Cite as: F. Obermeyer *et al.*,
Science 10.1126/science.abm1208 (2022).

Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness

Fritz Obermeyer[1,2]*†, Martin Jankowiak[1,2], Nikolaos Barkas[1], Stephen F. Schaffner[1,3,4], Jesse D. Pyle[1,5], Leonid Yurkovetskiy[6], Matteo Bosso[6], Daniel J. Park[1], Mehrtash Babadi[1], Bronwyn L. MacInnis[1,4,7], Jeremy Luban[1,6,7,8], Pardis C. Sabeti[1,3,4,7,9]‡, Jacob E. Lemieux[1,10]*‡

[1]Broad Institute of MIT and Harvard; 415 Main Street, Cambridge, MA 02142, USA. [2]Pyro Committee, Linux AI & Data Foundation; 548 Market St San Francisco, California 94104. [3]Department of Organismic and Evolutionary Biology, Harvard University; Cambridge, MA 02138, USA. [4]Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Harvard University; Boston, MA, USA. [5]Program in Molecular Medicine, University of Massachusetts Medical School; Worcester, MA 01605, USA. [6]Research Department, Molecular Enzymology Division, New England Biolabs, Ipswich, MA 01938, USA. [7]Massachusetts Consortium on Pathogen Readiness; Boston, MA 02115, USA. [8]Ragon Institute of MGH, MIT, and Harvard; 400 Technology Square, Cambridge, MA 02139, USA. [9]Howard Hughes Medical Institute; 4000 Jones Bridge Rd, Chevy Chase, MD 20815, USA. [10]Division of Infectious Diseases, Massachusetts General Hospital; Boston, MA, USA.

*Corresponding author. lemieux@broadinstitute.org (J.E.L.); fritz.obermeyer@gmail.com (F.O.)

†Present address: Generate Biomedicines, Inc.26 Landsdowne Street, Cambridge, MA 02139, USA.

‡These authors contributed equally to this work.

Repeated emergence of SARS-CoV-2 variants with increased fitness underscores the value of rapid detection and characterization of new lineages. We have developed PyR$_0$, a hierarchical Bayesian multinomial logistic regression model that infers relative prevalence of all viral lineages across geographic regions, detects lineages increasing in prevalence, and identifies mutations relevant to fitness. Applying PyR$_0$ to all publicly available SARS-CoV-2 genomes, we identify numerous substitutions that increase fitness, including previously identified spike mutations and many non-spike mutations within the nucleocapsid and nonstructural proteins. PyR$_0$ forecasts growth of new lineages from their mutational profile, ranks the fitness of lineages as new sequences become available, and prioritizes mutations of biological and public health concern for functional characterization.

The SARS-CoV-2 pandemic has been characterized by repeated waves of cases driven by the emergence of new lineages with higher fitness, where fitness encompasses any trait that affects the lineage's growth, including its basic reproduction number (R$_0$), ability to evade existing immunity, and generation time. Rapidly identifying such lineages as they emerge, and accurately forecasting their dynamics, is critical for guiding outbreak response. Doing so effectively would benefit from the ability to interrogate the entirety of the global SARS-CoV-2 genomic dataset. The large size (currently over 7.5 million virus genomes) and geographic and temporal variability of the available data present significant challenges that will become greater as more viruses are sequenced. Current phylogenetic approaches are computationally inefficient on datasets with more than ~5000 samples and take days to run at that scale. Ad hoc methods to estimate the relative fitness of particular SARS-CoV-2 lineages are a computationally efficient alternative (*1*–*3*), but have typically relied on models in which one or two lineages of interest are compared to all others and do not capture the complex dynamics of multiple co-circulating lineages.

Furthermore, estimates of relative fitness based on lineage frequency data alone (*2*, *4*, *5*) do not take advantage of additional statistical power that can be gained from analyzing the independent appearance and growth of the same mutation in multiple lineages. Performing a mutation-based analysis of lineage prevalence has the additional advantage of identifying specific genetic determinants of a lineage's phenotype, which is critically important both for understanding the biology of transmission and pathogenesis and for predicting the phenotype of new lineages. The SARS-CoV-2 pandemic has already been dominated by several genetic changes of functional and epidemiological importance, including the spike (S) D614G mutation that is associated with higher SARS-CoV-2 loads (*6*, *7*). Mutations found in Variants of Concern (VoC), such as S:N439R, S:N501Y, and S:E484K, have been linked, respectively, to increased transmissibility (*8*), enhanced binding to ACE2 (*9*), and antibody escape (*10*, *11*). Despite these successes, identifying functionally important mutations in the context of a large background of genetic variants of little or no phenotypic consequence remains challenging.

In modeling the relative fitness of SARS-CoV-2 lineages, we estimated their growth as a linear combination of the effects of individual mutations. To this end, we developed PyR$_0$, a hierarchical

Bayesian regression model that enables scalable analysis of the complete set of publicly available SARS-CoV-2 genomes, that can be applied to any viral genomic dataset and to other viral phenotypes. The model, which is summarized in fig. S1, and described in detail in the supplementary materials, avoids the complexity of full phylogenetic inference by first clustering genomes by genetic similarity (refining PANGO lineages (*12*)), and estimating the incremental effect on growth rate of each of the most common amino acid changes on the lineages in which they appear. By regressing growth rate on genome sequence, the model shares statistical strength among genetically similar lineages without explicitly relying on phylogeny. By modeling only the multinomial proportion of different lineages rather than the absolute number of samples for each lineage (*13*, *14*), and by doing so within 14-day intervals in 1,560 globally-distributed geographic regions, the model achieves robustness to a number of sources of bias that affect all lineages, across regions and over time, including differences in data collection and changes in transmission due to such factors as social behavior, public health policy, and vaccination.

We fit PyR$_0$ to 6,466,300 SARS-CoV-2 genomes available on GISAID (*15*) as of January 20, 2022, in a model that contained 3,000 clusters, derived from 1,544 PANGO lineages, and 2,904 nonsynonymous mutations. The output of the model is a posterior distribution for the relative fitness (exponential growth rate) of each lineage and for the contribution to the fitness from each mutation. Fitting this large model is computationally challenging, so we used stochastic variational inference, an approximate inference method that reduced our task to solving a 75-million-dimensional optimization problem on a GPU. Inference was implemented in the Pyro (*16*) probabilistic programming framework (see Supplemental Materials). The trained model can be used to infer lineage fitness, predict the fitness of completely new lineages, forecast future lineage proportions, and estimate the effects of individual mutations on fitness.

The model's lineage fitness estimates (Fig. 1B) show a modest upward trend over time among all lineages, interrupted by several lineages with much higher fitness. Sensitivity analyses revealed qualitative consistency of fitness estimates across spatial data subsets (fig. S2). The upward trend may in part reflect an upward bias caused by the lineage assignment process, as can be seen in simulation studies (fig. S3), but the high tail of the distribution exhibits elevated fitness values far in excess of this trend. The spread of the virus into human populations in late 2019 and early 2022 has been marked by periods of rapid evolution in fitness and waves of increase in case counts (Fig. 1). While PANGO lineages facilitate communication by providing a stable nomenclature, we observed some PANGO lineages with multiple successive peaks in some regions, suggesting that sublineages within them had differing fitnesses. We therefore algorithmically refined the 1,544 PANGO lineages into 3,000 finer clusters, and found that our model identified significant heterogeneity within some PANGO lineages (fig. S4). When we tested the model's predictive ability (fig. S5), we found that forecasts were reliable for 1-2 months into the future for variants of concern, but not necessarily other variants, when they tended to be disrupted by the emergence of a completely new strain (table S1, fig. S6). The accuracy of forecasts stabilized typically stabilized within two weeks after the emergence of a new competitive lineage in a region (fig. S6).

The model correctly infers WHO classification variant Omicron (PANGO BA.2) to have the highest fitness to date, 8.9-fold (95% CI, 8.6-9.2) higher than the original A lineage (Fig. 1 inset), accurately foreshadowing its rise in regions where it is circulating (fig. S7). Through systematic backtesting, we found that the model would have provided early warning and aided in the identification of VoCs had it been routinely applied to SARS-CoV-2 samples, confirming the importance for public health of timely publication of genomic data. For example, the elevated fitness of BA.2 was identified by mid-December 2021 on the basis of 76 reported sequences (fig. S8); sharing statistical strength over mutations enabled an earlier and more confident prediction that BA.2 was the fittest lineage yet observed (fig. S10). Likewise, PyR$_0$ would have forecast the dominance of B.1.1.7 in late November 2020 (fig. S9), AY.4 by May 2021 (fig. S10), and BA.1 by early December 2021 (fig. S8). While variant-specific models were accurate and useful in predicting the rise of these lineages (*2*), each modeling effort was specific to a particular lineage and geographic region. PyR$_0$'s global approach provides similar early detection while also offering automated, rapid, and standardized unbiased consideration of all variants and lineages, together with ranking based on relative fitness.

Compared to standard multinomial regression models, PyR$_0$ estimates of lineage fitness were similar (Pearson's R = 0.95, S11-S12), but including mutations in the model enables PyR$_0$ to infer elevated fitness of Omicron lineages BA.1 and BA.2 faster than the model without mutations (fig. S14). In contrast to non-hierarchical binomial logistic regression (fig. S13), PyR$_0$ estimates displayed less variability as data accumulated, benefitting from the sharing of information across regions and the regularizing effect of the priors. Lineage fitness estimates were also stable between our initial analysis of 2.1 million genomes in August 2021 (*17*), shortly after the emergence of Delta lineages, and before the emergence of Omicron (Spearman's rho = 0.78, fig. S15C). The correlation between individual amino acids in the two models was weaker than that for lineages (fig. S15D-E, rho = 0.48) but still significant (test of no association for rho, $p < 2 \times 10^{-16}$), reflecting both the inherent difficulty of estimating high-dimensional mutational coefficients observed indirectly through lineage counts (Supplementary Note 1), as well as the addition of 4.3 million sequences, including highly fit Omicron lineages distinguished by their enhanced immune escape.

By jointly modeling fitness estimates using lineage counts and individual mutations, PyR$_0$ harnesses convergent evolution (Table 1 and fig. S16) to infer the fitness of new constellations of mutations based on the trajectories of other lineages in which they have previously emerged. This predictive capability has the potential to aid public health efforts because the model has the potential to learn faster by incorporating mutations than it would by relying on lineage counts alone (fig. S14). To test the reliability of this kind of estimate, we fit leave-one-out estimators for PANGO lineages on subsets of the dataset with that entire lineage removed, based solely on the mutational content of the omitted lineage (fig. S17). These estimators showed excellent agreement with estimators based on the observed behavior of the lineages, and they were also more accurate than naive phylogenetic estimators that assume the fitness of each new strain is equal to its parent lineage's fitness (Pearson's R = 0.983, after correcting for parent fitness, fig. S17). Together, these analyses suggest that PyR_0 has the potential to aid genomic surveillance efforts by providing an automated early warning system on a similar time scale as sophisticated regional surveillance efforts (*18*, *19*).

Genome-wide estimates of the effect of SARS-CoV-2 mutations on fitness also provide a powerful tool for better understanding the biology of fitness. Our model allowed us to estimate the contribution of 2,904 amino acid substitutions (Fig. 2A and Table 1) to lineage fitness and to rank them by inferred statistical significance (fig. S18). Cross-validation confirmed that these results replicate qualitatively across different geographic regions (fig. S19).

The highest concentrations of fitness-associated mutations were found in the S, N, and the ORF1 polyprotein genes (ORF1a and ORF1b, Fig. 2, A and B, and figs. S20 and S21). Using spatial autocorrelation as a measure of spatial structure, we found evidence of functional hotspots in the S, N, ORF7a, ORF3a, and ORF1a genes (table S2). Within S, we confirmed three hotspots of fitness-enhancing mutations, each within a defined functional region: the N-terminal domain, the receptor-binding domain (RBD), and the furin-cleavage site (Fig. 2B). We assessed mutational enrichment in the top-ranked set of mutations and identified an enrichment for lysine to asparagine mutations in the S gene (fig. S22C). We visualized top scoring mutations within atomic structures for the spike protein (Fig. 2, D to E), the nucleocapsid's N-terminal domain (Fig. 2F), the polymerase (fig. S23), and two proteases (fig. S24). Many of the top mutations in the S gene occurred in the receptor binding domain (RBD) making direct contacts with the ACE2 receptor, including K417N/T and E484K (Fig. 2, D to E). Two top-ranked mutations, T478K and S477N, occur in a flexible loop adjacent to the S-ACE2 interface (Fig. 3E), suggesting that these mutations may affect the kinetics of receptor engagement or the Spike conformational changes that follow. Other mutations occurred in regions proximal to essential enzymatic active sites of the viral replication (fig. S15) or protein processing (fig. S16) machinery.

We tested several of the high-scoring mutations in single-cycle infectivity assays as done previously (*7*), focusing on the RBD (Fig. 3A). We found that while some individual mutations increased infectivity, on average, high-scoring RBD mutations did not promote infectivity per se. We considered an alternate possibility that fitness of Spike mutations is driven by immune escape. Using RBD-aggregated mutations as a proxy for immune escape, we found that the fitness effect of these Spike mutations correlates well with antibody escape estimates from Greaney et al. (*20*) (Fig. 3B). Together with the observed jump in fitness beginning in late 2021 (Fig. 3C) associated with Spike mutations, but not mutations elsewhere in the genome (Fig. 3E), these results suggest that immune escape is the dominant driver of current fitness increases. BA.1 and BA.2 had similar estimated fitness from Spike mutations, potentially consistent with similar Spike antibody neutralization of these variants (*21*), whereas PyR_0 inferred that the elevated fitness of BA.2 is attributed to non-Spike mutations (fig. S25). In contrast to mutations in Spike, those in the serine-arginine rich region of N were linked to increased efficiency of SARS-CoV-2 genomic RNA packaging (*22*). Within ORF1, we found fitness-associated mutations across all viral enzymes, and clusters within additional non-structural proteins (nsps). The highest concentration of fitness-associated mutations is found in nsp4, nsp6, and nsp12–14 (fig. S12B,S13C-D), suggesting unexplored function at those sites. For example, nsp4 and nsp6 have roles in assembly of replication compartments, and substitutions in these regions may influence the kinetics of replication (see Supplemental Note 3). We caution that while convergent evolution makes it possible to identify candidate functional mutations, observational data alone is insufficient to declare mutations as causal rather than merely correlated. Our uncertainty-ranked list of important mutations can be used to prioritize hits identified by our study for functional follow-up.

Some lineages increased in fitness more than others over the course of the pandemic (fig. S4). Notably, B.1.1 displayed the greatest variability among sublineages, followed by B.1. Fitness appeared to reach a plateau over time for most lineages (Fig. 1 and fig. S4). In contrast to Omicron sublineages, Alpha and Delta showed little variability in Spike-attributable fitness (fig. S25), suggesting that the propensity to acquire new Spike mutations depends on the constellation of mutations that comprise a lineage, consistent with epistasis. A limitation of PyR_0 is that it does not incorporate epistatic interactions between mutations (Supplemental Note 1); however, our results demonstrate the feasibility of inferring genetic determinants and lineage fitness using the simplest possible linear-additive model and provide a foundation for future research for more complex modeling that includes epistatic effects between mutations and migration across geographic regions.

In summary, PyR_0 provides a genome-wide, automated approach for detecting viral lineages with increased fitness. By combining a model-based assessment of lineage fitness with absolute case counts, our model provides a global picture of the events of the first two years of the pandemic. Because it assesses the contribution of individual mutations and aggregates across all lineages and geographic regions, it can identify mutations and gene regions that likely increase fitness, and mutation-level information may help detect fitter lineages earlier than case counts alone. Applied to the full set of publicly available SARS-CoV-2 genomes, it provides a genomic view of the mutations driving increased fitness of the virus, identifying experimentally established driver mutations in S and highlighting the key role of non-S mutations, particularly in N, ORF1b, and ORF1a, which have received relatively less research attention. By modeling millions of viral sequences across thousands of regions, PyR$_0$ yields mechanistic insight into viral fitness and offers a panoramic view of viral evolution, revealing a pattern whereby major circulating lineages fragment into sublineages with modest differences in fitness before they are collectively displaced by the sudden emergence of markedly fitter variants.

REFERENCES AND NOTES

1. N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. B. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, T. Wenseleers, A. Gimma, W. Waites, K. L. M. Wong, K. van Zandvoort, J. D. Silverman, K. Diaz-Ordaz, R. Keogh, R. M. Eggo, S. Funk, M. Jit, K. E. Atkins, W. J. Edmunds; CMMID COVID-19 Working Group; COVID-19 Genomics UK (COG-UK) Consortium, Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, eabg3055 (2021). doi:10.1126/science.abg3055 Medline

2. E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O'Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C. V. Ariani, O. Boyd, N. J. Loman, J. T. McCrone, S. Gonçalves, D. Jorgensen, R. Myers, V. Hill, D. K. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D. P. Kwiatkowski, S. Flaxman, O. Ratmann, S. Bhatt, S. Hopkins, A. Gandy, A. Rambaut, N. M. Ferguson; COVID-19 Genomics UK (COG-UK) consortium, Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021). Medline

3. P. Stefanelli, F. Trentini, G. Guzzetta, V. Marziano, A. Mammone, P. Poletti, C. M. Grané, M. Manica, M. del Manso, X. Andrianou, Others, Co-circulation of SARS-CoV-2 variants B. 1.1. 7 and P. 1. *medRxiv* (2021) (available at https://www.medrxiv.org/content/10.1101/2021.04.06.21254923v1.abstract).

4. P. Stefanelli, F. Trentini, G. Guzzetta, V. Marziano, A. Mammone, M. Sane Schepisi, P. Poletti, C. Molina Grané, M. Manica, M. Del Manso, X. Andrianou, M. Ajelli, G. Rezza, S. Brusaferro, S. Merler; COVID-19 National Microbiology Surveillance Study Group, Co-circulation of SARS-CoV-2 Alpha and Gamma variants in Italy, February and March 2021. *Euro Surveill.* **27**, (2022). doi:10.2807/1560-7917.ES.2022.27.5.2100429 Medline

5. H. S. Vöhringer, T. Sanderson, M. Sinnott, N. De Maio, T. Nguyen, R. Goater, F. Schwach, I. Harrison, J. Hellewell, C. V. Ariani, S. Gonçalves, D. K. Jackson, I. Johnston, A. W. Jung, C. Saint, J. Sillitoe, M. Suciu, N. Goldman, J. Panovska-Griffiths, E. Birney, E. Volz, S. Funk, D. Kwiatkowski, M. Chand, I. Martincorena, J. C. Barrett, M. Gerstung; Wellcome Sanger Institute COVID-19 Surveillance Team; COVID-19 Genomics UK (COG-UK) Consortium*, Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature* **600**, 506–511 (2021). Medline

6. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, C. McDanal, L. G. Perez, H. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, D. C. Montefiori, A. Angyal, R. L. Brown, L. Carrilero, L. R. Green, D. C. Groves, K. J. Johnson, A. J. Keeley, B. B. Lindsey, P. J. Parsons, M. Raza, S. Rowland-Jones, N. Smith, R. M. Tucker, D. Wang, M. D. Wyles; Sheffield COVID-19 Genomics Group, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812–827.e19 (2020). doi:10.1016/j.cell.2020.06.043 Medline

7. L. Yurkovetskiy, X. Wang, K. E. Pascal, C. Tomkins-Tinch, T. P. Nyalile, Y. Wang, A. Baum, W. E. Diehl, A. Dauphin, C. Carbone, K. Veinotte, S. B. Egri, S. F. Schaffner, J. E. Lemieux, J. B. Munro, A. Rafique, A. Barve, P. C. Sabeti, C. A. Kyratsous, N. V. Dudkina, K. Shen, J. Luban, Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell* **183**, 739–751.e8 (2020). doi:10.1016/j.cell.2020.09.032 Medline

8. X. Deng, M. A. Garcia-Knight, M. M. Khalid, V. Servellita, C. Wang, M. K. Morris, A. Sotomayor-González, D. R. Glasner, K. R. Reyes, A. S. Gliwa, N. P. Reddy, C. Sanchez San Martin, S. Federman, J. Cheng, J. Balcerek, J. Taylor, J. A. Streithorst, S. Miller, B. Sreekumar, P.-Y. Chen, U. Schulze-Gahmen, T. Y. Taha, J. M. Hayashi, C. R. Simoneau, G. R. Kumar, S. McMahon, P. V. Lidsky, Y. Xiao, P. Hemarajata, N. M. Green, A. Espinosa, C. Kath, M. Haw, J. Bell, J. K. Hacker, C. Hanson, D. A. Wadford, C. Anaya, D. Ferguson, P. A. Frankino, H. Shivram, L. F. Lareau, S. K. Wyman, M. Ott, R. Andino, C. Y. Chiu, Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cell* **184**, 3426–3437.e8 (2021). doi:10.1016/j.cell.2021.04.025 Medline

9. T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. D. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, N. P. King, D. Veesler, J. D. Bloom, Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295–1310.e20 (2020). doi:10.1016/j.cell.2020.08.012 Medline

10. B. Choi, M. C. Choudhary, J. Regan, J. A. Sparks, R. F. Padera, X. Qiu, I. H. Solomon, H.-H. Kuo, J. Boucau, K. Bowman, U. D. Adhikari, M. L. Winkler, A. A. Mueller, T. Y.-T. Hsu, M. Desjardins, L. R. Baden, B. T. Chan, B. D. Walker, M. Lichterfeld, M. Brigl, D. S. Kwon, S. Kanjilal, E. T. Richardson, A. H. Jonsson, G. Alter, A. K. Barczak, W. P. Hanage, X. G. Yu, G. D. Gaiha, M. S. Seaman, M. Cernadas, J. Z. Li, Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.* **383**, 2291–2293 (2020). doi:10.1056/NEJMc2031364 Medline

11. A. J. Greaney, T. N. Starr, P. Gilchuk, S. J. Zost, E. Binshtein, A. N. Loes, S. K. Hilton, J. Huddleston, R. Eguia, K. H. D. Crawford, A. S. Dingens, R. S. Nargi, R. E. Sutton, N. Suryadevara, P. W. Rothlauf, Z. Liu, S. P. J. Whelan, R. H. Carnahan, J. E. Crowe Jr., J. D. Bloom, Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* **29**, 44–57.e9 (2021). doi:10.1016/j.chom.2020.11.007 Medline

12. A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020). doi:10.1038/s41564-020-0770-5 Medline

13. H. S. Vöhringer, T. Sanderson, M. Sinnott, N. De Maio, T. Nguyen, R. Goater, F. Schwach, I. Harrison, J. Hellewell, C. Ariani, S. Gonçalves, D. Jackson, I. Johnston, A. W. Jung, C. Saint, J. Sillitoe, M. Suciu, N. Goldman, E. Birney, S. Funk, E. Volz, D. Kwiatkowski, M. Chand, I. Martincorena, J. C. Barrett, M. Gerstung, The Wellcome Sanger Institute Covid-19 Surveillance Team, The COVID-19 Genomics UK (COG-UK) Consortium, Genomic reconstruction of the SARS-CoV-2 epidemic across England from September 2020 to May 2021*bioRxiv* (2021), doi:10.1101/2021.05.22.21257633.

14. F. Campbell, B. Archer, H. Laurenson-Schafer, Y. Jinnai, F. Konings, N. Batra, B. Pavlin, K. Vandemaele, M. D. Van Kerkhove, T. Jombart, O. Morgan, O. le Polain de Waroux, Increased transmissibility and global spread of SARS-CoV-

2 variants of concern as at June 2021. *Euro Surveill.* **26**, (2021). doi:10.2807/1560-7917.ES.2021.26.24.2100509 Medline

15. S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **1**, 33–46 (2017). doi:10.1002/gch2.1018 Medline

16. E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, N. D. Goodman, Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.* **20**, 973–978 (2019).

17. F. Obermeyer, S. F. Schaffner, M. Jankowiak, N. Barkas, J. D. Pyle, D. J. Park, B. L. MacInnis, J. Luban, P. C. Sabeti, J. E. Lemieux, Analysis of 2.1 million SARS-CoV-2 genomes identifies mutations associated with transmissibility*medRxiv* (2021), doi:10.1101/2021.09.07.21263228.

18. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations*Virological* (2020) (available at 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).

19. R. Viana, S. Moyo, D. G. Amoako, H. Tegally, C. Scheepers, C. L. Althaus, U. J. Anyaneji, P. A. Bester, M. F. Boni, M. Chand, W. T. Choga, R. Colquhoun, M. Davids, K. Deforche, D. Doolabh, L. du Plessis, S. Engelbrecht, J. Everatt, J. Giandhari, M. Giovanetti, D. Hardie, V. Hill, N.-Y. Hsiao, A. Iranzadeh, A. Ismail, C. Joseph, R. Joseph, L. Koopile, S. L. Kosakovsky Pond, M. U. G. Kraemer, L. Kuate-Lere, O. Laguda-Akingba, O. Lesetedi-Mafoko, R. J. Lessells, S. Lockman, A. G. Lucaci, A. Maharaj, B. Mahlangu, T. Maponga, K. Mahlakwane, Z. Makatini, G. Marais, D. Maruapula, K. Masupu, M. Matshaba, S. Mayaphi, N. Mbhele, M. B. Mbulawa, A. Mendes, K. Mlisana, A. Mnguni, T. Mohale, M. Moir, K. Moruisi, M. Mosepele, G. Motsatsi, M. S. Motswaledi, T. Mphoyakgosi, N. Msomi, P. N. Mwangi, Y. Naidoo, N. Ntuli, M. Nyaga, L. Olubayo, S. Pillay, B. Radibe, Y. Ramphal, U. Ramphal, J. E. San, L. Scott, R. Shapiro, L. Singh, P. Smith-Lawrence, W. Stevens, A. Strydom, K. Subramoney, N. Tebeila, D. Tshiabuila, J. Tsui, S. van Wyk, S. Weaver, C. K. Wibmer, E. Wilkinson, N. Wolter, A. E. Zarebski, B. Zuze, D. Goedhals, W. Preiser, F. Treurnicht, M. Venter, C. Williamson, O. G. Pybus, J. Bhiman, A. Glass, D. P. Martin, A. Rambaut, S. Gaseitsiwe, A. von Gottberg, T. de Oliveira, Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature* **603**, 679–686 (2022). doi:10.1038/s41586-022-04411-y Medline

20. A. J. Greaney, T. N. Starr, J. D. Bloom, An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain. *Virus Evol.* **8**, veac021 (2022). doi:10.1093/ve/veac021 Medline

21. J. Yu, A. Y. Collier, M. Rowe, F. Mardas, J. D. Ventura, H. Wan, J. Miller, O. Powers, B. Chung, M. Siamatu, N. P. Hachmann, N. Surve, F. Nampanya, A. Chandrashekar, D. H. Barouch, Neutralization of the SARS-CoV-2 omicron BA.1 and BA.2 variants. *N. Engl. J. Med.* **386**, 1579–1580 (2022). doi:10.1056/NEJMc2201849 Medline

22. A. M. Syed, T. Y. Taha, T. Tabata, I. P. Chen, A. Ciling, M. M. Khalid, B. Sreekumar, P.-Y. Chen, J. M. Hayashi, K. M. Soczek, M. Ott, J. A. Doudna, Rapid assessment of SARS-CoV-2-evolved variants using virus-like particles. *Science* **374**, 1626–1632 (2021). doi:10.1126/science.abl6184 Medline

23. L. Ferretti, A. Ledda, C. Wymant, L. Zhao, V. Ledda, L. Abeler-Dörner, M. Kendall, A. Nurtay, H.-Y. Cheng, T.-C. Ng, H.-H. Lin, R. Hinch, J. Masel, A. M. Kilpatrick, C. Fraser, The timing of COVID-19 transmission*bioRxiv* (2020), doi:10.1101/2020.09.04.20188516.

24. *broadinstitute/pyro-cov: v0.2.1* (2022; https://zenodo.org/record/6399987).

25. Y. Turakhia, B. Thornlow, A. S. Hinrichs, N. De Maio, L. Gozashti, R. Lanfear, D. Haussler, R. Corbett-Detig, Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021). doi:10.1038/s41588-021-00862-7 Medline

26. J. McBroome, B. Thornlow, A. S. Hinrichs, A. Kramer, N. De Maio, N. Goldman, D. Haussler, R. Corbett-Detig, Y. Turakhia, A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. *Mol. Biol. Evol.* **38**, 5819–5824 (2021). doi:10.1093/molbev/msab264 Medline

27. S. Nersisyan, A. Zhiyanov, M. Shkurnikov, A. Tonevitsky, T-CoV: a comprehensive portal of HLA-peptide interactions affected by SARS-CoV-2 mutations*bioRxiv*, 2021.07.06.451227 (2021).

28. J. F. Crow, M. and Kimura, *An Introduction to Population Genetics Theory* (The Blackburn Press, 1970).

29. T. A. Hopf, C. P. I. Schärfe, J. P. G. L. M. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. J. J. Bonvin, D. S. Marks, Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**, e03430 (2014). doi:10.7554/eLife.03430 Medline

30. J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, D. S. Marks, Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021). doi:10.1038/s41586-021-04043-8 Medline

31. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch (2017) (available at https://openreview.net/pdf?id=BJJsrmfCZ).

32. M. Gorinova, D. Moore, M. Hoffman, in *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research. H. D. Iii, A. Singh, Eds. (PMLR, 2020), vol. 119, pp. 3648–3657.

33. R. M. Neal, Slice sampling. *Ann. Stat.* **31**, (2003). doi:10.1214/aos/1056562461

34. D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization*arXiv [cs.LG]* (2014) (available at https://arxiv.org/abs/1412.6980).

35. L. Cappello, J. Kim, S. Liu, J. A. Palacios, Statistical Challenges in Tracking the Evolution of SARS-CoV-2*arXiv [stat.AP]* (2021) (available at https://arxiv.org/abs/2108.13362).

36. Y. Cao, J. Wang, F. Jian, T. Xiao, W. Song, A. Yisimayi, W. Huang, Q. Li, P. Wang, R. An, J. Wang, Y. Wang, X. Niu, S. Yang, H. Liang, H. Sun, T. Li, Y. Yu, Q. Cui, S. Liu, X. Yang, S. Du, Z. Zhang, X. Hao, F. Shao, R. Jin, X. Wang, J. Xiao, Y. Wang, X. S. Xie, Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022). doi:10.1038/s41586-021-04385-3 Medline

37. D. Planas, N. Saunders, P. Maes, F. Guivel-Benhassine, C. Planchais, J. Buchrieser, W.-H. Bolland, F. Porrot, I. Staropoli, F. Lemoine, H. Péré, D. Veyer, J. Puech, J. Rodary, G. Baele, S. Dellicour, J. Raymenants, S. Gorissen, C. Geenen, B. Vanmechelen, T. Wawina-Bokalanga, J. Martí-Carreras, L. Cuypers, A. Sève, L. Hocqueloux, T. Prazuck, F. A. Rey, E. Simon-Loriere, T. Bruel, H. Mouquet, E. André, O. Schwartz, Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. *Nature* **602**, 671–675 (2022). doi:10.1038/s41586-021-04389-z Medline

38. Y. Weisblum, F. Schmidt, F. Zhang, J. DaSilva, D. Poston, J. C. Lorenzi, F. Muecksch, M. Rutkowska, H.-H. Hoffmann, E. Michailidis, C. Gaebler, M. Agudelo, A. Cho, Z. Wang, A. Gazumyan, M. Cipolla, L. Luchsinger, C. D. Hillyer, M. Caskey, D. F. Robbiani, C. M. Rice, M. C. Nussenzweig, T. Hatziioannou, P. D. Bieniasz, Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *eLife* **9**, e61312 (2020). doi:10.7554/eLife.61312 Medline

39. A. E. Lin, W. E. Diehl, Y. Cai, C. L. Finch, C. Akusobi, R. N. Kirchdoerfer, L. Bollinger, S. F. Schaffner, E. A. Brown, E. O. Saphire, K. G. Andersen, J. H. Kuhn, J. Luban, P. C. Sabeti, Reporter Assays for Ebola Virus Nucleoprotein

Oligomerization, Virion-Like Particle Budding, and Minigenome Activity Reveal the Importance of Nucleoprotein Amino Acid Position 111. *Viruses* **12**, 105 (2020). doi:10.3390/v12010105 Medline

40. A. M. Syed, T. Y. Taha, M. M. Khalid, T. Tabata, I. P. Chen, B. Sreekumar, P.-Y. Chen, J. M. Hayashi, K. M. Soczek, M. Ott, J. A. Doudna, Rapid assessment of SARS-CoV-2 evolved variants using virus-like particles*bioRxiv*, 2021.08.05.455082 (2021).

41. M. M. Angelini, M. Akhlaghpour, B. W. Neuman, M. J. Buchmeier, Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. *mBio* **4**, e00524-13 (2013). doi:10.1128/mBio.00524-13 Medline

42. R. L. Graham, A. C. Sims, S. M. Brockway, R. S. Baric, M. R. Denison, The nsp2 replicase proteins of murine hepatitis virus and severe acute respiratory syndrome coronavirus are dispensable for viral replication. *J. Virol.* **79**, 13399–13411 (2005). doi:10.1128/JVI.79.21.13399-13411.2005 Medline

43. I. Jungreis, R. Sealfon, M. Kellis, SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. *Nat. Commun.* **12**, 2642 (2021). doi:10.1038/s41467-021-22905-7 Medline

44. M. R. Islam, M. N. Hoque, M. S. Rahman, A. S. M. R. U. Alam, M. Akther, J. A. Puspo, S. Akter, M. Sultana, K. A. Crandall, M. A. Hossain, Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. *Sci. Rep.* **10**, 14004 (2020). doi:10.1038/s41598-020-70812-6 Medline

45. C. T. Cornillez-Ty, L. Liao, J. R. Yates 3rd, P. Kuhn, M. J. Buchmeier, Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. *J. Virol.* **83**, 10314–10318 (2009). doi:10.1128/JVI.00842-09 Medline

46. M. Gupta, C. M. Azumaya, M. Moritz, S. Pourmal, A. Diallo, G. E. Merz, G. Jang, M. Bouhaddou, A. Fossati, A. F. Brilot, D. Diwanji, E. Hernandez, N. Herrera, H. T. Kratochvil, V. L. Lam, F. Li, Y. Li, H. C. Nguyen, C. Nowotny, T. W. Owens, J. K. Peters, A. N. Rizo, U. Schulze-Gahmen, A. M. Smith, I. D. Young, Z. Yu, D. Asarnow, C. Billesbølle, M. G. Campbell, J. Chen, K.-H. Chen, U. S. Chio, M. S. Dickinson, L. Doan, M. Jin, K. Kim, J. Li, Y.-L. Li, E. Linossi, Y. Liu, M. Lo, J. Lopez, K. E. Lopez, A. Mancino, F. R. Moss, M. D. Paul, K. I. Pawar, A. Pelin, T. H. Pospiech, C. Puchades, S. G. Remesh, M. Safari, K. Schaefer, M. Sun, M. C. Tabios, A. C. Thwin, E. W. Titus, R. Trenker, E. Tse, T. K. M. Tsui, F. Wang, K. Zhang, Y. Zhang, J. Zhao, F. Zhou, Y. Zhou, L. Zuliani-Alvarez, QCRG Structural Biology Consortium, D. A. Agard, Y. Cheng, J. S. Fraser, N. Jura, T. Kortemme, A. Manglik, D. R. Southworth, R. M. Stroud, D. L. Swaney, N. J. Krogan, A. Frost, O. S. Rosenberg, K. A. Verba, CryoEM and AI reveal a structure of SARS-CoV-2 Nsp2, a multifunctional protein involved in key host processes. *bioRxiv* (2021), doi:10.1101/2021.05.10.443524.

47. Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L. W. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao, H. Yang, Structure of M^{pro} from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293 (2020). doi:10.1038/s41586-020-2223-y Medline

48. J. Osipiuk, S.-A. Azizi, S. Dvorkin, M. Endres, R. Jedrzejczak, K. A. Jones, S. Kang, R. S. Kathayat, Y. Kim, V. G. Lisnyak, S. L. Maki, V. Nicolaescu, C. A. Taylor, C. Tesar, Y.-A. Zhang, Z. Zhou, G. Randall, K. Michalska, S. A. Snyder, B. C. Dickinson, A. Joachimiak, Structure of papain-like protease from SARS-CoV-2 and its complexes with non-covalent inhibitors. *Nat. Commun.* **12**, 743 (2021). doi:10.1038/s41467-021-21060-3 Medline

49. H. S. Hillen, G. Kokic, L. Farnung, C. Dienemann, D. Tegunov, P. Cramer, Structure of replicating SARS-CoV-2 polymerase. *Nature* **584**, 154–156 (2020). doi:10.1038/s41586-020-2368-8 Medline

50. L. Yan, J. Ge, L. Zheng, Y. Zhang, Y. Gao, T. Wang, Y. Huang, Y. Yang, S. Gao, M. Li, Z. Liu, H. Wang, Y. Li, Y. Chen, L. W. Guddat, Q. Wang, Z. Rao, Z. Lou, Cryo-EM Structure of an Extended SARS-CoV-2 Replication and Transcription Complex Reveals an Intermediate State in Cap Synthesis. *Cell* **184**, 184–193.e10 (2021). doi:10.1016/j.cell.2020.11.016 Medline

51. J. Chen, B. Malone, E. Llewellyn, M. Grasso, P. M. M. Shelton, P. D. B. Olinares, K. Maruthi, E. T. Eng, H. Vatandaslar, B. T. Chait, T. M. Kapoor, S. A. Darst, E. A. Campbell, Structural Basis for Helicase-Polymerase Coupling in the SARS-CoV-2 Replication-Transcription Complex. *Cell* **182**, 1560–1573.e13 (2020). doi:10.1016/j.cell.2020.07.033 Medline

52. Y. Chen, H. Cai, J. Pan, N. Xiang, P. Tien, T. Ahola, D. Guo, Functional screen reveals SARS coronavirus nonstructural protein nsp14 as a novel cap N7 methyltransferase. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3484–3489 (2009). doi:10.1073/pnas.0808790106 Medline

53. Y. Huang, C. Yang, X.-F. Xu, W. Xu, S.-W. Liu, Structural and functional properties of SARS-CoV-2 spike protein: Potential antivirus drug development for COVID-19. *Acta Pharmacol. Sin.* **41**, 1141–1149 (2020). doi:10.1038/s41401-020-0485-4 Medline

54. J. Cubuk, J. J. Alston, J. J. Incicco, S. Singh, M. D. Stuchell-Brereton, M. D. Ward, M. I. Zimmerman, N. Vithani, D. Griffith, J. A. Wagoner, G. R. Bowman, K. B. Hall, A. Soranno, A. S. Holehouse, The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nat. Commun.* **12**, 1936 (2021). doi:10.1038/s41467-021-21953-3 Medline

55. Z. Chen, D. Pei, L. Jiang, Y. Song, J. Wang, H. Wang, D. Zhou, J. Zhai, Z. Du, B. Li, M. Qiu, Y. Han, Z. Guo, R. Yang, Antigenicity analysis of different regions of the severe acute respiratory syndrome coronavirus nucleocapsid protein. *Clin. Chem.* **50**, 988–995 (2004). doi:10.1373/clinchem.2004.031096 Medline

56. Alaa Abdel Latif, Julia L. Mullen, Manar Alkuzweny, Ginger Tsueng, Marco Cano, Emily Haag, Jerry Zhou, Mark Zeller, Emory Hufbauer, Nate Matteson, Chunlei Wu, Kristian G. Andersen, Andrew I. Su, Karthik Gangavarapu, Laura D. Hughes, and the Center for Viral Systems Biology, Spike:D614G Mutation Report.

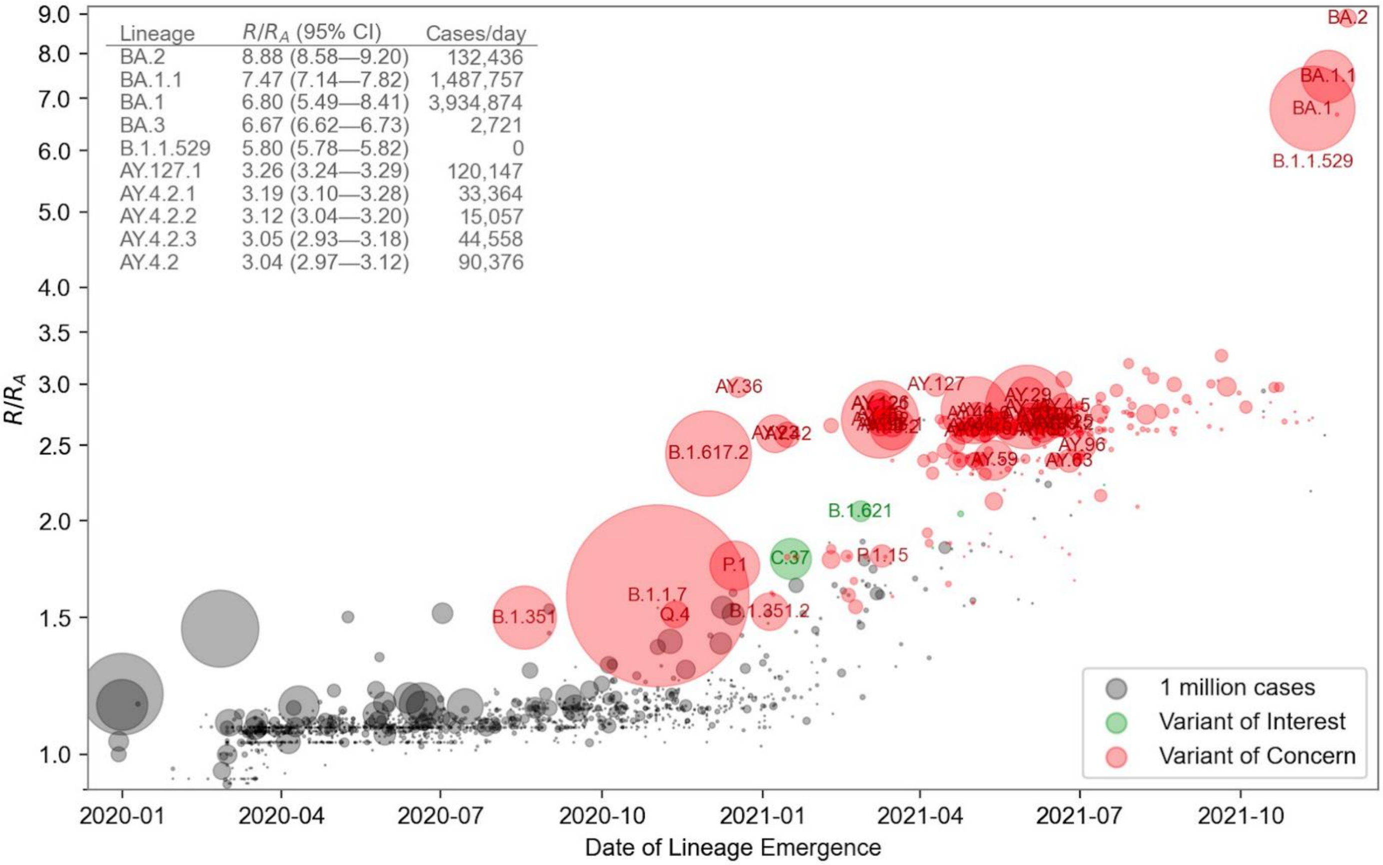
ACKNOWLEDGMENTS

We acknowledge crucial assistance in data preprocessing from A. Hinrichs. We thank T. Bedford and C. Roemer for visualizing the outputs of our model on nextstrain.org. We acknowledge helpful discussions and feedback from D. Phan, W. Hanage, C. Tomkins-Tinch, S. Weingarten-Gabbay, K. Siddle, S. Gosai, S. Reilly, E. Bingham, H. Soutter, D. Marks, N. Youssef, S. Gurev, and N. Thadani. We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared via GISAID genetic sequence data on which this research is based (Supplementary Data File 3). This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material. **Funding:** This work was sponsored by the U.S. Centers for Disease Control and Prevention (BAA 75D30120C09605 to B.L.M.), as well as support from the Doris Duke Charitable Foundation (J.E.L.), the Howard Hughes Medical Institute (P.C.S.), the National Institute of Allergy and Infectious Diseases R37AI147868 (J.L.) and U19AI110818 (P.C.S.), and the Evergrande COVID-19 Response Fund Award from the Massachusetts Consortium on Pathogen Readiness (J.L.). **Competing interests:** P.C.S. is a co-founder of and consultant to Sherlock Biosciences and a Board Member of Danaher Corporation, and holds equity in the companies. The authors declare no conflicts of interest. **Author contributions:** Conceptualization: F.O., S.F.S., J.E.L., M.J. Data curation: F.O., N.B. Formal Analysis: F.O., S.F.S, M.J., N.B., J.E.L. Funding acquisition: D.J.P., B.M., P.C.S, J.L., J.E.L. Investigation: all authors. Methodology: F.O., S.F.S, M.J., J.E.L., L.Y., M.B. Project administration: all authors. Software: F.O., N.B., M.J. Supervision: D.J.P., B.M., J.L., P.C.S., J.E.L. Validation: F.O., N.B., M.J., S.F.S. Visualization: F.O., J.E.L., N.B., J.P., S.F.S. Writing – original draft: F.O., S.F.S., B.M., P.C.S, J.E.L. Writing – review and editing: All authors. **Data and materials availability:** Code is available at (*24*). We gratefully acknowledge all data contributors, i.e., the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID initiative (*15*) on which this research is based. A total of 6,466,300 submissions are included in this study. A complete list of 6.4 million accession numbers is included as data S3.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm1208
Materials and Methods
Supplementary Text
Figures S1 to S34
Tables S1 to S5
References (*25*–*56*)
MDAR Reproducibility Checklist
Data S1 to S5
GISAID Acknowledgments table

26 August 2021; resubmitted 16 February 2022
Accepted 10 May 2022
Published online 24 May 2022
10.1126/science.abm1208



Lineage	R/R_A (95% CI)	Cases/day
BA.2	8.88 (8.58—9.20)	132,436
BA.1.1	7.47 (7.14—7.82)	1,487,757
BA.1	6.80 (5.49—8.41)	3,934,874
BA.3	6.67 (6.62—6.73)	2,721
B.1.1.529	5.80 (5.78—5.82)	0
AY.127.1	3.26 (3.24—3.29)	120,147
AY.4.2.1	3.19 (3.10—3.28)	33,364
AY.4.2.2	3.12 (3.04—3.20)	15,057
AY.4.2.3	3.05 (2.93—3.18)	44,558
AY.4.2	3.04 (2.97—3.12)	90,376

Fig. 1. Relative fitness versus date of lineage emergence. Circle size is proportional to cumulative case count inferred from lineage proportion estimates and confirmed case counts. Inset table lists the 10 fittest lineages inferred by the model. R/R_A is the fold increase in relative fitness over the Wuhan (**A**) lineage, assuming a fixed generation time of 5.5 days.

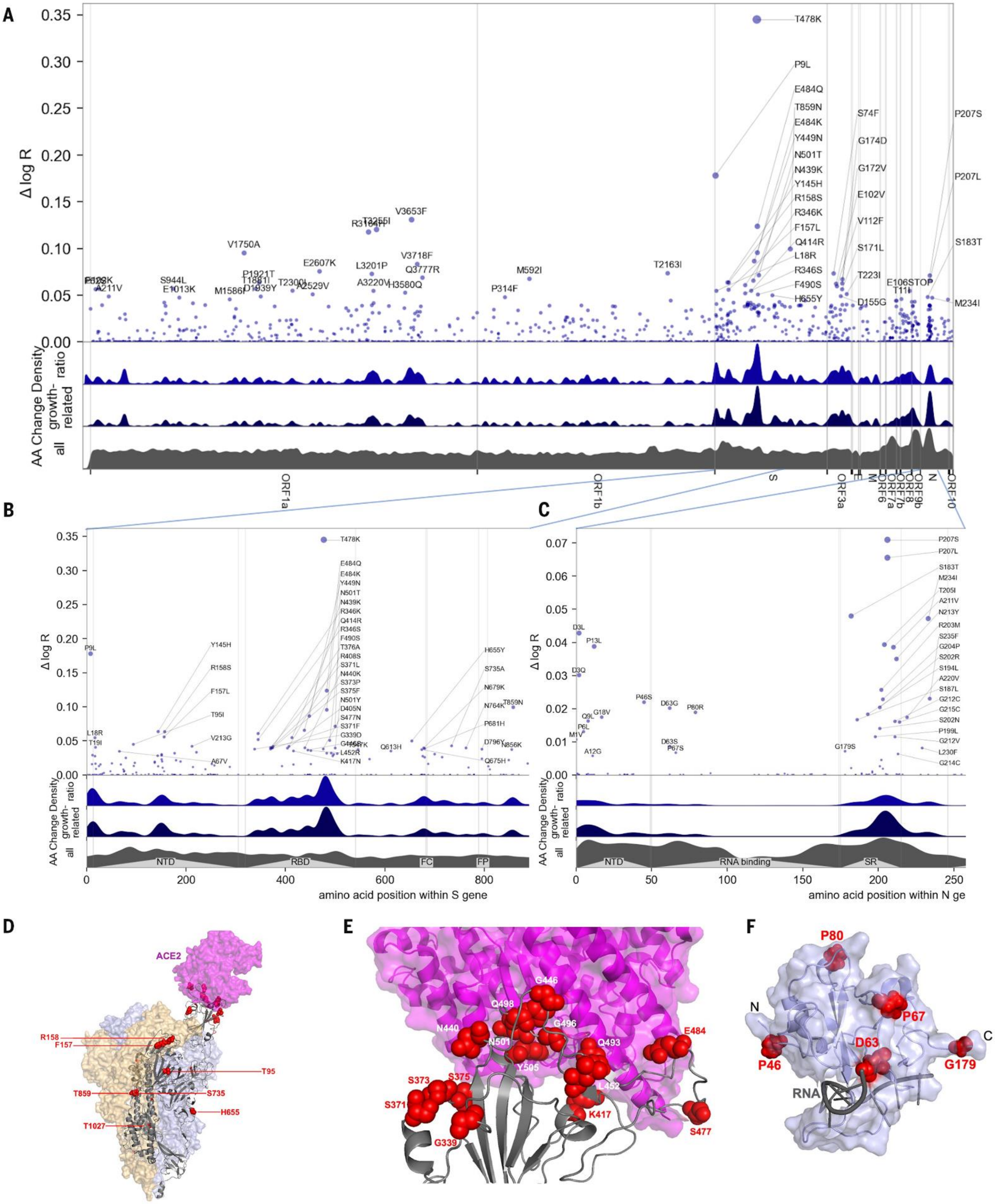
A
Δ log R
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
T478K
P9L
E484Q
T859N
E484K
Y449N
N501T
N439K
Y145H
R158S
R346K
F157L
Q414R
L18R
R346S
F490S
H655Y
S74F
G174D
G172V
E102V
V112F
S171L
T223I
D155G
E106STOP
T11I
P207S
P207L
S183T
M234I
V3653F
R3164P
T3255I
V3718F
Q3777R
H3580Q
A3220V
L3201P
E2607K
A2529V
T2300I
V1750A
P1921T
T1891I
D1939Y
M1586I
S944L
E1013K
P323K
F323K
A211V
M592I
P314F
T2163I
AA Change Density
ratio
growth-related
all
ORF1a
ORF1b
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
ORF9b
N
ORF10
B
Δ log R
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
T478K
E484Q
E484K
Y449N
N501T
N439K
R346K
Q414R
R346S
F490S
T376A
R408S
S371L
N440K
S373P
S375F
N501Y
D405N
S477N
S371F
G339D
G446S
T547K
L452R
K417N
P9L
L18R
T19I
Y145H
R158S
F157L
T95I
V213G
A67V
Q613H
H655Y
S735A
N679K
N764K
T859N
P681H
D796Y
N856K
Q675H
AA Change Density
ratio
growth-related
all
NTD
RBD
FC
FP
0
200
400
600
800
amino acid position within S gene
C
Δ log R
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
P207S
P207L
S183T
M234I
T205I
A211V
N213Y
R203M
S235F
G204P
S202R
S194L
A220V
S187L
G212C
G215C
S202N
P199L
G212V
L230F
G214C
G179S
D3L
P13L
D3Q
P46S
D63G
P80R
Q9L
G18V
P6L
M1V
A12G
D63S
P67S
AA Change Density
ratio
growth-related
all
NTD
RNA binding
SR
0
50
100
150
200
250
amino acid position within N ge
D
ACE2
R158
F157
T95
T859
S735
H655
T1027
E
G446
Q498
G496
N440
N501
Q493
E484
Y505
S373
S375
S371
L452
K417
G339
S477
F
P80
N
P67
P46
D63
G179
C
RNA

Fig. 2. Manhattan plot of amino acid changes assessed in this study. (**A**) Changes across the entire genome. (**B**) Changes in the first 850 amino acids of S. In each of (A) to (C) the y axis shows effect size Δ log R, the estimated change in log relative fitness due to each amino acid change. The bottom three axes show the background density of all observed amino acid changes, the density of those associated with growth (weighted by |Δ log R|), and the ratio of the two. The top 55 amino acid changes are labeled. See fig. S13 for detailed views of S, N, ORF1a, and ORF1b. (**C**)**.** Changes in the first 250 amino acids of N. (**D**) Structure of the spike-ACE2 complex (PDB: 7KNB). Spike subunits colored light blue, light orange, and gray. Top-ranked mutations are shown as red spheres. ACE2 is shown in magenta. (**E**) Close-up view of the RBD interface. (**F**) Top-ranked mutations in the N-terminal RNA-binding domain of N. Residues 44-180 of N (PDB: 7ACT) are shown in light blue. Amino acid positions corresponding to top mutations in this region are shown as red spheres. A 10-nt bound RNA is shown in gray.

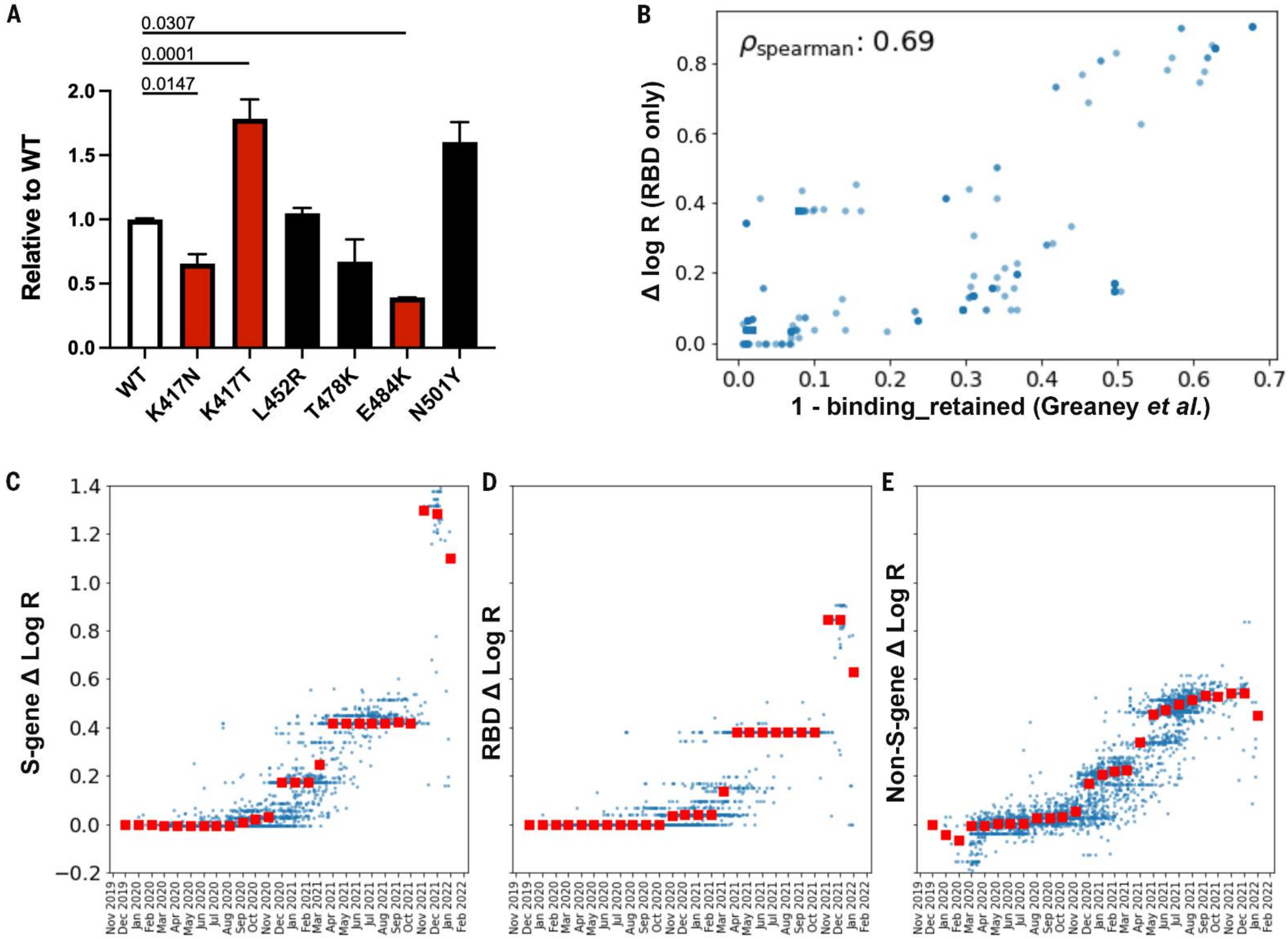


Fig. 3. (**A**) **Infectivity relative to WT of lentiviral vectors pseudotyped with the indicated Spike mutants.** Target cells were HEK293T cells expressing ACE2 and TMPRSS2 transgenes. The genetic background of the Spike was Wuhan-Hu-1 bearing D614G. Red bars were significantly different from WT (adjusted p values shown). Black bars were not significantly different from WT. (**B**) For the 1701 SARS-CoV-2 clusters with at least one amino acid substitution in the RBD domain we compare: i) the PyR$_0$ prediction for the contribution to Δ log R from RBD substitutions only; to ii) antibody binding computed using the antibody-escape calculator in (*20*). The escape calculator is based on an intuitive non-linear model parameterized using deep mutational scanning data for 33 neutralizing antibodies elicited by SARS-CoV-2. PyR$_0$ predictions exhibit high (Spearman) correlation with predictions from Greaney et al. (*20*) (**C to E**) We dissect PyR$_0$ Δ log R estimates into S-gene (C), RBD (D), and non-S-gene (E) contributions for 3000 SARS-CoV-2 clusters (blue dots). The horizontal axis corresponds to the date at which each cluster first emerged. Red squares denote the median Δ log R within each monthly bin. The increased importance of S-gene mutations (notably in the RBD) over non-S-gene mutations starting around November 2021 is apparent.

Table 1. Amino acid substitutions most significantly associated with increased fitness. Significance is defined as posterior mean / posterior standard deviation. Fitness is per 5.5 days (estimated generation time of the Wuhan (A) lineage (1, 23)). Final column: number of PANGO lineages in which each substitution emerged independently.

Rank	Gene	Substitution	Fold Increase in Fitness	Number of Lineages
1	S	H655Y	1.051	33
2	S	T95I	1.046	30
3	ORF1a	P3395H	1.039	5
4	S	N764K	1.04	6
5	ORF1a	K856R	1.039	2
6	S	S371L	1.041	3
7	E	T9I	1.04	5
8	S	Q954H	1.04	5
9	ORF9b	P10S	1.039	25
10	S	L981F	1.04	2
11	N	P13L	1.04	25
12	S	G339D	1.039	4
13	S	S375F	1.04	5
14	S	S477N	1.039	47
15	S	N679K	1.04	11
16	S	S373P	1.04	5
17	M	Q19E	1.039	5
18	S	D796Y	1.038	11
19	S	N969K	1.04	5
20	S	T547K	1.038	3