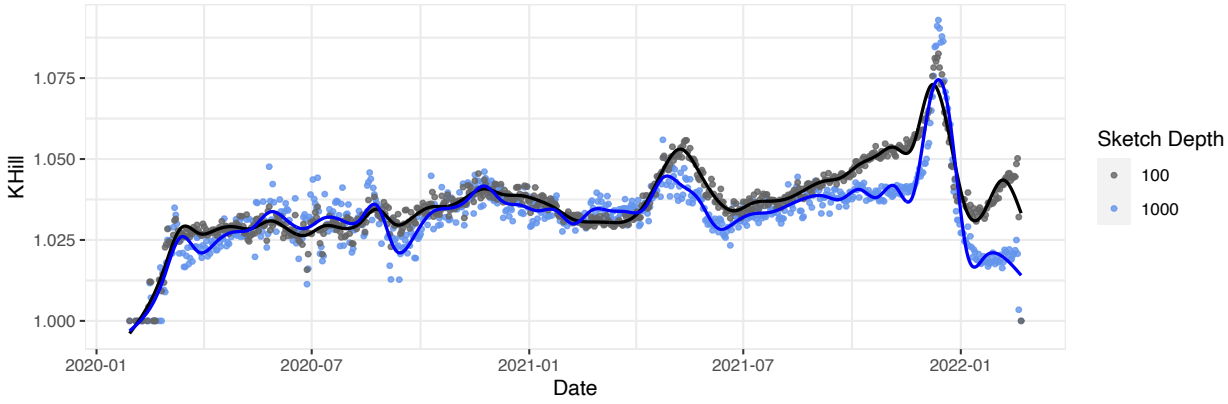
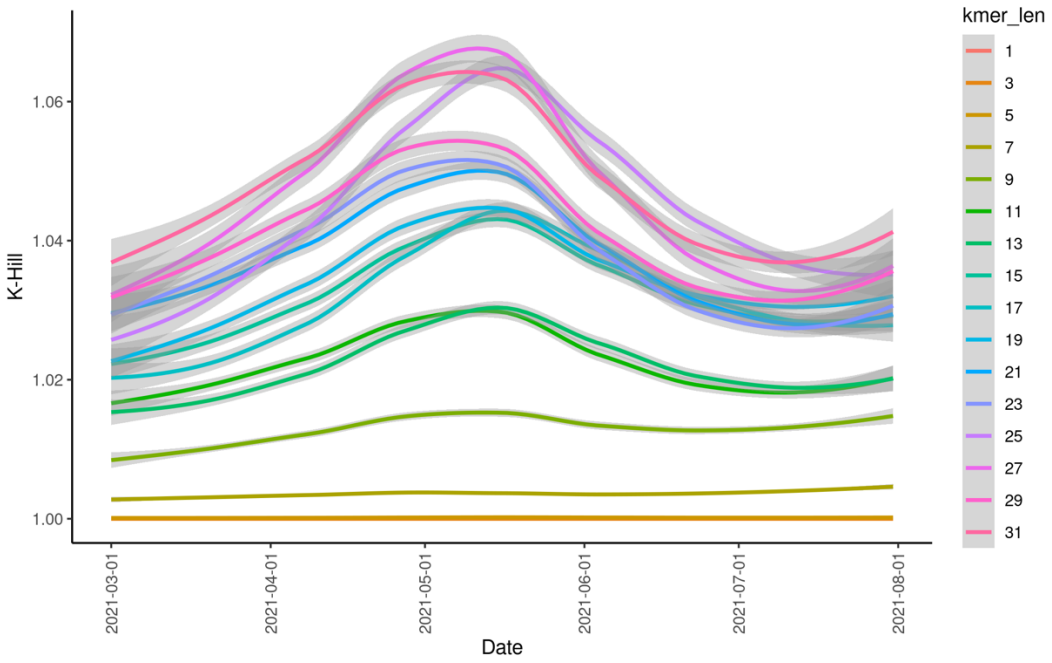
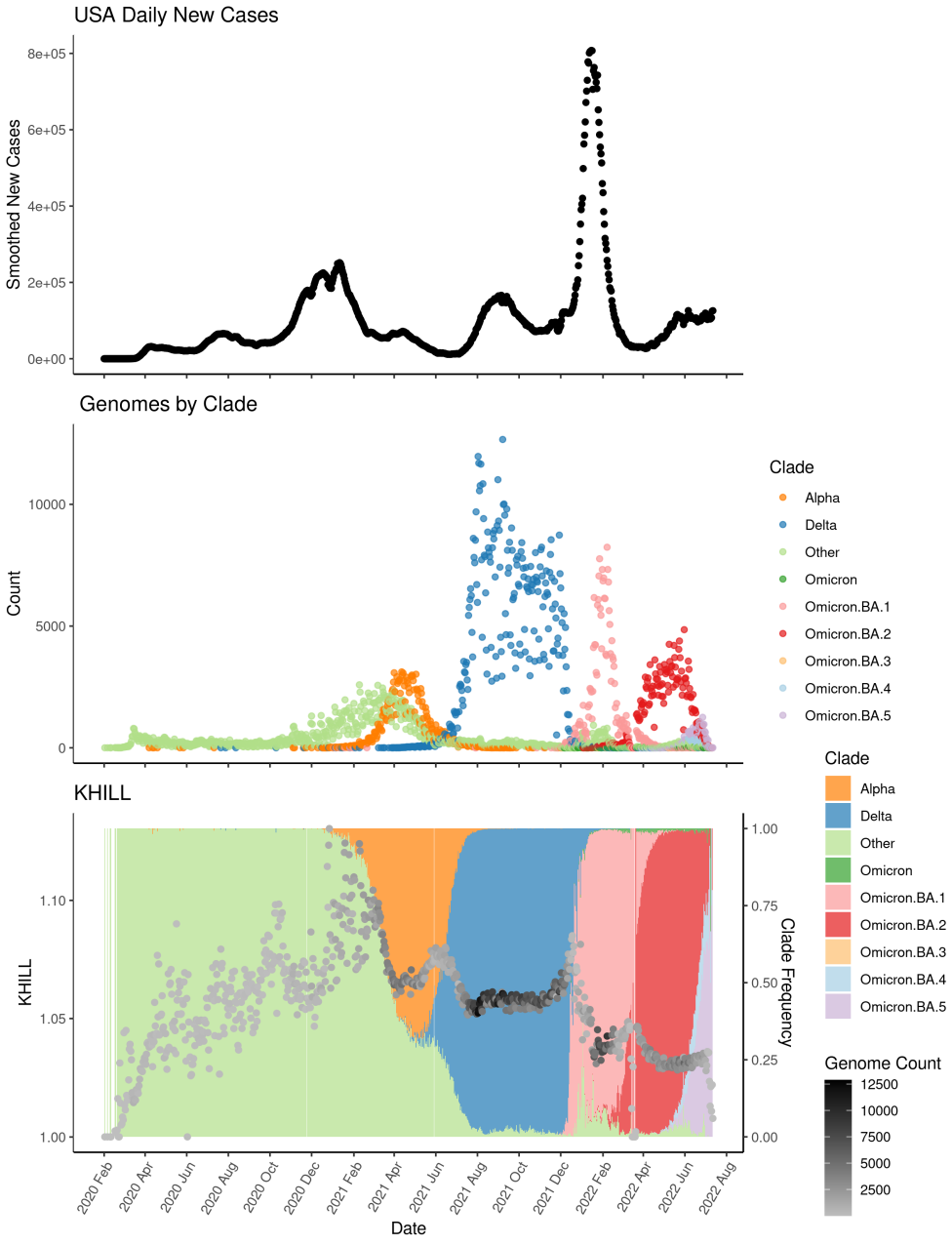
Supplemental Figures



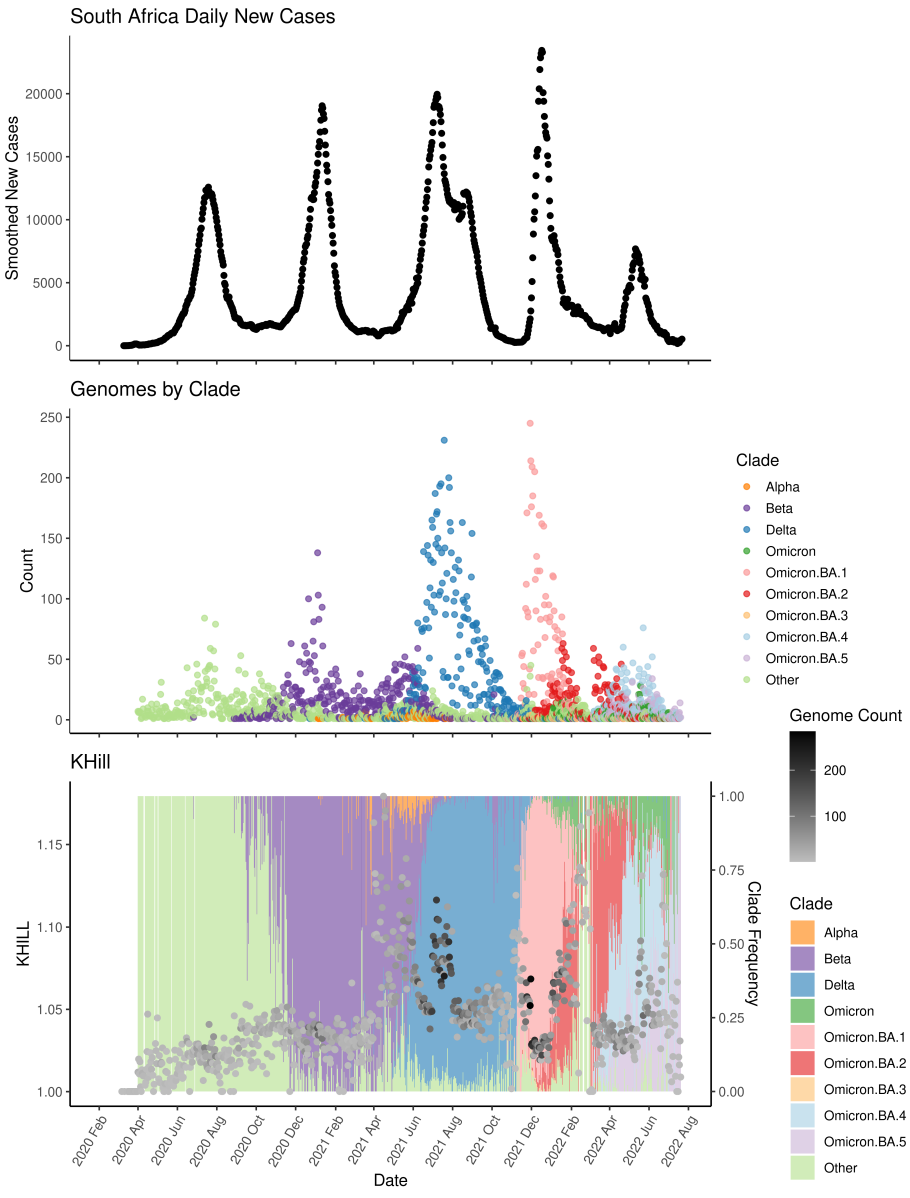
Supplemental Figure 1. Sketch depth and the detection of subvariants. We used a generalized additive model to fit a cubic spline (lines) with a kurtosis of 50 to the UK covid pandemic KHill statistics (y-axis, as a response to date on the x-axis), calculated with sketch rates of 100 and 1,000. Red arrows indicate the additional diversity captured with a deeper sketch during the Delta and Omicron waves.



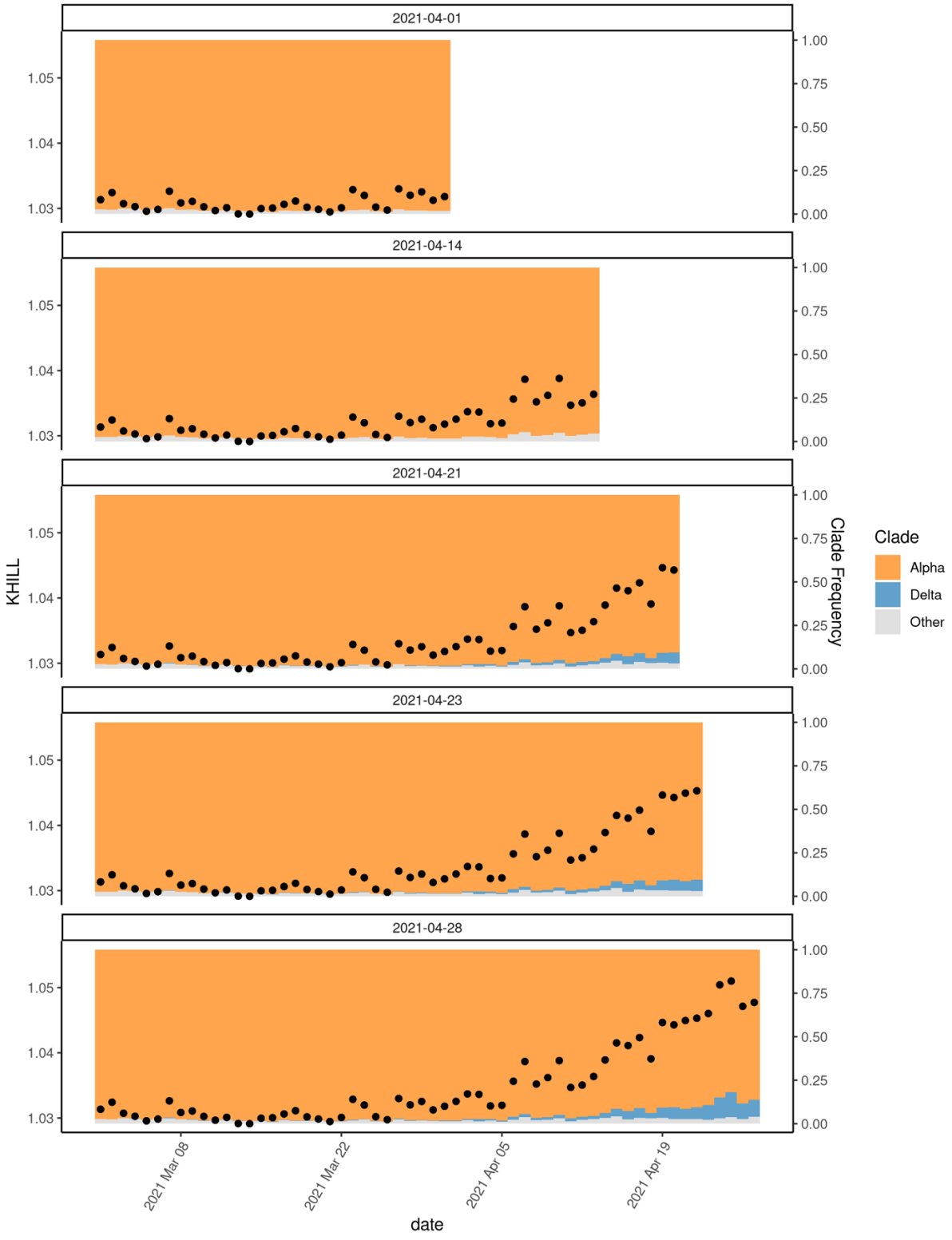
Supplemental Figure 2. The effect of k-mer size. We show the transition between Alpha and Delta in the UK pandemic sweeping across k-mer sizes from 1 to 31. For all but the smallest k-mer sizes, we observe the transition peak as occurring just after May 1st, 2021 with a Delta sweep complete by July 2021.



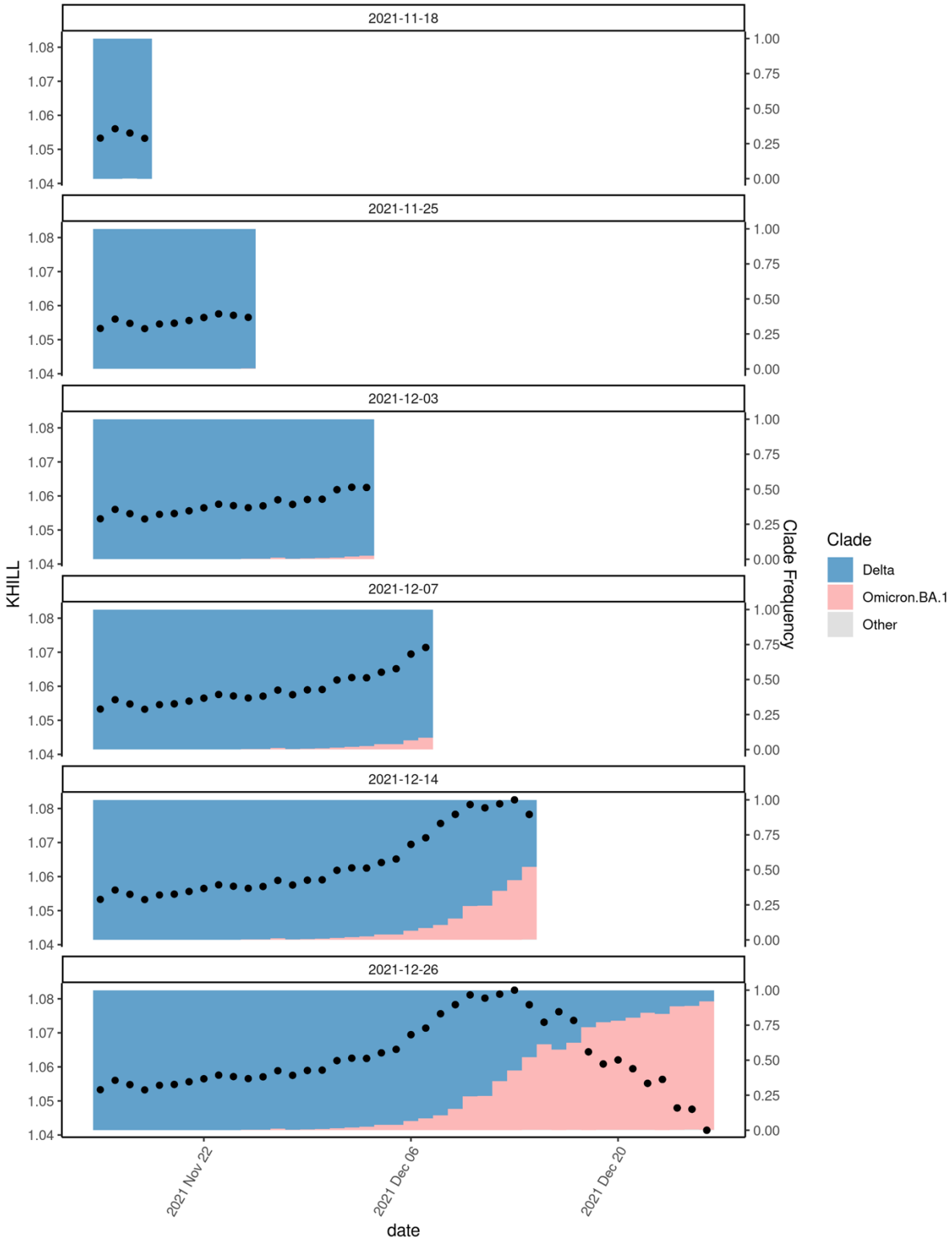
Supplemental Figure 3. The United States (US) COVID-19 pandemic. We show new case burden (Panel A) and gross phylogenetic classification (Panel B) for the US pandemic since its inception. The arrival of Alpha (March 2021) seems to have homogenized a previously variable US viral population. The Delta surge (June 2021) moderately increases complexity, while the arrival of Omicron (December 2021) is accompanied by a KHILL spike followed by a rapid descent after Delta is forced into near extinction (Panel C).



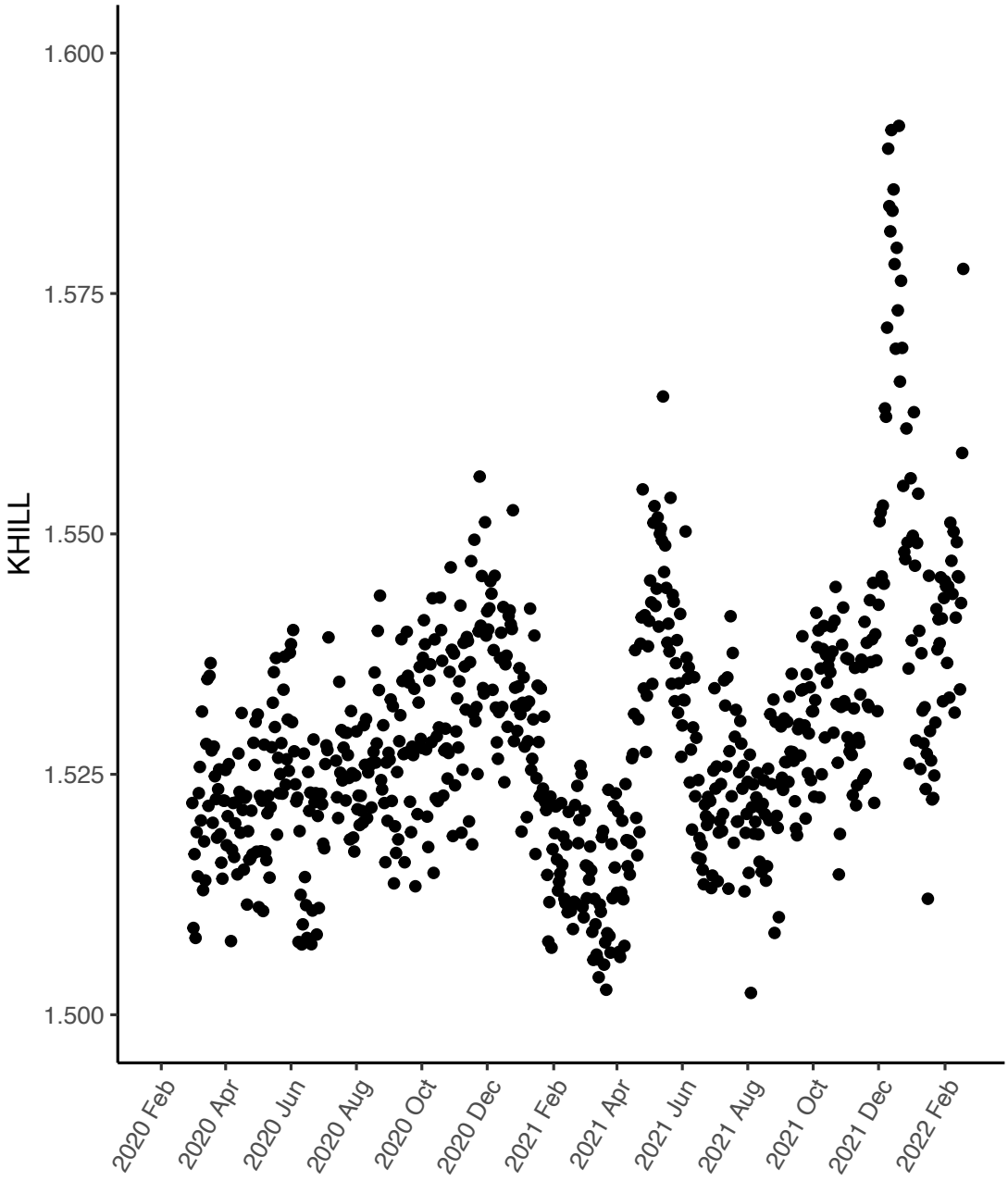
Supplemental Figure 4. The South Africa (SA) COVID-19 pandemic. We show new case burden (Panel A) and gross phylogenetic classification (Panel B) for the SA pandemic since its inception. The arrival of both Delta and Omicron lead to KHILL spikes. Because of sparser sequencing in this population, the SA data is considerably noisier than either the UK or USA. Note that we also include the Beta variant here as it was a significant variant in the SA pandemic.



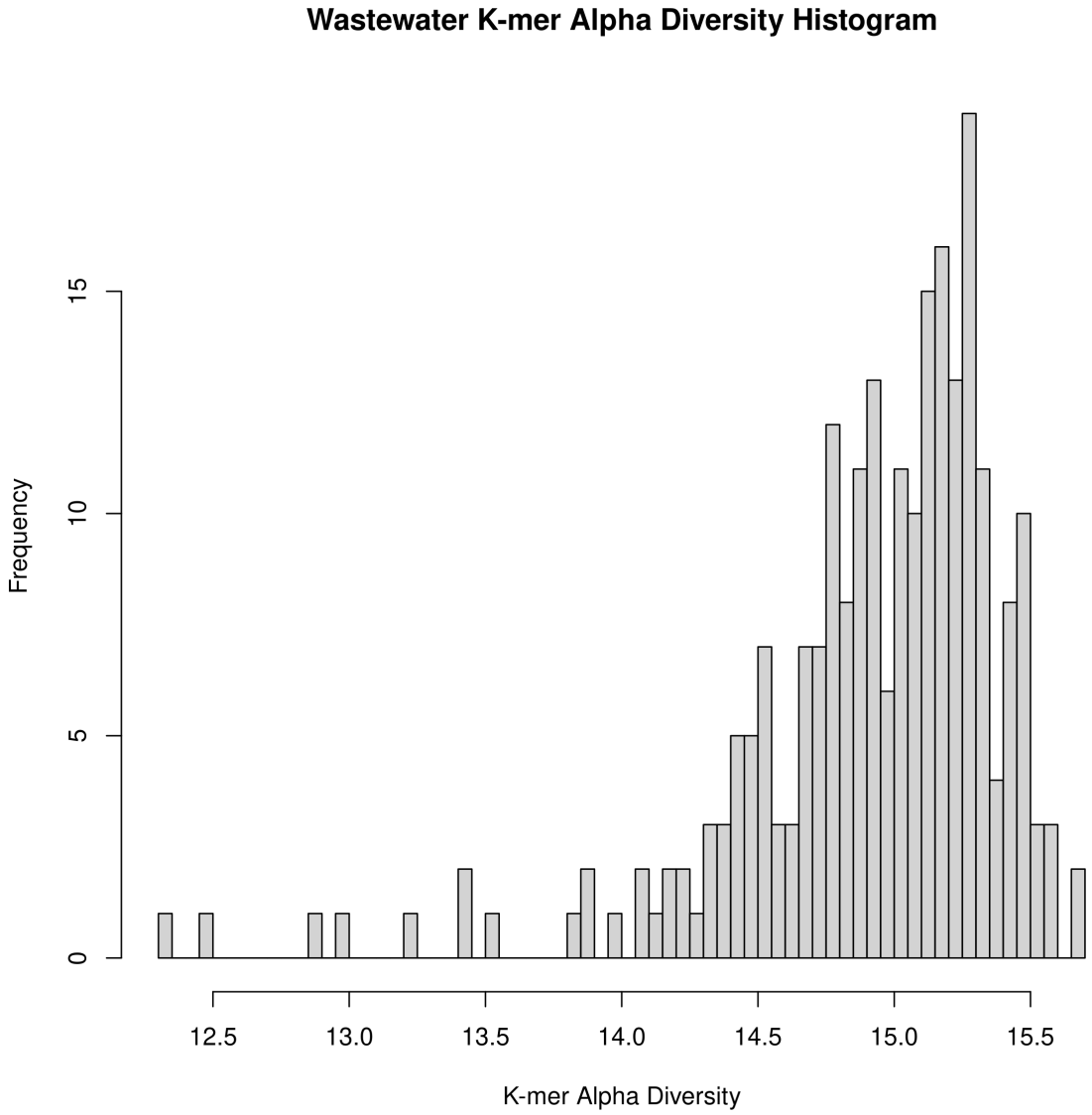
Supplemental Figure 5. The emergence of Delta. We detect the onset of Delta in the UK population in tandem with its Pangolin annotation as of 4/21/2021.



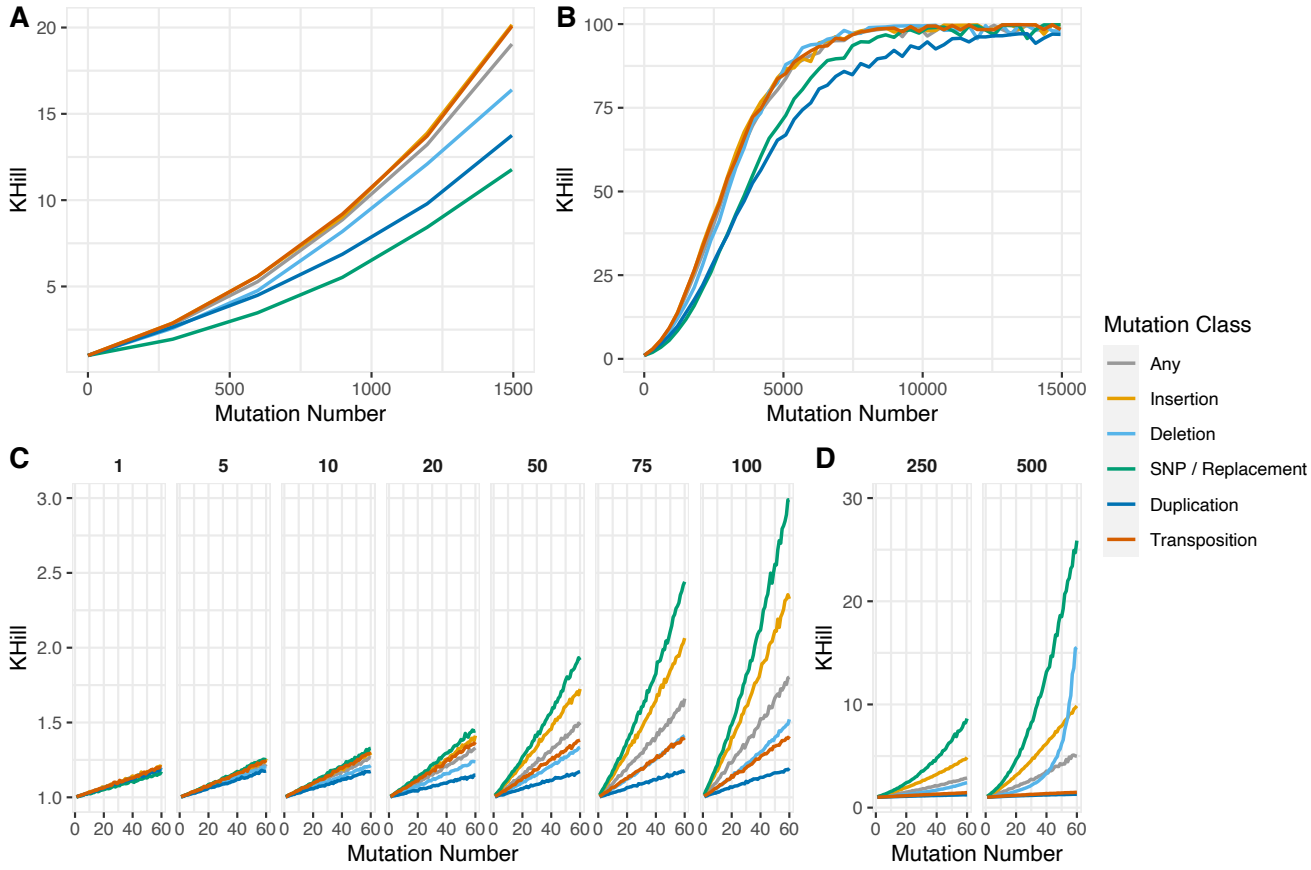
Supplemental Figure 6. The emergence of Omicron. We detect the onset of Omicron in the UK population in step with its Pangolin annotation as early as 12/3/2021.



Supplemental Figure 7. Simulated reads. We simulated 3 Mbases of Illumina sequence from each of 50 randomly selected genomes taken daily over the course of the UK pandemic. We show that the KHILL pandemic curve in this unassembled sequence mirrors that shown in Figure 2 for assembled genomes.

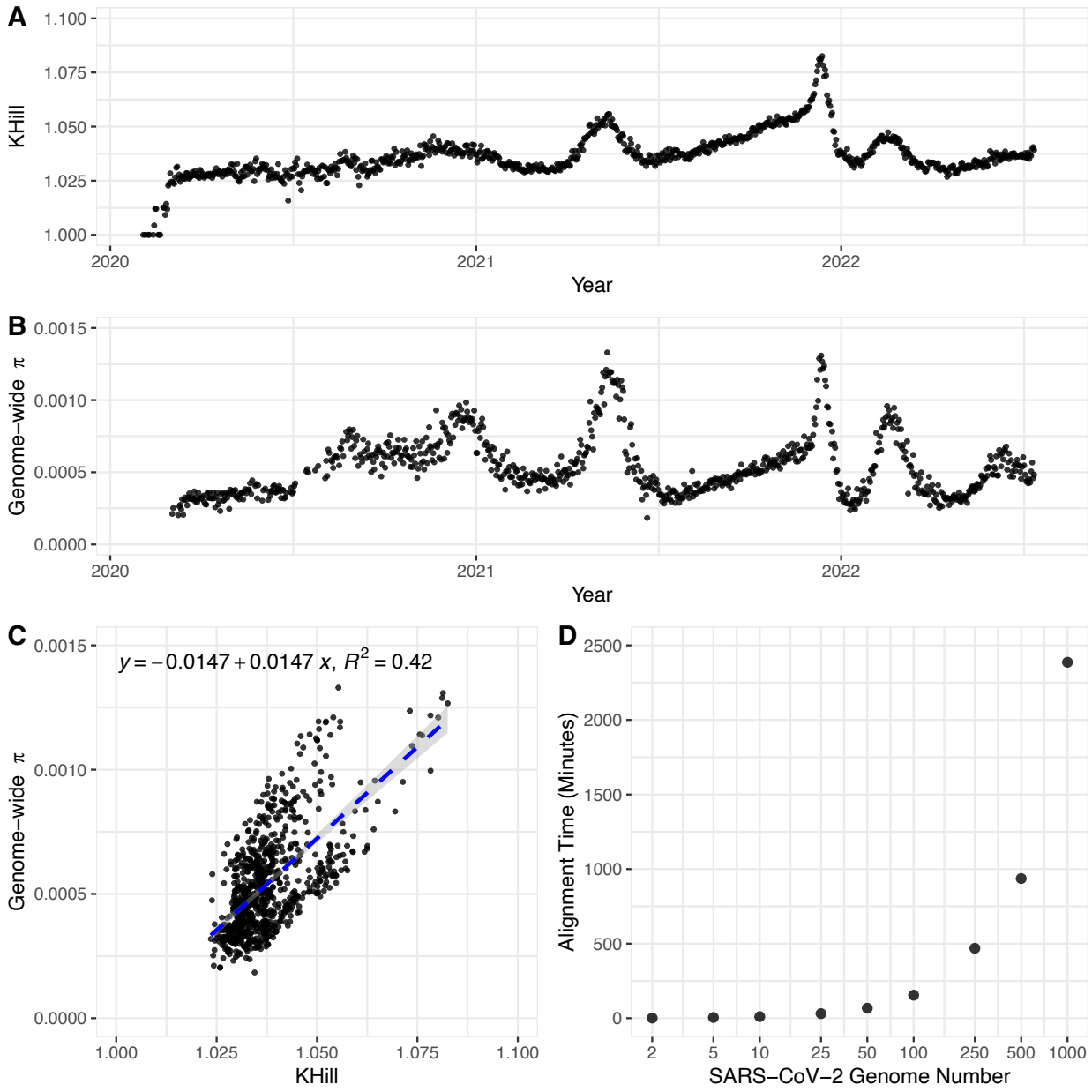


Supplemental Figure 8. Alpha diversity of wastewater k-mers. We plot the alpha diversity of sequence gathered from each day of the San Diego wastewater sample. Because the sample is amplicon based and enriched for SARS-CoV-2, higher alpha diversity indicates more unique viral k-mers sequenced at a higher depth, while lower diversity is likely characteristic of over-amplified, less even coverage. We remove these low complexity (and some high complexity) days by taking only those samples with sequence one standard deviation from the alpha diversity mean.

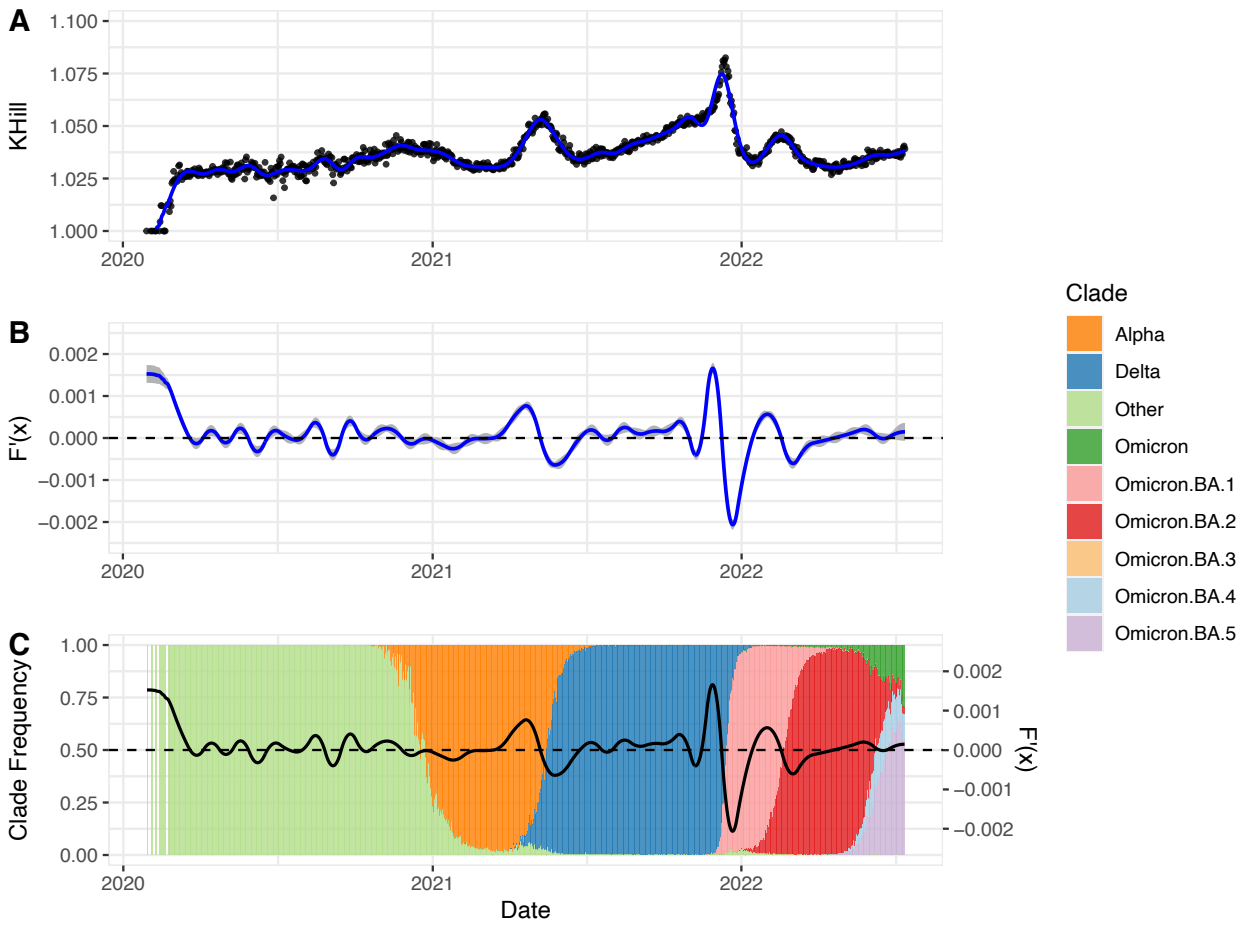


Supplemental Figure 9. Simulations. We simulated mutations in populations of 29.9Kb SARS-CoV-2 genomes using the EMBOSS tool *msbar* (24) and the SARS-CoV-2 reference (NCBI Reference Sequence: NC_045512.2) to understand how numbers of mutations and mutation classes impact the KHILL statistic. Each population differs in the number of mutations (x-axis), mutation class (legend), and the block size of each mutation (as below). Panels A and B report the KHILL statistic (y-axis) calculated from populations of 1,000 genomes with 1bp SNP mutations. Mutations were simulated to represent 1% - 50% of the genome in steps of 1% (i.e. 299 – 14,952 mutations). Panel A reports the KHILL statistic for zero to 1,500 mutations, while the maximum number of mutations in panel B is 14,952, or 50% of the genome. Panels C and D report KHILL statistics for populations of 100 genomes and are faceted and annotated (top) by

mutation block size (1, 5, 10, 20, 50, 75, and 100 bp in panel C, and 250 and 500 bp in panel D). We simulated mutation classes: insertion, deletion, SNP (or replacement for block mutations greater than 1 bp), duplication, transposition (copy/paste), or “any”, which is a random combination of all mutation forementioned classes.



Supplemental Figure 10. Comparison between KHILL and genome wide nucleotide diversity (π). We show that genome wide π of 100 subsampled genomes per day across the UK pandemic, closely approximates the more data heavy KHILL curve (Panels A and B). Genome wide π is well correlated with KHILL (Panel C), but its calculation is expensive for anything beyond 100 genomes per day because of the cost incurred by whole genome alignment (Panel D).



Supplemental Figure 11. Slopes. In Panel A we used a generalized additive model to fit a cubic spline (blue line) with a kurtosis of 50 to the UK covid pandemic KHill statistics (y-axis, as a response to date on x-axis). Panel B presents the first derivative F'(x), or slope, of the cubic spline presented in Panel A. Positive values indicated a positive slope (increasing KHill), while negative values indicated a negative slope (decreasing KHill). The grey shaded area represents the 95% confidence interval of F'(x). Panel C presents F'(x) (right y-axis) overlayed on the SARS-CoV-2 clade frequency (left y-axis) of the UK covid pandemic dataset. Alpha, Delta, and Omicron sub-variants have been collapsed (legend). In Panel B and C the horizontal dashed line at F'(x) = 0 indicates the transition point from positive to negative slopes of the fitted cubic spline.