Cell Host & Microbe, Volume *30*

Supplemental information

Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2

Kathryn E. Kistler, John Huddleston, and Trevor Bedford

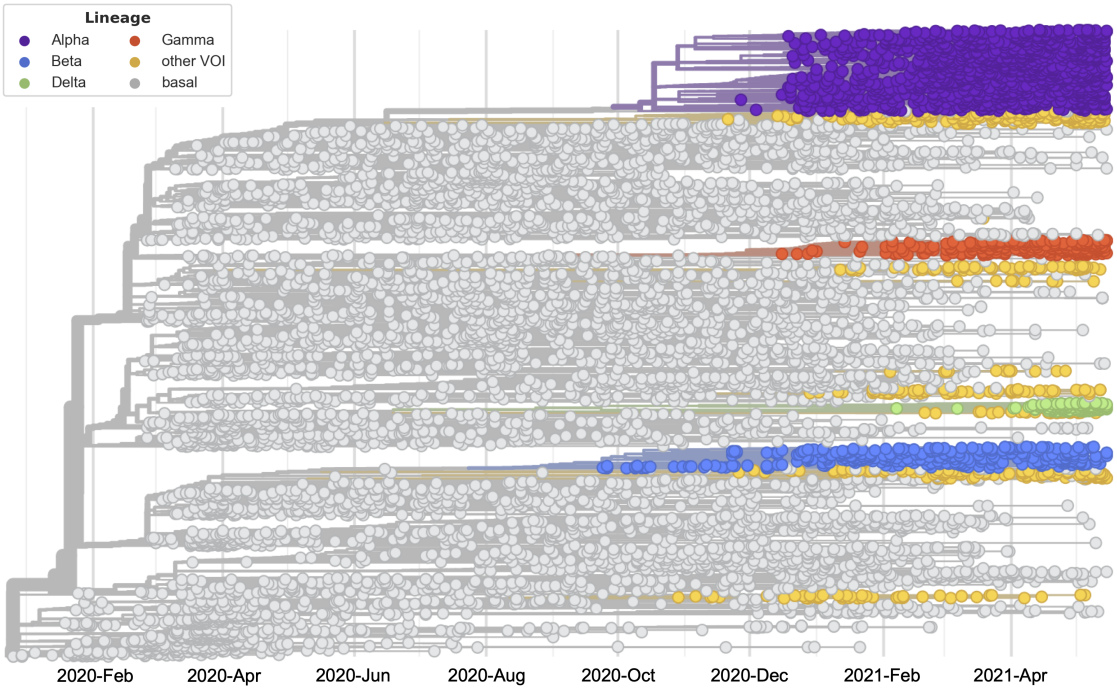

Figure S1. Phylogeny of 9544 SARS-CoV-2 genomes- related to Figure 1. Screenshot of the phylogeny used for the primary analyses in this manuscript. Tips and branches are colored according to viral lineage. The prominent Variant of Concern (VOC) lineages Alpha, Beta, Delta and Gamma are shown. The "other VOI" category includes 13 emerging lineages — WHO VOCS, VOIs, and prominent PANGO lineages (Rambaut et al., 2020). Colors match those used in Figure 1. An interactive version of this phylogeny can be accessed at nextstrain.org/groups/blab/ncov/adaptive-evolution/2021-05-15.

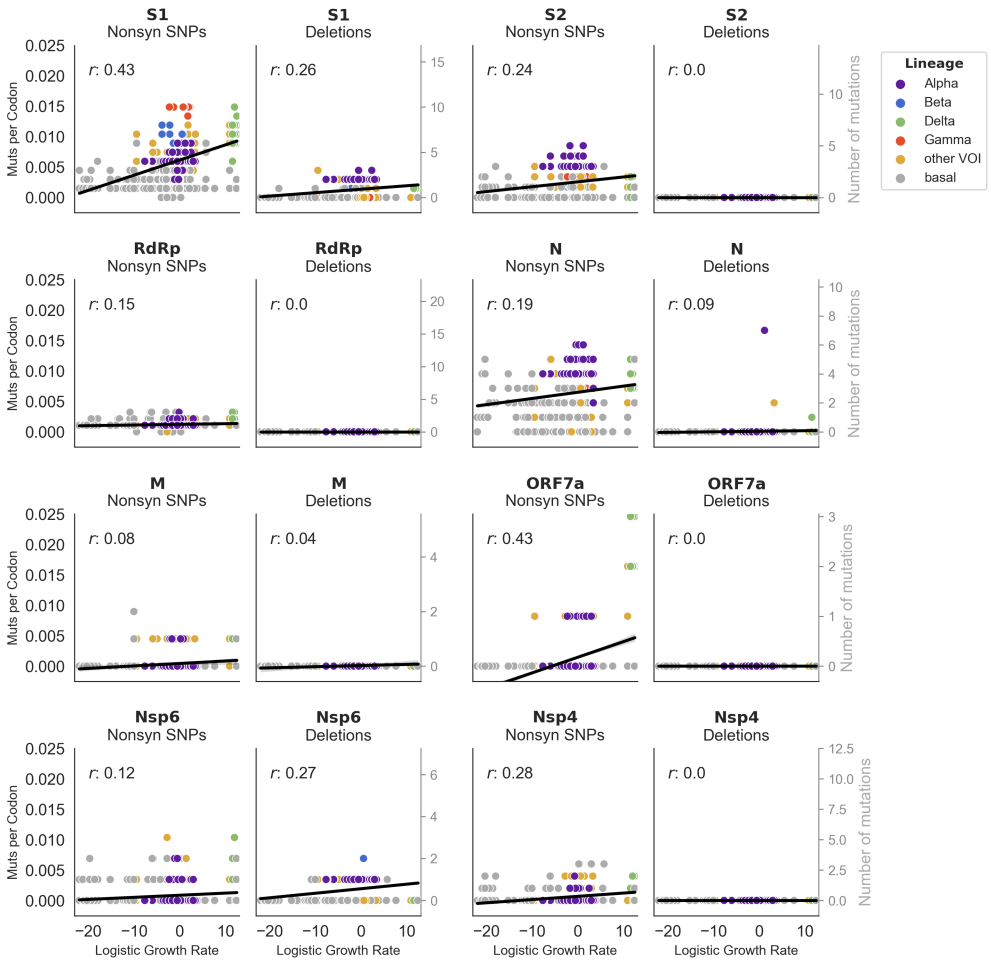

Figure S2. Deletions contribute to protein-coding changes in S1, N and Nsp6- related to Figure 1. For each gene nonsynonymous mutation accumulation is separated into nonsynonymous SNPs (left) and deletions (right). Accumulation of these mutations is plotted against logistic growth rate for 8 genes (or subunits), as in Figure 1B.

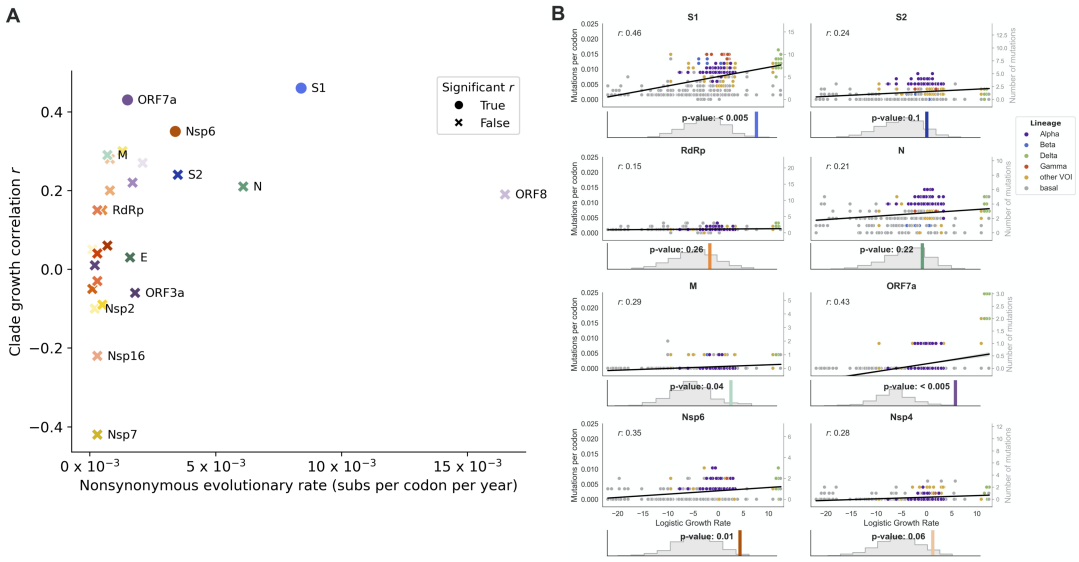

Figure S3. Correlation between nonsynonymous mutation accumulation and clade success is strongest in S1- related to Table 1. A) For every gene in the genome, the rate of nonsynonymous substitutions (and deletions) per codon per year is plotted against the correlation coefficient r of mutation accumulation with logistic growth. Circles indicate genes with significant r values at the p=0.01 level, and Xs indicate genes with insignificant r values. **B)** Nonsynonymous mutation accumulation (mutations per codon) is plotted against logistic growth rate for 8 genes (or subunits), as in Figure 1B. Histograms beneath each plot show the empirical correlation coefficient r (colored line) compared to the distribution of r coefficients from 1000 randomizations, as well as the p-value resulting from this comparison.

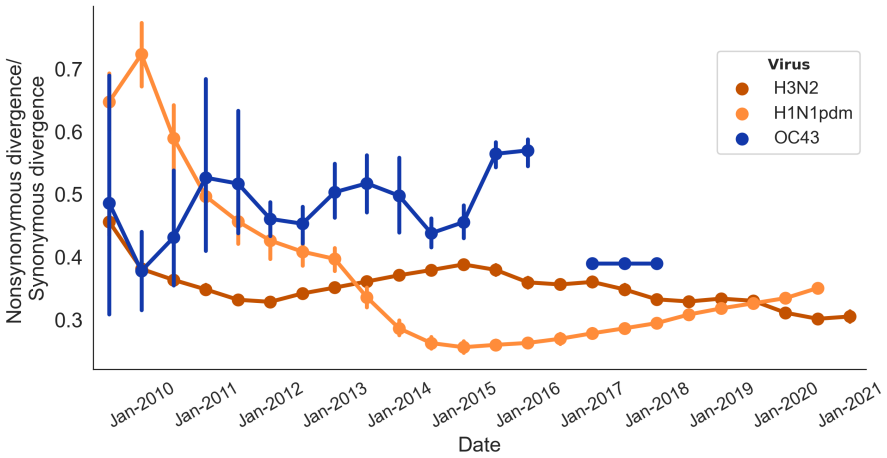

Figure S4. Ratio of nonsynonymous to synonymous divergence in HA1 subunit of seasonal influenza H3N2 and H1N1pdm and spike S1 subunit of seasonal coronavirus OC43- related to Figure 2. The mean and 95% confidence intervals for nonsynonymous/synonymous divergence ratios for the seasonal influenza HA1 subunits and seasonal coronavirus S1 subunit are shown over a 12-year period starting in January 2009. Divergence accumulation from the root is calculated as in Figure 2 except using windows of 1 year that overlap by half a year.

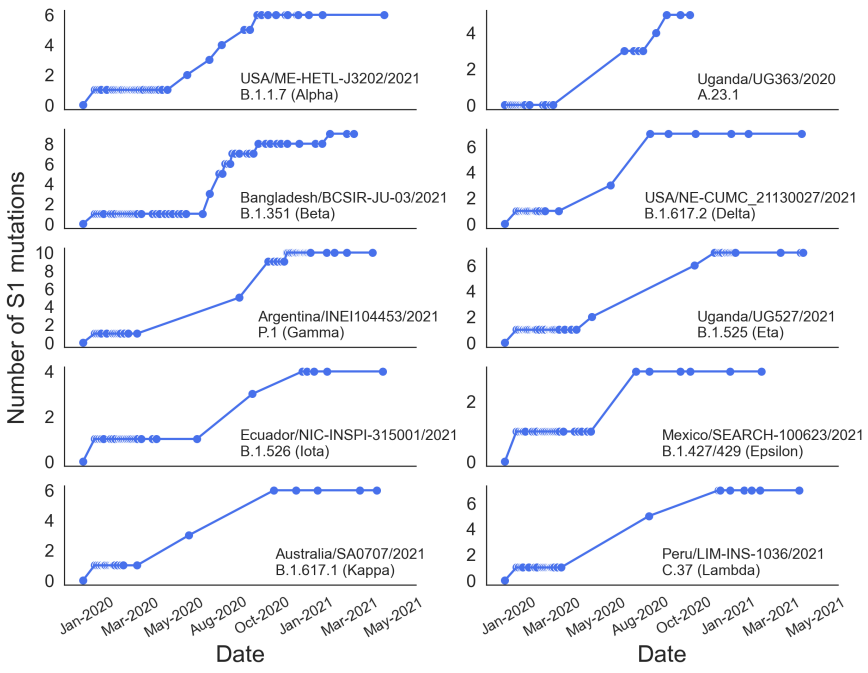

Figure S5. Temporal accumulation of S1 mutations on representative paths through the tree- related to Figure 3. The total number of accumulated S1 nonsynonymous mutations is counted at every branch along a path through the tree. This is plotted for 10 representative paths from the root to an isolate in an emerging lineage clade. The isolate and emerging lineage are labeled on each panel.

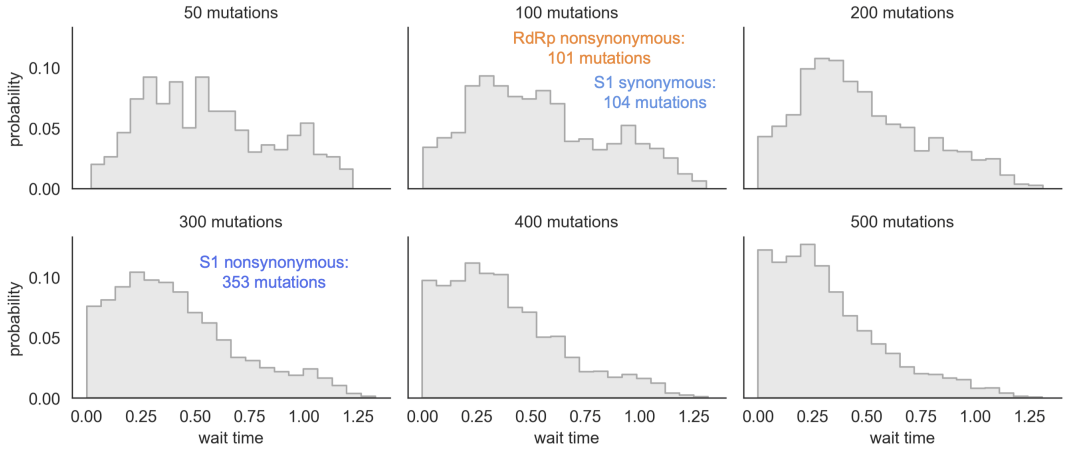

Figure S6. Distribution of expected wait times is affected by the number of mutations that occur across the phylogeny- related to Figure 3. The phylogeny was randomized with varying numbers of mutations to display the expected wait time distributions if 50, 100, 200, 300, 400 or 500 mutations occur on internal branches of the phylogeny. Each randomization is run for 10 iterations. The empirical number of S1 nonsynonymous, S1 synonymous, and RdRp nonsynonymous mutations observed on internal branches of the phylogeny are indicated.

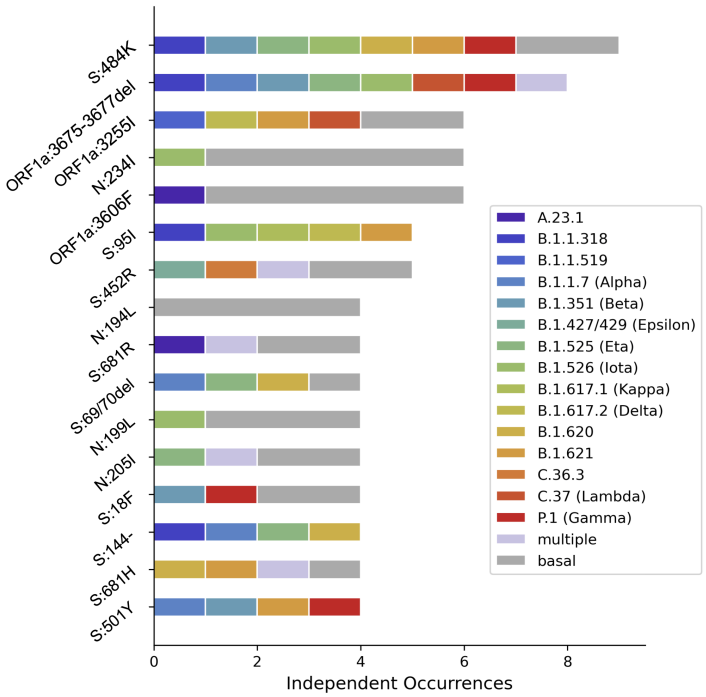

Figure S7. Every occurrence of the 3-amino acid deletion in Nsp6 resulted in an emerging lineage- related to Figure 4. Every occurrence of the convergently-evolved mutations is colored according to the emerging lineage it occurs at the base of. Multiple emerging lineages descending from the branch a mutation occurs on is represented by light purple.

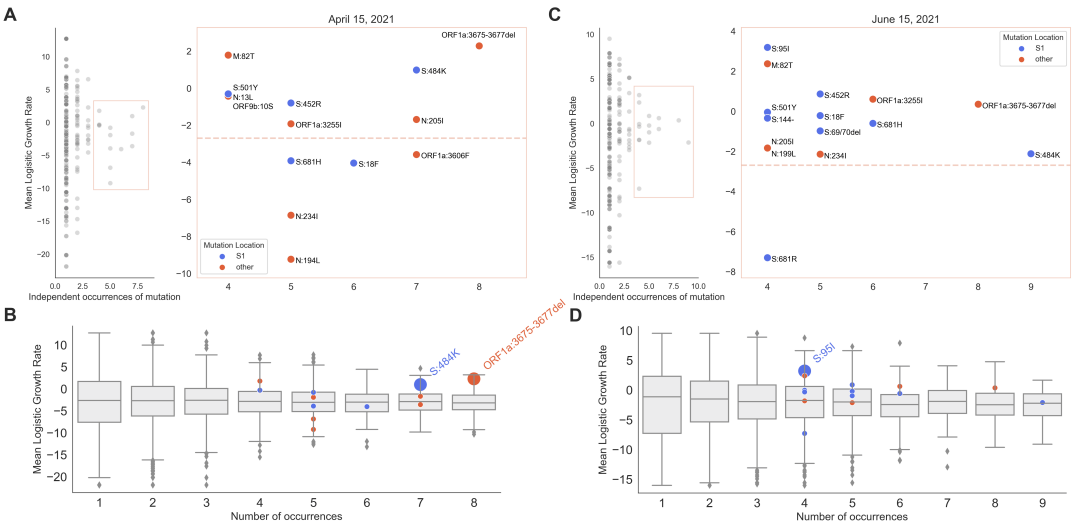

Figure S8. Analyses of convergent evolution shown 1 month before and 1 month after the primary analysis- related to Figure 4. A) Same as Figure 4A, completed using sequences up to April 15, 2021 (1 month before the primary analysis). **B)** Same as Figure 4B, completed using sequences up to April 15, 2021. **C)** Same as Figure 4A, completed using sequences up to June 15, 2021 (1 month after the primary analysis). **D)** Same as Figure 4B, completed using sequences up to June 15, 2021.

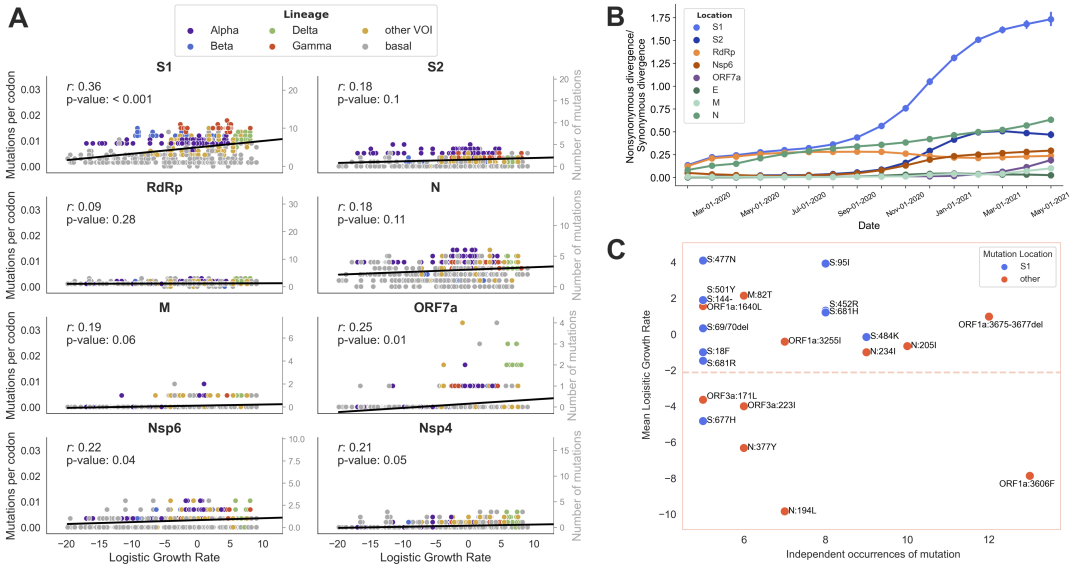

Figure S9. Primary results are reproduced using a phylogeny of 19,694 genomes-related to Figures 1, 2 and 4. Analyses presented in the primary figures were repeated using a tree containing twice as many samples nextstrain.org/groups/blab/ncov/adaptive-evolution/2021-05-15/20k. **A)** Correlation between nonsynonymous mutation accumulation and clade growth for 8 genes as in Figure S3. For each gene, the empirical correlation coefficient r is compared to a distribution of r coefficients from 1000 randomizations to determine the p-value. **B)** Nonsynonymous to synonymous divergence accumulation ratio over time as in Figure 2. **C)** Convergently-evolved mutations that occur 5 or more times independently and the mean growth rate of every clade containing that mutation are plotted, as in Figure 4A.

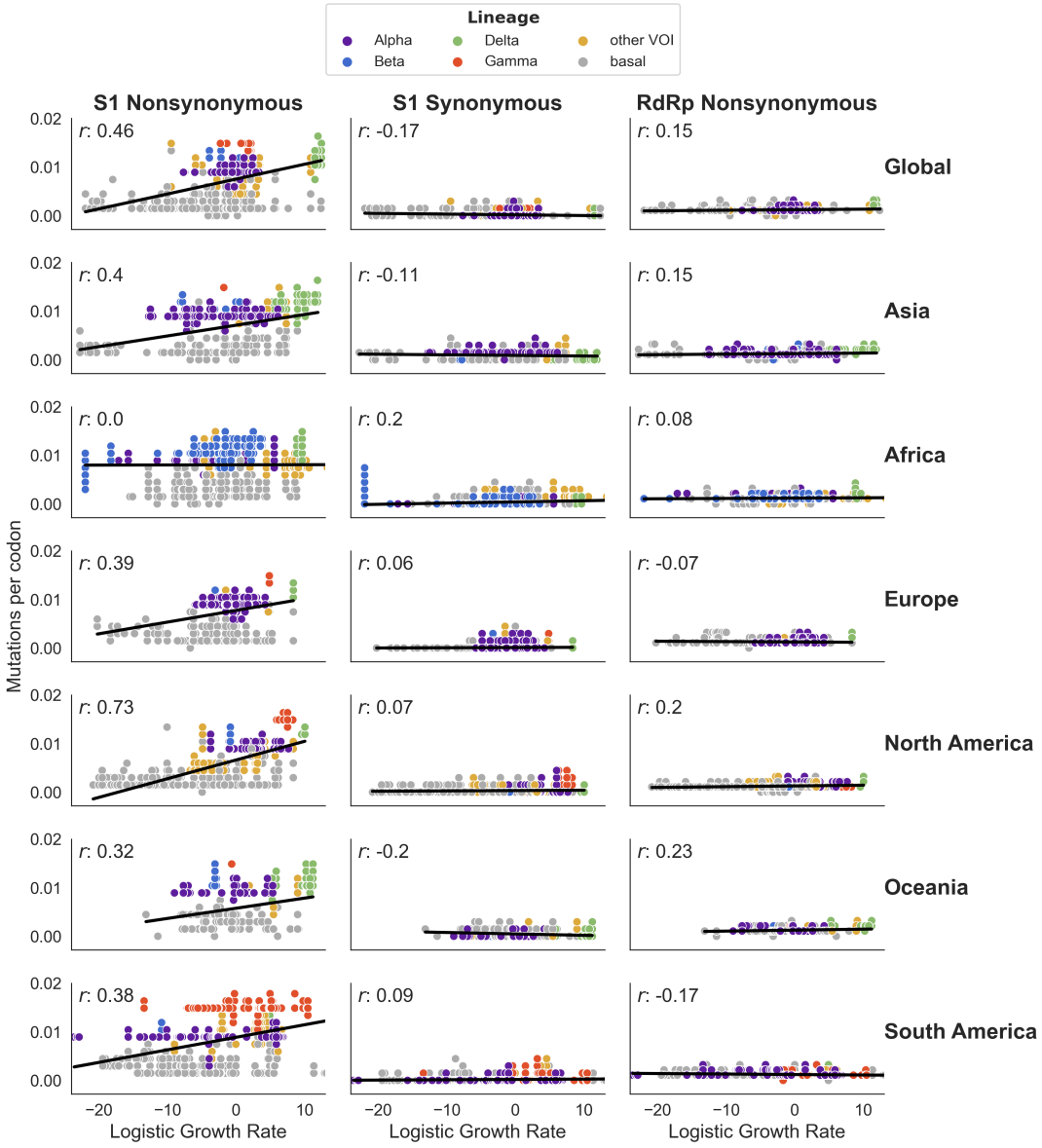

Figure S10. Correlation between S1 substitutions and clade growth rate is consistent across all geographic regions except one- related to Figure 1. Phylogenies were built using only around 10,000 samples from a single geographic region (Africa, Asia, Europe, North America, Oceania, and South America). The correlation between S1 nonsynonymous, S1 synonymous and RdRp nonsynonymous mutation accumulation and clade growth rate (as in Figure 1) is plotted for each geographic region.

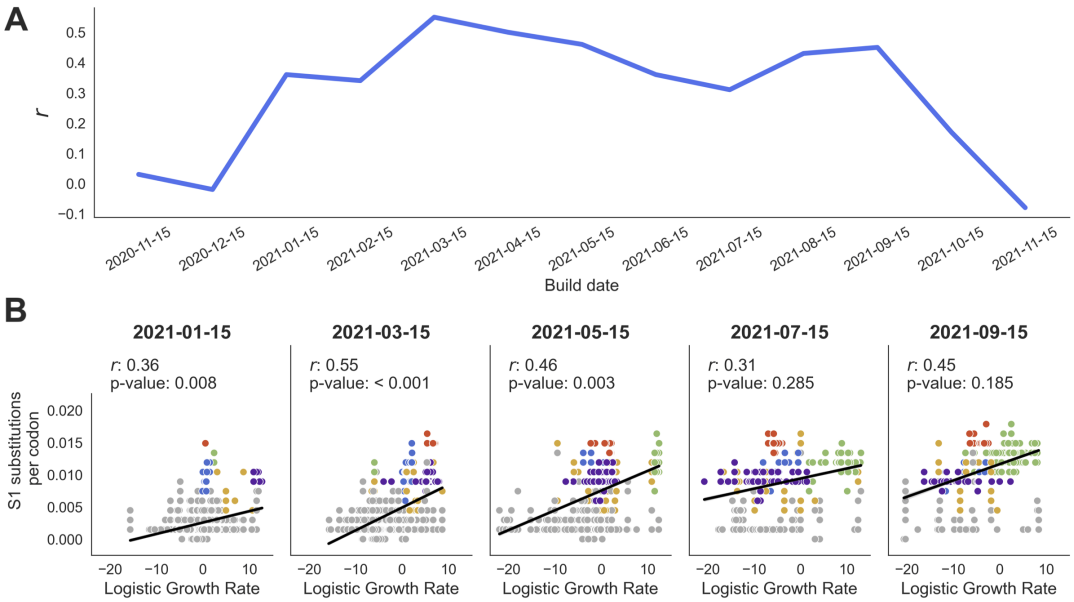

Figure S11. Correlation between S1 substitutions and clade success over time- related to Figure 1. A) The correlation coefficient r is calculated between logistic growth rate and S1 substitutions for every clade within 6 weeks preceding the build date (x-axis). **B)** Mutation accumulation is plotted against logistic growth rate and the points are fit by linear regression (as in Figure 1). p-values are computed by comparing the observed correlation coefficient r to the distribution of r coefficients from 1000 trees where S1 substitutions are randomized across branches according to a multinomial draw.