

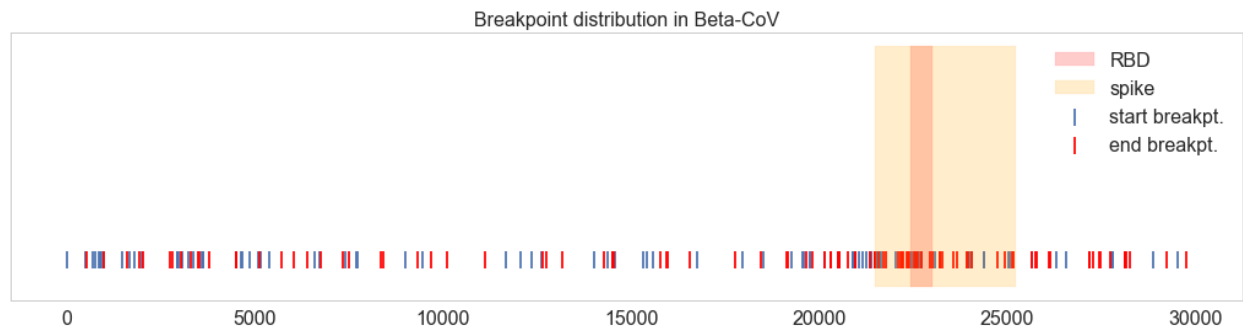


Fig S1a | Recombination breakpoints in betacoronaviruses. Start (blue) and end (red) breakpoints from all of the 103 detected recombination events in betacoronaviruses are shown, highlighting the Spike protein and the RBD.

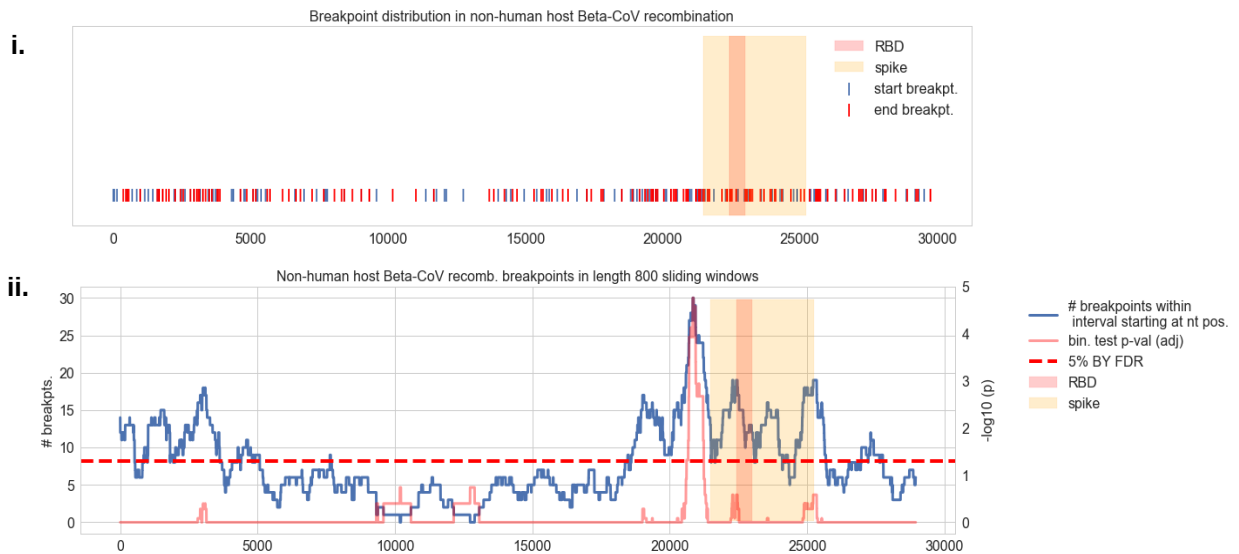


Fig S1b | Recombination breakpoints in betacoronaviruses (non-human host only). i.

Distribution of start and end breakpoints recombination events detected in betacoronaviruses infecting non-human hosts. The Spike protein and the RBD are highlighted. **ii.** Sliding window analysis shows the distribution of recombination breakpoints (either start or end) in 800 nucleotide (nt) length windows upstream (namely, in the 5' to 3' direction) of every nt position along the viral genome (blue curve). BY-adjusted p -values are shown in red.

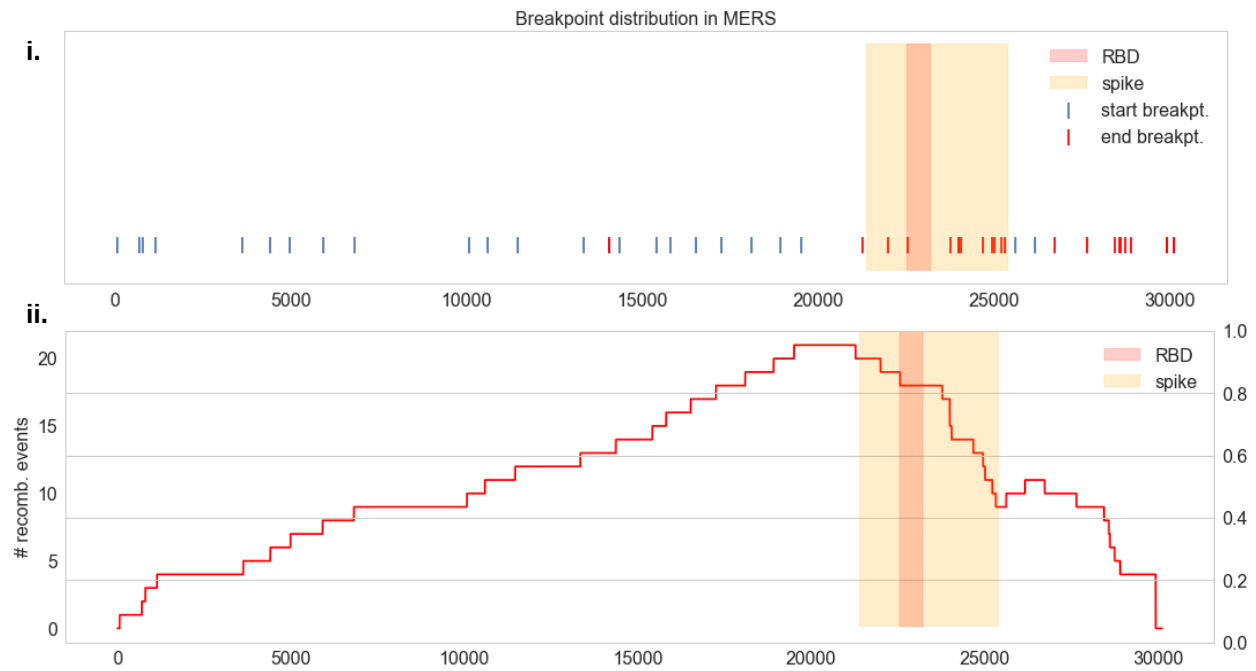


Fig S1c | Recombination breakpoints in MERS-CoV. **i.** Distribution of start and end breakpoints from 24 recombination events detected from MERS-CoV isolates. **ii.** Number of recombination events covering each nt position of the MERS-CoV genome. The Spike gene, and the n-terminus in particular, overlaps with the majority of the detected recombination segments.

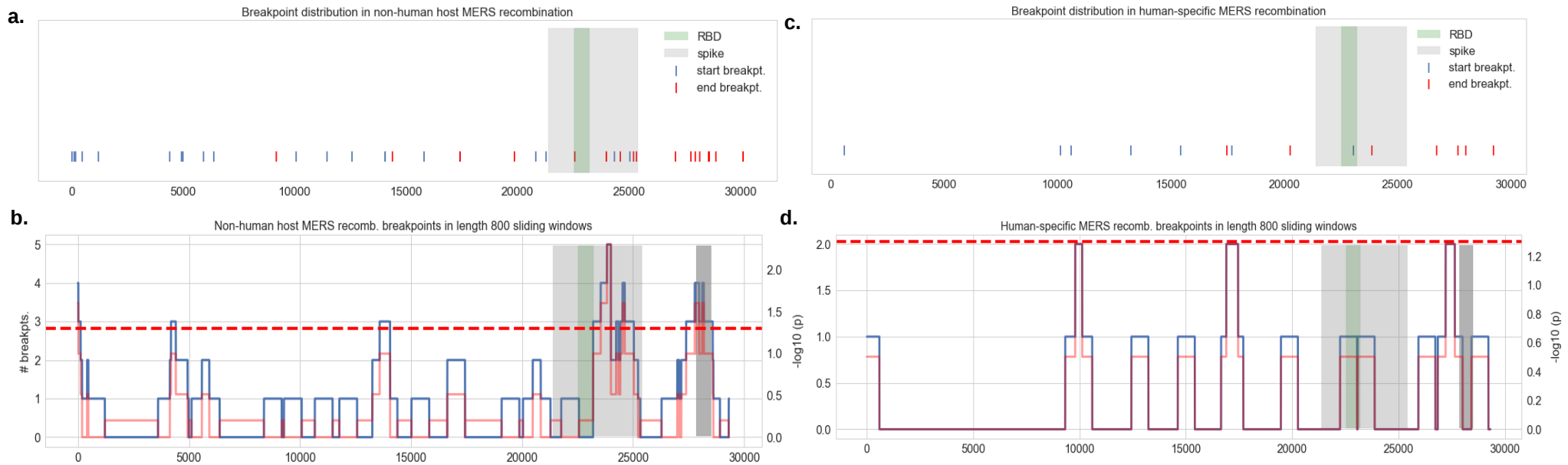


Fig S2 | Host-specific recombination breakpoints in MERS-CoV. **a, c.** Distribution of start and end recombination breakpoints in MERS-CoV isolates from non-human hosts (a) and human hosts (c). **b, d.** Sliding window analysis: 800 nt length windows for non-human (b) and human-specific recombination events (d); recombination breakpoint count in each interval shown in blue; binomial test nominal p-values in red. Spike protein (light gray), RBD of Spike (green) and membrane protein (dark gray) are highlighted.

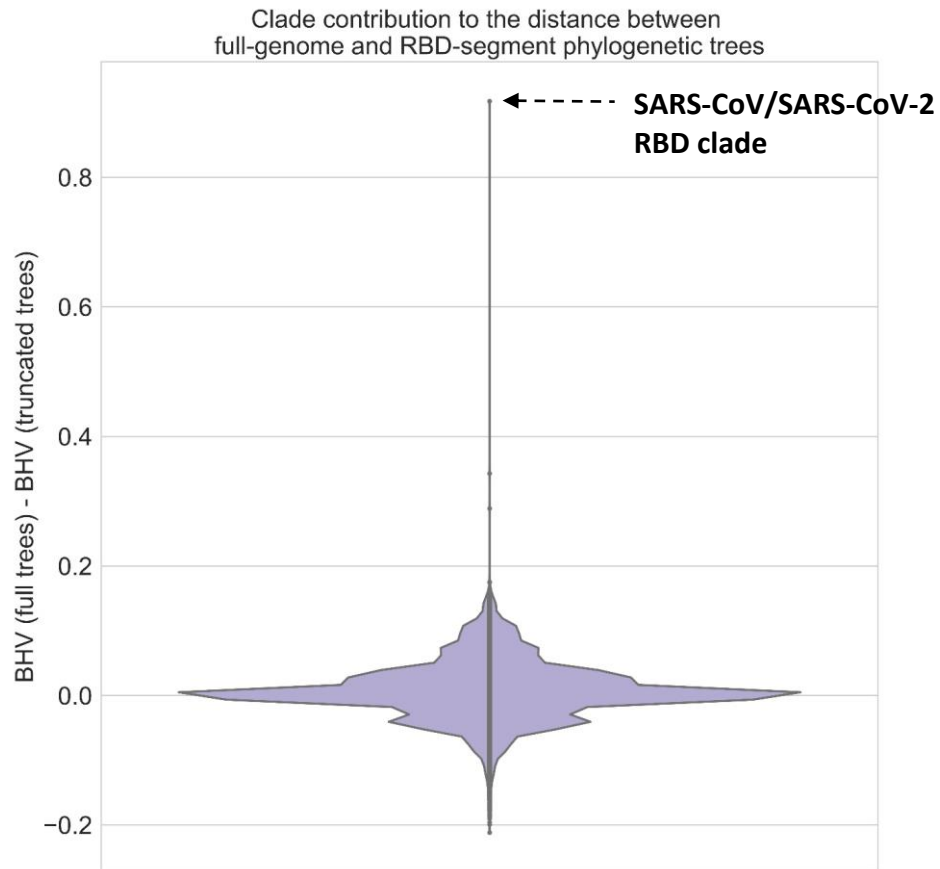
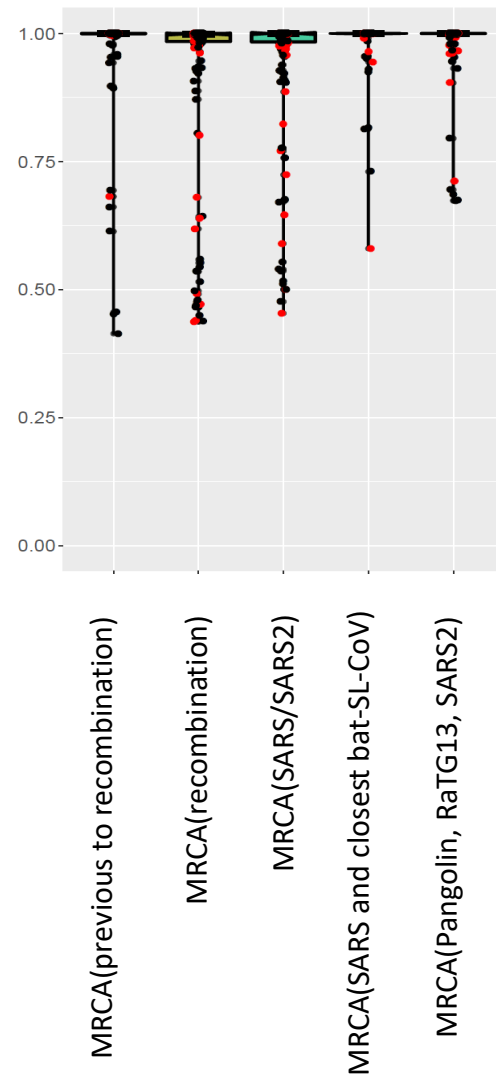


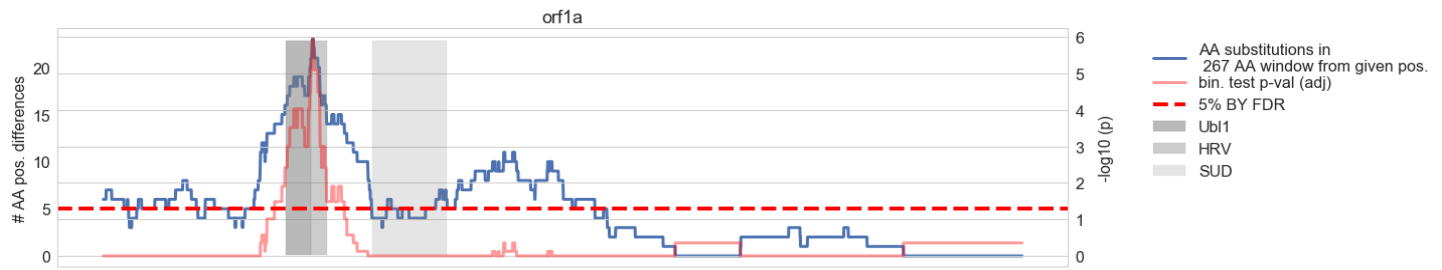
Fig S3 | SARS-CoV and SARS-CoV-2 lineages contribute significantly to genome vs RBD tree differences.

Violin plot showing the difference of BHV distances between the full-genome tree and the RBD-specific tree on the one hand, and the corresponding truncated trees on the other hand. Differences for 23780 random sets of sequences and clades are shown. The difference of BHV distances quantifies the contribution of the SARS-CoV/SARS-CoV-2 clade towards the overall divergence between the whole-genome and the RBD evolutionary histories, and this randomization test shows the probability that a randomly selected set of taxa contributes more towards the phylogenetic divergence of the full-genome and the RBD-specific trees than the SARS-CoV/SARS-CoV-2 clade. The SARS-CoV/SARS-CoV-2 clade ranks the highest among all the clades and random sets of taxa analyzed.

Fig S4 | Distribution of Likelihood values associated with the most likely amino acid inferred for the different MRCAs in the RBD ancestral reconstruction analyses. Red dots represent positions highlighted in our work (Spike 333, 359, 427, 436, 432-435, 460-472, 484, 505).



i.



ii.

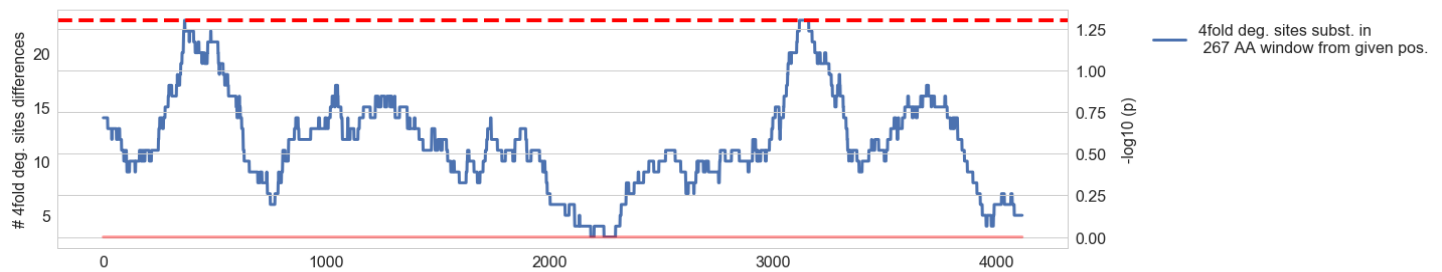


Fig S5a | Divergence between SARS-CoV-2 and the RaTG13 bat virus in orf1a. i. Sliding window analysis of 267 aa length windows showing non-synonymous substitutions where the two strains differ (blue). BY-adjusted p -values for each window shown in red. Several domains of non-structural protein three are highlighted in gray: ubiquitin-like domain 1 (Ubl1), hypervariable region (HRV) and SARS-unique domain (SUD). ii. Sliding window analysis (267 aa length windows) containing 4-fold degenerate sites that differ between SARS-CoV-2 and RaTG13. No windows are statistically significant after BY p -value adjustment (red line).

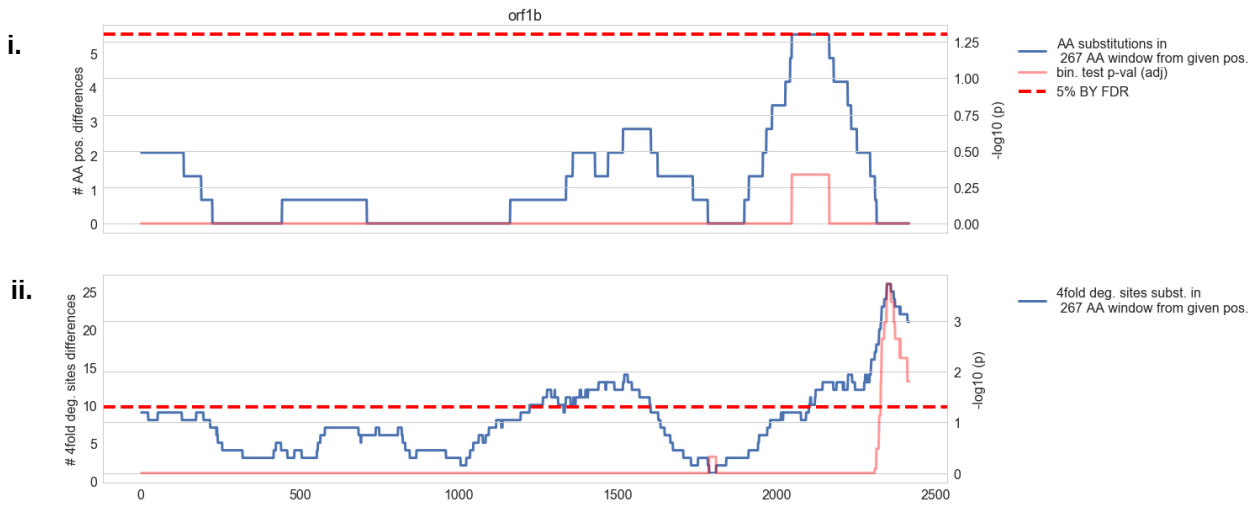


Fig S5b | Divergence between SARS-CoV-2 and the RaTG13 bat virus in orf1b. **i.** Sliding window analysis of 267 aa length windows showing non-synonymous substitutions where the two strains differ (blue). BY-adjusted p -values for each window shown in red. **ii.** Sliding window analysis (267 aa length windows) containing 4-fold degenerate sites that differ between SARS-CoV-2 and RaTG13.

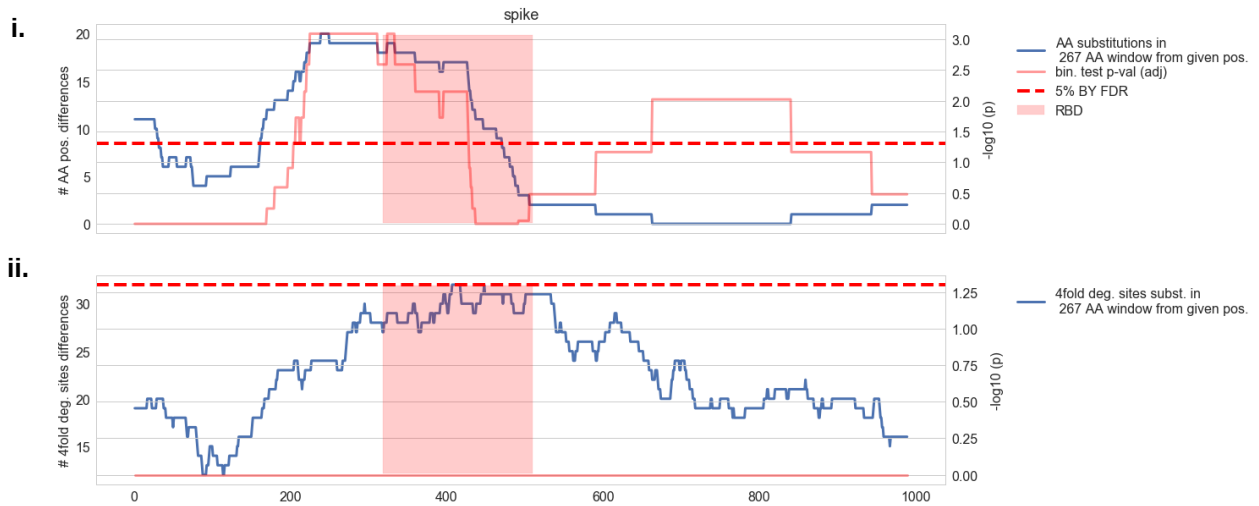


Fig S5c | Divergence between SARS-CoV-2 and the RaTG13 bat virus in Spike. **i.** Sliding window analysis of 267 aa length windows showing non-synonymous substitutions where the two strains differ (blue). BY-adjusted p -values for each window shown in red. RBD is highlighted. **ii.** Sliding window analysis (267 aa length windows) containing 4-fold degenerate sites that differ between SARS-CoV-2 and RaTG13.

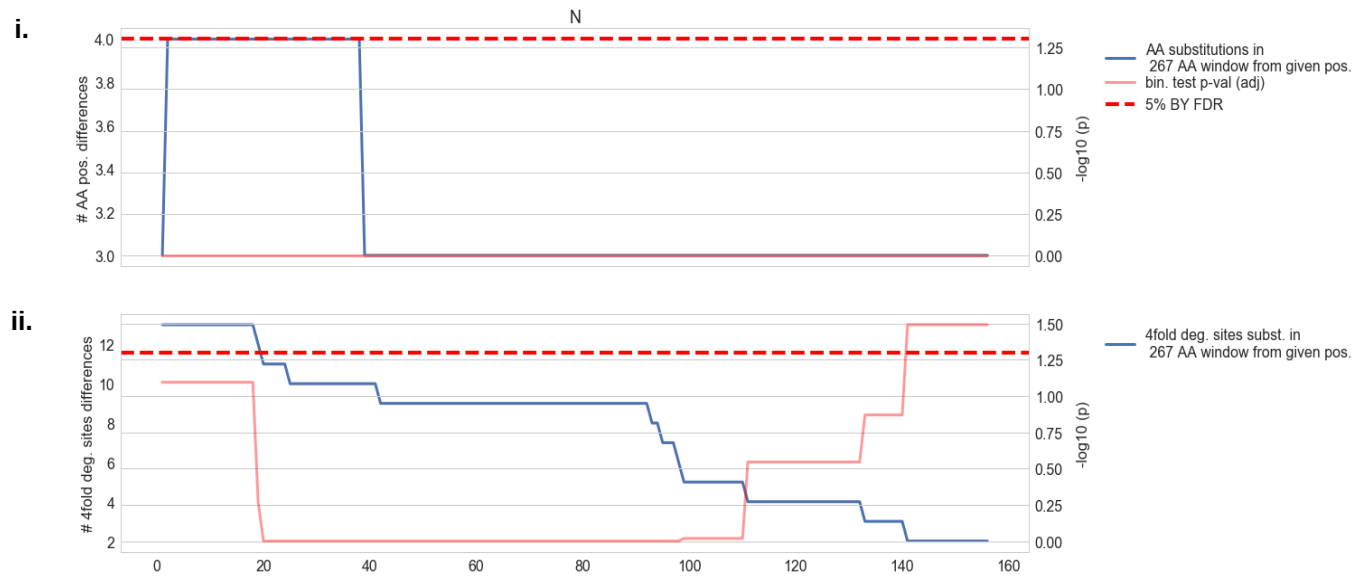


Fig S5d | Divergence between SARS-CoV-2 and the RaTG13 bat virus in the Nucleocapsid.

i. Sliding window analysis of 267 aa length windows showing non-synonymous substitutions where the two strains differ (blue). BY-adjusted p -values for each window shown in red. **ii.** Sliding window analysis (267 aa length windows) containing 4-fold degenerate sites that differ between SARS-CoV-2 and RaTG13.

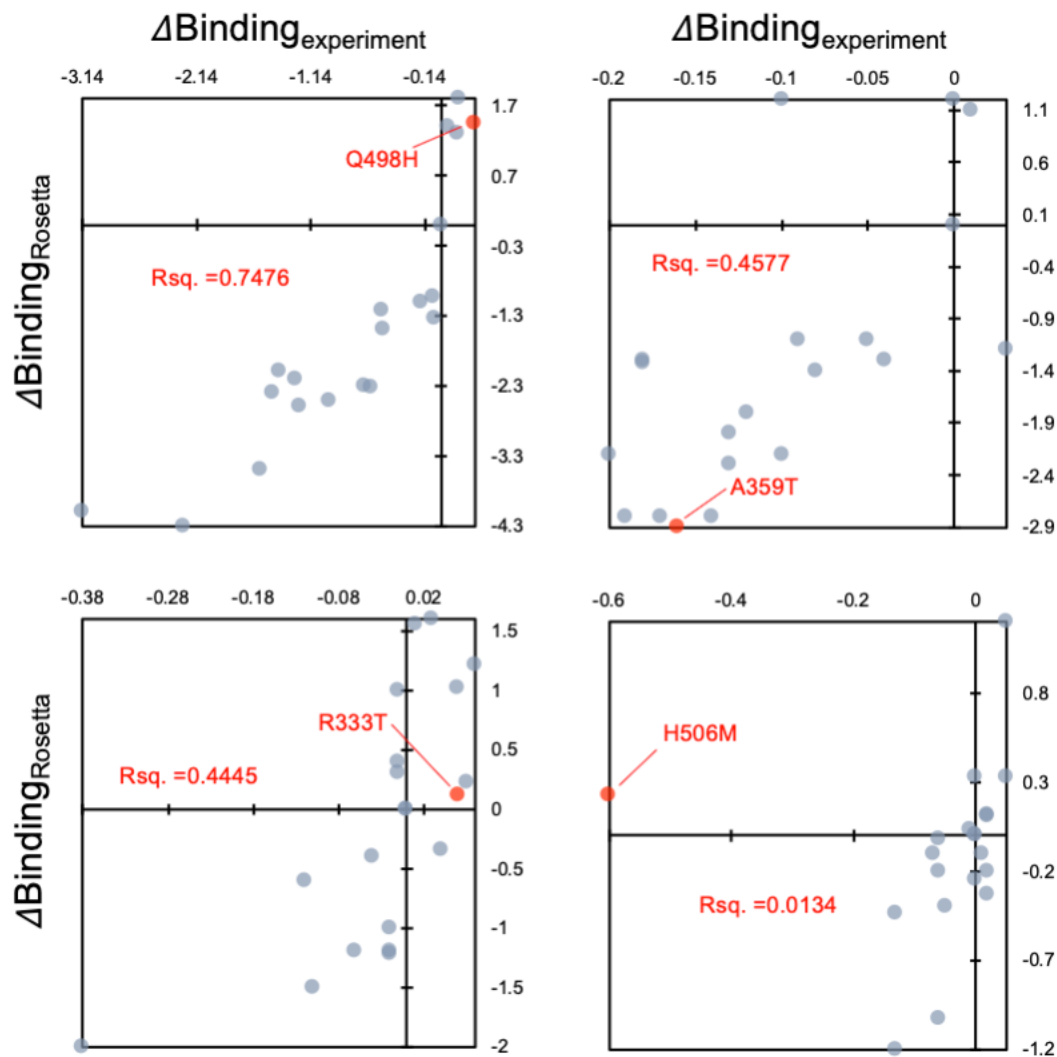
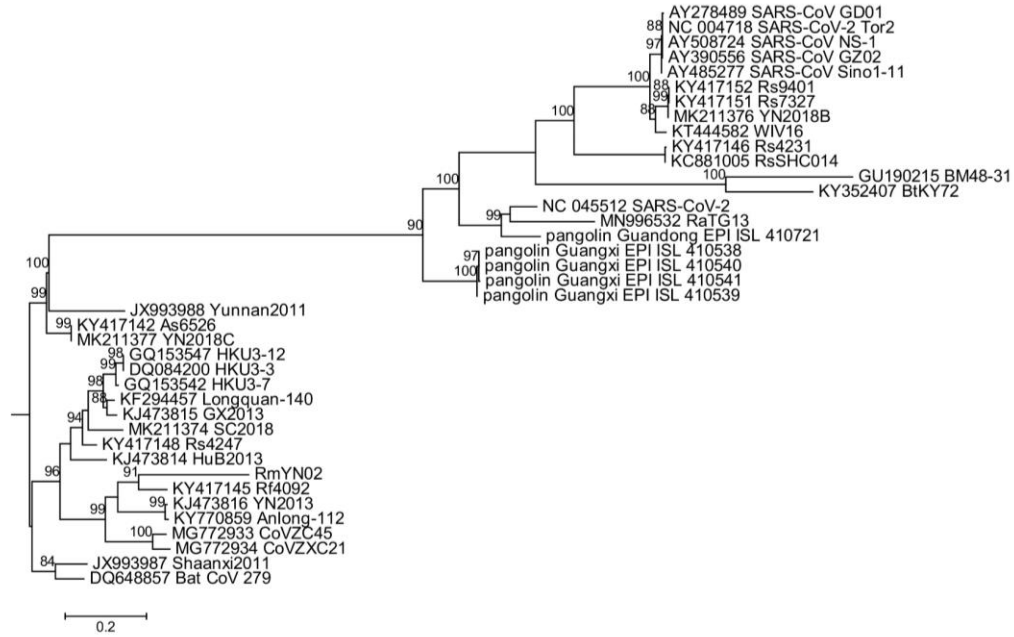


Fig S6. Analysis of ten independent Rosetta trajectories per mutant to report the median binding scores with hACE2 after imposing all nineteen amino acid point mutations at these four loci on the viral RBD structure. Each grey dot represents mutation to one of the 19 other amino acids for each position. The Y and X axes show change (Δ) in hACE2 binding score from the default amino acid type at the backbone loci. The dot at origin (0,0), thus, represents the unmutated state (i.e. Q484H etc.). The red dot represents the mutations we reported in the previous study. We see that the R-squared correlation values are better in $\frac{3}{4}$ cases – except H505M. Each dot represents the binding score averaged over 10 independent simulation trajectories performed using Rosetta. Each simulation trajectory involves (a) making the mutation, (b) minimizing the complex between mutated RBD and hACE2, and (c) interaction energy calculation.

a)



b)



Fig S7. Maximum Likelihood trees of Sarbecoviruses derived from the RBD recombination segment (a) and the rest of the genome (b). Numeric values of Shimodaira-Hasegawa-like supports higher than 0.80 are shown.