Within-host evolutionary dynamics and tissue compartmentalization during acute SARS-CoV-2 infection

Mireille Farjo[1], Katia Koelle[2], Michael A. Martin[2], Laura L. Gibson[3], Kimberly K.O. Walden[4], Gloria Rendon[4], Christopher J. Fields[4], Fadi G. Alnaji[1], Nicholas Gallagher[5], Chun Huai Luo[5], Heba H. Mostafa[5], Yukari C. Manabe[6,7], Andrew Pekosz[7], Rebecca L. Smith[8,9,10], David D. McManus[11], Christopher B. Brooke[1,8]

[1] Department of Microbiology, University of Illinois at Urbana-Champaign;
[2] Department of Biology, Emory University;
[3] Division of Infectious Diseases and Immunology, Departments of Medicine and Pediatrics, University of Massachusetts Medical School;
[4] High-Performance Biological Computing at the Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign;
[5] Division of Medical Microbiology, Department of Pathology, Johns Hopkins University School of Medicine;
[6] Division of Infectious Disease, Department of Medicine, Johns Hopkins School of Medicine;
[7] W. Harry Feinstone Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health;
[8] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign;
[9] Department of Pathobiology, University of Illinois at Urbana-Champaign;
[10] Carle Illinois College of Medicine, University of Illinois at Urbana-Champaign;
[11] Division of Cardiology, University of Massachusetts Medical School

Abstract:
The global evolution of SARS-CoV-2 depends in part upon the evolutionary dynamics within individual hosts with varying immune histories. To characterize the within-host evolution of acute SARS-CoV-2 infection, we deep sequenced saliva and nasal samples collected daily from immune and unvaccinated individuals early during infection. We show that longitudinal sampling facilitates high-confidence genetic variant detection and reveals evolutionary dynamics missed by less-frequent sampling strategies. Within-host dynamics in both naïve and immune individuals appeared largely stochastic; however, we identified clear mutational hotspots within the viral genome, consistent with selection and differing between naïve and immune individuals. In rare cases, minor genetic variants emerged to frequencies sufficient for forward transmission. Finally, we detected significant genetic compartmentalization of virus between saliva and nasal swab sample sites in many individuals. Altogether, these data provide a high-resolution profile of within-host SARS-CoV-2 evolutionary dynamics.

Introduction:
The large-scale sequencing and phylogenetic analyses of clinical samples during the SARS-CoV-2 pandemic have captured global evolutionary dynamics of the virus with unprecedented speed and resolution. However, our understanding of viral evolutionary dynamics within individual infected hosts remains limited. Most studies of SARS-CoV-2 within-host evolution have focused on chronic infections of immunocompromised individuals, as these patients are more amenable to repeated, longitudinal sampling. It has been hypothesized that chronic infections promote the emergence of novel viral variants by providing a combination of prolonged time for replication and relatively weak immune selection that promotes the emergence of variants with increased fitness to high frequency within the host (Baang et al. 2021; Corey et al. 2021; Avanzato et al. 2020). Persistent replication within

immunocompromised individuals treated with convalescent sera or therapeutic monoclonal antibodies has also been identified as a potential source of antigenically novel variants (Kemp et al. 2021; Choi et al. 2020; Truong et al. 2021).

Previous studies of SARS-CoV-2 within-host evolutionary dynamics during acute infection of immunocompetent hosts observed low within-host diversity in SARS-CoV-2 populations, with most specimens containing 15 or fewer intra-host single-nucleotide variants (iSNVs) (Valesano et al. 2021; Braun et al. 2021; Tonkin-Hill et al. 2021; Lythgoe et al. 2021). Studies of household transmission reaffirm that within-host diversity is low and that iSNVs are rarely transmitted between members of a household (Braun et al. 2021; Lythgoe et al. 2021; Valesano et al. 2021). Altogether, these data suggest that acute infections typically exhibit low overall levels of within-host genetic diversity and that the selection-driven emergence of iSNVs to high frequency during acute infection is likely rare. However, our understanding of within-host evolutionary dynamics has been hampered by the absence of high-resolution time course data within individuals.

The extent to which pre-existing immunity, elicited either through vaccination and/or prior infection, influences the within-host evolution of SARS-CoV-2 is poorly understood. For two-dose vaccinations, it remains unclear whether administration of a single dose without a follow-up may create an evolutionary sandbox where moderate immune selection in the absence of rapid clearance can drive the emergence of immune-escape variants (Saad-Roy et al. 2021; Cobey et al. 2021). A similar question has been raised by the emergence of new variants like Omicron that are able to efficiently replicate in vaccinated individuals where the virus may accumulate additional immune escape substitutions. Thus, it is important to characterize the extent of immune selection and potential for escape variant emergence during infections of immune-competent individuals at differing stages of vaccination.

To characterize viral evolutionary dynamics during acute SARS-CoV-2 infection, we sequenced longitudinal nasal swab and saliva samples collected from 32 students, faculty, and staff at the University of Illinois at Urbana-Champaign enrolled during the early stages of infection through an on-campus screening program (Ranoa et al. 2021). This cohort included 20 naive individuals and 12 individuals with presumed pre-existing immunity to SARS-CoV-2 resulting from vaccination or prior infection. By taking repeated measures of iSNV frequencies from two sample sites (mid-turbinate (MT) nasal swab and saliva) within individuals, we were able to generate high-resolution profiles of iSNV dynamics between tissue compartments and across time. Our results demonstrate that selection, genetic drift, and spatial compartmentalization all play important roles in shaping the within-host evolution of SARS-CoV-2 populations.

Results:

Sample collection

During the 2020-2021 school year, all students, faculty, and staff on the University of Illinois at Urbana-Champaign campus were required to undergo saliva-based PCR testing for SARS-CoV-2 at least twice a week (Ranoa et al. 2021). We enrolled individuals who were either (a) within 24 hours of their first positive test result, or (b) within 5 days of exposure to someone else who tested positive. Daily saliva samples and nasal swabs were collected from each enrolled participant for up to 14 days. Details on the dynamics of viral shedding in this cohort have been published previously (Smith et al. 2021; Ke, Martinez, Smith, Gibson, Mirza, et al. 2022; Ke, Martinez, Smith, Gibson, Achenbach, et al. 2022).

Optimization and validation of saliva sample sequencing protocol

The saliva-based PCR assay used in this study involves a 30-minute treatment at 95°C which partially degrades the viral RNA present in the sample and could potentially compromise sequencing quality (Ranoa et al. 2021; 2020). To address this concern and determine whether saliva Ct values are predictive of sequencing data quality, we examined sequencing depth across samples with a range of Ct values. Over a set of 10 samples that spanned a Ct range of 21.63 to 34.34, we observed a clear negative correlation between N gene Ct value and coverage depth (Adjusted R-squared = 0.4296, p = 0.02359) (**Fig 1A**). For Ct values below 28, we obtained average per-nucleotide read depths of over >10,000 reads, indicating that high quality sequence data can be obtained from heat-treated samples.

We next evaluated the relationship between Ct values and the reliability of iSNV detection in saliva samples. We generated control samples in which RNA isolated from a B.1.1.7 (Alpha) lineage sample was spiked into a B.1.2 lineage sample at defined frequencies of 50%, 10%, 5%, 2%, and 1%. We normalized both B.1.1.7 and B.1.2. samples to Ct values of 23.6, 26, or 28 based on Ct values prior to mixing. Spike-ins were then divided into replicate samples and deep-sequenced. We compared the measured frequencies of the 17 characteristic B.1.1.7 SNPs and indels between technical replicates (**Fig 1B**). We detected B.1.1.7-associated mutations at the expected frequencies for spike-ins at 5% or greater, but frequency estimates were much noisier for the 1% and 2% spike-ins. The correlation between technical replicates was also stronger at dilutions above 2% and in samples with a Ct of 23.59 or 26 than in samples with a Ct of 28. Based off these results, we set a variant calling threshold of 3% and a Ct cutoff of 28 for analyzing saliva samples.

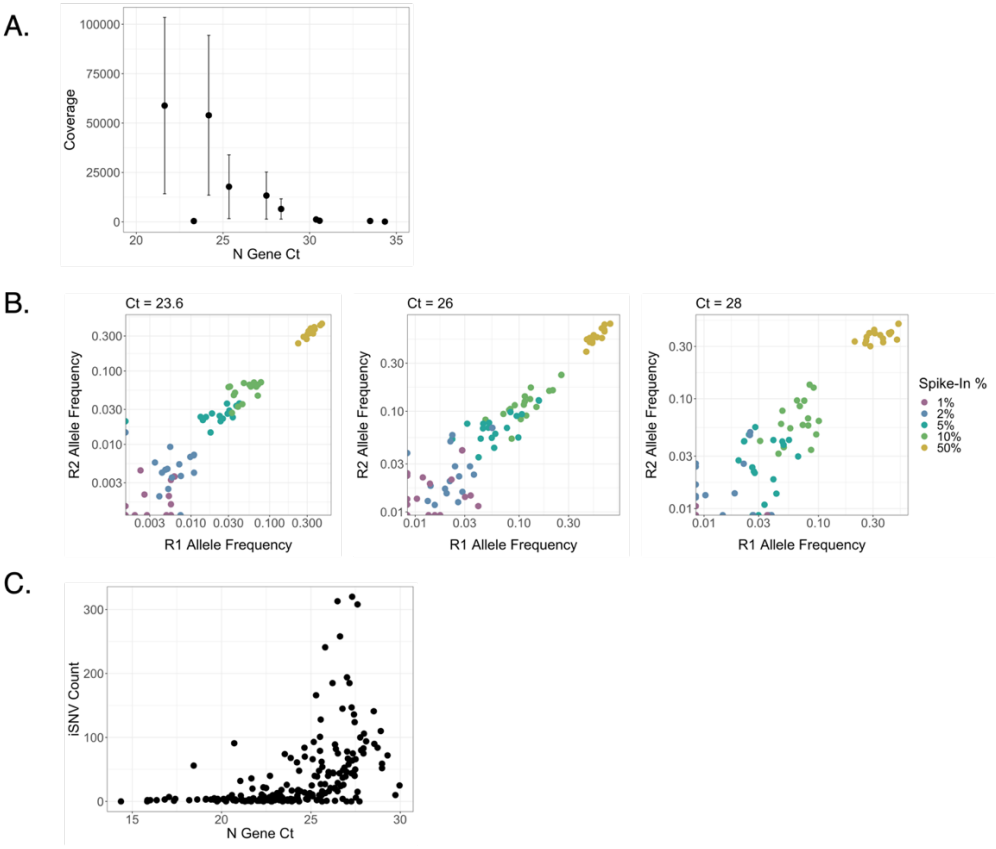


Figure 1: Relationship between saliva sample Ct values and sequence quality. (A) *Linear regression between Ct values of nucleocapsid (N) gene and mean sequence coverage depth. Error bars represent standard deviation.* ***(B)*** *Frequencies of characteristic B.1.1.7 SNPs at Ct values of 23.6, 26, and 28. B.1.1.7 RNA was spiked into B.1.2 RNA at final percentages ranging from 1% to 50% and divided between two replicates (R1 and R2).* ***(C)*** *Relationship between Ct of N and total iSNV count in resulting sequences (Spearman's rank correlation, rho = 0.6351, p < 0.001).*

We ranked study participants based on the number of saliva samples with Ct<28 and the range of time that these samples covered and selected 20 unvaccinated participants for further study. We also selected 12 study participants who were either vaccinated (fully or partially; definitions in methods) or reported a previous positive SARS-CoV-2 PCR test result and had at least one saliva sample with Ct<30. We refer throughout to this group as "immune" as we assume they mounted some sort of adaptive immune response to vaccination or infection; however, we were unable to empirically measure immune responses in this study. We chose a higher Ct threshold for these immune participants because Ct values overall were much higher in this group (Ke, Martinez, Smith, Gibson, Mirza, et al. 2022; Ke, Martinez, Smith, Gibson, Achenbach, et al. 2022). Across the entire cohort, the number of iSNVs per sample was correlated with the Ct value of the sample (Spearman's rank correlation, rho = 0.6351, $p < 0.001$), further demonstrating that high Ct values can contribute to noise in iSNV detection (**Fig 1C**).

Analysis of within-host diversity

We next examined the diversity within and between individual saliva samples, focusing on iSNVs and short insertions/deletions (indels) present at frequencies between 3-97%, with coverage depths of >1000 reads. The numbers of iSNVs that fit these criteria varied substantially between samples, generally spanning values between 1 and 100 at different points during infection (**Fig 2A,B**). To minimize false positive iSNV calls, we focused on iSNVs that appeared in at least two saliva samples collected from a given individual across different dates of infection (shared iSNVs). Numbers of shared iSNVs were similar between participants, averaging 6.31 shared iSNVs per individual, which aligns with previous assessments (Valesano et al. 2021; Braun et al. 2021; Tonkin-Hill et al. 2021).

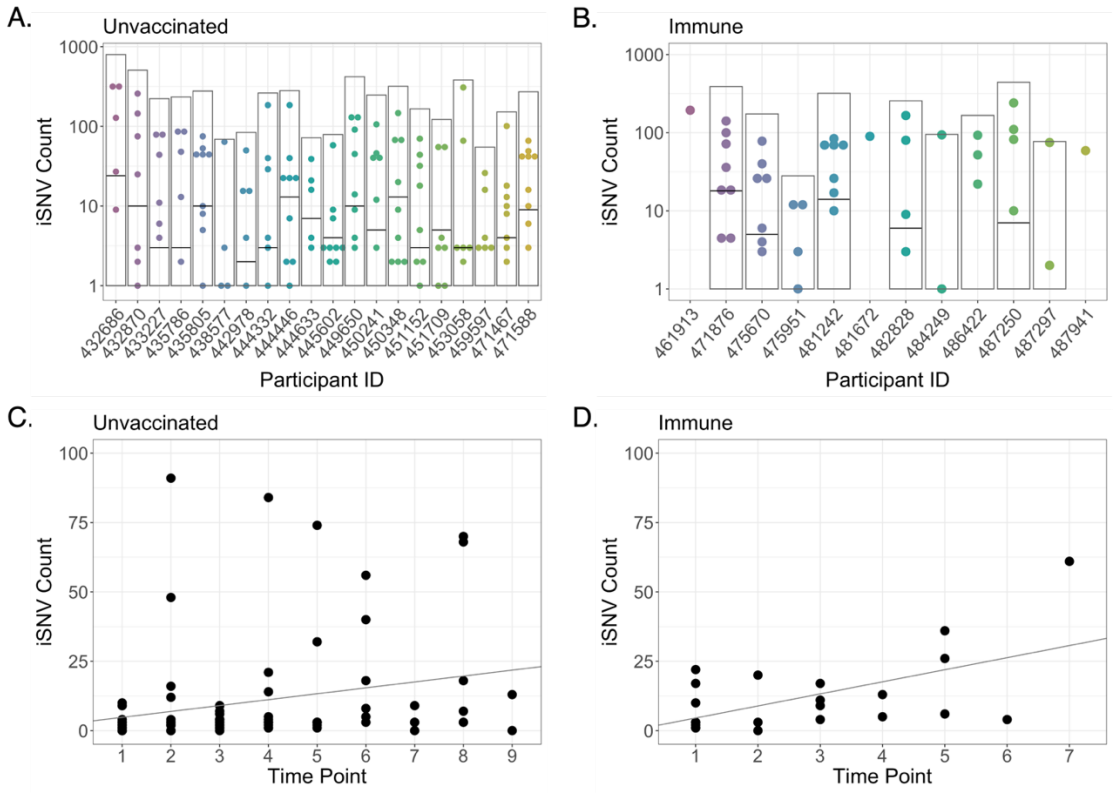


Figure 2: Intra-host single nucleotide variant (iSNV) diversity compared between samples and individuals. (A) *Total iSNV counts for each sample from each unvaccinated participant. Light grey boxes indicate total iSNV count for all samples and horizontal black lines indicate number of shared iSNVs for each participant.* ***(B)*** *iSNV counts for immune participants.* ***(C)*** *iSNV counts for individual samples with Ct < 25 from naïve participants as a function of number of days post-enrollment (Adjusted R-squared = 0.05007, p = 0.02255). Line represents linear regression.* ***(D)*** *iSNV counts for individual samples with Ct < 25 from immune participants as a function of number of days post-enrollment (Adjusted R-squared = 0.2857, p = 0.006359). Line represents linear regression.*

Naive participants had lower overall iSNV counts than immune participants, with average variant counts of 33.39 and 51.73 respectively (Welch two sample t-test, p = 0.05645) (**Fig 2A,B**). However, the higher iSNV counts in immune samples is likely due to higher Ct values, as indicated above (unvaccinated average = 23.79, immune average = 25.15, p = 0.007773). There was no significant difference in the number of shared iSNVs between the two groups (unvaccinated average = 6.65, immune average = 5.56, Welch two sample t-test, p = 0.6734). We also observed an upward trend in iSNV counts over time in both groups. To account for the impact of high Ct values on artifactual iSNV accumulation, we only considered samples with a Ct below 25 in our analysis (we did not restrict the analysis to only shared iSNVs, to avoid violating assumptions of independence between data points). We found a stronger correlation between iSNV counts and time of sample collection in immune individuals (Adjusted R-squared = 0.2857, p = 0.006359) than in naïve individuals (Adjusted R-squared = 0.05007, p = 0.02255), but the relationship was significant in both groups (**Fig 2C,D**). These data indicate similar overall levels of within-host diversity in naïve and immune individuals, but potentially a higher rate of mutant accumulation over time in immune individuals.

We next examined the distributions of shared iSNVs across the viral genome for both unvaccinated and immune populations. Since 3 immune individuals only had a single timepoint, and 4 others had no shared iSNVs, we excluded these 7 immune participants from this analysis. We also detected four frameshift mutations (at nucleotide positions 6696, 11074, 15965, and 29051) in many samples at low but consistent frequencies. Given the low likelihood of identical frameshift mutations repeatedly arising and persisting in multiple populations, we concluded that these variants are likely sequencing artifacts and we removed them from the dataset.

After the removal of these variants, we still observed several iSNVs and indels that recurred across multiple naïve individuals, including a t29760c substitution in the 3' UTR region present in 9/20 naive participants, and several coding substitutions in ORF1ab (**Fig 3A**). In immune individuals, a handful of mutations were shared by pairs of participants (**Fig 3B**). These included a P681H substitution at the S1/S2 cleavage site in the spike protein associated with the B.1.1.7 (Alpha) lineage and mutations (a G→A substitution and a G→GAACA insertion) at nucleotide position 28262, in the untranslated region between the E (Envelope) and N (Nucleocapsid) genes. From our data, we cannot determine whether shared mutations arose independently in multiple participants or were transmitted. Beyond these exceptions, the vast majority of iSNVs were only detected in a single study participant (**Fig S1**).

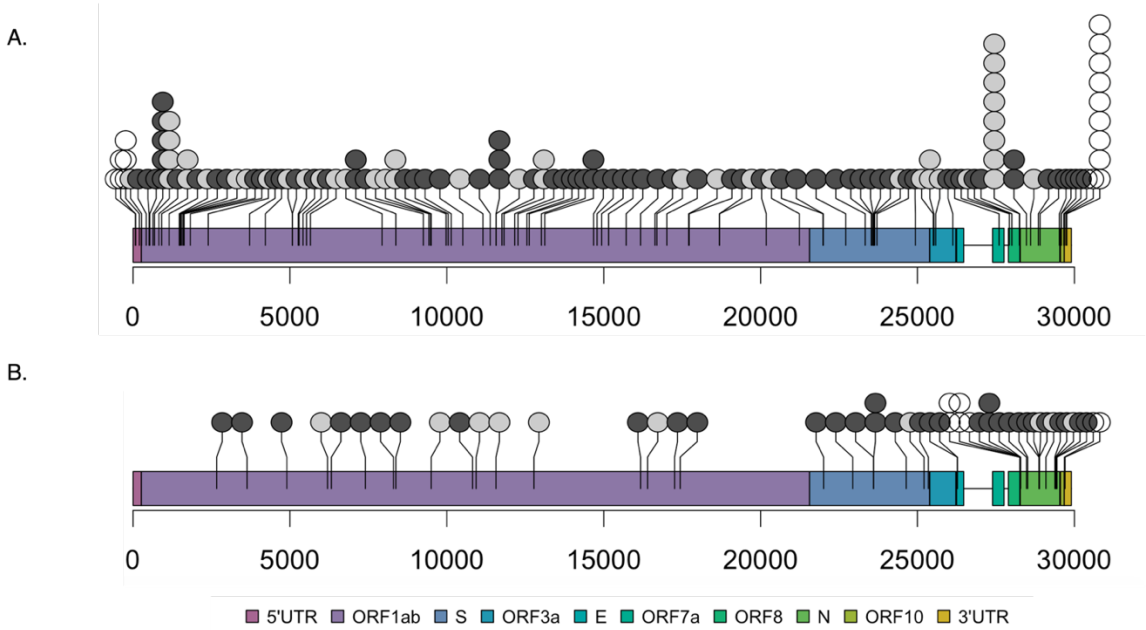


Figure 3: Locations of shared iSNVs across the SARS-CoV-2 genome. *Genome locations of shared iSNVs found in naive* ***(A)*** *and immune* ***(B)*** *participants. Number of dots at a locus indicate number of participants in which the shared iSNV was detected. Light grey dots indicate synonymous mutations, dark grey dots indicate nonsynonymous mutations, and white dots indicate UTR mutations.*

Mutations were not evenly distributed across the viral genome and clear hotspots of accumulation were easily observable. In naive participants, we observed 6 mutations between ORF1ab positions 402-457 and 4 mutations between spike positions 655-681, near the S1/S2 cleavage site (**Fig 3A**). As indicated above, S:P681H was also observed in two members of the

immune cohort. In immune individuals, we observed 3 clusters of mutations in the N gene (positions 3-80, 199-204, and 370-391; **Fig 3B**). Mutations in the N gene were enriched in immune participants – they made up 31.37% of shared variants in immune individuals but only 4.29% of shared variants in naïve participants (Fisher's exact test, $p < 0.001$). This enrichment was largely driven by samples from one vaccinated individual, who had 11 shared iSNVs in the N gene. However, when the individual was removed from the study set, immune individuals still exhibited higher proportions of N iSNVs than naïve individuals (Fisher's exact test, $p = 0.03669$). This was surprising because the N protein is not targeted by currently licensed vaccines. Despite similarities in overall levels of within-host diversity between the two groups, there appear to be differences between naive and immune individuals in the distribution of this diversity across the viral genome.

Compartmentalization between tissue environments

Previous studies revealed that SARS-CoV-2 replication dynamics can be highly discordant between saliva and nasal swab samples, suggesting strong compartmentalization of virus between different anatomical sites (Ke, Martinez, Smith, Gibson, Mirza, et al. 2022; Ke, Martinez, Smith, Gibson, Achenbach, et al. 2022). To directly evaluate the extent of compartmentalization between nasal and saliva-associated tissue sites, we compared iSNV frequencies between paired saliva and nasal swab samples over the course of infection in 13 individuals with high quality sequences for both saliva and nasal samples. We first simply compared the frequencies of shared iSNVs present at any frequency in saliva versus nasal swab samples (**Fig 4A**). In the absence of compartmentalization, we would expect iSNV frequencies to be highly correlated between sample sites. Instead, data points almost exclusively fell along the edges of the plot, consistent with substantial compartmentalization between sample sites.

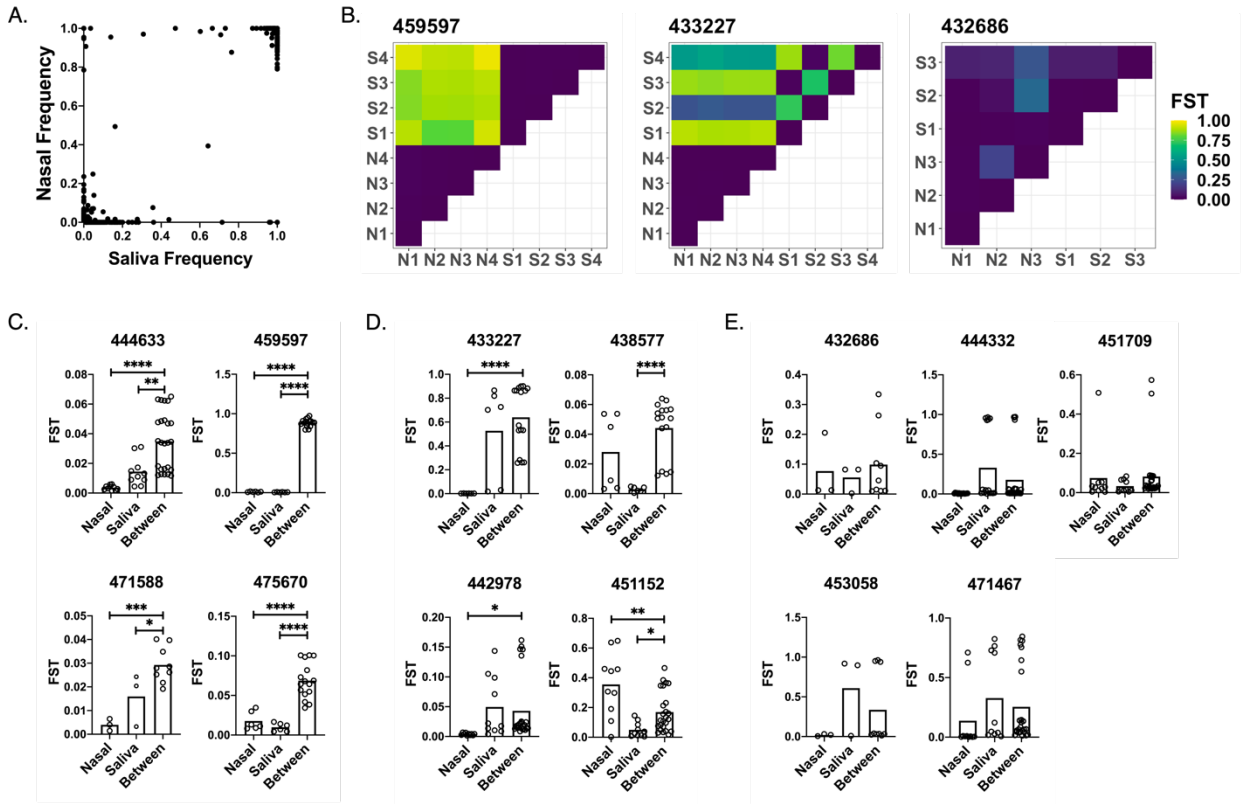


Figure 4: Quantification of genetic compartmentalization of virus between sample sites. ***(A)*** *Comparison of iSNV frequencies between matched samples in nasal and saliva environments.* ***(B)*** *Representative heatmaps exemplifying strong (459597), partial (433227), and insignificant (432686) compartmentalization. Maps show F_{ST} values between pairs of samples from nasal ("N") and/or saliva ("S") environments (numbered by order of sampling).* ***(C)*** *Participants exhibiting strong compartmentalization (within-nasal and within-saliva F_{ST} values are significantly lower than F_{ST} values from paired nasal-saliva samples).* ***(D)*** *Participants exhibiting partial compartmentalization (one set of within-environment F_{ST} values is lower than between-environment F_{ST} values).* ***(E)*** *Participants exhibiting no significant compartmentalization (neither set of within-environment F_{ST} values is lower than between-environment F_{ST} values). Asterisks indicate levels of significance (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$). P-values are derived from unpaired t-tests between each group.*

To quantify the extent of compartmentalization more precisely, we calculated fixation indices (F_{ST}) within and between environments (**Fig 4B,C,D**). The fixation index measures the ratio of allele frequency variation between sub-populations versus the variation in the total population. F_{ST} values range from 0 to 1, and values closer to 1 indicate higher levels of variation between populations. In 7 out of 13 individuals (and overall), the between-environment F_{ST} values were significantly higher than the within-nasal F_{ST} values, reflecting compartmentalization. In 6 out of 13 individuals (but not overall), the between-environment F_{ST} values were significantly higher than the within-saliva F_{ST} values (**Fig 4C,D,E**), again indicative of genetic compartmentalization between tissue compartments.

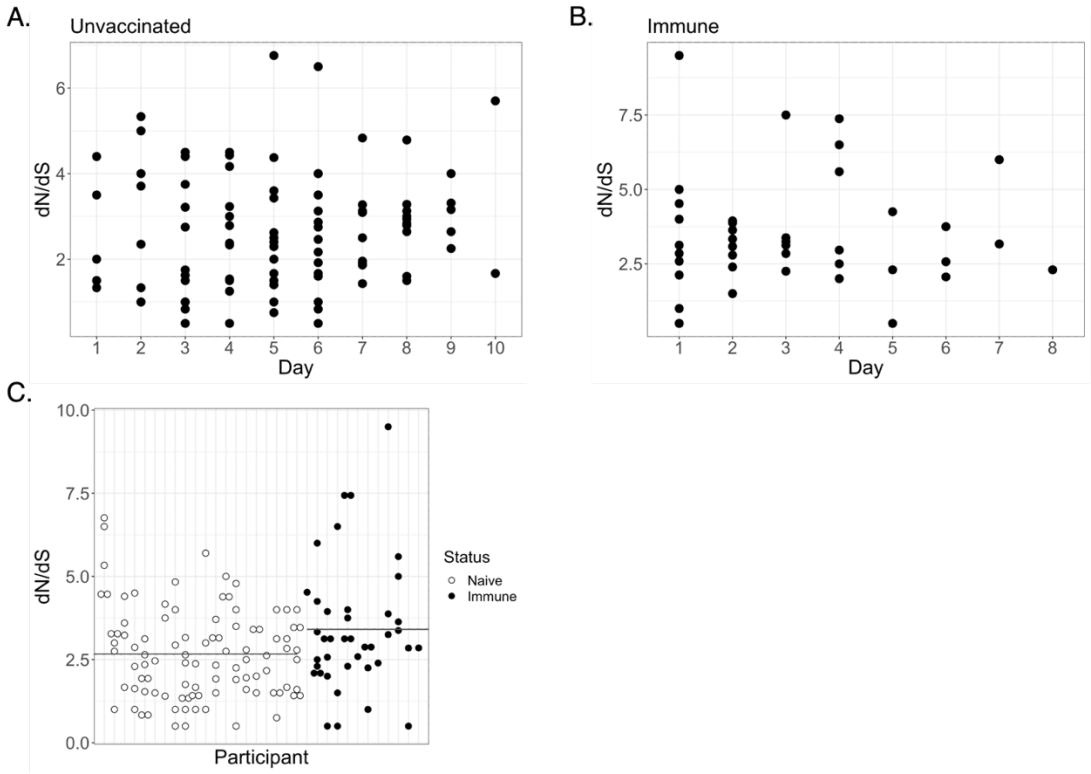
Study participants fell into three subcategories: (1) higher variation between nasal and saliva environments than within either environment, consistent with strong tissue compartmentalization (**Fig 4B,C**); (2) higher variation between environments than within one environment, consistent with partial compartmentalization (**Fig 4B,D**); and (3) no difference in between-environment versus within-environment variation, consistent with the absence of significant compartmentalization (**Fig 4B,E**). There was only one instance (participant 451152) where within-environment F_{ST} values were significantly higher than between-environment F_{ST} values (**Fig 4**). Our data suggest a significant degree of genetic compartmentalization between tissue environments present in most (8/13), but not all, participants examined.

Within-host evolutionary dynamics

Our dense longitudinal sampling allowed us to examine the evolutionary forces shaping SARS-CoV-2 populations over the course of acute infection. We first compared numbers of nonsynonymous to synonymous mutations (within a frequency range of 0.03 to 0.97, including both shared and unshared iSNVs) for each saliva sample from each participant (**Fig 5A,B**). We normalized nonsynonymous and synonymous mutation counts based on estimates of total numbers of nucleotide positions across the SARS-CoV-2 genome where a substitution would have a protein-coding effect or a silent effect, respectively. We did not detect any significant temporal trend in dN/dS ratios in either unvaccinated or immune individuals (**Fig 5A,B**). Overall, dN/dS ratios were significantly higher in immune individuals compared with naive individuals, with mean values of 3.411 and 2.664, respectively (Welch two sample t-test, $p = 0.02966$)(**Fig 5C**).

While there is not a clear relationship between dN/dS ratios obtained from individual related populations and the evolutionary forces acting on those populations (Kryazhimskiy and Plotkin

291 2008), our data clearly show an enrichment for nonsynonymous iSNVs within SARS-CoV-2
292 infected hosts that is more pronounced within immune individuals. This pattern may reflect
293 positive selection occurring within immune individuals; alternatively, it may reflect higher levels
294 of genetic drift in immune individuals (potentially due to lower overall viral loads) and the
295 resulting inability of purifying selection to act as efficiently in these individuals.
296



297
298
299 ***Figure 5: dN/dS ratios significantly differ between naïve and immune individuals. (A)***
300 *dN/dS values for all unvaccinated participants over time.* ***(B)*** *dN/dS values for all immune*
301 *participants over time.* ***(C)*** *Comparison between dN/dS values for unvaccinated and immune*
302 *participants (unvaccinated mean = 2.664, immune mean = 3.411, p = 0.02966).*
303
304 To look for signs of potential positive selection acting on specific sites in the viral genome, we
305 examined changes in the frequencies of recurring iSNVs over time. We plotted all detected
306 instances of these shared iSNVs, even if they fell outside of the frequency range of 3% to 97%
307 or fell below our chosen depth threshold of 1000 reads. Overall, the longitudinal dynamics of
308 many iSNVs in both unvaccinated and immune individuals appeared highly stochastic,
309 consistent with genetic drift and the absence of strong selection (**Figs 6,7; S2,S3**). Many iSNVs
310 detected at high frequency at one or more timepoints fell below the limit of detection (LOD) at
311 others within the same individual. In several of these cases, two or more iSNVs maintained
312 highly similar frequencies over the course of infection, suggesting linkage (e.g. ORF1ab:V3261F
313 and N:P67S in participant 451709; or UTR:t76a and UTR:t78g in participant 471588) (**Fig**
314 **6A,B**). The extreme fluctuations in frequency observed for some collection days may be
315 explained in part by variation in the quality of population sampling associated with sample
316 collection.
317

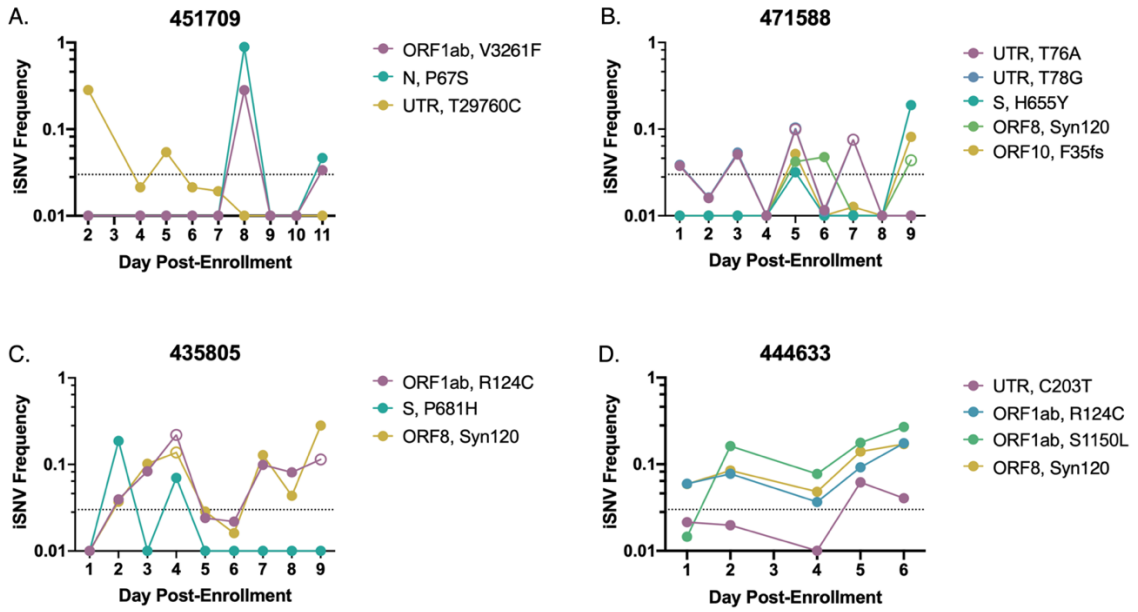


Figure 6: iSNV dynamics over time in saliva from naïve individuals. *Frequency tracking of selected iSNVs from unvaccinated participants 451709 (A), 471588 (B), 435805 (C), and 444633 (D). Dashed line marks frequency threshold of 0.03. Unfilled points mark iSNVs with read depths below the threshold of 1000 reads.*

While most recurrent iSNVs did not appear to be under strong selection in naïve individuals, we observed several examples of variants that exhibited consistent patterns of emergence or decline over the course of acute infection that could be indicative of selection. We observed two ORF1ab substitutions (R124C and S1150L) in saliva that exhibited dynamics consistent with positive selection (**Fig 6C,D**). In the case of ORF1ab:R124C, these dynamics were observed across multiple individuals (**Fig S2**). Additionally, in nasal samples, we observed a substitution at ORF1ab:P5402H that emerged to near-fixation over the course of infection (**Fig S4**). However, across the global SARS-CoV-2 tree, these substitutions are observed either sporadically or not at all, suggesting the absence of positive selection at the between-host scale (**Fig S5**).

In participant 471588, the P.1. lineage-associated substitution S:H655Y showed up just above the LOD at day 5, dropped below the LOD, then re-emerged at over 10% on day 9 post-enrollment (**Fig 6B**). Another spike substitution associated with multiple variants of concern (Saxena et al. 2022; “Preliminary Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in the UK Defined by a Novel Set of Spike Mutations - SARS-CoV-2 Coronavirus / NCoV-2019 Genomic Epidemiology” 2020), S:P681H, was observed to emerge in participant 435805 on days 2 and 4, before ultimately falling back below the LOD. The within-host transience of a mutation associated with increased fitness at the global scale highlights how within-host evolutionary trends can diverge from global trends. Further illustrating the strength of genetic drift in these populations, we observed several cases in which synonymous mutations (which we assume to be neutral) rose to higher frequencies over the course of infection: *e.g.* the ORF1ab:1717 substitution in participant 435786 and the ORF1ab:1668 substitution in participant 451152 (**Fig S2**).

Overall, these data suggest that in a subset of acute infections, positive selection may be driving the emergence of specific iSNVs, but not generally to high enough frequencies to reliably transmit given the narrow transmission bottlenecks observed across multiple studies (Valesano et al. 2021; Braun et al. 2021; Martin and Koelle 2021).

We did not observe any sweeps of antigenically significant spike substitutions in our small cohort of immune participants, suggesting that spike-based vaccination does not impose strong immune selection on the viral populations sampled over the course of acute infection (**Fig 7, S3**). The only recurrent, antigenically significant spike iSNV that we observed within our immune cohort resulted in a K458N substitution in the receptor binding domain (RBD), a residue that has been previously associated with monoclonal antibody (mAb) escape (**Fig 7A**) (Harvey et al. 2021; Z. Liu et al. 2021). This iSNV remained steady at a low frequency between 3% and 10% over the course of infection however, suggesting the absence of a strong selective advantage and low probability of forward transmission.

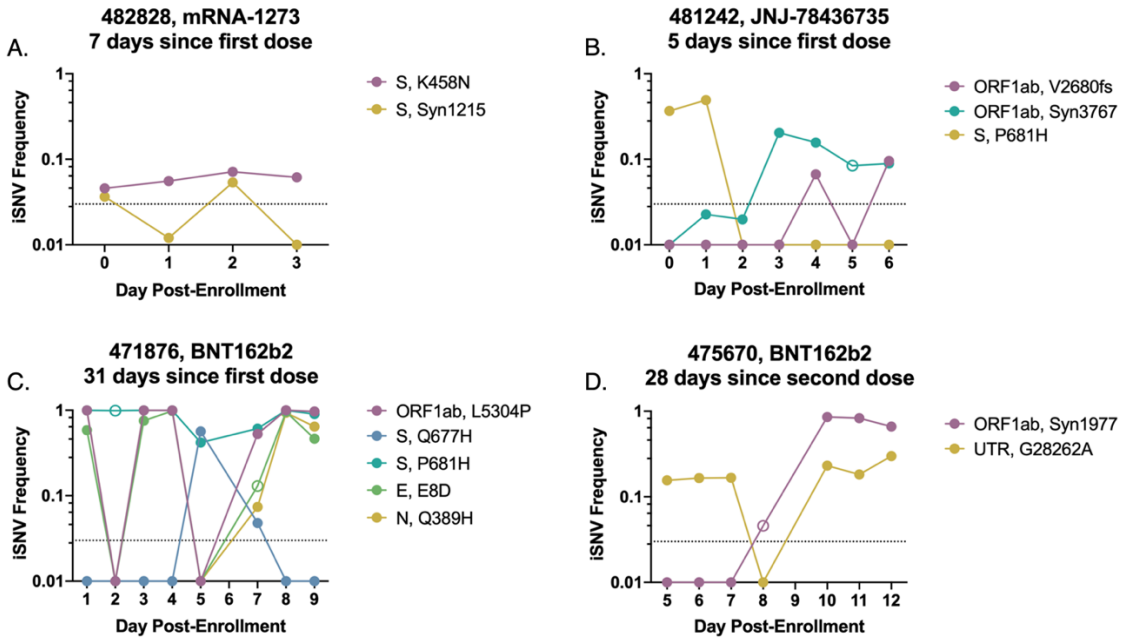


Figure 7: iSNV dynamics over time in saliva from vaccinated individuals. *Frequency tracking of selected iSNVs from immune participants 482828 (newly vaccinated) (A), 481242 (newly vaccinated) (B), 471876 (partially vaccinated) (C), and 475670 (fully vaccinated) (D). Dashed line marks frequency threshold of 0.03. Unfilled points mark iSNVs with read depths below the threshold of 1000 reads. Panel headings indicate vaccine received and time between enrollment and last vaccine dose.*

Outside of the established antigenic sites, we observed interesting dynamics near the S1/S2 cleavage site in immune individuals. In participant 481242, S:P681H was observed at middle frequencies on days 1 and 2 post enrollment but dropped below the LOD by day 3 and remained undetectable in later timepoints, suggesting a sweep by an S:P681 revertant (**Fig 7B**). In participant 471876, S:P681H was at or near fixation over the first four days of sampling, dropped in frequency on days 5 through 7, and then returned to near-fixation at day 8 (**Fig 7C**). On the two days where S:P681H dropped below 90%, a nearby spike substitution, Q677H,

381 emerged to high frequency before dropping back below the LOD on days 8 and 9. The co-
382 occurrence of the dip in S:P681H frequency with the emergence of S:Q677H and subsequent
383 reversal are consistent with competition between these two substitutions. Both substitutions
384 have proliferated at the global scale, suggesting that in some cases within-host and global
385 dynamics may be aligned (Hodcroft et al. 2021; Colson et al. 2022; Ghosh et al. 2021; Saxena
386 et al. 2022). Critically, S:Q677H peaked at a time (day 5 post-enrollment) when this participant
387 was still viral culture positive (see figure 1 in (Ke, Martinez, Smith, Gibson, Achenbach, et al.
388 2022)), indicating the potential for this *de novo* variant to be successfully transmitted.
389
390 Overall, we did not detect obvious signs of antibody-mediated immune selection within immune
391 individuals and found that iSNV frequencies often appeared to vacillate stochastically, like what
392 was observed in naïve individuals. In participant 471876, several iSNVs (ORF1ab:L5304P,
393 E:E8D, N:Q389H) fluctuated between fixation and frequencies below the LOD over the course
394 of infection (**Fig 7C**). Additionally, in participant 475670, we observed a synonymous mutation
395 (ORF1ab:Syn1977) rapidly rise to fixation (**Fig 7D**). However, it is hard to ascribe all observed
396 dynamics entirely to genetic drift. Our observations of wild-type reversion and competition
397 between iSNVs at the S1/S2 cleavage site (**Fig 7B,C**) suggest that selection may drive the
398 within-host fluctuations of iSNVs at non-antigenic sites during acute infection of some immune
399 individuals.
400
401 Finally, in keeping with our compartmentalization analysis, we found that frequencies of shared
402 iSNVs in nasal swab samples over the course of infection often varied from the dynamics
403 observed in saliva samples (**Fig S4**). Some of these differences arose from the lack of detection
404 of a mutation in a certain environment, and some resulted when an iSNV that fluctuated in one
405 environment was fixed in the other—for example, S:Y145del in participant 450241 was fixed in
406 all nasal samples and was therefore only called as an iSNV in saliva samples (**Fig S2,4**).
407 Following the trend observed in saliva samples, iSNV dynamics in nasal swab samples were
408 generally stochastic, with only rare instances of dynamics consistent with selection.
409
410 **Discussion:**
411 By analyzing longitudinal samples collected daily over the course of acute infection, we
412 captured a high-resolution temporal profile of SARS-CoV-2 within-host dynamics in humans. In
413 general, we observed little evidence of strong selection acting on within-host viral populations in
414 our cohort, consistent with previous reports (Valesano et al. 2021; Braun et al. 2021; Tonkin-Hill
415 et al. 2021). This was true even within our group of vaccinated or previously infected individuals,
416 mirroring a previous study of influenza virus that failed to detect the emergence of antigenic
417 variants during infection of immune individuals (Debbink et al. 2017). These data suggest that
418 respiratory viruses like SARS-CoV-2 and influenza virus may be able to replicate at some
419 mucosal sites with minimal restriction by neutralizing antibodies, even within individuals with
420 robust systemic antibody responses.
421
422 While signs of strong positive selection were rare in this cohort, we did identify a handful of
423 nonsynonymous substitutions (S:Q677H, N:P67S, ORF1ab:P5402H) that emerged from below
424 the limit of detection to high frequency over the course of infection. Importantly, S:Q677H
425 emerged to 56.5% frequency on a day when the associated study participant had detectable
426 infectious virus in a nasal swab (Ke, Martinez, Smith, Gibson, Achenbach, et al. 2022),
427 suggesting the potential for this iSNV to be transmitted forward. Substitutions at S:Q677
428 (including Q677H) have independently emerged in multiple viral sub-lineages around the world,
429 supporting that mutations at this site can be advantageous. We also observed signs of

430 competition between S:Q677H and S:P681H within the same individual, with S:Q677H briefly
431 emerging to a high frequency on a day at which the initially fixed S:P681H dipped in frequency.
432 However, the observed reversion to a S:P681H-only genotype after day 7 suggests that the
433 selective advantage conferred by S:P681H is greater than that of S:Q677H. This fitness
434 advantage is supported globally by the more widespread proliferation of S:P681H-containing
435 lineages in comparison to S:Q677H (Hadfield et al. 2018). Our data demonstrate how, in rare
436 cases, *de novo* generated variants can emerge to within-host frequencies sufficient for
437 transmission during acute infection.
438
439 In addition to the limited number of iSNVs that emerged to high frequency within single
440 individuals, we observed several iSNVs that arose above background in multiple participants.
441 Mapping the genomic locations of these shared mutations revealed several hotspots of non-
442 synonymous mutation accumulation that differed between naïve and immune individuals. In
443 naïve individuals, we identified hotspots at residues 402-457 in ORF1ab, and 655-681 in spike.
444 The latter is especially interesting, as this region is directly adjacent to the S1/S2 cleavage site.
445 Substitutions that modulate cleavage efficiency are important for transmission in ferrets
446 (Peacock et al. 2021) and replication in cell culture (Johnson, Xie, et al. 2021). S1/S2 cleavage
447 site substitutions are characteristic features of the Omicron, Delta, and Alpha lineages, and
448 have been shown to be responsible for Delta's increased relative fitness compared with Alpha
449 (Y. Liu et al. 2021), suggesting the importance of this domain in the adaptation of SARS-CoV-2
450 to humans. The enrichment of amino acid substitutions immediately upstream of the S1/S2
451 cleavage site in our cohort is further evidence that this region may be subject to stronger within-
452 host selection in humans.
453
454 We also observed a surprisingly high density of N gene substitutions in immune participants. An
455 observed hotspot of mutation accumulation in N:199-204 matches up with previous
456 observations of frequent changes at positions 201-205 in the serine-rich region of the gene,
457 several of which are characteristic of emerging lineages. R203K and G204R substitutions (both
458 of which we observed in our samples) can increase the relative fitness of the virus, potentially
459 through increased phosphorylation of the nucleocapsid (Johnson, Zhou, et al. 2021). These
460 substitutions have also been associated with the transcription of an alternate subgenomic
461 mRNA with anti-interferon activity (Mears et al. 2022). While spike protein substitutions are
462 clearly the primary drivers of SARS-CoV-2 adaptation to humans, our results are also consistent
463 with previous data suggesting an important role for the N gene during human adaptation.
464
465 Finally, several shared mutations occurred within untranslated regions of the viral genome. The
466 most frequent of these was a t29760c substitution (in the 3' UTR), which reoccurred across 9
467 different naïve individuals. We also observed recurring substitutions in the 5' UTR, and, in
468 immune individuals, a recurring substitution and insertion in the untranslated region preceding
469 the N gene. The untranslated regions of the coronavirus genome form secondary structures that
470 play a role in viral replication and translation (Yang and Leibowitz 2015). It remains to be seen
471 whether the recurring UTR mutations that we observe have appreciable effects on viral fitness.
472
473 Variant dynamics in multiple participants exhibited extreme fluctuations where iSNVs at or near
474 fixation abruptly fall below the limit of detection, only to return to high frequencies days later.
475 Given the abruptness of these fluctuations, it is doubtful that they were selection-driven. They
476 could potentially be explained if there is a significant degree of spatial structuring of within-host
477 viral genetic diversity, as has recently been described for influenza virus (Amato et al. 2021).
478 Spatial structuring could promote more extreme, drift-driven fluctuations in sampled iSNV

frequencies, due to bottleneck effects (Amato et al. 2021; Pfeiffer and Kirkegaard 2006; Orton et al. 2020). Alternatively, these fluctuations might be artifactual, potentially arising from poor quality sampling of the viral population. We think the latter is unlikely, due to the Ct value thresholds we used for including samples in our analyses, but we cannot formally rule this possibility out. Regardless, either explanation further emphasizes the advantages of longitudinal sampling, as single-timepoint snapshots of viral populations can present misleading views of the within-host landscape.

Supporting the possibility that stochastic within-host SNV dynamics may partially result from spatial structuring, we observed significant compartmentalization between the oral and nasal environments over the course of SARS-CoV-2 infection in a subset of individuals. iSNVs varied in frequency between the two environments, a finding that builds on previous observations that peaks in viral shedding are often offset by several days between saliva and nasal environments (Ke, Martinez, Smith, Gibson, Mirza, et al. 2022), and that shedding is sometimes limited to the saliva compartment in immune individuals (Ke, Martinez, Smith, Gibson, Achenbach, et al. 2022). These results suggest the potential for tissue-specific adaptation by the virus and reaffirm that sampling of a single tissue site may not provide a complete view of viral population diversity within a host.

A clear advantage of repeated longitudinal sampling is that it allows for higher confidence variant calling compared with single-timepoint sampling. Across individual samples, we measured iSNV counts ranging from zero to several hundred and found that these values could shift rapidly within an individual over short periods of time. However, the number of variants shared across multiple days was consistent across both cohorts and remained relatively low, with both groups exhibiting shared iSNV counts that align with previous assessments of within-host diversity (Valesano et al. 2021; Lythgoe et al. 2021).

Altogether, our results suggest that viral evolution is largely driven by stochastic forces during acute infections but that in rare occasions selection can drive the emergence of iSNVs capable of forward transmission. Furthermore, our recurrent detection of iSNVs that have been successful (or not) at the global scale indicate areas of alignment and discordance between within-host and between-host selective pressures and thus help shed light on the forces that shape global patterns of SARS-CoV-2 evolution.

Acknowledgments
This work has been generously supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health award 3U54HL143541-02S2) through the RADx-Tech program to D.D.M., L.L.G. and C.B.B., as well as additional funds generously provided by the Carl R. Woese Institute for Genomic Biology and the Department of Microbiology at the University of Illinois at Urbana-Champaign.

We are grateful to Alvaro Hernandez, Chris Wright, and the DNA services team at the Roy J. Carver Biotechnology Center for assistance in next generation sequencing.

Declaration of interests
The authors declare no competing interests.

Methods
Sample collection

528 To monitor on-campus COVID cases, students and employees at the University of Illinois at
529 Urbana-Champaign were required to submit biweekly saliva samples for RTqPCR testing.
530 Individuals who tested positive were given the option to enroll in a longitudinal sample collection
531 study within 1 day of receiving a positive result. Additionally, individuals who been in close
532 contact with a positive case were eligible to enroll in the same study within 5 days of their
533 exposure. Enrolled participants then provided saliva samples and mid-turbinate nasal swabs for
534 14 days after the date of their first positive test (This collection protocol is described in detail in
535 (Ke, Martinez, Smith, Gibson, Mirza, et al. 2022)) Within 12 hours of collection, RTqPCR was
536 performed on heat-inactivated saliva samples to assess viral load, as described in Ranoa et al.
537 (Ranoa et al. 2020). Nasal swab samples were stored in viral transport media at -80° C and
538 shipped to Johns Hopkins University for analysis.
539
540 Participants were designated as fully vaccinated if they had been infected at least 14 days after
541 receiving a single-dose vaccine (JNJ-78436735) or a second dose of a two-dose vaccine
542 (BNT162b2 or mRNA-1273). If at least 14 days had passed since receiving the first dose of a
543 two-dose vaccine, participants were designated as partially vaccinated, and if less than 14 days
544 had passed since receiving a dose of any vaccine, participants were designated as newly
545 vaccinated. Study enrollment was concluded prior to the approval of vaccine boosters.
546
547 *Participant selection*
548 After RTqPCR analysis, unvaccinated participants with fewer than three saliva samples under
549 the cycle threshold (Ct) cutoff value of 28 were filtered from the dataset. The remaining
550 participants were sorted and ranked based on their number of quality samples and the range of
551 dates covered by these samples. The top 20 participants were selected for further analysis. All
552 immune participants with saliva samples under a Ct value of 30 were retained, which resulted in
553 a study group of 12 individuals. From this combined cohort of naïve and immune individuals,
554 nasal samples from 14 individuals were chosen to evaluate environmental differences between
555 the oral and nasal cavities.
556
557 *RNA extraction and sequencing (Saliva samples)*
558 To extract viral RNA, a volume of 140 uL from each heat-inactivated saliva sample was
559 processed using the QIAamp viral RNA mini kit. Viral cDNA was generated from 100 ng of the
560 resulting RNA aliquots and sequencing libraries were prepared from the cDNA using the Swift
561 SNAP Amplicon SARS-CoV-2 kit. Deep sequencing was then performed on an Illumina
562 NovaSeq. Raw sequences were processed using the nf-core/viral-recon workflow, in order to
563 align sequences to the Wuhan-Hu-1 reference genome and extract frequencies and annotations
564 for variants at frequencies higher than 0.01. Lineages were assigned using *Pango* version
565 1.2.34.
566
567 *Analysis of variant dynamics (Saliva samples)*
568 Variants were extracted from sequences aligned to Wuhan-Hu-1 using *iVar*, and variant effects
569 were annotated with *SnpEff*. To focus on minor sequence variants, variants present above a
570 frequency of 0.97 were left out of the dataset. Variants at frequencies lower than 0.03 were also
571 removed to decrease the potential for error due to noise. A per-nucleotide depth threshold of
572 1000 reads was also applied to the dataset. For each participant, variants present across two or
573 more days of infection were extracted, and their frequencies tracked. Though the depth and
574 frequency cutoffs described above were used to identify these shared variants, frequency
575 tracking was performed on a dataset curated without thresholds, to avoid cases in which
576 variants crossing either threshold may erroneously appear to fall out of the dataset. Variants

577 with per-nucleotide coverage values below the cutoff were specially marked and plotted to
578 indicate their low depth values. *SnpEff* annotations were used to characterize shared variants
579 as synonymous, nonsynonymous, or untranslated, and to assign them to the appropriate region
580 of the SARS-CoV-2 genome. Genome positions of variants were visualized using *trackViewer*
581 package (version 1.28.1) in R (Ou and Zhu 2019). The ratio of nonsynonymous to synonymous
582 variants (dN/dS) was then calculated for each sample, using all variants above the depth
583 threshold of 1000. Variant counts were normalized to an estimate of the number of
584 nonsynonymous sites (9803) or synonymous sites (19606) in the genome, which were
585 calculated by estimating that each codon contains one synonymous site. Infinite and NaN
586 values were excluded from further analysis.
587
588 *RNA extraction and sequencing (Nasal swab samples)*
589 Mid-turbinate nasal swab specimen aliquots were maintained at -80°C prior to use. RNA was
590 extracted from 300 µl of clinical specimen using the Chemagic™ 360 system (Perkin Elmer)
591 according to the manufacturer's specifications. RNA was eluted with 60µl elution buffer and
592 stored at –80°C until use. cDNA synthesis was performed using Superscript IV reverse
593 transcriptase Kit (ThermoFisher Scientific) following the manufacturer's protocol. The
594 amplification of the genome was performed using Q5 Hot Start DNA Polymerase
595 (ThermoFisher) and two pools of primers, each containing unique non-overlapping binding sites
596 covering half of the SARS-CoV-2 genome. Library preparation was performed following the
597 protocol provided with the Illumina Nextera DNA Flex kit for sample inputs of 100-500ng.
598 Briefly, adapter sequences were ligated to genomic DNA fragments via Tagmentation, after
599 which mean amplicon size was determined via Agilent TapeStation 4500 and concentration was
600 determined using a Qubit Flex Fluorometer (Invitrogen). Size normalization was performed and
601 samples were diluted to a loading concentration of 1.2-1.3pM. Samples were sequenced using
602 Illumina MiniSeq High Output Reagents (150 cycles).
603
604 *Analysis of genetic compartmentalization between sample sites*
605 To account for sequencing differences between saliva samples and nasal swab samples (which
606 were processed at different facilities), we imposed varying quality control thresholds on each
607 dataset. To remove potentially artefactual iSNVs, we imposed a per-nucleotide depth cutoff of
608 500 reads in saliva samples and 200 reads in nasal swab samples. To remove samples with low
609 overall coverage, we imposed a mean coverage cutoff of 1000 on saliva samples and a median
610 coverage cutoff of 200 on nasal samples. We assigned SNPs present below a threshold of 1% a
611 frequency of 0, and those present above a threshold of 99% a frequency of 1. Because nasal
612 swab samples were sequenced using ARTIC primers, we also filtered out common SNP
613 artifacts that arise frequently with this primer set ("Issues with SARS-CoV-2 Sequencing Data -
614 SARS-CoV-2 Coronavirus / NCoV-2019 Genomic Epidemiology" 2020). In one individual
615 (451709), we also excluded samples from days 4 and 7, due to evidence of possible cross-
616 contamination. For each participant, pairwise F_{ST} values were calculated for all possible pairs of
617 samples, including pairs of saliva samples, pairs of nasal samples, and pairs of one saliva
618 sample and one nasal sample.
619
620 *Phylogenetic analysis*
621 The metadata file for all sequences present in the GISAID EpiCov database (10.2807/1560-
622 7917.ES.2017.22.13.30494) was downloaded on June 10th, 2022. This metadata file was was
623 filtered to include only entries from human hosts, only complete and high coverage entries, and
624 only those with complete sampling dates. The filtered metadata entries were downsampled to at
625 most 100 per month. Downsampling was conducted in Python v3.9.4 (Van Rossum and Drake

2009)using Pandas v1.1.4 (10.5281/zenodo.3509134) and Numpy v1.19.4 (10.1038/s41586-020-2649-2). The selected sequences were downloaded from GISAID EpiCov and aligned to Wuhan/WIV04 (EPI_ISL_402124) using MAFFT v7.464 (10.1093/molbev/mst010), removing any insertions to the reference. IQtree v2.1.3 (10.1093/molbev/msaa015) was used to infer a phylogenetic tree of the aligned sequences using a GTR+G4 substitution model, saving with Wuhan/WIV04 as the outgroup. TreeTime v0.8.0 (10.1093/ve/vex042) was used to filter sequences to include only those falling within four interquartile ranges of the best fit molecular clock, rooting at Wuhan/WIV04. Wuhan/WIV04 was forced to be included in the filtered tree and no other tips were identified as failing this filter.

For each of the amino acids of interest (ORF1ab R124, ORF1ab S1150, ORF1ab P5402), we first identified the corresponding nucleotide positions and then identified the nucleotide identity at each of those sites for each sequence in the alignment. These nucleotide identifies were used to infer the amino acid for each sequence at each position. Note that this method does not account for the presence of frame shift mutations, however, we expect these to be sufficiently rare as to not bias our results.

For each of the four substitutions we plotted the downsampled phylogenetic tree, labeling any tips with amino acids that did not match the reference. Any tips in which any of the nucleotides in the codon of interest were deleted or ambiguously genotyped were ignored. Visualization was done in Python using Matplotlib v3.5.1 (10.1109/MCSE.2007.55) and Baltic v0.1.5 [https://github.com/evogytis/baltic].

Substitution frequency analysis

The GISAID EpiCoV "MSA full" alignment was downloaded on June 2nd, 2022. All sequences in this file have been aligned to Wuhan/WIV04 using MAFFT, retaining any insertions relative to the reference. Full details on how this file were generated are available from GISAID.

For each of the four amino acid positions of interest we first identified the corresponding nucleotide positions in the gapped alignment and identified the nucleotides at each of these for each position. These nucleotides were used to infer the amino acid for each sequence in the full alignment at each position. Similarly to above, this method does not account for frameshift mutations.

For each amino acid position we identified the percentage of sequences harboring non-reference amino acids per month, ignoring any sequences in each one of the nucleotides was deleted or ambiguously genotyped. All amino acid identifies with a maximum monthly frequency less than or equal to 0.01% were grouped into an "Other" category. This analysis was conducted in Python using Pandas and visualized with Matplotlib.

References:

Amato, Katherine A., Luis A. Haddock, Katarina M. Braun, Victoria Meliopoulos, Brandi Livingston, Rebekah Honce, Grace A. Schaack, et al. 2021. “Influenza A Virus Undergoes Compartmentalized Replication in Vivo Dominated by Stochastic Bottlenecks.” https://doi.org/10.1101/2021.09.28.462198.

Avanzato, Victoria A., M. Jeremiah Matson, Stephanie N. Seifert, Rhys Pryce, Brandi N. Williamson, Sarah L. Anzick, Kent Barbian, et al. 2020. “Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised

674 Individual with Cancer." *Cell* 183 (7): 1901-1912.e9.
675 https://doi.org/10.1016/j.cell.2020.10.049.
676 Baang, Ji Hoon, Christopher Smith, Carmen Mirabelli, Andrew L Valesano, David M Manthei,
677 Michael A Bachman, Christiane E Wobus, et al. 2021. "Prolonged Severe Acute
678 Respiratory Syndrome Coronavirus 2 Replication in an Immunocompromised Patient."
679 *The Journal of Infectious Diseases* 223 (1): 23–27. https://doi.org/10.1093/infdis/jiaa666.
680 Braun, Katarina M., Gage K. Moreno, Cassia Wagner, Molly A. Accola, William M. Rehrauer,
681 David A. Baker, Katia Koelle, et al. 2021. "Acute SARS-CoV-2 Infections Harbor Limited
682 within-Host Diversity and Transmit via Tight Transmission Bottlenecks." *PLOS*
683 *Pathogens* 17 (8): e1009849. https://doi.org/10.1371/journal.ppat.1009849.
684 Choi, Bina, Manish C. Choudhary, James Regan, Jeffrey A. Sparks, Robert F. Padera, Xueting
685 Qiu, Isaac H. Solomon, et al. 2020. "Persistence and Evolution of SARS-CoV-2 in an
686 Immunocompromised Host." *The New England Journal of Medicine* 383 (23): 2291–93.
687 https://doi.org/10.1056/NEJMc2031364.
688 Cobey, Sarah, Daniel B. Larremore, Yonatan H. Grad, and Marc Lipsitch. 2021. "Concerns
689 about SARS-CoV-2 Evolution Should Not Hold Back Efforts to Expand Vaccination."
690 *Nature Reviews Immunology*, April, 1–6. https://doi.org/10.1038/s41577-021-00544-9.
691 Colson, Philippe, Jeremy Delerce, Emilie Burel, Mamadou Beye, Pierre-Edouard Fournier,
692 Anthony Levasseur, Jean-Christophe Lagier, and Didier Raoult. 2022. "Occurrence of a
693 Substitution or Deletion of SARS-CoV-2 Spike Amino Acid 677 in Various Lineages in
694 Marseille, France." *Virus Genes* 58 (1): 53–58. https://doi.org/10.1007/s11262-021-
695 01877-2.
696 Corey, Lawrence, Chris Beyrer, Myron S. Cohen, Nelson L. Michael, Trevor Bedford, and
697 Morgane Rolland. 2021. "SARS-CoV-2 Variants in Patients with Immunosuppression."
698 *New England Journal of Medicine* 385 (6): 562–66.
699 https://doi.org/10.1056/NEJMsb2104756.
700 Debbink, Kari, John T. McCrone, Joshua G. Petrie, Rachel Truscon, Emileigh Johnson, Emily K.
701 Mantlo, Arnold S. Monto, and Adam S. Lauring. 2017. "Vaccination Has Minimal Impact
702 on the Intrahost Diversity of H3N2 Influenza Viruses." *PLOS Pathogens* 13 (1):
703 e1006194. https://doi.org/10.1371/journal.ppat.1006194.
704 Ghosh, Asish Kumar, Marco Kaiser, Md. Maruf Ahmed Molla, Tasnim Nafisa, Mahmuda
705 Yeasmin, Rifat Hossain Ratul, Md. Mohiuddin Sharif, et al. 2021. "Molecular and
706 Serological Characterization of the SARS-CoV-2 Delta Variant in Bangladesh in 2021."
707 *Viruses* 13 (11): 2310. https://doi.org/10.3390/v13112310.
708 Hadfield, James, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton
709 Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. 2018. "Nextstrain:
710 Real-Time Tracking of Pathogen Evolution." *Bioinformatics* 34 (23): 4121–23.
711 https://doi.org/10.1093/bioinformatics/bty407.
712 Harvey, William T., Alessandro M. Carabelli, Ben Jackson, Ravindra K. Gupta, Emma C.
713 Thomson, Ewan M. Harrison, Catherine Ludden, et al. 2021. "SARS-CoV-2 Variants,
714 Spike Mutations and Immune Escape." *Nature Reviews Microbiology* 19 (7): 409–24.
715 https://doi.org/10.1038/s41579-021-00573-0.
716 Hodcroft, Emma B., Daryl B. Domman, Daniel J. Snyder, Kasopefoluwa Y. Oguntuyo, Maarten
717 Van Diest, Kenneth H. Densmore, Kurt C. Schwalm, et al. 2021. "Emergence in Late
718 2020 of Multiple Lineages of SARS-CoV-2 Spike Protein Variants Affecting Amino Acid
719 Position 677." medRxiv. https://doi.org/10.1101/2021.02.12.21251658.
720 "Issues with SARS-CoV-2 Sequencing Data - SARS-CoV-2 Coronavirus / NCoV-2019 Genomic
721 Epidemiology." 2020. Virological. May 5, 2020. https://virological.org/t/issues-with-sars-
722 cov-2-sequencing-data/473.

723 Johnson, Bryan A., Xuping Xie, Adam L. Bailey, Birte Kalveram, Kumari G. Lokugamage,
724 Antonio Muruato, Jing Zou, et al. 2021. "Loss of Furin Cleavage Site Attenuates SARS-
725 CoV-2 Pathogenesis." *Nature* 591 (7849): 293–99. https://doi.org/10.1038/s41586-021-
726 03237-4.
727 Johnson, Bryan A., Yiyang Zhou, Kumari G. Lokugamage, Michelle N. Vu, Nathen Bopp,
728 Patricia A. Crocquet-Valdes, Craig Schindewolf, et al. 2021. "Nucleocapsid Mutations in
729 SARS-CoV-2 Augment Replication and Pathogenesis."
730 https://doi.org/10.1101/2021.10.14.464390.
731 Ke, Ruian, Pamela P. Martinez, Rebecca L. Smith, Laura L. Gibson, Chad J. Achenbach, Sally
732 McFall, Chao Qi, et al. 2022. "Longitudinal Analysis of SARS-CoV-2 Vaccine
733 Breakthrough Infections Reveal Limited Infectious Virus Shedding and Restricted Tissue
734 Distribution." *Open Forum Infectious Diseases*, April, ofac192.
735 https://doi.org/10.1093/ofid/ofac192.
736 Ke, Ruian, Pamela P. Martinez, Rebecca L. Smith, Laura L. Gibson, Agha Mirza, Madison
737 Conte, Nicholas Gallagher, et al. 2022. "Daily Longitudinal Sampling of SARS-CoV-2
738 Infection Reveals Substantial Heterogeneity in Infectiousness." *Nature Microbiology* 7
739 (5): 640–52. https://doi.org/10.1038/s41564-022-01105-z.
740 Kemp, Steven A., Dami A. Collier, Rawlings P. Datir, Isabella A. T. M. Ferreira, Salma Gayed,
741 Aminu Jahun, Myra Hosmillo, et al. 2021. "SARS-CoV-2 Evolution during Treatment of
742 Chronic Infection." *Nature* 592 (7853): 277–82. https://doi.org/10.1038/s41586-021-
743 03291-y.
744 Kryazhimskiy, Sergey, and Joshua B. Plotkin. 2008. "The Population Genetics of DN/DS." *PLoS*
745 *Genetics* 4 (12): e1000304. https://doi.org/10.1371/journal.pgen.1000304.
746 Liu, Yang, Jianying Liu, Bryan A. Johnson, Hongjie Xia, Zhiqiang Ku, Craig Schindewolf, Steven
747 G. Widen, et al. 2021. "Delta Spike P681R Mutation Enhances SARS-CoV-2 Fitness
748 over Alpha Variant." https://doi.org/10.1101/2021.08.12.456173.
749 Liu, Zhuoming, Laura A. VanBlargan, Louis-Marie Bloyet, Paul W. Rothlauf, Rita E. Chen,
750 Spencer Stumpf, Haiyan Zhao, et al. 2021. "Identification of SARS-CoV-2 Spike
751 Mutations That Attenuate Monoclonal and Serum Antibody Neutralization." *Cell Host &*
752 *Microbe* 29 (3): 477-488.e4. https://doi.org/10.1016/j.chom.2021.01.014.
753 Lythgoe, Katrina A., Matthew Hall, Luca Ferretti, Mariateresa de Cesare, George MacIntyre-
754 Cockett, Amy Trebes, Monique Andersson, et al. 2021. "SARS-CoV-2 within-Host
755 Diversity and Transmission." *Science* 372 (6539): eabg0821.
756 https://doi.org/10.1126/science.abg0821.
757 Martin, Michael A., and Katia Koelle. 2021. "Comment on 'Genomic Epidemiology of
758 Superspreading Events in Austria Reveals Mutational Dynamics and Transmission
759 Properties of SARS-CoV-2.'" *Science Translational Medicine* 13 (617): eabh1803.
760 https://doi.org/10.1126/scitranslmed.abh1803.
761 Mears, Harriet V, George R Young, Theo Sanderson, Ruth Harvey, Margaret Crawford, Daniel
762 M Snell, Ashley S Fowler, et al. 2022. "Emergence of New Subgenomic MRNAs in
763 SARS-CoV-2," 19.
764 Orton, Richard J., Caroline F. Wright, Donald P. King, and Daniel T. Haydon. 2020. "Estimating
765 Viral Bottleneck Sizes for FMDV Transmission within and between Hosts and
766 Implications for the Rate of Viral Evolution." *Interface Focus* 10 (1): 20190066.
767 https://doi.org/10.1098/rsfs.2019.0066.
768 Ou, Jianhong, and Lihua Julie Zhu. 2019. "TrackViewer: A Bioconductor Package for Interactive
769 and Integrative Visualization of Multi-Omics Data." *Nature Methods* 16 (6): 453–54.
770 https://doi.org/10.1038/s41592-019-0430-y.

Peacock, Thomas P., Daniel H. Goldhill, Jie Zhou, Laury Baillon, Rebecca Frise, Olivia C. Swann, Ruthiran Kugathasan, et al. 2021. “The Furin Cleavage Site in the SARS-CoV-2 Spike Protein Is Required for Transmission in Ferrets.” *Nature Microbiology* 6 (7): 899–909. https://doi.org/10.1038/s41564-021-00908-w.

Pfeiffer, Julie K., and Karla Kirkegaard. 2006. “Bottleneck-Mediated Quasispecies Restriction during Spread of an RNA Virus from Inoculation Site to Brain.” *Proceedings of the National Academy of Sciences* 103 (14): 5520–25. https://doi.org/10.1073/pnas.0600834103.

“Preliminary Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in the UK Defined by a Novel Set of Spike Mutations - SARS-CoV-2 Coronavirus / NCoV-2019 Genomic Epidemiology.” 2020. Virological. December 18, 2020. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.

Ranoa, Diana Rose E., Robin L. Holland, Fadi G. Alnaji, Kelsie J. Green, Leyi Wang, Christopher B. Brooke, Martin D. Burke, Timothy M. Fan, and Paul J. Hergenrother. 2020. “Saliva-Based Molecular Testing for SARS-CoV-2 That Bypasses RNA Extraction.” bioRxiv. https://doi.org/10.1101/2020.06.18.159434.

Ranoa, Diana Rose E., Robin L. Holland, Fadi G. Alnaji, Kelsie J. Green, Leyi Wang, Richard L. Fredrickson, Tong Wang, et al. 2021. “Mitigation of SARS-CoV-2 Transmission at a Large Public University.” Preprint. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2021.08.03.21261548.

Saad-Roy, Chadi M., Sinead E. Morris, C. Jessica E. Metcalf, Michael J. Mina, Rachel E. Baker, Jeremy Farrar, Edward C. Holmes, et al. 2021. “Epidemiological and Evolutionary Considerations of SARS-CoV-2 Vaccine Dosing Regimes.” *Science* 372 (6540): 363–70. https://doi.org/10.1126/science.abg8663.

Saxena, Shailendra K., Swatantra Kumar, Saniya Ansari, Janusz T. Paweska, Vimal K. Maurya, Anil K. Tripathi, and Ahmed S. Abdel-Moneim. 2022. “Characterization of the Novel SARS-CoV-2 Omicron (B.1.1.529) Variant of Concern and Its Global Perspective.” *Journal of Medical Virology* 94 (4): 1738–44. https://doi.org/10.1002/jmv.27524.

Smith, Rebecca L, Laura L Gibson, Pamela P Martinez, Ruian Ke, Agha Mirza, Madison Conte, Nicholas Gallagher, et al. 2021. “Longitudinal Assessment of Diagnostic Test Performance Over the Course of Acute SARS-CoV-2 Infection.” *The Journal of Infectious Diseases* 224 (6): 976–82. https://doi.org/10.1093/infdis/jiab337.

Tonkin-Hill, Gerry, Inigo Martincorena, Roberto Amato, Andrew RJ Lawson, Moritz Gerstung, Ian Johnston, David K Jackson, et al. 2021. “Patterns of Within-Host Genetic Diversity in SARS-CoV-2.” Edited by Richard A Neher, Sara L Sawyer, Richard A Neher, and Adam S Lauring. *ELife* 10 (August): e66857. https://doi.org/10.7554/eLife.66857.

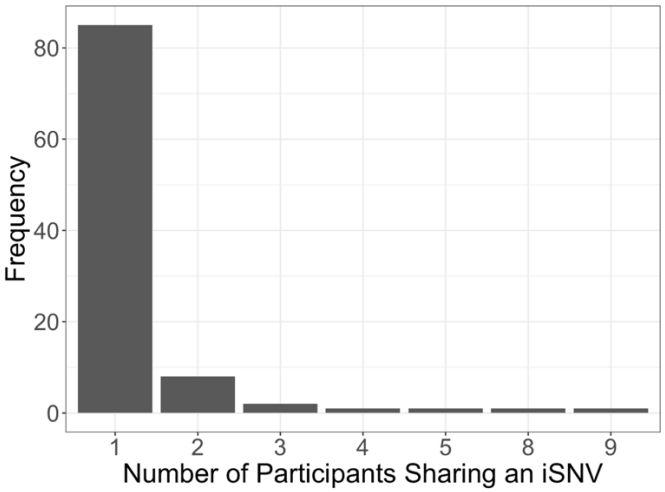
Truong, Thao T., Alex Ryutov, Utsav Pandey, Rebecca Yee, Lior Goldberg, Deepa Bhojwani, Paibel Aguayo-Hiraldo, et al. 2021. “Increased Viral Variants in Children and Young Adults with Impaired Humoral Immunity and Persistent SARS-CoV-2 Infection: A Consecutive Case Series.” *EBioMedicine* 67 (May): 103355. https://doi.org/10.1016/j.ebiom.2021.103355.

Valesano, Andrew L., Kalee E. Rumfelt, Derek E. Dimcheff, Christopher N. Blair, William J. Fitzsimmons, Joshua G. Petrie, Emily T. Martin, and Adam S. Lauring. 2021. “Temporal Dynamics of SARS-CoV-2 Mutation Accumulation within and across Infected Hosts.” *PLOS Pathogens* 17 (4): e1009499. https://doi.org/10.1371/journal.ppat.1009499.

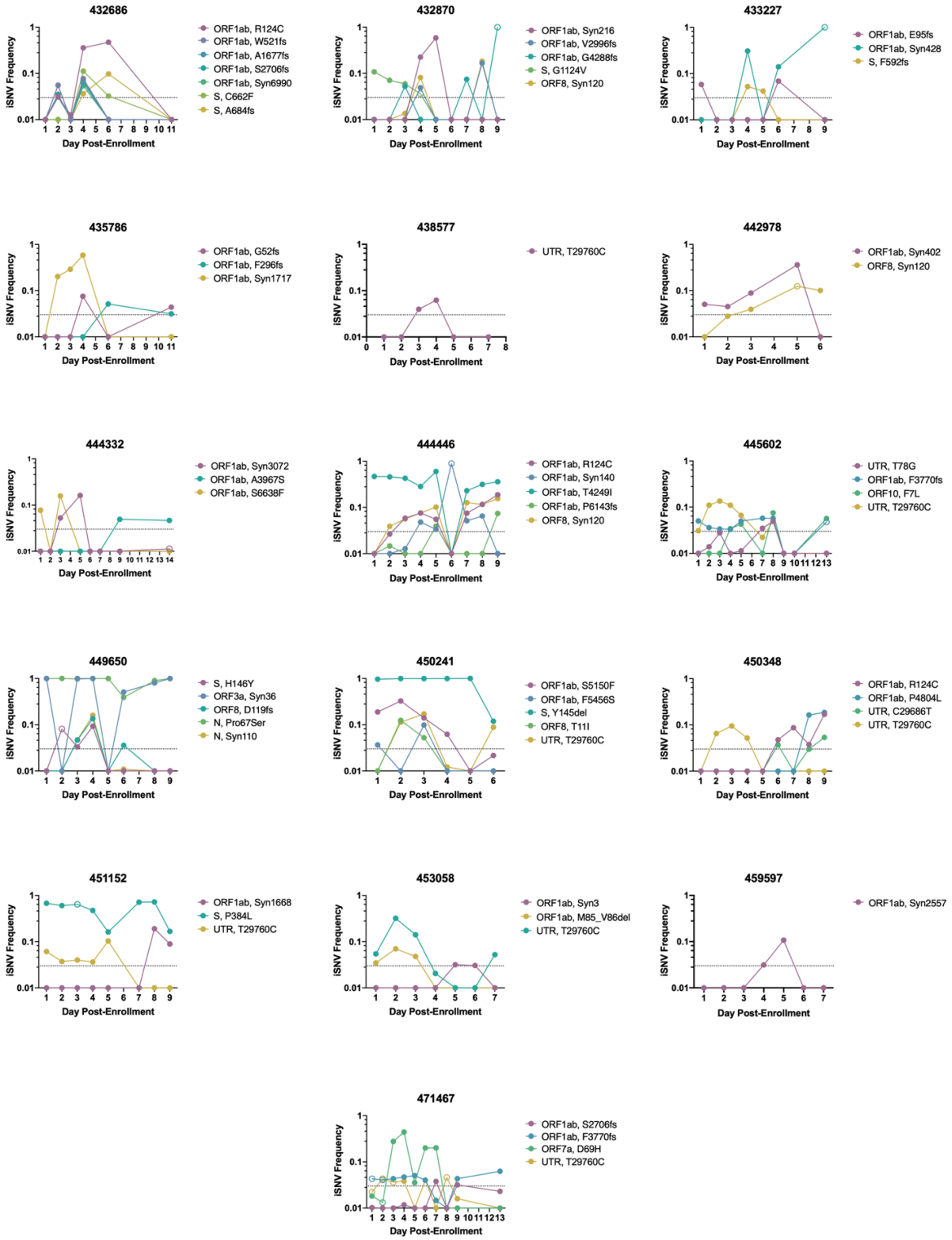
Van Rossum, G., and F.L. Drake. 2009. “Python 3 Reference Manual.” Scotts Valley, CA: CreateSpace.

819 Yang, Dong, and Julian L. Leibowitz. 2015. “The Structure and Functions of Coronavirus
820 Genomic 3′ and 5′ Ends.” *Virus Research*, Special Issue: Functions of the ends of
821 positive strand RNA virus genomes, 206 (August): 120–33.
822 https://doi.org/10.1016/j.virusres.2015.02.025.
823
824

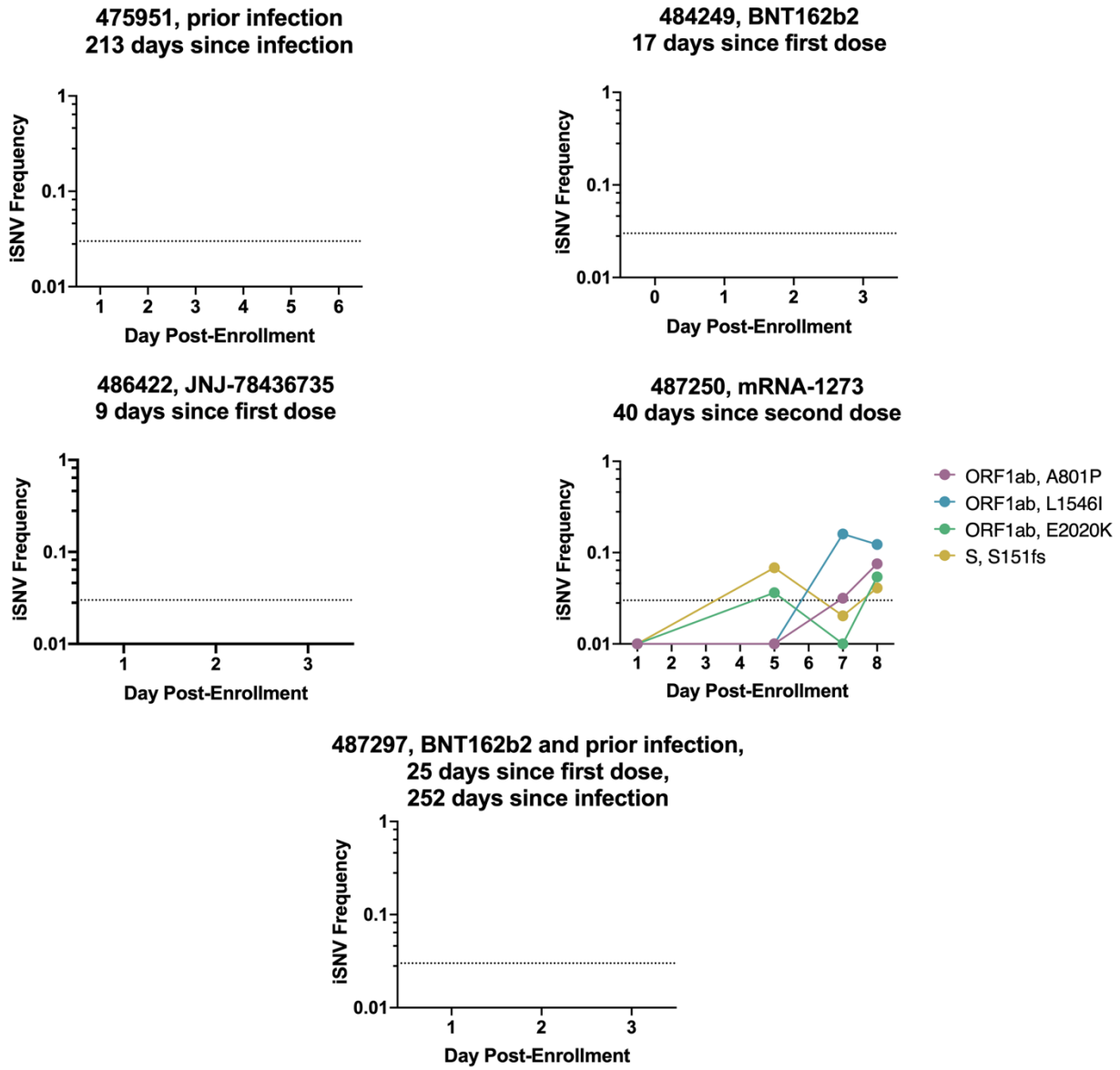
825 **Supplemental Figures:**
826



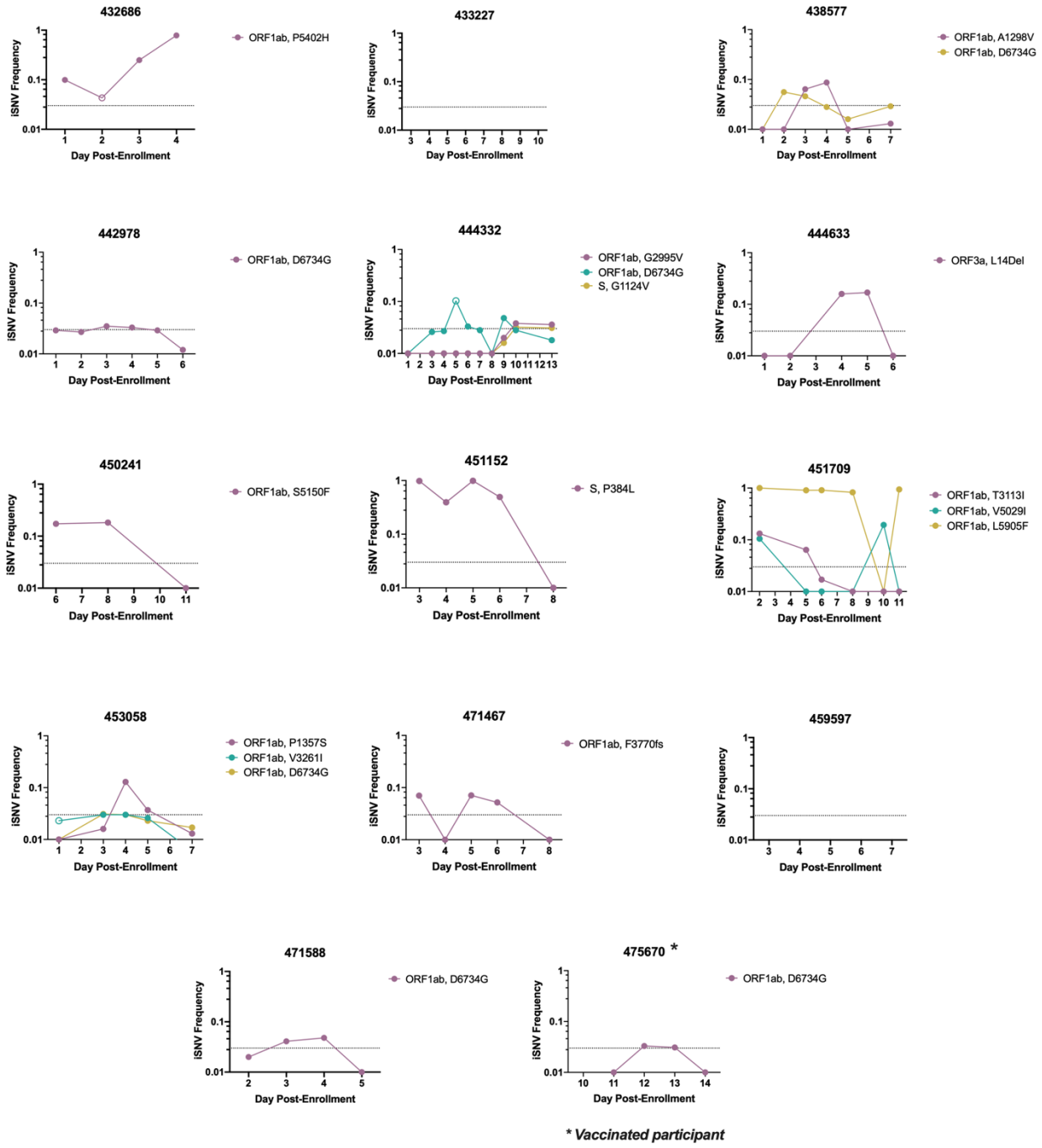
827
828
829 ***Supplemental Figure 1: Distribution of shared iSNVs found across multiple naive***
830 ***participants.*** *No shared iSNVs were detected in more than 2 vaccinated participants.*
831



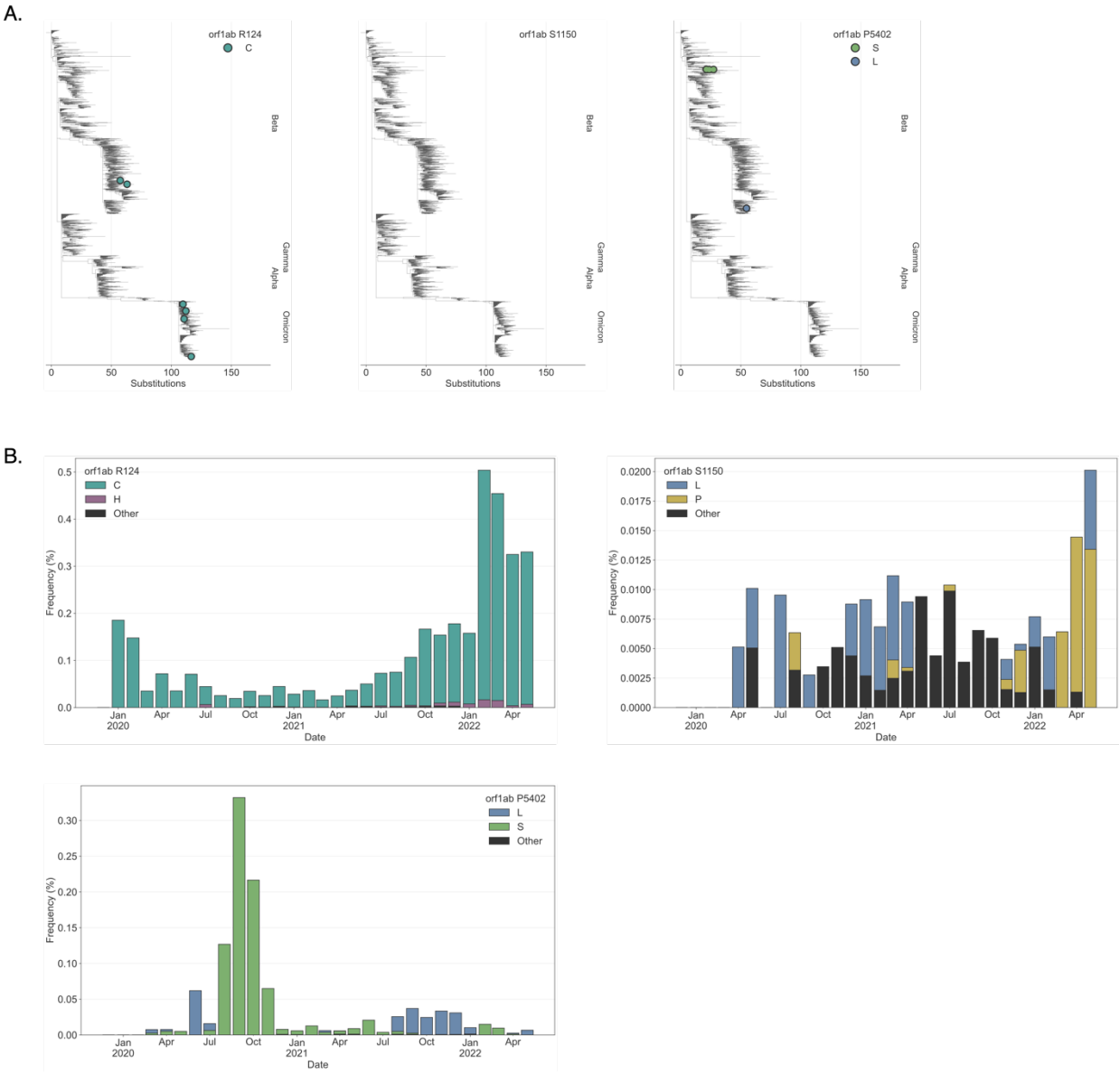
Supplemental Figure 2: iSNV dynamics over time from all unvaccinated participants. *Frequency tracking of selected iSNVs in unvaccinated participants. Dashed line marks frequency threshold of 0.03. Unfilled points mark iSNVs with read depths below the threshold of 1000 reads.*



Supplemental Figure 3: iSNV dynamics over time from all immune participants. *Frequency tracking of selected iSNVs in immune participants. Dashed line marks frequency threshold of 0.03. Unfilled points mark iSNVs with read depths below the threshold of 1000 reads.*



Supplemental Figure 4: iSNV dynamics over time from all nasal samples. *Frequency tracking of selected iSNVs from nasal swab samples. Dashed line marks frequency threshold of 0.03. Unfilled points mark iSNVs with read depths below the threshold of 1000 reads.*



***Supplemental Figure 5. Global occurrence of selected iSNVs**. **(A)** iSNV-associated amino acid changes plotted on a global phylogeny of SARS-CoV-2 sequences. **(B)** Frequency of iSNV-associated amino acid changes from January 2020 to April 2022.*