were used to represent Omicron BA.4/5 due to the absence of data for BA.4/5.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data files generated in this study were uploaded to the GitHub website (https://github.com/ipplol/SARS2EVO)[44].

Code availability

All custom scripts used in this study were uploaded to the GitHub website (https://github.com/ipplol/SARS2EVO)[44].

References

1. Ozono, S. et al. SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nat. Commun.* **12**, 848 (2021).
2. Liu, L. et al. Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602**, 676–681 (2022).
3. Telenti, A., Hodcroft, E. B. & Robertson, D. L. The evolution and biology of SARS-CoV-2 variants. *Cold Spring Harb. Perspect. Med.* https://doi.org/10.1101/cshperspect.a041390 (2022).
4. Markov, P. V., Katzourakis, A. & Stilianakis, N. I. Antigenic evolution will lead to new SARS-CoV-2 variants with unpredictable severity. *Nat. Rev. Microbiol.* **20**, 251–252 (2022).
5. Plante, J. A. et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116–121 (2021).
6. Liu, Y. et al. The N501Y spike substitution enhances SARS-CoV-2 infection and transmission. *Nature* **602**, 294–299 (2022).
7. Wang, P. et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* **593**, 130–135 (2021).
8. Baum, A. et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* **369**, 1014–1018 (2020).
9. Greaney, A. J. et al. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463–476.e466 (2021).
10. Greaney, A. J. et al. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* **29**, 44–57.e49 (2021).
11. Cao, Y. et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022).
12. Starr, T. N. et al. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science* **377**, 420–424 (2022).
13. Cao, Y. et al. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature* **608**, 593–602 (2022).
14. Cao, Y. et al. Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. *Nature* **614**, 521–529 (2022).
15. Focosi, D., Quiroga, R., McConnell, S., Johnson, M. C. & Casadevall, A. Convergent evolution in SARS-CoV-2 spike creates a variant soup from which new COVID-19 waves emerge. *Int. J. Mol. Sci.* **24**, 2264 (2023).
16. Ito, J. et al. Convergent evolution of SARS-CoV-2 Omicron subvariants leading to the emergence of BQ.1.1 variant. *Nat. Commun.* **14**, 2671 (2023).
17. Tuekprakhon, A. et al. Antibody escape of SARS-CoV-2 Omicron BA.4 and BA.5 from vaccine and BA.1 serum. *Cell* **185**, 2422–2433 e2413 (2022).
18. Spratt, A. N. et al. Continued complexity of mutations in Omicron sublineages. *Biomedicines* **10**, 2593 (2022).
19. Turakhia, Y. et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).
20. Dhar, M. S. et al. Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. *Science* **374**, 995–999 (2021).
21. van Dorp, L. et al. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun.* **11**, 5986 (2020).
22. Richard, M. et al. Factors determining human-to-human transmissibility of zoonotic pathogens via contact. *Curr. Opin. Virol.* **22**, 7–12 (2017).
23. Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
24. Tay, J. H., Porter, A. F., Wirth, W. & Duchene, S. The emergence of SARS-CoV-2 variants of concern is driven by acceleration of the substitution rate. *Mol. Biol. Evol.* **39**, msac013 (2022).
25. Hill, V. et al. The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. *Virus Evol.* **8**, veac080 (2022).
26. Mallapaty, S. Where did Omicron come from? Three key theories. *Nature* **602**, 26–28 (2022).
27. Robbiani, D. F. et al. Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature* **584**, 437–442 (2020).
28. McCallum, M. et al. N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332–2347 e2316 (2021).
29. Tonkin-Hill, G. et al. Patterns of within-host genetic diversity in SARS-CoV-2. *eLife* https://doi.org/10.7554/eLife.66857 (2021).
30. Deinhardt-Emmer, S. et al. Early postmortem mapping of SARS-CoV-2 RNA in patients with COVID-19 and the correlation with tissue damage. *eLife* https://doi.org/10.7554/eLife.60361 (2021).
31. Lauring, A. S. & Andino, R. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog.* **6**, e1001005 (2010).
32. Shen, Z. et al. Genomic diversity of severe acute respiratory syndrome-coronavirus 2 in patients with coronavirus disease 2019. *Clin. Infect. Dis.* **71**, 713–720 (2020).
33. Park, Y. J. et al. Imprinted antibody responses against SARS-CoV-2 Omicron sublineages. *Science* **378**, 619–627 (2022).

34. Martin, D. P. et al. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**, 5189–5200 e5187 (2021).
35. Yue, C. et al. ACE2 binding and antibody evasion in enhanced transmissibility of XBB.1.5. *Lancet Infect. Dis.* **23**, 278–280 (2023).
36. Ou, J. et al. V367F mutation in SARS-CoV-2 spike RBD emerging during the early transmission phase enhances viral infectivity through increased human ACE2 receptor binding affinity. *J. Virol.* **95**, e0061721 (2021).
37. Moulana, A. et al. Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA.1. *Nat. Commun.* **13**, 7011 (2022).
38. Yang, L., Liu, S., Tsoka, S. & Papageorgiou, L. G. Mathematical programming for piecewise linear regression analysis. *Expert Syst. Appl.* **44**, 156–167 (2016).
39. Gkioulekas, I. & Papageorgiou, L. G. Piecewise regression analysis through information criteria using mathematical programming. *Expert Syst. Appl.* **121**, 362–372 (2019).
40. Starr, T. N. et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310 e1220 (2020).
41. Otwinowski, J., McCandlish, D. M. & Plotkin, J. B. Inferring the shape of global epistasis. *Proc. Natl Acad. Sci. USA* **115**, E7550–E7558 (2018).
42. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
43. Starr, T. N. et al. Deep mutational scans for ACE2 binding, RBD expression, and antibody escape in the SARS-CoV-2 Omicron BA.1 and BA.2 receptor-binding domains. *PLoS Pathog.* **18**, e1010951 (2022).
44. Ma, W., Fu, H., Jian, F., Cao, Y. & Li, M. Immune evasion and ACE2 binding affinity contribute to SARS-CoV-2 evolution data and code. *Zenodo* https://zenodo.org/record/7954439 (2023).

Acknowledgements

We thank all the scientists around the globe for performing SARS-CoV-2 sequencing and surveillance analysis. This study was funded by the National Natural Science Foundation of China (grant no. 82161148009 to M.L.), the Strategic Priority Research Program of the Chinese Academy of Sciences (grant no. XDB38030400 to M.L.) and the Key Collaborative Research Program of the Alliance of International Science Organizations (grant no. ANSO-CR-KP-2022-09 to M.L.).

Author contributions

M.L. designed the study. W.M. and M.L. wrote the manuscript with input from all authors. W.M., H.F. and F.J. performed the bioinformatics analyses. Y.C. and F.J. generated and supervised the analysis of the DMS and neutralization data.

Competing interests

The authors declare no competing interests.

Additional information

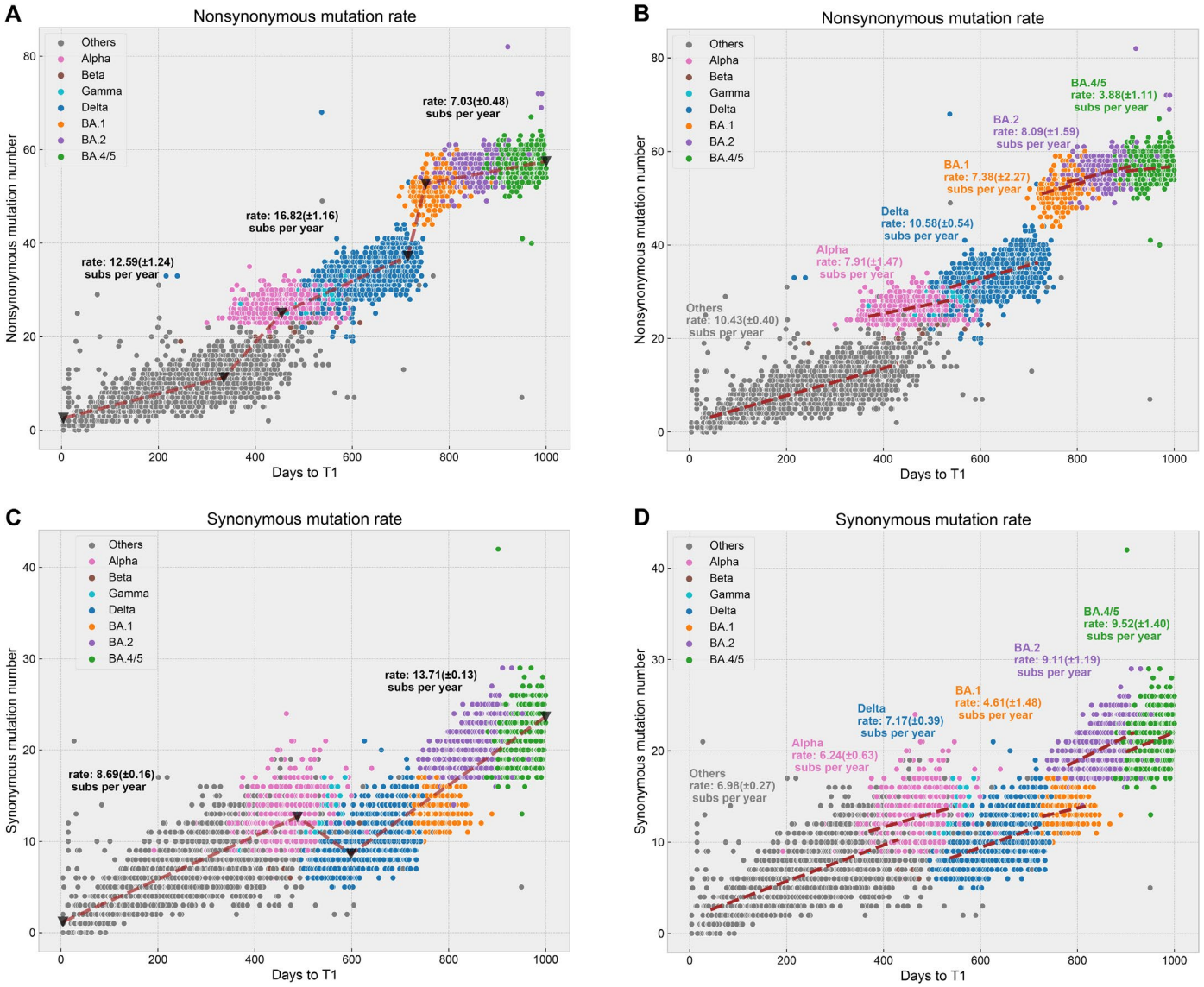
Extended data is available for this paper at https://doi.org/10.1038/s41559-023-02123-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-023-02123-8.

Correspondence and requests for materials should be addressed to Yunlong Cao or Mingkun Li.

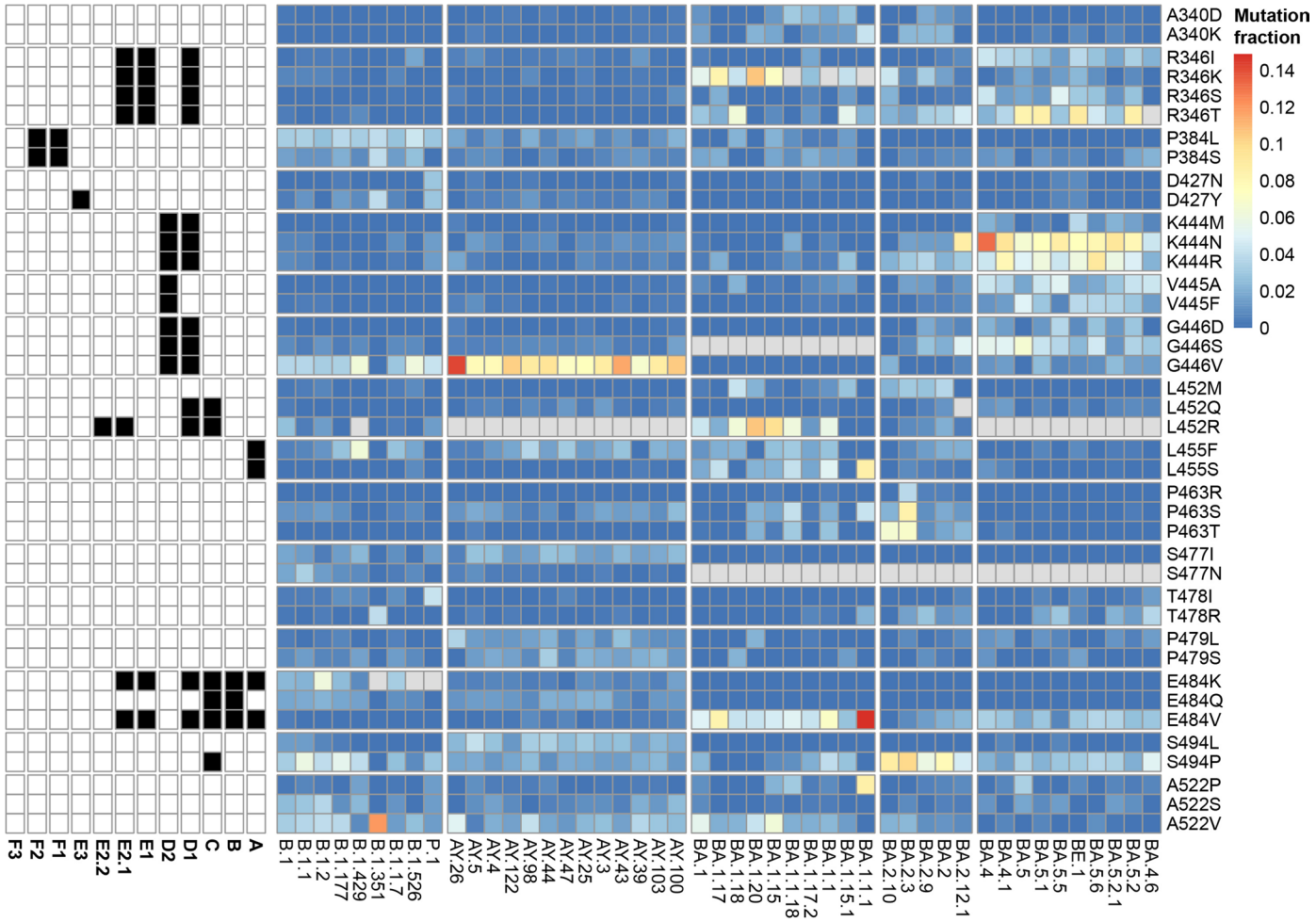
Peer review information *Nature Ecology & Evolution* thanks Oscar MacLean, Aiping Wu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41559-023-02123-8



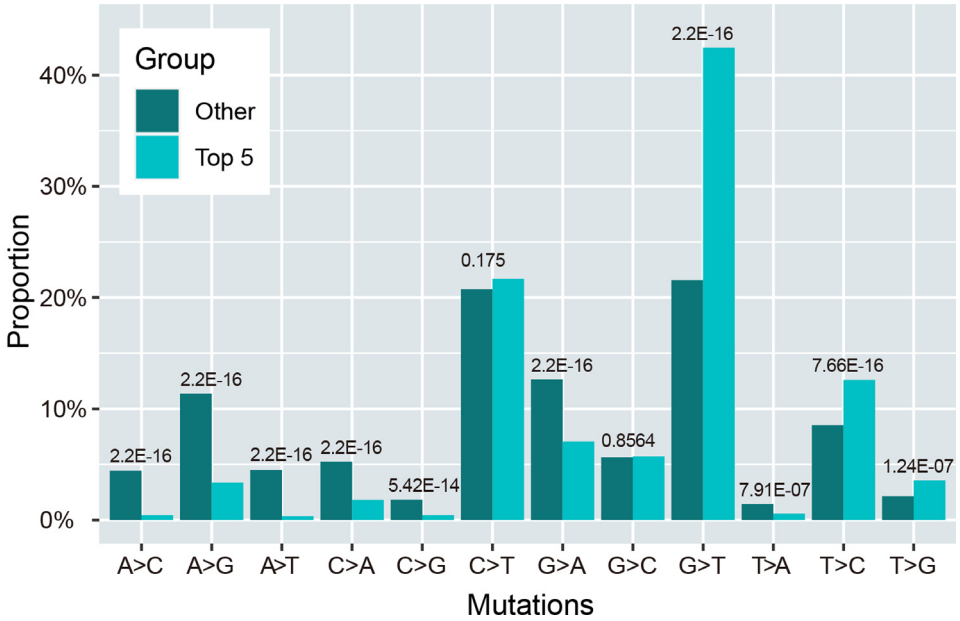
Extended Data Fig. 1 | The correlation between mutation count and collection date. a) The correlation between nonsynonymous mutation count and collection date. Two hundred sequences were randomly selected from each month based on the collection date. The segmented regression line was fitted using automatic piecewise linear regression, and the mutation rate was estimated as the slope of the regression line. We conducted 100 samplings and added the resulting median and standard deviation of the estimated mutation rate on top of the regression line. T1 represents Dec 24, 2019, which is the collection date of the first open-access SARS-CoV-2 sequence. **b**) The correlation between nonsynonymous mutation count and collection date for major lineages. **c**) The correlation between synonymous mutation count and collection date. D) The correlation between synonymous mutation count and collection date for major lineages.

https://doi.org/10.1038/s41559-023-02123-8



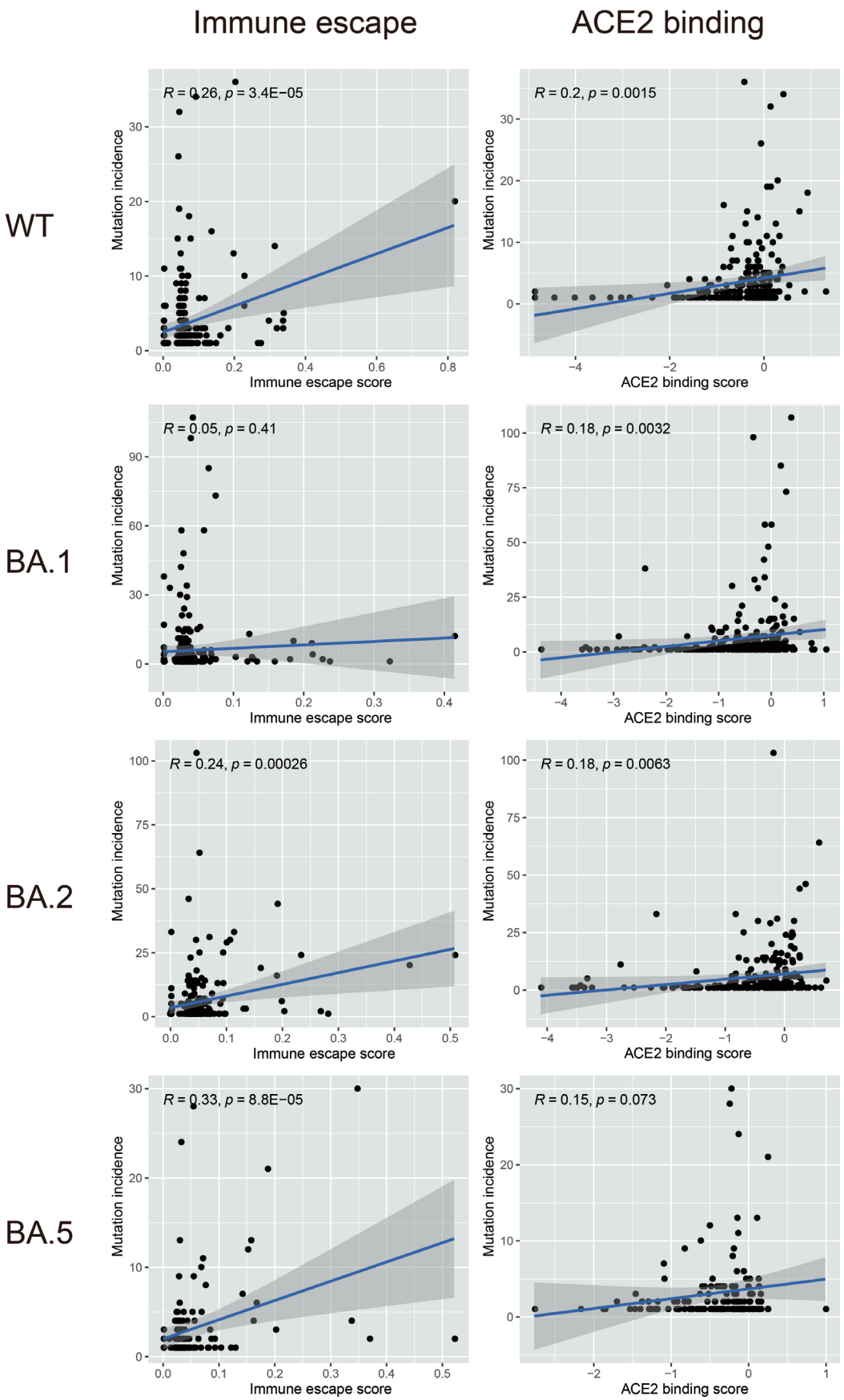
Extended Data Fig. 2 | The convergent evolution of the RBD region in the SARS-CoV-2 genome. The color denoted the ratio of the mutation's frequency to the frequency of all mutations in the particular lineage. The top five mutations occurring most frequently in each lineage are shown, while the sites with just one high-frequency mutation were excluded. The mutations that had been fixed in the lineage were labeled in grey. The antibody epitope groups that were evaded by the mutation in the right panel are labeled in the left panel.

https://doi.org/10.1038/s41559-023-02123-8



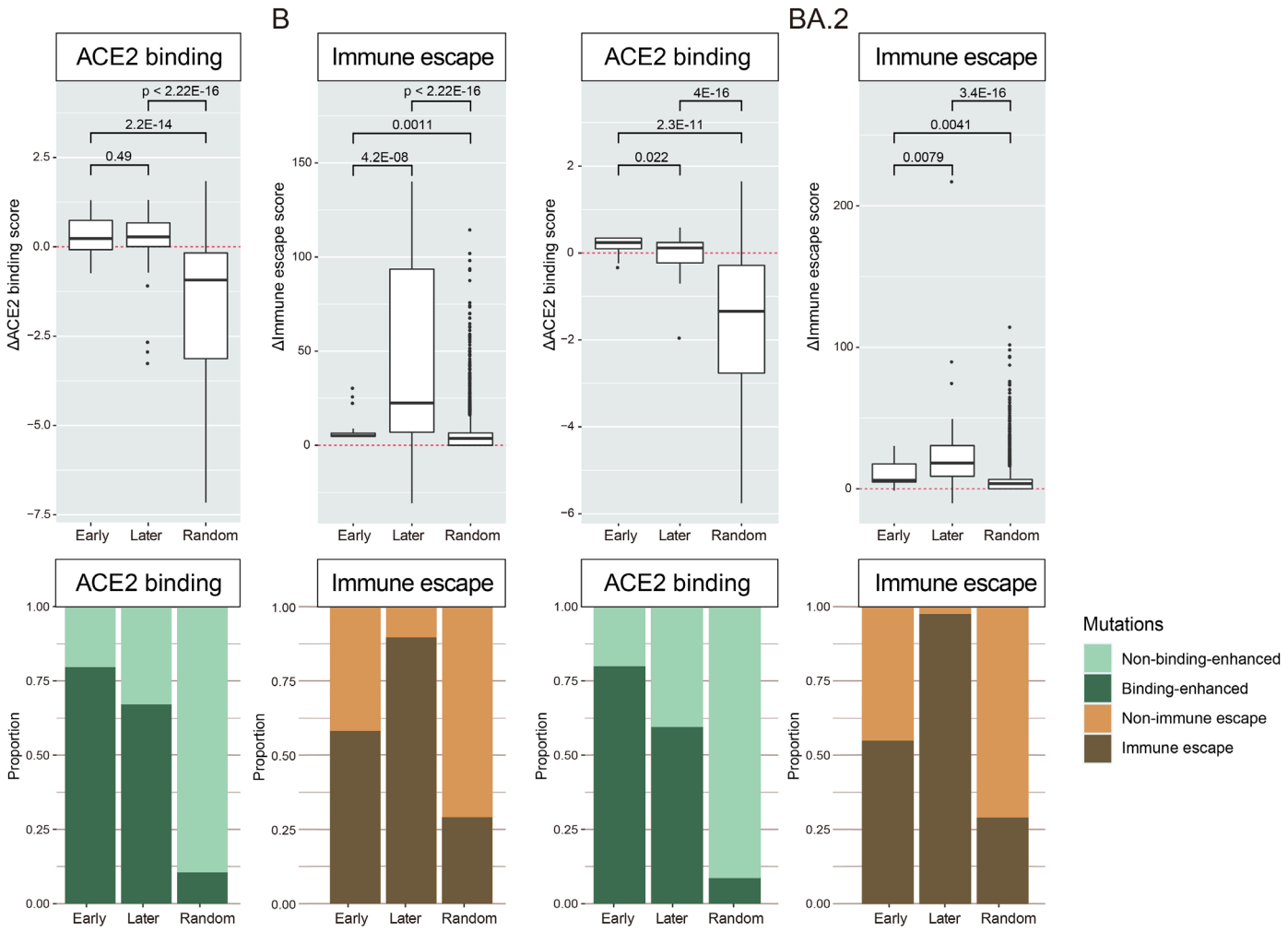
Extended Data Fig. 3 | The spectrum of non-synonymous mutations in the RBD region. The 'Top 5' denotes the top five most frequent mutations observed in at least one lineage while 'Other' denotes the rest mutations. The unadjusted p-values were obtained from the two-sided Fisher's exact test.

https://doi.org/10.1038/s41559-023-02123-8



Extended Data Fig. 4 | The correlation between the mutation incidence and immune escape score and ACE2 binding score at the individual mutation level in the SARS-CoV-2 RBD region. Each dot in the plot represents an individual mutation. The Pearson correlation coefficient and two-sided unadjusted p-value are shown in each plot. The shading represents the 99% confidence interval.

https://doi.org/10.1038/s41559-023-02123-8



Extended Data Fig. 5 | The comparison of early-stage mutations and later-stage mutations in the SARS-CoV-2 RBD region in B and BA.2 lineages. The early-stage mutations (Early, n=66 for B and n=34 for BA.2) were those identified in sequences collected at the first quartile of the time distribution of all mutations (the threshold for B macro-lineages: Oct 31, 2020; BA.2 lineage: May 12, 2022) and other mutations were defined as later-stage mutations (Later, n=66 for B and n=37 for BA.2). B lineage includes all its sub-lineages except the Omicron lineage. The upper part figures show the ACE2 binding score and immune escape score for different mutation types, while the bottom part figures show the proportion of ACE2 binding-enhancing and immune escape mutations for different mutation types. Random shows the background distribution of the metrics for different mutation types. The centre line denotes the median value, the black cross represents the mean value, the box represents the interquartile range (IQR), the whiskers extend to the furthest data point in each wing that is within 1.5 times the IQR value, and points represent outliers. The unadjusted p-value was obtained from the two-sided Wilcoxon test.

nature portfolio

Corresponding author(s): Mingkun Li

Last updated by author(s): May 21, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The sequence data and the mutation-annotated tree used in this work were downloaded from the UShER website (http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/UShER_SARS-CoV-2/). The Deep Mutation Screening (DMS) data and antibody neutralizing data, were obtained from previous studies as described in the Methods in the manuscript. No specific software was used to collect the data.
Data analysis	The data downloaded from the UShER website were processed using matUtils (v0.5.6). All data generated in this study and custom scripts used for this study have been uploaded to the Github website (https://github.com/ipplol/SARS2EVO) with DOI: 10.5281/zenodo.7954439.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We included a Data availability section and a Code availability section in the manuscript, and all files generated in the study and all custom scripts were upload to the GitHub website (https://github.com/ipplol/SARS2EVO) with an doi of https://zenodo.org/record/7954439

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable.
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	In this study, we analyzed the mutation present in all open-access SARS-CoV-2 genomes and correlated the mutation incidence and fitness change due to the mutation with its impact on immune evasion and ACE2 binding affinity.
Research sample	Our study is based on open-access SARS-CoV-2 genomes that are available from the UShER website (http://hgdownload.soe.ucsc.edu/goldenPath/wuhCor1/UShER_SARS-CoV-2/).
Sampling strategy	No sampling strategy was used except when we estimated the mutation rate of SARS-CoV-2; 200 sequences from each month were randomly selected based on their collection date to estimate the mutation rate, and this process was repeated 100 times, and the median and standard deviation of the mutation rate were calculated.
Data collection	A total number of 6,484,070 high-quality open-access SARS-CoV-2 sequences and corresponding metadata were downloaded from the UShER website on 23 November 2022.
Timing and spatial scale	All sequence included in the UShER mutation-annotated tree were used (submitted up to 23 November 2022). This data is a collection of viral sequence of the global COVID-19 pandemic.
Data exclusions	Incomplete and low-quality SARS-CoV-2 sequences were excluded from our analysis, the data downloaded from UShER has already been filtered.
Reproducibility	All results can be reproduced following our description in the Method section and using the code and data available at https://github.com/ipplol/SARS2EVO.
Randomization	Not applicable.

Blinding: Not applicable.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging