UNIVERSITY OF CALIFORNIA SAN DIEGO

The zoonotic origins and molecular epidemiology of SARS-CoV-2

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy

in

Bioinformatics & Systems Biology with a Specialization in Biomedical Informatics

by

Jonathan Pekar

Committee in charge:

Professor Joel O. Wertheim, Chair
Professor Siavash Mirarab, Co-Chair
Professor Justin Meyer
Professor Niema Moshiri
Professor Tetyana Vasylyeva

2023

Copyright

Jonathan Pekar, 2023

All rights reserved.

The Dissertation of Jonathan Pekar is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To *Mom*, *Dad*, and *Eric*.

To all the healthcare workers on the front lines of the COVID-19 pandemic, all the pandemics before, and all the pandemics to come.

EPIGRAPH

Ils se croyaient libres et personne ne sera jamais libre tant qu'il y aura des fléaux.
(They thought they were free and no one will ever be free as long as there are plagues.)

- Albert Camus, *La Peste* (1947)

TABLE OF CONTENTS

Dissertation Approval Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures xi

List of Tables xvi

Acknowledgements xviii

Vita xxi

Abstract of the Dissertation xxiii

Introduction 1

Chapter 1 Timing the SARS-CoV-2 Index Case in Hubei Province 5
1.1 Introduction 5
1.2 Results 7
1.3 Discussion 15
1.4 Acknowledgments 17
1.4.1 Author contributions 17
1.4.2 Declaration of interests 18
1.4.3 Data and materials availability 18

Chapter 2 The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2 19
2.1 Introduction 19
2.2 Results 21
2.2.1 Erroneous assignment of haplotypes intermediate to lineages A and B 21
2.2.2 Progenitor genome reconstruction 21
2.2.3 Reversions across the early pandemic phylogeny 22
2.2.4 Inferring the MRCA of SARS-CoV-2 22
2.2.5 Separate introductions of lineages A and B 25
2.2.6 Timing the introductions of lineages A and B 28
2.2.7 Minimal cryptic circulation of SARS-CoV-2 29
2.2.8 Additional introductions 30
2.2.9 Limitations 30
2.3 Discussion 31
2.4 Materials and Methods Summary 33
2.4.1 Sequence data 33
2.4.2 Progenitor genome reconstruction and reversion analysis 33
2.4.3 Phylodynamic inference and epidemic simulations 33

2.5 Acknowledgments . . . 34
2.5.1 Funding . . . 35
2.5.2 Author contributions . . . 36
2.5.3 Competing interests . . . 36
2.5.4 Data and materials availability . . . 36

Chapter 3 The recency and geographical origins of the bat viruses ancestral to SARS-CoV and SARS-CoV-2 . . . 37
3.1 Introduction . . . 38
3.2 Inference of non-recombinant regions . . . 39
3.3 Sarbecovirus genomes provide evidence for a very recent bat ancestor of SARS-CoV-1 and SARS-CoV-2 . . . 40
3.4 The recombinant common ancestors became more similar to SARS-CoV-1 and SARS-CoV-2 with increased sampling . . . 43
3.5 Phylogeography of bat sarbecoviruses related to SARS-CoV-1 and SARS-CoV-2 . . . 45
3.6 Location of the closest-inferred ancestors of SARS-CoV-1 and SARS-CoV-2 . . . 47
3.7 Sarbecoviruses spread at rates similar to their bat hosts . . . 47
3.8 Direct ancestors of the SARS-CoVs likely could not have reached sites of emergence via the bat reservoir alone . . . 49
3.9 Discussion . . . 50
3.10 Data and code availability . . . 53
3.11 Acknowledgments . . . 54
3.11.1 Author contributions . . . 55
3.11.2 Declaration of interests . . . 55

Epilogue . . . 56

Chapter A Supplemental Material for Chapter 1 . . . 58
A.1 Materials and Methods . . . 58
A.1.1 Sequence data . . . 58
A.1.2 Phylogenetic inference . . . 58
A.1.3 Epidemic Simulation . . . 59
A.1.4 Determining Stable Coalescence . . . 60
A.1.5 Sensitivity Analysis—Faster rate of infection . . . 61
A.1.6 Sensitivity Analysis—Slower rate of infection . . . 61
A.1.7 Sensitivity Analysis—Higher average degree . . . 61
A.1.8 Sensitivity Analysis—Lower average degree . . . 61
A.1.9 Sensitivity Analysis—Higher probability of hospitalization . . . 62
A.1.10 Sensitivity Analysis—Small village . . . 62
A.1.11 Sensitivity Analysis—Random contact network . . . 62
A.1.12 Sensitivity Analysis—Impact of subsampling on stable coalescence . . . 62
A.1.13 Two-Phase Epidemics in FAVITES . . . 63
A.1.14 Sensitivity Analysis—Varied transmission rates for Phase 1 . . . 63
A.1.15 Sensitivity Analysis—Higher ascertainment rate for Phase 1 . . . 63
A.1.16 Sensitivity Analysis—Higher two-phase probabilities of hospitalization . . . 63
A.1.17 Combining FAVITES and BEAST via Rejection Sampling . . . 64
A.1.18 Rejection Sampling — two-phase epidemics . . . 65

A.1.19 Sensitivity Analysis—ascertained and unascertained cases 65
A.1.20 Sensitivity Analysis—date of first COVID-19 case 65
A.1.21 Sensitivity Analysis—relaxed clock and coalescent priors 66
A.1.22 Sensitivity Analysis—shifting the tMRCA 66
A.2 Supplementary text 67
A.2.1 Relaxed molecular clock 67
A.2.2 Phylogenetic rooting 67
A.2.3 Time to stable coalescence in simulations 68
A.2.4 Epidemic extinction 68
A.2.5 Subsampling lineages at the stable coalescent 69
A.2.6 Varying the rate of transmission for Phase 1 of a two-phase epidemic 69
A.2.7 Varying the ascertainment rates for Phase 1 of the epidemic 69
A.2.8 Number of hospitalized individuals through time 70
A.3 Supplementary Tables 72
A.4 Supplementary Figures 76

Appendix B Supplemental Material for Chapter 2 94
B.1 Materials and Methods 94
B.1.1 Sequence data 94
B.1.2 Examining shared mutations between early C/C and T/T genomes and lineages A and B 95
B.1.3 Sequencing quality of early C/C and T/T genomes 95
B.1.4 C/C and T/T genomes in San Diego 95
B.1.5 T/T genomes in NYC 96
B.1.6 Constructing the recombinant common ancestor 96
B.1.7 recCA robustness analysis 97
B.1.8 Reversions early in the pandemic 97
B.1.9 Reversions throughout the pandemic 97
B.1.10 Phylogenetic inference 98
B.1.11 Random-effects substitution model 98
B.1.12 Quantifying the support for ancestral haplotypes 99
B.1.13 Phylogenetic inference with constrained roots and recCA 100
B.1.14 BEAST sensitivity analyses 100
B.1.15 Inferring empirical doubling times 101
B.1.16 Epidemic simulation 101
B.1.17 Doubling time inference 104
B.1.18 Sensitivity analysis–faster rate of infection 105
B.1.19 Sensitivity analysis–slower rate of infection 105
B.1.20 Sensitivity analysis–higher ascertainment rate 105
B.1.21 Sensitivity analysis–lower ascertainment rate 106
B.1.22 Determining stable coalescence 106
B.1.23 Determining the topological patterns of simulated phylogenies 106
B.1.24 Quantifying the support for two introductions against a single introduction of SARS-CoV-2 108
B.1.25 Combining epidemic simulations and BEAST via rejection sampling 111
B.1.26 Sensitivity analysis—earliest case date of 8 December 113
B.1.27 Sensitivity analysis—rejection sampling with hospitalization only 113

B.1.28 Sensitivity analysis—recCA and constrained roots . . . 113
B.1.29 Rejection sampling for lineages A and B . . . 113
B.1.30 Sensitivity analysis—earlier lineage B COVID-19 index case dates . . . 114
B.1.31 Simulating cross-species transmissions to achieve two successful introductions 114
B.2 Supplementary text . . . 115
B.2.1 Reduced genomic diversity in the early pandemic . . . 115
B.2.2 C/C and T/T genomes through 14 February 2020 . . . 115
B.2.3 T/T genomes aboard the Diamond Princess . . . 116
B.2.4 T/T genomes in New York City . . . 116
B.2.5 T/T and C/C genomes in San Diego . . . 117
B.2.6 The recombinant common ancestor (recCA) . . . 117
B.2.7 Justification for a non-reversible substitution model . . . 118
B.2.8 Sensitivity analyses of ancestral haplotype inference with unconstrained rooting . . . 118
B.2.9 Ancestral haplotype inference with outgroup rooting . . . 119
B.2.10 The tMCRA of SARS-CoV-2 is consistent across explored ancestral haplotypes . . . 119
B.2.11 Doubling time of the early pandemic . . . 121
B.2.12 Ascertainment rate of the early pandemic and parameterizing the simulations 123
B.2.13 Synthesizing evidence for multiple introductions of SARS-CoV-2 . . . 123
B.2.14 Index case dates of SARS-CoV-2 . . . 126
B.2.15 Robustness of lineage B primary case inference to date of index case . . . 127
B.2.16 Epidemiological dynamics of a single SARS-CoV-2 introduction scenario . 127
B.2.17 Consistent timing of the primary case when conditioning only on hospitalization . . . 128
B.2.18 Minimal cryptic circulation before December 2019 . . . 128
B.2.19 Similarities to WA1 and WA outbreak clades . . . 129
B.2.20 MRCA of multiple introductions and the possibility of intermediate genomes 129
B.2.21 Lineages of SARS-CoV-2 introductions . . . 130
B.3 Supplementary Tables . . . 132
B.4 Supplementary Figures . . . 139

Appendix C Supplemental Material for Chapter 3 . . . 170
C.1 Materials and Methods . . . 170
C.1.1 Dataset compilation . . . 170
C.1.2 Recombination analysis . . . 171
C.1.3 Comparison of breakpoint positions . . . 172
C.1.4 Molecular clock calibration . . . 173
C.1.5 Bayesian divergence time estimation . . . 174
C.1.6 Bayesian divergence time estimation sensitivity analyses . . . 175
C.1.7 Constructing the recombinant common ancestor . . . 175
C.1.8 Bayesian phylogeographic inference . . . 175
C.1.9 Ranking the dispersal velocity of the lineage descending from the closest-inferred ancestor . . . 177
C.1.10 Modeling of bat distributions . . . 177
C.1.11 Data and Code Availability . . . 179
C.2 Supplementary Tables . . . 180

C.3 Supplementary Figures 187

Bibliography 205

LIST OF FIGURES

Figure 1.1. Posterior distribution for the time of the most recent common ancestor (tMRCA) of 583 sampled SARS-CoV-2 genomes circulating in China between December 2019 and April 2020. 8

Figure 1.2. Hypothetical coalescent scenarios depicting how the time between index case infection and time of stable coalescence can vary based on stochastic extinction events of basal viral lineages. 9

Figure 1.3. Forward simulations estimating the timing of the index case in Hubei province. 11

Figure 1.4. Epidemic growth in compartmental simulations. 14

Figure 2.1. Maximum likelihood phylogeny of the early SARS-CoV-2 pandemic, showing nucleotide reversions and putative candidates for the ancestral haplotype at the most common recent ancestor (MRCA). 23

Figure 2.2. Probability of phylogenetic structures arising from a single introduction of SARS-CoV-2 in epidemic simulations. 26

Figure 2.3. Comparison of the tMRCA and primary case dates for lineage A and lineage B across rooting strategies. 27

Figure 2.4. Dynamics of simulated SARS-CoV-2 epidemics resulting from separate introductions of lineages A and B. 29

Figure 3.1. Recombination patterns and dating of the closest-inferred bat virus ancestor of SARS-CoV-1 and SARS-CoV-2 for each NRR. 41

Figure 3.2. The recCA (recombinant common ancestor) and the most closely related sarbecovirus over time. 44

Figure 3.3. Estimated evolutionary history of SARS-CoV-1-like and SARS-CoV-2-like viruses. 48

Figure A.1. Posterior distribution for the tMRCA of 583 sampled SARS-CoV-2 genomes circulating in China between December 2019 and April 2020 using different clock models 76

Figure A.2. Example of how coalescence of all sampled genomes can shift forward as basal lineages are lost and cease to propagate 77

Figure A.3. Shifting of tMRCA in empirical molecular clock analyses 78

Figure A.4. FAVITES robustness analyses 79

Figure A.5. Combined simulation and phylogenetic workflows to estimate the timing of the Hubei index case in a one-phase epidemic 80

Figure A.6. Illustration of the interplay rejection sampling scheme and the SAPHIRE compartmental model 81

Figure A.7. Robustness analysis for timing of SARS-CoV-2 index case in Hubei province 82

Figure A.8. Example phylogeny of the two-phase epidemic simulations 83

Figure A.9. Combined simulation and phylogenetic workflows to estimate the timing of the Hubei index case in a two-phase epidemic 84

Figure A.10. Robustness analysis for timing of SARS-CoV-2 index case in Hubei province in the two-phase epidemic simulations 85

Figure A.11. Robustness analysis for the number of people infected SARS-CoV-2 based on the SAPHIRE model in late 2019 86

Figure A.12. Robustness analysis for the number of people infected SARS-CoV-2 based on the SAPHIRE model in late 2019 using two-phase simulations 87

Figure A.13. Number of people hospitalized with SARS-CoV-2 based on the SAPHIRE model in late 2019 88

Figure A.14. Number of people hospitalized with SARS-CoV-2 based on the SAPHIRE model in late 2019 using two-phase simulations 89

Figure A.15. Bayesian skyline plot reconstruction 90

Figure A.16. Stable coalescence in forward compartmental simulations 91

Figure A.17. Impact of subsampling taxa from FAVITES epidemic simulations 92

Figure A.18. Impact of relative fitness and virulence of Phase 1 variant on the timing of the index case in Hubei province during two-phase epidemic simulations . . . 93

Figure B.1. Mutation map of SARS-CoV-2 intermediate C/C genomes and their shared mutations within lineages A and B. 139

Figure B.2. Maximum likelihood phylogeny of SARS-CoV-2 genomes from the Diamond Princess outbreak. 140

Figure B.3. Subtree showing the placement of T/T genomes in the NYC Public Health Laboratories dataset. 141

Figure B.4. Subtree showing the placement of T/T genomes in the San Diego SEARCH dataset. 142

Figure B.5. Subtree showing the placement of C/C genome in the San Diego SEARCH dataset. 143

Figure B.6. Reconstructing the recombinant common ancestor (recCA) of SARS-CoV-2 infecting a non-human animal. 144

Figure B.7. Simplot of closely related sarbecoviruses and recCA with Hu-1 as the reference. 145

Figure B.8. Maximum likelihood tree of non-recombinant region 2 with branches colored based on the nucleotide at position 2416. 146

Figure B.9. Maximum likelihood tree of non-recombinant region 8 with branches colored based on the nucleotide at position 19524. 147

Figure B.10. Maximum likelihood tree of non-recombinant region 11 with branches colored based on the nucleotide at position 23929. 148

Figure B.11. Maximum likelihood tree of non-recombinant region 5 with branches colored based on the nucleotide at position 8782. 149

Figure B.12. Maximum likelihood tree of non-recombinant region 8 with branches colored based on the nucleotide at position 18060. 150

Figure B.13. Maximum likelihood tree of non-recombinant region 14 with branches colored based on the nucleotide at position 28144. 151

Figure B.14. Maximum likelihood tree of non-recombinant region 15 with branches colored based on the nucleotide at position 29095. 152

Figure B.15. Maximum likelihood phylogenies of variants of concern (VOC) and variants of interest (VOI) with branches containing reversions colored. 153

Figure B.16. Subsampled global phylogeny showing reversions. 154

Figure B.17. Substitution process random-effects for the unconstrained rooting model. . . 155

Figure B.18. Schematic depicting the rooting strategies used in different phylodynamic models. 156

Figure B.19. SARS-CoV-2 maximum likelihood tree rooted on lineage A (n=787 taxa, through 14 February 2020). 157

Figure B.20. Substitution counts of SARS-CoV-2 genomes through 14 February 2020 from the root of the maximum likelihood tree when rooted on lineage A). 158

Figure B.21. Growth of the early pandemic. 159

Figure B.22. Inferred doubling times of simulated epidemics. 160

Figure B.23. Early SARS-CoV-2 introductions into Washington state. 161

Figure B.24. Dynamics of COVID-19 hospitalizations resulting from separate introductions of lineages A and B. 162

Figure B.25. The timing of the MRCA and primary case for lineage A and lineage B with a shorter doubling time. 163

Figure B.26. The timing of the MRCA and primary case for lineage A and lineage B with a longer doubling time. 164

Figure B.27. Dynamics of SARS-CoV-2 resulting from separate introductions of lineages A and B and a shorter doubling time. 165

Figure B.28. Dynamics of SARS-CoV-2 resulting from separate introductions of lineages A and B and a longer doubling time. 166

Figure B.29. Single-introduction timing of the MRCA and primary case and subsequent epidemic growth. 167

Figure B.30. Schematic depicting the multiple zoonotic origin of SARS-CoV-2. 168

Figure B.31. Rooting orientations of observed SARS-CoV-2 phylogenies resulting from different MRCAs and multiple introductions from the intermediate host. . . . 169

Figure C.1. Schematics illustrating the relationship between a SARS-CoV, the direct ancestor, and the closest-inferred ancestor 187

Figure C.2. Workflow to perform phylogenetic and phylogeographic analyses 188

Figure C.3. Mean and 90% confidence intervals of the log-normal breakpoint pair distance distributions for the observed SARS-CoV-1-like and SARS-CoV-2-like estimated breakpoints and the simulated randomly sampled breakpoint sets 189

Figure C.4. Substitution rates for the SARS-CoV-1-like and SARS-CoV-2-like viruses . . 190

Figure C.5. Time of the closest-inferred ancestor sensitivity analysis 191

Figure C.6. Genetic similarity of the NRRs of the recCA to each SARS-CoV over time . 192

Figure C.7. Prisoner of War (PoW) transformation comparison 193

Figure C.8. Time to the most recent common ancestor (tMRCA) after PoW transformation 194

Figure C.9. Continuous phylogeographic reconstruction of the dispersal history of viral lineages for the SARS-CoV-1-like viruses 195

Figure C.10. Continuous phylogeographic reconstruction of the dispersal history of viral lineages for the SARS-CoV-1-like viruses 196

Figure C.11. Continuous phylogeographic reconstruction of the dispersal history of viral lineages for the SARS-CoV-2-like viruses 197

Figure C.12. Continuous phylogeographic reconstruction of the dispersal history of viral lineages for the SARS-CoV-2-like viruses 198

Figure C.13. Distribution maps for species of the *Rhinolophus* genus 199

Figure C.14. Distribution maps for species of the *Rhinolophus* genus 200

Figure C.15. Distribution maps for species of the *Rhinolophus*, *Ascelliscus*, and *Chaerephon* genera 201

Figure C.16. Comparison of the continuous phylogeographic reconstructions for two selected non-recombinant regions (NRRs) when not considering or when considering the human and pangolin sequences in the analyses 202

Figure C.17. Comparison of the estimated position of the recCA according to continuous phylogeographic reconstructions that did not consider or did consider the human and pangolin sequences in the analyses 203

Figure C.18. Ranking of dispersal velocity for the closest-inferred bat virus ancestor to reach the province of emergence 204

LIST OF TABLES

Table 2.1. Posterior probabilities of inferred ancestral haplotype at the MRCA of SARS-CoV-2 24

Table A.1. Simulation parameters 72

Table A.2. Parameterization and success/failure of one-phase compartmental epidemic simulations in primary analysis and robustness analyses 72

Table A.3. Estimates for existence of lineages in one- and two-phase epidemics. 72

Table A.4. Number infected estimates for one-phase primary and robustness analyses . . 73

Table A.5. Number infected estimates for two-phase primary and robustness analyses . . 74

Table A.6. Parameterization and success/failure of two-phase compartmental epidemic simulations in primary analysis and robustness analyses. 74

Table A.7. FAVITES parameter values for the primary analysis 75

Table B.1. Nucleotide variant calls at positions 8782 and 28144 for three SARS-CoV-2 genomes with intermediate T/T haplotypes[1]. 132

Table B.2. Inferred tMRCAs for SARS-CoV-2, lineage B, and lineage A under different rooting strategies. 132

Table B.3. Simulation parameters, with all parameters except b and h based on Hao *et al.*[132]. The b value listed is for the main analysis. The h parameter is not included in the Hao *et al.* model. 132

Table B.4. The inferred doubling time for each transmission rate (b) and ascertainment rate (r) combination used in the simulations. 133

Table B.5. Frequencies of observed topologies in epidemic simulations and corresponding Bayes factor in favor of multiple introductions versus a single introduction across varying doubling times, varying ascertainment rate, minimum polytomy size, and phylogenetic rooting method. 133

Table B.6. Number of days the timing of the primary case of lineage A occurs after the timing of the primary case of lineage B. 134

Table B.7. Time of the lineage B primary case under different robustness analyses for different lineage B index case and hospitalization dates and conditioning only on hospitalization dates. 134

Table B.8. Time of the lineage B primary case under different robustness analyses for different doubling times and ascertainment rates. 135

Table B.9. Time of the lineage A primary case results when conditioning with either both the index and hospitalization dates or just the hospitalization date. . . . 135

Table B.10. Time of the lineage A primary case under different doubling times and ascertainment rates. 136

Table B.11. Number of estimated infections on 1 December 2019 for lineage A, lineage B, lineages A and B combined, and single introduction simulations. 136

Table B.12. Number of estimated infections on 15 December 2019 for lineage A, lineage B, lineages A and B combined, and single introduction simulations. 136

Table B.13. Number of infections at the tMRCA and hospitalizations on 1 December 2019. 137

Table B.14. Time of the primary case of a hypothetical single-introduction scenario under different robustness analyses for different index case dates and conditioning on only hospitalization dates. 137

Table B.15. Time of the primary case of a hypothetical single-introduction scenario under different doubling times and ascertainment rates. 138

Table C.1. Posterior clock rates across each NRR from the clock calibration using SARS-CoV-1 genomes and Bayesian divergence time analyses of SARS-CoV-1-like viruses 180

Table C.2. Posterior clock rates across each NRR from the clock calibration using SARS-CoV-2 genomes and Bayesian divergence time analyses of SARS-CoV-2-like viruses 182

Table C.3. Posterior times (median and 95% HPD) of the closest-inferred ancestor of SARS-CoV-1 across each NRR 184

Table C.4. Posterior times (median and 95% HPD) of the closest-inferred ancestor of SARS-CoV-2 across each NRR 185

Table C.5. Comparison of the weighted diffusion coefficient and isolation-by-distance signal estimated for different data sets of viral genomes 186

ACKNOWLEDGEMENTS

First and foremost, my deepest thanks to *Mom*, *Dad*, and *Eric*, for their unwavering love, support, and patience.

I want to thank all the mentors I have had along the way in graduate school, starting with my doctoral advisor, *Joel Wertheim*. When I began my PhD, I was unsure what direction I wanted my research to go in. From my first meeting with him, my direction and passion for research began to crystallize. He provided a supportive and intellectually rigorous environment, pushing me to identify what topics were not merely interesting, but also of social and global importance. He has since spent countless hours mentoring me, whether on how to properly communicate a singular idea or in how to be an ethical and thoughtful scientist. I could not be the researcher I am today without him.

Aside from Joel, I have been fortunate to study under the guidance of several other individuals. My research career began in the lab of *Jonathan Snow* while still at Williams College, and through his guidance, I eventually applied to Masters programs, landing at the Bioengineering program at UC San Diego. I am thankful to *Garrett Graham*, *Jeff Hasty*, *Jonathan Monk*, and *Bernhard Palsson* for teaching me when I knew almost nothing about computational biology. After transitioning to the Bioinformatics and Systems Biology program, I had the privilege of working with *Lucila Ohno-Machado* and *Tsung-Ting Kuo*, both of whom took the time to instill in me the importance of understanding the most critical stakeholder in my research: the public.

I would like to thank my doctoral committee: *Siavash Mirarab*, *Niema Moshiri*, *Tanya Vasylyeva*, and *Justin Meyer*. Each of you has provided invaluable support and insights. To *Michael Worobey*, my academic grandfather, for always pushing me to simplify, contextualize, and never lose sight of the bigger picture. I will forever aspire to your clarity of thought and sense of obligation toward understanding pandemic origins. I have also been fortunate to collaborate with many individuals around the world, more than I can name here. To *Maha Farhat*, *Kristian Andersen*, *Marc Suchard*, *Simon Dellicour*, *Philippe Lemey*, *Sébastien Calvignac-Spencer*, and *Fabian Leendertz*, for exposing me to and teaching me so much about other fields. I am grateful to be able to work with such thoughtful and brilliant researchers. Special thank you to *Jennifer Havens*, who, along with *Tanya*, helped create a wonderful environment in our (often virtual)

lab. I would also like to thank *Cyrus Maher* and *Amalio Telenti*, who supervised me during my internship at Vir Biotechnology. And to all the individuals whom I have had the opportunity to mentor as well; thank you for trusting me with this responsibility and for teaching me so much in return.

Thank you to the Siebel Foundation and Merkin Fellowship for their financial support and opportunities they provided.

To the people along the way at UC San Diego, friends in Bioinformatics and Systems Biology, Biomedical Informatics, and Bioengineering. Also, to the many wonderful people–friends, colleagues, and mentors–I met while visiting the Helmholtz Institute for One Health, in Greifswald, Germany, and especially *Hattie Herridge*, for making it feel like a home. I am lucky to have met and become friends with all of you.

Nafeesa Khan, *Gladys Ornelas*, *Clara Posner*, *Victor Cantu*, *Srihita Rudraraju*, *Sabah Ul-Hasan*, *Haley Gilbert*, and *Jingyi Liu*, for continued support, thoughtful discussions, and culinary adventures. *Dorothy Wang* and *Mary Channen Caldwell*, for your encouragement ever since I was at Williams. *Jamin Guo* and *Yu Wang*, for being great housemates. *Julie Kawasaki*, for teaching me about compassion and care, in all contexts.

And, lastly, to all the healthcare workers, researchers, and clinicians who were on the front lines of the COVID-19 pandemic. None of this research would have been possible without you. May we all be so courageous.

Chapter 1, in full, is a reformatted reprint of the material as it appears in "Timing the SARS-CoV-2 Index Case in Hubei Province"; Jonathan E. Pekar, Michael Worobey, Niema Moshiri, Konrad Scheffler, and Joel O. Wertheim. *Science*, 2021. The dissertation author was the primary investigator and first author of this paper.

Chapter 2, in full, is a reformatted reprint of the material as it appears in "The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2"; Jonathan E. Pekar, Andrew Magee, Edyth Parker, Niema Moshiri, Katherine Izhikevich, Jennifer L. Havens, Karthik Gangavarapu, Lorena Mariana Malpica Serrano, Alexander Crits-Christoph, Nathaniel L. Matteson, Mark Zeller, Joshua I. Levy, Jade C. Wang, Scott Hughes, Jungmin Lee, Heedo Park, Man-Seong Park, Katherine Ching Zi Yan, Raymond Tzer Pin Lin, Mohd Noor Mat Isa, Yusuf Muhammad Noor,

Tetyana I. Vasylyeva, Robert F. Garry, Edward C. Holmes, Andrew Rambaut, Marc A. Suchard, Kristian G. Andersen, Michael Worobey, and Joel O. Wertheim. *Science*, 2022. The dissertation author was the primary investigator and first author of this paper.

Chapter 3, in full, is a reformatted reprint of the material as it appears in "The recency and geographical origins of the bat viruses ancestral to SARS-CoV and SARS-CoV-2"; Jonathan E. Pekar, Spyros Lytras, Mahan Ghafari, Andrew F. Magee, Edyth Parker, Jennifer L. Havens, Aris Katzourakis, Tetyana I. Vasylyeva, Marc A. Suchard, Alice C. Hughes, Joseph Hughes, David L. Robertson, Simon Dellicour, Michael Worobey, Joel O. Wertheim, and Philippe Lemey. *bioRxiv*, 2023. The dissertation author was the primary investigator and co-first author of this paper.

VITA

2010–2014 B.A. in Biology
Williams College, Massachusetts

2016–2018 M.S. in Bioengineering
University of California San Diego

2018–2023 Ph.D. in Bioinformatics & Systems Biology with a Specialization in Biomedical Informatics
University of California San Diego

PUBLICATIONS

Author names marked with † indicate shared first co-authorship.
Publications marked with △ are included in this text.

Jonathan E. Pekar†, Spyros Lytras†, Mahan Ghafari, Andrew F. Magee, Edyth Parker, Jennifer L. Havens, Aris Katzourakis, Tetyana I. Vasylyeva, Marc A. Suchard, Alice C. Hughes, Joseph Hughes, David L. Robertson, Simon Dellicour, Michael Worobey, Joel O. Wertheim, and Philippe Lemey. "The recency and geographical origins of the bat viruses ancestral to SARS-CoV and SARS-CoV-2," *bioRxiv*, 2023. △

Andrew F Magee, Andrew J Holbrook, **Jonathan E. Pekar**, Itzue W Caviedes-Solis, Fredrick A Matsen IV, Guy Baele, Joel O Wertheim, Xiang Ji, Philippe Lemey, and Marc A Suchard. "Random-effects substitution models for phylogenetics via scalable gradient approximations," *arXiv*, 2023.

Evan A Boyle, Gabriela Goldberg, Jonathan C Schmok, Jillybeth Burgado, Fabiana Izidro Layng, Hannah A Grunwald, Kylie M Balotin, Michael S Cuoco, Keng-Chi Chang, Gertrude Ecklu-Mensah, Aleena KS Arakaki, Noorsher Ahmed, Ximena Garcia Arceo, Pratibha Jagannatha, **Jonathan Pekar**, Mallika Iyer, DASL Alliance, and Gene W Yeo. "Junior scientists spotlight social bonds in seminars for diversity, equity, and inclusion in STEM," *PlosOne*, 2023.

Alexander Crits-Christoph, Joshua I Levy, **Jonathan E. Pekar**, Stephen A Goldstein, Reema Singh, Zach Hensel, Karthik Gangavarapu, Matthew B Rogers, Niema Moshiri, Robert F Garry, Edward C Holmes, Marion PG Koopmans, Philippe Lemey, Saskia Popescu, Andrew Rambaut, David L Robertson, Marc A Suchard, Joel Wertheim, Angela L Rasmussen, Kristian G Andersen, Michael Worobey, and Florence Débarre. "Genetic tracing of market wildlife and viruses at the epicenter of the COVID-19 pandemic," *bioRxiv*, 2023.

Jonathan E. Pekar, Andrew Magee, Edyth Parker, Niema Moshiri, Katherine Izhikevich, Jennifer L. Havens, Karthik Gangavarapu, Lorena Mariana Malpica Serrano, Alexander Crits-Christoph, Nathaniel L. Matteson, Mark Zeller, Joshua I. Levy, Jade C. Wang, Scott Hughes, Jungmin Lee, Heedo Park, Man-Seong Park, Katherine Ching Zi Yan, Raymond Tzer Pin Lin,

Mohd Noor Mat Isa, Yusuf Muhammad Noor, Tetyana I. Vasylyeva, Robert F. Garry, Edward C. Holmes, Andrew Rambaut, Marc A. Suchard, Kristian G. Andersen, Michael Worobey, and Joel O. Wertheim. "The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2," *Science*, 2022. △

Michael Worobey, Joshua I Levy, Lorena Malpica Serrano, Alexander Crits-Christoph, **Jonathan E. Pekar**, Stephen A Goldstein, Angela L Rasmussen, Moritz UG Kraemer, Chris Newman, Marion PG Koopmans, Marc A Suchard, Joel O Wertheim, Philippe Lemey, David L Robertson, Robert F Garry, Edward C Holmes, Andrew Rambaut, and Kristian G Andersen. "The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic," *Science*, 2022.

Joel O Wertheim, Jade C Wang, Mindy Leelawong, Darren P Martin, Jennifer L Havens, Moinuddin A Chowdhury, **Jonathan E. Pekar**, Helly Amin, Anthony Arroyo, Gordon A Awandare, Hoi Yan Chow, Edimarlyn Gonzalez, Elizabeth Luoma, Collins M Morang'a, Anton Nekrutenko, Stephen D Shank, Stefan Silver, Peter K Quashie, Jennifer L Rakeman, Victoria Ruiz, Lucia V Torian, Tetyana I Vasylyeva, Sergei L Kosakovsky Pond, and Scott Hughes. "Detection of SARS-CoV-2 intra-host recombination during superinfection with Alpha and Epsilon variants in New York City," *Nature Communications*, 2022.

Jonathan E. Pekar, Michael Worobey, Niema Moshiri, Konrad Scheffler, and Joel O. Wertheim. "Timing the SARS-CoV-2 index case in Hubei province," *Science*, 2021. △

Michael Worobey, **Jonathan E. Pekar**, Brendan B Larsen, Martha I Nelson, Verity Hill, Jeffrey B Joy, Andrew Rambaut, Marc A Suchard, Joel O Wertheim, and Philippe Lemey. "The emergence of SARS-CoV-2 in Europe and North America," *Science*, 2020.

Britt Skaathun, Heather A Pines, Thomas L Patterson, Shirley J Semple, **Jonathan E. Pekar**, Alicia Harvey-Vera, Gudelia Rangel, and Sanjay R Mehta. "Recent HIV Infection among men who have sex with men and transgender women in Tijuana," *Revista de Saúde Pública*, 2020.

Henrique Machado, Liam L Weng, Nicholas Dillon, Yara Seif, Michelle Holland, **Jonathan E. Pekar**, Jonathan M Monk, Victor Nizet, Bernhard O Palsson, and Adam M Feist. "Strain-Specific Metabolic Requirements Revealed by a Defined Minimal Medium for Systems Analyses of *Staphylococcus aureus*," *Applied and Environmental Microbiology*, 2019.

Jonathan E. Pekar, Patrick Phaneuf, Richard Szubin, Bernhard Palsson, Adam Feist, and Jonathan M Monk. "Gapless, unambiguous genome sequence for Escherichia coli C, a workhorse of industrial biology," *Microbiology Resource Announcements*, 2018.

ABSTRACT OF THE DISSERTATION

The zoonotic origins and molecular epidemiology of SARS-CoV-2

by

Jonathan Pekar

Doctor of Philosophy in Bioinformatics & Systems Biology with a Specialization in Biomedical Informatics

University of California San Diego, 2023

Professor Joel O. Wertheim, Chair
Professor Siavash Mirarab, Co-Chair

Pandemics pose a persistent and substantial threat to human health, with consequences that include widespread mortality and suffering. The 21st century has thus far been most dramatically affected by the coronavirus disease 2019 (COVID-19) pandemic, and understanding its origins is critical in preventing future epidemics and pandemics. This dissertation investigates the emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of COVID-19. First, I present a novel approach coupling phylogenetics and epidemic modeling to infer when the primary case of the COVID-19 pandemic was infected. I then expand on this approach, incorporating additional epidemiological data and interrogating likely phylogenetic structures, to determine that the COVID-19 pandemic is likely the result of at least

two separate cross-species transmission events of SARS-CoV-2 into humans. Finally, I examine sarbecoviruses closely related to SARS-CoV and SARS-CoV-2 to reconstruct the temporal and geographic history of these viruses. Taken together, this dissertation outlines the events both preceding and at the start of the COVID-19 pandemic, identifying SARS-CoV-2 as a zoonotic pathogen, and implicating the wild and farmed animal trade in the emergence of SARS-CoV and SARS-CoV-2.

Introduction

In late December 2019, cases of coronavirus disease 2019 (COVID-19), the disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), started being documented in the city of Wuhan, China. This marked the beginning of a global pandemic that has resulted in over six million deaths as of 2023[1;2]. In the years since the first identified case of COVID-19, countless efforts have gone toward understanding this disease and the virus that causes it. A great deal has since been learned about the nature of this pandemic, including its incipience.

Understanding the origins of a pandemic is crucial for preventing future epidemics and pandemics. Modern human history has been marked by multiple zoonotic viral pandemics, including those caused by influenza A virus[3–5], human immunodeficiency virus (HIV)[6–9], and SARS-CoV (henceforth referred to as SARS-CoV-1)[10–13]. The exact events surrounding the start of these pandemics have historically been unknown, especially in the first months and years after they started[11;14;15], despite an increased consensus in recent years that a cross-species transmission of the virus from a non-human animal to a human was involved. The origins of these pandemics are often repeatedly analyzed as more data and improved methods become available, leading to a more complete understanding of their emergence.

The advent of sequencing technologies and bioinformatic methods has significantly improved our ability to understand the emergence and persistence of pathogens, with clinicians and researchers also becoming able to more rapidly identify pathogens responsible for outbreaks. Subsequent analyses of data collected at the onset of an outbreak and information related to the identified responsible pathogen allow for precise inferences regarding the evolutionary history of the pathogen. Although this has continued to prove challenging for most virus pandemics, there has been substantially more data produced regarding the COVID-19 pandemic, allowing for more detailed analyses of SARS-CoV-2 than any pathogen prior.

Phylogenetics has been instrumental in unraveling the evolutionary history of human viruses, including shedding light on their origins. However, phylogenetics is inherently unable to account for unsampled genomic diversity[16;17]. As virus genomes are less likely to be sampled and sequenced at the start of a pandemic, phylogenetics is limited in its ability to determine when exactly a virus might have spilled over into a human population. Rather, phylogenetics provides the ability to infer when the most recent common ancestor (MRCA) of sequenced virus genomes existed[18]. Meanwhile, epidemiological investigations of a virus can deduce its fundamental transmission dynamics without relying on any genomic data[19]. Such information can also serve as an additional data source in reconstructing the events at the start of the pandemic: number of cases, when the cases started to rapidly increase, etc. A basic understanding of the transmission dynamics of a virus can then be used for modeling its spread[20].

In Chapter 1, I discuss a novel approach integrating phylogenetics and epidemic modeling to infer when the first infection of SARS-CoV-2 occurred. I use Bayesian phylogenetics to first determine the time of the most recent common ancestor (tMRCA) of SARS-CoV-2 and then epidemic simulations to reconstruct the transmission history of SARS-CoV-2 immediately after its emergence. By subsequently linking the tMRCA with the epidemic simulations, I show that, despite the tMRCA of SARS-CoV-2 being in late November-to-early December, the primary case[1] of SARS-CoV-2 was likely infected in November. I then examine the effects of transmissibility and virulence on the spread of SARS-CoV-2 and demonstrate that a substantial selective event or fitness increase was not necessary for SARS-CoV-2 to persist in the human population after it began spreading.

As more data is collected and retrospective surveys are completed, additional data streams can be used to inform our understanding of the emergence of a novel pathogen. Retrospective investigations, such as those performed by individual research labs[20;23–25] as well as the World Health Organization[26], have helped curate case counts and virus sequencing data related to SARS-CoV-2 more accurately. In particular, early SARS-CoV-2 genomes were found to have

[1]In Chapter 1, I refer to this individual as the index case, as that is how we referred to it when we published the paper. The index case, however, is the first case identified by health authorities, whereas the primary case is the infected individual that brings the disease to a larger group, leading to subsequent transmissione[21]. In Chapter 2, I correctly refer to this individual as the primary case. Occasionally, the index case is the primary case, but the primary case will often not be found. The index case of SARS-CoV-2 is likely an individual from the Huanan Seafood Market whose symptoms started on 8 December, 2019[22].

erroneous mutations, potentially affecting the inference of the tMRCA and, therefore, the timing of the primary case[26].

In Chapter 2, I discuss my examination of the early genomic SARS-CoV-2 data and how the genomes intermediate to lineages A and B of SARS-CoV-2 are likely identified as such due to erroneous base calls as well. This finding led me to investigate whether there were, in fact, multiple introductions of SARS-CoV-2 into the human population. I again couple Bayesian phylogenetics and epidemic simulations to reconstruct the early dynamics of the COVID-19 pandemic, extending our prior approach to more fully account for hospitalization and phylogenetic patterns. By comparing simulated phylogenies resulting from the epidemic modeling with the observed phylogeny of SARS-CoV-2, I demonstrate that the observed phylogeny is likely the product of at least two introductions of SARS-CoV-2 into humans. My results additionally indicate there was minimal cryptic transmission of SARS-CoV-2 before the earliest identification of COVID-19 cases, with the primary cases of the two introductions likely occurring in November 2019.

Importantly, surveillance systems for infectious disease tracking must extend beyond human pathogens. Investigating the evolutionary history of a zoonotic pathogen requires sampling closely related pathogens found in animals, a process that can easily take years, if not decades, especially if either the ultimate reservoir or likely intermediate hosts are unknown. Fortunately, recent years have seen an increase in research dedicated to discovering such pathogens in non-human animals. For example, coronaviruses[27;28], orthopoxviruses[29], and retroviruses[30;31] closely related to those found in humans have been identified, contributing significantly to our ability to reconstruct their history and evolutionary patterns.

Horseshoe bats are the ultimate reservoir of sarbecoviruses[32] and, likely via zoonotic transmission through a transient intermediate species, were the source of SARS-CoV-1 and SARS-CoV-2[33–35] (jointly referred to as the SARS-CoVs). Since the emergence of SARS-CoV-1 in Guangdong Province and SARS-CoV-2 in Hubei Province, there has also been increased sampling of sarbecoviruses closely related to the SARS-CoVs, allowing for more fine-grained analyses of the geographic and temporal history of these viruses. In Chapter 3, I examine the SARS-CoVs and sarbecoviruses closely related to them to reconstruct their deeper evolutionary

history. Employing recombination-aware Bayesian phylogenetics, I reconstruct when the MRCA of each SARS-CoV and its most closely related bat virus, or set of viruses, circulated. We refer to this MRCA as the 'closest-inferred' ancestor and this inference sheds light on how recently an ancestor of each SARS-CoV circulated in bats prior its emergence, based on the available data. I then perform phylogeographic analyses to understand the historical circulation of these viruses and show that the ancestors of SARS-CoV-1-like viruses and SARS-CoV-2-like viruses are tens-to-hundreds of thousands of years old. Lastly, by examining where the closest-inferred ancestors of SARS-CoV-1 and SARS-CoV-2 circulated, along with how quickly they would have had to have traveled to reach the provinces of emergence for SARS-CoV-1 and SARS-CoV-2, respectively, I demonstrate that these virus lineages likely could not have reached the provinces of emergence via bat movement alone. These findings, coupled with the results from Chapter 2, implicate the wild and farmed animal trade in the emergence of SARS-CoV-2, aligning with the established understanding of SARS-CoV-1.

The integration of diverse data streams, including case counts, hospitalization rates, serology surveys, and genomics, enables a comprehensive analysis of pandemic origins. Because of the vast amount of data produced, both in real-time and retrospectively, our understanding of the evolutionary history of SARS-CoV-2 is arguably greater than that of any other virus. Using a combination of genomics and epidemiology, I first show how we can identify when SARS-CoV-2 likely infected the primary case of the pandemic. By then folding in hospitalization rates, refined case counts, and interrogating whether the observed phylogeny of SARS-CoV-2 can be recapitulated by simulated phylogenies resulting from a single introduction of SARS-CoV-2, I demonstrate that the COVID-19 pandemic was likely the result of at least two, if not more, zoonotic transmissions of SARS-CoV-2 and that there was minimal cryptic circulation before the identification of the first COVID-19 cases. And lastly, by analyzing SARS-CoV-1, SARS-CoV-2, and sarbecoviruses closely related to them, I reconstruct the geographic and temporal history of sarbecoviruses more broadly and that the emergence of the human SARS-CoVs was likely linked to the wild and farmed animal trade. These findings have elucidated, in particular, the emergence and evolutionary history of SARS-CoV-2 and contributed to our understandings of pandemic emergence more broadly.

Chapter 1

Timing the SARS-CoV-2 Index Case in Hubei Province

Understanding when SARS-CoV-2 emerged is critical to evaluating our current approach to monitoring novel zoonotic pathogens and understanding the failure of early containment and mitigation efforts for COVID-19. We employed a coalescent framework to combine retrospective molecular clock inference with forward epidemiological simulations to determine how long SARS-CoV-2 could have circulated prior to the time of the most recent common ancestor. Our results define the period between mid-October and mid-November 2019 as the plausible interval when the first case of SARS-CoV-2 emerged in Hubei province. By characterizing the likely dynamics of the virus before it was discovered, we show that over two-thirds of SARS-CoV-2-like zoonotic events would be self-limited, dying out without igniting a pandemic. Our findings highlight the shortcomings of zoonosis surveillance approaches for detecting highly contagious pathogens with moderate mortality rates.

1.1 Introduction

In late-December of 2019, the first cases of COVID-19, the disease caused by SARS-CoV-2, were described in the city of Wuhan in Hubei province, China[1;2]. The virus quickly spread within China[36]. The cordon sanitaire that was put in place in Wuhan on 23 January 2020 and mitigation efforts across China eventually brought about an end to sustained local transmission. In March and April 2020, restrictions across China were relaxed[37]. By then, however, COVID-19 was a pandemic[38].

A concerted effort has been made to determine when the virus first began transmitting among humans by retrospectively diagnosing the earliest cases of COVID-19. Both epidemiological and phylogenetic approaches suggest an emergence of the pandemic in Hubei province at some point in late-2019[2;39;40]. The first described cluster of COVID-19 was associated with the Huanan Seafood Wholesale Market in late-December 2019, and the earliest sequenced SARS-CoV-2 genomes came from this cluster[41;42]. However, this market cluster is unlikely to have marked the beginning of the pandemic, as COVID-19 cases from early December lacked connections to the market[40]. The earliest such case in the scientific literature is from an individual retrospectively diagnosed on 01 December 2019[39]. Notably, however, newspaper reports document retrospective COVID-19 diagnoses recorded by the Chinese government going back to 17 November 2019 in Hubei province[43]. In fact, these reports detail daily retrospective COVID-19 diagnoses through the end of November, suggesting that SARS-CoV-2 was actively circulating for at least a month before it was discovered.

Molecular clock phylogenetic analyses have inferred the time of most recent common ancestor (tMRCA) of all sequenced SARS-CoV-2 genomes to be in late November or early December 2019, with uncertainty estimates typically ranging back to October 2019[40;44;45]. Crucially, though, this tMRCA is not necessarily equivalent to the date of zoonosis or index case infection[16;17], because coalescent processes can prune basal viral lineages before they have the opportunity to be sampled, potentially pushing SARS-CoV-2 tMRCA estimates forward in time from the index case by days, weeks, or months. For a point of comparison, consider the zoonotic origins of the HIV-1 pandemic, whose tMRCA in the early 20th century coincides with the urbanization of Kinshasa, in the Congo[6;7], but whose cross-species transmission from a chimpanzee reservoir occurred in southeast Cameroon, likely predating the tMRCA of sampled HIV-1 genomes by many years[46]. Despite this important distinction, the tMRCA has been frequently conflated with the date of the index case infection in the SARS-CoV-2 literature[40;47;48].

Here, we combine retrospective molecular clock analysis in a coalescent framework with a forward compartmental epidemiological model to estimate the timing of the SARS-CoV-2 index case in Hubei province. The inferred dynamics during these unobserved early days of SARS-CoV-2 highlight challenges in detecting and preventing nascent pandemics.

1.2 Results

We first explored the evolutionary dynamics of the first wave of SARS-CoV-2 infections in China. We used Bayesian phylodynamics[49] to reconstruct the underlying coalescent processes using a Bayesian Skyline approach for 583 SARS-CoV-2 complete genomes, sampled in China between when the virus was first discovered at the end of December 2019 and the last of the non-reintroduced circulating virus in April 2020. Applying a strict molecular clock, we inferred an evolutionary rate of 7.90×10^{-4} substitutions/site/year (95% highest posterior density [HPD]: 6.64×10^{-4}–9.27×10^{-4}). The tMRCA of these circulating strains was inferred to fall within a 34-day window with a mean of 09 December 2019 (95% HPD: 17 November–20 December) (Fig. 1.1). This estimate accounts for the many disparate inferred rooting orientations [see Supplementary Material and described previously[50]]. Notably, 78.7% of the posterior density post-dates the earliest published case on 01 December, and 95.1% post-dates the earliest reported case on 17 November. Relaxing the molecular clock provides a similar tMRCA estimate, as does applying a Skygrid coalescent approach (Fig. A.1). The recency of this tMRCA estimate in relation to the earliest documented COVID-19 cases obliges us to consider the possibility that this tMRCA does not capture the index case and that SARS-CoV-2 was circulating in Hubei province prior to the inferred tMRCA.

If the tMRCA post-dates the earliest documented cases, then the earliest diverged SARS-CoV-2 lineages must have gone extinct (Fig. 1.2). As these early basal lineages disappeared, the tMRCA of the remaining lineages would move forward in time (Fig. A.2). Thus, we interrogated the posterior trees sampled from the phylodynamic analysis to determine if this time of coalescence had stabilized prior to the sequencing of the first SARS-CoV-2 genomes on 24 December 2019 or if this process of basal lineage loss was ongoing in late-December/early-January. Importantly, these basal lineages need not be associated with specific mutations, as the phylodynamic inference reconstructs the coalescent history, not the mutational history[49].

We find only weak evidence for basal lineage loss between 24 December 2019 and 13 January 2020 (Fig. A.3A). The root tMRCA is within 1 day of the tMRCA of virus sampled on or after 01 January 2020 in 78.5% of posterior samples (Fig. A.3B). The tMRCA of genomes

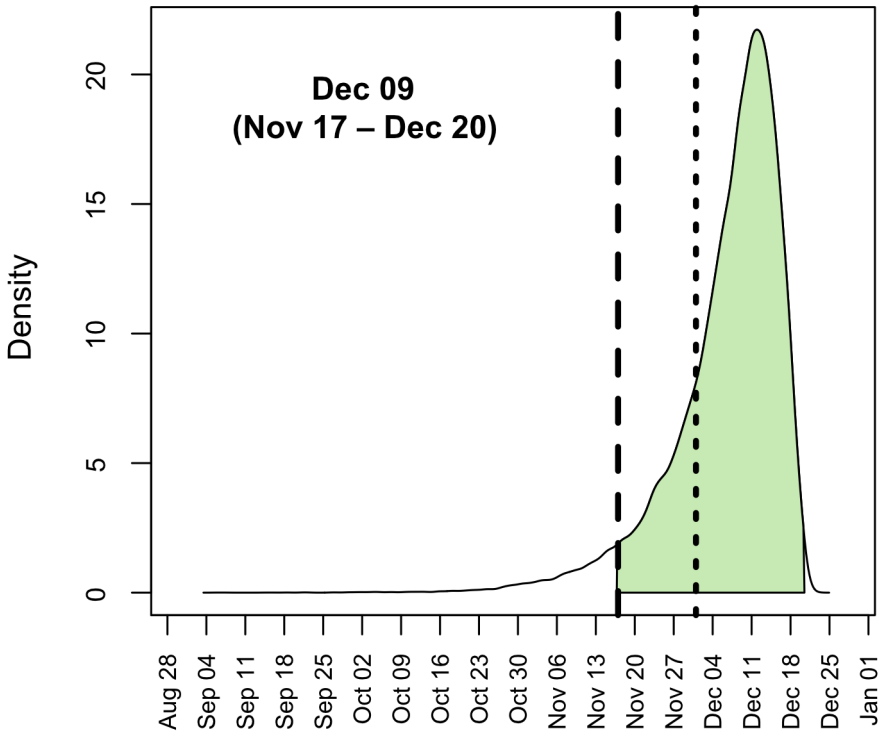


Figure 1.1. Posterior distribution for the time of the most recent common ancestor (tMRCA) of 583 sampled SARS-CoV-2 genomes circulating in China between December 2019 and April 2020. Inference was performed using a strict molecular clock and a Bayesian Skyline coalescent prior. Shaded area denotes 95% HPD. Long-dashed line is 17 November 2019, and short-dashed line is 01 December 2019.

sampled on or after 01 January 2020 is 3 days younger than the tMRCA of all sampled genomes. In contrast, the mean tMRCA does not change when considering genomes sampled on or after 01 January 2020 versus on or after 13 January 2020. This consistency indicates a stabilization of coalescent processes at the start of 2020 when there had been an estimated 1000 total people infected with SARS-CoV-2 in Wuhan[20]. Nonetheless, to account for the weak signal of a delay in reaching a stable coalescence (i.e., the point in time at which basal lineages cease to be lost), we identified the tMRCA for all viruses sampled on or after 01 January 2020 (i.e., at the time of stable coalescence) for each tree in the posterior sample.

Phylogenetic analysis alone cannot tell us how long SARS-CoV-2 could have circulated in Hubei province before the tMRCA. To answer this question, we performed forward epidemic simulations[51]. These simulations were initiated by a single index case using a compartmental

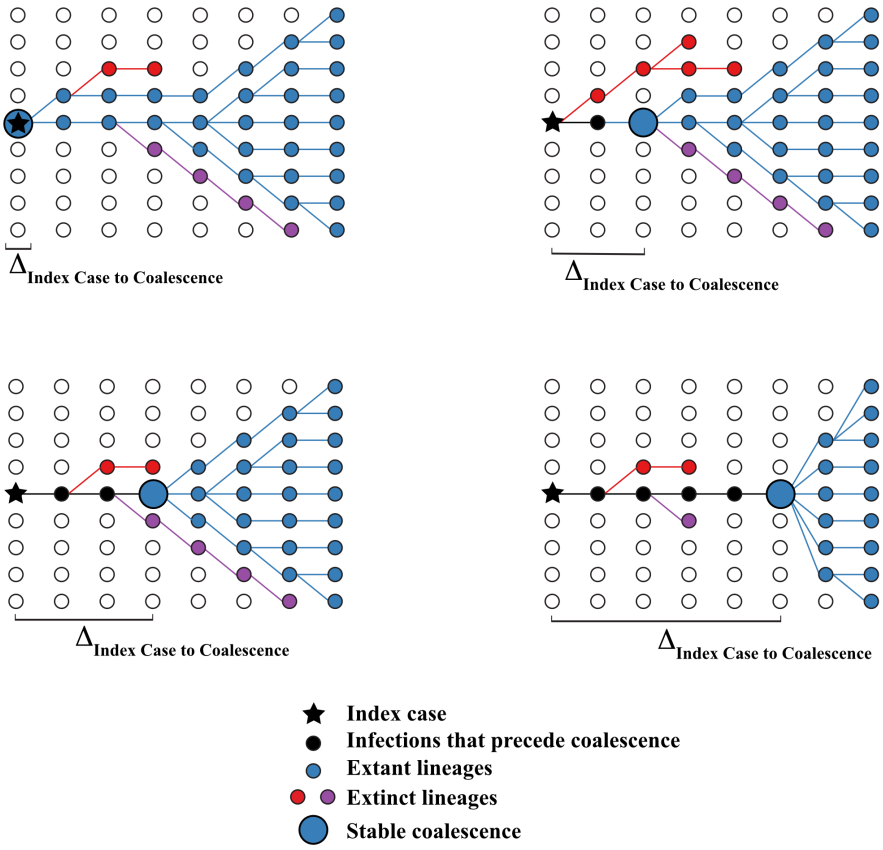


Figure 1.2. Hypothetical coalescent scenarios depicting how the time between index case infection and time of stable coalescence can vary based on stochastic extinction events of basal viral lineages. Coalescence can occur within or contemporaneously with the index case (upper left) or in cases infected later in the course of the epidemic with one (upper right) or more (lower left) basal lineages going extinct. In extreme cases, the epidemic can persist at low levels for a long time before stable coalescence (lower right).

epidemiological model across scale-free contact networks (mean number of contacts=16). This compartmental model was previously developed to describe SARS-CoV-2 transmission dynamics in Wuhan[20]. This model, termed SAPHIRE, includes compartments for susceptible (S), exposed (E), presymptomatic (P), unascertained (A), ascertained (I), hospitalized (H), and removed (R) individuals. Our simulations used parameters from the time period prior to COVID-19 mitigation efforts, from 01 January through 22 January 2020 (Table A.1), based on Hao et al.[20]. We analyzed 1000 epidemic simulations that resulted in ≥1000 total infected people. These simulated epidemics had a median doubling time of 4.1 days (95% range across simulations: 2.7-6.7), matching pre-mitigation incidence trends in Wuhan (Table A.2).

We simulated coalescent processes across the transmission network to determine the tMRCA of the virus at the end of the simulation. This approach allowed us to determine the distribution of the expected number of days between index case infection and the stable coalescence (i.e., tMRCA) (Fig. 1.2). The median number of days between index case infection and this tMRCA was 8.0 days (95% range: 0.0 and 41.5 days) (Fig. 1.3A). The median time between index case infection and the first person exiting the presymptomatic phase (i.e., ascertained or unascertained infection) was 5.7 days (95% range: 0.9 to 15.7 days).

As a robustness check, we also simulated epidemics with more densely and more sparsely connected contact networks (mean 26 and 10 contacts, respectively), rescaling the per-contact transmission rate to maintain empirical epidemic growth dynamics. We also explored the effects of faster (mean: 3.1 days; 95% range: 2.0 to 5.1 days) and slower (mean: 5.3 days; 95% range: 3.6 to 7.5 days) epidemic doubling times (Table A.2). Slower transmission rates led to more days between the index case and the stable coalescence, but modifying the density of the contact network had minimal effect on this time interval (Fig. A.4).

To estimate the date of infection for the index case in Hubei province, we combined the retrospective molecular clock analysis with the forward epidemic simulations (Fig. A.5). We identified the stable tMRCA in the posterior trees as an anchor to the real-world calendar dates and then extended this date back in time according to the number of days between the index case infection and the time of stable coalescence from the compartmental epidemic simulations. However, a random sample of tMRCAs and days from index case infection to coalescence will not produce epidemiologically meaningful results, because many of these combinations do not precede the earliest dates of reported COVID-19 cases. Therefore, we implemented a rejection sampling-based approach to generate a posterior distribution of dates of infection for the Hubei index case, conditioning on at least one individual who has progressed past the pre-symptomatic stage in the simulated epidemic before the date of the first reported COVID-19 case (see Methods; Fig. A.6).

In our primary analysis, we assume that 17 November represents the first documented case of COVID-19 (ascertained or unascertained in the SAPHIRE model). Under this assumption, the median number of days between index case infection and stable coalescence after rejection

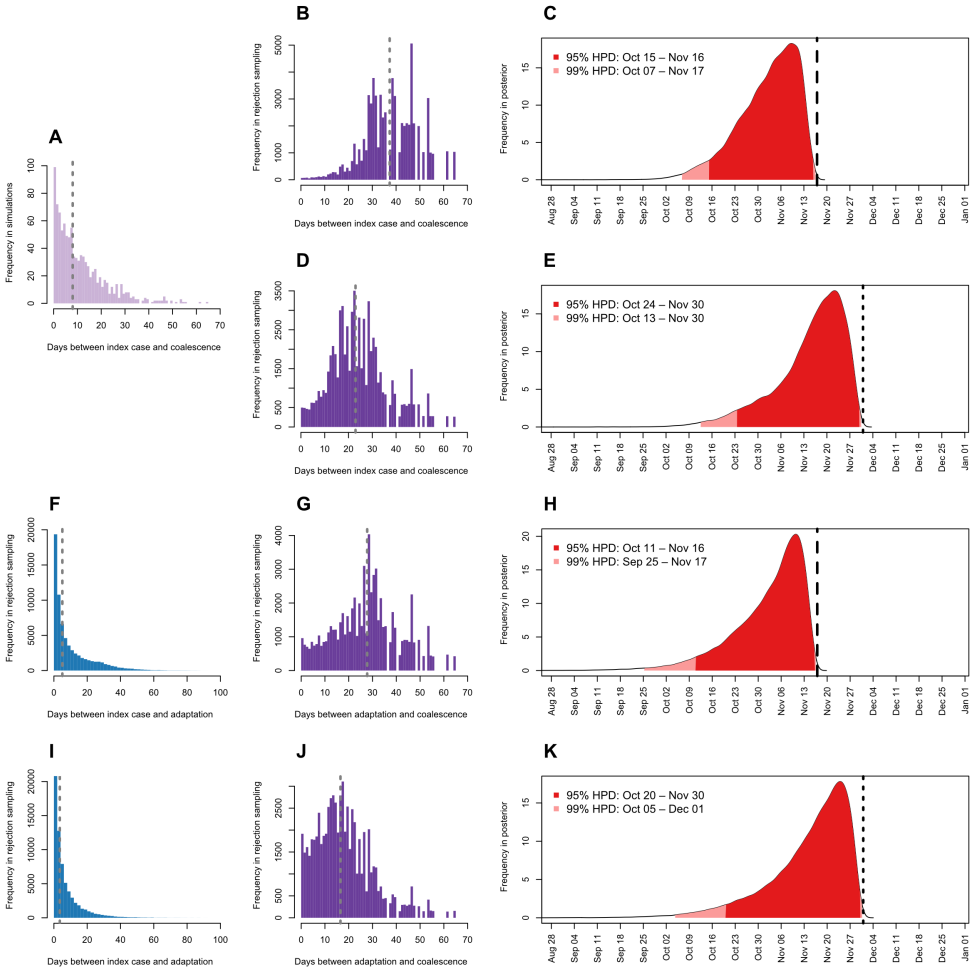


Figure 1.3. Forward simulations estimating the timing of the index case in Hubei province. (**A**) Days between index case infection and stable coalescence in forward compartmental epidemic simulations (n=1000). (**B**) Days between index case infection and stable coalescence after rejection sampling and (**C**) posterior distribution for date of index case infection, conditioned on an ascertained case by 17 November 2019 denoted by long-dashed line. (**D-E**) Epidemic simulation, conditioned on an ascertained case by 01 December 2019 denoted by short-dashed line. Two-phase epidemic (**F**) days between index case carrying less fit variant and adaptation (n=2000), (**G**) days between adaptation and stable coalescence (n=1000), and (**H**) posterior distribution for date of index case infection, conditioned on an ascertained case by 17 November 2019. (**I-K**) Two-phase epidemic conditioned on an ascertained case by 01 December 2019. Grey dashed lines indicate median estimates.

sampling is 37 days (95% HPD: 12 to 55 days) (Fig. 1.3B). Consequently, the index case in Hubei likely contracted SARS-CoV-2 on or around 04 November 2019 (95% upper HPD on 15 October; 99% upper HPD on 07 October) (Fig. 1.3C).

This timeframe for the Hubei index case is robust (Fig. A.7). Epidemic simulations

with faster or slower transmission rates and more or less densely connected contact networks produce similar date estimates. Further, using the root tMRCA from the sampled posterior trees, rather than adjusting for the shifting coalescence between 24 December 2019 and 01 January 2020, produces a median date of 03 November (95% upper HPD 14 October; 99% upper HPD 04 October). Incorporating a relaxed molecular clock or adjusting the coalescent prior assumption (Skyline vs. Skygrid) also had minimal effect.

If we enforce that there be at least one ascertained case in our simulations prior to 17 November 2019, the median date of the Hubei index case is pushed back about a week to 28 October (95% upper HPD 11 October; 99% upper HPD 05 October) (Fig. A.7). However, the distinction between ascertained and unascertained in the original SAPHIRE model was meant to reflect the probability of missed diagnoses in January 2020 of the Wuhan epidemic and does not account for the investigations that resulted in retrospective diagnoses in November and December 2019.

If we discount the reported evidence of retrospective COVID-19 diagnoses throughout the end of November and instead take 01 December as representing the first confirmed case of COVID-19, then the median time between index case infection and stable coalescence after rejection sampling is 23 days (95% range: 1 to 47 days) (Fig. 1.3D). Under this scenario, the index case in Hubei would have contracted SARS-CoV-2 on or around 17 November 2019 (95% upper HPD 24 October; 99% upper HPD 13 October) (Fig. 1.3E). Similar dates are inferred with a relaxed molecular clock and conditioning on an ascertained infection by 01 December (Fig. A.7).

It is reasonable to postulate that the variant of SARS-CoV-2 that first emerged was less fit than the variant that spread through China and that evolutionary adaptation was critical to its establishment in humans[45]. Therefore, we simulated two-phase epidemics wherein the index case was infected with a less fit variant (i.e., half as transmissible) that went extinct but not before giving rise to a mutant strain matching the transmission dynamics estimated in Wuhan (Fig. A.8, A.9; Supplementary Material). If we condition on an ascertained or unascertained COVID-19 case (due to either the original or adapted variant) by 17 November, the original variant transmits for a median of 5 days (95% HPD: 0-36 days) before the adapted strain emerges

(Fig. 1.3F); this adapted strain then circulates for a median of 28 days (95% HPD: 0-52 days) before reaching a stable coalescence (Fig. 1.3G). In this scenario, the index case in Hubei would have likely contracted an unobserved variant of SARS-CoV-2 on 05 November 2019 (95% upper HPD on 11 October; 99% upper HPD on 25 September) (Fig. 1.3H). If we again discount the reported evidence COVID-19 in November and take 01 December as the first confirmed case, then the virus would spend less time in humans across both phases of the early epidemic (Fig. 1.3I to J) and the index case in Hubei would have acquired SARS-CoV-2 on 18 November 2019 (95% upper HPD on 20 October; 99% upper HPD on 05 October) (Fig. 1.3K). As in the primary analysis, the inferred date of the index case in the two-phase epidemic was robust to varying model assumptions (Fig. A.10), including the amount by which viral fitness differed between the two phases (Fig. A.18; Supplementary Material).

These one-phase and two-phase epidemic simulations both suggest that the tMRCA is not representative of the emergence of SARS-CoV-2. In the primary analysis, the index case remains infected at the tMRCA (as in Fig. 1.2, upper right panel) in less than 1.1% of simulated epidemics (Table A.3). In the two-phase epidemics, the index case is still infected at the tMRCA in 2.2% of simulations, likely due to an increased variance in the time between index case and tMRCA. The initial, less fit variant persisted until the tMRCA in 17% of simulated epidemics and until 01 January 2020 in 3.7% of simulated epidemics. However, when this less fit variant persisted, it was represented by only a single infected individual at the tMRCA; this low frequency suggests that even if this less fit variant did exist, it could have easily been missed in early genome sequencing efforts. In the two-phase epidemic, the first ascertained (or unascertained) case was due to the less fit variant in around two-thirds of simulated epidemics.

By anchoring our epidemic simulations to specific tMRCA estimates, we can reconstruct a plausible range for the number of SARS-CoV-2 infections prior to the discovery of the virus (Fig. 1.4A). The median number of individuals infected with SARS-CoV-2 in our primary analysis is below 1 until 04 November. On 17 November, the median number infected is 4 individuals (95% HPD: 1-13) and reaches 9 (95% HPD: 2-26) on 01 December. These values are generally robust to model specifications, molecular clock method, and date of first COVID-19 case (Table A.4; Fig. A.11). The two-phase epidemics tend to exhibit similar growth patterns (Fig. 1.4B;

Fig. A.12; Table A.5). Notably, we do not see any evidence for an increase in hospitalizations until mid-to-late December, even when we increase the virulence of the less-fit variant in the two-phase epidemics or increase the probability of hospitalization prior to the stable coalescence (Supplementary Material; Fig. A.13, A.14).

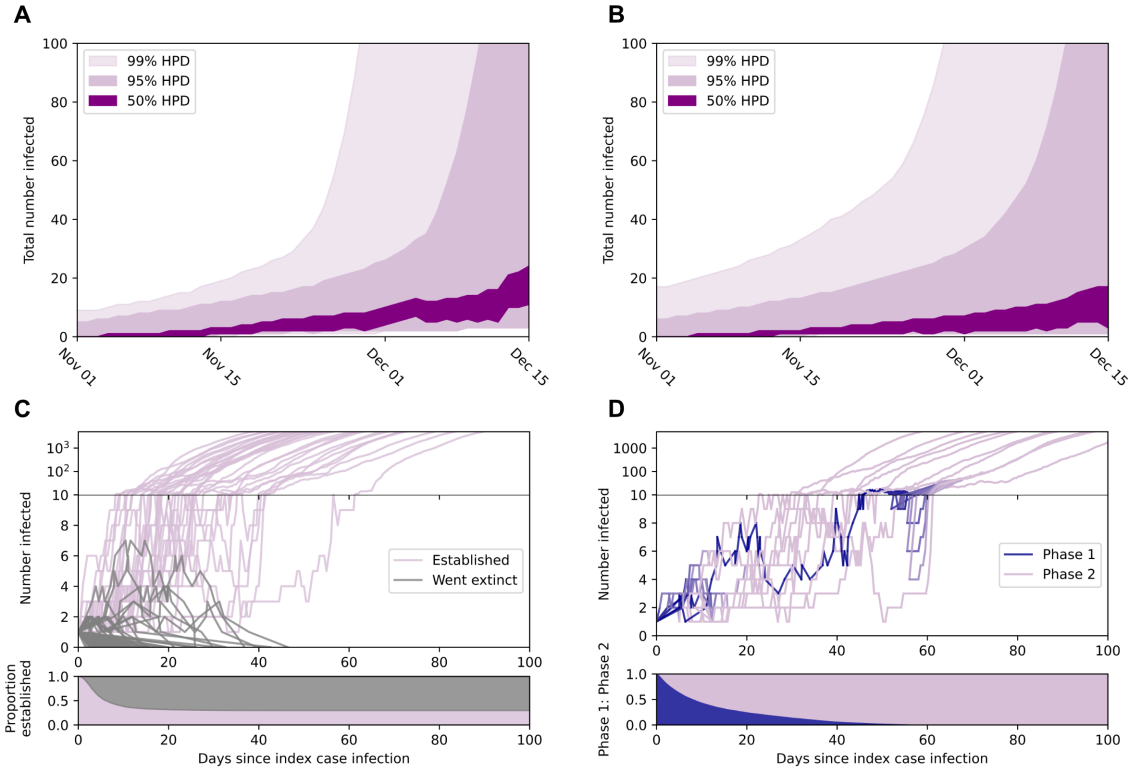


Figure 1.4. Epidemic growth in compartmental simulations. (**A**) Total estimated people infected in late-2019. Dark purple shading is central 50% HPD, intermediate purple shading is central 95% HPD, and light purple is central 99% HPD. (**B**) Total estimated people infected in late-2019 for a two-phase epidemic. (**C**) The number of people infected over time in a sample of epidemic simulations that established (purple; n=30) and went extinct (grey; n=70). The y-axis transitions to log-scale once ≥10 people are infected at any given time. The lower panel shows the proportion of simulations that still have ≥1 infected individual over time (persisting epidemics in purple; extinct epidemics in grey). (**D**) A sample (n=10) of two-phase epidemic simulations transitioning from less-fit Phase 1 (blue) to more-fit Phase 2 (purple). Each line represents a single simulation and its transition over time. The lower panel shows the average proportion of Phase 1 to Phase 2 infected individuals over time.

Empirical observation throughout the SARS-CoV-2 pandemic has shown the outsized role of superspreading events in the propagation of SARS-CoV-2[52–55], wherein the average infected person does not transmit the virus. Our results suggest the same dynamics likely influenced the

initial establishment of SARS-CoV-2 in humans, as only 29.7% of simulated epidemics from the primary analysis went on to establish self-sustaining epidemics. The remaining 70.3% of epidemics went extinct (Fig. 1.4C). Epidemics that went extinct typically produced only 1 infection (95% range: 1-9) and never more than 44 infections total or 14 infections at any given time (Table A.2). The median failed epidemic went extinct by day 8. As the contact network became more or less densely connected, the number of epidemics that went extinct was similar: 68.3% and 69.4%, respectively (Table A.2). However, the percentage of extinct epidemics increased as the transmission rate decreased (80.5%) and decreased as the transmission rate increased (53.6%). In the two-phase epidemic, this original less fit variant went extinct by day 9 (95% HPD: 2-52) and produced a median of 1 infection (95% HPD: 1-13) (Fig. 1.4D; Table A.6).

The overdispersed nature of SARS-CoV-2 transmission patterns favors its persistence, as epidemics simulated over random contact networks[55] that are not characterized by superspreading events tended to go extinct more frequently: 83.7% of the time. Furthermore, the large and highly connected contact networks characterizing urban areas seem critical to the establishment of SARS-CoV-2. When we simulated epidemics where the number of connections was reduced by 50% or 75% (without rescaling per-contact transmissibility), to reflect emergence in a rural community, the epidemics went extinct 94.5% or 99.6% of the time, respectively.

1.3 Discussion

Our results highlight the unpredictable dynamics that characterized the earliest days of the COVID-19 pandemic. The successful establishment of SARS-CoV-2 post-zoonosis was far from certain, as more than two-thirds of simulated epidemics quickly went extinct. It is highly probable that SARS-CoV-2 was circulating in Hubei province at low levels in early-November 2019 and possibly as early as October 2019, but not earlier. Nonetheless, the inferred prevalence of this virus was too low to permit its discovery and characterization for weeks or months. By the time COVID-19 was first identified, the virus had firmly established itself in Wuhan. This delay highlights the difficulty in surveillance for novel zoonotic pathogens with high transmissibility and moderate mortality rates.

The high extinction rates we inferred suggest that spillover of SARS-CoV-2-like viruses

may be frequent, even if pandemics are rare[56]. Further, the same dynamics that characterized the establishment of SARS-CoV-2 in Hubei province may have played out all over the world as the virus was repeatedly introduced, but only occasionally took hold[57;58]. The reports of cases in December 2019 and January 2020 in France and California that did not establish sustained transmission fit this pattern[59–61]. However, our results suggest that PCR evidence of SARS-CoV-2 in wastewater outside of China before November 2019 is unlikely to be valid[62] and the suggestion of international spread in mid-November or early-December 2019 should be viewed with skepticism[63–65], given that our results suggests fewer than 20 people infected with SARS-CoV-2 at this time (Table A.4; Fig. A.11). Our results also refute claims[66] of large numbers of patients requiring hospitalization due to COVID-19 in Hubei province prior to December 2019 (Fig. A.13, A.14). Nevertheless, SARS-CoV-2 may be detectable should archived wastewater samples or other biomaterials from Hubei province exist from early-to-mid November 2019, and incorporating these types of data in our model could further refine our timing estimates. Moreover, wastewater detection may present the best chance of early detection of future pandemics during the early phase of spread where we estimate very low numbers of infections[67].

Even though all of the earliest documented cases of COVID-19 were found in Hubei province, we cannot discount the possibility that the index case initially acquired the virus elsewhere. Nonetheless, our dating inference is insensitive to geography. Further, our results suggest that if the virus first emerged in a rural community, it would have needed to migrate to an urban setting to avoid extinction. The lack of reports of COVID-19 elsewhere in China in November and early-December suggest Hubei province is the location where human-to-human transmission chains were first established.

The circumstances surrounding the emergence of SARS-CoV-2 in Hubei province remain shrouded. Although SARS-CoV-2 is repeatedly adapting to spread among humans[68;69], our findings do not reveal whether the virus that first emerged was less fit than the virus that spread throughout China. Nevertheless, the inferred timing of the index case is generally similar in both of these scenarios, because less-fit viruses in our simulations that went extinct tended to do so very quickly. Importantly, it is yet unknown whether the virus emerged directly from its

animal reservoir, presumably horseshoe bats[70;71], or first circulated in and possibly adapted to an intermediate host. Our estimates for the timing of the Hubei index case further distance this individual from the outbreak at the Huanan Seafood Wholesale Market. Finding the animal reservoir, or hypothetical intermediate host, will help to further narrow down the date, location, and circumstances of the original SARS-CoV-2 infection in humans. However, even in the absence of that information, coalescent-based approaches permit us to look back beyond the tMRCA and towards the earliest days of the COVID-19 pandemic. Although there was a pre-tMRCA fuse to the COVID-19 pandemic, it was almost certainly very short. This brief period of time suggests that future pandemics with similar characteristics to the COVID-19 pandemic permit only a narrow window for preemptive intervention.

1.4 Acknowledgments

We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared via GISAID the viral genomic sequence data on which this research is based. A complete list acknowledging the authors who submitted data analyzed in this study can be found in Data S1, which is available at https://www.science.org/doi/10.1126/science.abf8003. We thank Jennifer Havens, Mike Sanderson, and Betsy Wertheim for fruitful conversations and insight.

Chapter 1, in full, is a reformatted reprint of the material as it appears in "Timing the SARS-CoV-2 Index Case in Hubei Province"; Jonathan E. Pekar, Michael Worobey, Niema Moshiri, Konrad Scheffler, and Joel O. Wertheim. *Science*, 2021. The dissertation author was the primary investigator and first author of this paper.

1.4.1 Author contributions

Conceptualization: J.O.W. Methodology: J.E.P., N.M., M.W., K.S., and J.O.W. Software: J.E.P., N.M., and J.O.W. Validation: J.E.P. and J.O.W. Formal analysis: J.E.P. and J.O.W. Investigation: J.E.P. and J.O.W. Resources: J.O.W. Data Curation: J.O.W. Writing - original draft preparation: J.E.P. and J.O.W. Writing - review and editing: J.E.P., N.M., M.W., K.S., and J.O.W. Visualization: J.E.P. and J.O.W. Supervision: J.O.W. Project administration: J.O.W.

Funding acquisition: J.O.W., M.W.

1.4.2 Declaration of interests

JOW has received funding from Gilead Sciences, LLC (completed) and the CDC (ongoing) via grants and contracts to his institution unrelated to this research.

1.4.3 Data and materials availability

All data used in this analysis are free to access via GISAID. BEAST XML input, FAVITES JSON, and results files are available at DataDryad.org (https://doi.org/10.5061/dryad.4f4qrfjbm)[72]. 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.

Chapter 2

The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2

Understanding the circumstances that lead to pandemics is important for their prevention. We analyzed the genomic diversity of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) early in the coronavirus disease 2019 (COVID-19) pandemic. We show that SARS-CoV-2 genomic diversity before February 2020 likely comprised only two distinct viral lineages, denoted "A" and "B." Phylodynamic rooting methods, coupled with epidemic simulations, reveal that these lineages were most probably the result of at least two separate cross-species transmission events into humans. The first zoonotic transmission likely involved lineage B viruses around 18 November 2019 (23 October to 8 December), and the separate introduction of lineage A likely occurred within weeks of this event. These findings indicate that it is unlikely that SARS-CoV-2 circulated widely in humans before November 2019 and define the narrow window between when SARS-CoV-2 first jumped into humans and when the first cases of COVID-19 were reported. As with other coronaviruses, SARS-CoV-2 emergence likely resulted from multiple zoonotic events.

2.1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is responsible for the coronavirus disease 2019 (COVID-19) pandemic that caused more than 5 million confirmed deaths in the 2 years after its detection at the Huanan Seafood Wholesale Market (hereafter the "Huanan market") in December 2019 in Wuhan, China[73–75]. As the original outbreak spread to other countries, the diversity of SARS-CoV-2 quickly increased and led to the emergence of multiple

variants of concern, but the beginning of the pandemic was marked by two major lineages denoted "A" and "B"[76]. Lineage B has been the most common throughout the pandemic and includes all 11 sequenced genomes from humans directly associated with the Huanan market, including the earliest sampled genome, Wuhan/IPBCAMS-WH-01/2019, and the reference genome, Wuhan/Hu-1/2019 (hereafter "Hu-1")[77], sampled on 24 and 26 December 2019, respectively. The earliest lineage A viruses, Wuhan/IME-WH01/2019 and Wuhan/WH04/2020, were sampled on 30 December 2019 and 5 January 2020, respectively[78]. Lineage A differs from lineage B by two nucleotide substitutions, C8782T and T28144C, which are also found in related coronaviruses from *Rhinolophus* bats[76], the presumed host reservoir[79]. Lineage B viruses have a "C/T" pattern at these key sites (C8782 and T28144), whereas lineage A viruses have a "T/C" pattern (C8782T and T28144C). The earliest lineage A genomes from humans lack a direct epidemiological connection to the Huanan market but were sampled from individuals who lived or had recently stayed close to the market[22]. It has been hypothesized that lineages A and B emerged separately[80], but "C/C" and "T/T" genomes intermediate to lineages A and B present a challenge to that hypothesis because their existence suggests within-human evolution of one lineage toward the other by way of a transitional form. Questions about these lineages remain: If lineage B viruses are more distantly related to sarbecoviruses from *Rhinolophus* bats, then (i) why were lineage B viruses detected earlier than lineage A viruses, and (ii) why did lineage B predominate early in the pandemic? Answering these questions requires determining the ancestral haplotype, the genomic sequence characteristics of the most recent common ancestor (MRCA) at the root of the SARS-CoV-2 phylogeny. In this study, we combined genomic and epidemiological data from early in the COVID-19 pandemic with phylodynamic models and epidemic simulations. We eliminated many of the haplotypes previously suggested as the MRCA of SARS-CoV-2 and show that the pandemic most likely began with at least two separate zoonotic transmissions starting in November 2019.

2.2 Results

2.2.1 Erroneous assignment of haplotypes intermediate to lineages A and B

There are 787 near-full-length genomes available from lineages A and B sampled by 14 February 2020 (data S1 and S2). However, there are also 20 genomes of intermediate haplotypes from this period that contain either T28144C or C8782T but not both mutations: C/C or T/T, respectively. We identified numerous instances of C/C and T/T genomes sharing rare mutations with lineage A or lineage B viruses, often sequenced in the same laboratory, indicating that these intermediate genomes are likely artifacts of contamination or bioinformatics[81], similar to findings from our analysis of the emergence of SARS-CoV-2 in North America (fig. B.1 and supplementary text)[82]. We confirmed that a C/C genome from South Korea sharing three such mutations had low sequencing depth at position 28144 ($\leq$10×), a T/T genome sampled in Singapore had low coverage at both 8782 and 28144 ($\leq$10×), and three T/T genomes sampled in Wuhan had low sequencing depth and indeterminate nucleotide assignment at position 8782 (table B.1). Further, the authors of 11 C/C genomes sampled in Wuhan and Sichuan confirmed that low sequencing depth at position 8782 led to the erroneous assignment of intermediate haplotypes. C/C and T/T genomes continue to be observed throughout the pandemic as a result of convergent evolution, including T/T in the Diamond Princess cruise ship outbreak and subsequent COVID-19 waves in New York City and San Diego (figs. B.2-B.5 and supplementary text). Instances of convergent evolution are identifiable because SARS-CoV-2 phylogenies exist in "near-perfect" tree space, in which topology can be inferred with high accuracy[83]. These findings cast doubt on the claim that transitional C/C or T/T haplotypes between lineages A and B circulated in humans, reopening the door to the hypothesis that lineages A and B represent separate introductions.

2.2.2 Progenitor genome reconstruction

To better understand SARS-CoV-2 mutational patterns, we reconstructed the genome of a hypothetical progenitor of SARS-CoV-2. Using maximum likelihood ancestral state reconstruction across 15 nonrecombinant regions of SARS-CoV-2 and closely related sarbecovirus genomes sampled from bats and pangolins[84], we inferred the genome of this recombinant common ancestor (recCA) (figs. B.6, B.7 and supplementary text). The recCA differed from Hu-1 by just

381 substitutions, including C8782T and T28144C. It is more informative than an outgroup sarbecovirus because it accounts for the closest relative across all recombinant segments (figs. B.8 to B.14 and supplementary text)[85] and, as an internal node on the phylogeny, is more genetically similar to SARS-CoV-2 than any extant sarbecovirus.

2.2.3 Reversions across the early pandemic phylogeny

The ubiquity of SARS-CoV-2 reversions (mutations from Hu-1 toward the recCA) indicates that genetic similarity to related viruses is a poor proxy for the ancestral haplotype. We observe 23 distinct reversions and 631 distinct substitutions (excluding reversions) across the SARS-CoV-2 phylogeny from the COVID-19 pandemic up to 14 February 2020 (Fig. 2.1). Substitutions were overrepresented at the 381 sites separating the recCA from Hu-1 (23 of 381, 6.04%), compared with substitutions at all other sites (631 of 29,134, 2.17%).

Most reversions were C-to-T mutations (19 of 23, 82.6%), matching the mutational bias of SARS-CoV-2[86–88]. Genomes with C-to-T reversions can be found within lineage A, including C18060T (lineage A.1; for example, WA1) and C29095T (for example, 20SF012), as well as C24023T, C25000T, C4276T, and C22747T in mid-late January and February 2020. Hence, triple revertant genomes, such as WA1 and 20SF012, are neither unique nor rare. We also identified a lineage A genome (Malaysia/MKAK-CL-2020-6430/2020), sampled on 4 February 2020 from a Malaysian citizen traveling from Wuhan whose only four mutations from Hu-1 are all reversions (lineage A.1+T6025C) (Fig. 2.1). Therefore, no highly revertant haplotype can automatically be assumed to represent the MRCA of SARS-CoV-2, especially when these reversions are most often the result of C-to-T mutations. We continue to observe these reversion patterns throughout the pandemic, including in the emergence of World Health Organization (WHO)–named variants (figs. B.15 and B.16.

2.2.4 Inferring the MRCA of SARS-CoV-2

To infer the ancestral SARS-CoV-2 haplotype, we developed a nonreversible, random-effects substitution process model in a Bayesian phylodynamic framework that simultaneously reconstructs the underlying coalescent processes and the sequence of the MRCA of the SARS-CoV-2 phylogeny. The random-effects substitution model captures the C-to-T transition and

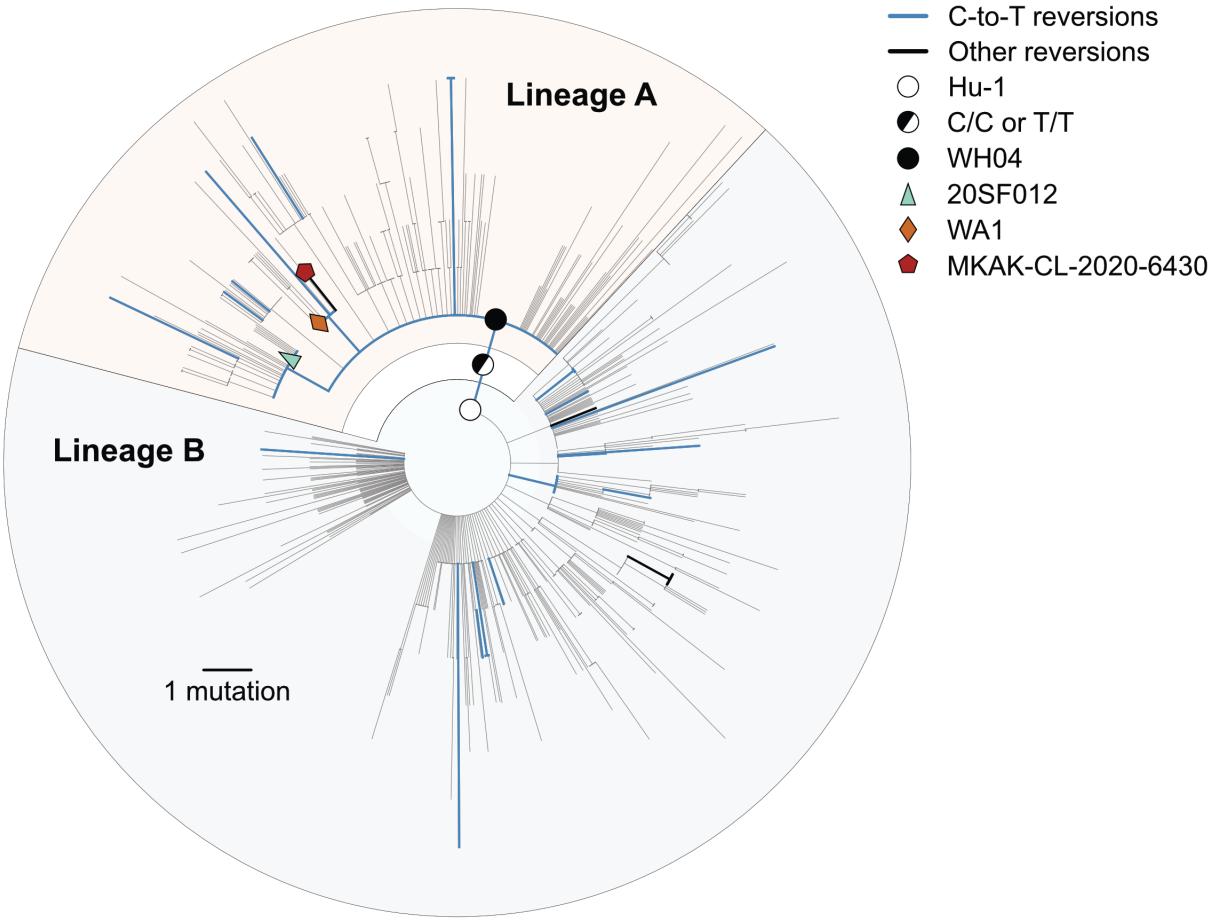


Figure 2.1. Maximum likelihood phylogeny of the early SARS-CoV-2 pandemic, showing nucleotide reversions and putative candidates for the ancestral haplotype at the most common recent ancestor (MRCA). Putative ancestral haplotypes are identified with colored shapes. Reversions from the Hu-1 reference genotype to the recCA are colored. Blue represents C-to-T reversions and black indicates all other reversions. The tree is rooted on Hu-1 to show reversion dynamics to the recCA.

G-to-T transversion biases (fig. B.17 and supplementary text). Using this model, referred to as the unconstrained rooting (fig. B.18A), we inferred the ancestral haplotype of the 787 lineage A and B genomes sampled by 14 February 2020. Our unconstrained rooting strongly favors a lineage B or C/C ancestral haplotype and shows that a lineage A ancestral haplotype is inconsistent with the molecular clock [Bayes factor (BF) = 48.1] (table 2.1). Lineage B exhibits more divergence from the root of the tree than would be expected if lineage A were the ancestral virus in humans (figs. B.19 and B.20). The T/T ancestral haplotype was also disfavored (BF >10), likely because of the C-to-T transition bias (fig. B.17). We acknowledge that the timing of

the earliest sampled lineage B genomes associated with the Huanan market could bias rooting inference toward lineage B haplotypes; however, lineage A was still disfavored after excluding all market-associated genomes (BF = 11.0).

Even though sequence similarity to closely related sarbecoviruses alone is insufficient to determine the SARS-CoV-2 ancestral haplotype, this similarity can inform phylodynamic inference. Rather than rely on outgroup rooting (fig. B.18B)[89], we developed a rooting method that assigns the recCA as the progenitor of the inferred SARS-CoV-2 MRCA (fig. B.18C). As opposed to the unconstrained rooting, the recCA root favored a lineage A haplotype over lineage B (BF = 9.4), although support for C/C was unchanged (table 2.1). Our results were insensitive to the method of breakpoint identification in the recCA (supplementary text).

Table 2.1. Posterior probabilities of inferred ancestral haplotype at the MRCA of SARS-CoV-2. Positions 8782 and 28144 are indicated in parentheses. Representative genome is genome with sequence matching the haplotype. "No market" excludes 15 market-associated genomes (13 lineage B genomes associated with the Huanan market plus one lineage A and one lineage B genome not associated with the Huanan market). *BF >3.2; **BF >10; ***BF >100. BFs are in favor of hypothesis rejection. †Haplotype with greatest posterior probability; reference for BF.

Haplotype	Mutations from Hu-1 reference	Representative genome	Phylodynamic analysis		
			Unconstrained (%)	No market (%)	recCA (%)
B (C/T)	N/A	Hu-1	80.85†	62.96†	8.18*
A (T/C)	C8782T + T28144C	WH04	1.68**	5.73**	77.28†
C/C	T28144C	N/A	10.32*	23.02	10.49*
T/T	C8782T	N/A	0.92**	1.68**	3.71**
A+C29025T (T/C)	C8782T + T28144C + C29095T	20SF012	<0.01***	<0.01***	0.20**
A.1 (T/C)	C8782T + T28144C + C18060T	WA1	<0.01***	<0.01***	0.04***

The A.1 and A+C29095T proposed ancestral haplotypes were strongly rejected by all the phylodynamic analyses, even when rooting with recCA or bat sarbecovirus outgroups, which include both C18060T and C29095T (table 2.1 and data S3). Hence, WA1-like and 20SF012-like haplotypes cannot plausibly represent the MRCA of SARS-CoV-2 as previously suggested[90–92]; the similarity of these genomes to the recCA is due to C-to-T reversions. Haplotypes not reported in table 2.1 were similarly rejected (data S3). We inferred the time of MRCA (tMRCA) for SARS-CoV-2 to be 11 December 2019 [95% highest posterior density (HPD) interval, 25 November to 12 December] by using unconstrained rooting. It has been suggested that a phylogenetic root in lineage A would produce an older tMRCA than would a lineage B rooting[92]. Therefore,

we developed an approach to assign a haplotype as the SARS-CoV-2 MRCA (A, B, C/C, A.1, or A+C29095T) and inferred the tMRCA (fig. B.18D). The tMRCA was consistent with the recCA-rooted and fixed ancestral haplotype analyses (table B.2 and supplementary text). We infer only three plausible ancestral haplotypes: lineage A, lineage B, and C/C. However, the inability to reconcile the molecular clock at the outset of the COVID-19 pandemic with a lineage A ancestor without information from related sarbecoviruses (such as the recCA) requires us to question the assumption that both lineages A and B resulted from a single introduction.

2.2.5 Separate introductions of lineages A and B

We next sought to determine whether a single introduction from one of the plausible ancestral haplotypes (lineage A, lineage B, or C/C) is consistent with the SARS-CoV-2 phylogeny. We simulated SARS-CoV-2–like epidemics[93;94] with a doubling time of 3.47 days [95% highest density interval (HDI) across simulations, 1.35 to 5.44][95–97] to account for the rapid spread of SARS-CoV-2 before it was identified as the etiological agent of COVID-19 (figs. B.21 and B.22, tables B.3 and B.4, and supplementary text). We then simulated coalescent processes and viral genome evolution across these epidemics to determine how frequently we recapitulated the observed SARS-CoV-2 phylogeny. Lineages A and B comprise 35.2 and 64.8% of the early SARS-CoV-2 genomes, respectively, and each lineage is characterized by a large polytomy (many sampled lineages descending from a single node on the phylogenetic tree), with the base of lineages A and B being the two largest polytomies observed in the early pandemic (Fig. 2.1). Furthermore, large polytomies are characteristic of SARS-CoV-2 introductions into geographical regions at the start of the pandemic (for example, fig. B.23)[82;98–100] and would similarly be expected to occur after a successful introduction of SARS-CoV-2 into humans. Congruently, the most common topology in our simulations is a large basal polytomy (with $\geq$100 descendent lineages), which is present in 47.5% of simulated epidemics (Fig. 2.2A).

By contrast, a topology corresponding to a single introduction of an ancestral C/C haplotype—characterized by two clades, each comprising $\geq$30% of the taxa, possessing a large polytomy at the base, and separated from the MRCA by one mutation (Fig. 2.2B)—was only observed in 0.0% of our simulations. Further, a topology corresponding to a single introduction of

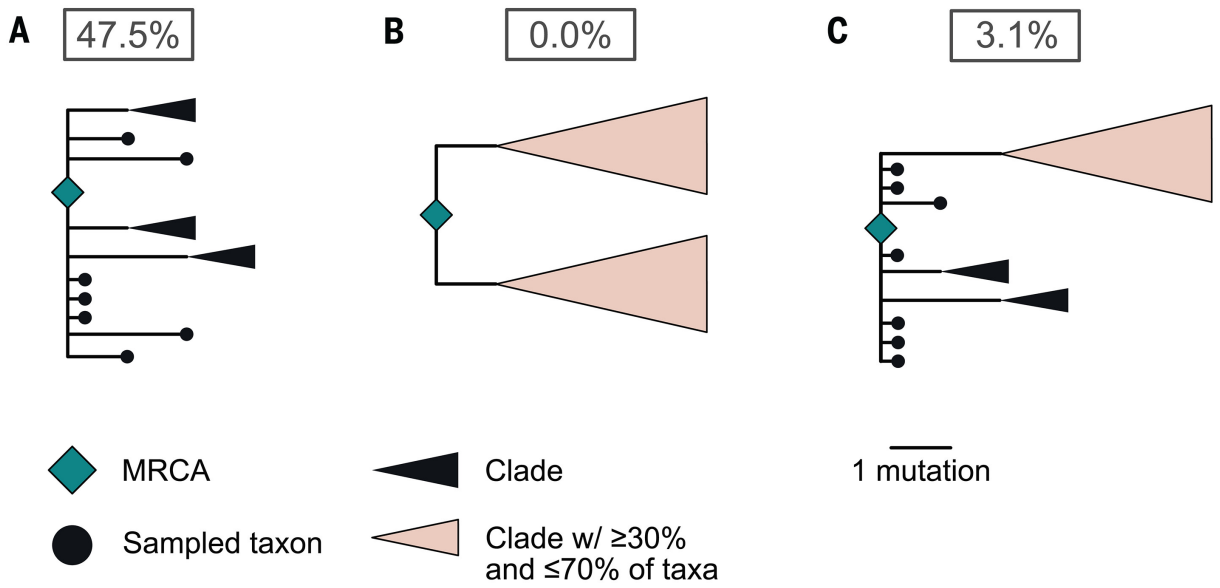


Figure 2.2. Probability of phylogenetic structures arising from a single introduction of SARS-CoV-2 in epidemic simulations. (**A**) A large polytomy of at least 100 descendent lineages, consistent with the base of both lineages A and B. (**B**) Topology matching a C/C ancestral haplotype: two clades each one mutation from the ancestor, both with polytomies of at least 100 descendent lineages. (**C**) Topology matching either a lineage A or lineage B ancestral haplotype: a basal polytomy with at least 100 descendent lineages including a large clade separated by two mutations, also possessing a polytomy of at least 100 descendent lineages. Basal taxa have short branch lengths for clarity. The probability of each phylogenetic structure after a single introduction is reported in the respective boxes.

an ancestral lineage A or lineage B haplotype—characterized by a large basal polytomy and a large clade, comprising between 30 and 70% of taxa, two mutations from the root with no intermediate genomes—was observed in only 3.1% of our simulations (Fig. 2.2C and supplementary text). Our epidemic simulations do not support a single introduction of SARS-CoV-2 giving rise to the observed phylogeny. We therefore quantified the relative support for two introductions resulting in the empirical topology. By synthesizing posterior probabilities of inferred ancestral haplotypes, frequencies of topologies in epidemic simulations, and the expected relationships between these haplotypes and topologies, we inferred substantial support favoring separate introductions of lineages A and B (BF = 4.3 and BF = 4.2 by using the recCA and unconstrained rooting, respectively) [supplementary materials (SM), materials and methods]. This support is robust across shorter and longer doubling times, varying ascertainment rates, and minimum polytomy size (tables B.4 and B.5). If lineages A and B arose from separate introductions, then the

MRCA of SARS-CoV-2 was not in humans, and it is the tMRCAs of lineages A and B that are germane to the origins of SARS-CoV-2 (not the timing of their shared ancestor). Rooting with the recCA, we inferred the median tMRCA of lineage B to be 15 December (95% HPD, 5 December to 23 December) and the median tMRCA of lineage A to be 20 December (95% HPD, 5 December to 29 December) (Fig. 2.3A). The tMRCA of lineage B consistently predates the tMRCA of lineage A (Fig. 2.3B). These results are robust to using unconstrained rooting, fixing the ancestral haplotype, and excluding market-associated genomes (Fig. 2.3, A and B; table B.2; and supplementary text).

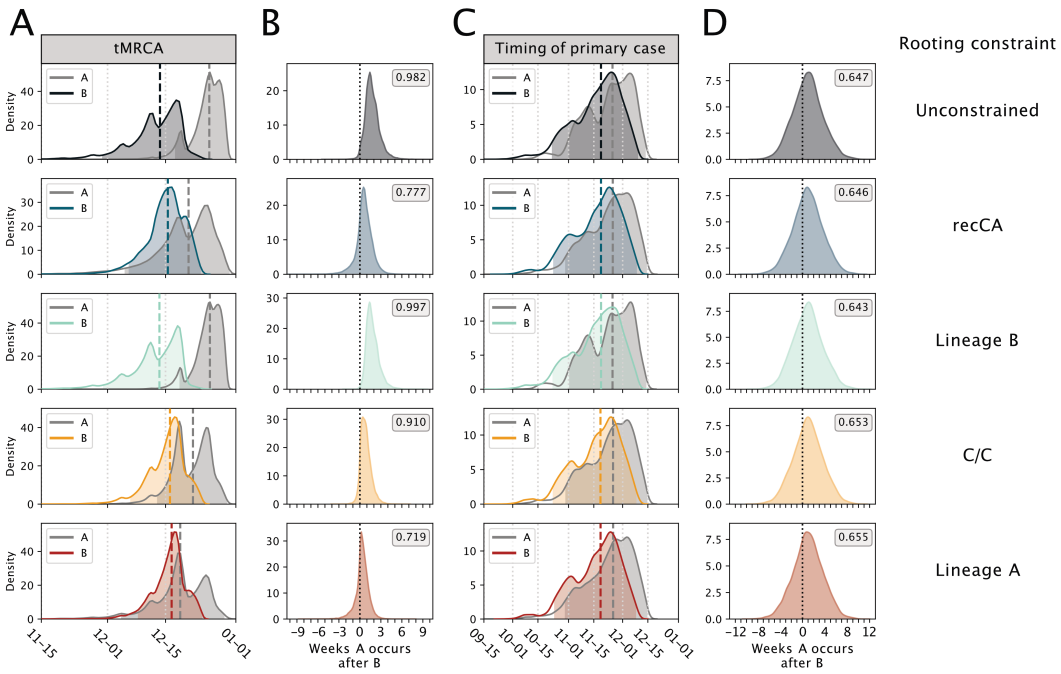


Figure 2.3. Comparison of the tMRCA and primary case dates for lineage A and lineage B across rooting strategies. Each row represents a different rooting constraint in phylodynamic analysis, with the bottom three representing a fixed ancestral haplotype. (**A**) The tMRCA for lineages A and B. (**B**) The number of weeks the tMRCA of lineage A occurs after the tMRCA of lineage B. (**C**) The timing of the primary case for lineages A and B. (**D**) The number of weeks the time of the primary case of lineage A occurs after the time of the primary case of lineage B. Long dashed lines indicate the median and shading represents the 95% HPD for each distribution. Short dashed lines indicate 0 weeks difference between lineages A and B. Posterior probability that lineage A originated after lineage B is reported in the grey box.

2.2.6 Timing the introductions of lineages A and B

The primary case, the first human infected with a virus in an outbreak, could precede the tMRCA if basal lineages went extinct during cryptic transmission[21;94;101]. The index case, the first identified case, is rarely also the primary case[102;103]. We next used an extension of our previously published framework that combines epidemic simulations and phylodynamic tMRCA inference (SM materials and methods)[21;94;101] to infer the timing of the lineage B and lineage A primary cases, accounting for both the index case symptom onset date and earliest documented COVID-19 hospitalization date. The earliest unambiguous case of COVID-19, with symptom onset on 10 December and hospitalization on 16 December, was a seafood vendor at the Huanan market. Unfortunately, no published genome is available for this case[22]. Nonetheless, we can reasonably assume that this individual had a lineage B virus (supplementary text) because an environmental sample (EPI_ISL_408512) from the stall this vendor operated was lineage B. The earliest lineage A genome (IME-WH01) is from a familial cluster for which the earliest symptom onset is 15 December and earliest hospitalization is 25 December[26]. Accounting for these dates and using the recCA rooting, we inferred the infection date of the lineage B primary case to be 18 November (95% HPD, 23 October to 8 December) and the infection date of the primary case of lineage A to be 25 November (95% HPD, 29 October to 14 December). The lineage B primary case predated that of lineage A in 64.6% of the posterior sample, by a median of 7 days (Fig. 2.3D and table B.6). Our lineage A and B primary case inference is robust to rooting strategy and fixing the plausible ancestral haplotype to lineage A, lineage B, or C/C, as well as different index case dates, accounting for only hospitalization dates and varying growth rates and ascertainment rates (tables B.7 to B.10 and supplementary text). Therefore, our results indicate that lineage B was introduced into humans no earlier than late October and likely in mid-November 2019, and the introduction of lineage A occurred within days to weeks of this event. We then inferred the number of ascertained infections and hospitalizations arising from these separate introductions. We found that an earlier introduction of lineage B led to a faster rise in lineage B–associated infections, dominating the simulated epidemics (Fig. 2.4) and recapitulating the predominance of lineage B observed in China in early 2020[104]. Similarly, simulated lineage B hospitalizations are more common than those from lineage A through January 2020 (Fig. B.24). We observed

these patterns regardless of rooting strategy (unconstrained or recCA), ancestral haplotype (B, A, or C/C) (Fig. 2.4 and tables B.11 and B.12), and doubling time (Figs. B.25-B.28).

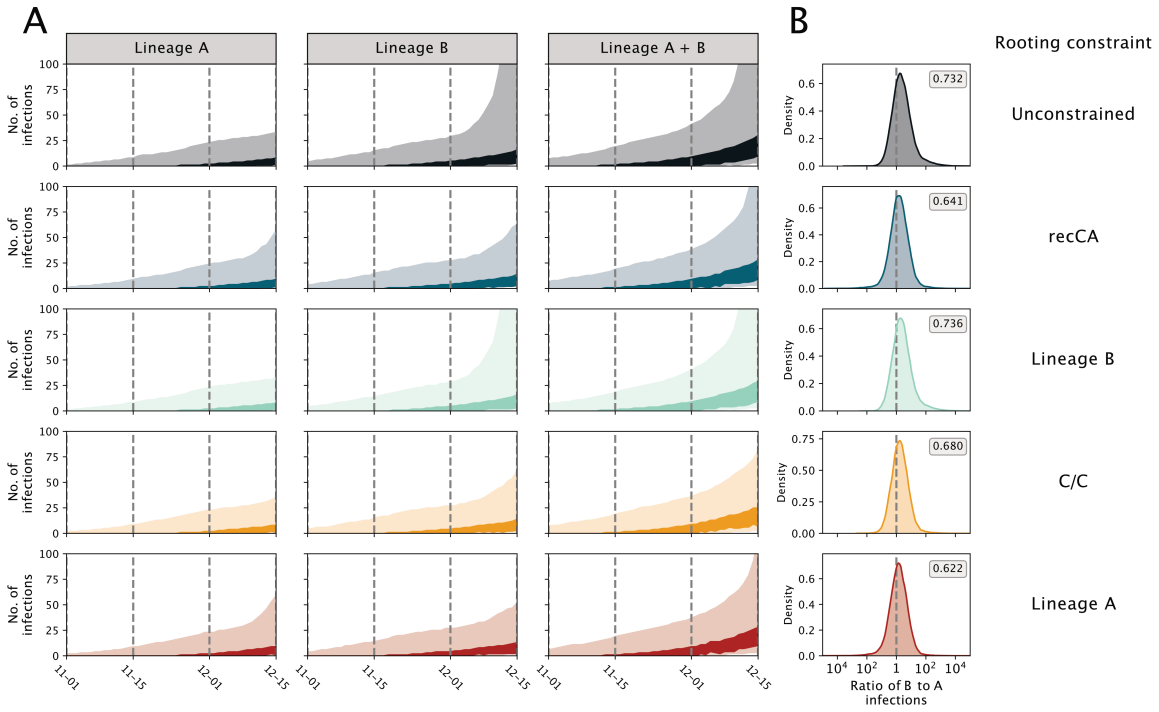


Figure 2.4. Dynamics of simulated SARS-CoV-2 epidemics resulting from separate introductions of lineages A and B. Each row represents a different rooting constraint in phylodynamic analysis, with the bottom three representing a fixed ancestral haplotype. (**A**) Estimated number of infections. The header of each column indicates whether the number of infections are caused by lineage A, lineage B, or the two lineages combined. Darker and lighter shading represent the 50% and 95% HPD, respectively. (**B**) The log ratio of lineage B to lineage A infections on 15 December 2019. Posterior probability of having more lineage B infections than lineage A reported in the grey box.

2.2.7 Minimal cryptic circulation of SARS-CoV-2

We do not see evidence for substantial cryptic circulation before December 2019 (Fig. 2.4), even if we assume a single introduction (fig. B.29 and supplementary text). Our simulated epidemics have a median of three (95% HPD, 1 to 18) cumulative infections at the tMRCA, with 99% of simulated epidemics resulting in at most 33 infections (table B.13 and supplementary text). Further, it is unlikely that there were any COVID-19–related hospitalizations before December [105] because the simulated epidemics show a median of zero (95% HPD, 0 to 2) hospitalizations by 1 December 2019. These results are in accordance with the lack of a single SARS-CoV-2–positive

sample among tens of thousands of serology samples from healthy blood donors from September to December 2019[106] and thousands of specimens obtained from influenza-like illness patients at Wuhan hospitals from October to December 2019[26]. Therefore, there was likely extremely low prevalence of SARS-CoV-2 in Wuhan before December 2019. Even when we simulated epidemics with a longer doubling time, resulting in an earlier timing of the primary cases (tables B.8 and B.10), there were still few infections before December 2019 (table B.13).

2.2.8 Additional introductions

The extinction rate of our simulated epidemics (simulations that did not produce self-sustaining transmission chains) indicate that there were likely multiple failed introductions of SARS-CoV-2. Similar to our previous findings[94], 77.8% of simulated epidemics went extinct. These failed introductions produced a mean of 2.06 infections and 0.10 hospitalizations; hence, failed introductions could easily go unnoticed. If we treat each SARS-CoV-2 introduction, failed or successful, as a Bernoulli trial and simulate introductions until we see two successful introductions, we estimate that eight (95% HPD, 2 to 23) introductions led to the establishment of both lineage A and B in humans.

2.2.9 Limitations

Our analysis of the putative intermediate haplotypes suggests that there remain lineage assignment errors between lineages A and B, particularly of genomes sampled in January and February of 2020, which could influence the precision of the phylogenetic topology and tMRCA inference. We lack direct evidence of a virus closely related to SARS-CoV-2 in nonhuman mammals at the Huanan market or its supply chain. The genome sequence of a virus directly ancestral to SARS-CoV-2 would provide more precision regarding the timing of the introductions of SARS-CoV-2 into humans and the epidemiological dynamics before its discovery. Although we simulated epidemics across a range of plausible epidemiological dynamics, our models represent a time frame before the ascertainment of COVID-19 cases and sequencing of SARS-CoV-2 genomes and thus before when these models could be empirically validated.

2.3 Discussion

The genomic diversity of SARS-CoV-2 during the early pandemic presents a paradox. Lineage A viruses are at least two mutations closer to bat coronaviruses, indicating that the ancestor of SARS-CoV-2 arose from this lineage. However, lineage B viruses predominated early in the pandemic, particularly at the Huanan market, indicating that this lineage began spreading earlier in humans. Further complicating this matter is the molecular clock of SARS-CoV-2 in humans, which rejects a single-introduction origin of the pandemic from a lineage A virus. We resolved this paradox by showing that early SARS-CoV-2 genomic diversity and epidemiology are best explained by at least two separate zoonotic transmissions, in which lineage A and B progenitor viruses were both circulating in nonhuman mammals before their introduction into humans (figs. B.30 and B.31). The most probable explanation for the introduction of SARS-CoV-2 into humans involves zoonotic jumps from as-yet-undetermined, intermediate host animals at the Huanan market[26;107;108]. Through late 2019, the Huanan market sold animals that are known to be susceptible to SARS-CoV-2 infection and capable of intraspecies transmission[109–111]. The presence of potential animal reservoirs, coupled with the timing of the lineage B primary case and the geographic clustering of early cases around the Huanan market[108], support the hypothesis that SARS-CoV-2 lineage B jumped into humans at the Huanan market in mid-November 2019. In a related study[108], we show that the two earliest lineage A cases are more closely positioned geographically to the Huanan market than expected compared with other COVID-19 cases in Wuhan in early 2020, despite having no known association with the market. This geographic proximity is consistent with a separate and subsequent origin of lineage A at the Huanan market in late November 2019. The presence of lineage A virus at the Huanan market was confirmed by Gao *et al.*[112] from a sample taken from discarded gloves. The high extinction rate of SARS-CoV-2 transmission chains, observed in both our simulations and real-world data[113], indicates that the two zoonotic events that established lineages A and B may have been accompanied by additional, cryptic introductions. However, such introductions could easily be missed, particularly if their subsequent transmission chains quickly went extinct or the introduced viruses had a lineage A or B haplotype. Failed introductions of intermediate haplotypes are also possible.

Critically, we have no evidence of subsequent zoonotic introductions in late December leading up to the closure of the Huanan market on 1 January 2020. By then, the susceptible host animals that had been documented at the market during the previous months were no longer found in the Huanan market[26]. Other coronavirus epidemics and outbreaks in humans—including SARS-CoV-1, Middle East respiratory syndrome coronavirus (MERS-CoV), and most recently, porcine deltacoronavirus in Haiti—have been the result of repeated introductions from animal hosts[114–116]. These repeated introductions were easily identifiable because human viruses in these outbreaks were more closely related to viruses sampled in the animal reservoirs than to other human viruses. However, the genomic diversity within the putative SARS-CoV-2 animal reservoir at the Huanan market was likely shallower than that seen in SARS-CoV-1 and MERS-CoV reservoirs[114;115;117]. Hence, even though lineages A and B had nearly identical haplotypes, their MRCA likely existed in an animal reservoir. The ability to disentangle repeated introductions of SARS-CoV-2 from a shallow genetic reservoir has previously been shown in the early SARS-CoV-2 epidemic in Washington state, where two viruses, separated by two mutations, were independently introduced from, and shared an MRCA in, China (figs. B.23 and B.30 and supplementary text)[82]. Successful transmission of both lineage A and B viruses after independent zoonotic events indicates that evolutionary adaptation within humans was not needed for SARS-CoV-2 to spread[118]. We now know that SARS-CoV-2 can readily spread after reverse-zoonosis to Syrian hamsters (*Mesocricetus auratus*), American mink (*Neovison vison*), and white-tailed deer (*Odocoileus virginianus*), indicating its host generalist capacity[119–124]. Furthermore, once an animal virus acquires the capacity for human infection and transmission, the only remaining barrier to spillover is contact between humans and the pathogen. Thereafter, a single zoonotic transmission event indicates that the conditions necessary for spillovers have been met, which portends additional jumps. For example, there were at least two zoonotic jumps of SARS-CoV-2 into humans from pet hamsters in Hong Kong[124] and dozens from minks to humans on Dutch fur farms[121;122]. We show that it is highly unlikely that SARS-CoV-2 circulated widely in humans earlier than November 2019 and that there was limited cryptic spread, with at most dozens of SARS-CoV-2 infections in the weeks leading up to the inferred tMRCA, but likely far fewer. By late December, when SARS-CoV-2 was identified as the etiological agent

of COVID-19[22], the virus had likely been introduced into humans multiple times as a result of persistent contact with a viral reservoir.

2.4 Materials and Methods Summary

Materials and methods described in full detail can be found in the supplementary materials.

2.4.1 Sequence data

We queried the GISAID database[125], GenBank, and National Genomics Data Center of the China National Center for Bioinformatics (CNCB) for complete high-coverage SARS-CoV-2 genomes collected by 14 February 2020, resulting in a dataset of 787 taxa belonging to lineages A and B and 20 taxa with C/C or T/T haplotypes. Genomes were aligned by using MAFFT v7.453[126] to the SARS-CoV-2 reference genome (Wuhan/Hu-1/2019), and 388 sites were masked at the 5' and 3' ends and at sites based on De Maio et al.[127]. All genome accessions are available in data S1 and S2.

2.4.2 Progenitor genome reconstruction and reversion analysis

We reconstructed the progenitor of SARS-CoV-2, the the recCA. We (i) inferred a maximum likelihood tree of 31 sarbecovirus genomes (SARS-CoV-2 and 30 closely related sarbecoviruses sampled from bats and pangolins) across 15 predefined nonrecombinant regions[84] with IQ-TREE v2.0.7[128], (ii) inferred the sequence of the ancestor of SARS-CoV-2 in each tree with TreeTime v0.8.1[129], and (iii) concatenated the resulting sequences. We next inferred a maximum likelihood tree of the 787 SARS-CoV-2 taxa with IQ-TREE and performed ancestral state reconstruction with TreeTime to identify substitutions that were reversions from Wuhan-Hu-1 to the recCA across the SARS-CoV-2 phylogeny.

2.4.3 Phylodynamic inference and epidemic simulations

We performed phylodynamic inference using BEAST v1.10.5[130] with the 787-taxon dataset to infer the ancestral haplotype and the tMRCA of SARS-CoV-2 (and the tMRCAs of lineages A and B), using a nonreversible random-effects substitution model and exploring

unconstrained rooting, recCA-rooting, fixing the ancestral haplotype as a root, and outgroup rooting. SARS-CoV-2–like epidemics were simulated with FAVITES-COVID-Lite v0.0.1 [93;131] using a scale-free network of 5 million individuals and a customized extension of the SAPHIRE model [132], producing coalescent trees on which we simulated mutations. We calculated the BF comparing the support of two introductions of SARS-CoV-2 with one introduction by considering the posterior probabilities of the four most likely ancestral haplotypes from the phylodynamic inference (lineage A, lineage B, C/C, and T/T), the frequencies of the phylogenetic structures associated with introductions of these haplotypes in the epidemic simulations, and equal prior probabilities for each ancestral haplotype and the number of introductions. We connected the phylodynamic inference and epidemic simulations by means of a rejection sampling–based approach [94], accounting for the tMRCAs of lineages A and B and the earliest documented COVID-19 illness onset and hospitalization dates. We then inferred the timing of the introductions of lineages A and B and the infections and hospitalizations for each lineage. The proportion of epidemic simulations that went extinct (no onward transmission by the end of the simulation) was used to approximate the number of SARS-CoV-2 introductions needed to result in two introductions with sustained onward transmission.

2.5 Acknowledgments

We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared through GISAID the viral genomic sequences and metadata on which this research is based (data S1) [126]. We are greatly appreciative toward L. Chen, D. Liu, and Y. Yan for providing insight into the putative intermediate genomes and clarification regarding the relative sequencing depth at positions 8782 and 28144, M. Eloit and S. Temmam for sharing their sarbecovirus dataset and recombination analysis results, and M. Kuehnert for general feedback. Figure B.30 was created with Biorender.com.

Chapter 2, in full, is a reformatted reprint of the material as it appears in "The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2"; Jonathan E. Pekar, Andrew Magee, Edyth Parker, Niema Moshiri, Katherine Izhikevich, Jennifer L. Havens, Karthik Gangavarapu, Lorena Mariana Malpica Serrano, Alexander Crits-Christoph, Nathaniel L. Matteson, Mark

Zeller, Joshua I. Levy, Jade C. Wang, Scott Hughes, Jungmin Lee, Heedo Park, Man-Seong Park, Katherine Ching Zi Yan, Raymond Tzer Pin Lin, Mohd Noor Mat Isa, Yusuf Muhammad Noor, Tetyana I. Vasylyeva, Robert F. Garry, Edward C. Holmes, Andrew Rambaut, Marc A. Suchard, Kristian G. Andersen, Michael Worobey, and Joel O. Wertheim. *Science*, 2022. The dissertation author was the primary investigator and first author of this paper.

2.5.1 Funding

This project has been funded in whole or in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), Department of Health and Human Services, under contract 75N93021C00015 (M.W.). J.E.P. acknowledges support from NIH (T15LM011271). N.M. acknowledges support from the National Science Foundation (NSF) (NSF-2028040). J.I.L. acknowledges support from NIH (5T32AI007244-38). J.O.W. acknowledges support from NIH (R01AI135992 and R01AI136056). R.F.G. is supported by NIH (R01AI132223, R01AI132244, U19AI142790, U54CA260581, U54HG007480, and OT2HL158260), the Coalition for Epidemic Preparedness Innovation, the Wellcome Trust Foundation, Gilead Sciences, and the European and Developing Countries Clinical Trials Partnership Programme. M.A.S. and A.R. acknowledge the support from the Wellcome Trust (Collaborators Award 206298/Z/17/Z–ARTIC network), the European Research Council (grant agreement 725422–ReservoirDOCS), and NIH (R01AI153044). K.G.A. is supported by NIH (U19AI135995, U01AI151812, and UL1TR002550). E.C.H. is funded by an Australian Research Council Laureate Fellowship (FL170100022). J.L., H.P., and M.-S.P. acknowledge support from the National Research Foundation of Korea, funded by the Ministry of Science and Information and Communication Technologies, Republic of Korea (NRF-2017M3A9E4061995 and NRF-2019R1A2C2084206). T.I.V. acknowledges support from the Branco Weiss Fellowship. We thank AMD for the donation of critical hardware and support resources from its HPC Fund that made this work possible. This work was supported (in part) by the Epidemiology and Laboratory Capacity (ELC) for Infectious Diseases Cooperative Agreement [grant ELC DETECT (6NU50CK000517-01-07)] funded by the Centers for Disease Control and Prevention (CDC). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC or the Department of Health and Human Services.

2.5.2 Author contributions

Conceptualization: J.E.P., M.A.S., K.G.A., M.W., and J.O.W. Methodology: J.E.P., A.M., N.M., M.A.S., K.G.A., M.W., and J.O.W. Software: J.E.P., A.M., N.M., K.G., and M.A.S. Validation: J.E.P., A.M., K.I., K.G., and M.A.S. Formal analysis: J.E.P., A.M., E.P., K.I., J.L.H., K.G., and J.O.W. Investigation: J.E.P., A.M., E.P., K.I., J.L.H., K.G., and J.O.W. Resources: M.A.S., K.G.A., and J.O.W. Data curation: J.E.P., E.P., K.G., M.Z., J.C.W., S.H., J.L., H.P., M.-S.P., K.C.Z.Y., R.T.P.L., M.N.M.I., Y.M.N., and J.O.W. Writing – original draft preparation: J.E.P., M.W., and J.O.W. Writing – review and editing: All authors. Visualization: J.E.P., J.L.H., K.G., and L.M.M.S.; Supervision: M.A.S, K.G.A., M.W., and J.O.W.; Project administration: M.A.S., K.G.A., M.W., and J.O.W.; Funding acquisition: M.A.S., K.G.A., M.W., and J.O.W.

2.5.3 Competing interests

J.O.W. has received funding from the CDC (ongoing) through contracts or agreements to his institution unrelated to this research. M.A.S. receives contracts and grants from the US Food and Drug Administration, the US Department of Veterans Affairs, and Janssen Research and Development unrelated to this research. R.F.G. is cofounder of Zalgen Labs, a biotechnology company developing countermeasures to emerging viruses. M.W., E.C.H., A.R., M.A.S., J.O.W., and K.G.A. have received consulting fees and/or provided compensated expert testimony on SARS-CoV-2 and the COVID-19 pandemic.

2.5.4 Data and materials availability

Genome accessions are available in data S1 and S2, and raw data for two genomes were deposited to NCBI SRA (PRJNA806767 and PRJNA802993). Code is available on Zenodo[133]. The following data are available on Zenodo[134]: recCA sequence, BEAST phylogenetic inference output, and simulation and rejection sampling output for the primary analysis. Data S1-S3 is available at https://www.science.org/doi/full/10.1126/science.abp8337.

Chapter 3

The recency and geographical origins of the bat viruses ancestral to SARS-CoV and SARS-CoV-2

The emergence of SARS-CoV in 2002 and SARS-CoV-2 in 2019 has led to increased sampling of related sarbecoviruses circulating primarily in horseshoe bats. These viruses undergo frequent recombination and exhibit spatial structuring across Asia. Employing recombination-aware phylogenetic inference on bat sarbecoviruses, we find that the closest-inferred bat virus ancestors of SARS-CoV and SARS-CoV-2 existed just 1–3 years prior to their emergence in humans. Phylogeographic analyses examining the movement of related sarbecoviruses demonstrate that they traveled at similar rates to their horseshoe bat hosts and have been circulating for thousands of years in Asia. The closest-inferred bat virus ancestor of SARS-CoV likely circulated in western China, and that of SARS-CoV-2 likely circulated in a region comprising southwest China and northern Laos, both a substantial distance from where they emerged. This distance and recency indicate that the direct ancestors of SARS-CoV and SARS-CoV-2 could not have reached their respective sites of emergence via the bat reservoir alone. Our recombination-aware dating and phylogeographic analyses reveal a more accurate inference of evolutionary history than performing only whole-genome or single gene analyses. These results can guide future sampling efforts and demonstrate that viral genomic fragments extremely closely related to SARS-CoV and SARS-CoV-2 were circulating in horseshoe bats, confirming their importance as the reservoir species for SARS viruses.

3.1 Introduction

Horseshoe bats (*Rhinolophus spp.*) are the main hosts of the *Sarbecovirus* subgenus[32] (genus *Betacoronavirus*, family *Coronaviridae*) and, likely via virus transmission through a transient intermediate species, were the source of SARS-CoV (referred to hereafter as SARS-CoV-1, now extinct) and SARS-CoV-2[33–35]. Since the emergence of SARS-CoV-1 in Guangdong Province in 2002 and SARS-CoV-2 (jointly referred to as the SARS-CoVs) in Hubei Province in 2019, there has been increased sampling of sarbecoviruses in horseshoe bats, which has contributed to our understanding of sarbecovirus diversity and their geographical spread.

Genome-wide sequence identity is typically used to compare bat sarbecoviruses to the SARS-CoVs, but because coronaviruses frequently recombine[135;136], whole genome identity of these bat viruses and the SARS-CoVs does not adequately reflect their complex evolutionary histories. For example, the most recent common ancestor (MRCA) of SARS-CoV-2 and a closely-related bat virus genome sequence might be estimated at 20 years before 2019. However, when analyzing subgenomic, non-recombinant regions (NRRs) we will inevitably find some NRRs with older and some with younger MRCAs than this whole-genome estimate, which is effectively a weighted average of NRR tMRCAs. Because recombination is so rampant among bat sarbecoviruses, we likely would have had to have sampled the sarbecovirus from the individual bat involved in each cross-species transmission event leading to the SARS-CoVs (which we refer to as the "direct ancestor"; Fig. C.1), to have sampled a bat virus that had no recombination in its genome relative to these SARS-CoVs. What we have, instead, is a relatively small number of bat sarbecovirus genome sequences, gathered across different times and places. Hence, it is necessary to identify the NRRs of sarbecoviruses—genomic regions within which there is little to no detectable signal of recombination among the analyzed genomes and representing, to the extent possible, a single evolutionary history. These regions provide the clearest insights into the evolutionary history of SARS-CoV-1 and SARS-CoV-2 and the origins of their respective outbreaks.

Sarbecoviruses exhibit substantial geographic structuring linked to the ranges of their host horseshoe bat species[33–35;135;136], but where and when their ancestors circulated is poorly

understood. The sarbecoviruses have an ancient origin[137–139], whose timing is difficult to infer due to substitution saturation, the repeated occurrence of substitutions at the same site resulting in biased estimation of older divergence in viral phylogenies[140;141]. This feature of viruses complicates dating estimates of deep nodes in the sarbecovirus phylogenies and could falsely suggest that their ancestors circulated more recently than they actually did[142].

Here, we separately analyze the recombination patterns and respective evolutionary histories of SARS-CoV-1 and its closely related viruses (referred to as SARS-CoV-1-like viruses) and SARS-CoV-2 and its closely related viruses (referred to as SARS-CoV-2-like viruses). We estimate the separate sets of NRRs for the SARS-CoV-1-like and SARS-CoV-2-like viruses. For each NRR we utilize a molecular clock rate specific for that portion of the genome and determine the MRCA of the SARS-CoV and the most closely related bat virus, or set of viruses, among currently sampled sarbecoviruses for the NRR. We refer to this MRCA as the "closest-inferred ancestor" for that NRR because it is the parent node of the SARS-CoV in the phylogeny (Fig. C.1) and is limited by the current sample set of bat sarbecoviruses.

We show that the available sequence data provide evidence of the closest-inferred bat virus ancestors of both SARS-CoV-1 and SARS-CoV-2 circulating in bats in the years immediately preceding the emergence of each SARS-CoV. By performing a phylogeographic analysis of these sarbecoviruses and accounting for substitution saturation[137;138], we demonstrate that the sarbecoviruses spread at a rate similar to their bat hosts and that, based on current sampling, both SARS-CoV-1 and SARS-CoV-2 likely emerged far from their direct bat virus ancestors.

3.2 Inference of non-recombinant regions

To identify the recombination patterns of the sarbecoviruses, we aligned the available 167 full-genome sarbecovirus sequences (Supplementary Data File 1), including SARS-CoV-1 and SARS-CoV-2. We used GARD[143] to independently identify recombination breakpoints on genomes closely related to SARS-CoV-1 and SARS-CoV-2 (Fig. C.2). These breakpoints identify 31 putative NRRs for the SARS-CoV-1-like viruses, with a median length of 866 nucleotides (nt) and a range of 309 to 2195 nt, and 27 putative NRRs for the SARS-CoV-2-like viruses, with a median length of 970 nt and a range of 252 to 2836 nt (Fig. 3.1a). To examine whether the

inferred positions of recombination breakpoints of the SARS-CoV-1-like and SARS-CoV-2-like NRR analyses were closer to each other than expected, we simulated randomly distributed recombination breakpoint sets along the alignment for the SARS-CoV-1-like and SARS-CoV-2-like viruses (30 and 26 breakpoints, respectively) and compared the distribution of distances between the two breakpoint sets. The empirical breakpoint sets had a mean genomic distance of 307 nt (90% CIs: 204-463 nt) separating the SARS-CoV-1 and SARS-CoV-2 breakpoints, whereas the simulated distances had a mean of 540 nt (90% CIs: 334-872 nt) (Fig. C.3). Additionally, there were six breakpoint pairs between the two virus sets that were in the smallest 10% of the observed distance distribution (15 to 78 nt apart; Fig. 3.1a), indicating that hotspots of recombination are likely shared between the two sarbecovirus clades (e.g., 5' end of nsp3, upstream of Spike, and other accessory genes).

3.3 Sarbecovirus genomes provide evidence for a very recent bat ancestor of SARS-CoV-1 and SARS-CoV-2

The evolutionary history of sarbecoviruses can only be accurately inferred when focusing on NRRs rather than entire genomes or specific genes because these viruses, as well as their ancestors, have undergone extensive recombination[35;136;144]. Here, we examine when the closest-inferred bat viruses ancestral to SARS-CoV-1 and SARS-CoV-2 circulated. These ancestors are represented by the MRCA of each SARS-CoV and its most closely related bat virus (or viruses), which is akin to the parent node of each SARS-CoV in its respective phylogeny.

There is insufficient temporal signal when calibrating a molecular clock using tip dating with sarbecoviruses sampled from bats and pangolins (Manis javanica)[136], likely as a consequence of limited sampling across space and time. Therefore, we used SARS-CoV-1 genomes to identify a suitable rate prior for the molecular clock in our time-scaled phylogenetic analyses (Fig. C.1; see Methods). The inferred substitution rates served as NRR-specific rate priors for subsequent Bayesian phylogenetic inference of SARS-CoV-1 and the 139 SARS-CoV-1-like viruses (Fig. C.4, table C.1). The median time of the closest-inferred bat virus ancestor of SARS-CoV-1 within each NRR varied by several decades (Fig. 3.1b), and the median of these medians is 1996. However, the most recent time of the closest-inferred ancestor across all NRRs of SARS-CoV-1 was 2001

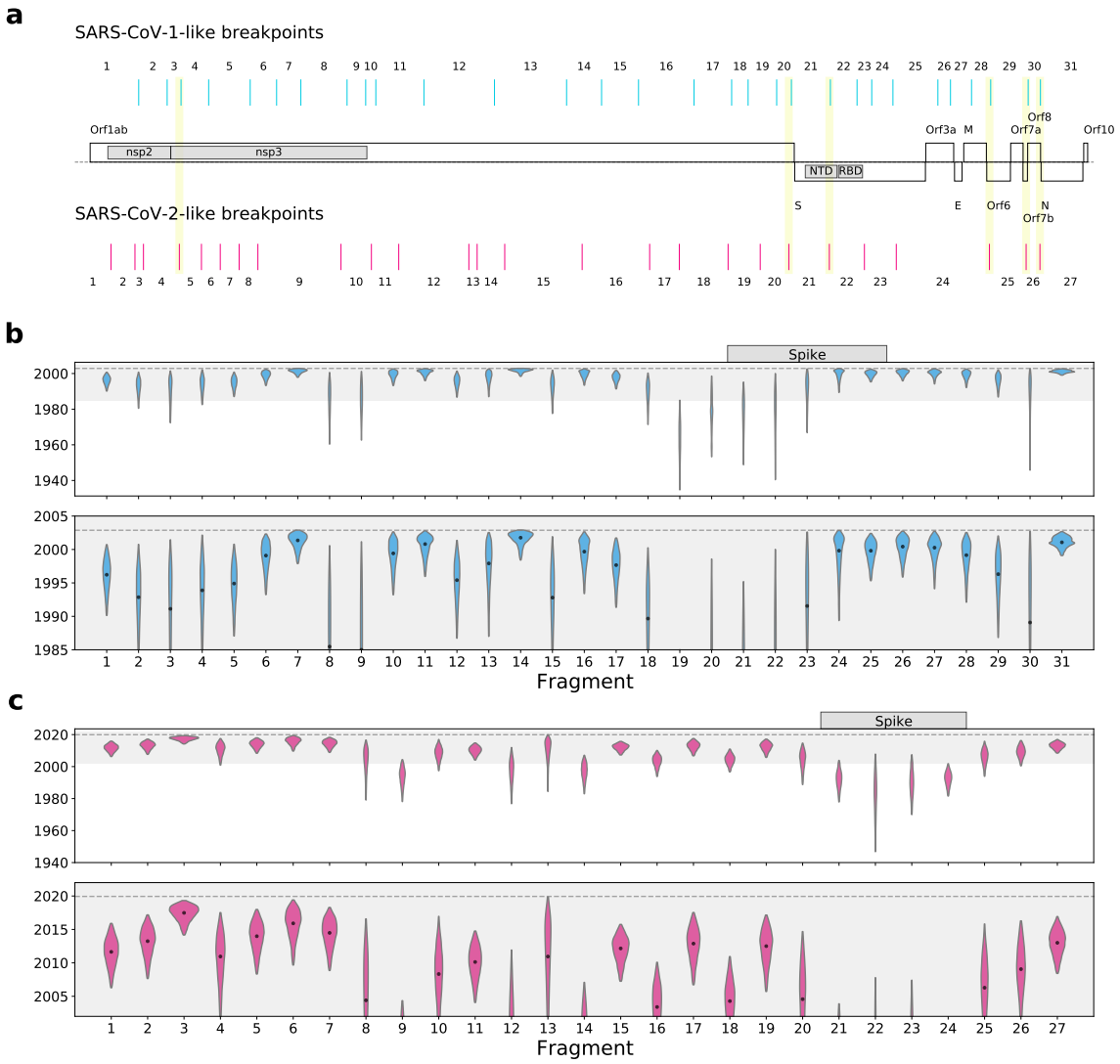


Figure 3.1. Recombination patterns and dating of the closest-inferred bat virus ancestor of SARS-CoV-1 and SARS-CoV-2 for each NRR. a, Distribution of SARS-CoV-1-like and SARS-CoV-2-like breakpoints across the sarbecovirus genome alignment. Schematic of major open-reading frames is displayed in the middle of the figure. Gene domains of interest are also shown within the schematic. Breakpoint pairs with significantly short distances (¡10% point of distribution) are highlighted in yellow. Numbers indicate NRRs resulting from breakpoints. The alignment is 31,243 nt long. **b,** Time of the closest-inferred bat virus ancestor of SARS-CoV-1 across the 31 NRRs of the SARS-CoV-1-like viruses. The 20 years before the emergence of SARS-CoV-1 is shown in gray in the upper panel and zoomed-in upon in the lower panel. **c,** Time of the closest-inferred bat virus ancestor of SARS-CoV-2 across the 27 NRRs of the SARS-CoV-2-like viruses The 20 years before the emergence of SARS-CoV-2 is shown in gray in the upper panel and zoomed-in upon in the lower panel. The dashed line in each panel is the date of the earliest sampled human case of each SARS-CoV (16 November 2002 for SARS-CoV-1; 24 December 2019 for SARS-CoV-2). The violin plots correspond to the 95% highest posterior density (HPD) intervals and the dot within each violin plot indicates the median. NRRs corresponding to the Spike gene are indicated in the top panels of b-c.

(95% highest posterior density [HPD]: 1998–2002; NRR 14 [Orf1b]; 1065 nt), only one year before the emergence of SARS-CoV-1 in 2002. The viruses most similar to SARS-CoV-1 within this NRR were YNLF_31C and YNLF_34C, sampled in Yunnan, China in 2013. Molecular clock inference informed by narrower rate priors resulted in times of the closest-inferred ancestors for several NRRs closer to the emergence of SARS-CoV-1 (Fig. C.5a, table C.3). Our findings were consistent when we included a more divergent human SARS-CoV-1 genome along with a masked palm civet (Paguma larvata) SARS-CoV-1 genome (Fig. C.5a, table C.3).

We similarly used the evolutionary rate of SARS-CoV-2 across the inferred NRRs of the SARS-CoV-2-like viruses (Fig. C.4, table C.2) as priors for the phylogenetic analysis of the NRRs of the SARS-CoV-2-like viruses, comprising SARS-CoV-2 and the 26 SARS-CoV-2-like genomes. The time of the closest-inferred bat virus ancestor of SARS-CoV-2 ranged from several years to several decades preceding late 2019 — when SARS-CoV-2 was introduced into humans — across the 27 NRRs (Fig. 3.1c). Although the median time of the closest-inferred ancestor of SARS-CoV-2 was in 2009, the most recent time of the closest-inferred ancestor across all NRRs was in 2017 (95% HPD: 2014–2019; NRR 3 [Orf1a]; 259 nt), only two years prior to its introduction into humans[145]. The genome most similar to SARS-CoV-2 within this NRR is RmYN02, sampled in Yunnan, China in 2019. We note that although NRR 3 is only 259 nt, NRR 6 is 573 nt and has a similar time of the closest-inferred ancestor of 2015 (95% HPD: 2009–2019). Within NRR 6, the genome most related to SARS-CoV-2 is BANAL-20-103, sampled in Laos in 2020. A narrower clock rate prior results in similar inferred dates (Fig. C.5b, table C.4).

The times of closest-inferred ancestor of SARS-CoV-1 across the NRRs are nearer to the date of human emergence than those for SARS-CoV-2 (Fig. 3.1b,c): 14 NRRs of the SARS-CoV-1-like viruses have a median time of the closest-inferred ancestor of SARS-CoV-1 within 5 years of the emergence of SARS-CoV-1, whereas only 3 NRRs of the SARS-CoV-2-like viruses have a closest-inferred ancestor of SARS-CoV-2 within 5 years of the emergence of SARS-CoV-2. Both sets of viruses consistently have more divergent fragments across the Spike gene (Fig. 3.1b). The oldest time of the closest-inferred bat virus ancestor within a single NRR of SARS-CoV-1 and SARS-CoV-2 was, respectively, 1963 (95% HPD: 1934–1985; NRR 19 [Orf1b]) and 1982 (95% HPD: 1946–2007; NRR 22 [Spike]), with each approximately 40 years prior to the emergence of

their corresponding SARS-CoV. Notably, only NRR 19 among the SARS-CoV-1-like viruses has a median time of the closest-inferred ancestor more than 50 years earlier than the most recently sampled SARS-CoV-1-like virus.

Accurate estimation of the ages of deeper nodes in the SARS-CoV phylogenies require complex models accommodating high levels of substitution saturation[137;141]. We restricted our analyses to the most recent nodes of the tree. Our inferred ages for the closest-inferred ancestors of each of the SARS-CoVs across the NRRs indicate that several of the published sarbecovirus genomes include non-recombinant fragments that are descendant from viruses that circulated likely only a few years before the emergence of SARS-CoV-1 and SARS-CoV-2.

3.4 The recombinant common ancestors became more similar to SARS-CoV-1 and SARS-CoV-2 with increased sampling

We quantified the extent to which the publication of more viruses closely related to SARS-CoV-1 and SARS-CoV-2 has contributed toward our understanding of the ancestor of each of these human viruses. The respective recombinant common ancestor ("recCA") of SARS-CoV-1 and SARS-CoV-2 can be understood as the aggregate of the closest-inferred ancestors of each of the SARS-CoVs across each of its NRRs[145], with each closest-inferred ancestor existing at different points in time and with potentially different viruses most closely related to the SARS-CoV. The recCA therefore accounts for the closest relative(s) across all non-recombinant segments. We inferred the genome of the recCA of each SARS-CoV over time, progressively adding published genomes to the reconstruction based on their publication date. Then, we calculated the similarity of each recCA to its respective SARS-CoV as a function of the publication date to examine how additional closely related genomes can affect ancestral reconstruction and reflect our increasing understanding of sarbecoviruses over time (Fig. 3.2a).

We observe that, as more genomes were published, the SARS-CoV-1 recCA and SARS-CoV-2 recCA became more similar to SARS-CoV-1 and SARS-CoV-2, respectively (Fig. 3.2). The full bat virus genomes most similar to SARS-CoV-1 and SARS-CoV-2 are, respectively, YN2020E (96.2% genetic identity to SARS-CoV-1) and BANAL-20-52 (96.8% genetic identity

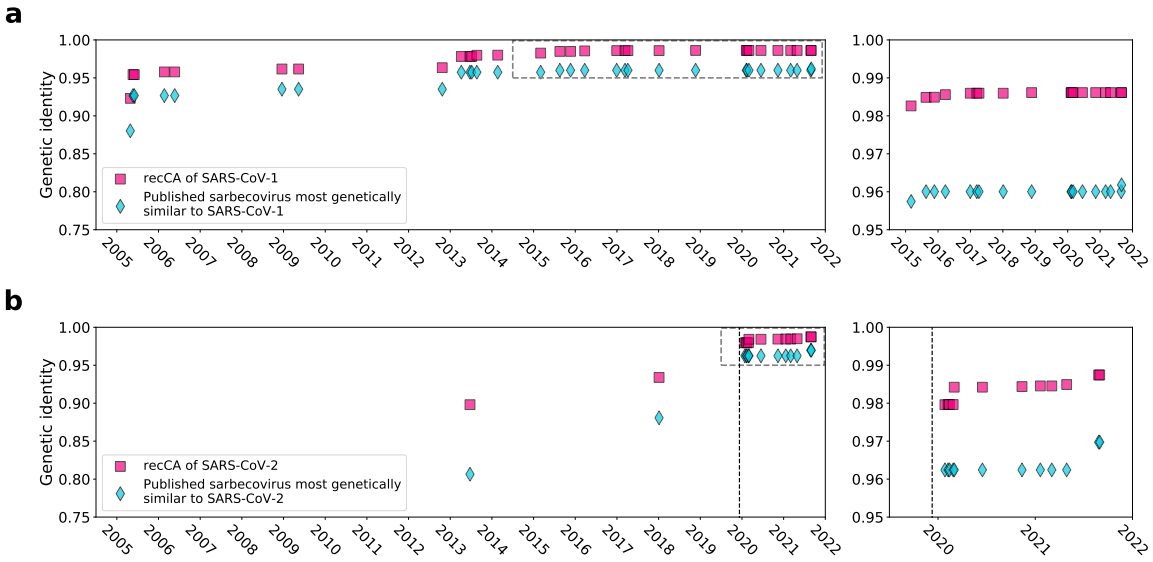


Figure 3.2. The recCA (recombinant common ancestor) and the most closely related sarbecovirus over time. a, The similarity of the recCA of SARS-CoV-1 and the most closely related sarbecovirus in the SARS-CoV-1 viruses to SARS-CoV-1 as a function of time. **b,** The similarity of the recCA of SARS-CoV-2 and the most closely related sarbecovirus in the SARS-CoV-2 viruses to SARS-CoV-2 as a function of time. The right panels for **a** and **b** are zoomed-in panels of the dashed box in the left panels. The dashed vertical line in **b** is the sampling date of the earliest sampled genome of SARS-CoV-2.

to SARS-CoV-2), both published in 2021. The genetic identity of the recCA of SARS-CoV-1 with SARS-CoV-1 exceeded the genetic identity of YN2020E with SARS-CoV-1 by 2013. The genetic identity of the recCA of SARS-CoV-2 with SARS-CoV-2 exceeded the genetic identity of BANAL-20-52 with SARS-CoV-1 in 2020, coinciding with the publication of the full genome of RaTG13 (96.2% genetic identity to SARS-CoV-2) in January 2020. The recCA of each SARS-CoV therefore exceeded the genetic similarity of the most closely related full genome before it was actually sampled.

Furthermore, the recCAs continued to become more similar to their corresponding SARS-CoV in the subsequent years. For example, the recCA of SARS-CoV-2 continued to increase in similarity after RaTG13 and before BANAL-20-52 were published, despite the interim published genomes sharing less overall genetic identity with SARS-CoV-2 than RaTG13 (Fig. 3.2b). This increase in similarity is the result of these genomes containing fragments that descend from a more closely related ancestor despite the rest of the genome being more divergent from

SARS-CoV-2 than RaTG13. These results can therefore be observed at the level of individual NRRs: sequentially published genomes increase the genetic identity of some of the NRRs of the recCA to the matching regions of SARS-CoV-2 (Fig. C.6). However, not all NRRs of the recCA exhibit increases in genetic identity to SARS-CoV-2 with additionally published genomes, and occasionally there is no increase of genetic identity to SARS-CoV-2 across any of the NRRs of the recCA when a new genome is published. Regardless, our results show that, by the end of 2022, the inferred recCA of each SARS-CoV is very similar to these human viruses, with the recCA of SARS-CoV-2 sharing 98.8% genetic identity with SARS-CoV-2 and the recCA of SARS-CoV-1 sharing 98.6% genetic identity with SARS-CoV-1.

3.5 Phylogeography of bat sarbecoviruses related to SARS-CoV-1 and SARS-CoV-2

The phylogenies of the sarbecoviruses have an appreciable degree of geographic structuring linked to their hosts' ranges (Fig. 3.3)[136]. To infer the spatiotemporal spread of the SARS-CoV-1-like and SARS-CoV-2-like viruses, we simultaneously fitted a spatially-explicit phylogeographic model to each NRR. As the sampling locations of SARS-CoV-1, SARS-CoV-2, and the pangolin sarbecoviruses likely do not represent where their direct bat virus ancestors circulated[146–149], we excluded their locations from the primary phylogeographic analyses to avoid the impact of dispersal in non-bat hosts (Fig. C.2).

Due to substitution saturation, the deeper evolutionary history of rapidly evolving viruses, including sarbecoviruses, are underestimated using standard molecular clock approaches[138;141;150;151]. To account for the effects of substitution saturation on the SARS-CoV-1-like and SARS-CoV-2-like phylogenies when trying to understand the long-term dispersal patterns of sarbecoviruses, we applied the Prisoner of War (PoW) model[138]—a mechanistic molecular clock model that estimates divergence times while accounting for time-dependent evolutionary rates—to the posterior phylogenies with phylogeographic estimates scaled in units of genetic distance (Fig. C.1; see Methods). We find that substitution saturation primarily affected the inferred time of divergence events at least 50 years before the most recently sampled virus in our dataset, with younger internal nodes having relatively unchanged ages and indicating that our inferences of the time of

the closest-inferred ancestors are unbiased (Fig. C.7). The PoW-transformed phylogenies had tMRCAs that were at least several hundred to tens of thousands of years older (Fig. C.8).

To investigate the dispersal dynamics and history of the two sets of viruses, we computed their associated lineage diffusion coefficients[152] (Fig. 3.3e,f and Fig. C.2) and separately represented the PoW-transformed phylogeny of each NRR across Asia (Fig. 3.3g,h, Fig. C.9-C.12). Both sets of viruses appear to have a few long lineage dispersal events, but the number of these dispersals is likely an underestimate due to undersampling of sarbecoviruses from various regions in Asia. The inferred dispersal history of viral lineages looks consistent among NRRs (Fig. 3.3g,h, Fig. C.9-C.12): we observe a similar spatial distribution of the sets of viruses and their connections on the different NRR maps. Moreover, the phylogeographic patterns appear to be spatially structured, with phylogenetically similar viruses being generally sampled in geographically similar regions and almost no back and forth travel of viral lineages across relatively large regions of the study area. Our analyses reveal a moderate to strong correlation between the patristic distance—the sum of the branch lengths connecting two nodes on a tree—and great-circle geographic distance computed for each pair of bat virus samples: a Spearman correlation coefficient of 0.45 for SARS-CoV-1-like viruses (95% HPD: 0.36-0.59) and of 0.66 for SARS-CoV-2-like viruses (95% HPD: 0.12-0.83) (see table C.5 for a comparison with other viruses).

Either the root (Fig. 3.3d) or a deep ancestral node (see Data and Code Availability) of the SARS-CoV-2-like phylogenies consistently corresponds to the split between Rc-o319, sampled in Japan in 2013, and virus relatives from China and Southeast Asia for most NRRs. Our PoW-calibrated analysis gives a mean date of 8259 BCE across all NRR tMRCA median heights for this node (range: 23,777 BCE–1381 CE), providing further support for the ancient origins of the SARS-CoV-2-like viruses and therefore sarbecoviruses as a whole[137;138]. Conversely, although most NRRs of each set of viruses had a tMRCA of at least several thousands of years old, NRRs 20 (Orf1b) and 31 (N, Orf10) of the SARS-CoV-1-like viruses had tMRCAs of 1429 CE (95% HPD: 970 CE–1642 CE) and 1730 CE (95% HPD: 1625 CE–1828 CE), respectively, suggesting that these NRRs experienced selective sweeps, reducing genomic diversity in these genomic regions in the relatively recent past.

3.6 Location of the closest-inferred ancestors of SARS-CoV-1 and SARS-CoV-2

The PoW transformation shows that the timing of divergence events in the last 50 years inferred from standard molecular clock models are unbiased (Fig. C.7). However, as the PoW model enforces contemporaneous tips, it can produce unreliable results for the most recent inferred ages. Therefore, we performed a separate tip-dated phylogeographic analysis to infer where the closest-inferred ancestors of each SARS-CoV circulated in the past 50 years (Fig. C.2). We find that the closest-inferred ancestors of SARS-CoV-1 likely circulated in western China (i.e., Yunnan, Sichuan, or Guizhou). The closest-inferred ancestors of SARS-CoV-2 likely circulated in Yunnan, China or northern Laos, overlapping with a set of contiguous cave structures extending through these regions[153;154]. Our results indicate that both SARS-CoV-1 and SARS-CoV-2 emerged in humans over a thousand kilometers from where their closest-inferred bat virus ancestors likely circulated. The inferred geographic location of the closest-inferred ancestors of each SARS-CoV additionally overlapped with regions with moderate Rhinolophid (horseshoe bat) species richness (Fig. 3.3b and Fig. C.13–C.15), including in regions known to harbor *R. affinis* and *R. ferrumequinum* (Fig. C.13), and the closest-inferred ancestor of SARS-CoV-2 additionally overlapped with known locations of *R. malayanus* circulation (Fig. C.14). Including the location of the human and pangolin SARS-CoVs as a robustness analysis resulted in the inferred regions where the closest-inferred ancestors of both SARS-CoV-1 and SARS-CoV-2 circulated to include where these viruses also emerged (Fig. C.16, C.17); however, these inferences are likely biased by the non-bat travel (e.g. intermediate hosts associated with wild and farmed animal supply chains) of the ancestors of SARS-CoV-1 and SARS-CoV-2 and trafficking[148] of the pangolin hosts.

3.7 Sarbecoviruses spread at rates similar to their bat hosts

To investigate how fast the SARS-CoV-1-like and SARS-CoV-2-like viruses recently spread across Asia, we inferred the weighted diffusion coefficient, a measure of the diffusivity of viral transmission, for the 50 years before the most recently sampled genome in each dataset using our tip-dated phylogenies (Fig. C.2). We found that the SARS-CoV-1-like viruses spread

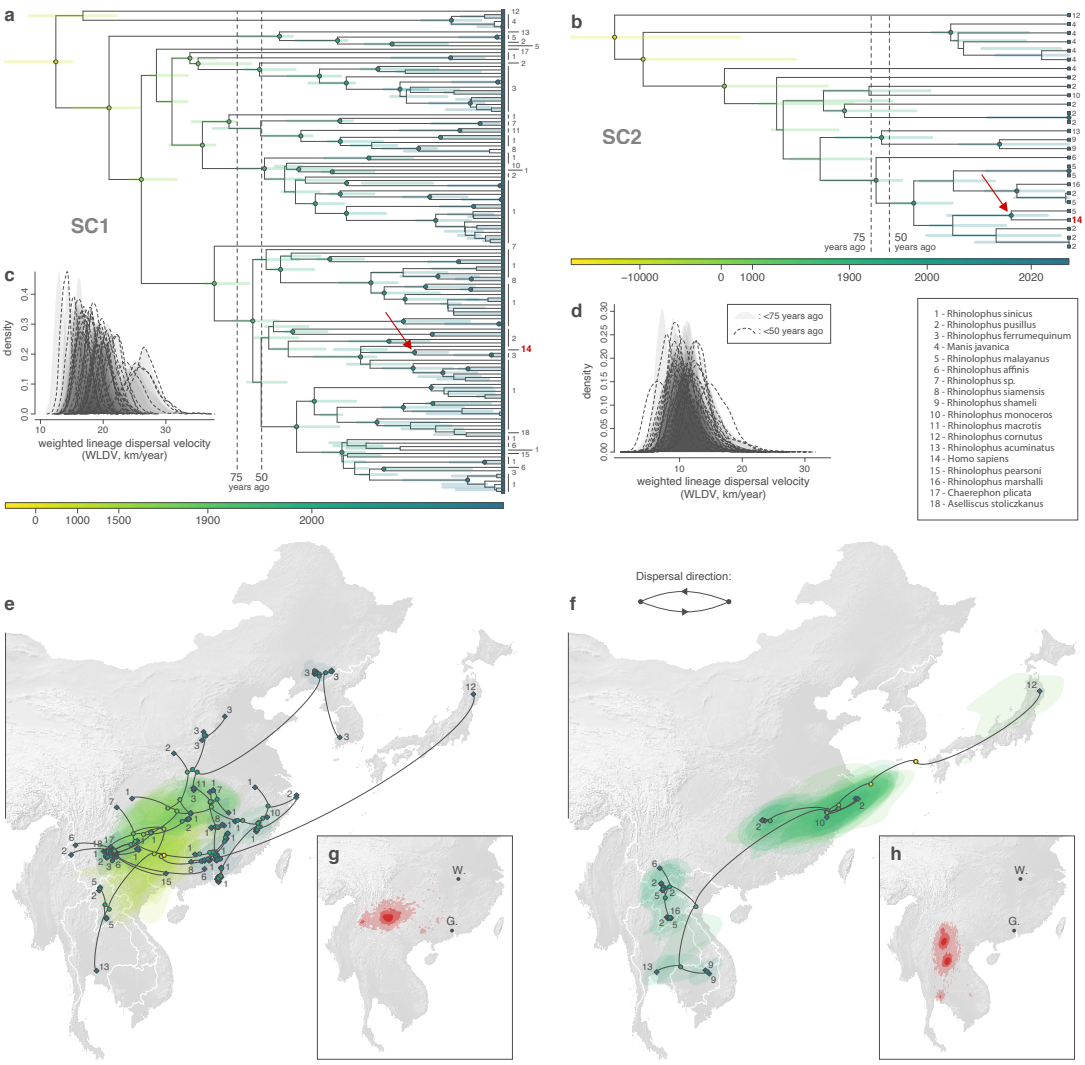


Figure 3.3. Estimated evolutionary history of SARS-CoV-1-like (SC1, on the left) and SARS-CoV-2-like (SC2, on the right) viruses. a, Global map with the study area highlighted in dark gray. **b,** Rhinolophid (horseshoe bat) species richness with the scale bar showing the number of co-occurring Rhinolophid species per km^2. **c-d,** time-scaled maximum clade credibility (MCC) phylogenies obtained after correction for substitution saturation, shown on a log-transformed time scale. Horizontal segments displayed at each internal node indicate the 95% HPD interval of the estimated node age. Closest-inferred ancestor indicated by a red arrow. **e-f,** weighted diffusion coefficient estimates based on phylogenetic branches occurring less than 75 (plain light gray curves) and 50 (dashed curves) years ago. **g-h,** reconstruction of the dispersal history of viral lineages for two NRRs (NRR14 for SARS-CoV-1-like viruses, and NRR3 for SARS-CoV-2-like viruses) obtained through continuous phylogeographic inference (see Fig. C.9-C.12 for all other NRRs). Nodes are coloured from yellow (oldest most recent common ancestor) to blue (most recent collection date) and are superimposed on 80% HPD polygons coloured according to the same log-transformed time scale, reflecting the uncertainty of the Bayesian phylogeographic inference. 80% HPD polygons were only computed and reported for the last 1,000 years. Dispersal direction (anti-clockwise) of viral lineages is indicated by the edge curvature. In panels **c-d-g-h**, tip nodes are displayed as squares (as opposed to dots for internal ones) and are associated with a number indicating the host species from which the virus was sampled. The host species are numbered by decreasing order of sampling frequency. **i-j,** estimated position of the closest-inferred bat virus ancestors of SARS-CoV-1 (**i**) and SARS-CoV-2 (**j**) based on the continuous phylogeographic reconstruction of all NRRs. We display the 95%, 75% and 50% HPD regions with an increasing red color darkness.

with a weighted diffusion coefficient of 2884 km^2/yr (95% HPD: 1654–4358; Fig. 3.3e), and the SARS-CoV-2-like viruses circulated with a weighted diffusion coefficient of 2050 km^2/yr (95% HPD: 233–5470; Fig. 3.3f). The SARS-CoV-1-like viruses were likely inferred to spread, on average, slightly more quickly than the SARS-CoV-2-like viruses because of the recent dispersal events to South Korea. These results were robust to the inclusion of the locations of the human and pangolin SARS-CoVs and were similar to the diffusion coefficients calculated for the 75 years preceding the most recently sampled genomes (table C.5). Horseshoe bats, the primary hosts of sarbecoviruses, tend to forage within around 2-3 km of their roost[155–158] and have a reported diffusion coefficient[159] of 1999 km^2/yr. Therefore, both the SARS-CoV-1-like and SARS-CoV-2-like viruses (in the context of host infections) traveled at a similar rate as their hosts, suggesting that these viruses are able to efficiently explore new niches, likely due to the density of Rhinolophid roosts.

3.8 Direct ancestors of the SARS-CoVs likely could not have reached sites of emergence via the bat reservoir alone

To understand whether the bat virus ancestors of each SARS-CoV could have spread to where the SARS-CoVs emerged in humans via dispersal of SARS-related coronaviruses across bat populations, we determined the minimum distance from the location of the closest-inferred bat virus ancestor to the province of emergence (Fig. 3.3i,j), and given the time between emergence and this ancestor (Fig. 3.1b,c), we calculated the dispersal velocity this implies. We compare this rate to the dispersal velocity of bat virus lineages to obtain a posterior rank distribution (Fig. C.18). For the closest-inferred bat virus ancestor of SARS-CoV-1 to have reached Guangdong Province, it would have needed a dispersal velocity greater than 95% of the dispersal velocities of bat virus lineages across 8 NRRs, including NRR 14 (Fig. C.18a). Similarly, for the direct bat virus ancestor of SARS-CoV-2 to have reached Hubei Province, it would have needed a dispersal velocity greater than 95% of the dispersal velocities of bat virus lineages across 7 NRRs, including NRRs 3 and 6 (Fig. C.18b) and NRRs with a time of the closest-inferred ancestor as early as 2009 (Fig. 3.1c). Considering the large distance separating the closest-inferred bat virus ancestors from where SARS-CoV-1 and SARS-CoV-2 each emerged (Fig. 3.3i,j) and the

high dispersal velocities necessary to traverse the aforementioned distance (Fig. C.18), it is very unlikely that the lineages descending from the closest-inferred bat virus ancestors of SARS-CoV-1 and SARS-CoV-2 reached, respectively, Guangdong Province and Hubei Province solely via dispersal of these viruses through their bat reservoirs.

3.9 Discussion

Our study reveals viral genomic sequence fragments closely related to the very recent ancestors of SARS-CoV-1 and SARS-CoV-2 are present in published genomes sampled from Rhinolophid (horseshoe) bats. By reconstructing the evolutionary history of the non-recombinant fragments of sarbecoviruses, we show that the ancestors of SARS-CoV-1 and SARS-CoV-2 circulated in bat populations hundreds to thousands of kilometers away from the sites of, and as recently as 1–3 years prior to, the emergence of these viruses in humans.

Sarbecoviruses undergo extensive recombination and tend to share breakpoints at specific genome regions, so-called recombination hotspots[35;160]. Although whole-genome comparisons previously suggested there may have been decades of separation between the closest-inferred bat virus ancestor and each of the SARS-CoVs[161], the recency of the closest-inferred ancestor in multiple subgenomic regions indicate that whole-genome comparisons are an unreliable measure of evolutionary divergence. It is, therefore, the non-recombinant fragments that must be analyzed to properly understand the emergence of the human SARS-CoVs[135;136]. We additionally show that adding new genomes to the recombination-aware ancestor reconstruction brought the sequence identity of the recombinant common ancestor of each SARS-CoV to its respective human virus, closer to 100%. This was true even though the new genomes did not always share a higher whole-genome identity to the SARS-CoVs than previously published sarbecovirus genomes.

The phylogeographic approach implemented here provides detailed insights into the sarbecoviruses' extensive historical movement across China and Southeast Asia. We find that the viral lineages diffuse at a similar rate to their primary hosts: small horseshoe bats. These bats typically have small home ranges (2-3 km^2 per night), are rarely migratory[157;162], and have a limited dispersal capacity with scant evidence of longer migrations (and limited to species in temperate rather than tropical regions)[163;164]. After accounting for recombination, the closest-

inferred bat virus ancestors of SARS-CoV-1 and SARS-CoV-2 are inferred to have circulated in regions spanning from Northern Laos to Western China.

Considering that the closest-inferred bat virus ancestors circulated in these regions almost immediately preceding the emergence of SARS-CoV-1 and SARS-CoV-2, our results suggest that there would not have been enough time for the direct bat virus ancestor to reach the location of emergence of the human SARS-CoVs via normal dispersal through bat populations alone. SARS-CoV-1 is thought to have been moved from bats in Yunnan province to Guangdong province via intermediate mammalian hosts[146]. The movement of palm civets and raccoon dogs as part of the wild and farmed animal trade has been implicated in the emergence of SARS-CoV-1[165], suggesting that human-mediated transport of these animals can allow for viral lineages to move faster than their bat hosts (Fig. C.18). Given (i) the presumed emergence of SARS-CoV-1 through the animal trade, (ii) the recency and location of the closest-inferred ancestor relative to SARS-CoV-2, and (iii) the clear evidence that the epicenter of the SARS-CoV-2 pandemic at one of only four markets in Wuhan that sold live wildlife from plausible intermediate host mammal species[166;167], the direct ancestor of SARS-CoV-2 likely moved from an area in or around Yunnan province, to Hubei province, via the wild and farmed animal trade.

The similar rates of diffusion among viral lineages and horseshoe bats suggest these viruses are able to rapidly and efficiently explore their environmental niches. The sarbecoviruses are therefore being transmitted among dense bat populations of the same and different species, as is also evidenced by the high levels of recombination in bat sarbecoviruses. For example, viral movement is likely to take place through horseshoe bats regularly moving among roosts, especially as roost switching tends to happen during breeding season or if bats travel or migrate to hibernacula[158;168]. Furthermore, in hibernacula horseshoe bats will roost in very dense aggregations with different species[169], providing the potential for spillover between bat species whilst bodily functions are downregulated and body temperatures can reach very low levels[169;170]. Heterothermy during hibernation (and in some species even sleep) and variable temperatures among species may also impact arousal patterns and vulnerability to pathogens[171;172].

Despite limited evidence for long-distance travel of horseshoe bats, there are several relatively long-distance viral lineage dispersals between nodes in the phylogenies. These geo-

graphically distant nodes therefore likely represent a lack of sarbecovirus sampling from areas between the distant locations rather than long-range bat movements, though it could also indicate dispersal from more migratory conspecific bat species, such as *Chaerephon* which can share roosts. For example, internal branches of the SARS-CoV-2-like tree represent movement from the east of China (Zhejiang) to the west (Yunnan), to Thailand and Cambodia (Fig. 3.3h). There are likely many more, still-undiscovered SARS-CoV-2-like virus populations that would bridge the geographic gap between these regions (e.g., in northern Vietnam, Guizhou, and Guangxi) as diversity of bats remains high across these regions (Figure 3i).

By accommodating time-dependent rate variation in our phylogeography analysis, we are able to infer the deeper evolutionary history of sarbecoviruses more accurately than previous efforts. For example, one ancestral node connecting distant regions is that between the Chinese and Southeast Asian SARS-CoV-2-like viruses and the Japanese sarbecoviruses (Rc-o319 represented in our dataset, but more viruses similar to SARS-CoV-2 have been recently sampled in Japan[173]), and we infer an average estimate of 8,259 BCE for this node (the oldest median estimate—NRR 21—being earlier than 20,000 years BCE). Rhinolophid bats will not fly over the open sea deliberately[174–177], but can be blown over narrow ocean channels or travel on ships[178]. However, during the Last Glacial Maximum (LGM), until approximately 8,000 BCE, Japan was likely connected to the mainland through land bridges both in the north (connecting Hokkaido to Siberia) and south (connecting Kyushu to China)[179]. This most recent time when Japan was connected to the mainland aligns well with our dating estimates of when the Japanese bat sarbecovirus clade separated from the mainland clade. Additionally, there is evidence of mammals moving through the land bridges during the LGM[180]. The bat species endemic to Japan known to harbour sarbecoviruses, *R. cornutus*, and its phylogenetically closest mainland species infected by SARS-CoV-2-like viruses, *R. pusillus*, are estimated to have shared a common ancestor 13.3 million years ago[181]. This much more ancient separation between the host species indicates that the viruses did not co-evolve with the bats. Instead, mainland bats likely transmitted the ancestral SARS-CoV-2-like viruses to the Japanese bat populations through movement via land bridges during the LGM.

Whilst horseshoe bat ranges extend across most of China, diversity decreases at increasing

latitudes, and virus sampling in bats has largely been limited to higher diversity areas, including Yunnan province in China and Laos. Additionally, other large areas with high diversity, such as Northern Vietnam, remain unsampled and host a similar community of bat species. In the absence of bat viruses sampled closer to where SARS-CoV-1 and SARS-CoV-2 emerged, the inference of the location of the closest-inferred bat virus ancestors is biased toward where published, closely related viruses were sampled. The SARS-CoV-1-like bat viruses are better represented (n=134) and therefore have more phylogeographic resolution than the SARS-CoV-2-like bat viruses (n=20). Although there are viruses closely related to the SARS-CoVs across most of the NRRs, the viruses are relatively more divergent across the NRRs that overlap with the Spike gene (Fig. 3.1, Fig. C.6). This relatively greater divergence suggests that Spike genes that can bind non-bat entry receptors circulate in lower frequencies in wild horseshoe bat populations and, therefore, are less likely to be sampled. Further sampling of closely related sarbecoviruses, methodological improvements in disentangling complex recombination patterns, and improved molecular clock calibration can allow for the inference of more detailed spatiotemporal trends and recombination patterns, as well as more complete evolutionary histories of each NRR.

Since the emergence of SARS-CoV-1 and SARS-CoV-2 in humans, the descendants of the bat viruses that gave rise to them have experienced years of evolution and almost certainly recombined with yet-unsampled lineages of the sarbecovirus tree. Hence, we should not expect to find the direct ancestor of SARS-CoV-1 or SARS-CoV-2 circulating in wild bats in future sampling. The search for, and analyses of, sarbecoviruses related to SARS-CoV-1 and SARS-CoV-2 have often focused on partial genomes, typically the RdRp protein[182–187], but future sequencing efforts should aim for whole genome sequences to detect and characterize all genomic fragments that descend from closely related ancestors. The non-recombinant segments of these mosaic genomes are pieces of a complex puzzle that hold the key to understanding the intricate mechanisms behind the emergence of past and future SARS-CoVs into the human population.

3.10 Data and code availability

Supplementary Data Files 1 and 2 are available at https://www.biorxiv.org/content/10.1101/2023.07.12.548617v1. The XML files, code, and newick trees are available at https:

//github.com/phylogeography/sarbecovirus_phylogeography.

3.11 Acknowledgments

S.D. acknowledges support from the *Fonds National de la Recherche Scientifique* (F.R.S.-FNRS, Belgium; grant n°F.4515.22) and from the Research Foundation - Flanders ('Fonds voor Wetenschappelijk Onderzoek - Vlaanderen', FWO, Belgium; grant n°G098321N). S.D. and P.L. acknowledge support from the European Union Horizon 2020 project MOOD (grant agreement n°874850). P.L., A.M. and M.A.S. acknowledge support from the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422-ReservoirDOCS), from the Wellcome Trust through project 206298/Z/17/Z and from the National Institutes of Health grants R01 AI153044, R01 AI162611 and U19 AI135995. P.L. also acknowledges support from the Research Foundation - Flanders ('Fonds voor Wetenschappelijk Onderzoek - Vlaanderen', G0D5117N and G051322N). T.I.V. acknowledges support from the Branco Weiss Fellowship. D.L.R. and J.H. acknowledge funding from the UK Medical Research Council (MRC, MC_UU_12014/12, MC_UU_00034/5 and MR/V01157X/1). S.L. is funded by an MRC studentship. M.G. is supported by the Biotechnology and Biological Science Research Council (BBSRC) grant number BB/M011224/1. J.E.P. acknowledges support from NIH (T15LM011271) and the UC San Diego Merkin Fellowship. This work was funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), Department of Health and Human Services (contract no. 75N93021C00015 to M.W.). J.O.W. acknowledges support from NIH (R01AI135992). We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared through GISAID the viral genomic sequences and metadata on which this research is based (Supplementary Data File 2). We thank Kristian Andersen, Alexander Crits-Christoph, Fabian Leendertz, Niema Moshiri, Richard Neher, and Andrew Rambaut for fruitful conversations and insight. We gratefully acknowledge support from Advanced Micro Devices, Inc. with the donation of parallel computing resources used for this research.

Chapter 3, in full, is a reformatted reprint of the material as it appears in "The recency and geographical origins of the bat viruses ancestral to SARS-CoV and SARS-CoV-2"; Jonathan

E. Pekar, Spyros Lytras, Mahan Ghafari, Andrew F. Magee, Edyth Parker, Jennifer L. Havens, Aris Katzourakis, Tetyana I. Vasylyeva, Marc A. Suchard, Alice C. Hughes, Joseph Hughes, David L. Robertson, Simon Dellicour, Michael Worobey, Joel O. Wertheim, and Philippe Lemey. *bioRxiv*, 2023. The dissertation author was the primary investigator and co-first author of this paper.

3.11.1 Author contributions

J.E.P., S.L., D.L.R., M.W., J.O.W., and P.L. conceived and planned the research. J.E.P., S.L., S.D., M.G., A.M., E.P., and P.L. analyzed and processed the data. A.M., T.I.V., M.A.S., J.H., S.D., J.O.W., and P.L. advised on statistical methodology. J.E.P., S.L., S.D., M.G., and P.L. designed and implemented genomic data processing pipelines. A.C.H modelled bat distributions. J.E.P., S.L., S.D., and P.L. wrote the first draft of the manuscript. All authors contributed to editing and interpreting the results. D.L.R., S.D., M.W., J.O.W, and P.L. jointly supervised the work.

3.11.2 Declaration of interests

S.L. is a scientific consultant for EcoHealth Alliance. M.A.S. receives contracts and grants from the US Food and Drug Administration, the US Department of Veterans Affairs, and Janssen Research and Development unrelated to this research. J.O.W. has received funding from the CDC (ongoing) through contracts or agreements to his institution unrelated to this research. J.E.P., M.A.S., M.W., and J.O.W. have received consulting fees and/or provided compensated expert testimony on SARS-CoV-2 and the COVID-19 pandemic.

Epilogue

Les fléaux, en effet, sont une chose commune, mais on croit difficilement aux fléaux lorsqu'ils vous tombent sur la tête. Il y a eu dans le monde autant de pestes que de guerres. Et pourtant, pestes et guerres trouvent les gens toujours aussi dépourvus.
(Everybody knows that pestilences have a way of recurring in the world, yet somehow we find it hard to believe in ones that crash down on our heads from a blue sky. There have been as many plagues as wars in history, yet always plagues and wars take people equally by surprise.)

- Albert Camus, *La Peste* (1947)

The aim of this dissertation was to understand the origins of the COVID-19 pandemic and the events leading up to it. Starting with developing a novel approach combining phylogenetics and epidemic simulations, I identified when the primary case of the COVID-19 pandemic was likely infected. I next extended this approach and rigorously examined the early SARS-CoV-2 genomic data to determine that the COVID-19 pandemic was likely the result of repeated introductions of SARS-CoV-2 into the human population. Lastly, I analyzed bat viruses closely related to SARS-CoV-1 and SARS-CoV-2 to determine when and where their recent ancestors circulated and demonstrated that bat movement alone could not have led to the emergence of SARS-CoV-1 and SARS-CoV-2 in their respective provinces. In additional work not expanded upon here, I contributed toward identifying the Huanan market as the epicenter of COVID-19 pandemic[167] and tracing the genomic diversity of potential intermediate hosts, including raccoon dogs (*Nyctereutes procyonoides*) and masked palm civets (*Paguma larvata*) (both susceptible to SARS-CoV-1, with the latter also implicated in its emergence[10]), of SARS-CoV-2 at the Huanan market[188]. Altogether, this body of work points toward a zoonotic origin of SARS-CoV-2 and implicates the wild and farmed animal trade in its emergence.

Although this dissertation has primarily focused on the emergence of SARS-CoV-2, the ideas present here extend to epidemics and pandemics more broadly. For example, the epidemic

modeling described here identified that a densely populated network is required for the sustained transmission of a pathogen with transmission dynamics similar to those of SARS-CoV-2. This is concordant with SARS-CoV-1 and SARS-CoV-2 first emerging in, respectively, Foshan and Wuhan, two densely populated cities in China, rather than in a remote or rural area. The finding that there were repeated introductions of SARS-CoV-2 into the human population at the start of the pandemic matches the observation of repeated zoonosis of other coronavirus epidemics and outbreaks in humans, including SARS-CoV-1, Middle East respiratory syndrome coronavirus (MERS-CoV), and porcine deltacoronavirus in Haiti[114–116]. In fact, SARS-CoV-2 has also repeatedly spilled over into humans after reverse-zoonosis to Syrian hamsters (*Mesocricetus auratus*)[124], American mink (*Neovison vison*)[121;122], and likely white-tailed deer (*Odocoileus virginianus*)[123;189]. Once a virus is able to infect and transmit among humans, sustained contact with the reservoir of the virus is likely to result in repeated zoonotic transmissions. In essence, the details relevant to one viral pandemic often apply to other viral zoonoses, epidemics, and pandemics.

People living in the vicinity of horseshoe bat species (*Rhinolophus spp.*) are more likely to show serological evidence of infection by SARS-related CoVs[56;102;190]. This observation underscores the continued threat of emergence of novel viruses from animal reservoirs, with wildlife markets posing a particularly high risk[10;191–193]. Although zoonoses leading to sustained epidemics will inevitably be discovered, those with little to no onward transmission are likely to occur without leaving any epidemiological trace[194], particularly with viruses associated with asymptomatic infections or mild disease, like SARS-CoV-2. This dissertation highlights the imperative of establishing intensive virus surveillance architectures. Such architectures will involve early detection of unexplained disease in humans, but must be complemented by a focus on wild, farmed, and traded animals, as well as additional sites of the human-animal interface, where the risk of transmission to humans is greatest. The ability to more rapidly identify spillovers and novel pathogens will improve our capacity to contain pathogens with pandemic potential.

Appendix A

Supplemental Material for Chapter 1

A.1 Materials and Methods

A.1.1 Sequence data

We queried the GISAID database[195] for SARS-CoV-2 viral genome alignment for sequences from mainland China that were not annotated as travel associated, as of 16 July 2020. We restricted our dataset to genomes that (i) were complete ≥29,000 nt, (ii) had high coverage with ≤0.5% unique amino acid mutations, (iii) had fewer than 1% 'N's, (iv) were not identified as potentially problematic via NextStrain, and (v) had a day-month-year sampling date reported. The first 200 and last 299 nucleotides were removed due to poor evidence of homology. The final alignment comprised 583 taxa. List of GISAID IDs are available in Data S1.

A.1.2 Phylogenetic inference

The phylogenetic history of SARS-COV-2 in China was first inferred in a maximum likelihood framework in IQ-TREE 2[196] using a GTR+F+I model, selected by model testing. Molecular clock analysis was conducted using a Bayesian Markov chain Monte Carlo (MCMC) approach in BEAST v1.10.4[49]. For the primary analysis, we employed a GTR+F+I substitution model, a strict molecular clock, a Bayesian skyline coalescent prior. To facilitate convergence, (i) a hard lower bound of $1x10^{-5}$ substitutions/site/year was placed on the clock rate and (ii) we initiated the MCMC using the maximum likelihood phylogeny that had been transformed into a chronogram via TempEst v1.5.3[197]. Four independent chains of 500 million generations were run, sampling every 25 thousand, and the first 15% were discarded as burnin. Convergence and mixing was assessed in Tracer v1.7.1[198] and chains were combined in LogCombiner, such that all

ESS values were >2070. The resulting posterior distribution comprised over 70,000 sampled trees to facilitate fuller exploration in the rejection sampling (see below). Evidence for shifting root tMRCAs after excluding the earliest sampled SARS-CoV-2 genomes was explored using TreeStat, a part of the BEAST package. Robustness analysis was conducted using a relaxed clock with an uncorrelated lognormal distribution (ULD) and a Bayesian Skygrid coalescent prior[199], in which 4 and 2 independent chains of 500 million generations were run, respectively. We did not explore the effect of highly structured coalescent priors (e.g., constant size, exponential growth), because the Skyline dynamics from China depict a complicated history with stable, exponential, and decreasing effective population sizes (Fig. A.15).

A.1.3 Epidemic Simulation

To explore the evolutionary dynamics at play during the beginning of the COVID-19 pandemic, we performed a series of epidemic simulations using FAVITES v1.2.6[51]. First, we generated static contact networks in FAVITES under a preferential-attachment model using the Barabási–Albert algorithm[200]. We used this network algorithm, because its scale-free properties recapitulate infectious disease spread. We chose to simulate a static contact network, because our focus is on the number of people infected at the beginning of the epidemic. For the primary simulations, we selected an intermediate value of 16 contacts per day (mean degree), based on Mossong *et al.*[201], within a contact network comprising 100,000 individuals (nodes).

Across this contact network, we performed a forward simulation SAPHIRE (Susceptible-Ascertained-Presymptomatic-Hospitalized-Not Ascertained (I)-Removed-Exposed) model[20] to generate a viral transmission network using GEMF[202]. We did not include the travel component of the original SAPHIRE model (i.e., individuals flying into and out of Wuhan), because our focus was on the early dynamics of the pandemic before its spread. Simulated epidemics started with a single seed infection among our 100,000 susceptible individuals. The epidemic was propagated using the parameters determined by Hao *et al.*[20](see Table A.1 for values) for 100 days. Our epidemics had a median doubling time of 4.1 days (95% range: 2.7–6.7 days), corresponding to epidemic growth in Wuhan between 01 January 2020 and 23 January 2020[20].

For the primary analysis, we ran 5000 epidemic simulations. Of these simulations, 29.7%

successfully established epidemics, defined as those simulations in which ≥1000 people had become infected and ≥1 person was still infectious at the end of the simulation. Failed epidemics were simulations that did not become established (i.e., 0 infectious people at the end of the simulation) or had fewer than 1000 people infected over the entire simulation; 70.3% of simulations failed to reach this epidemic threshold after 100 days. For the successfully established simulations, we constructed the coalescent history using the VirusTreeSimulator[203]package in FAVITES. For each infected individual in these simulations, a single viral lineage was randomly sampled uniformly across the duration of infection to represent viral genotype sampling. If a simulation failed to converge, it was rerun until it successfully achieved coalescence. The final output generated by FAVITES is the viral time-based phylogeny (part 2b in Fig. A.5). 1000 successfully established simulations were then randomly selected for further analysis. All inputs for the FAVITES primary analysis can be found in Table A.7 and the JSON input files are available at https://github.com/pekarj/SC2_Index_Case.

Each FAVITES simulation with the SAPHIRE model produced an output documenting when individuals transitioned from one compartment to another throughout the entire simulation. We used these to determine the amount of individuals in a given compartment (e.g., total infections, ascertained infections, unascertained infections, and hospitalized individuals) across each day in the simulation.

A.1.4 Determining Stable Coalescence

Once we had the time-based phylogenies, we labeled the internal nodes using the FAVITES helper script label_internal_nodes.py. We then extracted the tMRCA of infected individuals every day across each simulation using TreeSwift 1.1.14[204]. This tMRCA was calculated for each day of the 100 days or until 10,000 individuals had been infected, whichever came first. We chose not to explore dynamics after 10,000 infections due to a slowing in exponential growth arising from the saturation of the contact network.

We defined the stable coalescence as the coalescence that does not shift forward in time by more than one day, even as new individuals become infected and previously infected individuals recover. Therefore, the stable coalescence is reached the first day that the coalescence for the

currently infected individuals is within one day of the time of coalescence after the 100 day simulation or once 10,000 total individuals have been infected.

A.1.5 Sensitivity Analysis—Faster rate of infection

The same methods for epidemic simulations were performed to evaluate the dynamics of a more rapidly spreading virus. We used the aforementioned parameters, except the edge-based rates of transmission were increased (Table A.2). We produced 5000 simulations to generate at least 1000 established replicates with at least 1000 infected individuals each. Increasing the infectiousness coefficient to 0.55 per day produced a median epidemic doubling time of 3.1 days (95% range: 1.6-5.1).

A.1.6 Sensitivity Analysis—Slower rate of infection

The same methods for epidemic simulations were performed to evaluate the dynamics of a more slowly spreading virus. We used the aforementioned parameters, except the edge-based rates of transmission were increased (Table A.2). We produced 7000 simulations to generate at least 1000 established replicates with at least 1000 infected individuals each. Decreasing the infectiousness coefficient to 0.3 per day produced a median epidemic doubling time of 5.3 days (95% range: 3.6-7.5)(Table A.2).

A.1.7 Sensitivity Analysis—Higher average degree

The same methods for epidemic simulations were performed to evaluate the dynamics of a more densely connected network. We used the aforementioned parameters, except the average degree of the contact network was 26, and we rescaled the per-contact transmissibility in the SAPHIRE parameters by the mean degree such that the overall doubling time was maintained (Table A.2). We produced 5000 simulations to generate at least 1000 established replicates with at least 1000 infected individuals each.

A.1.8 Sensitivity Analysis—Lower average degree

The same methods for epidemic simulations were performed to evaluate the dynamics of a less densely connected network. We used the aforementioned parameters, except the average

degree of the contact network was 10, and we rescaled the per-contact transmissibility in the SAPHIRE parameters by the mean degree such that the overall doubling time was maintained (Table A.2). We produced 5000 simulations to generate at least 1000 established replicates with at least 1000 infected individuals each.

A.1.9 Sensitivity Analysis—Higher probability of hospitalization

The same methods for epidemic simulations were performed to evaluate dynamics involving higher probabilities of hospitalization. We used the aforementioned parameters, except with either 2-times or 10-times the probability of hospitalization (D_h^{-1} in the SAPHIRE model) in the main analysis. We produced 5000 simulations to generate at least 1000 established replicates with at least 1000 infected individuals each.

A.1.10 Sensitivity Analysis—Small village

The same methods for epidemic simulations in FAVITES were performed to evaluate the dynamics of a less densely connected network, except without rescaling the transmission rate with the average degree. We performed two sets of 5000 simulations, decreasing the mean degree to 8 (50% reduction) and 4 (75% reduction). We kept constant all other parameters from the primary analysis.

A.1.11 Sensitivity Analysis—Random contact network

The same methods for primary epidemic simulations in FAVITES were performed to evaluate the extinction dynamics across a random network. We used the aforementioned parameters from the primary analysis, except transmission was propagated across an Erdős-Rényi network (mean degree = 16), to produce 5000 simulations.

A.1.12 Sensitivity Analysis—Impact of subsampling on stable coalescence

For each of the FAVITES epidemic simulations from the primary analysis, we randomly subsampled 50% of the lineages circulating at the time of stable coalescence and calculated the new time of coalescence. This process was repeated 100 times for each epidemic simulation, and we used the median time of coalescence. This approach was also applied using 5% and 25% of

the circulating lineages at the stable coalescence.

A.1.13 Two-Phase Epidemics in FAVITES

We also used FAVITES to explore the evolutionary dynamics of a two-phase epidemic, in which the index case was infected with a less fit (i.e., less transmissible) variant that eventually went extinct (termed Phase 1 of the epidemic); however, before this less-fit variant went extinct, it gave rise to an adapted variant that became dominant in Wuhan and spread through China (termed Phase 2 of the epidemic). To simulate the first phase of this epidemic, we applied the same methods for epidemics simulations and parameters as in the primary analysis, except with a 50% lower transmission rate (b in Table A.1). We generated 5000 simulations in FAVITES and randomly selected 2000 epidemics that do not establish (i.e., went extinct) for further analysis. These simulations (Phase 1) were combined via rejection sampling (see below) with the primary established simulations (Phase 2), which would persist until reaching a stable coalescence.

A.1.14 Sensitivity Analysis—Varied transmission rates for Phase 1

The same methods for the two-phase simulations were performed to evaluate dynamics involving higher and lower rates of transmission for Phase 1. We used the aforementioned parameters, except with transmission rates of 1/6, 2/6, 4/6, and 5/6 the rate of transmission during Phase 2. We produced 5000 Phase 1 simulations, each of these lower transmission rates.

A.1.15 Sensitivity Analysis—Higher ascertainment rate for Phase 1

The same methods for the epidemic simulations were performed to evaluate dynamics involving higher rates of ascertainment (Table A.1; Table A.6). Increased ascertained is a proxy for increased virulence, as ascertained individuals are more likely to transmit the virus and become hospitalized. We used the aforementioned parameters, except with 2-times, 4-times, or 6-times the rate of ascertainment for Phase 1, relative to Phase 2, to produce 5000 simulations.

A.1.16 Sensitivity Analysis—Higher two-phase probabilities of hospitalization

The same methods for the epidemic simulations were performed to evaluate dynamics involving higher probabilities of hospitalization for either just the Phase 1 epidemic, or the

combined Phase 1 and Phase 2 epidemic. We used the aforementioned parameters, except with either 2-times or 10-times the probability of hospitalization for Phase 1 and Phase 2 simulations and 4-times the rate of ascertainment as the primary Phase 1 simulations.

A.1.17 Combining FAVITES and BEAST via Rejection Sampling

Our aim is to obtain a posterior distribution for the date of the index case in Hubei province, conditioned on X both the available sequencing data D_s and the date of the first reported COVID-19 case D_c. We do this in a Bayesian framework by marginalizing over the date Y of the first ascertained or unascertained COVID-19 case (see below) and the date Z of the tMRCA as follows:

$$P(X|D_S, D_C) = \int_Z P(X|Z, D_s, D_s)P(Z|D_s, D_s)\, dZ \tag{A.1}$$

We assume that the sequencing data are informative only for the tMRCA, i.e. given Z, X does not depend on D_S: $P(X|Z, D_S, D_C) = P(X|Z, D_C)$. We also assume that the first reported COVID-19 case data are not informative for the tMRCA: $P(Z|D_S, D_C) = P(Z|D_S)$. This gives:

$$P(X|D_S, D_C) = \int_Z P(X|Z, D_C)P(Z|D_S)\, dZ \tag{A.2}$$

We further note that $P(X|Z, D_c) = \int P(X, Y|Z, D_c)dY$, where we model $P(X, Y|Z, D_c)$ as proportional to $C(Y)P(X, Y|Z)$, where $C(Y)$ is a "consistency function" with a value of 1 when Y is consistent with D_c and 0 otherwise. This approach allows us to sample from the posterior distribution of Equation 2. The BEAST analysis provides values of Z sampled from the distribution $P(Z|D_s)$. For each sampled value of Z, we sample corresponding values of X and Y from the distribution $P(X, Y|Z, D_c)$ using the FAVITES simulation, providing samples from the distribution $P(X, Y|Z)$ in conjunction with a simple rejection sampling-based strategy: sample values from $P(X, Y|Z)$ until a sample is obtained for which $C(Y) = 1$. The resulting set of sample values for X then follow the posterior distribution $P(X|D_s, D_c)$.

We assign Y as the first instance of an ascertained (SAPHIRE stage: I) or unascertained

(SAPHIRE stage: A) (I/A: Fig. A.6) case and D_c as 17 November 2019, where $C(Y) = 1$ when Y precedes D_c. However, we note that the first ascertained/unascertained case will often be the index case themselves, unless a secondary or tertiary case progresses faster through the course of infection. Importantly, the rate at which cases were ascertained in the SAPHIRE model is based on real-time patterns in COVID-19 diagnosis from 01 through 22 January 2020 and may not reflect the actions that led to the retrospective diagnosis of earliest cases of COVID-19. Further, coalescence can happen any time after the index case is first infected, and there is no requirement for coalescence to occur after the first ascertained and unascertained individuals.

A.1.18 Rejection Sampling — two-phase epidemics

Rejection sampling for the two-phase epidemics involved randomly sampling (i) a Phase 1 and (ii) a Phase 2 epidemic simulations and combining them at (iii) a time point during Phase 1, representing the moment at which a more fit variant arose. This time when the adaptive variant evolved was chosen uniformly across the Phase 1 epidemic, weighted by the number of circulating lineages (Fig. A.9-2B). The first ascertained or unascertained case can come from either Phase 1 or Phase 2—whichever occurs first—and is used for the consistency function. The $C(Y)$ remainder of the rejection sampling method remains the same for the two-phase epidemic.

A.1.19 Sensitivity Analysis—ascertained and unascertained cases

The primary rejection sampling analysis was conditioned on an individual having an ascertained (SAPHIRE stage: I) or unascertained (SAPHIRE stage: A) infection prior to the date of the first reported case of COVID-19 on 17 November 2019. However, we also performed sensitivity analyses whereby the minimum date for the earliest case must be an ascertained case (Fig. A.6). We performed rejection sampling using the first ascertained or unascertained case (I/A; Fig. A.6) or solely the first ascertained case (I-only; Fig. A.6).

A.1.20 Sensitivity Analysis—date of first COVID-19 case

Sensitivity Analysis—date of first COVID-19 case. We also explored the sensitivity of the rejection sampling approach to the date before which an ascertained or unascertained case must have existed. We performed rejection sampling using 17 November 2019 or 01 December

2019 as the minimum date for the earliest case.

A.1.21 Sensitivity Analysis—relaxed clock and coalescent priors

We explored the sensitivity of the rejection sampling approach to different molecular clocks in the BEAST inference. We performed separate rejection sampling analyses using tMRCAs inferred under a relaxed ULD clock and a Skygrid coalescent prior.

A.1.22 Sensitivity Analysis—shifting the tMRCA

We explored the sensitivity of the rejection sampling approach to the stability of the timing of coalescence in the BEAST analysis (Fig. A.3A). The primary analysis used the inferred tMRCA of all viruses sampled on or after 01 January 2020, at which point the tMRCA had stabilized. Sensitivity analysis was conducted using the tMRCA of all sampled viruses (i.e., the root tMRCA). This tMRCA is within 1 calendar day in 78.5% of sampled trees (Fig. A.3B) and represented by the same node in 74.6% of sampled trees (Fig. A.3C).

A.2 Supplementary text

A.2.1 Relaxed molecular clock

Relaxed molecular clock. Strict molecular clocks can produce overly-precise tMRCA estimates when evolution occurred under a relaxed clock. Relaxing the molecular clock for SARS-CoV-2 in China using an uncorrelated lognormal distribution of rates (ULD) did produce a slightly wider tMRCA estimate, with a mean of December 6th (95% HPD: November 9th–December 22nd) (Fig. A.1A) with a comparable rate of 8.45×10^{-4} (95% HPD: 7.05×10^{-4}–9.89×10^{-4}). However, the standard deviation of the ULD was 0.0009 (95% HPD: 0.00005–0.00013), suggesting strong (strict) clock-like evolution across the SARS-CoV-2 phylogeny in China. Furthermore, 91.7% of the posterior tMRCA estimate post-dated the earliest reported case on 17 November 2019.

A.2.2 Phylogenetic rooting

The position of the root in the Bayesian phylodynamic inference under a strict molecular clock was ambiguous. In 63.5% of the posterior trees, the root fell within the basal polytomy of 79 identical genomes, exemplified by the Wuhan-Hu-1 reference genome (GenBank Accession MN908947). This position aligns with the assumption of the NextStrain algorithm (https://nextstrain.org/ncov/global). In 21.8% of posterior trees, the root fell on a branch which led to a single virus: the earliest sampled genome IPBCAMS-WH-01 (GenBank Accession MT019529) in 15.0% of trees and another early genome WH01 (GenBank Accession MT291826) in 2.4% of trees. This hypothetical root orientation corresponds with the plurality of estimates from Pipes *et al.*[50], though their approach did not make use of the molecular clock. Notably, we find very little support for a root position on branches corresponding to the T28114C (3.9%) or C8782T (1.4%) mutations, as previously suggested by Zhang *et al.*[40].

Rooting configurations were similar when using a relaxed ULD clock. The root fell among viruses with the Wuhan-Hu-1 reference genome sequence in 61.7% of the posterior sample, on the branch leading to IPBCAMS-WH-01 in 18.1% of samples, and on the branch leading to WH01 in 3.7% of samples. Again, there was little support for a root position on branches corresponding to the T28114C (2.5%) or C8782T (0.4%) mutations.

Rooting configurations were also similar when using a Skygrid coalescent prior. The root

fell among viruses with the Wuhan-Hu-1 reference genome sequence in 68.2% of the posterior sample, on the branch leading to IPBCAMS-WH-01 in 9.7% of samples, and on the branch leading to WH01 in 1.9% of samples. Again, there was little support for a root position on branches corresponding to the T28114C (3.3%) or C8782T (0.7%) mutations.

A.2.3 Time to stable coalescence in simulations

Across the simulated epidemics, a stable coalescence (i.e., non-shifting tMRCA from that point in time until the end of the simulated epidemic) was established after a median of 16 days after the index case was first infected; 95% of simulations reached a stable coalescence by day 53 post-index case infection (Fig. A.16A), indicating that the 100-day simulations were sufficient to determine time to stable coalescence. Importantly, the time at which stable coalescence was achieved represents when the tMRCA of the currently circulating lineages is within one day of the tMRCA of the lineages remaining at the end of the simulation. This point in time is when basal lineages ceased to be lost in the coalescent tree, not the tMRCA of the remaining viral lineages (Fig. A.16B). At the time a stable coalescence was reached, a median of 16 people had been infected; 95% of simulations reached this stable coalescence after 1194 people had become infected. Recall that the empirical phylogenetic analysis from the Chinese epidemic appears to have reached a stable coalescence by 01 January 2020, when around 1000 people are believed to have been infected[20].

A.2.4 Epidemic extinction

Most epidemics that go extinct, that is they do not establish themselves in the population, do so after a single infection by day 9, the average duration of a single infection in the SAPHIRE model (Fig. 1.4C, Table A.2). As the probability of epidemic establishment decreases, the maximum number of total infections during the simulated epidemic tends to increase. This pattern holds in networks with variable transmission rates and network connectivity (Table A.2). As the probability of viral transmission decreases, more infections can occur before an epidemic goes extinct.

A.2.5 Subsampling lineages at the stable coalescent

To match the real-world scenario of unseen lineages when calculating a tMRCA, we subsampled the circulating lineages in the FAVITES simulations at the stable coalescent to determine any potential effects on the timing of the index case for a one-phase epidemic. The subsampling was done at 5%, 25%, and 50%. The subsampled lineages were used to determine a new time of stable coalescence for the FAVITES simulations, and then rejection sampling was performed as in the primary analysis. The point of coalescence shifted slightly forward in time as sampling completeness decreased (Fig. A.17A), but the timing of the index case was robust and closely matched the primary analysis without subsampling (Fig. A.17B).

A.2.6 Varying the rate of transmission for Phase 1 of a two-phase epidemic

Although a less-fit variant would have been less transmissible than the variant that spread through China in late-2019, it is unclear how much less transmissible it would have been. Our two-phase analysis (Fig. 1.3F to K) assumed a relative fitness of one-half. To explore the robustness of inferred date of the index case to this assumption, we explored additional two-phase simulations varying the relative fitness of this Phase 1 variant by sixths: from 1/6 through 5/6 the transmission rate of the Phase 2 variant (Table A.1). The extinction probability varied substantially across these fitness variants (Table A.6). When the relative fitness in Phase 1 was 5/6 of that in Phase 2, 77.3% of simulated epidemics ended in extinction; however, when the relative fitness was 1/6 of that in Phase 2, 4999 out of 5000 simulated epidemics went extinct. Remarkably, the inferred timing of the index case was robust across all these relative fitness values (Fig. A.18). This consistency likely arises because our index case inference relied on only Phase 1 epidemics that went extinct, and the time from index case to extinction was similar across relative fitness values (Table A.6).

A.2.7 Varying the ascertainment rates for Phase 1 of the epidemic

It is possible that the less-fit variant during Phase 1 of a two-phase epidemic was also more virulent. Therefore, we explored this scenario by increasing the ascertainment rate of the Phase 1 variant to 2- and 4-times that of the ascertainment rate of the Phase 2 variant (Table

A.1). The results of the index case timing were similar across all these ascertainment rates (Table A.6; Fig. A.18). The proportion of simulations that went extinct decreased slightly as the ascertainment rate increased, because ascertained cases have a higher transmission rate than unascertained cases.

A.2.8 Number of hospitalized individuals through time

There has also been speculation that hospitals in Hubei province were inundated with COVID-19 patients in October and November 2019[66]. However, our primary simulation analysis conditioning on the first case being identified by 17 November 2019 suggests that an increase in hospitalizations due to COVID-19 would not have been notable until mid- to late-December 2019. In fact, we do not observe enough total infected people, let alone ascertained of hospitalized people, in October and November 2019 to overwhelm hospitals. We note that our estimate of 0 to 36 hospitalizations as of 01 January 2020 in the primary analysis is less than the 42 hospitalizations that had been previously reported[39]; however, this value is contained within the 95% HPD of many robustness analyses (Fig. A.13). To explore if increased probability of hospitalization early in the epidemic (due to different clinical practices or a more virulent virus), we explored increasing by two- and ten-fold the probability of hospitalization. The simulations also produced the 42 hospitalizations on 01 January 2020 in the 95% HPD (Fig. A.13). However, even these extreme deviations from observed parameters failed to produce high rates of hospitalization prior to December 2019, even in the 99% HPD (Fig. A.13). Similar results were seen in two-phase epidemic simulations, including a scenario in which the Phase 1 variant had both higher rates of ascertainment and hospitalization. To test the most extreme case, we simulated a Phase 1 epidemic with 4-times the ascertainment rate as Phase 2 and a ten-fold higher probability of hospitalization for Phases 1 and 2, but there are consistently fewer than 10 hospitalized individuals prior to 01 December 2019 (Fig. A.14). It is important to acknowledge that our SAPHIRE model is based on parameters from Hao *et al.* during the period between 01 January and 22 January 2020, which occurred after SARS-CoV-2 was discovered and while there was still limited understanding of COVID-19 dynamics[20]. Therefore, even though we tested a two-phase epidemic, higher rates of ascertainment, and higher probabilities of hospitalization,

we have little confidence in the precision of our estimates regarding the number of hospitalized patients in late-2019. Nonetheless, even across the various robustness analyses we explored, we never observed a substantial number of hospitalized patients in November 2019, even in the 99% extreme of our estimates (Fig. A.13, S14).

A.3 Supplementary Tables

Table A.1. Simulation parameters, with all parameters except for b based on Hao et al[20]. The b value listed is for the primary analysis.

Parameter	Meaning	Value
b	Transmission rate of ascertained cases	0.385
r	Ascertainment rate	0.15
α	Ratio of transmission for unascertained	0.55
D_e	Latent period (days)	2.9
D_p	Presymptomatic infectious period (days)	2.3
D_i	Symptomatic infectious period (days)	2.9
D_q	Duration from illness onset to isolation (hospitalization) (days)	21
D_h	Isolation (hospitalization) period (days)	30

Table A.2. Parameterization and success/failure of one-phase compartmental epidemic simulations in primary analysis and robustness analyses

Parameter	Primary analysis	Robustness analyses						
		Less infectious	More infectious	Densely connected	Sparsely connected	Small village 1	Small village 2	Erdos-Renyi
Mean No. of edges per node (degree)	16	16	16	10	26	8	4	16
Infectiousness (b)	0.385	0.3	0.55	0.385	0.385	0.385	0.385	0.385
Doubling time of established simulations[a]	4.1 (2.7-6.7)	5.3 (3.6-7.5)	3.1 (1.6-5.1)	4.2 (2.7-7.0)	4.2 (2.7-7.1)	8.5 (7.8-16.4)	11.4 (6.5-32.3)	13.8 (10.7-18.5)
Percentage failed	70.3	80.5	53.6	69.4	68.3	94.5	99.6	83.7
Days to extinction of failed simulations[a]	7.9 (1.5-40.4)	8.3 (1.6-45.7)	7.2 (1.4-34.2)	8.0 (1.5-42.7)	7.6 (1.4-40.1)	8.6 (1.7-51.3)	7.9 (1.8-39.9)	10.4 (1.7-75.0)
No. of infections of failed simulations[a]	1 (1-9)	1 (1-11)	1 (1-7)	1 (1-9)	1 (1-9)	1 (1-15)	1 (1-7)	2 (1-33)
Maximum No. of infections in failed simulations	44	69	20	29	33	110	565	101

[a]Median and 95% range in parentheses

Table A.3. Estimates for existence of lineages in one- and two-phase epidemics.

Type of Epidemic				One-phase	Two-phase	Two-phase	Two-phase	Two-phase
Analysis Framework				Index case infected at tMRCA (%)	Index case infected at tMRCA (%)	First adapted case infected at tMRCA (%)	Phase 1 infections exist at tMRCA (%)	Earliest case due is Phase 1 infection (%)
Earliest Case	Clock	Ascertained	Modification					
Nov 17	Strict	I/A	None	1.1	2.2	12	17	61
Nov 17	Strict	I-only	None	1	4	20	33	70
Dec 01	Strict	I/A	None	7.9	8	25	22	63
Dec 01	Strict	I-only	None	6.9	9.1	30	39	74

Table A.4. Number infected estimates for one-phase primary (first row) and robustness analyses, with the number infected reported on 17 November, 01 December, and 15 December 2019.

Analysis framework				Median # infected (95% HPD)		
Earliest Case	**Clock**	**Ascertained**	**Modification**	**17 Nov 2019**	**01 Dec 2019**	**15 Dec 2019**
Nov 17	Strict	I/A	None	4 (1-13)	9 (2-26)	21 (3-281)
Nov 17	Strict	I-only	None	7 (1-15)	13 (4-31)	24 (8-271)
Nov 17	Strict	I/A	No coalescent shift	4 (1-15)	10 (2-53)	23 (3-1052)
Nov 17	Strict	I/A	Faster transmission	3 (1-12)	7 (2-24)	17 (3-909)
Nov 17	Strict	I/A	Slower transmission	4 (1-17)	11 (2-38)	26 (4-241)
Nov 17	Strict	I/A	Densely connected	4 (1-15)	9 (2-29)	21 (4-268)
Nov 17	Strict	I/A	Sparsely connected	3 (1-16)	9 (2-32)	21 (4-345)
Nov 17	Strict	I/A	2x Hospitalization Probability	6 (1-18)	10 (3-26)	23 (6-250)
Nov 17	Strict	I/A	10x Hospitalization Probability	7 (2-23)	13 (4-40)	27 (8-390)
Nov 17	UCL	I/A	None	4 (1-13)	9 (2-29)	18 (3-402)
Nov 17	UCL	I-only	None	7 (1-15)	12 (4-33)	21 (8-401)
Dec 01	Strict	I/A	None	0 (0-9)	4 (1-20)	14 (2-276)
Dec 01	Strict	I-only	None	1 (0-12)	8 (1-25)	18 (2-325)
Dec 01	Strict	I/A	2x Hospitalization	1 (0-14)	6 (1-24)	18 (3-328)
Dec 01	Strict	I/A	10x Hospitalization	1 (0-14)	8 (1-31)	21 (3-322)
Dec 01	UCL	I/A	None	0 (0-9)	4 (1-24)	13 (2-392)
Dec 01	UCL	I-only	None	1 (0-12)	8 (1-28)	17 (2-451)

Table A.5. Number infected estimates for two-phase primary (first row) and robustness analyses, with the number infected reported on 17 November, 01 December, and 15 December 2019.

Earliest Case	Clock	Ascertained	Phase 1 ascertainment rate	Hospitalization probability ($1/D_h$)		Median # infected (95% HPD)		
				Phase 1	Phase 2	17 Nov 2019	01 Dec 2019	15 Dec 2019
Nov 17	Strict	I/A	1	1	1	3 (1-15)	7 (1-30)	17 (1-254)
Nov 17	Strict	I-only	1	1	1	5 (1-26)	10 (1-46)	20 (1-252)
Nov 17	Strict	I/A	4	1	1	3 (1-14)	7 (1-27)	17 (1-254)
Nov 17	Strict	I/A	4	2	1	3 (1-14)	8 (1-28)	17 (1-253)
Nov 17	Strict	I/A	4	2	2	3 (1-15)	8 (1-28)	18 (1-270)
Nov 17	Strict	I/A	4	10	1	3 (1-14)	8 (1-27)	18 (1-262)
Nov 17	Strict	I/A	4	10	10	3 (1-16)	8 (1-30)	19 (1-254)
Dec 01	Strict	I/A	1	1	1	0 (0-9)	3 (1-22)	12 (1-246)
Dec 01	Strict	I-only	1	1	1	1 (0-16)	5 (1-33)	14 (1-234)
Dec 01	Strict	I/A	4	1	1	0 (0-9)	3 (1-21)	12 (1-231)
Dec 01	Strict	I/A	4	2	1	0 (0-9)	3 (1-20)	12 (1-238)
Dec 01	Strict	I/A	4	2	2	0 (0-8)	3 (1-22)	12 (1-251)
Dec 01	Strict	I/A	4	10	1	0 (0-9)	3 (1-21)	12 (1-236)
Dec 01	Strict	I/A	4	10	10	0 (0-9)	3 (1-24)	13 (1-244)

Table A.6. Parameterization and success/failure of two-phase compartmental epidemic simulations in primary analysis and robustness analyses.

Parameter	Infectiousness robustness analysis					Ascertainment rate robustness analysis		
Infectiousness	1/6	2/6	3/6	4/6	5/6	3/6	3/6	3/6
Ascertainment rate	1	1	1	1	1	2	4	6
Percentage failed	>99.9	98.6	93.9	85.5	77.3	91.8	89.7	87.2
Days to extinction of failed simulations[a]	7.8 (1.7-30.8)	8.2 (1.7-46.9)	8.6 (1.8-51.1)	8.4 (1.5-48.8)	8.1 (1.6-41.6)	8.7 (1.7-56.3)	8.5 (1.7-62.3)	9.0 (1.5-71.7)
Number of infections of failed simulations[a]	1 (1-4)	1 (1-12)	1 (1-13)	1 (1-13)	1 (1-10)	1 (1-13)	1 (1-11)	1 (1-11)

[a]Median and 95% range in parentheses

Table A.7. FAVITES parameter values for the primary analysis. Rate values and end time are adjusted for years and density of contact network as needed.

Parameter[a]	**Interpretation**	**Value**
num_cn_nodes	Total number of nodes in network	100,000
num_seeds	Number of seeds (individuals in compartments E,P,I,A) at time 0	1
saphire_freq_s	Frequency of S individuals at time 0	99,999
saphire_freq_e	Frequency of E individuals at time 0	1
saphire_freq_p	Frequency of P individuals at time 0	0
saphire_freq_i	Frequency of I individuals at time 0	0
saphire_freq_a	Frequency of A individuals at time 0	0
saphire_freq_h	Frequency of H individuals at time 0	0
saphire_freq_r	Frequency of R individuals at time 0	0
saphire_s_to_e_seed	Target infection rate (S→E) from outside the contact network (i.e., seed infection)	0
saphire_e_to_p	Target rate of becoming presymptomatic (E→P)	365/2.9
saphire_p_to_i	Target rate of becoming ascertained (P→I)	0.15*365/2.3
saphire_p_to_a	Target rate of becoming unascertained (P→A)	(1-0.15)*365/2.3
saphire_i_to_h	Target rate of becoming hospitalized (I→H)	365/21
saphire_i_to_r	Target rate of becoming removed/recovered (I→R)	365/2.9
saphire_a_to_r	Target rate of becoming removed/recovered (A→R)	365/2.9
saphire_h_to_r	Target rate of becoming removed/recovered (H→R)	365/30
saphire_s_to_e_by_p	Target infection rate (S→E) by P	0.55*0.385*365/16
saphire_s_to_e_by_i	Target infection rate (S→E) by I	0.385*365/16
saphire_s_to_e_by_a	Target infection rate (S→E) by A	0.55*0.385*365/16
end_time	Time at which to end the transmission simulation	100/365
vts_model	Intrahost viral population growth model to use	constant
vts_growthRate	Target effective population size growth rate (used in exponential and logistic models)	0
vts_max_attempts	Maximum number of attempts to coalesce a single tree (e.g. 100) before FAVITES kills VirusTreeSimulator	1000
vts_n0	Target effective population size at time zero (used in all models)	1
vts_t50	Time point, relative to the time of infection in backwards time, at which the population is equal to half its final asymptotic value, in the logistic model	-99999

[a]FAVITES module choices and parameter values for robustness analyses can be found in the json files at https://github.com/pekarj/SC2_Index_Case.

A.4 Supplementary Figures

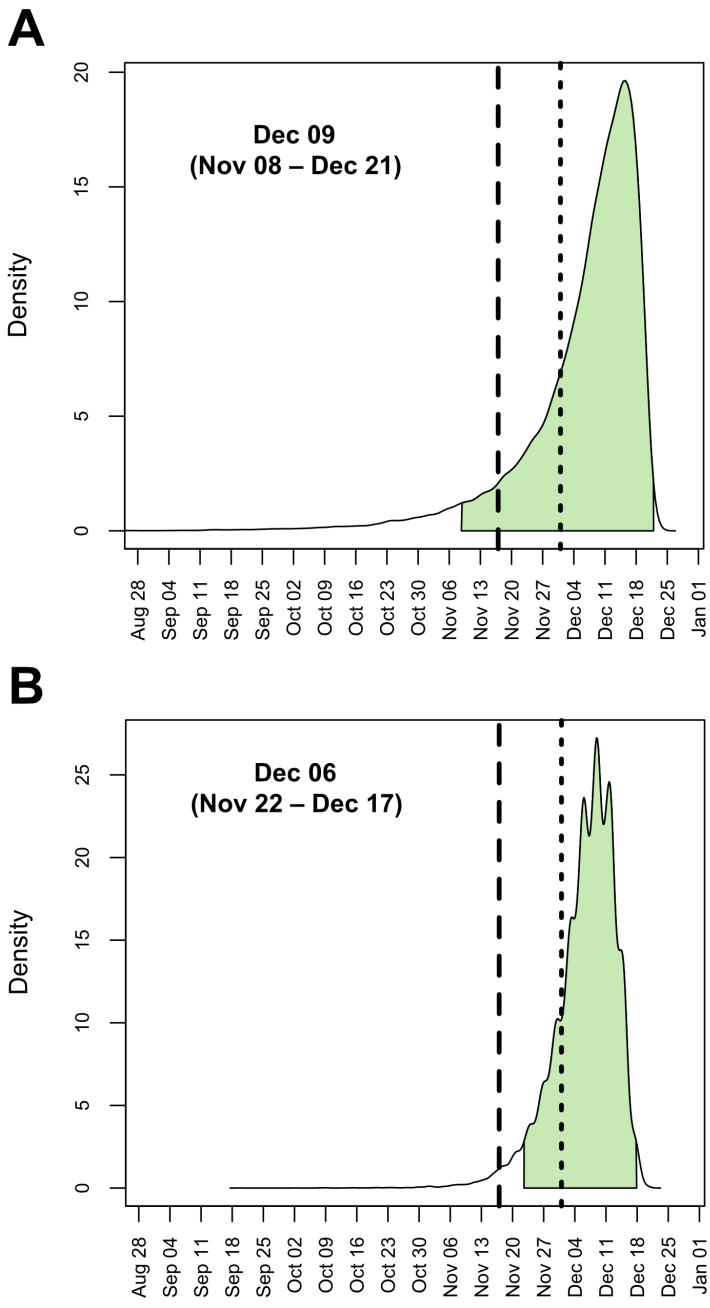


Figure A.1. Posterior distribution for the tMRCA of 583 sampled SARS-CoV-2 genomes circulating in China between December 2019 and April 2020 using (A) a relaxed molecular clock (ULD) and Skyline coalescent prior and (B) a strict clock model and Skygrid coalescent prior. Shaded area denotes 95% HPD. Long-dashed line is 17 November 2019, and short-dashed line is 01 December 2019.

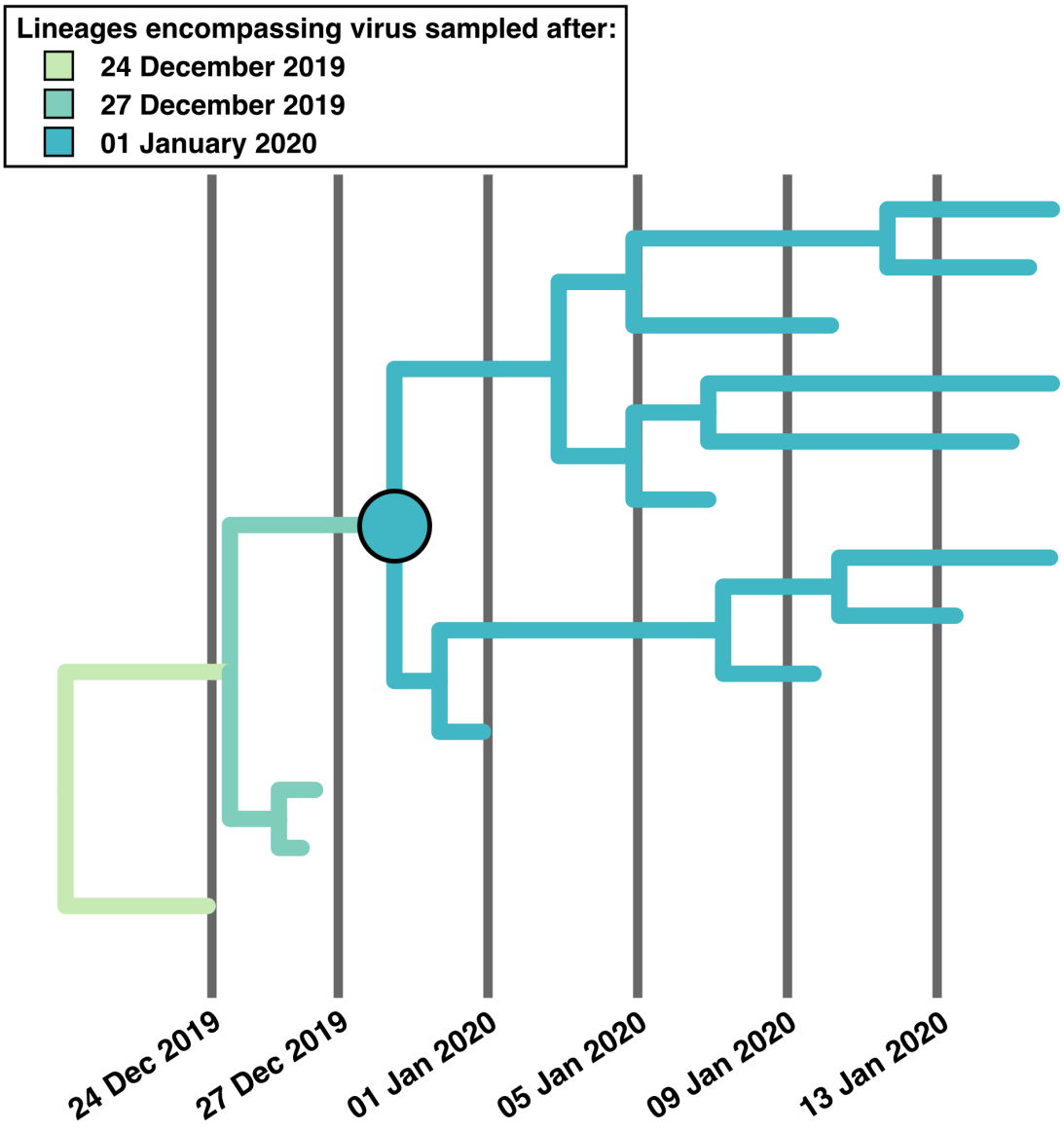


Figure A.2. Example of how coalescence of all sampled genomes can shift forward when analyzing viral genomes sampled in late-December 2019 and January 2020 as basal lineages are lost and cease to propagate. In this example, viruses sampled after 01 January 2020 have a stable tMRCA, indicated by the large circle.

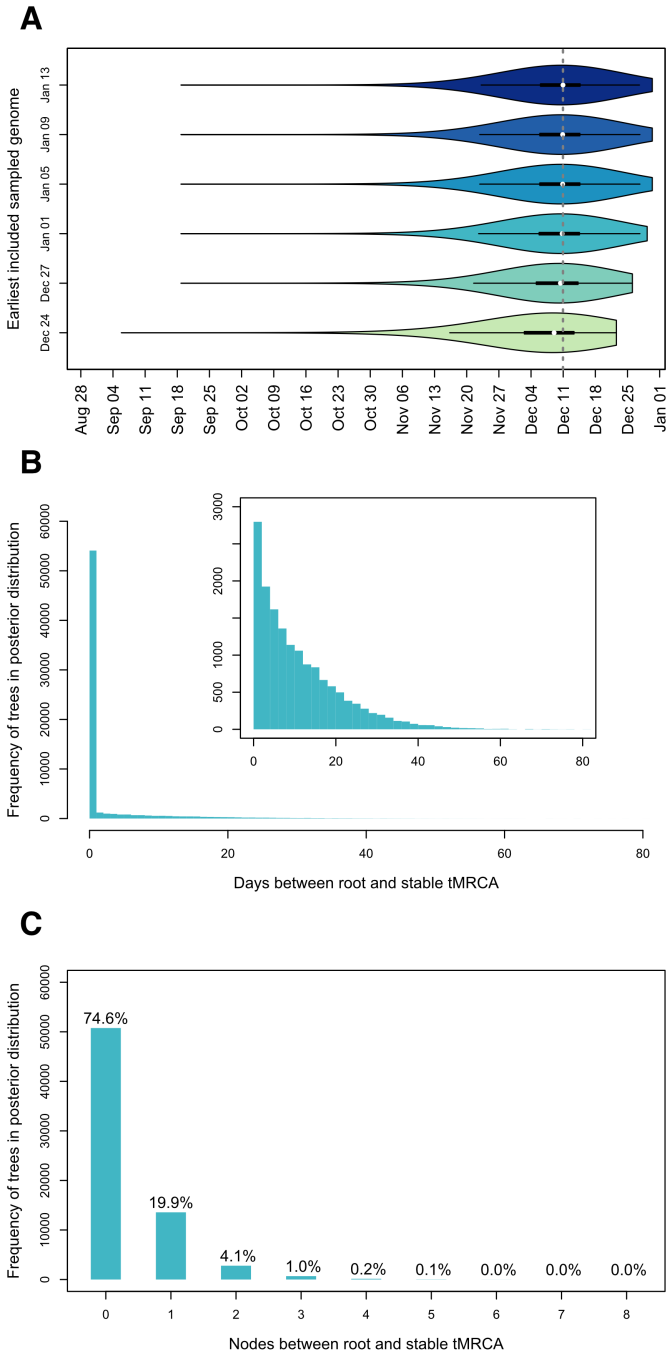


Figure A.3. Shifting of tMRCA in empirical molecular clock analyses. (A) Violin plots of tMRCA estimates upon excluding the earliest sampled genomes between 24 December 2019 and 13 January 2020 (darker colors exclude progressively more early sampled genomes). Mean tMRCAs are depicted with white dots. The dashed grey line represents the mean estimate when including only genotypes sampled on 13 January 2020 or later. (B) The number of days between the tMRCA of all genomes (i.e., root) and the tMRCA of genomes sampled on or after 01 January 2020. Inset excludes BEAST samples with <1 day shift. (C) Number of nodes between the tMRCA of all genomes (i.e., root) and the tMRCA of genomes sampled on or after 01 January 2020. Percentage denotes the number of BEAST samples represented in column.

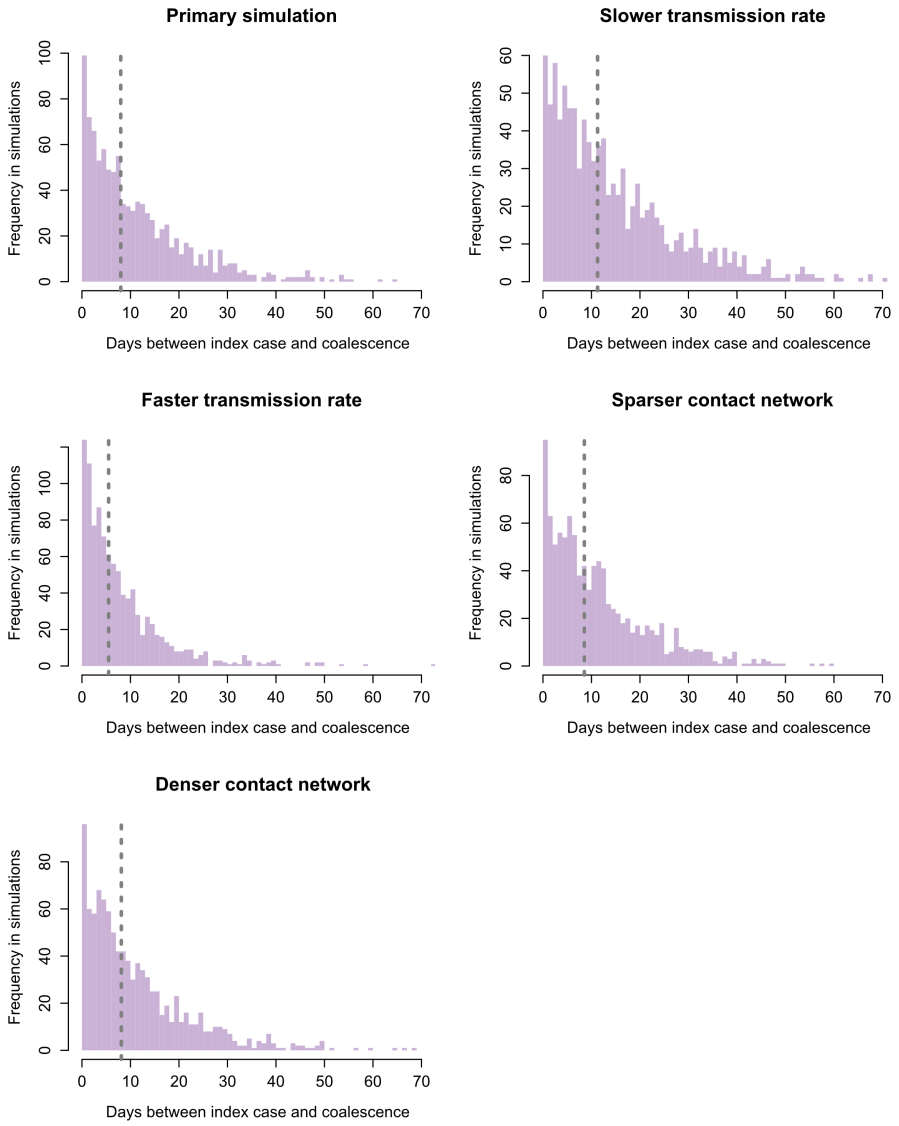


Figure A.4. FAVITES robustness analyses. Number of days between the index case the coalescence when simulated transmission rate is slower and faster and when the contact network has fewer (mean=10) and more (mean=26) contacts.

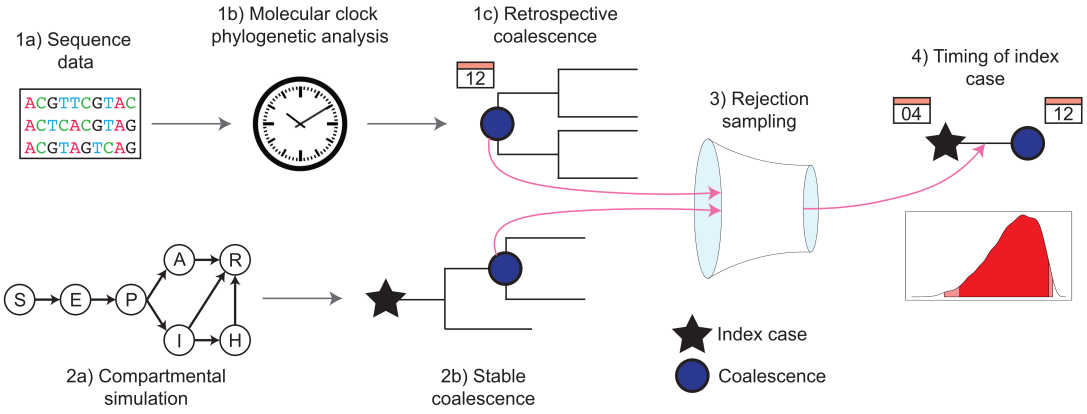


Figure A.5. Combined simulation and phylogenetic workflows to estimate the timing of the Hubei index case in a one-phase epidemic. (1a) Using sequence and epidemiological data, (1b) BEAST performs a phylodynamic molecular clock analysis to (1c) determine the tMRCA. (2a) FAVITES simulates the epidemic in Hubei using a SAPHIRE compartmental model (22) and (2b) estimates a prior distribution for the time from index case to the stable coalescence. The results of (1) and (2) are combined via rejection sampling (3; Fig. A.6) to (4) determine the timing of the index case and its posterior distribution.

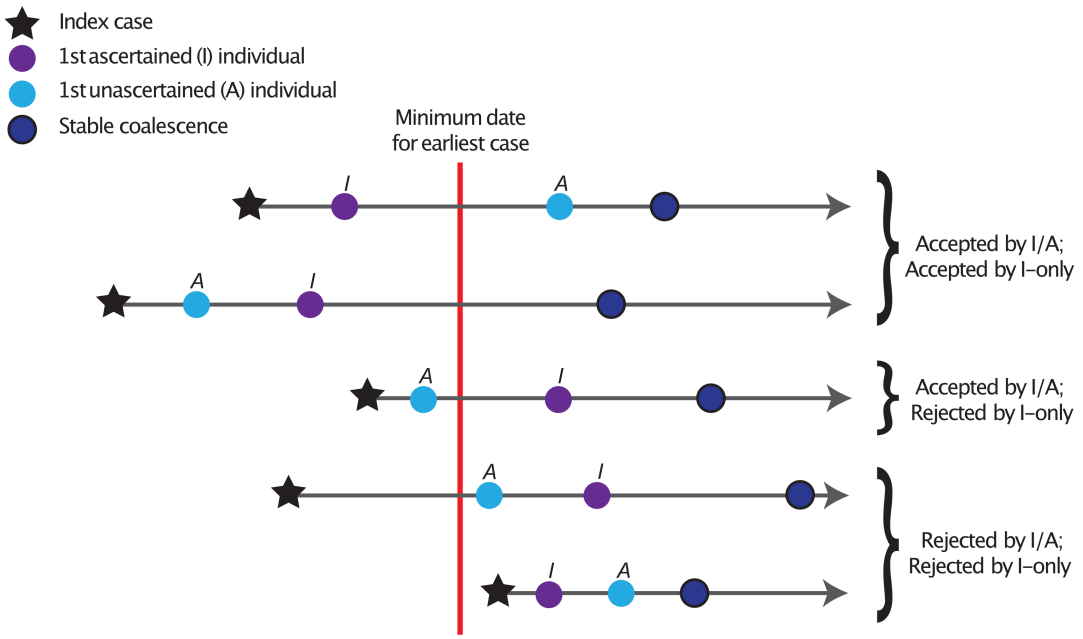


Figure A.6. Illustration of the interplay rejection sampling scheme and the SAPHIRE compartmental model. The top two examples would be accepted by both I/A and I-only rejection sampling, because the first ascertained (I) case came before the minimum date for the earliest case In contrast, the middle example would only be accepted by I/A sampling, because the first unascertained (A) case came before the minimum date for the earliest case, but the first ascertained (I) case did not. The bottom two examples would be always be rejected, because both the first ascertained and unascertained cases came after the minimum date for the earliest case Note that coalescence can happen any time after the index case is first infected, and there is no requirement for coalescence to occur after the first ascertained and unascertained individuals. The index case begins in the exposed (E) compartment and can be the individual that first transitions into the A or I compartments; though other subsequently infected individuals can progress to the A or I compartments before index case. This methodology applies to both one-phase and two-phase epidemic simulations, where ascertained and unascertained cases can come from either Phase 1 or Phase 2 in the latter.

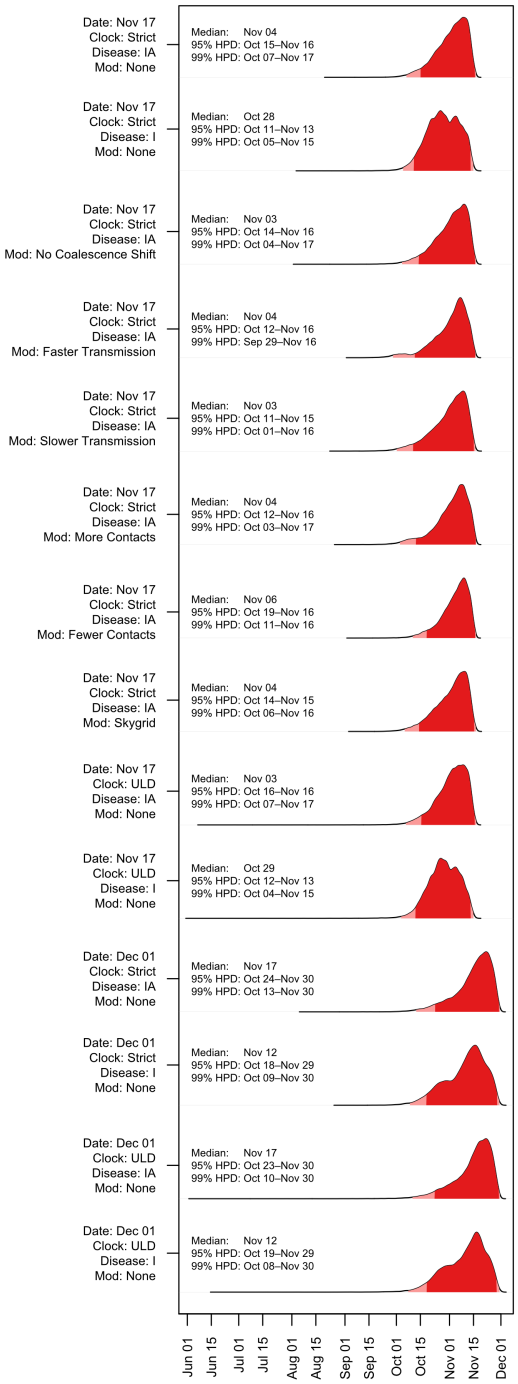


Figure A.7. Robustness analysis for timing of SARS-CoV-2 index case in Hubei province. The 95HPD is shown in dark red, and the 99% HPD is shown in light red. Primary analysis is shown on top. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling. 'Mod' denotes if any robustness modifications were explored.

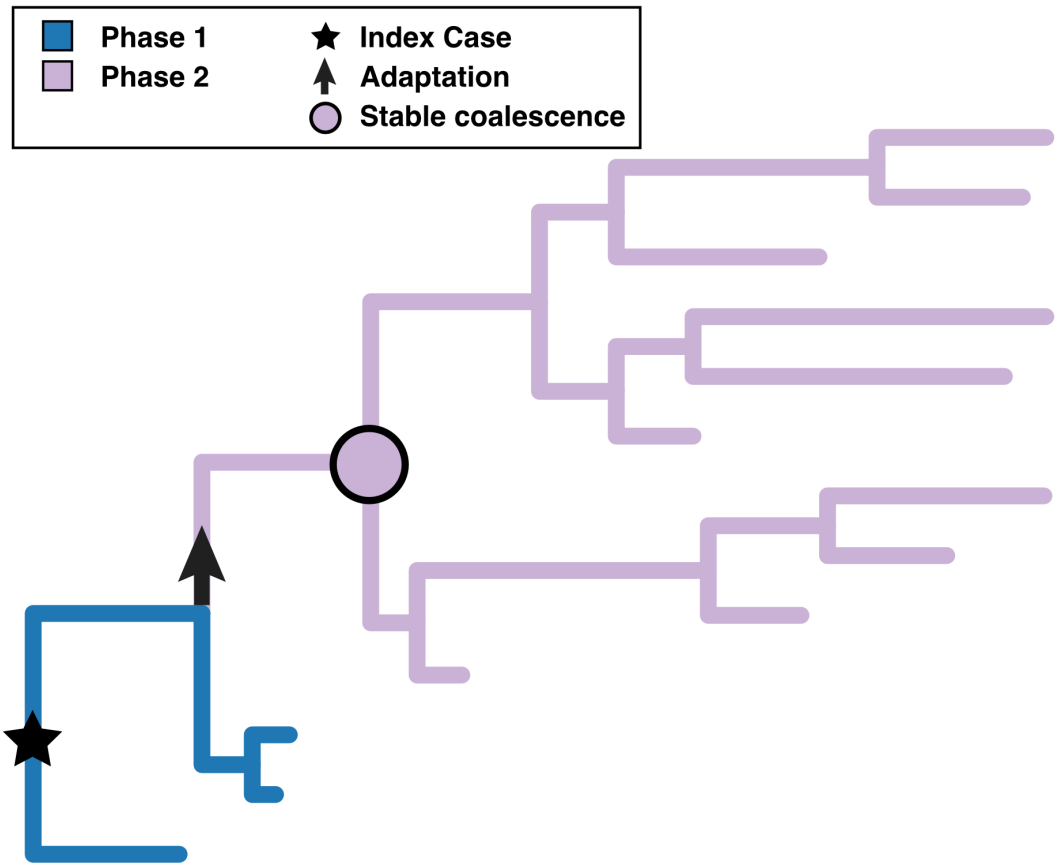


Figure A.8. Example phylogeny of the two-phase epidemic simulations. The index case occurs during Phase 1, characterized by a less-fit viral variant (shown in blue). At some point before Phase 1 goes extinct, an adapted variant arises (arrow), giving rise to Phase 2, characterized by the variant observed in Hubei province (shown in light purple). The stable coalescence (circle) occurs during Phase 2.

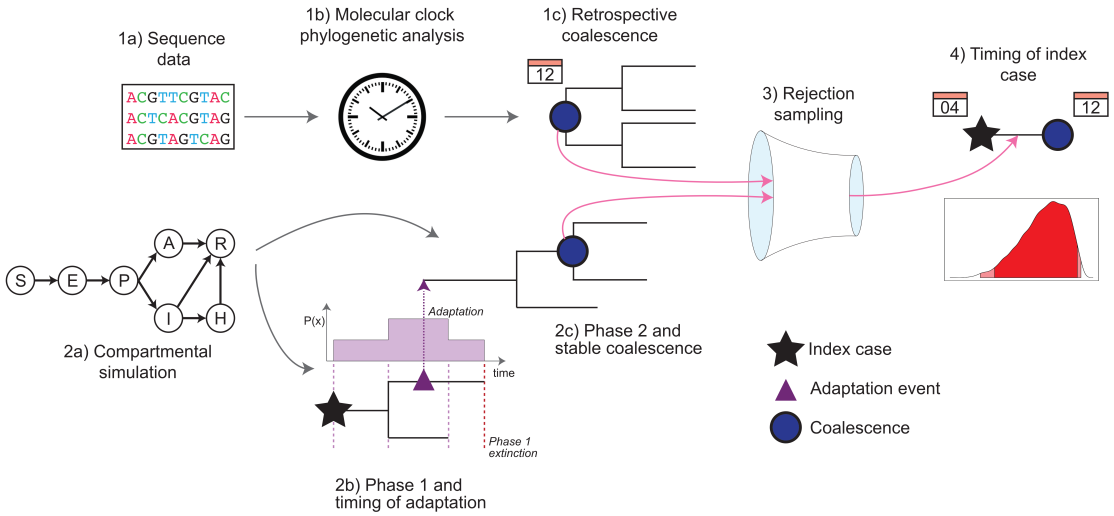


Figure A.9. Combined simulation and phylogenetic workflows to estimate the timing of the Hubei index case in a two-phase epidemic. (1a) Using sequence and epidemiological data, (1b) BEAST performs a phylodynamic molecular clock analysis to (1c) determine the tMRCA. (2a) FAVITES simulates the epidemic in Hubei using a SAPHIRE compartmental model, (2b) estimates a prior distribution for the length of Phase 1 of the epidemic, and (2c) estimates a prior distribution for the time from adaptation to the stable coalescence for Phase 2 of the epidemic. The timing of adaptation (2b) is calculated based on the number of circulating lineages at a given time in Phase 1 (see Methods). The results of (1) and (2) are combined via rejection sampling (3; Fig. A.6) to (4) determine the timing of the index case and its posterior distribution.

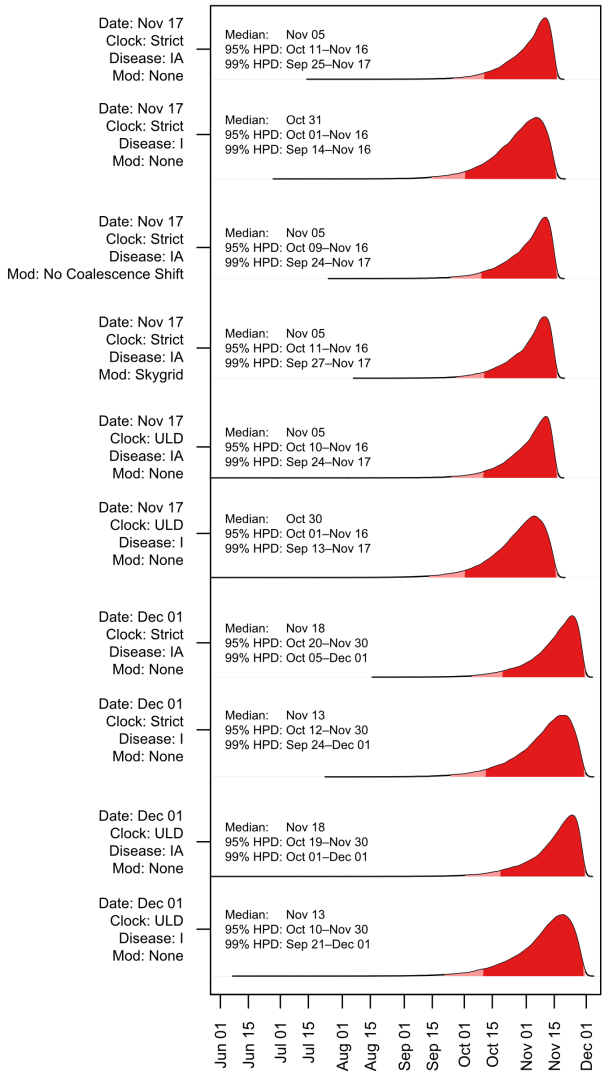


Figure A.10. Robustness analysis for timing of SARS-CoV-2 index case in Hubei province in the two-phase epidemic simulations. Primary two-phase analysis (from Fig. 1.3H) is shown on top. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling. 'Mod' denotes if any robustness modifications were explored. The 95% HPD is shown in dark red, and the 99% HPD is shown in light red.

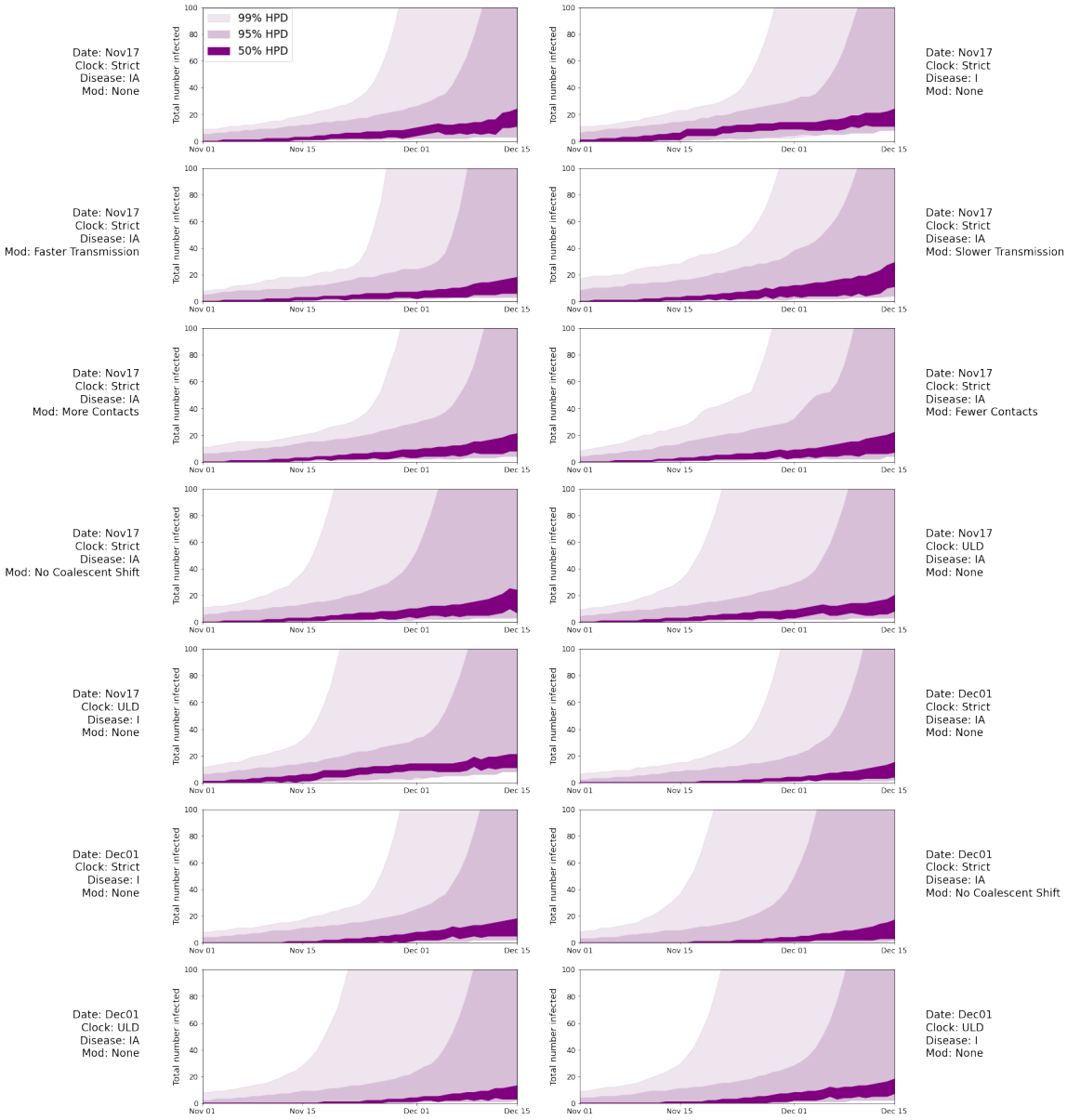


Figure A.11. Robustness analysis for the number of people infected SARS-CoV-2 based on the SAPHIRE model in late 2019. Innermost shading is 50% HPD, middle shading is 95% HPD, and outer shading is 99% HPD. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling.

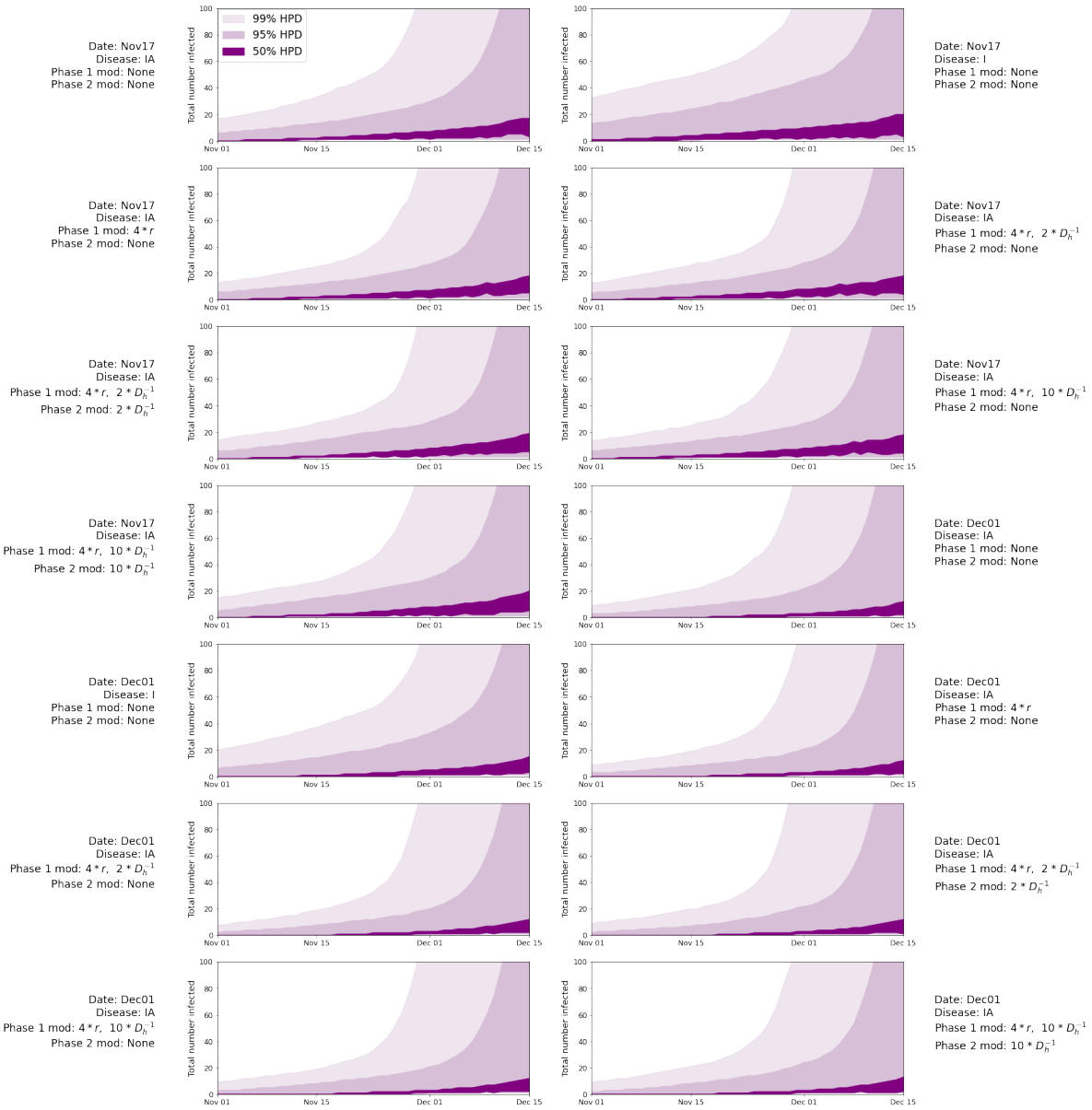


Figure A.12. Robustness analysis for the number of people infected SARS-CoV-2 based on the SAPHIRE model in late 2019 using two-phase simulations. Innermost shading is 50% HPD, middle shading is 95% HPD, and outer shading is 99% HPD. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling. D_h is the average time to hospitalization and r is the ascertainment rate, with the same values as in the primary analysis (Table A.1).

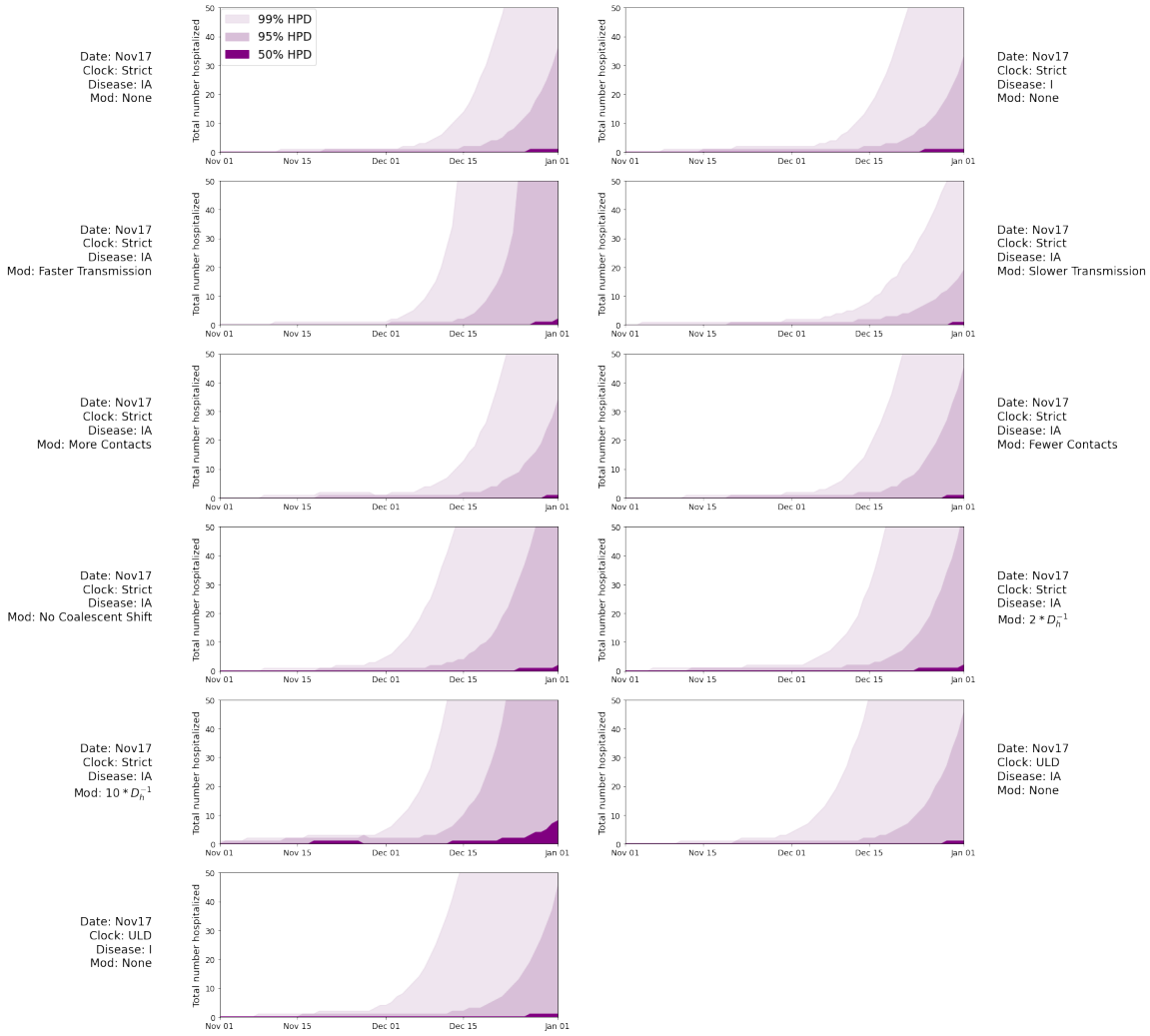


Figure A.13. Number of people hospitalized with SARS-CoV-2 based on the SAPHIRE model in late 2019. Innermost shading is 50% HPD, middle shading is 95% HPD, and outer shading is 99% HPD. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling.

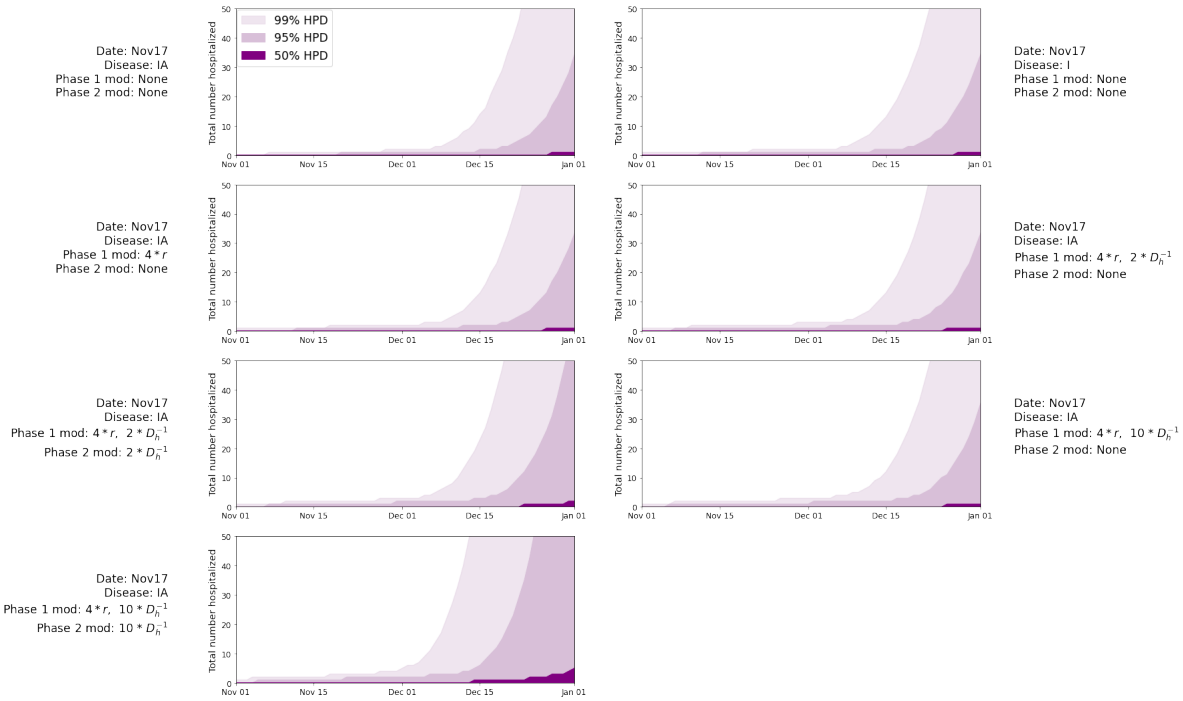


Figure A.14. Number of people hospitalized with SARS-CoV-2 based on the SAPHIRE model in late 2019 using two-phase simulations. Innermost shading is 50% HPD, middle shading is 95% HPD, and outer shading is 99% HPD. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling. D_h is the average time to hospitalization and r is the ascertainment rate, with the same values as in the primary analysis (Table A.1).

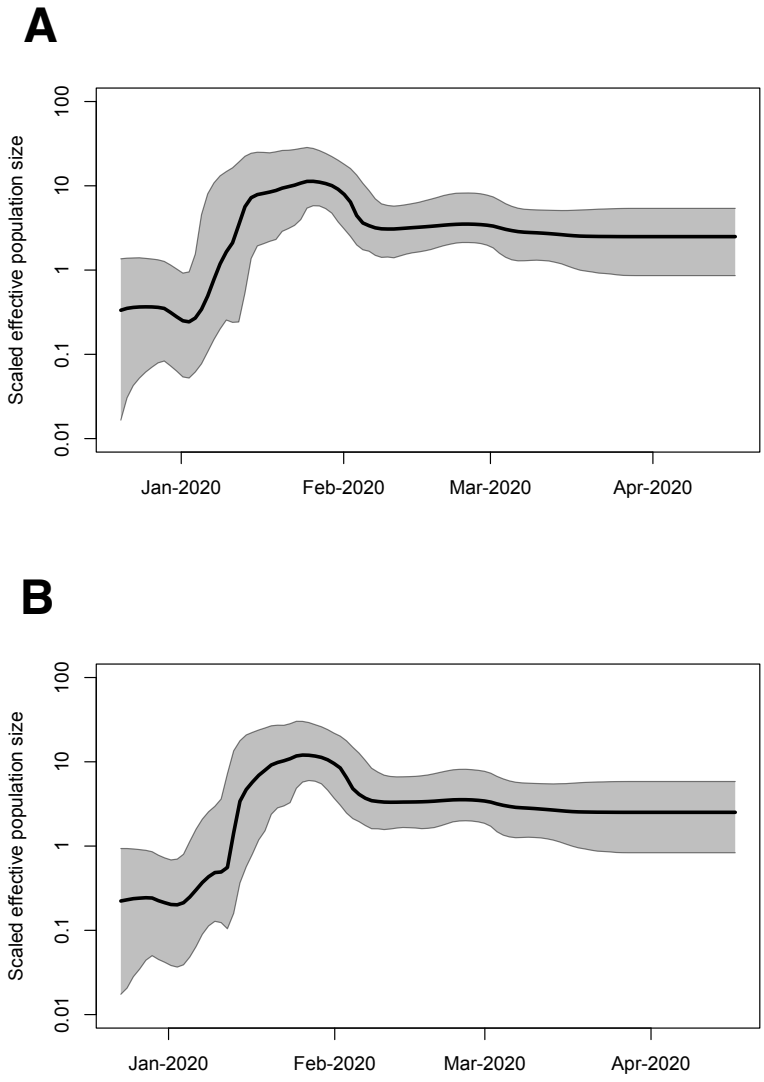


Figure A.15. Bayesian skyline plot reconstruction. (A) Strict molecular clock analysis and (B) ULD relaxed molecular clock. Solid line is the median estimate. Shaded area is the 95% HPD.

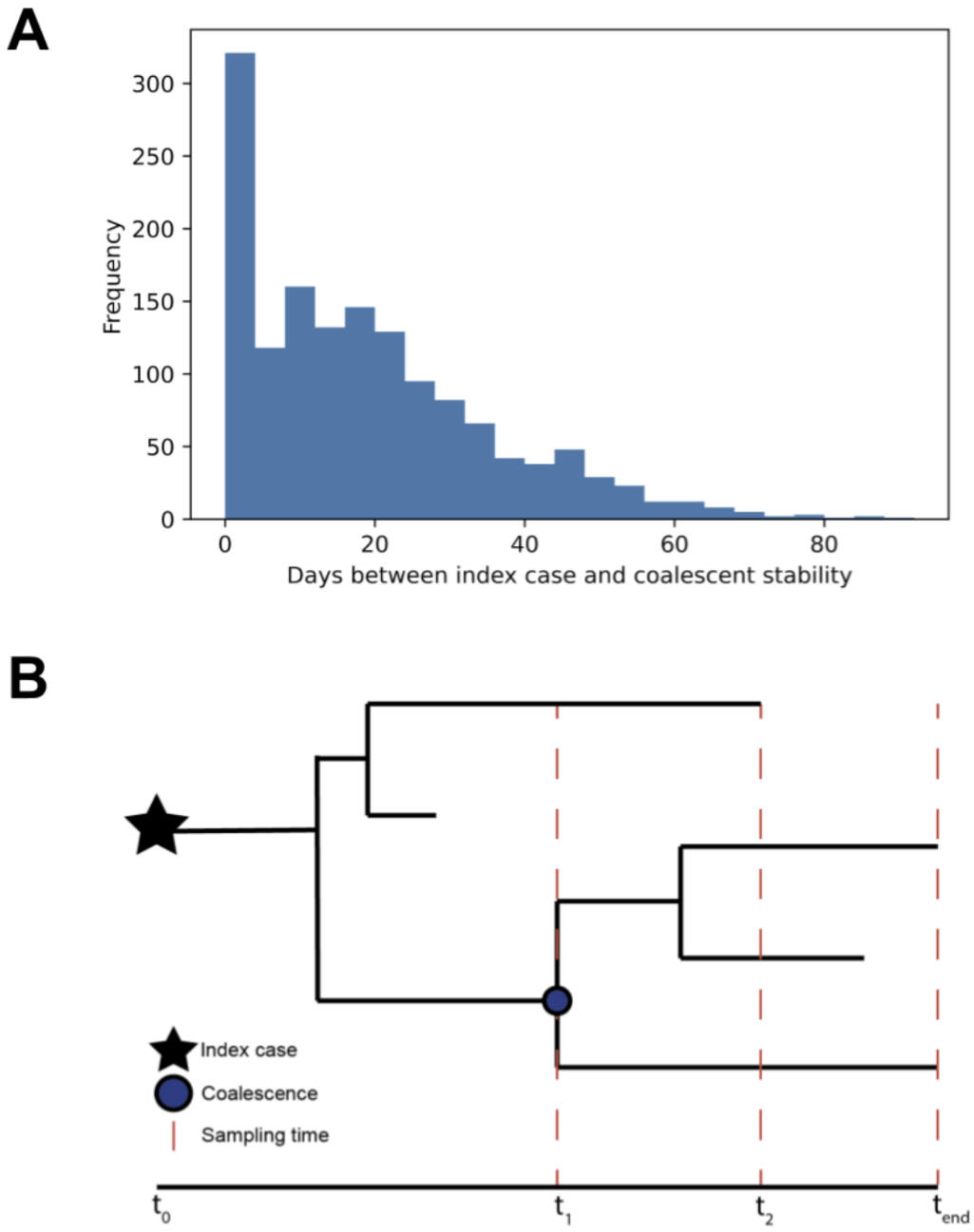


Figure A.16. Stable coalescence in forward compartmental simulations. (A) Days between index case infection and time stable coalescence is achieved in primary simulations. (B) Distinction between the date of index case infection (at t_0), the tMRCA of surviving lineages (at t_1), and the time at which stable coalescence is achieved when the last basal lineage goes extinct (at t_2).

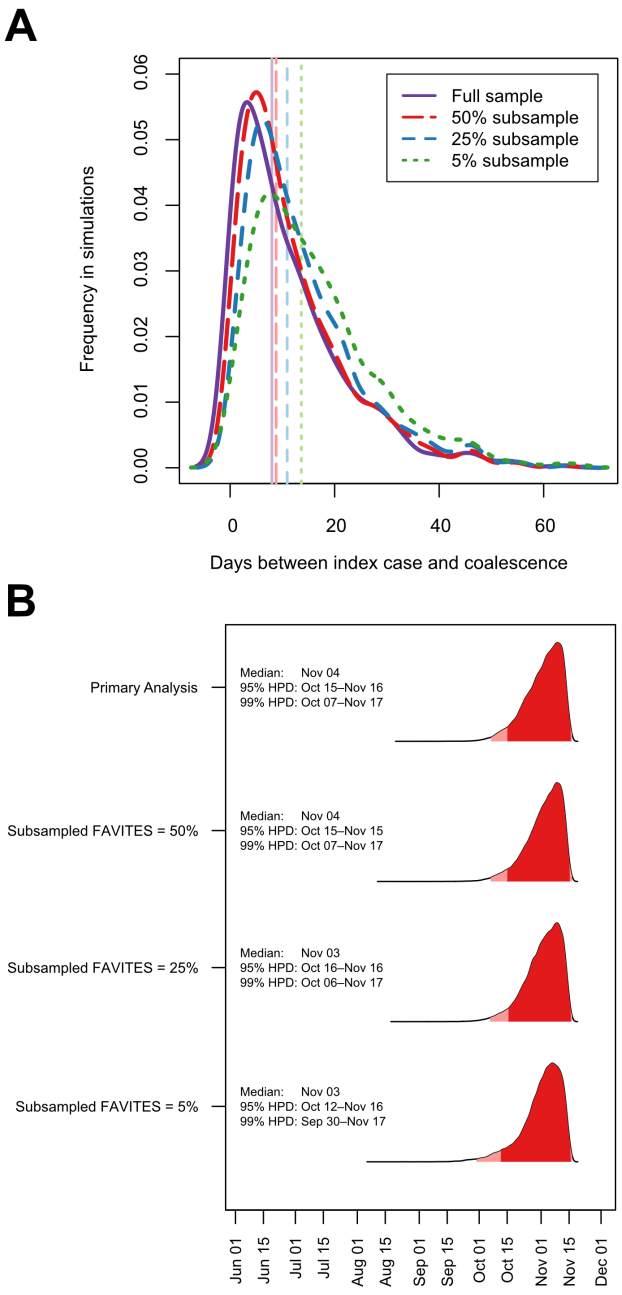


Figure A.17. Impact of subsampling taxa from FAVITES epidemic simulations. (A) Days between index case and stable coalescent with varying amounts of subsampling. (B) Timing of the index case in Hubei province upon subsampling.The primary analysis (from Fig. 1.3C) is shown on top. Inference is conditioned on an ascertained or unascertained case (I or A in SAPHIRE model) by 17 November using a strict clock. The 95% HPD is shown in dark red, and the 99% HPD is shown in light red.

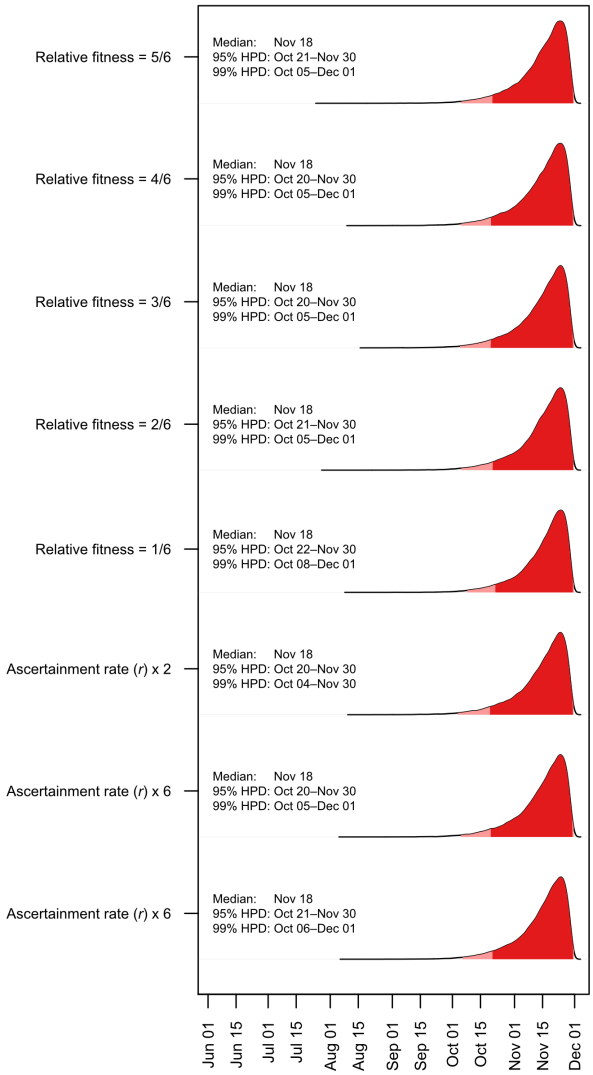


Figure A.18. Impact of relative fitness and virulence of Phase 1 variant on the timing of the index case in Hubei province during two-phase epidemic simulations. The primary two-phase analysis described in the main text (Fig. 1.3F-H) used a relative fitness of 3/6 (or 0.5). Virulence is modulated by the ascertainment rate (r in the SAPHIRE model). Inference is conditioned on an ascertained or unascertained case (I or A in SAPHIRE model) by 17 November using a strict clock. The 95% HPD is shown in dark red, and the 99% HPD is shown in light red.

Appendix B

Supplemental Material for Chapter 2

B.1 Materials and Methods

B.1.1 Sequence data

We queried the GISAID database SARS-CoV-2 viral genome alignment for sequences collected by 14 February 2020[125]. We selected this date to have a data set whose size is appropriate for Bayesian phylodynamic analyses (i.e., under 1000 genomes). We restricted our data set to sequences that (i) were ≥29,000 nucleotides, (ii) had high coverage with ≤0.5% unique amino acid mutations, (iii) had fewer than 1% 'N's, (iv) were not identified as potentially problematic via NextStrain[205], and (v) had a year-month-day sampling date reported. We additionally queried for the National Genomics Data Center, part of the China National Center for Bioinformatics (CNCB), for additional genomes collected by 14 February 2020 not represented on GISAID. Genomes described in the WHO report[26]to have erroneous mutations were updated, and if the virus was sequenced multiple times, the corrections belonging to the genome with the highest coverage were used. The first 88 and last 195 nucleotides of each genome were masked due to poor evidence of homology, and an additional 105 sites were masked based on the work by De Maio *et al.*[127], leading to a total of 388 masked sites. Genomes with an ambiguous nucleotide (e.g., Y or N) at site 8782 or 28144 were excluded. We excluded an additional 20 genomes from the primary phylodynamic analyses, because these sequences contained either C8782T or T28144C, but not both, for reasons described in the main text and De Maio *et al.* (2020)[81]. The final dataset from the early COVID-19 pandemic comprised 787 taxa. A list of GISAID accessions are available in Data S1, and a list of CNCB and GenBank accessions are available in S2. All GISAID accessions

are also available through GISAID when using the identifier *EPI_SET_20220305ud*.

B.1.2 Examining shared mutations between early C/C and T/T genomes and lineages A and B

The 787 SARS-CoV-2 genome dataset and the 20 genomes with either C8782T or T28144C were aligned with MAFFT v7.453[126](options –auto –keeplength –addfragments) to reference genome Hu-1 (GenBank accession MN908947.3; GISAID accession EPI_ISL_402125). We then identified pairs of genomes comprising an intermediate genome (C/C or T/T) and a major lineage (lineage A or B) that shared derived mutations.

We additionally examined 83 genomes from the Diamond Princess outbreak, aligned to reference genome Hu-1 with MAFFT. We inferred a maximum likelihood tree using IQ-TREE 2 v2.0.7[128]under a GTR+F+I substitution model, which was visualized using FigTree[206].

B.1.3 Sequencing quality of early C/C and T/T genomes

We aligned reads FASTQ files belonging to EPI_ISL_413017, a C/C genome from South Korean, and EPI_ISL_462306, a T/T genome from Singapore, using Minimap2[207], sorted the subsequent SAM files using samtools[208], and called variants using iVar[209]. The variant calls were then manually inspected for depth and indeterminacy at specific sites, including 8782 and 28144. The raw data for EPI_ISL_413017 and EPI_ISL_462306 are available at https://www.ncbi.nlm.nih.gov/sra, with project IDs PRJNA806767 and PRJNA802993, respectively.

B.1.4 C/C and T/T genomes in San Diego

The SEARCH consortium has sequenced over 35,000 genomes from San Diego during the course of the pandemic, by 3 February 2022. We generated a multiple sequence alignment of these genomes using Minimap2 and gofasta[210]. We queried this alignment for consensus genomes with both C8782T and T28144, or both C8782 and T28144C, and validated these mutations by checking the read depth and allele frequency in the original alignment files.

The consensus genomes and associated BAM files are publicly available at https://github.com/andersen-lab/HCoV-19-Genomics, and the genome accessions are listed in Data S2. The tree figures were rendered using baltic[211].

B.1.5 T/T genomes in NYC

Molecular surveillance conducted by the New York City Public Health Laboratory, part of the Department of Health and Mental Hygiene, has sequenced >5000 SARS-CoV-2 genomes through the end of 2021. We queried these data for genomes with both C8782T and T28144 in the consensus sequence and validated the consensus sequence by checking the read depth and allele frequency from the primer removed BAM files. The genome accessions are listed in Data S1. The tree figure was rendered using baltic.

B.1.6 Constructing the recombinant common ancestor

We used the aligned sarbecovirus genomes (5' non-coding and poly-A tail excluded; includes the entirety of the SARS-CoV-2 genome; genome accession IDs available in Data S1 and S2) and 14 breakpoints inferred from GARD[212] and provided by Sarah Temmam and Marc Eloit[84] to infer the phylogenetic history of 15 non-recombinant regions. We inferred a maximum likelihood tree of the animal viruses in the alignment and Hu-1 for each non-recombinant region using IQ-TREE 2 under a GTR+F+G+I substitution model. We midpoint-rooted each maximum likelihood tree and used TreeTime v0.8.1[129] to perform ancestral sequence reconstruction for each fragment. The genome belonging to the parent node of Hu-1 for each non-recombinant region was inferred, and these inferred regions were concatenated to construct the recombinant common ancestor (recCA) of SARS-CoV-2 in the sarbecovirus clade. The mid-point root of each tree is always separated from the parent node of SARS-CoV-2 by at least several internal nodes, indicating that the recCA inference would not be sensitive to rooting.

There are 382 substitutions between Hu-1 and the recCA, and one of these sites is masked in the phylodynamic analyses. Ignoring the 387 masked sites of the remaining 29,521 sites, there are 29,134 sites identical between Hu-1 and the recCA. When we used a different SARS-CoV-2 genome (e.g., WH04, WA1) to construct the recCA, the recCA sequence was consistent, indicating the recCA can reliably be used as an ancestor of SARS-CoV-2.

We created a simplot of the sarbecovirus genomes and the recCA against Hu-1 using RDP4[213], and our phylogenies were visualized using FigTree. The genome accessions are listed in Data S1 for genomes from GISAD and Data S2 for genomes from GenBank.

B.1.7 recCA robustness analysis

To examine how sensitive the recCA and downstream analyses were to the recombination inference method, we constructed a second recCA. Using the alignment provided by Sarah Temmam and Marc Eloit, we re-inferred breakpoints with 3SEQ[214] with an approach based on Boni *et al.*[215]. Briefly, after running 3SEQ on the 31-genome alignment, we (1) collected all inferred breakpoints into a single set, (2) complemented this set to generate the set of non-breakpoints, (3) grouped non-breakpoints into continuous breakpoint free regions, (4) reran 3SEQ on the all BFRs >3,000 nucleotides to examine them for mosaicism, (5) pooled the breakpoints from the first and second 3SEQ runs, and (6) collapsed all breakpoints within 100 nucleotides of each other into the most upstream (5') breakpoint. This resulted in 21 breakpoints and 22 BFRs. We next constructed a recCA using the same methods as above.

B.1.8 Reversions early in the pandemic

Here, we define mutations away from the Hu-1 reference genome toward the recCA, such as C8782T and T28144C, as reversions. The 787 genomes sampled by 14 February 2020 were aligned with MAFFT to Hu-1 (GISAID accession EPI_ISL_402125). The phylogenetic history of SARS-COV-2 in China was first inferred in a maximum likelihood framework in IQ-TREE 2 using a GTR+F+I model. We used TreeTime to perform ancestral state reconstruction on the maximum likelihood tree of 787 genomes, rooted on Hu-1. We determined which branches had reversions. Each unique reversion and non-reversion substitution was only counted once to account for phylogenetic uncertainty. The tree figure was rendered using baltic.

B.1.9 Reversions throughout the pandemic

We extended the reversion analysis from above to the following variants: Alpha (PANGO lineage B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), Epsilon 1 (B.1.427), Epsilon 2 (B.1.429), Zeta (P.2), Iota (B.1.526), and Kappa (B.1.617.1). To match the 52-day window for sequence collection from the early pandemic, we found the earliest 52-day window in GISAID for each variant that contained at least 750 sequences. The genomes for each lineage were then aligned using MAFFT to reference genome Hu-1. We next used IQ-TREE 2 and TreeTime as

above to generate a phylogeny and perform ancestral state reconstruction. We followed the same protocol as above with the recCA to determine which branches had reversions. The tree figures were rendered using baltic.

To examine reversions across a subsample of the entire pandemic, we extracted the global tree and associated public genomes from Nextstrain[205] on 14 January 2022 and constrained it to sequences before November 2021. We aligned the genomes using MAFFT as above, used TreeTime with the NextStrain tree and alignment to perform ancestral state reconstruction, and determined reversions with the recCA. The genomes used for the clade and global pandemic reversion analysis can be found in Data S1 and S2, respectively.

B.1.10 Phylogenetic inference

Molecular clock phylodynamic inference was conducted using a Bayesian approach in BEAST v1.10.5[130]. For the primary analysis, we developed and employed a non-reversible, random-effects substitution model (described below), a strict molecular clock, a non-parametric skygrid prior with 20 grid points and a cut off of 0.37, which translates to 5 October 2019. To facilitate Markov chain Monte Carlo (MCMC) chain convergence, (i) we used our previous results of 7.9×10^{-4} and 6.8×10^{-5} substitutions/site/year as the mean and standard deviation, respectively, of a normal prior for the clock rate, and (ii) we initiated the MCMC sampling using the maximum likelihood phylogeny that had been transformed into a chronogram via TempEst v1.5.3[216]. We ran four independent chains of 400 million generations, sub-sampling every 25 thousand iterations to continuous parameter log files, 100 thousand iterations for the tree file, and 100 thousand iterations for the ancestral state reconstruction of the most recent common ancestor (MRCA); the first 15% of the chains were discarded as burnin. Convergence and mixing was assessed in Tracer v1.7.1[217]and all 4 chains were combined in LogCombiner, such that all relevant effective sample size (ESS) values were >200. The accession IDs can be found in Data S1 and S2, and the XML files can be found at[133]. The outputs can be found at doi:10.6076/D1Q59K[134].

B.1.11 Random-effects substitution model

To accommodate the mutational bias in SARS-CoV-2 for C-to-T transitions[86–88], we developed a random-effects substitution model. This model employs a standard phylogenetic

substitution model as a base model, incorporated as fixed effects, while the random effects allow each individual mutation rate (e.g., C-to-T, separate from T-to-C) to be elevated (or decreased) relative to that model. Note that this makes the model non-reversible. We use HKY as the base model, which is defined by the parameters κ (governing the relative rate of transitions versus transversions) and $\boldsymbol{\pi}$ (governing the root and stationary frequencies). Working on the log-scale and denoting the random effect ϵ_{ij}, our HKY+RE model gives the log of the substitution rate matrix entries as,

$$\log(q_{ij}) = \log(\pi_j) + \log(\kappa) + \epsilon_{ij} + \log(\xi) \quad \text{if } i > j \text{ is a transition}$$

or,

$$\log(q_{ij}) = \log(\pi_j) + \epsilon_{ij} + \log(\xi) \quad \text{if } i > j \text{ is a transversion,}$$

where ξ is a normalizing constant.

To accommodate among-site rate variation, we employ a proportion of invariable sites in the model, with a Uniform(0,1) prior. We place independent and identical Normal(mean=0,SD=σ_ϵ) priors on ϵ_{ij}, an improper infinite uniform prior on σ_ϵ, and a Normal(mean=0,SD=1.25) prior on $\log(\kappa)$. We fix $\boldsymbol{\pi}$ to the frequencies observed in the alignment.

B.1.12 Quantifying the support for ancestral haplotypes

We assume equally likely prior probabilities for each sequence in the ancestral state reconstruction posterior. Therefore, the Bayes factor (BF) in favor of sequence S_1 against another sequence S_0, given the data D, can be expressed as follows:

$$B_{10} = \frac{P(S_1|D)}{P(S_0|D)}$$

where $P(S|D)$ is the posterior probability of a sequence. Note that all BFs were calculated with the sequence comprising the highest posterior probability as S_1, and BFs for each phylodynamic model were calculated separately.

B.1.13 Phylogenetic inference with constrained roots and recCA

In a standard phylogenetic model, the sequences at all internal nodes (and the root) are integrated out. A prior distribution is required for the root sequence in order to compute the likelihood or sample ancestral states at nodes. Typically, the prior distribution for the root sequence assumes every site is drawn identically and independently from some multinomial distribution defining the prior probability of an A, C, G, or T nucleotide at any (and every) site. Here, we consider two novel approaches.

The first approach constrains the character state of the MRCA of all human SARS-CoV-2 sequences to be identical to a specific hypothesized ancestral haplotype (Fig B.18D). We consider 6 ancestral haplotypes that match the sequences belonging to Wuhan Hu-1, the C/C haplotype (Hu-1 with T28144C), the T/T haplotype (Hu-1 with C8782T), WH04 (lineage A, or Hu-1 with C8782T and T28144C), 20SF2012 (WH04 with C29095T), and WA1 (lineage A.1, or WH04 with C18060T). Note that in this model, the root of the tree is the MRCA of all 787 human SARS-CoV-2 sequences, and all 787 genomes, including the genomes that are identical to the ancestral haplotype, remain taxa—the tree is not rooted on them.

The second approach places a per-site prior distribution on the ancestral haplotype. Specifically, we add a branch ancestral to the MRCA and fix the sequence at the root to be the sequence of the recCA (Fig B.18C). This approach places a prior distribution on the ancestral haplotype, as the root and MRCA are distinct in this model. At each site, this prior is determined by the nucleotide present in the recCA, the length of the branch leading to the MRCA, and the substitution model parameters. Our primary analysis uses this approach with the recCA based on the 14 breakpoints from GARD. We performed a robustness analysis by using this approach with the recCA based on the 21 breakpoints from 3SEQ.

B.1.14 BEAST sensitivity analyses

We performed a series of sensitivity analyses for the Bayesian phylogenetic inference: (i) unconstrained rooting with a GTR+F+I model; (ii) unconstrained rooting with unmasked sequences; (iii) unconstrained rooting excluding 15 market-associated genomes; (iv) unconstrained rooting excluding all genomes sampled from Wuhan; (v) unconstrained rooting including the

20 “intermediate” genomes sampled before 14 February 2020; (vi) unconstrained rooting but including the 16 C/C genomes sampled before 14 February 2020; (vii) unconstrained rooting but including the 4 T/T genomes sampled before 14 February 2020; (viii) recCA rooting using the recCA based on the 14 breakpoints from GARD and with a GTR+F+I model; (ix) outgroup rooting with RaTG13 as an outgroup; and (x) outgroup rooting with BANAL-20-52 as an outgroup.

The 7 July 2020 sampling date of BANAL-20-52, which post-dates the last included SARS-CoV-2 genome in the dataset by several months, creating a problematic months-long lineage for the coalescent skygrid prior. To accomodate this issue for sensitivity analysis rooted on BANAL-20-52, we truncate the branch leading to BANAL-20-52 at 14 February 2020. Compared to the overall duration of this branch (roughly 17 years), the effect of the truncation is negligible to the likelihood.

B.1.15 Inferring empirical doubling times

To determine the appropriate doubling time for the epidemic simulations, we inferred the doubling time of the early pandemic. We extracted the daily number of cases from the WHO report[26]through the end of December 2019 and coupled it with the data from Li *et al.*[23]. Because the case data included early cases that were later found to not associated with SARS-CoV-2 infection, we (i) removed the case from 2 December 2019 [as per the WHO report[26]] and (ii) shifted the case from 8 December to 16 December [as per Worobey[22]]. Additionally, we (iii) added a case to 10 December [as per Worobey[22]]. We used EpiNow2 and its default parameters with the daily case count to infer the doubling time of SARS-CoV-2 through 15 January 2020[218].

B.1.16 Epidemic simulation

To explore the evolutionary dynamics during the beginning of the COVID-19 pandemic, we developed FAVITES-COVID-Lite[131], a simplified individual-based simulation pipeline based on FAVITES[93], and performed a series of epidemic simulations. First, we generated static contact networks comprising 5 million individuals (nodes) using NiemaGraphGen[219]under a preferential-attachment model using the Barabási-Albert algorithm[220]. We used this algorithm to generate the contact networks, because its scale-free properties recapitulate infectious disease

spread[220], including the superspreading dynamics of SARS-CoV-2[82;94]. We chose to simulate a static contact network because our focus is on the number of people infected at the beginning of the epidemic, and we used an intermediate value of 16 contacts per day (mean degree), based on Mossong *et al.*[221].

We extended the SAPHIRE [Susceptible (S)-Ascertained (I)-Presymptomatic (P)-Hospitalized (H)-Not Ascertained (A)-Removed (R)-Exposed (E)] model developed by Hao *et al.*[132], and implemented in our previous study on the timing of the primary case[94], to have two ascertained compartments. Individuals from the first ascertained compartment can either enter the recovered compartment or an "ascertained-pre-hospitalization" (I_H) compartment, where they eventually transition to the hospitalized compartment. We extended the model with the I_H compartment to decouple the proportion of people hospitalized with the amount of time until hospitalization. We did not include the travel component of the original SAPHIRE model (i.e., individuals flying into and out of Wuhan), because our focus was on the early dynamics of the pandemic before its spread outside of Wuhan. The dynamics of these compartments across time (t) are described by the following set of ordinary differential equations:

$$\frac{dS}{dt} = -\frac{bS(\alpha P + \alpha A + I + I_h)}{N}$$

$$\frac{dE}{dt} = \frac{bS(\alpha P + \alpha A + I + I_h)}{N} - \frac{E}{D_e}$$

$$\frac{dP}{dt} = \frac{E}{D_e} - \frac{P}{D_p}$$

$$\frac{dA}{dt} = \frac{(1-r)P}{D_p} - \frac{A}{D_i}$$

$$\frac{dI}{dt} = \frac{rP}{D_p} - \frac{I}{D_i}$$

$$\frac{dI_h}{dt} = \frac{hI}{D_i} - \frac{I_h}{D_q - D_i}$$

$$\frac{dH}{dt} = \frac{I_h}{D_q - D_i} - \frac{H}{D_h}$$

$$\frac{dR}{dt} = \frac{A + (1-h)I}{D_i} + \frac{H}{D_h}$$

We note that we rescale the transmission rate (b) reported in B.3 and used for each analysis to the average degree centrality in the network[87].

We performed forward simulations using this extended model to generate a viral transmission network using GEMF[222]. Simulated epidemics started with a single seed infection among our 5 million susceptible individuals, and infected nodes infect other nodes via Poisson processes. The parameters were primarily determined by Hao *et al.*[132](B.3), except we required half the ascertained population to become hospitalized with an average of 11 days between symptom onset and hospitalization, matching reports of hospitalization of the early pandemic in Wuhan[24;223;224]. Each simulation was run for 100 days and produced an output documenting when individuals transitioned from one compartment to another throughout the entire simulation. We used these outputs to determine the number of individuals in a given compartment (e.g., total infections, ascertained infections, unascertained infections, and hospitalized individuals) across each day in the simulation.

Once the forward simulations were complete, we subsampled the first 50,000 infected individuals. If there were fewer than 50,000 infections, we used each infected individual for subsequent analyses. Then, we sample a sequencing time for each ascertained individual from a uniform distribution that starts when they enter the first ascertained compartment and ends when they were recovered. To match real-world data from December 2019, we include only individuals with genome sampling times that occur after the first hospitalization of the simulated epidemic. Additionally, the primary case (i.e., first infected individual in the simulation) is sampled regardless of ascertainment status. We sample the primary case regardless of ascertainment or

hospitalization status to properly determine stable coalescence (described below). Unascertained individuals are not sampled.

Lastly, we provide the genome sampling times and transmission network to CoaTran[219], which uses a coalescent process to generate time trees. We then use a constant substitution rate of 9.2×10^{-4} substitutions/site/year (inferred from our primary BEAST results) to convert the branch lengths from years to substitutions per site to generate mutation trees.

For the primary analysis, we ran successive epidemic simulations until we reached 1100 successful simulations, defined as those simulations in which $\geq$400 people had become infected and $\geq$1 person was still infectious at the end of the simulation. Failed epidemics were those simulations that did not become established (i.e., 0 infectious people at the end of the simulation) or had fewer than 400 people infected over the entire simulation; 3857 (77.8%) simulations failed to reach this epidemic threshold after 100 days. Our epidemics had a median doubling time of 3.47 days (95% range: 1.35-5.44), slightly shorter than the doubling times from Hao *et al.*[222]to match empirical estimates of the growth rate from before 1 January 2020 in Wuhan[95;96]. Twelve of the 4957 simulations (0.2%) did not reach 400 infections but were still persisting at the end of the 100-day simulation. Given the rarity of these 12 simulations, it is highly improbable that SARS-CoV-2 was characterized by their growth dynamics.

All input parameters for the primary and sensitivity analyses can be found in[133], the output files and post-processed results can be found at doi[134].

B.1.17 Doubling time inference

To characterize the growth dynamics of our epidemic simulations and compare them to empirical growth dynamics, we estimated the doubling time of the primary epidemic simulations. For our primary simulations, we calculated both the cumulative doubling time and 14-day doubling time at each day in the simulation (prior to day 14, the 14-day doubling time is identical to the cumulative doubling time). However, because of the stochasticity inherent in these epidemic simulations, the simulations can be more easily compared to one another when the same number of individuals are infected. Therefore, we also estimated the cumulative and 14-day doubling times at certain numbers of infections (e.g., 100 infections).

Based on the doubling time results (Fig. B.22), we reason that at 1,000 infections, there is sustained transmission, the variability in doubling times across simulations has slightly decreased, and the 5 million node network has not yet started to saturate. Therefore, unless otherwise specified, the doubling times we report for each set of epidemic simulations is the 14-day doubling time when 1,000 individuals are infected.

B.1.18 Sensitivity analysis–faster rate of infection

We simulated epidemics for a more rapidly spreading virus using the same parameters from the primary analysis, except the transmission rate was increased from 0.28 to 0.38 per day (1.36x) and the simulation time was decreased from 100 to 70 days (0.70x). We produced 1100 successful simulations with at least 400 infected individuals and a median epidemic doubling time of 2.65 days (95% range: 1.50-4.10), and 2204 (66.7%) simulations failed to reach the epidemic threshold.

B.1.19 Sensitivity analysis–slower rate of infection

We simulated epidemics for a more slowly spreading virus using the same parameters from the primary analysis, except the transmission rate was decreased from 0.28 to 0.22 per day (0.79x) and the simulation time was increased from 100 to 150 days (1.50x). We produced 1100 successful simulations with at least 400 infected individuals and a median epidemic doubling time of 4.45 days (95% range: 1.50-7.44), and 7202 (92.8%) simulations failed to reach the epidemic threshold.

B.1.20 Sensitivity analysis–higher ascertainment rate

We simulated epidemics with a higher ascertainment rate using the same parameters from the primary analysis, except the ascertainment rate was increased from 0.15 to 0.25 (1.67x) and the simulation time was increased from 100 to 120 days (1.20x). Additionally, we decreased the transmission rate from 0.28 to 0.255 (0.91x) to keep the doubling time consistent with the primary analysis. We produced 1100 successful simulations with at least 400 infected individuals and a median epidemic doubling time of 3.52 days (95% range: 1.38-5.64), and 3698 (77.1%) simulations failed to reach the epidemic threshold.

B.1.21 Sensitivity analysis–lower ascertainment rate

We simulated epidemics with a lower ascertainment rate using the same parameters from the primary analysis, except the ascertainment rate was decreased from 0.15 to 0.05 (0.33x) and the simulation time was increased from 100 to 120 days (1.20x). Additionally, we increased the transmission rate from 0.28 to 0.295 (1.05x) to keep the doubling time consistent with the primary analysis. We produced 1100 successful simulations with at least 400 infected individuals and a median epidemic doubling time of 3.50 days (95% range: 1.51-5.65), and 4632 (80.8%) simulations failed to reach the epidemic threshold.

B.1.22 Determining stable coalescence

As in our previous study[94], we define the stable coalescence as the tMRCA that does not shift forward in time by more than one day, even as new individuals become infected and previously infected individuals recover (i.e., the time to the most recent common ancestor [tMRCA]). The stable coalescence is reached the first day that the coalescence for the currently infected individuals is within one day of the time of MRCA after the simulation is complete or once 50,000 total individuals have been infected. Therefore, the stable coalescence ignores the preceding basal lineages that have gone extinct.

We extracted the tMRCA of infected and sampled individuals every day across each simulation using TreeSwift 1.1.14[225]. This tMRCA was calculated for each day of the 100 days or until 50,000 individuals had been infected, whichever came first. We chose not to explore dynamics after 50,000 infections due to a slowing in exponential growth arising from the saturation of the contact network.

B.1.23 Determining the topological patterns of simulated phylogenies

We examined the topology of the phylogenetic trees resulting from the epidemic simulations. For each tree, after determining the point of stable coalescence, we extracted the subtree rooted at the internal node at stable coalescence (i.e., the MRCA of the sampled genomes). We then simulated mutations down the branches of the subtree using a substitution rate of $9x2x10^{-4}$ substitutions/site/year (the inferred substitution rate from the unconstrained Bayesian

phylogenetic analysis). We then counted (i) the number of descendent lineages (i.e., the size of the basal polytomy, including basal taxa, tips with unique mutations, and clades) from the subtree root, (ii) the clade size of each descendant lineage that was one or two mutations from the root, and (iii) the number of lineages descending from each one- or two-mutation derived clade from (ii) (i.e., the size of the polytomy of the large derived clades). Note that (i) is represented in Fig. 2.2A (and the basal polytomy in Fig. 2.2C), and (ii) and (iii) describe the large clades in Fig. 2.2B and 2C. For clarity, we display the subtree rooted at stable coalescence (ignoring the preceding lineages that have gone extinct) in Fig. 2.2.

We use the results of i–iii above to quantify the frequencies of different topologies in the simulated epidemics (τ).

If there were at least n descendant lineages from the subtree root, the simulated phylogeny had a polytomy. We refer to this topology as τ_P.

When there were only two clades, each one mutation from the subtree root, the simulation matched the observed topology if there were a single introduction of the C/C or T/T ancestral haplotype. We refer to this topology as τ_{2C}. We first further constrained τ_{2C} to require each clade to be $\geq$30% and $\leq$70% of the simulated taxa, approximating lineage A and lineage B comprising 35.2% and 64.8% of all taxa, respectively. We constrained τ_{2C} once more to require at least n lineages descending from the root of the clade, to match the polytomies at the bases of the lineage A and lineage B clades.

When there was a polytomy at the base of the simulated phylogeny and a large clade two mutations from the root with a large portion of the taxa, the topology matches the observed topology had there been a single introduction of the lineage A or B ancestral haplotype. We refer to this topology as τ_{1C}. As with τ_{2C}, we constrain τ_{1C} to require the two-mutation clade to comprise $\geq$30% of the taxa. We constrained τ_{1C} once more to require at least n lineages descending from the large two-mutation clade (i.e., a polytomy, as with lineage B if lineage A were the single introduction and vice-versa).

The empirical phylogeny has 787 taxa and two large polytomies: the 108-descendant lineage A polytomy and the 231-descendant lineage B polytomy. Our primary results therefore require the polytomy size n to be 100 lineages, reflecting the empirical data. Importantly, this

is a conservative requirement because 98.4% of the simulated phylogenies had more than 1,000 taxa.

However, since 1.5% of the simulated phylogenies had fewer than 787 taxa, we also performed sensitivity analyses requiring n to be either 20 or 50 lineages. Additionally, since 96.7% of the simulated phylogenies had more than 5,000 taxa, we also performed sensitivity analyses requiring n to be 200 or 500 lineages.

B.1.24 Quantifying the support for two introductions against a single introduction of SARS-CoV-2

Here we synthesize the posterior probabilities of inferred ancestral haplotypes, frequencies of topologies in the forward epidemic simulations, and the expected relationships between these haplotypes and topologies to quantify the support for a two-introduction scenario over a single-introduction scenario. Our goal is to obtain $P(I_n|\mathbf{Y})$, the probability of n introductions (I_n) given our sequence data $\boldsymbol{Y}$, so that we can compare the probabilities of single- and multiple-introduction hypotheses.

We consider three phylogenetic topologies, or root-shapes: τ_P for a large polytomy at the root (stable coalescence), τ_{1C} for one well-defined clade two mutations from the root, and τ_{2C} for two well-defined clades each one mutation from the root, with no other lineages descending from the root. We define S_{MRCA} as the ancestral haplotype of all sampled SARS-CoV-2 genomes, and we specifically consider the four most likely ancestral haplotypes (Table 2.1): S_A for lineage A, S_B for lineage B, $S_{T/T}$ for the intermediate T/T haplotype, and $S_{C/C}$ for the intermediate C/C haplotype.

If there are n introductions, there are n distinct trees with n distinct root-shapes. We will use the vector $\boldsymbol{\tau}$ to track these. In the case that there was one introduction, this is a single root-shape; two introductions have two root-shapes; and we do not consider more than two introductions here.

We can rewrite the ratio of posterior probabilities as a ratio of joint probabilities,

$$\frac{P(I_2|\boldsymbol{Y})}{P(I_1|\boldsymbol{Y})} = \frac{P(I_2, \boldsymbol{Y})}{P(I_1, \boldsymbol{Y})}$$

Marginalizing over S_{MRCA} and assuming that In only affects Y through S_{MRCA},

$$\begin{aligned} P(I_n, \boldsymbol{Y}) &= \sum_{S_{MRCA}} P(I_n, \boldsymbol{Y}, S_{MRCA}) \\ &= \sum_{S_{MRCA}} P(\boldsymbol{Y}|I_n, S_{MRCA}) P(I_n, S_{MRCA}) \\ &= \sum_{S_{MRCA}} P(\boldsymbol{Y}|S_{MRCA}) P(I_n, S_{MRCA}) \end{aligned}$$

Therefore we get,

$$\begin{aligned} \frac{P(I_2, \boldsymbol{Y})}{P(I_1, \boldsymbol{Y})} &= \frac{\sum_{S_{MRCA}} P(\boldsymbol{Y}|S_{MRCA}) P(I_2, S_{MRCA})}{\sum_{S_{MRCA}} P(\boldsymbol{Y}|S_{MRCA}) P(I_1, S_{MRCA})} \\ &= \frac{\sum_{S_{MRCA}} \frac{P(S_{MRCA}|\boldsymbol{Y})P(\boldsymbol{Y})}{P(S_{MRCA})} P(I_2, S_{MRCA})}{\sum_{S_{MRCA}} \frac{P(S_{MRCA}|\boldsymbol{Y})P(\boldsymbol{Y})}{P(S_{MRCA})} P(I_1, S_{MRCA})} \\ &= \frac{\sum_{S_{MRCA}} \frac{P(S_{MRCA}|\boldsymbol{Y})}{P(S_{MRCA})} P(I_2, S_{MRCA})}{\sum_{S_{MRCA}} \frac{P(S_{MRCA}|\boldsymbol{Y})}{P(S_{MRCA})} P(I_1, S_{MRCA})} \end{aligned}$$

As we did when computing the Bayes Factors above, we assume that all ancestral haplotypes are *a priori* equally probable, leading to

$$\frac{P(I_2, \boldsymbol{Y})}{P(I_1, \boldsymbol{Y})} = \frac{\sum_{S_{MRCA}} P(S_{MRCA}|\boldsymbol{Y}) P(I_2, S_{MRCA})}{\sum_{S_{MRCA}} P(S_{MRCA}|\boldsymbol{Y}) P(I_1, S_{MRCA})} \tag{B.1}$$

The Bayesian phylogenetic inference provides $P(S_{MRCA}|\mathbf{Y})$, the probability of the ancestral haplotype of all human sequences given the sequence data.

We can then marginalize the joint probability $P(S_{MRCA}, I_n)$ over the root-shape vector and apply conditional probability:

$$P(S_{MRCA}, I_n) = \sum_{\boldsymbol{\tau}} P(\boldsymbol{\tau}, S_{MRCA}, I_n)$$
$$= \sum_{\boldsymbol{\tau}} P(S_{MRCA}|\boldsymbol{\tau}, I_n) P(\boldsymbol{\tau}, I_n)$$

Since the length of $\boldsymbol{\tau}$ depends on the number of introductions, In does not provide any additional information regarding the conditional probability of S_{MRCA}. Therefore,

$$P(S_{MRCA}, I_n) = \sum_{\boldsymbol{\tau}} P(S_{MRCA}|\boldsymbol{\tau}) P(\boldsymbol{\tau}, I_n)$$
$$= \sum_{\boldsymbol{\tau}} P(S_{MRCA}|\boldsymbol{\tau}) P(\boldsymbol{\tau}|I_n) P(I_n)$$

The maximum likelihood tree (Fig. 2.1), likely ancestral haplotypes (Table 2.1), and forward epidemic simulations (Fig. 2.2) allow for the following compatibility statements:

- $\boldsymbol{\tau} = \tau_{2C}$ is compatible with $S_{MRCA} \in \{S_{C/C}, S_{T/T}\}$
- $\boldsymbol{\tau} = \tau_{1C}$ is compatible with $S_{MRCA} \in \{S_A, S_B\}$
- $\boldsymbol{\tau} = (\tau_P, \tau_P)$ is compatible with $S_{MRCA} \in \{S_A, S_B, S_{C/C}, S_{T/T}\}$
- $\boldsymbol{\tau} = (\tau_P, \tau_{1C})$ is compatible with $S_{MRCA} \in \{S_A, S_B, S_{C/C}, S_{T/T}\}$
- $\boldsymbol{\tau} = (\tau_{1C}, \tau_P)$ is compatible with $S_{MRCA} \in \{S_A, S_B, S_{C/C}, S_{T/T}\}$
- $\boldsymbol{\tau} = (\tau_{1C}, \tau_{1C})$ is compatible with $S_{MRCA} \in \{S_A, S_B, S_{C/C}, S_{T/T}\}$

We can then define $P(S_{MRCA}|\boldsymbol{\tau})$ to be proportional to the vector of compatibilities:

$$P(S_{MRCA}|\boldsymbol{\tau}) \propto \begin{cases} 1 & \text{if } \boldsymbol{\tau} \text{ and } S_{MRCA} \text{ are compatible} \\ 0 & \text{otherwise} \end{cases}$$

In the case where multiple root-shapes are compatible with an ancestral haplotype, we must renormalize the compatibility vector such that it sums to 1 to describe probabilities.

However, note that some root-shapes are not compatible with any ancestral haplotypes. In this case, $P(S_{MRCA}|\boldsymbol{\tau})$ equals 0 for each ancestral haplotype.

The simulations provide $P(\boldsymbol{\tau}|I_1)$, the probability of the root-shape(s) given one introduction. We assume each introduction is independent, allowing us to generalize this probability to $P(\boldsymbol{\tau}|I_n)$. For example, $P(\boldsymbol{\tau} = \tau_{1C}|I_n = I_1) = 0.051$, and $P(\boldsymbol{\tau} = (\tau_{1C}, \tau_{1C})|I_n = I_2) = P(\boldsymbol{\tau} = \tau_{1C}|I_n = I_1)^2 = 0.0025$.

Lastly, we assume equal prior probabilities for one and two introductions of SARS-CoV-2, allowing us to cancel out $P(I_1)$ and $P(I_2)$ when calculating equation (1). We can then use $P(S_{MRCA}|Y)$, $P(S_{MRCA}|\boldsymbol{\tau})$, and $P(\boldsymbol{\tau}|I_1)$ to calculate the posterior odds of equation (1). Although we solved for the posterior odds in support of two introductions, the prior odds comparing two introductions to a single introduction is 1. Therefore, the posterior odds are equivalent to a Bayes factor in support of two introductions. The code used to calculate the Bayes factor comparing two introductions to a single introduction is available at https://github.com/sars-cov-2-origins/multi-introduction [133].

B.1.25 Combining epidemic simulations and BEAST via rejection sampling

We estimated the timing of the lineage B and lineage A primary cases. Our previous analysis [94] inferring the timing of the SARS-CoV-2 primary case incorporated the date of index case ascertainment; however, uncertainty regarding the true index case persists [22;26;226] (supplementary text). To overcome this uncertainty, we extend our previously published approach, which combines the epidemic simulations and phylodynamics tMRCA inference (described in the above two sections), to condition the timing of the primary case on both the index case symptom onset date and earliest documented COVID-19 hospitalization date. We included hospitalization dates because they are less susceptible to recall bias than date of symptom onset.

Our aim is to obtain a posterior distribution for the date X of the primary case (the first case resulting from a SARS-CoV-2 cross-species transmission) in Wuhan, conditioned on the available sequencing data D_S, the date of the first reported COVID-19 case D_C, and the date of the first hospitalization D_H due to COVID-19. We do this in a Bayesian framework by marginalizing over the date T of the tMRCA as follows:

$$P(X|D_S, D_C, D_H) = \int_T P(X|T, D_S, D_C, D_H) P(T|D_S, D_C, D_H)\, dT \tag{B.2}$$

We assume that the sequencing data are informative only for the tMRCA; i.e., given T, X does not depend on D_S: $P(X|T, D_S, D_C, D_H) = P(X|T, D_C, D_H)$. We also assume that the first reported COVID-19 case and hospitalization data are not informative for the tMRCA: $P(T|D_S, D_C, D_H) = P(T|D_S)$. This gives:

$$P(X|D_S, D_C, D_H) = \int_T P(X|T, D_C, D_H) P(T|D_S)\, dT \tag{B.3}$$

We further note that

$$P(X|T, D_C, D_H) = \int_{Y_H} \int_{Y_C} P(X, Y_C, Y_H|T, D_C, D_H)\, dY_C dY_H$$

where Y_C and Y_H are the first simulated COVID-19 ascertained case and hospitalization, respectively. We model $P(X, Y_C, Y_H|T, D_C, D_H)$ as proportional to $I(Y_C \leq D_C, Y_H \leq D_H) P(X, Y_C, Y_H|T)$, where $I(Y_C \leq D_C, Y_H \leq D_H)$ is an indicator function with a value of 1 when Y_C and Y_H are consistent with D_C and D_H, respectively, and 0 otherwise. This approach allows us to sample from the posterior distribution of equation (3). The BEAST analysis provides values of T sampled from the distribution $P(T|D_S)$, which we can use in conjunction with FAVITES-COVID-Lite to sample corresponding values of X, Y_C, and Y_H from the distribution $P(X, Y_C, Y_H|T)$. We use a simple rejection sampling approach to continue sampling from $P(X, Y_C, Y_H|T)$ until a sample is obtained for which $I(Y_C \leq D_C, Y_H \leq D_H) = 1$. The resulting set of sample values for X then follow the posterior distribution $P(X|D_S, D_C, D_H)$.

We require the first simulated case to be ascertained (SAPHIRE stage: I) and assign D_C as 10 December 2019. However, we note that this first ascertained case can be the primary case, unless a secondary or tertiary case progresses faster through the course of infection. We assign D_H as 16 December 2019. Importantly, the rate at which cases were ascertained in the SAPHIRE model is based on real-time patterns in COVID-19 diagnosis from 1 through 22 January 2020 and may not reflect the actions that led to the retrospective diagnosis of the

earliest cases of COVID-19. Further, stable coalescence (i.e., the MRCA) can happen any time after the primary case is infected, and there is no requirement for stable coalescence to occur after the first ascertained and unascertained individuals. Justifications for dates used here and in the sensitivity analyses are discussed in the supplementary text.

B.1.26 Sensitivity analysis—earliest case date of 8 December

We also condition single-introduction analyses on an 8 December case date, which was previously discounted by the WHO and 16 December hospitalization date (see supplementary text for full discussion).

B.1.27 Sensitivity analysis—rejection sampling with hospitalization only

We remove the requirement for the first simulated case to be ascertained by a given date, and then condition analyses only on the tMRCA and date of the first hospitalization.

B.1.28 Sensitivity analysis—recCA and constrained roots

We explored the sensitivity of the timing of the primary case to the phylodynamic model choice. We applied rejection sampling to the inferred phylogenies constrained by the recCA and SARS-CoV-2 ancestral haplotypes (e.g., lineage A.1) and the primary forward epidemic simulations, conditioning on the same dates as above.

B.1.29 Rejection sampling for lineages A and B

We apply the above method to the tMRCA, first ascertained case date, and first hospitalized case date for each lineage to infer the timing of the primary case for each lineage. For lineage B, the earliest case and hospitalization dates are 13 December and 16 December, respectively. For lineage A, the earliest case date is 15 December, and the earliest hospitalization date is 25 December (see supplementary text for full discussion). After performing rejection sampling for both lineages, we combine the number of individuals in each compartment for each day in the dated simulations.

B.1.30 Sensitivity analysis—earlier lineage B COVID-19 index case dates

We performed rejection sampling for lineage B using a case date and hospitalization date of 10 December and 16 December, matching the SARS-CoV-2 index case, which does not have an associated genome. We also performed rejection sampling for lineage B using a case date and hospitalization date of 8 December and 18 December, respectively, under the alternate assumption that the earliest lineage B case occurred on 8 December (see supplementary text for full discussion).

B.1.31 Simulating cross-species transmissions to achieve two successful introductions

Our epidemic simulations had a success rate of approximately 22.2% (1100 successful introductions; 3857 failed introductions). To simulate the number of cross-species transmissions needed to achieve two successful introductions, we treated successful introductions as Bernoulli trials, with a success rate of 22.2%, and simulated trials until there were two successful trials.

B.2 Supplementary text

B.2.1 Reduced genomic diversity in the early pandemic

Although the root of the SARS-CoV-2 phylogeny has often been inferred with Bayesian and maximum likelihood methods to fall on the branch leading to IPBCAMS-WH-01 (Lineage B) or other early genomes[89;94;104;227;228], reanalysis of sequence data from the earliest sampled viruses found that three previously reported mutations in IPBCAMS-WH-01 were spurious, and the genome was, in fact, identical to the Hu-1 reference genome[26]. Upon reexamination by WHO investigators, other early genomes were also found to have spurious mutations[26], thereby decreasing the overall genetic diversity of early SARS-CoV-2 sequence data. This decreased diversity suggests that prior studies, including our own[94], may have incorrectly rooted the SARS-CoV-2 phylogeny.

B.2.2 C/C and T/T genomes through 14 February 2020

Of the 16 C/C genomes in our data set, we found four that share nucleotide substitutions—other than T28144C—found in some lineage A genomes (Fig. B.1A). If the C/C intermediates actually existed, 11 of the 19 additional unique mutations in the C/C genomes would need to be homoplasies: identical mutations arising separately in the C/C and lineage A genomes. For example, a C/C genome from Anhui province (EPI_ISL_1069206) shares A11430G with 6 lineage A genomes sampled across China, and a genome from Sichuan province (EPI_ISL_451325) shares C1342T and C18060T with three lineage A genomes. The authors of the latter example confirmed that low sequencing depth at position 8782 led to the erroneous calling of the reference genome nucleotide at this position in this genome (L. Chen Personal Communication). Furthermore, these authors confirmed that incorrect base calls, often due to low sequencing depth, led to erroneous assignment of 11 additional C/C genomes sampled in Wuhan and Sichuan province (four of which share substitutions with lineage B genomes, see below).

A similar pattern was observed in the five C/C genomes sharing substitutions found within lineage B (Fig. B.1B), including a South Korean genome (EPI_ISL_413017) sharing G26640T, G26144T, and T26677C with another lineage B genome from South Korea. In this

instance, we confirmed that low sequencing depth at position 28144 (<10x) resulted in this erroneous assignment. Critically, therefore, we are able to explain all C/C genomes as artifactual, with the exception of two genomes sampled in Beijing in late January and early February, whose additional mutations were not observed in early lineage A or B genomes and whose underlying data was not available.

Unlike the C/C genomes, none of the four T/T genomes shared additional mutations with lineage A or B genomes that would clarify their veracity. However, we confirmed that the T/T genome sampled in Singapore on 14 February 2020 (EPI_ISL_462306) had low coverage at both 8782 and 28144 (≤10x). Moreover, the 3 T/T genomes sampled in Wuhan on 26 January (EPI_ISL_493179, EPI_ISL_493182, EPI_ISL_493180) had low sequencing depth and indeterminate C/T nucleotide assignment at position 8782 (B.1). These findings suggest all T/T genomes sampled by 14 February 2020 are similarly artifactual.

B.2.3 T/T genomes aboard the Diamond Princess

Two high coverage T/T genomes sampled after 14 February 2020 were from the Diamond Princess cruise ship outbreak. These two T/T genomes possessed T11083G, a mutation defining the Diamond Princess outbreak, as well as G11410T, and were therefore identifiably descendants of the lineage B virus that initiated this outbreak[229;230](Fig. B.2).

B.2.4 T/T genomes in New York City

We found 3 SARS-CoV-2 genomes with both C8782T and T28144 (T/T) in the NYC Public Health Laboratory surveillance data set. We placed these genomes on a global tree of 3 million genomes (v2022-01-21) using UShER[231], and all 3 appear to be descendants of lineage B. Two of these genomes (EPI_ISL_8953704 and EPI_ISL_8953705) belong to the B.1.526 Iota lineage and have identical sequences, suggesting that a T/T genome may have been transmitted locally (Fig. B.3). We note that although these genomes were sampled from different individuals in different parts of the city, their genomes were sequenced on the same sequencing plate. The third NYC T/T genome (EPI_ISL_1447116) belongs to the B.1.2 lineage and falls on a different part of the phylogeny, indicating an independent C8782T mutation. This third genome was sequenced on a separate run than the other two NYC T/T genomes. All three genomes at site

8782 and 28144 have read depth >4,000x, with coverage in both directions and 100% of the sequencing reads support the T allele.

B.2.5 T/T and C/C genomes in San Diego

We found 24 T/T genomes in the San Diego SEARCH data set, with collection dates between December 2020 and December 2021 and 1 C/C genome that was collected on 15 January 2021. We placed these genomes on a global tree of 3 million genomes using UShER. Eight of the T/T genomes were classified as Delta sublineages: AY.26, AY.40, and AY.44; the remaining T/T genomes were classified as B.1 and its other descendant lineages (Fig. B.4). The C/C genome was classified as B.1.1.432 (Fig. B.5). Therefore, these and other T/T and C/C haplotypes are the result of convergent evolution at 8782 and 28144.

The common occurrence of C/C and T/T genomes arising due to convergence provides further evidence that early intermediate genomes could have also been due to convergent evolution and do not represent transitional genomes.

B.2.6 The recombinant common ancestor (recCA)

The recCA differed from Hu-1 by just 381 reversions, including C8782T and T28144C. In this manner, lineage A (exemplified by Wuhan/WH04/2020) has two reversions, and hence has two fewer substitutions separating it from the recCA than lineage B. (Note that although these mutations are nominally 'reversions', if the true MRCA of SARS-CoV-2 were a lineage A virus, those lineage B to lineage A reversions would not actually have occurred). Additional reversions, C18060T and C29095T, have been separately identified in USA/WA1/2020 and Guangdong/20SF012/2020, respectively, and it has been argued that these haplotypes are the ancestral form of SARS-CoV-2[90;92]. We find that repeated substitutions at sites 8782, 18060, and 28144, are common among closely related sarbecoviruses (Fig. B.11-SB.13). In contrast, 29095 is strongly conserved among these sarbecoviruses but highly polymorphic in humans (Fig. B.14). Absent from the recCA are two mutations, C2416T and C23929T, previously suggested to have been present in the immediate ancestor of SARS-CoV-2[92]; these mutations occur on the branch to the related bat sarbecovirus RaTG13 (Fig. B.8, B.10)[232].

To ensure our downstream analyses are robust to the breakpoint inference and subsequent

construction of the recCA, we reconstructed a second recCA using 3SEQ, a recombination detection algorithm which examines triplets (two parents and one offspring sequence)[214]. We then constructed this second recCA with the 21 new breakpoints, and this recCA differed from Hu-1 at 371 sites. The two recCAs were identical at sites previously suggested to have mutations present in the immediate ancestor (i.e., 2416, 8782, 18060, 19524, 23929, 28144, and 29095).

B.2.7 Justification for a non-reversible substitution model

We developed a random-effects non-reversible substitution model for our phylogenetic inference because of the substantial C-to-T transition bias[86–88]and frequent C-to-T reversions (described in Main Text). To compare the ancestral haplotype inference between the random-effects model and a standard reversible substitution model, we performed Bayesian phylodynamic inference with a GTR substitution model with both the unconstrained and recCA-rooted analyses. We find that C/C and T/T ancestral haplotypes were less common under the GTR model than the random-effects model (Data S3). Notably, the difference in posterior support for C/C and T/T ancestral haplotypes was negligible under the unconstrained GTR model, indicating the increased level of biological realism reflected in the random-effects substitution model inference. Importantly, ancestral haplotypes such as A.1 (BF>150) and A+C29095T (BF>50) were poorly supported under the GTR model, just as in the unconstrained and the recCA random-effects model.

B.2.8 Sensitivity analyses of ancestral haplotype inference with unconstrained rooting

We extended the analysis that excluded the 15 market-associated genomes to account for any potential ascertainment bias of early sampling, particularly of lineage B, by excluding all 96 genomes from Wuhan (59 lineage B, 37 lineage A). Upon excluding these genomes from Wuhan, the posterior support for a lineage B ancestral haplotype decreased, while the support for C/C and lineage A ancestral haplotypes increased, with all other lineages still unsupported (Data S3). Therefore, lineage A is only a plausible ancestral haplotype (BF<10) of SARS-CoV-2 under the unconstrained model if we do not include all the early genomes from Wuhan, which represent a substantial portion of the early genomic diversity.

To understand the impact of excluding the C/C and T/T "intermediate" genomes, we performed Bayesian phylogenetic inference of SARS-CoV-2 with the original 787 genomes (see Methods) and the 20 excluded intermediate genomes, with unconstrained rooting. When including all 20 "intermediate" genomes—16 C/C and 4 T/T genomes—the results were similar to the main unconstrained analysis (Data S3). These results were also consistent when we included only the C/C or the T/T genomes separately, indicating the posterior support for a C/C or T/T ancestral haplotype does not increase when including intermediates, even if only including genomes with just one of the two intermediate ancestral haplotypes (Data S3).

When we included the previously masked 388 genomic sites for ancestral haplotype inference of the 787 genomes, our results were consistent with our primary masked unconstrained rooting approach (Data S3).

B.2.9 Ancestral haplotype inference with outgroup rooting

We used outgroup rooting with individual bat sarbecoviruses as a comparison for the recCA rooting inferences. We performed the Bayesian inference with either RaTG13 or BANAL-20-52 as an outgroup. Relative to the recCA rooting results, posterior support for a lineage B and C/C ancestral haplotypes increased, whereas the support for the lineage A ancestral haplotype decreased (Data S3). All other previously suggested or inferred haplotypes, including lineage A.1, were rejected, consistent with the other rooting approaches (Data S3).

These results indicate that decreasing the branch length from the MRCA of SARS-CoV-2 and its ancestor (i.e., the MRCA of SARS-CoV-2 and the outgroup, or the recCA) increases support for lineage A as the ancestral haplotype (when the recCA and the outgroup are identical at key sites—here, 8782 and 28144). These findings indicate rooting with the recCA as opposed to an outgroup can help account for the effects of rate variation along the long branch leading to SARS-CoV-2.

B.2.10 The tMCRA of SARS-CoV-2 is consistent across explored ancestral haplotypes

It has been suggested that a phylogenetic root in lineage A would produce older tMRCA estimates than a lineage B rooting[92]. However, we find that SARS-CoV-2 tMRCA inference is

generally robust to the rooting model and ancestral haplotype.

The unconstrained rooting (Fig. B.29A), which favored a lineage B ancestral haplotype (B.2), and produced a median tMRCA of 11 December 2019 [95% highest posterior density (HPD): 25 November to 20 December] and a mean substitution rate of 9.2x10^{-4} substitutions/site/year (95% HPD: 8.1x10^{-4} to 1.0x10^{-3}). These tMRCA estimates are similar to our previous inference[94], although the substitution is slightly faster. This elevated rate is expected, given that shorter sampling windows are associated with the inference of a more rapid substitution rate in SARS-CoV-2[94;233;234]. The recCA-constrained rooting (Fig. B.18C), which favored a lineage A ancestral haplotype (Table 2.1), produced a median tMRCA of 6 December 2019 (95% HPD: 15 November to 19 December) and a mean substitution rate of 9.2x10^{-4} substitutions/site/year (95% HPD: 8.0x10^{-4} to 1.0x10^{-3}).

To explicitly explore the effect of ancestral haplotype on the SARS-CoV-2 tMRCA, we employed our novel phylodynamic framework that fixes the MRCA of the SARS-CoV-2 phylogeny to ancestral haplotypes (Fig. B.18D, see methods), rather than using sampled taxon (e.g., Hu-1), an outgroup (e.g., RaTG13), or their inferred ancestor (e.g., recCA). We explored the plausible ancestral haplotypes (lineage A, lineage B, and C/C), as well A.1 (WA1) and A.1 + C29095T (20SF012)(see methods). The resulting tMRCAs were consistent across the ancestral haplotypes (B.2), indicating the ancestral haplotype has minimal impact on tMRCA inference.

The tMRCA of lineage B predates the tMRCA of lineage A when excluding market-associated genomes. We considered the possibility that the predominance of lineage B viruses in the beginning of the pandemic, particularly at the Huanan market, was biasing the earlier inference of the lineage B tMRCA. However, when we excluded all market-associated genomes, the median tMRCA of lineage B was 17 December (95% HPD: 29 November to 26 December), still earlier than the median tMRCA of lineage A: 25 December (95% HPD: 15 December to 4 January). Therefore, the predominance of lineage B at the Huanan market is not biasing its tMRCA to predate the tMRCA of lineage A (B.2).

B.2.11 Doubling time of the early pandemic

To properly parameterize the epidemic simulations, we characterized the epidemic growth dynamics of the early pandemic. We first extracted the daily case counts from the WHO report for December 2019[26] and then adjusted the case counts reported in Li *et al.*[23] to properly reflect the earliest cases identified (Fig. B.21A; see Methods). We used EpiNow2[218] to calculate the daily doubling time of SARS-CoV-2 through mid-January 2020, which resulted in a median doubling time of SARS-CoV-2 and 50% HPD consistently below 5 days through the end of December 2019 (Fig. B.21B). On 15 December 2019, five days after the earliest known COVID-19 case, the median doubling time was 3.74 days (50% HPD: 3.37–4.24; 95% HPD: 2.80–5.43). However, we are interested in the epidemic dynamics prior to the discovery of the earliest COVID-19 cases and the identification of SARS-CoV-2 as the etiological agent of what was then a pneumonia of unknown etiology, when viral spread was least inhibited by human behavior.

The inferred doubling time is stochastic and dependent on which day it is estimated (Fig. B.21B); it is therefore imprecise to report a single doubling time when describing general epidemic dynamics. In our simulations, we measure doubling time in four ways (Fig. B.22): (i) cumulative doubling time since the start of the simulation, (ii) 14-day doubling time from day 14 until the end of the simulation, with cumulative doubling time reported prior to day 14, (iii) cumulative doubling time once a certain number of individuals are infected (e.g., the cumulative doubling time at the 100th infection), and (iv) 14-day doubling time once a certain number of individuals are infected, with cumulative doubling time reported if that number of infections occurred before day 14 in the simulation. The 50% and 95% highest density intervals (HDIs) of the doubling time across simulations at cumulative infection counts [(iii) and (iv) above] narrows as more individuals are infected in the simulations (Fig. B.22, bottom panels), although HDIs of the doubling time reported daily [(i) and (ii) above] in the simulation (Fig. B.22, top panels) remain wide.

These results suggest doubling time based on the number of individuals infected [(iii) and (iv) above] is a better metric for understanding simulation growth dynamics than daily doubling times [(i) and (ii) above], because the simulations will typically reach this number of infections and therefore be at a similar "point" in the epidemic. For example, the 14-day doubling time

at the thousandth case in each simulation has a narrower HDI than the 14-day doubling time reported on the same day across each simulation, regardless of the day. We therefore focus on the doubling time once a specific number of infections has been reached, and as doubling time is proportional to the epidemic speed[235]at a specific time point, we report the 14-day doubling time. Specifically, we report the 14-day doubling time once there have been one thousand infections, as the epidemic network has not yet been saturated, but the doubling time has a narrower HDI than earlier in the simulations.

We note that although the doubling time of SARS-CoV-2 was initially reported to be longer than 5 days[25;235], the case counts of the early pandemic have become more thorough[23], particularly with the release of the WHO report[26]. A subsequent analysis of the doubling time in Wuhan used high resolution travel data and early case reports outside of Hubei province to avoid potential biases in reporting and case confirmation in Wuhan; their inferred doubling time was below 3.5 days[97]. Additionally, doubling times reported for well-characterized early outbreaks outside of China, before widespread non-pharmaceutical interventions were implemented, were often below four days[98;236;237].

Following from the empirical doubling times, we parameterize our primary simulations to have a median doubling time of 3.47 days (50% HDI: 2.71–4.15; 95% HDI: 1.35–5.44), slightly shorter than the inferred doubling time on 15 December 2019. We additionally perform sensitivity analyses with simulation parameterization resulting in a median doubling time of 2.65 days (50% HDI: 2.04–2.96; 95% HDI: 1.50–4.10) and 4.45 days (50% HDI: 3.30–5.35; 95% HDI: 1.50–7.44).

Importantly, the doubling times of the simulated data indicate that it is possible to infer a doubling time of greater than 5 days, even when the underlying parameters typically produce a shorter doubling time (Fig. B.22). Similarly, our empirical doubling time estimates (Fig. B.21) and the results from the literature suggest that it is possible to infer a doubling time of greater than 5 days[235;238]even when the underlying biology would typically produce a shorter doubling time[97](Fig. B.22), especially if cases are particularly undersampled. Coupled with our results indicating the doubling time was likely below four days in the early pandemic in Wuhan, these analyses of epidemic doubling times both in and outside of Wuhan suggest our simulations likely capture the doubling time of SARS-CoV-2 before its identification as the etiological agent of

COVID-19.

B.2.12 Ascertainment rate of the early pandemic and parameterizing the simulations

Our epidemic simulations use an ascertainment rate of 15% (every 15/100 infections are ascertained) based on Hao *et al.*[222]. Their ascertainment rate was informed by confirmed cases exported from Wuhan to Singapore. This ascertainment rate falls in the interquartile range (12.7–35.8%) reported in Chinazzi *et al.*[239]and is close to the 14% reported in *Li et al.*[240].

However, some of the earliest inferred ascertainment rates were lower: Wu *et al.*[235]calculate an ascertainment rate of 1.8% (0.9–3.3%). This value was likely reflective of the data available at the time: Wu *et al.* only reported 44 confirmed cases by 3 January 2020. However, more complete data from Li *et al.*[23]reports 392 confirmed infections by 3 January 2020. With 392 confirmed infections, a crude calculation leads to an ascertainment rate approximately 10-fold higher: 392/(44/0.018)= 16.0%. Therefore, some of the earliest estimations of ascertainment rates were likely too low for our simulation parameterization because we base our simulations on inferences of more recent and complete data[23;26].

Nonetheless, as there are lower and higher ascertainment rates reported in the literature describing the early pandemic, we performed sensitivity analyses of our epidemic simulations using ascertainment rates of 5% and 25%. When doing so, we slightly adjusted the transmission factor for the simulations to keep the doubling time centered on 3.5 days. Specifically, we performed one set of simulations with an ascertainment rate of 5% and median doubling time of 3.57 days (50% HDI: 2.96–3.55; 95% HDI: 1.38–5.64), and another set of simulations with an ascertainment rate of 25% and median doubling time of 3.50 days (50% HDI: 2.78–4.16; 95% HDI: 1.51–5.65).

B.2.13 Synthesizing evidence for multiple introductions of SARS-CoV-2

Both lineages A and B are characterized by large polytomies: many sampled lineages descending from a single node on the phylogenetic tree. There are 108 and 231 lineages (including basal taxa) descendent from the base of lineages A and B, respectively (Fig. 2.1). To match the empirical data, we first examined the simulations for polytomies with at least 100 descendant

lineages, and these were present in 47.5% of the simulated epidemics when the doubling time is 3.47 days (95% HDI: 1.35–5.44). (Fig. 2.2A). However, this is, in fact, a conservative estimate of the polytomy size in our simulated phylogenies, because 98.4% of these simulated phylogenies had more than 1,000 taxa.

If C/C is the ancestral haplotype, then SARS-CoV-2 is characterized by two clades: lineages A and B, each one mutation from the root with no transitional genomes (Fig. 2.2B). This topology, where there are only two clades of any size, each one mutation from the root, was present in 10.5% of phylogenies from our simulated epidemics. However, both lineages A and B are large clades, comprising 35.2% and 64.8% of the early SARS-CoV-2 genomes, respectively, and the smaller clade in these simulations was rarely this large. If we require our simulated clades to more realistically comprise at least 1% of the taxa, only 6.7% of the simulations match the C/C topology. If we require both clades to comprise $\geq$30% of the taxa—better reflecting empirical genomic diversity—only 1.5% of the simulations match the C/C topology. Finally, both lineages A and B comprise large polyomties. When we require each of these clades to have a basal polytomy of at least 100 descendant lineages—a conservative reflection of the 108- and 231-lineage polytomies characterizing lineages A and B, respectively—none of the simulations still match the C/C topology. These results indicate that a single introduction of C/C virus would not be expected to give rise to lineages A and B with no surviving ancestral C/C lineages.

If lineage A or B is the ancestral haplotype, then SARS-CoV-2 is characterized by a large basal polytomy with the largest clade in the tree separated by two mutations from the root (lineage B is the descendant clade if lineage A is the root, and vice-versa)(Fig. 2.2C). Importantly, our simulations permit these two mutations to occur either within a single individual or during successive infected hosts[241], reflective of multiple mutations of SARS-CoV-2 occuring within the serial interval between transmission partners[242]. We see a large clade comprising a substantial fraction of the sampled taxa (i.e., between 30% and 70%, reflecting either lineage A or B prevalence) in 10.8% of the epidemic simulations. When we require the large clade separated by at least two mutations from the basal polytomy of at least 100 descendant lineages, we observe this topology in 4.1% of epidemic simulations. However, if we also require the large clade to have at least a 100-lineage polytomy at its base, only 0.5% of the simulations match the topology

if there were a single introduction of lineage A or B without any surviving transitional C/C lineages.

We then quantified the support for two introductions versus a single introduction resulting in the observed phylogeny: two large polytomies separated by two mutations. There was substantial support for two introductions with our primary analysis (BF=4.3 and BF=4.2 with the recCA and unconstrained rooting, respectively; see Methods), as well as with sensitivity analyses with varying transmission and ascertainment rates (B.5). We observe that sensitivity analyses with shorter doubling times increase the support for multiple introductions (B.5)

Importantly, since 96.7% of the simulated phylogenies had more than 5,000 taxa, our primary analysis requiring polytomies to have at least 100 descendant lineages is quite conservative. We therefore performed sensitivity analyses requiring the polytomies to have either 200 or 500 descendant lineages. These increased polytomy sizes are more realistic when considering the lineage A polytomy and the empirical phylogeny. However, both polytomy sizes are still proportionally less than the 108 descendant lineages of lineage A from our empirical 787-taxa phylogeny and are therefore still conservative. Regardless, the support for two introductions increased with both of these analyses (B.5), again indicating the rarity of two polytomies occurring in rapid succession.

Additionally, since 1.5% of the simulated phylogenies had fewer than 787 taxa, we also performed sensitivity analyses requiring the polytomies to have either 20 or 50 descendant lineages. Regardless of the polytomy size we conditioned on, there was still support for two introductions (B.5).

In sum, there is consistent support for two introductions of SARS-CoV-2 across multiple rooting constraints, longer and shorter doubling times, higher and lower ascertainment rates, and increased and decreased minimum polytomy sizes. Therefore, two introductions, rather than a single introduction, of SARS-CoV-2 are more likely to have produced the two polytomies near the base of the SARS-CoV-2 phylogeny.

B.2.14 Index case dates of SARS-CoV-2

The date of the index case informs our understanding of when the pandemic began, and we use this date when inferring the time of the primary case. An early report suggested an index case (i.e., first identified case) dating to 17 November 2019, with at least one case being reported each day thereafter[243]. A different report suggested an index case date of 1 December 2019[25]. However, the World Health Organization (WHO)-China report did not find evidence to support the veracity of these cases and identified the earliest case as having an illness onset of 8 December (case S01 from Table 6 in the WHO report)[26]. A subsequent review of the earliest COVID-19 cases suggested that the '8 December' patient actually became ill on 16 December and concluded that the index case was a vendor from the Huanan Seafood Market who became ill on 10 December and was hospitalized on 16 December[22]. This shift in index case dates necessitates reexamining the timing of the primary case (i.e., first case resulting from a cross-species transmission) of SARS-CoV-2, as case data is crucial to timing the first SARS-CoV-2 infection[94].

When we perform rejection sampling to infer the date of the primary cases of lineages A and B separately, we use case and hospitalization dates associated with each lineage. The SARS-CoV-2 index case from 10 December does not have an associated published genome. However, every genome associated with the Huanan market and collected before 30 December was lineage B. Additionally, an environmental sample from the stall this vendor operated (EPI_ISL_408512) was also lineage B. We therefore assume this individual likely had a lineage B virus, and we used their illness onset and hospitalization dates to time the primary case of lineage B.

The earliest case and hospitalization dates for lineage A belong to 'Cluster 1' from the WHO report (Annex E2)[26]. The lineage A genome belongs to the individual in the cluster with illness onset on 26 December (case S13 from Table 6 in the WHO report; IME-WH01; GISAID accession EPI_ISL_529213). However, the spouse of this individual was infected as well, becoming symptomatic on 15 December and hospitalized on 25 December. We can therefore reasonably assume the spouse was also infected with a lineage A virus and subsequently infected the earliest confirmed lineage A case. We thus use the dates belonging to the spouse when timing the primary case of lineage A.

B.2.15 Robustness of lineage B primary case inference to date of index case

The earliest lineage B case with a published genome (IPBCAMS-WH-01; GISAID accession EPI_ISL_402123; case S02 from Table 6 in the WHO report) became symptomatic and was hospitalized on 13 and 18 December, respectively[74]. Although this individual was initially reported to have an illness onset date of 15 December, the WHO report subsequently determined he had an earlier onset of 13 December[26]. The timing of the lineage B primary case is robust to these later symptom onset and hospitalization dates, occurring on 21 November (95% HPD: 25 October–11 December) when rooting with recCA.

Because the '8 December' patient was lineage B[26], we examined the effect of conditioning on 8 December as an index case date. The timing of the lineage B primary case was robust to this earlier index case date (Table B.7), occurring on 17 November (95% HPD: 22 October to 6 December) when rooting with recCA. Therefore, even if the 8 December case were the index case, the first lineage B SARS-CoV-2 virus still likely jumped into humans at a time similar to our inference with the 10 December index case date.

B.2.16 Epidemiological dynamics of a single SARS-CoV-2 introduction scenario

As a counterfactual scenario, we examined the timing of the primary case if SARS-CoV-2 were the result of only a single introduction. Here, we condition on the tMRCA of all SARS-CoV-2 (Table B.2) and the ascertainment (10 December) and hospitalization (16 December) dates belonging to the SARS-CoV-2 index case (the seafood vendor). The timing of the primary case in a single introduction scenario is similar to that of lineage B (Table B.7, B.14), even when using lineage A.1 or lineage A + C29095T as the ancestral haplotype, and is also robust to higher and lower ascertainment rates (Table B.15). The number of infections and hospitalizations at specific dates resulting from a single introduction are similar to the combined infections and hospitalizations from lineages A and B (Fig. 2.4A, B.24A, B.29; B.11, B.12), even when using the ancestral haplotypes producing the oldest tMRCA estimates (lineage A.1 and lineage A + C29095T). These results indicate that there was likely not substantial cryptic spread even if lineage A.1 or A + C29095T were the sole introduction of SARS-CoV-2, and that standard

epidemic modeling that does not account for phylogenetics will not be able to distinguish the number of introductions based on case-counts or hospitalizations alone.

B.2.17 Consistent timing of the primary case when conditioning only on hospitalization

Due to the controversy that remains about the true SARS-CoV-2 index case[22], we performed sensitivity analyses on our primary case timing by conditioning just on the tMRCA and the date of the earliest hospitalization. The timing of the primary case was robust to the exclusion of the index case date in our rejection sampling approach for both single- and multi-introduction scenarios, indicating that index case dates did not bias our results toward earlier dates (Table B.7, B.9, B.14). Had there been extensive cryptic spread of SARS-CoV-2 in Wuhan, it would be reflected in earlier hospitalization dates. The consistency of our results when conditioning on (i) both the index case ascertainment and earliest hospitalization, versus (ii) only the earliest hospitalization, suggests there was a very limited period of cryptic spread before people began to be hospitalized for COVID-19 in late-2019.

B.2.18 Minimal cryptic circulation before December 2019

Although we do not see any evidence for substantial cryptic circulation before December 2019 with the epidemic simulations (Fig. 2.4), we can quantify the expected number of infections before the tMRCA. We calculate the cumulative number of infections in the epidemic simulations once they reach stable coalescence, the point in time at which basal lineages cease to be lost. Importantly, the time to stable coalescence is the equivalent to the tMRCA for the epidemic simulations. We observed, at most, 63 cumulative infections by the time of stable coalescence in the primary simulated epidemics, and 99% of simulated epidemics reached stable coalescence by 19 cumulative infections. With the tMRCA of lineage B, likely the first lineage of SARS-CoV-2 introduced into humans, estimated to 13 December (95% HPD: 29 November to 23 December), we would not expect more than a few dozen infections before 10 December, the date of the SARS-CoV-2 index case.

B.2.19 Similarities to WA1 and WA outbreak clades

To understand the phylogenetic signal of a hypothetical singular introduction of SARS-CoV-2 into humans, it is helpful to seek an analogy with the earliest introductions of SARS-CoV-2 into a new location, such as Washington State, Louisiana, or Lombardy, each of which had a polytomy[82;99;100]. Here, we consider the earliest introductions of SARS-CoV-2 into North America. The first confirmed case of SARS-CoV-2 in the U.S. was associated with a virus strain ('WA1') isolated in Washington State from a traveler who returned from Wuhan, China, on 15 January 2020. There was subsequently an outbreak (henceforth, 'WA outbreak clade') in Washington State, with cases confirmed starting in February 2020[98]. As we have previously shown, although the MRCA of the WA outbreak clade differed from WA1 by only two substitutions, the WA1 WA outbreak clades were, in fact, separate introductions into Washington State[82]. The WA outbreak clade showcases a basal polytomy, and although the WA1 introduction was contained, onward transmission would have likely led to a basal polytomy as well, as shown by the hypothetical polytomy in Fig. B.23. Therefore, this pattern is remarkably similar to that seen with lineages A and B, with the exception of a successful prevention of onward transmission from the WA1 case: the MRCA of WA1 and the WA outbreak clades was in China, and as we have shown here, the MRCA of lineages A and B was likely in the intermediate host reservoir (Fig. B.30). Both scenarios show introductions of SARS-CoV-2 from a prior location: China in the case of WA1 and the WA outbreak clades, and the intermediate host reservoir in the case of lineages A and B. Similarly, both scenarios lead to (or would lead to, in the case of WA1) basal polytomies from the onward transmission. Therefore, lineages A and B look like separate introductions because introductions with sustained onward transmission result in large basal polytomies.

B.2.20 MRCA of multiple introductions and the possibility of intermediate genomes

Although the unconstrained and recCA-rooted phylogenetic inferences most strongly support a lineage B and lineage A ancestral haplotype, respectively, the moderate support for a C/C ancestral haplotype from both inferences, the repeated observation of C8782T convergent

evolution, and the C-to-T mutational bias suggest a C/C ancestral haplotype is also plausible. However, we acknowledge that if the MRCA were in an animal where we have no evidence of a C-to-T mutational bias, the T/T ancestral haplotype would also be possible.

If one of the "intermediate" genomes that shares additional mutations with lineage A or B (i.e., at sites besides 8782 and 28144) is resequenced and validated as an intermediate C/C or T/T genome, shared synapomorphies should allow us to identify whether this intermediate haplotype is the result of convergent evolution. Phylogenetic placement can confirm whether the intermediate genome potentially represents the transition from lineage A to B (or B to A). Additionally, finding a true transitional intermediate genome would not resolve the conundrum with the molecular clock: (i) the unconstrained rooting model disfavors a lineage A root, whereas the recCA- and outgroup-rooted models favor a lineage A root, and (ii) lineage A exhibits less divergence from the root than would be expected if it were the sole ancestral virus in humans.

B.2.21 Lineages of SARS-CoV-2 introductions

The lineages introduced into humans are dependent on the viral diversity in the intermediate host, and the inferred ancestral haplotypes do not necessarily need to match the genomes of the introduced viruses. For example, lineages A and B are observed in humans, but the introduced viruses could have been of lineage A and C/C, with C/C quickly mutating into lineage B before leaving behind any descendant lineages. However, the simplest explanation would be lineage A and B progenitors circulating in animals and then these two lineages are separately introduced into humans, but other combinations of lineage introductions are plausible.

Importantly, these scenarios do not preclude an intermediate C/C or T/T haplotype from being introduced. If lineages A and B were present in the animal reservoir, a C/C or T/T haplotype could have circulated among the animals and then been introduced into humans as well. Considering the genomic data and high frequency of failed introductions in the epidemic simulations, the intermediate haplotype could have spread briefly among humans and then gone extinct. Therefore, it is not unreasonable to assume that an intermediate genome could appear in an environmental sample or an additional cross-species transmission.

Lastly, if a virus identical to or descendant from lineage B (or A) jumped into humans

after the initial jump of B (or A), we would likely not have the phylogenetic resolution to detect this event as a separate introduction. Therefore, although the data are consistent with two introductions, it is possible that even more introductions into humans occurred at the Huanan market.

B.3 Supplementary Tables

Table B.1. Nucleotide variant calls at positions 8782 and 28144 for three SARS-CoV-2 genomes with intermediate T/T haplotypes[1].

GISAID accession	8782									28144								
	Depth	Count				Proportion				Depth	Count				Proportion			
		A	C	G	T	A	C	G	T		A	C	G	T	A	C	G	T
EPI_ISL_493179	64	0	39	1	24	0.000	0.609	0.016	0.375	61361	121	3784	195	57261	0.002	0.062	0.003	0.933
EPI_ISL_493180	40	0	24	1	15	0.000	0.600	0.025	0.375	95374	226	5709	293	89146	0.002	0.060	0.003	0.935
EPI_ISL_493182	29	0	10	0	19	0.000	0.345	0.000	0.655	69369	153	4051	232	64933	0.002	0.058	0.003	0.936

[1]Variant calls and depths provided by Di Liu and Yi Yan.

Table B.2. Inferred tMRCAs for SARS-CoV-2, lineage B, and lineage A under different rooting strategies.

Phylodynamic analysis	SARS-CoV-2[1]	Lineage B[1]	Lineage A[1]
Unconstrained	12-11 (11-25 to 12-20)	12-13 (11-29 to 12-23)	12-25 (12-17 to 12-30)
recCA	12-06 (11-15 to 12-19)	12-15 (12-05 to 12-23)	12-20 (12-05 to 12-29)
Lineage B	12-12 (11-27 to 12-19)	12-13 (11-29 to 12-21)	12-25 (12-18 to 12-29)
C/C	12-08 (11-19 to 12-19)	12-16 (12-06 to 12-23)	12-21 (12-12 to 12-29)
T/T	12-08 (11-19 to 12-19)	12-15 (12-06 to 12-23)	12-21 (12-12 to 12-29)
Lineage A	12-07 (11-18 to 12-19)	12-16 (12-08 to 12-23)	12-18 (12-04 to 12-28)
Lineage A + C29095T	12-05 (11-17 to 12-19)	12-17 (12-10 to 12-23)	12-05 (11-17 to 12-19)
Lineage A.1	12-04 (11-16 to 12-18)	12-16 (12-10 to 12-23)	12-04 (11-16 to 12-19)
No markets	12-13 (11-25 to 12-26)	12-17 (11-29 to 12-26)	12-25 (12-15 to 01-04)

[1]Median and 95% HPD in parentheses.

Table B.3. Simulation parameters, with all parameters except b and h based on Hao *et al.*[132]. The b value listed is for the main analysis. The h parameter is not included in the Hao *et al.* model.

Parameter	Definition	Value
b	Transmission rate of ascertained cases	0.38
r	Ascertainment rate	0.15
α	Ratio of transmission for unascertained	0.55
h	Hospitalization rate	0.5
D_e	Latent period (days)	2.9
D_p	Presymptomatic infectious period (days)	2.3
D_i	Symptomatic infectious period (days)	2.9
D_q	Duration from illness onset to isolation (hospitalization) (days)	11
D_h	Isolation (hospitalization) period (days)	30

Table B.4. The inferred doubling time for each transmission rate (b) and ascertainment rate (r) combination used in the simulations.

b	***r***	**Doubling time**[2]
0.22	0.15	4.45 (1.50-7.44)
0.28[1]	0.15	3.47 (1.35-5.44)
0.38	0.15	2.65 (1.50-4.10)
0.295	0.05	3.50 (1.51-5.65)
0.255	0.25	3.52 (1.38-5.64)

[1]Primary analysis.
[2]Median and 95% HDI in parentheses

Table B.5. Frequencies of observed topologies in epidemic simulations and corresponding Bayes factor in favor of multiple introductions versus a single introduction across varying doubling times, varying ascertainment rate, minimum polytomy size, and phylogenetic rooting method.

Analysis			Topology			Bayes factor	
DT	Asc	Min. polytomy size	C/C	A/B	Polytomy	Unconstrained	recCA
2.65	0.15	100	0.0	3.4	58.6	5.8	6.0
3.47[1]	0.15	100	0.0	3.1	47.5	4.2	4.3
4.45	0.15	100	0.1	3.5	43.1	3.0	3.0
3.50	0.05	100	0.0	3.4	45.7	3.5	3.6
3.52	0.25	100	0.2	3.5	47.3	3.6	3.7
3.47	0.15	20	0.1	5.1	60.7	4.1	4.2
3.47	0.15	50	0.1	3.9	53.6	4.2	4.3
3.47	0.15	200	0.0	2.1	40.7	4.5	4.6
3.47	0.15	500	0.0	1.1	31.7	5.2	5.4

DT, Median doubling time
Asc, Ascertainment rate
Min., Minimum
[1]Primary analysis

Table B.6. Number of days the timing of the primary case of lineage A occurs after the timing of the primary case of lineage B.

Phylodynamic analysis	Primary analysis[1,2]	Robustness analysis[1]			
	DT: 3.47 Asc: 15%	**DT: 2.65 Asc: 15%**	**DT: 4.45 Asc: 15%**	**DT: 3.50 Asc: 5%**	**DT: 3.52 Asc: 25%**
Unconstrained	6.7 (-28.9 to 44.4)	7.2 (-45.3 to 59.0)	7.0 (-20.5 to 35.1)	6.5 (-53.1 to 61.3)	7.0 (-36.4 to 48.1)
recCA	6.6 (-30.4 to 43.5)	6.6 (-44.9 to 58.9)	5.9 (-22.8 to 33.7)	6.2 (-48.7 to 63.4)	6.6 (-34.5 to 48.8)
Lineage B	6.6 (-28.4 to 44.4)	7.2 (-43.6 to 60.3)	7.2 (-20.3 to 34.9)	6.6 (-52.9 to 61.5)	6.9 (-35.9 to 48.9)
C/C	6.9 (-30.3 to 43.0)	7.2 (-45.5 to 59.4)	6.3 (-21.6 to 34.3)	6.5 (-47.5 to 65.8)	7.2 (-33.4 to 50.6)
Lineage A	6.9 (-30.2 to 43.5)	7.1 (-44.2 to 59.3)	5.7 (-22.7 to 33.8)	6.1 (-50.6 to 62.3)	6.9 (-35.7 to 49.2)

DT, Median doubling time

Asc, Ascertainment rate

[1]Median and 95% HPD in parentheses.

[2]See Fig. 2.3D for graphical representation of the full distribution.

Table B.7. Time of the lineage B primary case under different robustness analyses for different lineage B index case and hospitalization dates and conditioning only on hospitalization dates.

Phylodynamic analysis	Primary analysis[1]	Robustness analysis[1]		
	Case: Dec 10 Hosp: Dec 16	**Case: Dec 8 Hosp: Dec 16**	**Case: Dec 13 Hosp: Dec 18**	**Hosp: Dec 16**
Unconstrained	11-18 (10-22 to 12-09)	11-18 (10-22 to 12-06)	11-21 (10-24 to 12-11)	11-20 (10-22 to 12-11)
recCA	11-18 (10-23 to 12-08)	11-17 (10-22 to 12-06)	11-21 (10-25 to 12-11)	11-20 (10-22 to 12-10)
Lineage B	11-19 (10-22 to 12-08)	11-17 (10-21 to 12-06)	11-21 (10-24 to 12-11)	11-20 (10-22 to 12-11)
C/C	11-18 (10-22 to 12-07)	11-17 (10-22 to 12-06)	11-21 (10-25 to 12-11)	11-20 (10-23 to 12-11)
Lineage A	11-18 (10-23 to 12-08)	11-17 (10-23 to 12-06)	11-21 (10-24 to 12-11)	11-20 (10-23 to 12-10)

Hosp, Hospitalization

[1]Median and 95% HPD in parentheses.

Table B.8. Time of the lineage B primary case under different robustness analyses for different doubling times and ascertainment rates.

Phylodynamic analysis	Primary analysis[1]	Robustness analysis[1]			
	DT: 3.47 Asc: 15%	**DT: 2.65 Asc: 15%**	**DT: 4.45 Asc: 15%**	**DT: 3.5 Asc: 5%**	**DT: 3.52 Asc: 25%**
Unconstrained	11-18 (10-22 to 12-09)	11-25 (11-04 to 12-08)	11-14 (10-06 to 12-08)	11-17 (10-07 to 12-07)	11-21 (10-21 to 12-09)
recCA	11-18 (10-23 to 12-08)	11-26 (11-05 to 12-08)	11-14 (10-05 to 12-07)	11-17 (10-06 to 12-08)	11-21 (10-21 to 12-09)
Lineage B	11-19 (10-22 to 12-08)	11-25 (11-04 to 12-08)	11-14 (10-06 to 12-08)	11-17 (10-08 to 12-08)	11-21 (10-21 to 12-09)
C/C	11-18 (10-22 to 12-07)	11-26 (11-05 to 12-08)	11-14 (10-05 to 12-07)	11-17 (10-06 to 12-08)	11-21 (10-21 to 12-09)
Lineage A	11-18 (10-23 to 12-08)	11-26 (11-05 to 12-08)	11-14 (10-06 to 12-07)	11-17 (10-06 to 12-09)	11-21 (10-21 to 12-09)

DT, Median doubling time
Asc, Ascertainment rate
[1]Median and 95% HPD in parentheses.

Table B.9. Time of the lineage A primary case results when conditioning with either both the index and hospitalization dates or just the hospitalization date.

Phylodynamic analysis	Primary analysis[1]	Robustness analysis[1]
	Case: Dec 15 Hosp: Dec 25	**Hosp: Dec 25**
Unconstrained	11-25 (10-31 to 12-13)	11-28 (10-31 to 12-19)
recCA	11-25 (10-29 to 12-14)	11-29 (10-30 to 12-19)
Lineage B	11-25 (11-01 to 12-13)	11-28 (11-01 to 12-19)
C/C	11-25 (10-30 to 12-13)	11-29 (10-31 to 12-20)
Lineage A	11-25 (10-29 to 12-14)	11-29 (10-30 to 12-19)

Hosp, Hospitalization
[1]Median and 95% HPD in parentheses.

Table B.10. Time of the lineage A primary case under different doubling times and ascertainment rates.

Phylodynamic analysis	Primary analysis[1]	Robustness analysis[1]			
	DT: 3.47 Asc: 15%	**DT: 2.65 Asc: 15%**	**DT: 4.45 Asc: 15%**	**DT: 3.5 Asc: 5%**	**DT: 3.52 Asc: 25%**
Unconstrained	11-25 (10-31 to 12-13)	12-02 (11-12 to 12-13)	11-21 (10-13 to 12-13)	11-24 (10-12 to 12-14)	11-28 (10-29 to 12-14)
recCA	11-25 (10-29 to 12-14)	12-02 (11-10 to 12-14)	11-21 (10-12 to 12-13)	11-23 (10-13 to 12-13)	11-28 (10-29 to 12-14)
Lineage B	11-25 (11-01 to 12-13)	12-02 (11-13 to 12-14)	11-21 (10-14 to 12-13)	11-24 (10-11 to 12-14)	11-28 (10-29 to 12-14)
C/C	11-25 (10-30 to 12-13)	12-02 (11-11 to 12-14)	11-21 (10-13 to 12-13)	11-24 (10-13 to 12-12)	11-28 (10-29 to 12-14)
Lineage A	11-25 (10-29 to 12-14)	12-01 (11-09 to 12-14)	11-21 (10-13 to 12-13)	11-23 (10-14 to 12-12)	11-28 (10-29 to 12-14)

DT, Median doubling time
Asc, Ascertainment rate
[1]Median and 95% HPD in parentheses.

Table B.11. Number of estimated infections on 1 December 2019 for lineage A, lineage B, lineages A and B combined, and single introduction simulations.

Phylodynamic analysis	Introductions[1]			
	Lineage A	**Lineage B**	**Lineages A and B combined**	**Single introduction**
Unconstrained	2 (0, 23)	4 (0, 27)	8 (0, 38)	5 (0, 31)
recCA	2 (0, 24)	4 (0, 25)	8 (0, 36)	6 (0, 85)
Lineage B	2 (0, 23)	4 (0, 27)	8 (0, 38)	5 (0, 29)
C/C	2 (0, 23)	4 (0, 25)	8 (0, 35)	5 (0, 51)
Lineage A	2 (0, 23)	4 (0, 25)	8 (0, 35)	6 (0, 58)

[1]Median and 95% HPD in parentheses.

Table B.12. Number of estimated infections on 15 December 2019 for lineage A, lineage B, lineages A and B combined, and single introduction simulations.

Phylodynamic analysis	Introductions[1]			
	Lineage A	**Lineage B**	**Lineages A and B combined**	**Single introduction**
Unconstrained	8 (1, 33)	16 (2, 251)	28 (3, 273)	17 (2, 517)
recCA	9 (1, 57)	14 (2, 63)	26 (3, 153)	31 (2, 3094)
Lineage B	8 (1, 33)	16 (2, 246)	28 (3, 259)	16 (2, 303)
C/C	8 (1, 35)	14 (2, 62)	25 (3, 81)	24 (2, 1640)
Lineage A	9 (1, 59)	13 (2, 52)	26 (3, 117)	26 (2, 2095)

[1]Median and 95% HPD in parentheses.

Table B.13. Number of infections at the tMRCA and hospitalizations on 1 December 2019.

	Time	**Primary analysis**[1]	**Robustness analysis**[1]			
		DT: 3.47 Asc: 15%	**DT: 2.65 Asc: 15%**	**DT: 4.45 Asc: 15%**	**DT: 3.50 Asc: 5%**	**DT: 3.52 Asc: 25%**
Infections	**tMRCA**	3 (1-18)	3 (1-12)	5 (1-39)	5 (1-29)	3 (1-18)
Hospitalizations	**1 Dec 2019**	0 (0–2)	0 (0–1)	1 (0–3)	0 (0–2)	0 (0–3)

DT, Median doubling time

Asc, Ascertainment rate

[1]Median and 95% HPD in parentheses.

Table B.14. Time of the primary case of a hypothetical single-introduction scenario under different robustness analyses for different index case dates and conditioning on only hospitalization dates.

Phylodynamic analysis	**Primary analysis**[1]	**Robustness analysis**[1]	
	Case: Dec 10 Hosp: Dec 16	**Case: Dec 8 Hosp: Dec 16**	**Hosp: Dec 16**
Unconstrained	11-18 (10-22 to 12-08)	11-18 (10-21 to 12-06)	11-20 (10-21 to 12-11)
recCA	11-18 (10-20 to 12-08)	11-17 (10-19 to 12-06)	11-19 (10-19 to 12-11)
Hu-1	11-19 (10-23 to 12-09)	11-18 (10-22 to 12-06)	11-20 (10-22 to 12-11)
C/C	11-18 (10-20 to 12-08)	11-17 (10-20 to 12-06)	11-20 (10-20 to 12-11)
T/T	11-18 (10-20 to 12-08)	11-17 (10-20 to 12-07)	11-20 (10-20 to 12-11)
Lineage A	11-18 (10-20 to 12-08)	11-17 (10-19 to 12-06)	11-19 (10-20 to 12-11)
Lineage A + C29095T	11-18 (10-19 to 12-08)	11-17 (10-19 to 12-07)	11-19 (10-19 to 12-10)
Lineage A.1	11-18 (10-19 to 12-08)	11-17 (10-19 to 12-07)	11-19 (10-18 to 12-09)

Hosp, Hospitalization

[1]Median and 95% HPD in parentheses.

Table B.15. Time of the primary case of a hypothetical single-introduction scenario under different doubling times and ascertainment rates.

Phylodynamic analysis	**Primary analysis**[1]	**Robustness analysis**[1]			
	DT: 3.47 Asc: 15%	**DT: 2.65 Asc: 15%**	**DT: 4.45 Asc: 15%**	**DT: 3.5 Asc: 5%**	**DT: 3.52 Asc: 25%**
Unconstrained	11-18 (10-22 to 12-08)	11-25 (11-03 to 12-08)	11-14 (10-05 to 12-08)	11-17 (10-08 to 12-07)	11-21 (10-21 to 12-09)
recCA	11-18 (10-20 to 12-08)	11-23 (10-30 to 12-08)	11-13 (10-03 to 12-06)	11-15 (10-09 to 12-07)	11-19 (10-20 to 12-08)
Lineage B	11-19 (10-23 to 12-09)	11-25 (11-03 to 12-08)	11-14 (10-05 to 12-08)	11-17 (10-08 to 12-08)	11-21 (10-21 to 12-09)
C/C	11-18 (10-20 to 12-08)	11-24 (11-01 to 12-08)	11-13 (10-04 to 12-07)	11-16 (10-09 to 12-07)	11-20 (10-20 to 12-09)
T/T	11-18 (10-20 to 12-08)	11-24 (11-01 to 12-08)	11-13 (10-03 to 12-07)	11-16 (10-10 to 12-07)	11-20 (10-20 to 12-09)
Lineage A	11-18 (10-20 to 12-08)	11-24 (10-31 to 12-08)	11-13 (10-04 to 12-07)	11-16 (10-11 to 12-07)	11-20 (10-20 to 12-08)
Lineage A + C29095T	11-18 (10-19 to 12-08)	11-23 (10-31 to 12-08)	11-13 (10-03 to 12-06)	11-16 (10-10 to 12-07)	11-19 (10-20 to 12-08)
Lineage A.1	11-18 (10-19 to 12-08)	11-23 (10-30 to 12-08)	11-13 (10-04 to 12-07)	11-16 (10-10 to 12-07)	11-19 (10-20 to 12-08)

DT, Median doubling time
Asc, Ascertainment rate
[1]Median and 95% HPD in parentheses.

B.4 Supplementary Figures

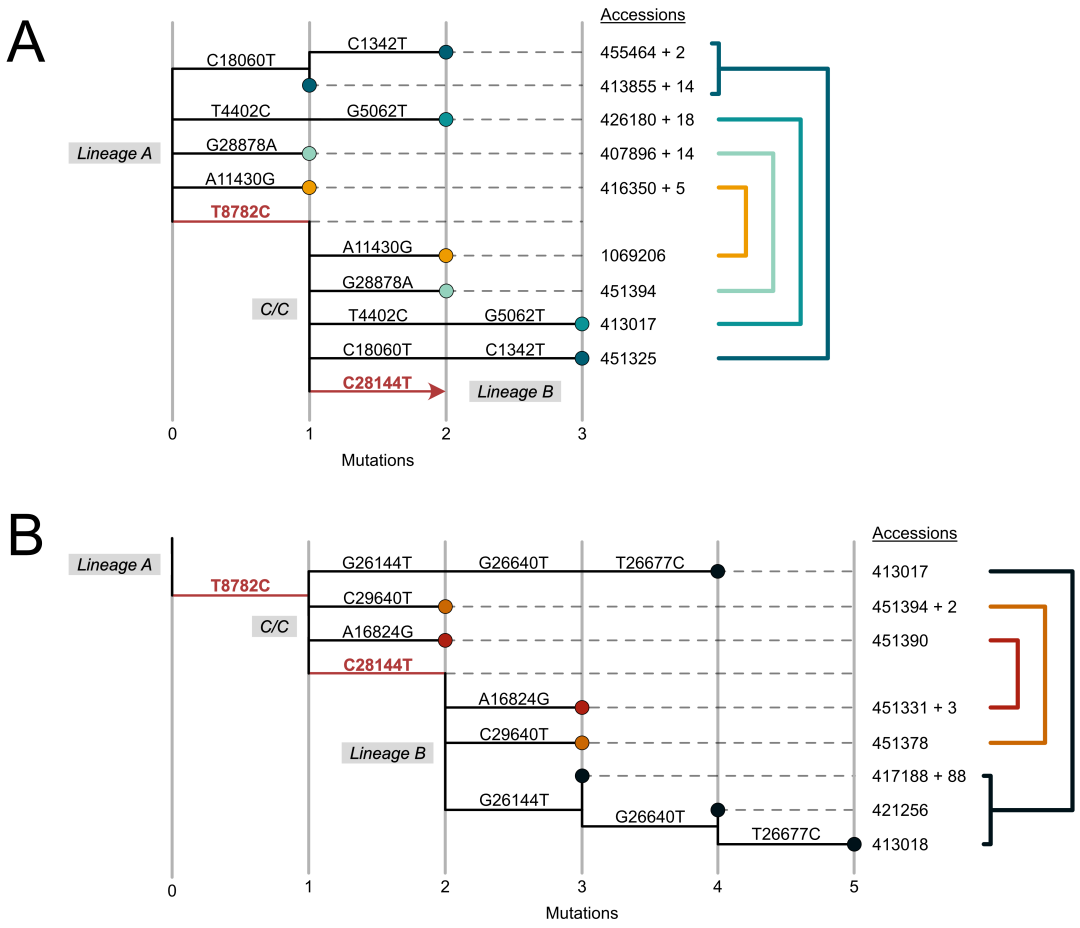


Figure B.1. Mutation map of SARS-CoV-2 intermediate C/C genomes and their shared mutations within lineages A and B. (A) Shared mutations across lineage A and C/C. (B) Shared mutations across lineage B and C/C. Mutations relative to the Hu-1 reference genome are shown above each branch. Lineage-defining mutations (8782 and 28144) are colored in red. Derived mutations not shared by both lineages are excluded. The taxon names are GISAID accession numbers, and the total number of additional matching homoplastic sequences are indicated. Sequences that share derived mutations are connected by the lines on the right, and brackets indicate that a group of sequences share the derived mutations that cannot be individually resolved.

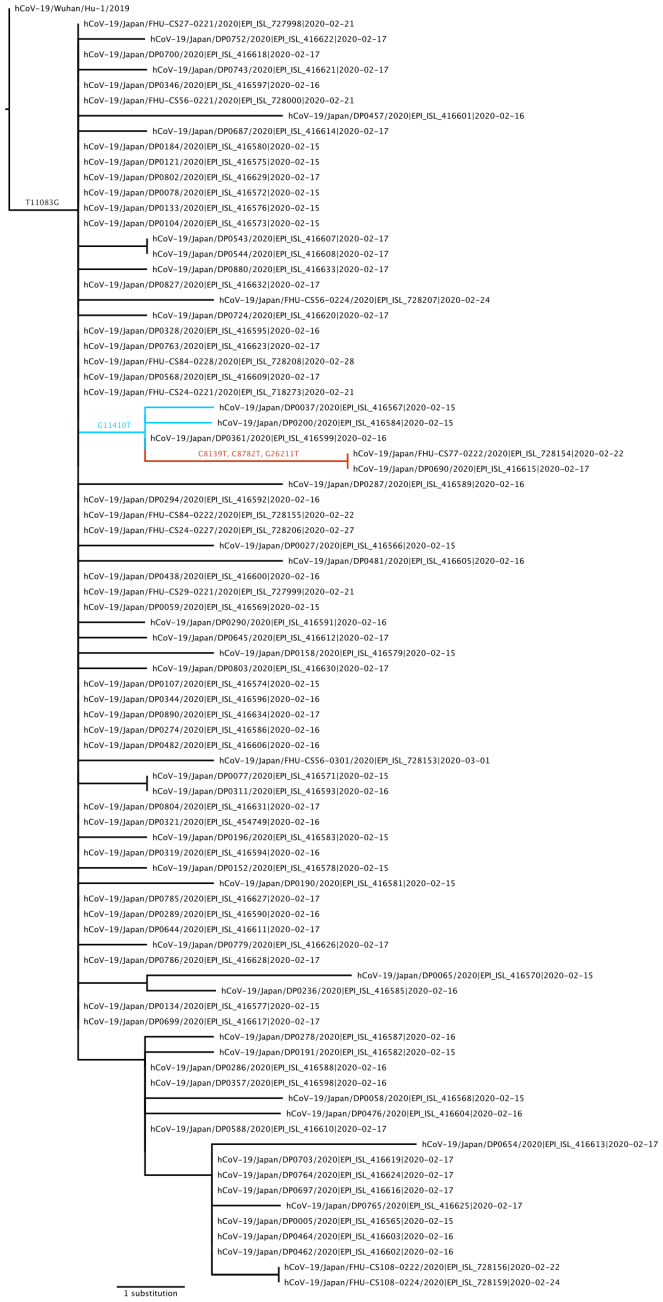


Figure B.2. Maximum likelihood phylogeny of SARS-CoV-2 genomes from the Diamond Princess outbreak. The tree is rooted on Hu-1. Substitutions found in T/T genomes relative to Hu-1 annotated on branches. The G11410T clade is colored blue, with the branch leading to the T/T genomes colored red.

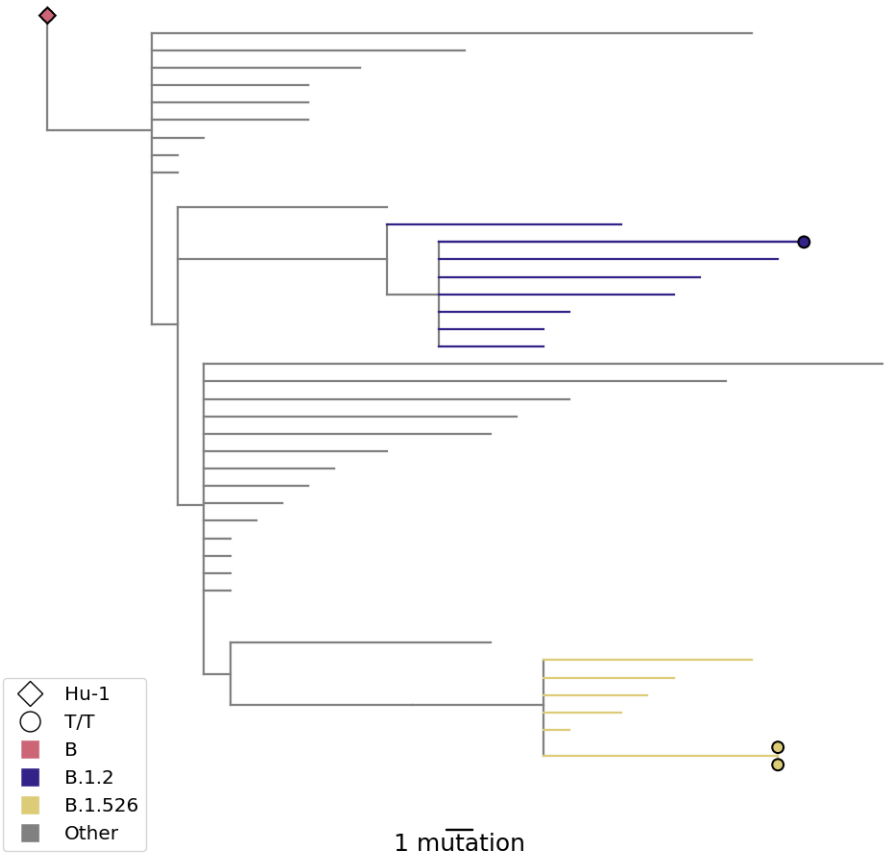


Figure B.3. Subtree showing the placement of T/T genomes in the NYC Public Health Laboratories dataset. 3 T/T genomes were placed on a global tree of 3 million genomes (v2022-01-21) using UShER. The node branches are colored by the assigned PANGO lineage. The T/T genomes are highlighted using circles.

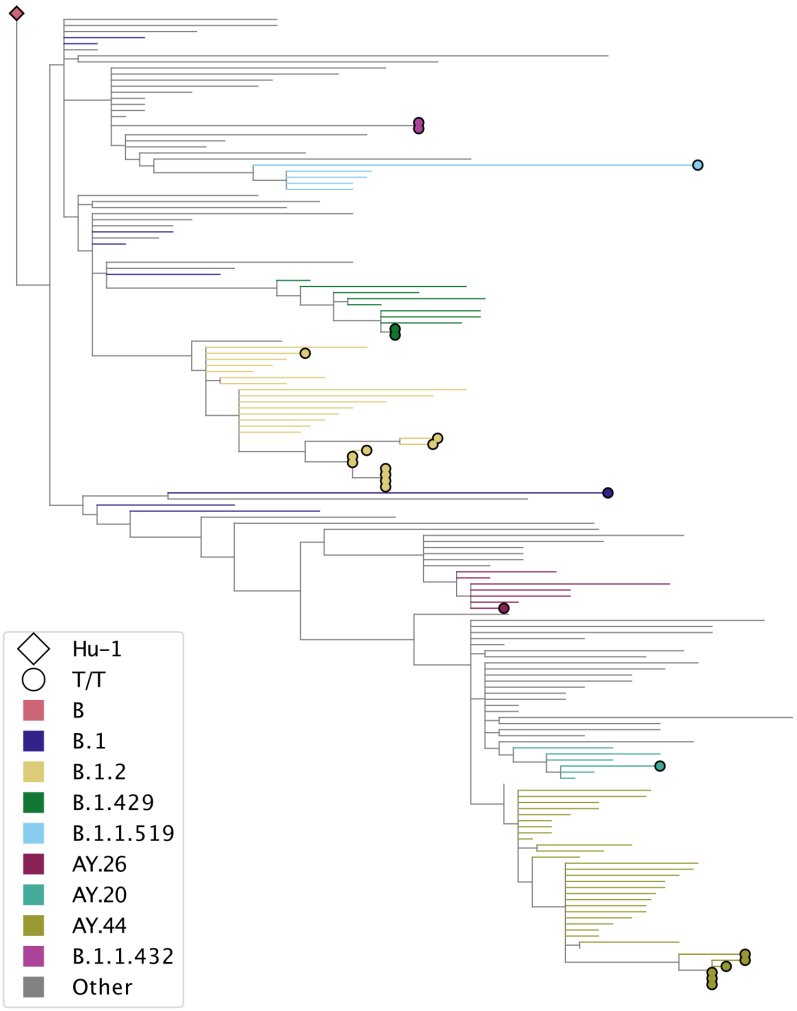


Figure B.4. Subtree showing the placement of T/T genomes in the San Diego SEARCH dataset. 24 T/T genomes were placed on a global tree of 3 million genomes (v2022-01-21) using UShER. The node branches are colored by the assigned PANGO lineage. The T/T genomes are highlighted using circles.

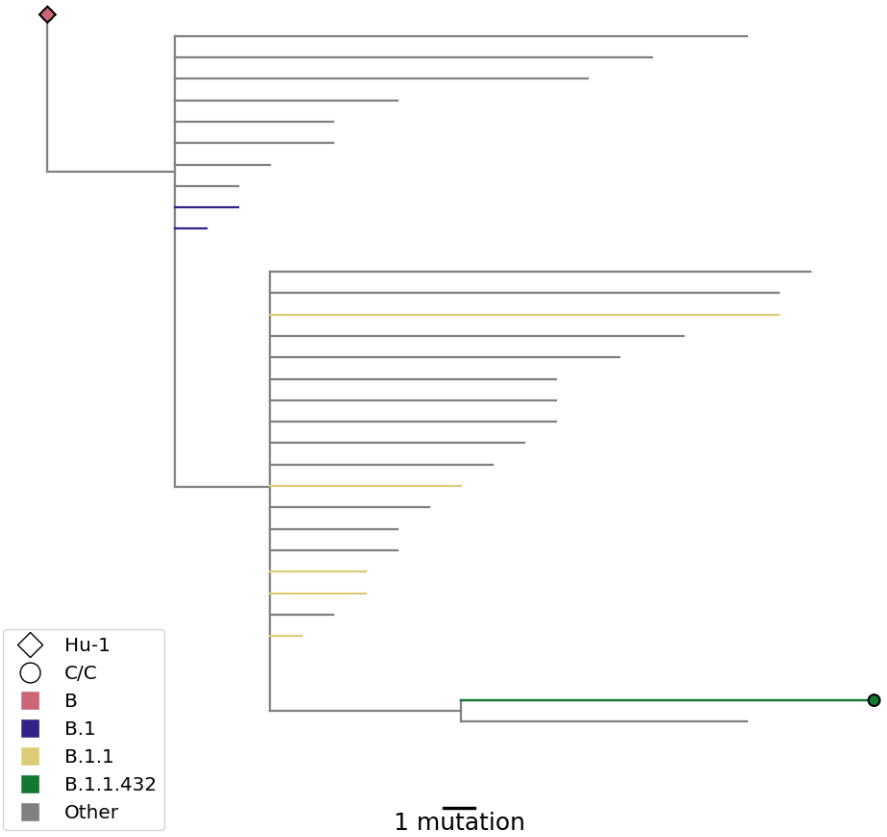


Figure B.5. Subtree showing the placement of C/C genome in the San Diego SEARCH dataset. 1 C/C genome was placed on a global tree of 3 million genomes (v2022-01-21) using UShER. The node branches are colored by the assigned PANGO lineage. The C/C genome is highlighted using a circle.

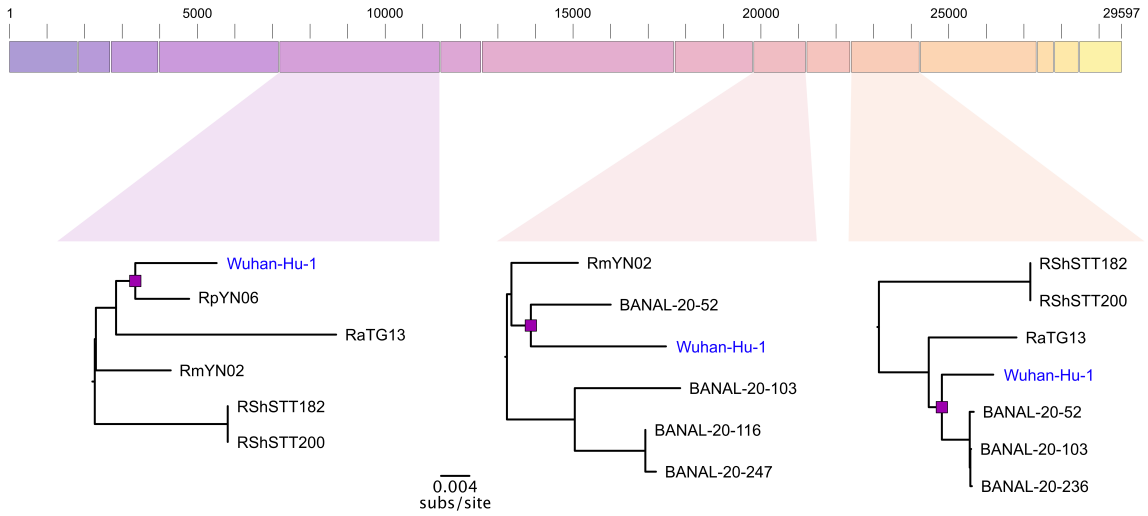


Figure B.6. Reconstructing the recombinant common ancestor (recCA) of SARS-CoV-2 infecting a non-human animal. The figure identifies 15 non-recombinant regions of SARS-CoV-2-like sarbecovirus genomes. Subtrees from sarbecovirus maximum likelihood phylogenies of example regions 5, 9, and 11 show the genomes most closely related to SARS-CoV-2. Ancestral state reconstruction at the MRCA (purple square) of SARS-CoV-2 (Wuhan-Hu-1) and the most closely related sarbecovirus for each of the 15 fragments is concatenated to construct the recCA.

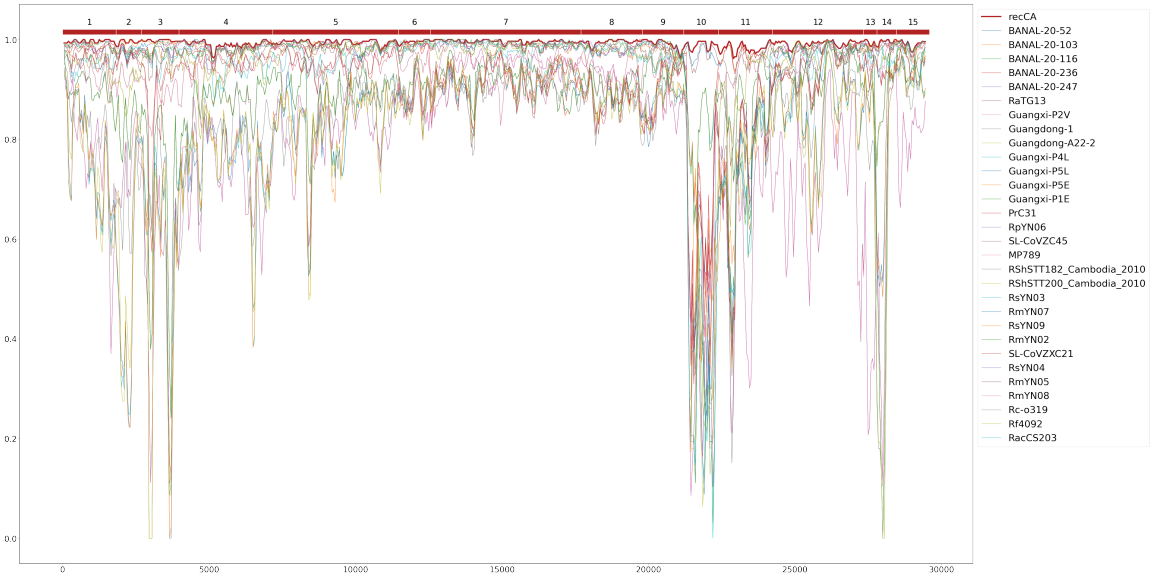


Figure B.7. Simplot of closely related sarbecoviruses and recCA with Hu-1 as the reference.

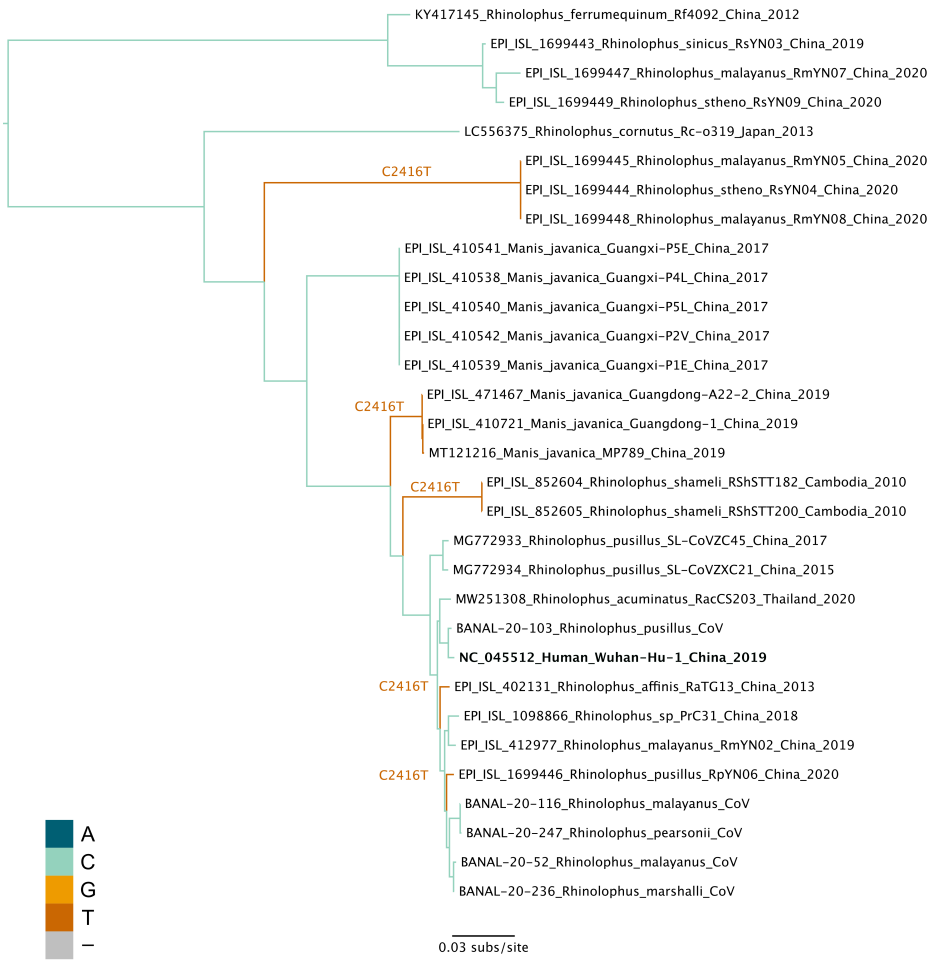


Figure B.8. Maximum likelihood tree of non-recombinant region 2 with branches colored based on the nucleotide at position 2416. Some substitution labels shifted for clarity.

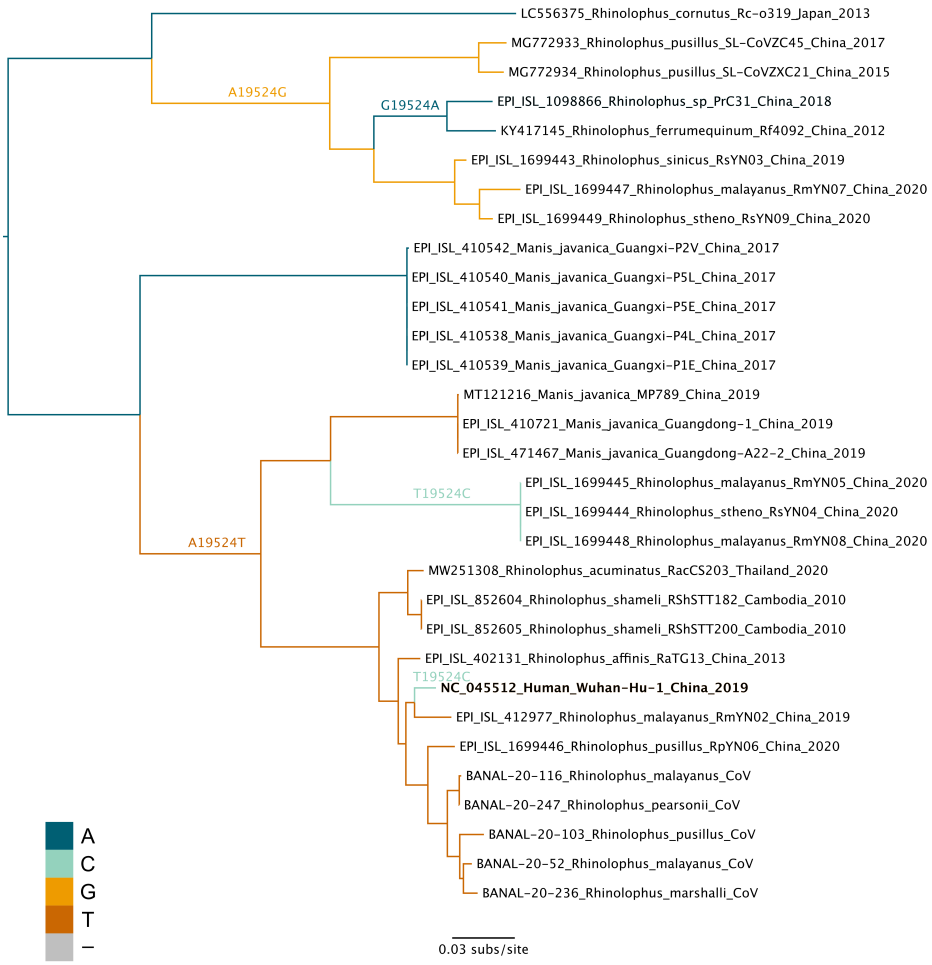


Figure B.9. Maximum likelihood tree of non-recombinant region 8 with branches colored based on the nucleotide at position 19524. Some substitution labels shifted for clarity.

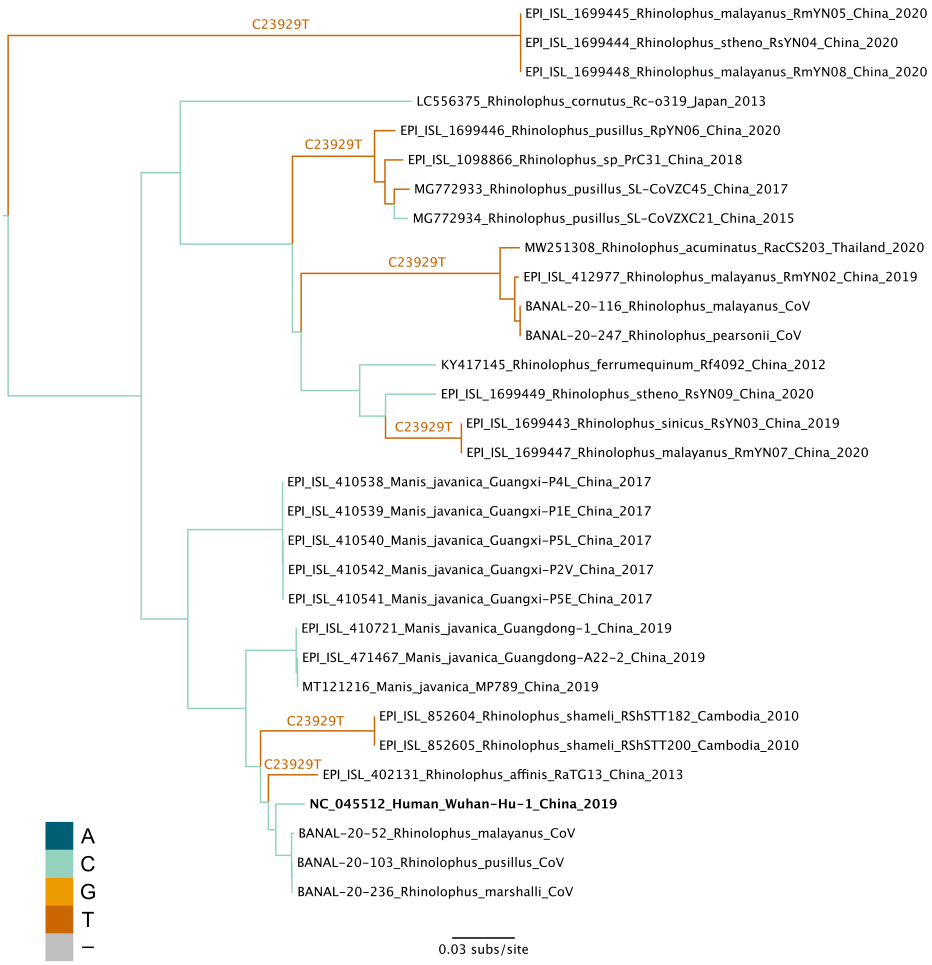


Figure B.10. Maximum likelihood tree of non-recombinant region 11 with branches colored based on the nucleotide at position 23929. Some substitution labels shifted for clarity.

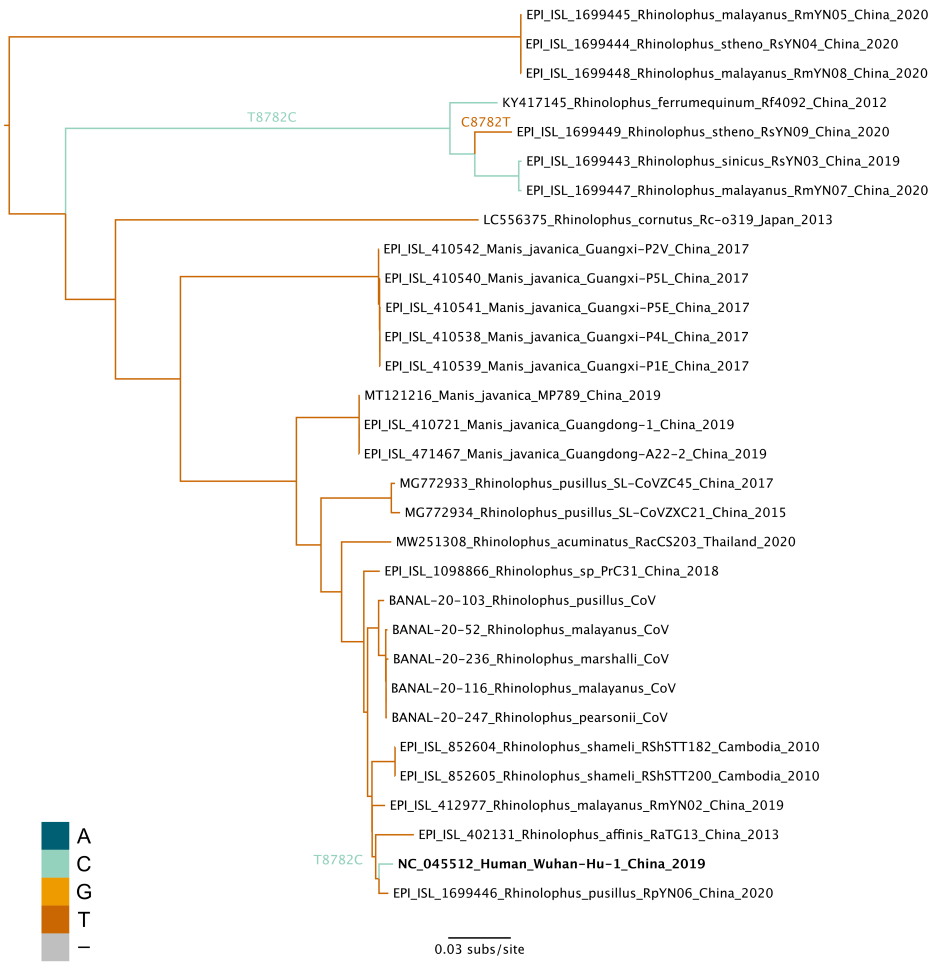


Figure B.11. Maximum likelihood tree of non-recombinant region 5 with branches colored based on the nucleotide at position 8782. Some substitution labels shifted for clarity.

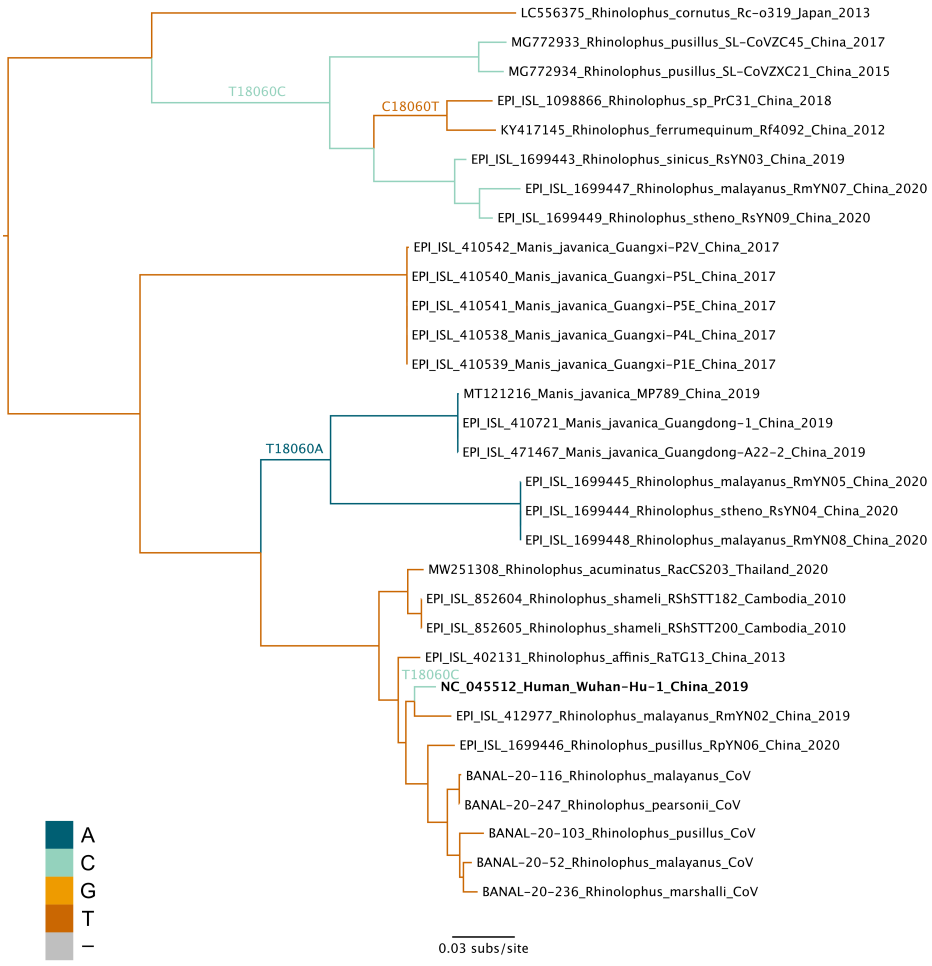


Figure B.12. Maximum likelihood tree of non-recombinant region 8 with branches colored based on the nucleotide at position 18060.

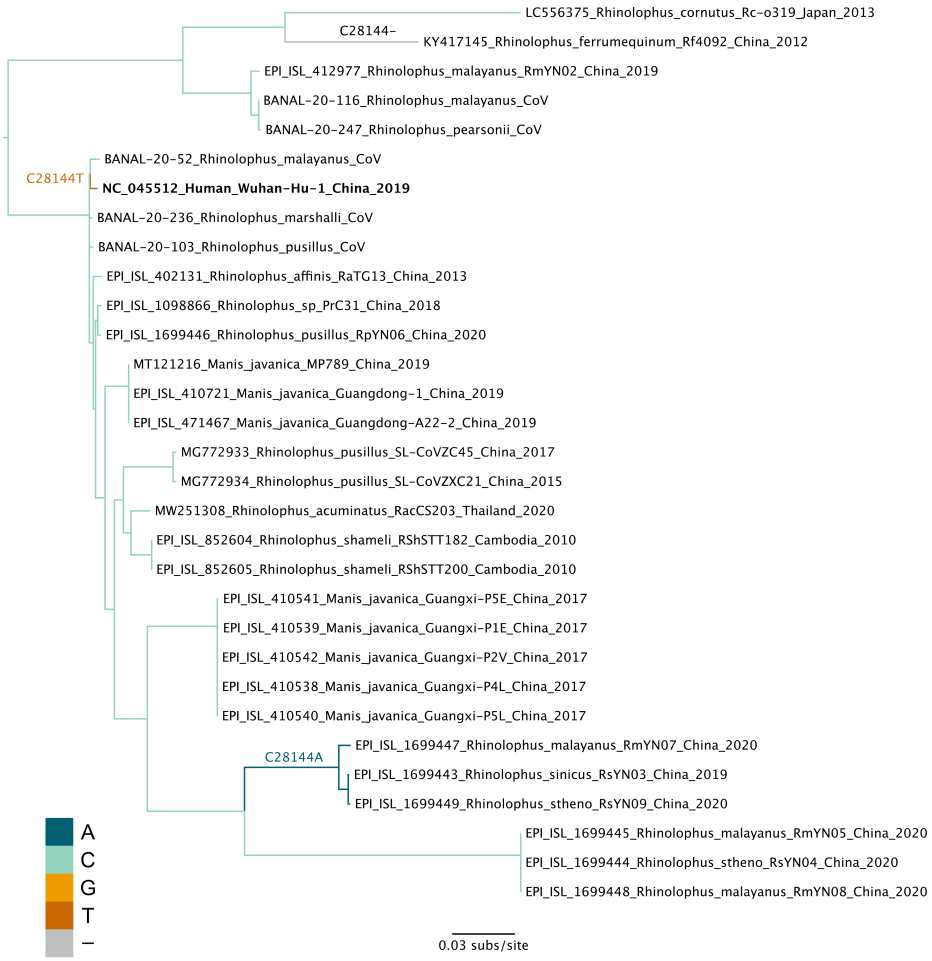


Figure B.13. Maximum likelihood tree of non-recombinant region 14 with branches colored based on the nucleotide at position 28144. Some substitution labels shifted for clarity.

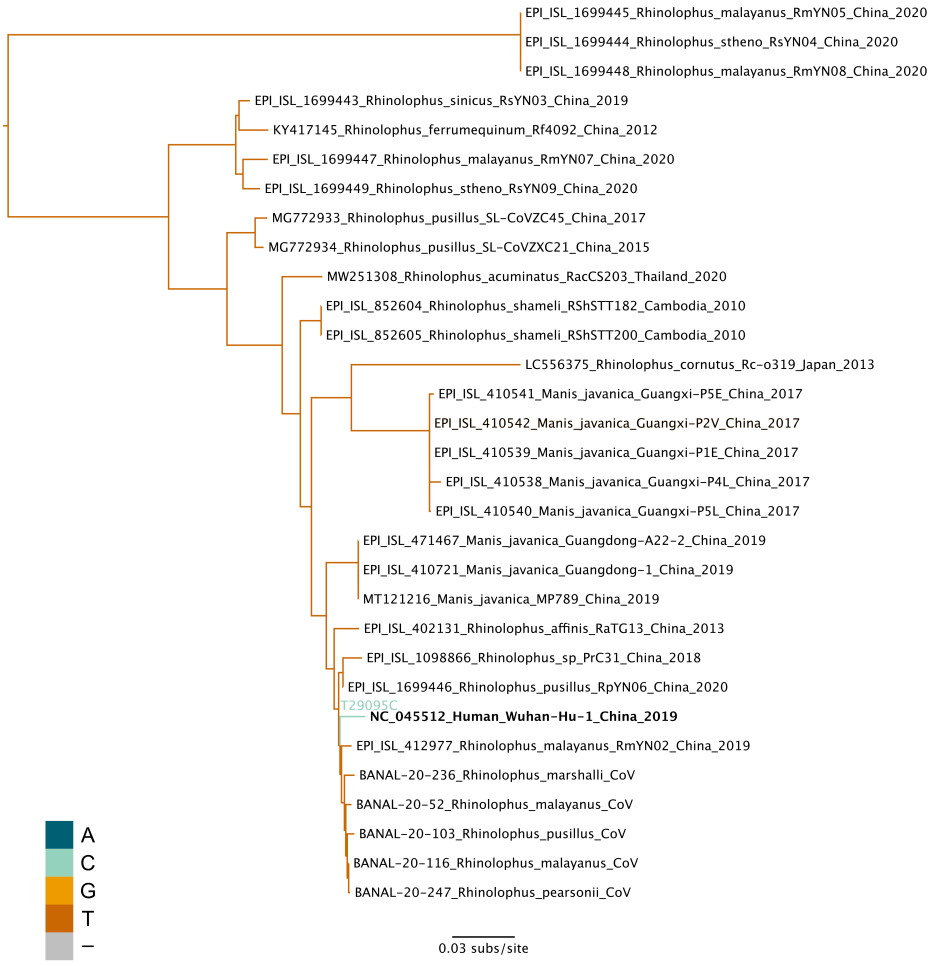


Figure B.14. Maximum likelihood tree of non-recombinant region 15 with branches colored based on the nucleotide at position 29095.

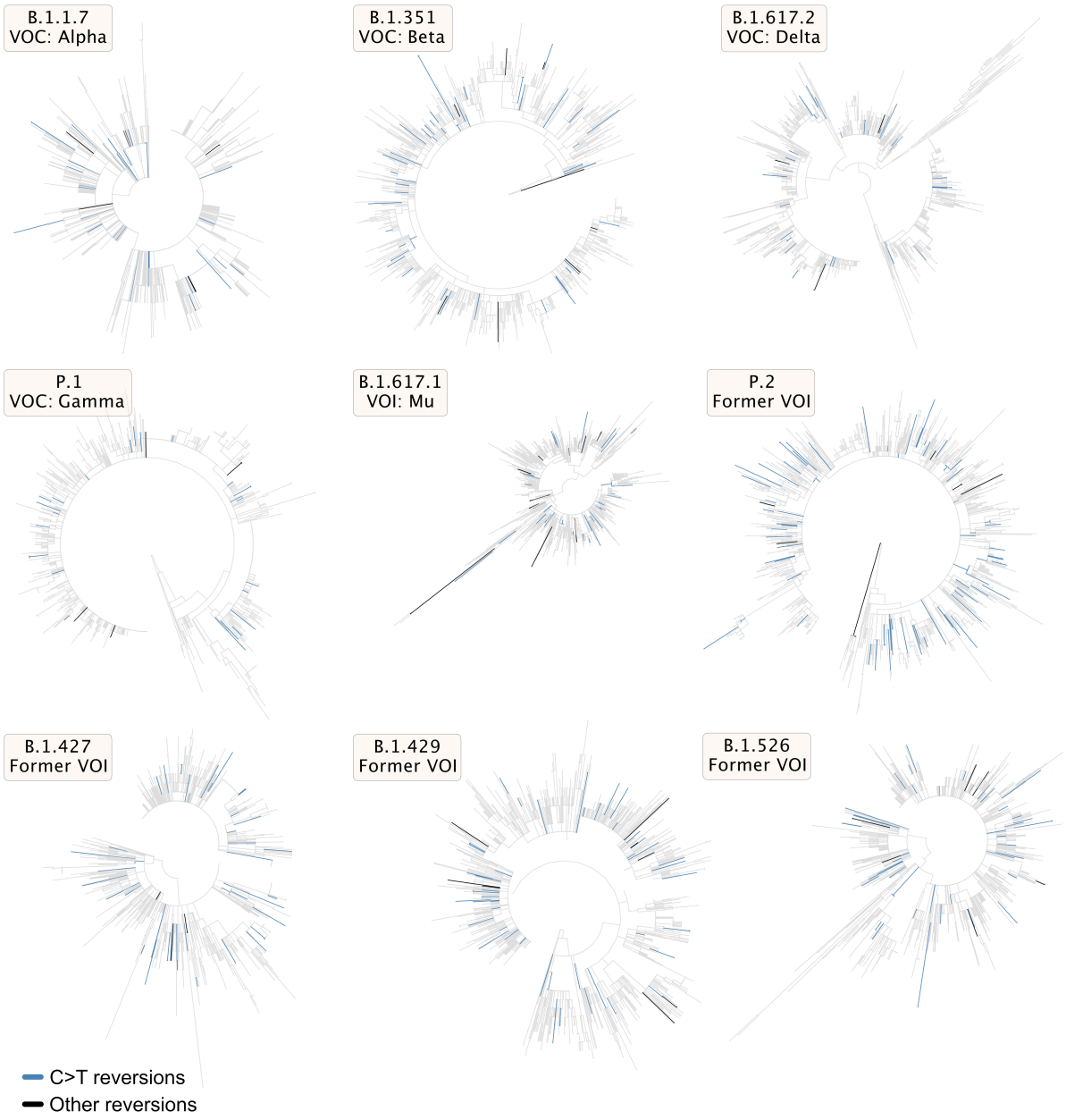


Figure B.15. Maximum likelihood phylogenies of variants of concern (VOC) and variants of interest (VOI) with branches containing reversions colored.

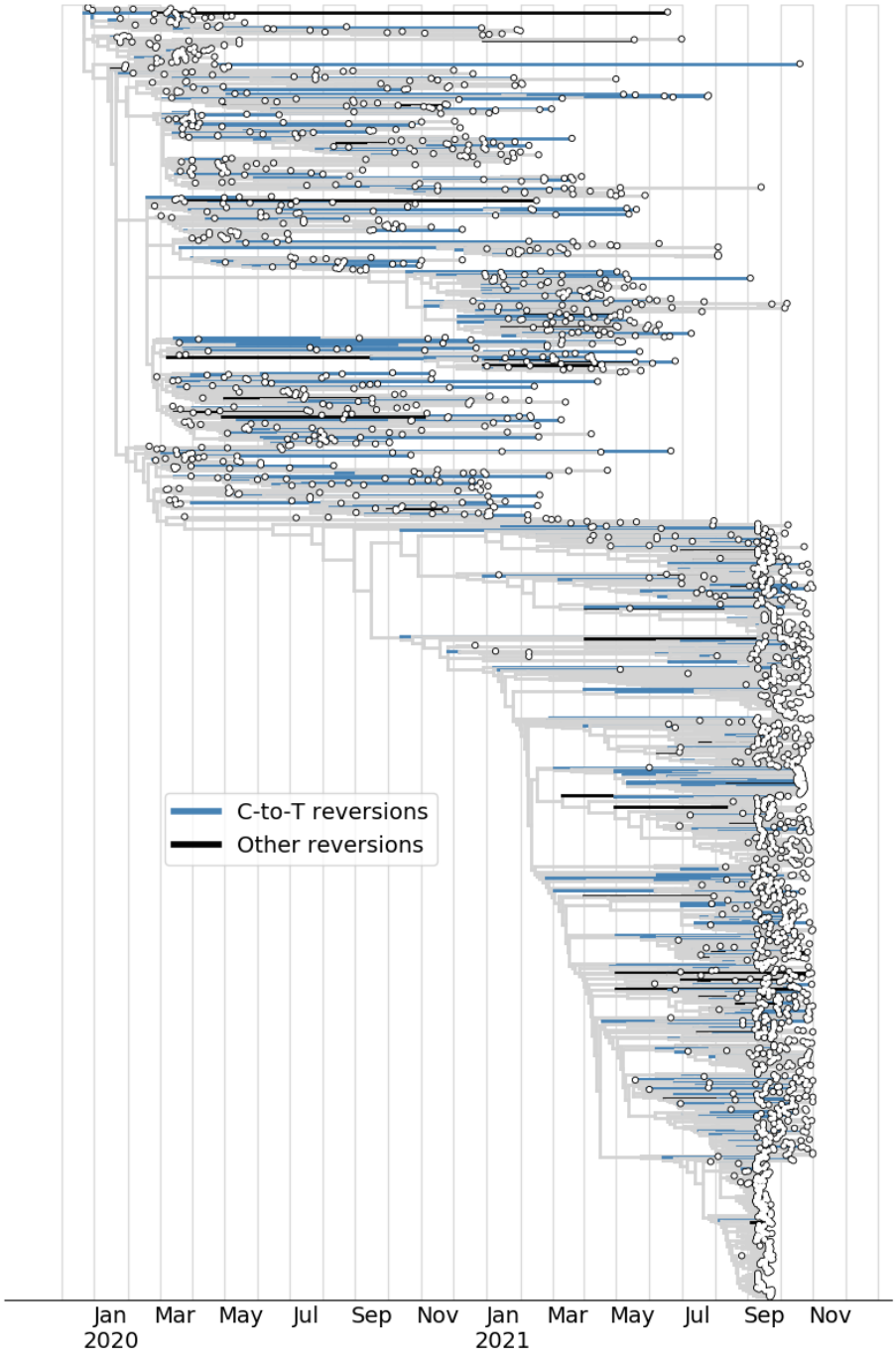


Figure B.16. Subsampled global phylogeny showing reversions. Subsampled SARS-CoV-2 time-resolved phylogeny from Nextstrain, with reversions colored blue if a C-to-T reversion and black otherwise.

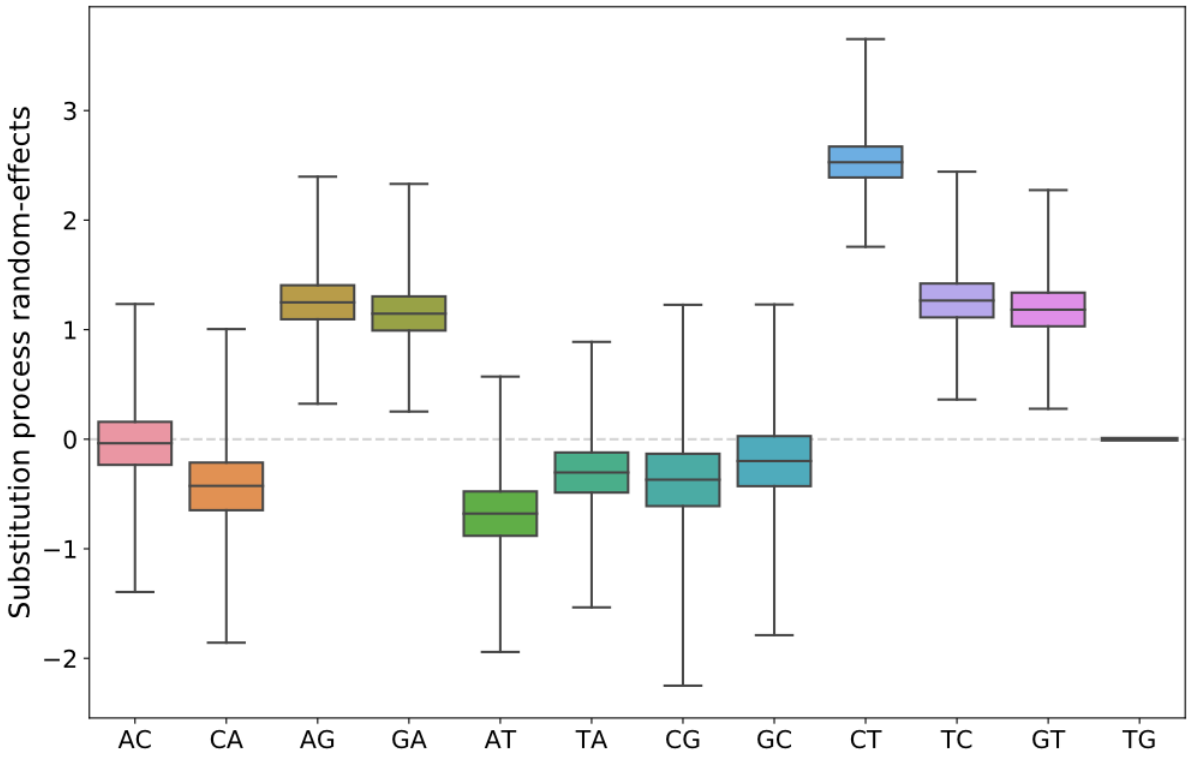


Figure B.17. Substitution process random-effects for the unconstrained rooting model. The random effects for transitions were rescaled with κ, and then all random effects were made relative to T-to-G (fixed to 0). The posterior probabilities that $e^{C-to-T} > e^{T-to-C}$ and $e^{G-to-T} > e^{T-to-G}$ is 1.00 for both, indicating the C-to-T transition and G-to-T transversion biases were present in every sample in the posterior.

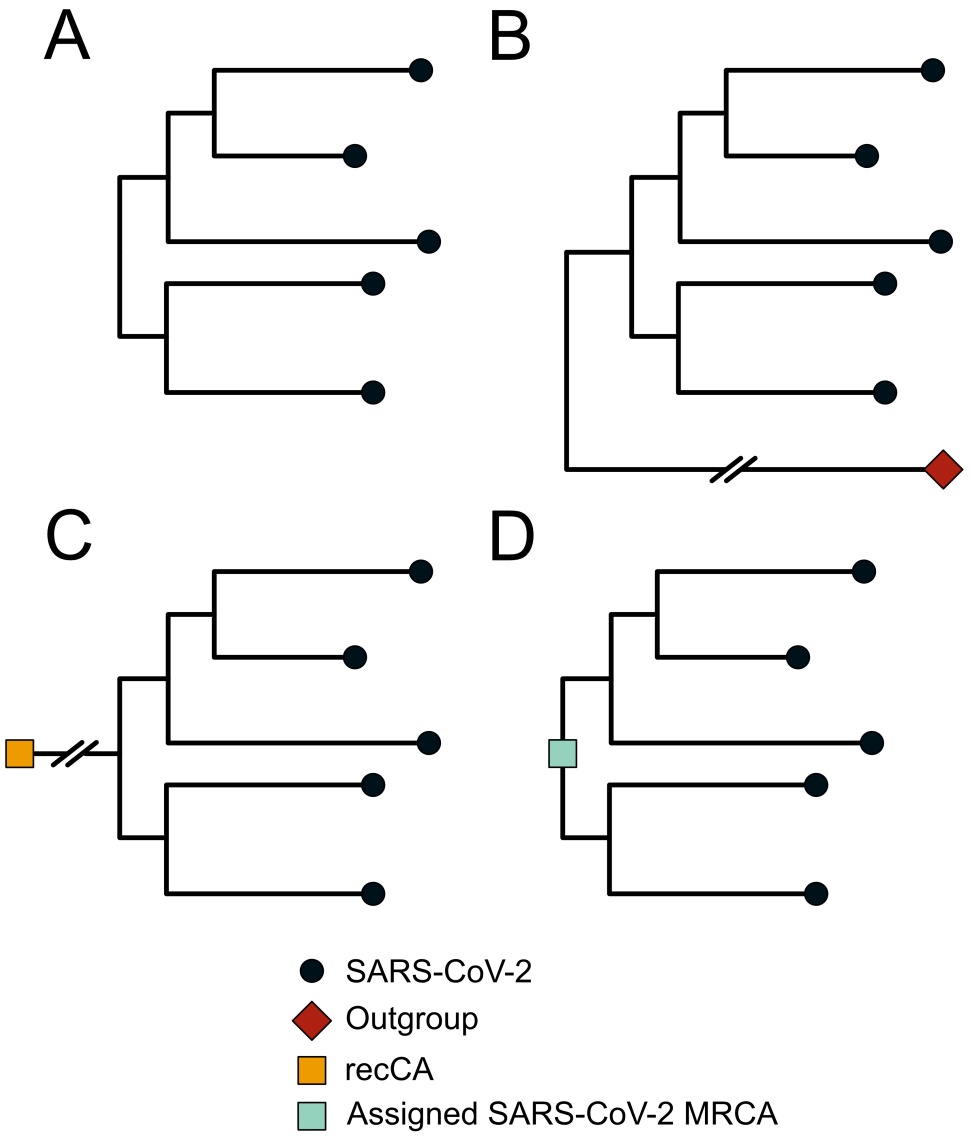


Figure B.18. Schematic depicting the rooting strategies used in different phylodynamic models. (A) An unconstrained rooting model with only SARS-CoV-2 where the root is inferred from the molecular clock calibrated using SARS-CoV-2 sampling dates. (B) An unconstrained rooting with SARS-CoV-2 and a sarbecovirus outgroup. (C) A constrained model where the ancestor of SARS-CoV-2 is constrained to be the recombinant common ancestor (recCA). (D) A constrained model with only SARS-CoV-2, but the MRCA forced to be a pre-specified haplotype.

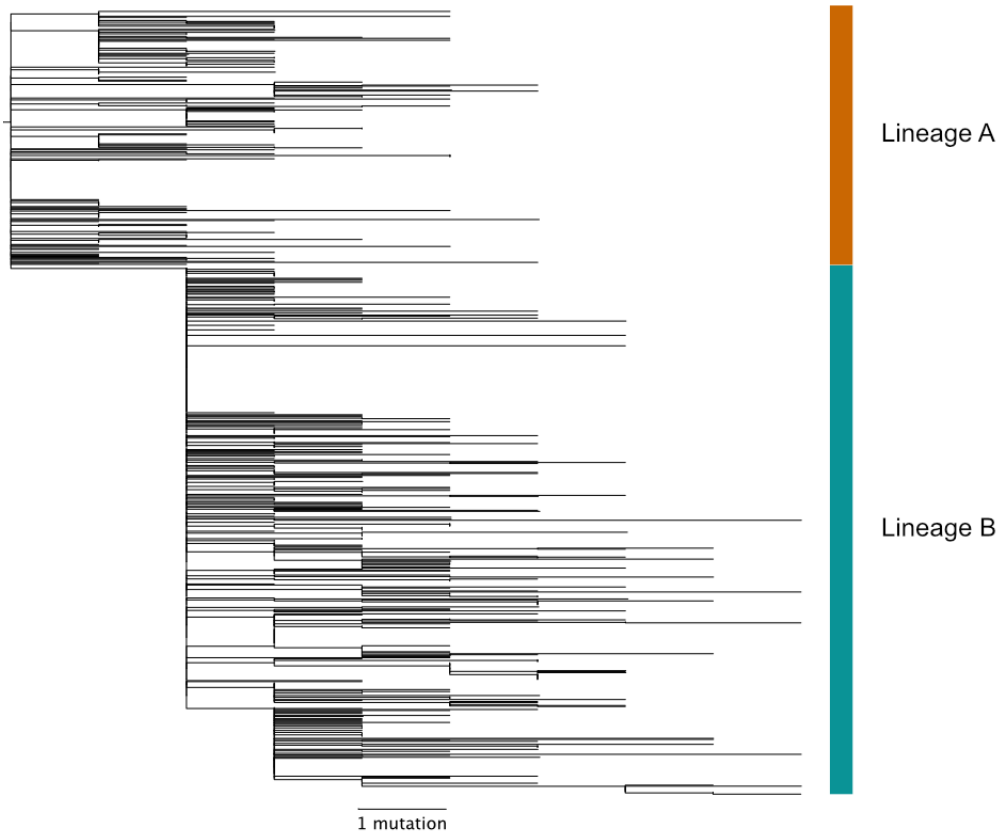


Figure B.19. SARS-CoV-2 maximum likelihood tree rooted on lineage A (n=787 taxa, through 14 February 2020).

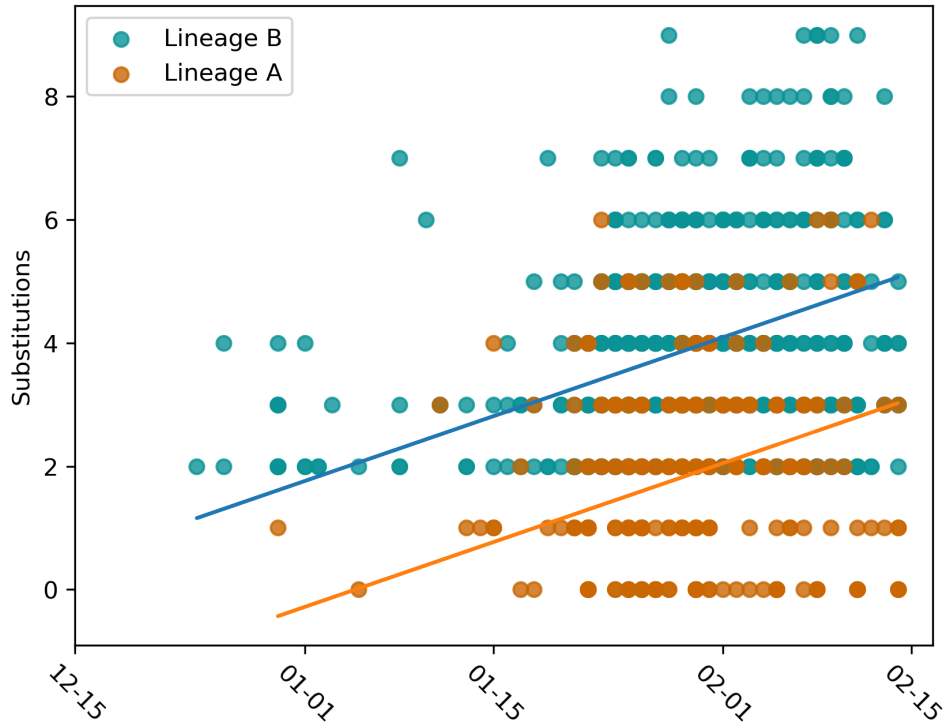


Figure B.20. Substitution counts of SARS-CoV-2 genomes through 14 February 2020 from the root of the maximum likelihood tree when rooted on lineage A (Fig. B.19). The plotted lines have a slope of 27.51 substitutions/year, are fit to their respective lineages, and are separated by 2.04 substitutions, showcasing the greater divergence of lineage B than lineage A when the tree is rooted on lineage A.

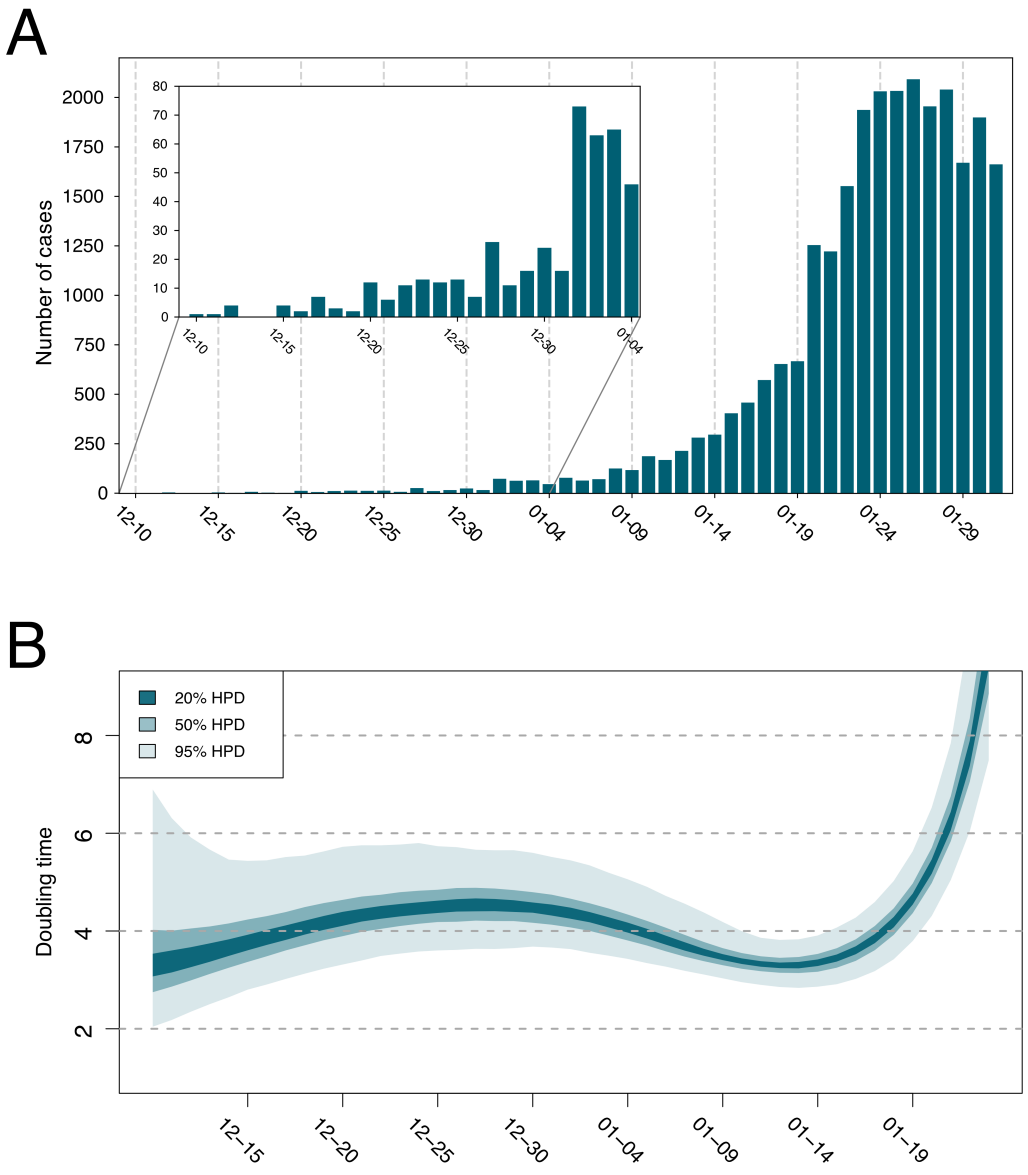


Figure B.21. Growth of the early pandemic. (A) Daily case count after combining the data from the WHO report[26] and Li *et al.*[23] through January 2020. Inset shows daily case count through 4 January 2020. See Methods for how data were combined. (B) Inferred doubling times of the pandemic in December 2019 and January 2020.

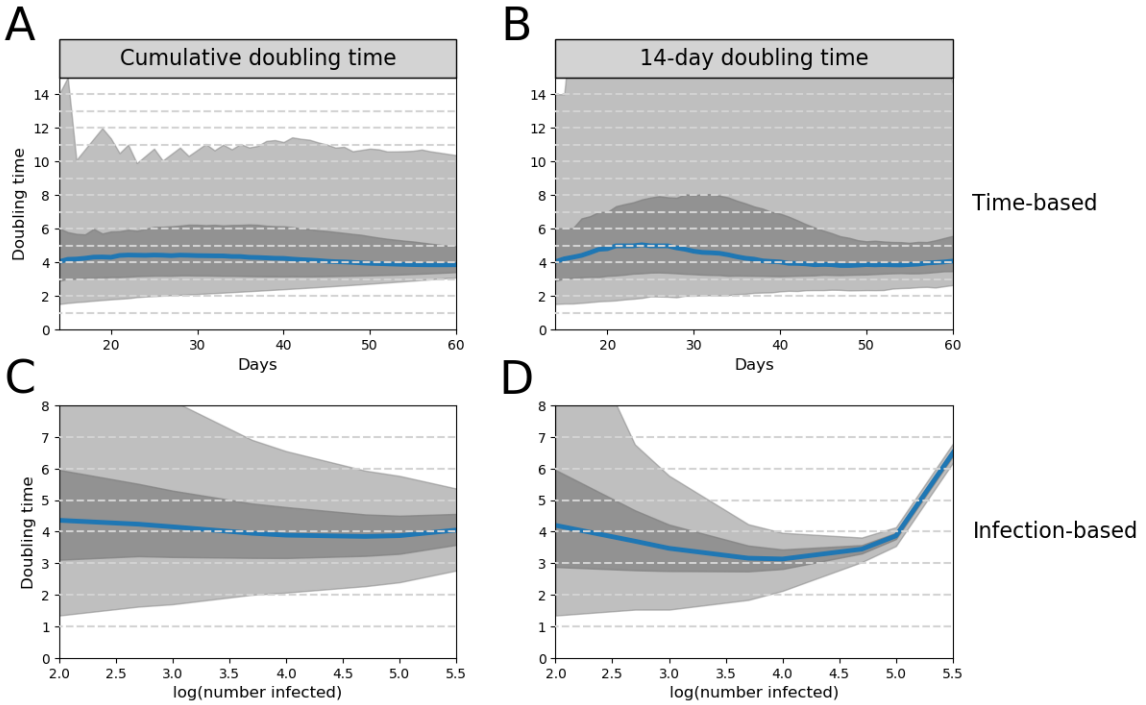


Figure B.22. Inferred doubling times of simulated epidemics. Inferred doubling times of the 1100 primary simulations. (A) Cumulative doubling time since the start of the simulation. (B) 14-day doubling time from day 14 until the end of the simulation, with cumulative doubling time reported prior to day 14. (C) cumulative doubling time once a certain number of individuals are infected (e.g., the cumulative doubling time at the 100th infection). (D) 14-day doubling time once a certain number of individuals are infected, with cumulative doubling time reported if that number of infections occurred before day 14 in the simulation. The center blue line represents the median doubling time across the simulations. Darker and lighter shading represent the 50% and 95% HDI, respectively.

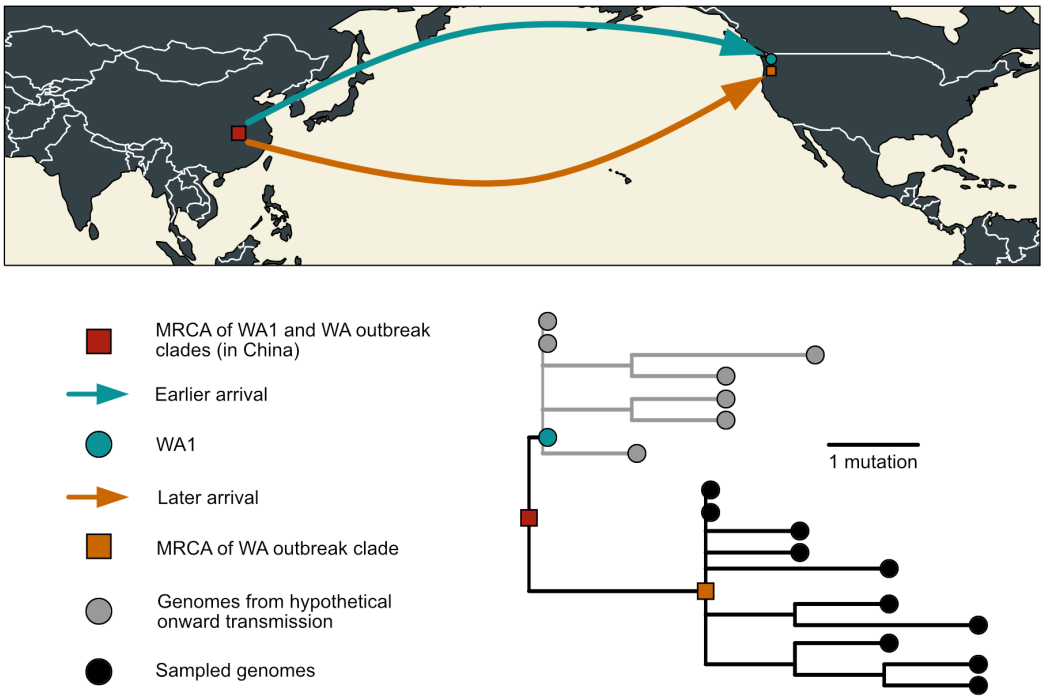


Figure B.23. Early SARS-CoV-2 introductions into Washington state. Similar phylogenetic structure to the origins of SARS-CoV-2 in Wuhan is observed in Washington state, with two separate introductions of SARS-CoV-2 from China differing by two mutations (with no intermediate genomes). Refer to the supplementary text for a discussion comparing the introductions to Washington State with the origins of SARS-CoV-2.

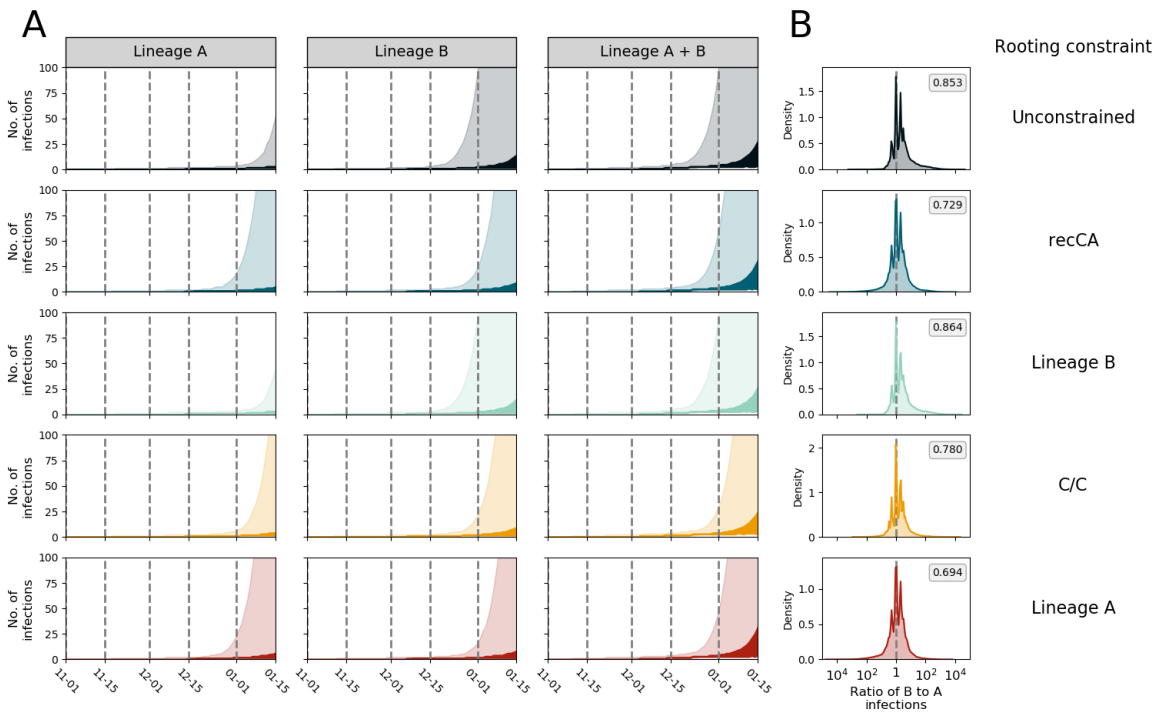


Figure B.24. Dynamics of COVID-19 hospitalizations resulting from separate introductions of lineages A and B. Each row represents a different rooting constraint in phylodynamic analysis, with the bottom three representing a fixed ancestral haplotype. (A) Estimated number of hospitalizations. The header of each column indicates whether the number of infections are caused by lineage A, lineage B, or the two lineages combined. Darker and lighter shading represent the 50% and 95% HPD, respectively. (B) The log ratio of lineage B to lineage A infections on 1 January 2020. Posterior probability of having more lineage B hospitalizations than lineage A reported in the grey box.

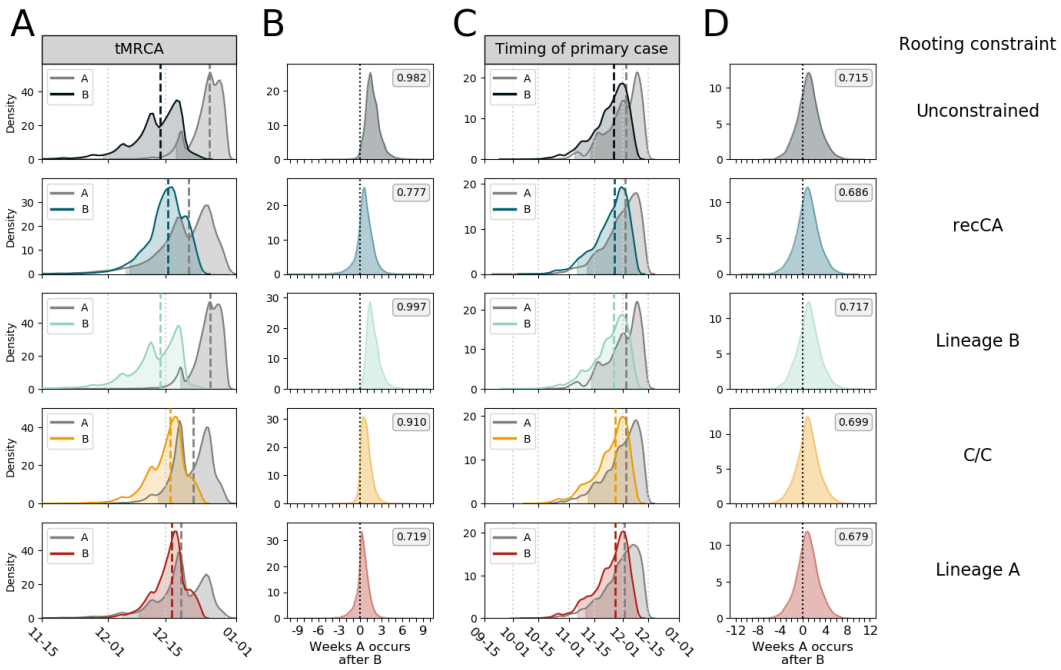


Figure B.25. The timing of the MRCA and primary case for lineage A and lineage B with a shorter doubling time. The simulations used here have a doubling time of 2.65 days (95% HDI: 1.50-4.10). Each row represents a different rooting constraint in phylodynamic analysis, with the bottom three representing a fixed ancestral haplotype. (A) The tMRCA for lineages A and B. (B) The number of weeks the tMRCA of lineage A occurs after the tMRCA of lineage B. (C) The timing of the primary case for lineages A and B. (D) The number of weeks the time of the primary case of lineage A occurs after the time of the primary case of lineage B. Long dashed lines indicate the median and shading represents the 95% HPD for each distribution. Short dashed lines indicate 0 weeks difference between lineages A and B. Posterior probability that lineage A originated after lineage B is reported in the grey box.

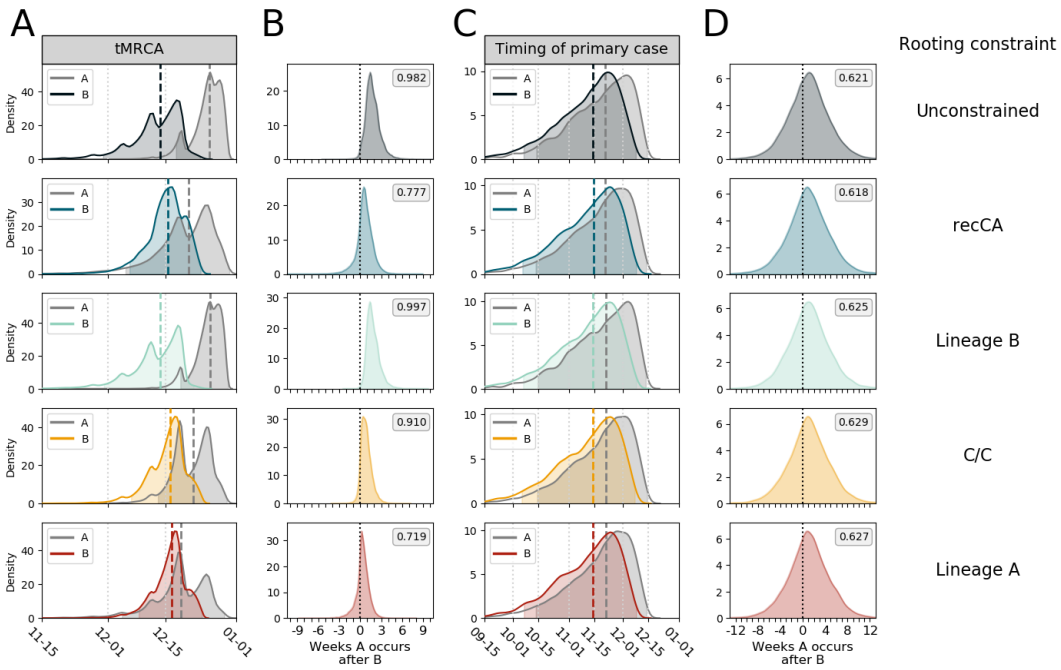


Figure B.26. The timing of the MRCA and primary case for lineage A and lineage B with a longer doubling time. The simulations used here have a doubling time of 4.45 days (95%: HDI: 1.50-7.44). Each row represents a different rooting constraint in phylodynamic analysis, with the bottom three representing a fixed ancestral haplotype. (A) The tMRCA for lineages A and B. (B) The number of weeks the tMRCA of lineage A occurs after the tMRCA of lineage B. (C) The timing of the primary case for lineages A and B. (D) The number of weeks the time of the primary case of lineage A occurs after the time of the primary case of lineage B. Long dashed lines indicate the median and shading represents the 95% HPD for each distribution. Short dashed lines indicate 0 weeks difference between lineages A and B. Posterior probability that lineage A originated after lineage B is reported in the grey box.

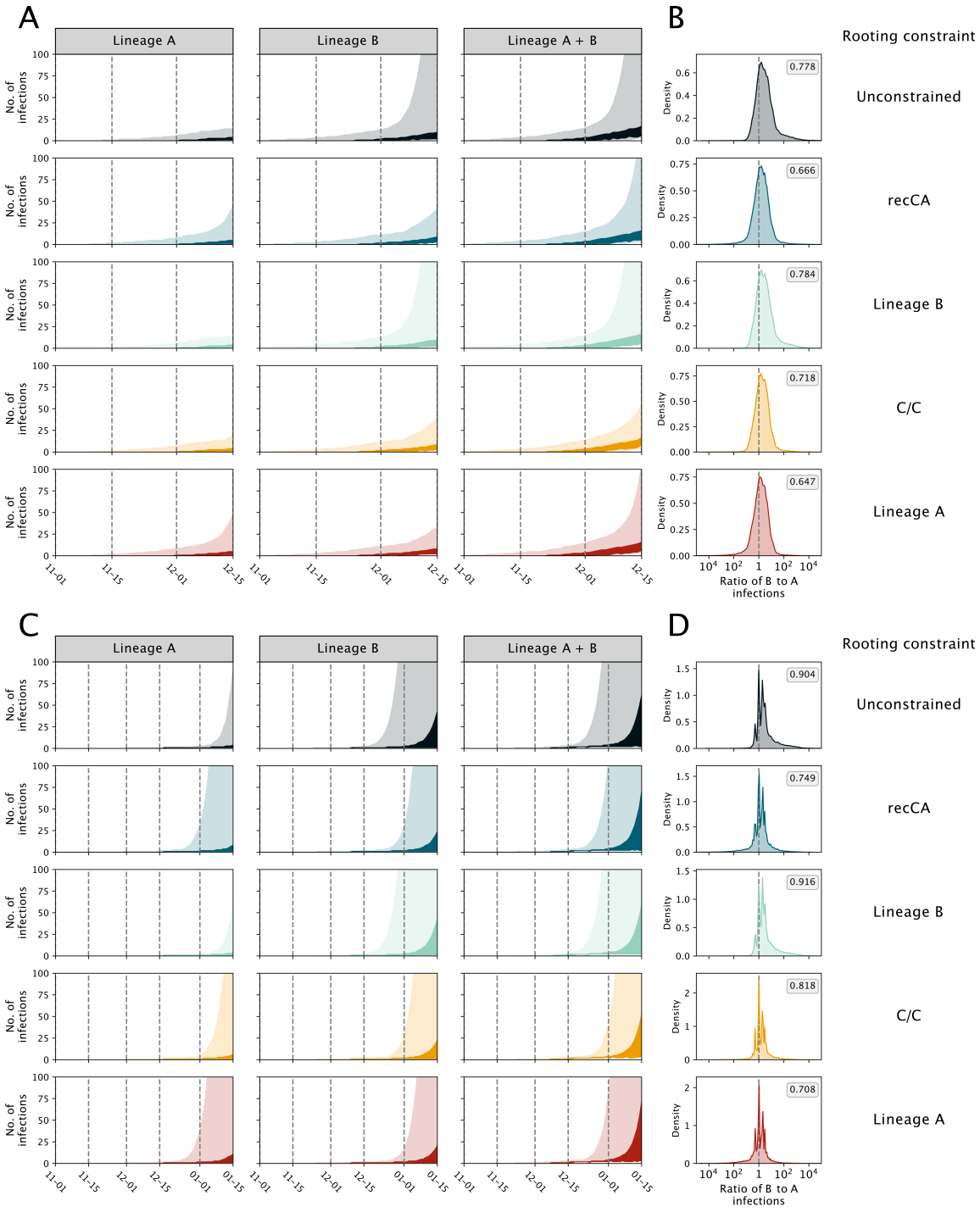


Figure B.27. Dynamics of SARS-CoV-2 resulting from separate introductions of lineages A and B and a shorter doubling time. The simulations used here have a doubling time of 2.65 days (95% HDI: 1.50-4.10). Each row represents a different phylodynamic analysis, with the bottom three representing an enforced ancestral haplotype. (A) Estimated number of infections. The header of each column indicates whether the infections are caused by lineage A, lineage B, or the two lineages combined. Darker and lighter shading represent the 50% and 95% HPD, respectively. (B) The log ratio of lineage B to lineage A infections on 15 December 2019. (C) Estimated number of hospitalizations, with column headers and shading identical to (A). (D) The log ratio of lineage B to lineage A hospitalizations on 1 January 2020. The proportion of the posterior with more lineage B infections or hospitalizations than lineage A in (B, D) is reported in the grey box.

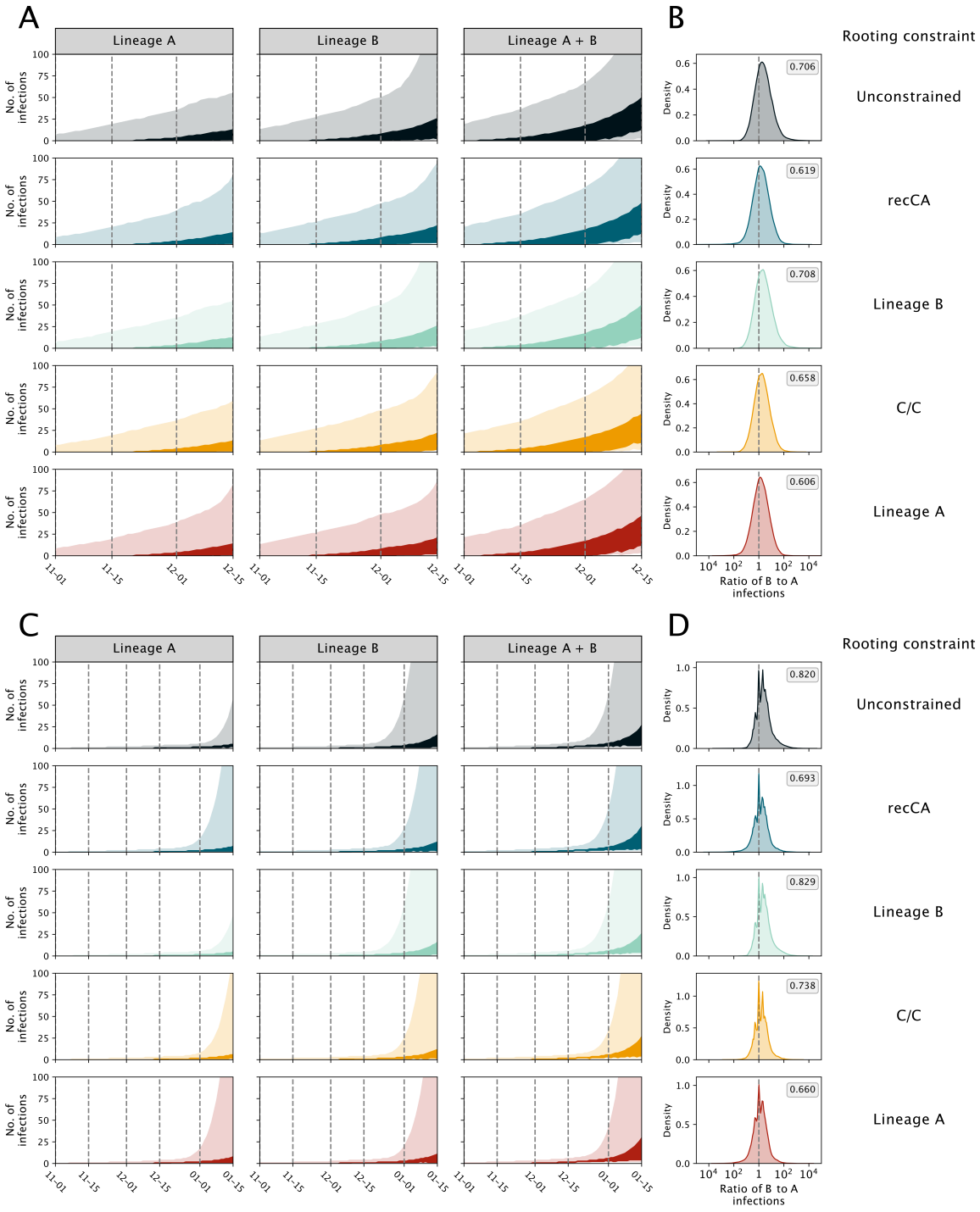


Figure B.28. Dynamics of SARS-CoV-2 resulting from separate introductions of lineages A and B and a longer doubling time. The simulations used here have a doubling time of 4.45 days (95%: HDI: 1.50-7.44). Each row represents a different phylodynamic analysis, with the bottom three representing an enforced ancestral haplotype. (A) Estimated number of infections. The header of each column indicates whether the infections are caused by lineage A, lineage B, or the two lineages combined. Darker and lighter shading represent the 50% and 95% HPD, respectively. (B) The log ratio of lineage B to lineage A infections on 15 December 2019. (C) Estimated number of hospitalizations, with column headers and shading identical to (A). (D) The log ratio of lineage B to lineage A hospitalizations on 1 January 2020. The proportion of the posterior with more lineage B infections or hospitalizations than lineage A in (B, D) is reported in the grey box.

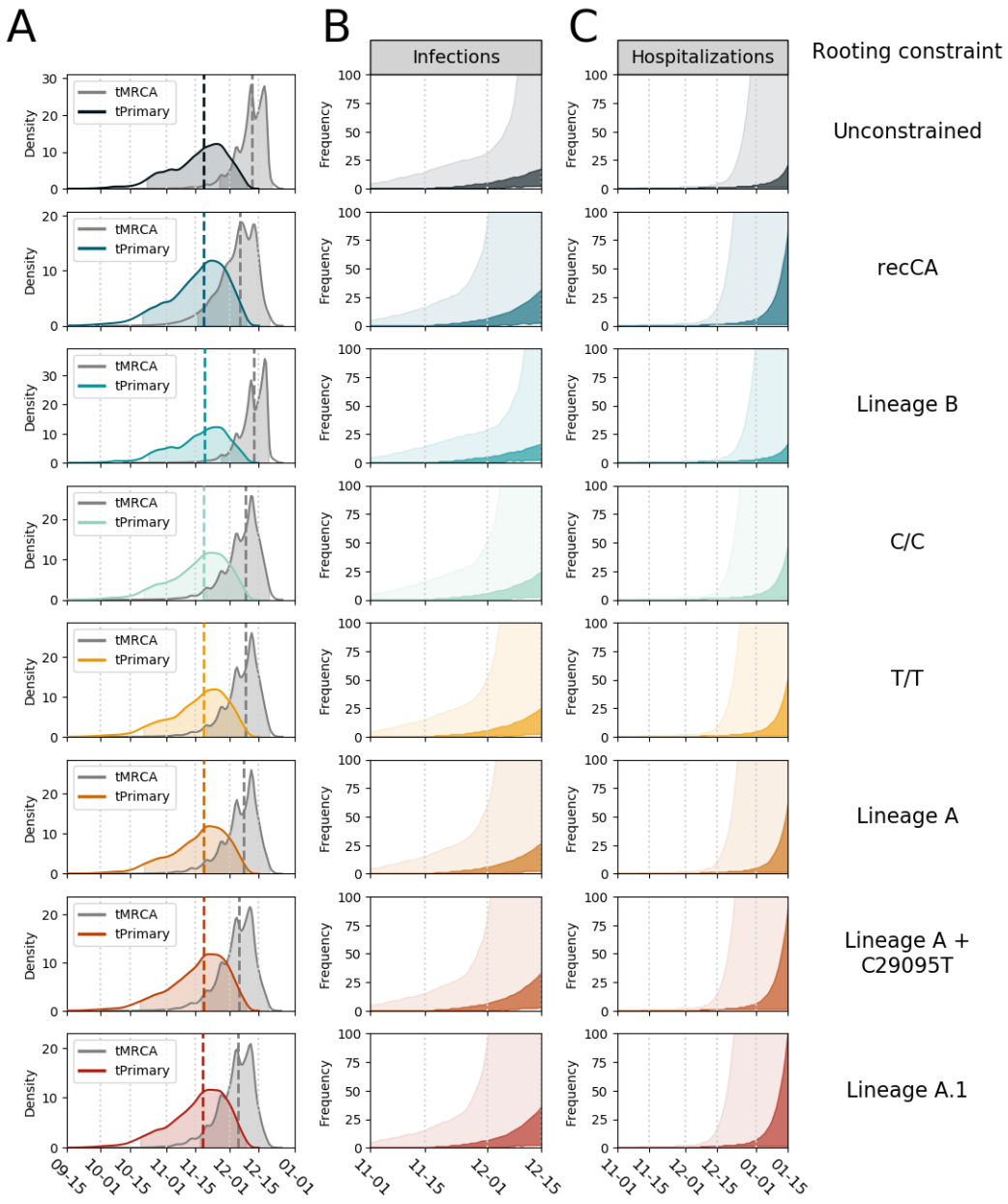


Figure B.29. Single-introduction timing of the MRCA and primary case and subsequent epidemic growth. (A) Posterior distributions of the timing of the MRCA (tMRCA) and primary case (tPrimary), with dashed lines indicating the median and shading representing the 95% HPD for each distribution. (B) Estimated number of infections in late 2019. Darker shading represents 50% HPD; lighter shading represents 95% HPD. (C) Estimated number of hospitalizations in late 2019. The legend indicates the phylodynamic model used: the unconstrained model uses just the SARS-CoV-2 genomes; the recCA-constrained model constrains the ancestor of the MRCA of SARS-CoV-2 as the recCA; the remaining models constrain the MRCA of SARS-CoV-2 as a particular sequence (Fig. B.20; see methods).

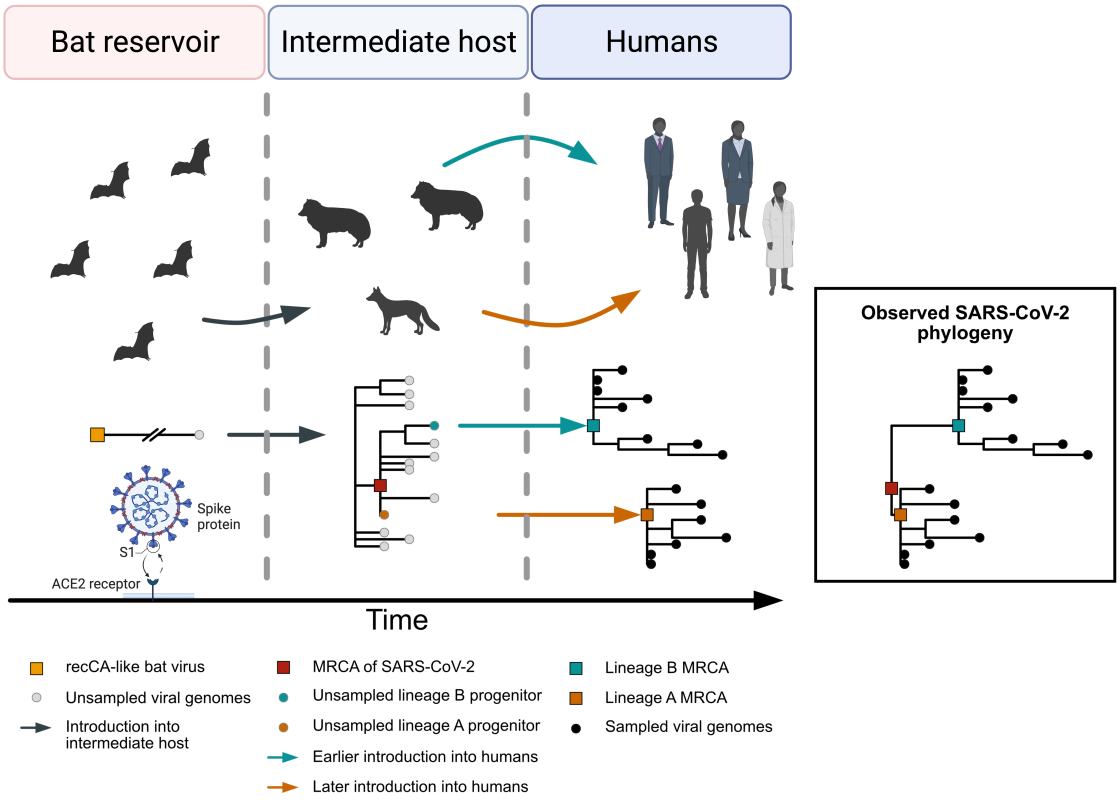


Figure B.30. Schematic depicting the multiple zoonotic origin of SARS-CoV-2. A recCA-like virus was circulating in bats, and likely after gaining the ability to bind ACE2, jumped into an intermediate host. Therein, lineages A and B appeared and were separately introduced into humans shortly thereafter. An example phylogeny of viruses in the intermediate host is depicted, leading to separate phylogenies for lineages A and B. The resulting SARS-CoV-2 phylogeny from the combined lineage A and B viruses is presented in the black box. This scenario depicts a lineage A ancestral haplotype. See Figure B.31 for intermediate and lineage B ancestral haplotypes.

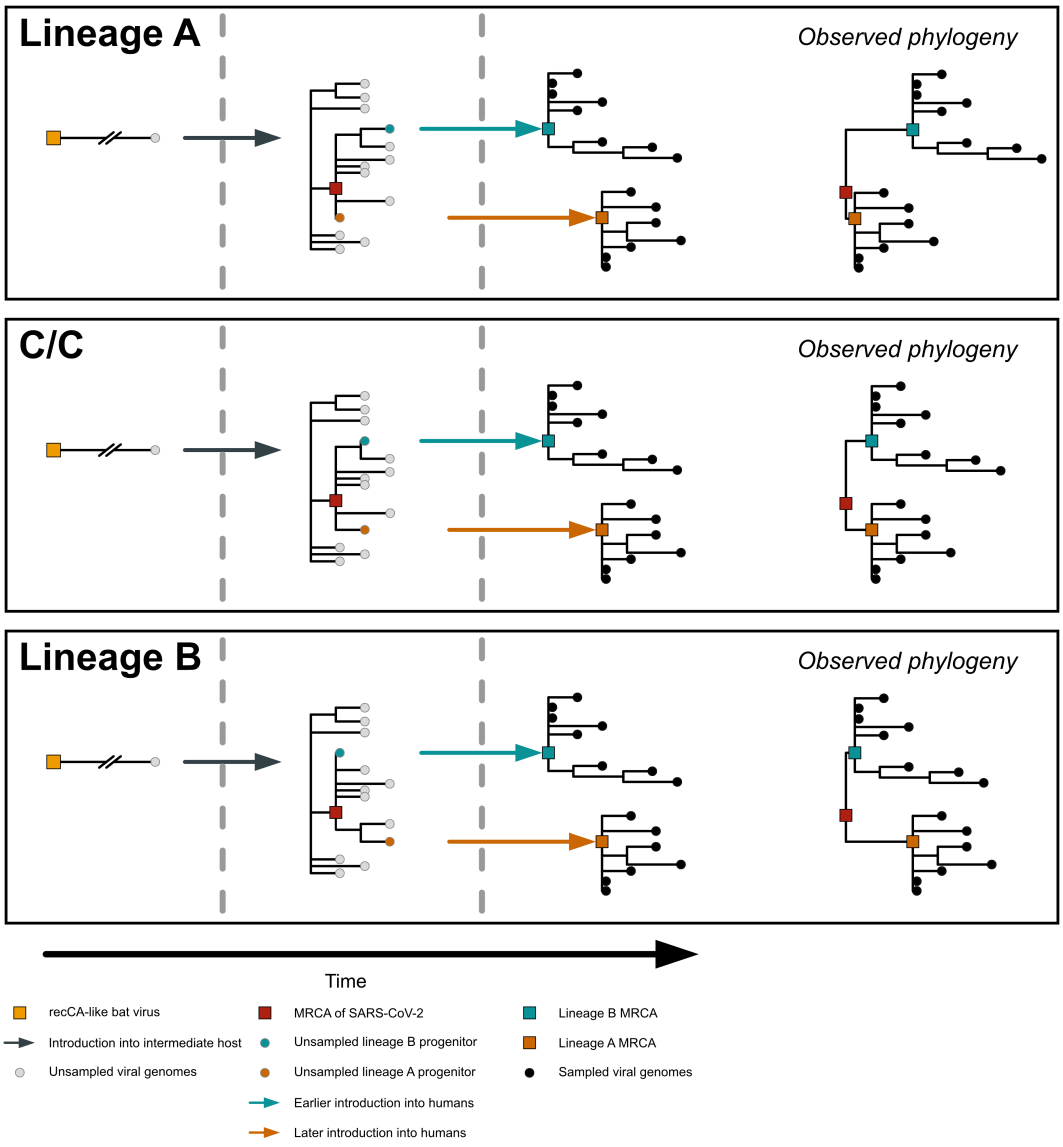


Figure B.31. Rooting orientations of observed SARS-CoV-2 phylogenies resulting from different MRCAs and multiple introductions from the intermediate host. See Figure B.30 for host depictions. The haplotype of the MRCA (red square in the left-center panel) is depicted in the upper left of each box.

Appendix C

Supplemental Material for Chapter 3

C.1 Materials and Methods

C.1.1 Dataset compilation

For the animal sarbecovirus dataset we compiled a set of 167 whole-genome sequences of the subgenus Sarbecovirus (available as of March 2022) sampled primarily in horseshoe bat hosts, including reference genomes of the human viruses SARS-CoV-1 and SARS-CoV-2 and six pangolin-infecting viruses. All sequences were publicly available as of October 2022 (Supplementary Data File 1). Collection date, sampling location and host metadata associated with the sequences were retrieved from the corresponding sequence databases and manually checked with the published papers describing the sequences. For sequences where exact city/cave sampling location information was provided a single longitude and latitude data point corresponding to that location was recorded. For cases where only province-specific sampling information were available a six point geographic location range encompassing each province was recorded. The metadata for these 167 genomes are available in Supplementary Data File 1.

For the SARS-CoV-2 dataset, we used the 787 genomes and sampled 1,000 trees from the posterior tree distribution from Pekar *et al.*[145]. For the SARS-CoV-1 dataset, we downloaded the 98 publicly available whole genomes of SARS-CoV-1 from humans and civets from Genbank, with dates and locations matched by metadata or literature review: https://github.com/andersen-lab/SARS-CoV-1_Evolution/blob/main/metadata_extended.csv. A list of all accessions used in this study are available in Supplementary Data File 2. All GISAID accessions are also available through GISAID when using the identifier EPI_SET_230410yb.

C.1.2 Recombination analysis

The 167 whole-genome sequences were aligned using MAFFT (v7.453, localpair option)[244] and manually curated using BioEdit[245]. To explore the recombination patterns of SARS-CoV-1-like and SARS-CoV-2-like sarbecoviruses separately we selected a number of representatives from each set of genomes on which to perform recombination analysis using the Genetic Algorithm for Recombination Detection (GARD)[143]. The genome selection was made by first performing a preliminary phylogenetic analysis of the relatedness between these genomes using the recombination breakpoints presented in Lytras *et al.* (2022)[35]. Based on the preliminary analysis, clusters of genomes grouping with SARS-CoV-1 and SARS-CoV-2 for at least one of their genomic segments and showing the same overall recombination pattern were manually identified and one representative genome was selected from each cluster for the corresponding downstream GARD analyses. For the SARS-CoV-2-like set, 19 genomes including the SARS-CoV-2 reference Wuhan-Hu-1 were used in the GARD analysis. For the SARS-CoV-1-like set, 38 genomes including the SARS-CoV-1 HSZ-Cc genome were used in the analysis (Supplementary Data File 1). The representative sequences for each set were retrieved from the full alignment and the extended 3' and 5' ends were trimmed from each resulting subset alignment to reduce sequence noise in the recombination analysis. GARD was performed for each subset alignment using the HyPhy software suite v.2.5.29[246] under a general time reversible (GTR) substitution model and a 3-class general discrete distribution (GDD) to account for site-to-site rate variation. The SARS-CoV-2-like GARD analysis finished with a total of 72 consecutive breakpoint models, while the SARS-CoV-1-like analysis finished with 56. Based on these models, we accepted positions as confident recombination breakpoints if they were present in at least 1/3 of the consecutive models and breakpoints closer than 100 alignment positions to one another were merged. The resulting breakpoint positions were mapped back to the full whole-genome alignment which was subsequently split into the corresponding non-recombinant regions (NRR) for each set of breakpoints. To avoid uninformative alignment regions biasing phylogenetic inferences, all NRR alignments were processed using HMMCleaner[247] before performing any downstream analyses. The GARD breakpoint results for the SARS-CoV-2-like set have previously been used in Martin *et al.* (2022)[248].

To separate all genomes into SARS-CoV-1-like and SARS-CoV-2-like viruses we first inferred maximum likelihood phylogenies for every NRR alignment using IQ-TREE (v.1.4.4)[249] under a GTR+I+F+G4 substitution model, assessing node support by performing 10,000 replicates of the ultrafast bootstrap approximation for each inference[250]. All trees were rooted by the outgroup clade of sarbecovirus genomes sampled in Europe and Africa (accessions: NC_014470, MZ190137, KY352407, MT726043, MT726044, MT726045, MZ190138, MW719567) and nodes with ultrafast bootstrap support below 70 were collapsed using the 'ete3' python package[251]. Based on the resulting topologies, the viruses clustering monophyletically with SARS-CoV-2 but not with SARS-CoV-1 for at least one of the SARS-CoV-2-like GARD NRR trees were placed in the SARS-CoV-2-like viruses and vice versa for the SARS-CoV-1-like virus set. This led to two non-independent sets of 27 SARS-CoV-2-like and 140 SARS-CoV-1-like genomes (each including SARS-CoV-2 and SARS-CoV-1 respectively) used in the downstream dating and phylogeographic analyses presented in the paper.

C.1.3 Comparison of breakpoint positions

Based on the GARD recombination analysis, 30 and 26 breakpoint positions were detected for the two virus subgroups. Only one genome was used in both GARD sets (Supplementary Data File 1), so overall breakpoint inferences are expected to represent independent recombination patterns of known SARS-CoV-1-like and SARS-CoV-2-like viruses respectively. We calculated the distances between the closest position pairs between the two breakpoint sets, resulting in 26 pair distances (each SARS-CoV-2-like position matched with its closest SARS-CoV-1-like position). To compare whether the pairs were more or less closely located than one would expect by chance, we performed 100 simulations of position sets randomly sampled across the same alignment length without replacement (one set with 26 positions and one with 30 for each simulation to match the empirical data on the number of breakpoints). We then quantified the distances between closest pairs in the simulated sets as done with the observed breakpoints. Both empirical and simulated distances formed distributions with log-normal shapes, hence 90% confidence intervals and means were estimated using Cox's method for inferring log-normal confidence intervals[252]. Similarly, the significantly closest observed breakpoint pairs were defined as those with distances below the

one-tailed lower 90% confidence interval of the observed log-normal distance distribution.

C.1.4 Molecular clock calibration

There is insufficient data to inform a molecular clock when using tip dating with sarbecoviruses sampled from bats and pangolins[136]. To identify a suitable rate prior for each NRR identified from the recombination analysis, we conducted a phylogenetic analysis of SARS-CoVs sampled from humans.

Molecular clock calibration across the 27 inferred NRRs of the SARS-CoV-2-like viruses was conducted using a Bayesian approach with BEAST v1.10.5[253]. We inferred the substitution rate of SARS-CoV-2 across each of these NRRs by using an empirical tree distribution (n=1,000) from our previously published Bayesian phylogenetic analysis of 787 early pandemic genomes (Supplementary Data File 2), sharing the topologies and substitution models across all the NRRs, but assuming an independent and estimable strict clock rate under an uninformative continuous-time Markov chain (CTMC) rate prior[254] for each NRR. We ran one Markov chain Monte Carlo (MCMC) chain of one million generations, subsampling every one thousand iterations to a continuous parameter log file. The first 15% of the chains were discarded as burn-in. Convergence and mixing was assessed in Tracer v1.7.1[255] and all relevant effective sample size (ESS) values were >200.

To calibrate the molecular clock for the SARS-CoV-1-like viruses analysis, we first performed Bayesian phylogenetic inference on 98 human and civet SARS-CoV-1 genomes (Supplementary Data File 2), using an HKY85 substitution model with a gamma site heterogeneity model and 4 gamma categories (G4), a strict molecular clock, and an exponential growth coalescent prior. We ran a single chain of 100 million generations, subsampling every 10 thousand iterations to a continuous parameter log and tree files. We next inferred the substitution rate of SARS-CoV-1 across the 31 inferred NRRs of the SARS-CoV-1-like viruses by using an empirical tree distribution (n=1,000) from these results, sharing the topologies and substitution models across all the NRRs, with independent and estimable clock rates under CTMC rate priors. We ran one chain of one million generations, subsampling every one hundred iterations to a continuous parameter log file. For each step of the inference, the first 15% of the chains were

discarded as burn-in. Convergence and mixing was assessed in Tracer and all relevant effective sample size (ESS) values were >200.

C.1.5 Bayesian divergence time estimation

We used BEAST to perform phylogenetic inference and ancestral state reconstruction across the 27 NRRs of the SARS-CoV-2-like viruses. For the primary analysis (Fig. C.2), we used all of the 27 SARS-CoV-2-like genomes (including SARS-CoV-2), an uncorrelated relaxed clock using calibrated region-specific normally-distributed substitution rate priors for each NRR based on SARS-CoV-2 sampled from humans, a GTR+G4 substitution model, a non-parametric skygrid prior[256] with 49 grid points and a cutoff of 1,000 (which translates to 1022 CE), and Hamiltonian Monte Carlo gradient-based sampling for node ages and branch rates[257;258]. Specifically, we set the mean and standard deviation of the rate prior for each NNR to equal the posterior mean and standard deviation estimated during molecular clock calibration with SARS-CoV-2 genomes. We tracked the stem age of SARS-CoV-2 (i.e., where SARS-CoV-2 connects to the rest of the phylogeny, also referred to as the time of the closest-inferred ancestor, with the parent node of SARS-CoV-2 referred to as the closest-inferred ancestor). For each NRR, we ran one chain of 10 million generations, subsampling every one thousand iterations to continuous parameter log and tree files. The first 15% of the chain was discarded as burn-in for most of the analyses; for several analyses we had to discard a greater percentage of the chain to account for poor initial mixing. Convergence and mixing was assessed in Tracer, and all relevant ESS values were >200.

We performed the same analysis for the SARS-CoV-1-like viruses for its 31 NRRs, using all of the 140 SARS-CoV-1-like genomes (including SARS-CoV-1). The BEAST parameterization was identical to the analysis for the SARS-CoV-2-like viruses, except we[32] tracked the stem age of SARS-CoV-1 (i.e., the time of the closest-inferred ancestor of SARS-CoV-1),[33] used the relevant calibrated molecular clock priors for the 31 NRRs of the SARS-CoV-1-like viruses, and[34] ran one chain of 100 million generations, sub-sampling every 10 thousand iterations to continuous parameter log and tree files.

C.1.6 Bayesian divergence time estimation sensitivity analyses

We performed a series of sensitivity analyses for the Bayesian divergence time estimation: (i) dividing the standard deviation of the molecular clock prior by 5, (ii) dividing the standard deviation of the molecular clock prior by 10, and (iii) including an additional human SARS-CoV-1 genome and a civet SARS-CoV-1 genome in the SARS-CoV-1-like viruses analyses. For (iii) , SZ3 (AY304486) was selected as representative of the earliest sampled detection in the intermediate animal reservoir (market-based civets). HZS2-C (AY394992) was selected to represent the early stages of the human epidemic from late January 2003 onwards.

C.1.7 Constructing the recombinant common ancestor

For each NRR of the SARS-CoV-2-like viruses, we reconstructed the most likely nucleotide sequence belonging to the parent node of SARS-CoV-2 based on the tree distribution from the primary Bayesian divergence time estimation (Fig. C.2). The most likely nucleotide sequences for the 27 NRRs are then concatenated to form the recombinant common ancestor (recCA) of SARS-CoV-2.

To analyze how the recCA changed as more genomes were released, we constructed the recCA of SARS-CoV-2 as the genomes of closely related sarbecoviruses were chronologically published. We performed ancestral state reconstruction using the posterior trees with all 27 taxa. We then identified the inferred sequence of the most recent common ancestor of SARS-CoV-2 and its most closely related taxa based on the published genomes as new genomes were sequenced. We then examined the genetic similarity of the recCA, as well as the most closely related published genome, to SARS-CoV-2 as a function of the publication date. The same analysis was performed for SARS-CoV-1 and the SARS-CoV-1-like viruses.

C.1.8 Bayesian phylogeographic inference

We inferred the evolutionary and geographic history of the 27 NRRs of the SARS-CoV-2-like viruses with BEAST and using the BEAGLE library v3 to increase computational performance[259]. In order to optimally inform a flexible phylogeographic reconstruction approach, we jointly fit a bivariate continuous diffusion model using a shared estimable precision matrix

across all independent random NRRs phylogenies, rescaling this matrix across branches in each tree to model a relaxed random walk (RRW)[260]. All branch-specific RRW rate scalars are drawn from the same log-normal distribution with a mean of 1 and an estimable standard deviation. In the continuous phylogeographic diffusion model, we exclude the sampling locations for the human and pangolin viruses and integrate over the uncertainty for locations that lack precision. As a tree prior, we specify a joint non-parametric skygrid prior with 49 grid points and a cutoff of 1,000 (which translates to 1022 CE) or a cutoff of 0.5 substitutions per site for trees scaled in units of genetic distance (see below). We further specify a joint HKY85 nucleotide substitution model across NRRs and independent strict molecular clock models; these relatively simple sequence evolutionary parameterizations were chosen to meet the requirements for the Prisoner of War (PoW) model[138] to rescale the evolutionary histories (see below). We perform two versions of this analysis (Fig. C.2): one with the clock rate priors also used in the previously described divergence time estimation to estimate the trees in units of time and one with clock rates fixed to 1 and contemporaneous tips in order to estimate trees in substitution units for PoW transformation. We ran two chains of 500 million generations, subsampling every 10 thousand iterations to continuous parameter log and tree files; the first 10% of the chain was discarded as burn-in. Convergence and mixing was assessed in Tracer, and all relevant ESS values were >200.

We performed the same analysis for the SARS-CoV-1-like virus set for its 31 NRRs. The BEAST parameterization was identical to the analysis for the SARS-CoV-2-like viruses, but here we use the rate priors with standard deviations divided by 5. We ran 10 chains for $>$ 200 million generations, subsampling every 500 thousand iterations to continuous parameter log and tree files; the first 10% of the chain was discarded as burn-in. Convergence and mixing was assessed in Tracer, and all relevant ESS values were >200.

To account for the time-dependent decline in the evolutionary rate estimates of sarbecoviruses, we applied the PoW model, a mechanistic evolutionary method for estimating virus substitution rates over long evolutionary timescales. For this, we used the post-burn-in substitution trees inferred from the previous step along with the posterior rate distributions for each NRR of the SARS-CoV-1-like viruses and the SARS-CoV-2-like viruses to rescale substitution trees into time trees using the PoW transformation (see ref Ghafari *et al.*[138] for more details on

the method).

We used TreeAnnotator 1.10[253] to obtain and annotate a maximum clade credibility (MCC) tree for each continuous phylogeographic reconstruction. Finally, we used the R package 'seraphim'[261] to extract the spatiotemporal information embedded within 100 trees sampled from each post-burn-in posterior distribution and to visualize the different phylogeographic reconstructions. For each NRR, we estimated the weighted diffusion coefficients[152] with the 'spreadStatistics' function of the R package 'seraphim'[261].

As a sensitivity analysis, we repeated these phylogeographic analyses using the sampled locations of the pangolin and human viruses.

C.1.9 Ranking the dispersal velocity of the lineage descending from the closest-inferred ancestor

Using the posterior tree distribution from the tip-dated phylogeography that excluded the human and pangolin viruses, we calculated the dispersal velocity across each branch of each post-burn-in posterior tree in the 50 years preceding the most recent tip, excluding the branches leading to the human or pangolin viruses (or their clades, if the clade was entirely composed of pangolin and human viruses). These dispersal velocities were pooled for each NRR, resulting in a dispersal velocity distribution per NRR. We next computed the minimum distance between the closest-inferred ancestor and Guangdong province or Hubei province for SARS-CoV-1 or SARS-CoV-2, respectively. The necessary dispersal velocity for the closest-inferred ancestor to reach either Guangdong or Hubei province was calculated using this minimum distance and the branch leading to, respectively, SARS-CoV-1 or SARS-CoV-2 (i.e., the time between closest-inferred ancestor and the SARS-CoV). We then computed the rank of this dispersal velocity by comparing it against the dispersal velocity distribution of the given NRR, resulting in a posterior distribution of dispersal velocity ranks for the lineages descending from the closest-inferred ancestors of SARS-CoV-1 and SARS-CoV-2 to reach their provinces of emergence.

C.1.10 Modeling of bat distributions

Our bat distribution database comprised of the data used in recently published studies[262], recent Global Diversity Information Facility (GBIF) data[263;264], and the Darkcide database[265].

The Darkcide database included recent data collected by bioacoustic surveys in Chiang Mai and Kanchanaburi in Thailand by A.C. Hughes using an echometer Pro, and multiple published datasets[262–264;266–274]. The data was integrated as we attempt to keep the database complete, even if not all data is used within our particular study. All Rhinolophid species, as well as *Chaerephon plicatus* and *Aselliscus stoliczkanus* were extracted from the dataset, and then the dataset was clipped to Southeast Asia (India to the West, Mongolia to the North, Japan to the East and Malaysia to the South) to buffer the region of interest with distribution data. Once all duplicate records had been removed, the clipped data for Rhinolophids included 10,254 distribution points for all species across the region, in addition to 235 records for *A. stoliczkanus* and 108 for *C. plicatus*. For environmental data, we based variable selection on those used in Zhou *et al.* (2021)[262] and known from our previous work to be important to bats in the region[275;276]. All analysis was conducted in ArcMap 10.8 (ESRI) at a resolution of 0.008o, which is approximately 1km2, and all layers were clipped and subsampled to the same resolution using a mask. Climate variables were downloaded from Chelsa[277](available at https://chelsa-climate.org/downloads/) based on a selection from our previous analyses (the bioclimatic variables: Annual Mean Temperature, Mean Diurnal temperature Range, Isothermality, Temperature Seasonality, Maximum and Minimum Temperature, Annual Precipitation, Precipitation seasonality, Maximum and Minimum precipitation)[275;276]. From Envirem (https://envirem.github.io/) we used the Thornthwaite Aridity index, climate moisture index, continentality, embergers pluviothermic quotient, growing degree days above 0, and both annual mean potential evapotranspiration and seasonality of potential evapotranspiration. In addition we used actual evapotranspiration[278](Trabusco & Zomer 2019), a vegetation canopy height layer[279], forest canopy height[280], tree density[281], distance to bedrock[282](the bdticm and bdricm aggregated grids available at https://github.com/ISRICWorldSoil/SoilGrids250m/), and bulk carbon stock[282]. These were all trimmed to the same region that bat data was clipped to and then species distributions were modelled using Maxent[283] using the GUI following the same approach as Hughes *et al.* (2012)[275]. For each species, three replicates were run and the average was used. These probabilistic distribution maps were then reclassified using the 10th percentile cumulative logistic threshold to produce a binary presence-absence map with ArcMap 10.8, and all models had an AUC exceeding 0.85. The

models were then clipped to the study region, and any Rhinolophids with no modelled suitable habitat in the clipped study area were removed from analysis. As the projections show suitable habitat but do not account for species biogeography we then cross-checked the 40 remaining Rhinolophid maps against the IUCN database (https://www.iucnredlist.org/), and species not listed were checked against relevant publications on the species. Based on the IUCN mapped ranges species endemic to Japan, or species not found in Japan or at similar latitudes (areas of the South Korean Peninsula are climatically suitable for a number of species, but lack continuous habitat to reach these regions if largely distributed in Southeast Asia) were then clipped to the biogeographically suitable range. The binary maps were then stacked using the Mosaic to New Raster function in ArcMap 10.8 to find the maximum number of species that may be co-distributed.

C.1.11 Data and Code Availability

All genome accessions are available in Supplementary Data File 2. Supplementary Data Files 1 and 2 are available at https://www.biorxiv.org/content/10.1101/2023.07.12.548617v1. The XML files, code, and newick trees are available at https://github.com/phylogeography/sarbecovirus_phylogeography.

C.2 Supplementary Tables

Table C.1. Posterior clock rates across each NRR from the clock calibration using SARS-CoV-1 genomes and Bayesian divergence time analyses of SARS-CoV-1-like viruses. The clock calibration posterior results are used as a prior for the Bayesian divergence time analyses, with the standard deviation of the clock calibration results varying across the analyses.

NRR	**Clock calibration posterior rate**	**Bayesian divergence time estimation**		
		Primary analysis	**Robustness analysis**	
		(Clock calibration-centred prior)	**(Clock calibration-centred prior with stdev / 5)**	**(Clock calibration-centred prior with stdev / 10)**
1	2.06e-03 (3.81e-04)	1.35e-03 (1.01e-03, 1.72e-03)	2.00e-03 (1.86e-03, 2.14e-03)	2.04e-03 (1.97e-03, 2.12e-03)
2	1.21e-03 (3.74e-04)	8.36e-04 (4.88e-04, 1.23e-03)	1.18e-03 (1.04e-03, 1.31e-03)	1.20e-03 (1.13e-03, 1.27e-03)
3	1.22e-03 (5.34e-04)	1.09e-03 (6.95e-04, 1.54e-03)	1.20e-03 (1.02e-03, 1.40e-03)	1.22e-03 (1.12e-03, 1.32e-03)
4	2.55e-03 (6.45e-04)	1.36e-03 (9.69e-04, 1.84e-03)	2.41e-03 (2.18e-03, 2.65e-03)	2.51e-03 (2.38e-03, 2.64e-03)
5	1.61e-03 (3.77e-04)	1.19e-03 (8.63e-04, 1.57e-03)	1.57e-03 (1.43e-03, 1.70e-03)	1.60e-03 (1.53e-03, 1.67e-03)
6	2.04e-03 (5.27e-04)	1.06e-03 (7.43e-04, 1.45e-03)	1.93e-03 (1.74e-03, 2.13e-03)	2.01e-03 (1.91e-03, 2.11e-03)
7	2.23e-03 (6.06e-04)	1.35e-03 (9.29e-04, 1.80e-03)	2.12e-03 (1.90e-03, 2.34e-03)	2.20e-03 (2.09e-03, 2.32e-03)
8	1.85e-03 (4.05e-04)	1.07e-03 (7.53e-04, 1.43e-03)	1.79e-03 (1.65e-03, 1.95e-03)	1.84e-03 (1.76e-03, 1.91e-03)
9	2.53e-03 (6.94e-04)	2.03e-03 (1.29e-03, 2.77e-03)	2.46e-03 (2.21e-03, 2.74e-03)	2.51e-03 (2.38e-03, 2.65e-03)
10	2.09e-03 (8.46e-04)	1.54e-03 (9.50e-04, 2.22e-03)	2.01e-03 (1.71e-03, 2.31e-03)	2.06e-03 (1.91e-03, 2.23e-03)
11	2.34e-03 (4.53e-04)	1.16e-03 (8.38e-04, 1.55e-03)	2.25e-03 (2.09e-03, 2.43e-03)	2.31e-03 (2.23e-03, 2.40e-03)
12	1.15e-03 (2.49e-04)	6.43e-04 (4.57e-04, 8.42e-04)	1.10e-03 (1.00e-03, 1.19e-03)	1.13e-03 (1.09e-03, 1.18e-03)
13	1.16e-03 (2.50e-04)	8.09e-04 (5.40e-04, 1.09e-03)	1.13e-03 (1.04e-03, 1.23e-03)	1.15e-03 (1.11e-03, 1.20e-03)
14	9.28e-04 (2.96e-04)	7.70e-04 (4.91e-04, 1.09e-03)	9.08e-04 (8.04e-04, 1.02e-03)	9.23e-04 (8.66e-04, 9.81e-04)
15	1.03e-03 (3.03e-04)	9.13e-04 (5.93e-04, 1.27e-03)	1.03e-03 (9.19e-04, 1.15e-03)	1.03e-03 (9.70e-04, 1.09e-03)
16	1.61e-03 (3.32e-04)	1.13e-03 (8.28e-04, 1.50e-03)	1.57e-03 (1.45e-03, 1.70e-03)	1.60e-03 (1.54e-03, 1.66e-03)

Table C.1, cont. Posterior clock rates across each NRR from the clock calibration using SARS-CoV-1 genomes and Bayesian divergence time analyses of SARS-CoV-1-like viruses. The clock calibration posterior results are used as a prior for the Bayesian divergence time analyses, with the standard deviation of the clock calibration results varying across the analyses.

NRR	**Clock calibration posterior rate**	**Bayesian divergence time estimation**		
		Primary analysis	**Robustness analysis**	
		(Clock calibration-centred prior)	**(Clock calibration-centred prior with stdev / 5)**	**(Clock calibration-centred prior with stdev / 10)**
17	1.87e-03 (4.33e-04)	1.56e-03 (1.13e-03, 2.04e-03)	1.83e-03 (1.68e-03, 2.00e-03)	1.87e-03 (1.78e-03, 1.95e-03)
18	1.08e-03 (4.71e-04)	1.18e-03 (7.70e-04, 1.66e-03)	1.13e-03 (9.54e-04, 1.31e-03)	1.09e-03 (1.00e-03, 1.18e-03)
19	1.99e-03 (5.10e-04)	1.13e-03 (7.41e-04, 1.59e-03)	1.91e-03 (1.72e-03, 2.09e-03)	1.97e-03 (1.88e-03, 2.07e-03)
20	2.21e-03 (7.49e-04)	1.42e-03 (9.18e-04, 2.04e-03)	2.11e-03 (1.85e-03, 2.39e-03)	2.18e-03 (2.04e-03, 2.32e-03)
21	3.38e-03 (6.44e-04)	1.56e-03 (1.18e-03, 2.02e-03)	3.24e-03 (3.02e-03, 3.50e-03)	3.35e-03 (3.22e-03, 3.47e-03)
22	3.38e-03 (7.33e-04)	1.74e-03 (1.18e-03, 2.34e-03)	3.24e-03 (2.99e-03, 3.54e-03)	3.34e-03 (3.21e-03, 3.49e-03)
23	4.00e-03 (1.05e-03)	2.98e-03 (2.05e-03, 4.22e-03)	3.86e-03 (3.47e-03, 4.26e-03)	3.97e-03 (3.77e-03, 4.16e-03)
24	2.08e-03 (5.97e-04)	1.70e-03 (1.09e-03, 2.45e-03)	2.04e-03 (1.82e-03, 2.27e-03)	2.07e-03 (1.96e-03, 2.19e-03)
25	2.80e-03 (5.11e-04)	1.49e-03 (1.14e-03, 1.94e-03)	2.70e-03 (2.52e-03, 2.90e-03)	2.78e-03 (2.68e-03, 2.87e-03)
26	3.15e-03 (9.71e-04)	1.75e-03 (1.21e-03, 2.40e-03)	2.96e-03 (2.62e-03, 3.31e-03)	3.10e-03 (2.92e-03, 3.29e-03)
27	3.78e-03 (8.66e-04)	2.07e-03 (1.43e-03, 2.89e-03)	3.66e-03 (3.35e-03, 4.00e-03)	3.75e-03 (3.59e-03, 3.92e-03)
28	2.41e-03 (6.81e-04)	1.83e-03 (1.18e-03, 2.59e-03)	2.36e-03 (2.11e-03, 2.61e-03)	2.40e-03 (2.27e-03, 2.53e-03)
29	3.22e-03 (8.02e-04)	2.02e-03 (1.42e-03, 2.69e-03)	3.10e-03 (2.81e-03, 3.41e-03)	3.19e-03 (3.04e-03, 3.35e-03)
30	6.63e-03 (1.71e-03)	2.68e-03 (1.72e-03, 3.77e-03)	6.28e-03 (5.66e-03, 6.91e-03)	6.54e-03 (6.21e-03, 6.87e-03)
31	2.88e-03 (4.98e-04)	1.81e-03 (1.34e-03, 2.34e-03)	2.81e-03 (2.62e-03, 3.00e-03)	2.87e-03 (2.77e-03, 2.96e-03)

Table C.2. Posterior clock rates across each NRR from the clock calibration using SARS-CoV-2 genomes and Bayesian divergence time analyses of SARS-CoV-2-like viruses. The clock calibration posterior results are used as a prior for the Bayesian divergence time analyses, with the standard deviation of the clock calibration results varying across the analyses.

NRR	**Clock calibration posterior rate**	**Bayesian divergence time estimation**		
		Primary analysis	**Robustness analysis**	
		(Clock calibration-centred prior)	**(Clock calibration-centred prior with stdev / 5)**	**(Clock calibration-centred prior with stdev / 10)**
1	1.35e-03 (3.26e-04)	1.27e-03 (8.12e-04, 1.82e-03)	1.34e-03 (1.22e-03, 1.47e-03)	1.35e-03 (1.29e-03, 1.41e-03)
2	1.06e-03 (2.65e-04)	1.16e-03 (7.24e-04, 1.77e-03)	1.07e-03 (9.63e-04, 1.17e-03)	1.06e-03 (1.01e-03, 1.11e-03)
3	2.03e-03 (7.07e-04)	2.06e-03 (1.07e-03, 3.36e-03)	2.03e-03 (1.78e-03, 2.31e-03)	2.03e-03 (1.90e-03, 2.17e-03)
4	1.23e-03 (2.49e-04)	1.26e-03 (8.19e-04, 1.79e-03)	1.24e-03 (1.14e-03, 1.33e-03)	1.23e-03 (1.18e-03, 1.28e-03)
5	1.38e-03 (3.62e-04)	1.16e-03 (6.66e-04, 1.74e-03)	1.37e-03 (1.24e-03, 1.52e-03)	1.38e-03 (1.31e-03, 1.45e-03)
6	1.01e-03 (3.14e-04)	1.15e-03 (5.63e-04, 2.06e-03)	1.01e-03 (8.91e-04, 1.14e-03)	1.01e-03 (9.47e-04, 1.07e-03)
7	8.65e-04 (2.77e-04)	9.62e-04 (5.17e-04, 1.62e-03)	8.70e-04 (7.64e-04, 9.74e-04)	8.66e-04 (8.13e-04, 9.22e-04)
8	6.21e-04 (2.18e-04)	6.95e-04 (2.91e-04, 1.27e-03)	6.24e-04 (5.39e-04, 7.11e-04)	6.22e-04 (5.81e-04, 6.67e-04)
9	7.17e-04 (1.33e-04)	7.55e-04 (5.30e-04, 1.04e-03)	7.19e-04 (6.64e-04, 7.69e-04)	7.18e-04 (6.93e-04, 7.43e-04)
10	9.02e-04 (2.24e-04)	9.22e-04 (5.52e-04, 1.40e-03)	9.04e-04 (8.20e-04, 9.95e-04)	9.02e-04 (8.59e-04, 9.44e-04)
11	1.21e-03 (2.79e-04)	1.04e-03 (6.73e-04, 1.50e-03)	1.20e-03 (1.09e-03, 1.31e-03)	1.21e-03 (1.15e-03, 1.26e-03)
12	4.99e-04 (1.07e-04)	5.23e-04 (3.30e-04, 7.64e-04)	5.00e-04 (4.58e-04, 5.41e-04)	4.99e-04 (4.78e-04, 5.20e-04)
13	1.08e-03 (4.80e-04)	1.23e-03 (4.32e-04, 2.49e-03)	1.09e-03 (9.07e-04, 1.28e-03)	1.08e-03 (9.93e-04, 1.18e-03)
14	7.29e-04 (2.10e-04)	6.33e-04 (3.65e-04, 9.72e-04)	7.24e-04 (6.41e-04, 8.08e-04)	7.28e-04 (6.89e-04, 7.70e-04)
15	8.75e-04 (1.47e-04)	8.52e-04 (5.89e-04, 1.15e-03)	8.75e-04 (8.18e-04, 9.32e-04)	8.75e-04 (8.45e-04, 9.02e-04)
16	8.80e-04 (1.65e-04)	8.67e-04 (5.71e-04, 1.22e-03)	8.80e-04 (8.16e-04, 9.44e-04)	8.80e-04 (8.49e-04, 9.11e-04)

Table C.2, cont. Posterior clock rates across each NRR from the clock calibration using SARS-CoV-2 genomes and Bayesian divergence time analyses of SARS-CoV-2-like viruses. The clock calibration posterior results are used as a prior for the Bayesian divergence time analyses, with the standard deviation of the clock calibration results varying across the analyses.

NRR	**Clock calibration posterior rate**	**Bayesian divergence time estimation**		
		Primary analysis	**Robustness analysis**	
		(Clock calibration-centred prior)	**(Clock calibration-centred prior with stdev / 5)**	**(Clock calibration-centred prior with stdev / 10)**
17	1.18e-03 (2.79e-04)	1.05e-03 (6.46e-04, 1.52e-03)	1.17e-03 (1.07e-03, 1.29e-03)	1.18e-03 (1.12e-03, 1.23e-03)
18	8.73e-04 (1.76e-04)	8.15e-04 (5.44e-04, 1.13e-03)	8.70e-04 (8.05e-04, 9.39e-04)	8.72e-04 (8.39e-04, 9.08e-04)
19	8.59e-04 (2.21e-04)	9.32e-04 (5.41e-04, 1.40e-03)	8.60e-04 (7.74e-04, 9.50e-04)	8.59e-04 (8.19e-04, 9.04e-04)
20	7.75e-04 (2.17e-04)	8.62e-04 (4.15e-04, 1.39e-03)	7.77e-04 (6.97e-04, 8.68e-04)	7.76e-04 (7.33e-04, 8.19e-04)
21	1.12e-03 (2.26e-04)	1.15e-03 (7.77e-04, 1.65e-03)	1.12e-03 (1.03e-03, 1.21e-03)	1.12e-03 (1.07e-03, 1.16e-03)
22	9.35e-04 (2.23e-04)	9.35e-04 (5.56e-04, 1.44e-03)	9.36e-04 (8.52e-04, 1.03e-03)	9.35e-04 (8.92e-04, 9.77e-04)
23	1.28e-03 (2.81e-04)	1.21e-03 (7.72e-04, 1.73e-03)	1.28e-03 (1.17e-03, 1.38e-03)	1.28e-03 (1.23e-03, 1.33e-03)
24	1.19e-03 (1.79e-04)	1.25e-03 (8.86e-04, 1.62e-03)	1.19e-03 (1.12e-03, 1.26e-03)	1.19e-03 (1.15e-03, 1.22e-03)
25	1.38e-03 (3.73e-04)	1.58e-03 (9.14e-04, 2.41e-03)	1.39e-03 (1.25e-03, 1.55e-03)	1.38e-03 (1.31e-03, 1.46e-03)
26	1.52e-03 (4.54e-04)	1.28e-03 (7.00e-04, 1.94e-03)	1.50e-03 (1.35e-03, 1.69e-03)	1.52e-03 (1.43e-03, 1.61e-03)
27	1.74e-03 (2.88e-04)	1.69e-03 (1.22e-03, 2.25e-03)	1.74e-03 (1.63e-03, 1.85e-03)	1.74e-03 (1.69e-03, 1.80e-03)

Table C.3. Posterior times (median and 95% HPD) of the closest-inferred ancestor of SARS-CoV-1 across each NRR. The SARS-CoV-1 clock calibration posterior serves as a prior for the analyses, with the standard deviation of the clock calibration results varying across the analyses and one analysis having an one additional civet SARS-CoV-1 genome and one additional human SARS-CoV-1 genome.

NRR	Primary analysis	Robustness analyses		
	Clock calibration-centred prior	Clock calibration-centred prior with stdev / 5	Clock calibration-centred prior with stdev / 10	Clock calibration-centred prior with additional genomes
1	1996 (1990-2000)	1998 (1994-2001)	1998 (1994-2001)	1995 (1989-2000)
2	1992 (1980-2000)	1996 (1991-2001)	1997 (1991-2001)	1991 (1978-1999)
3	1991 (1972-2001)	1991 (1974-2001)	1991 (1975-2001)	1987 (1969-1999)
4	1993 (1982-2002)	1998 (1991-2002)	1998 (1991-2002)	1993 (1981-2000)
5	1994 (1987-2000)	1997 (1990-2001)	1997 (1991-2001)	1994 (1986-2000)
6	1999 (1993-2002)	2000 (1997-2002)	2001 (1997-2002)	2001 (1997-2002)
7	2001 (1997-2002)	2001 (1998-2002)	2001 (1998-2002)	1999 (1994-2002)
8	1985 (1960-2000)	1999 (1994-2002)	1999 (1994-2002)	1984 (1960-1999)
9	1984 (1962-2001)	1987 (1960-2002)	1988 (1971-2001)	1976 (1952-1996)
10	1999 (1993-2002)	2000 (1994-2002)	2000 (1994-2002)	1997 (1990-2002)
11	2000 (1995-2002)	2001 (1998-2002)	2001 (1998-2002)	1999 (1994-2002)
12	1995 (1986-2001)	1999 (1994-2002)	1999 (1994-2002)	1997 (1989-2001)
13	1997 (1986-2002)	2000 (1995-2002)	2000 (1995-2002)	1997 (1987-2001)
14	2001 (1998-2002)	2001 (1999-2002)	2001 (1999-2002)	2000 (1995-2002)
15	1992 (1977-2001)	1969 (1930-1998)	1974 (1930-2001)	1970 (1925-1997)
16	1999 (1993-2002)	2000 (1997-2002)	2000 (1995-2002)	1998 (1991-2002)
17	1997 (1991-2001)	1998 (1994-2002)	1992 (1980-2001)	1991 (1977-2000)
18	1989 (1971-2000)	1986 (1961-1999)	1988 (1969-1999)	1989 (1972-1999)
19	1963 (1934-1985)	1980 (1963-1992)	1980 (1964-1991)	1963 (1934-1984)
20	1978 (1953-1998)	1985 (1966-2001)	1986 (1965-2001)	1979 (1952-1997)
21	1977 (1948-1995)	1989 (1967-2001)	1989 (1968-2001)	1975 (1947-1995)
22	1979 (1940-2000)	1988 (1963-2001)	1989 (1962-2002)	1980 (1952-1997)
23	1991 (1966-2002)	1995 (1975-2002)	1994 (1975-2002)	1989 (1941-2002)
24	1999 (1989-2002)	2000 (1992-2002)	2000 (1992-2002)	1996 (1969-2002)
25	1999 (1995-2002)	2000 (1996-2002)	2000 (1996-2002)	1998 (1994-2001)
26	2000 (1995-2002)	2001 (1997-2002)	2001 (1997-2002)	1996 (1986-2002)
27	2000 (1994-2002)	2000 (1996-2002)	2000 (1997-2002)	1998 (1989-2001)
28	1999 (1992-2002)	1999 (1994-2002)	1999 (1994-2002)	1998 (1990-2002)
29	1996 (1986-2002)	1998 (1990-2002)	1998 (1991-2002)	1994 (1984-2001)
30	1989 (1945-2002)	1996 (1971-2002)	1996 (1969-2002)	1982 (1863-2001)
31	2001 (1999-2002)	2001 (2000-2002)	2001 (2000-2002)	2000 (1997-2002)

Table C.4. Posterior times (median and 95% HPD) of the closest-inferred ancestor of SARS-CoV-2 across each NRR. The SARS-CoV-2 clock calibration posterior serves as a prior for the analyses, with the standard deviation of the clock calibration results varying across the analyses.

NRR	Primary analysis	Robustness analyses	
	Clock calibration-centred prior	Clock calibration-centred prior with stdev / 5	Clock calibration-centred prior with stdev / 10
1	2011 (2006-2015)	2012 (2007-2015)	2012 (2008-2015)
2	2013 (2007-2017)	2012 (2007-2016)	2012 (2008-2016)
3	2017 (2014-2019)	2017 (2014-2019)	2017 (2014-2019)
4	2010 (2000-2017)	2010 (2002-2016)	2010 (2002-2017)
5	2013 (2008-2017)	2014 (2010-2018)	2014 (2010-2017)
6	2015 (2009-2019)	2015 (2009-2018)	2015 (2010-2019)
7	2014 (2008-2018)	2013 (2009-2017)	2013 (2009-2017)
8	2004 (1979-2016)	2002 (1984-2014)	2003 (1984-2014)
9	1994 (1978-2004)	1992 (1981-2001)	1992 (1981-2001)
10	2008 (1997-2016)	2008 (1999-2016)	2008 (1997-2015)
11	2010 (2004-2014)	2011 (2007-2014)	2011 (2007-2014)
12	1997 (1976-2011)	1996 (1979-2009)	1996 (1979-2009)
13	2010 (1984-2019)	2010 (1989-2019)	2010 (1991-2019)
14	1997 (1982-2007)	1999 (1991-2005)	1999 (1991-2005)
15	2012 (2007-2015)	2012 (2008-2015)	2012 (2008-2015)
16	2003 (1993-2010)	2003 (1996-2009)	2003 (1995-2008)
17	2012 (2006-2017)	2013 (2008-2017)	2013 (2008-2017)
18	2004 (1996-2010)	2005 (2000-2009)	2005 (1999-2009)
19	2012 (2005-2017)	2011 (2006-2016)	2011 (2006-2016)
20	2004 (1988-2014)	2002 (1991-2011)	2002 (1991-2011)
21	1992 (1977-2003)	1991 (1980-2000)	1991 (1980-1999)
22	1982 (1946-2007)	1982 (1955-2004)	1982 (1955-2004)
23	1989 (1969-2007)	1991 (1978-2006)	1990 (1977-2005)
24	1992 (1981-2001)	1991 (1984-1997)	1991 (1983-1997)
25	2006 (1993-2015)	2004 (1993-2013)	2004 (1994-2013)
26	2009 (2000-2016)	2010 (2004-2016)	2010 (2004-2016)
27	2013 (2008-2016)	2013 (2008-2016)	2013 (2008-2016)

Table C.5. Comparison of the weighted diffusion coefficient and isolation-by-distance signal estimated for different data sets of viral genomes. The isolation-by-distance (IBD) signal has been estimated by computing the Spearman correlation coefficient (rS) between the patristic and great-circle geographic distances computed for each pair of virus samples. For each data set and metric, we report both the posterior median estimate and the 95% HPD interval. (*) estimates based on the combined analysis of lineages L1 and L3. (**) estimates obtained while only considering phylogenetic branches occurring less than 50 years ago and bat samples, but similar results are obtained when considering branches occurring less than 75 years ago (2839 km2/year [1595, 4558] for SARS-CoV-1-like viruses, and 2536 km2/year [638, 6089] for SARS-CoV-2-like viruses) or when considering all samples for SARS-CoV-1-like viruses (3030 km2/year [1965, 4668]). When considering all samples for SARS-CoV-2-like viruses, we however obtain a higher estimate for the weighted diffusion coefficient, but with heavily overlapping confidence intervals and the same order of magnitude (3817 km2/year [1824, 6929]).

Data set	**Weighted diffusion coefficient (km2/year)**	**IBD signal (rS)**	**Number of samples**	**Reference**
Porcine deltacoronavirus, China	26093 [17312, 37020]	-0.02 [-0.07, 0.05]	97	He *et al.* (2020)
Getah virus, China	24930 [17612, 35504]	0.06 [0.00, 0.14]	125	Zhao *et al.* (2023)
West Nile virus, North America	58757 [55115, 62060]	0.00 [-0.02, 0.03]	801	Dellicour *et al.* (2020)
Lassa virus, segment S, Africa	42 [38, 49]	0.68 [0.64, 0.72]	254	Klitting *et al.* (2022)
Rabies virus (bats)*, Peru	1416 [1105, 1852]	0.73 [0.55, 0.83]	260	Streicker *et al.* (2016)
Rabies virus (skunks), USA	580 [475, 676]	0.49 [0.48, 0.60]	241	Kuzmina *et al.* (2013)
SARS-CoV-1-like viruses (bats)	2884 [1654, 4358]**	0.45 [0.36, 0.59]	134	(present study)
SARS-CoV-2-like viruses (bats)	2050 [233, 5470]**	0.66 [0.12, 0.83]	20	(present study)

C.3 Supplementary Figures

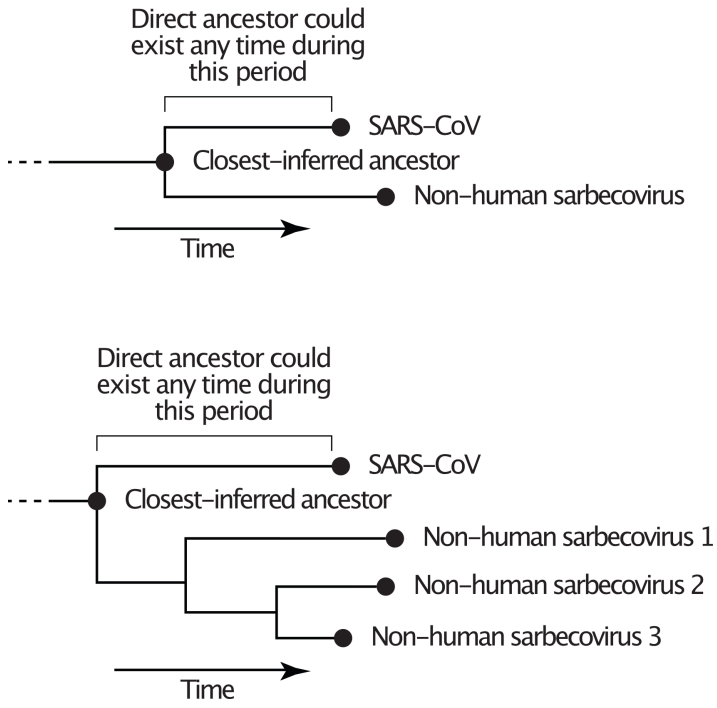


Figure C.1. Schematics illustrating the relationship between a SARS-CoV, the direct ancestor, and the closest-inferred ancestor. The closest-inferred ancestor could be the tMRCA of a given SARS-CoV and one non-human sarbecovirus (top) or the tMRCA of the SARS-CoV and multiple non-human sarbecoviruses (bottom); the closest-inferred ancestor is the parent node of the SARS-CoV. The direct ancestor is the bat sarbecovirus that was involved in each cross-species transmission event leading to the SARS-CoV and could have existed at any point on the branch leading from the closest-inferred ancestor to the SARS-CoV.

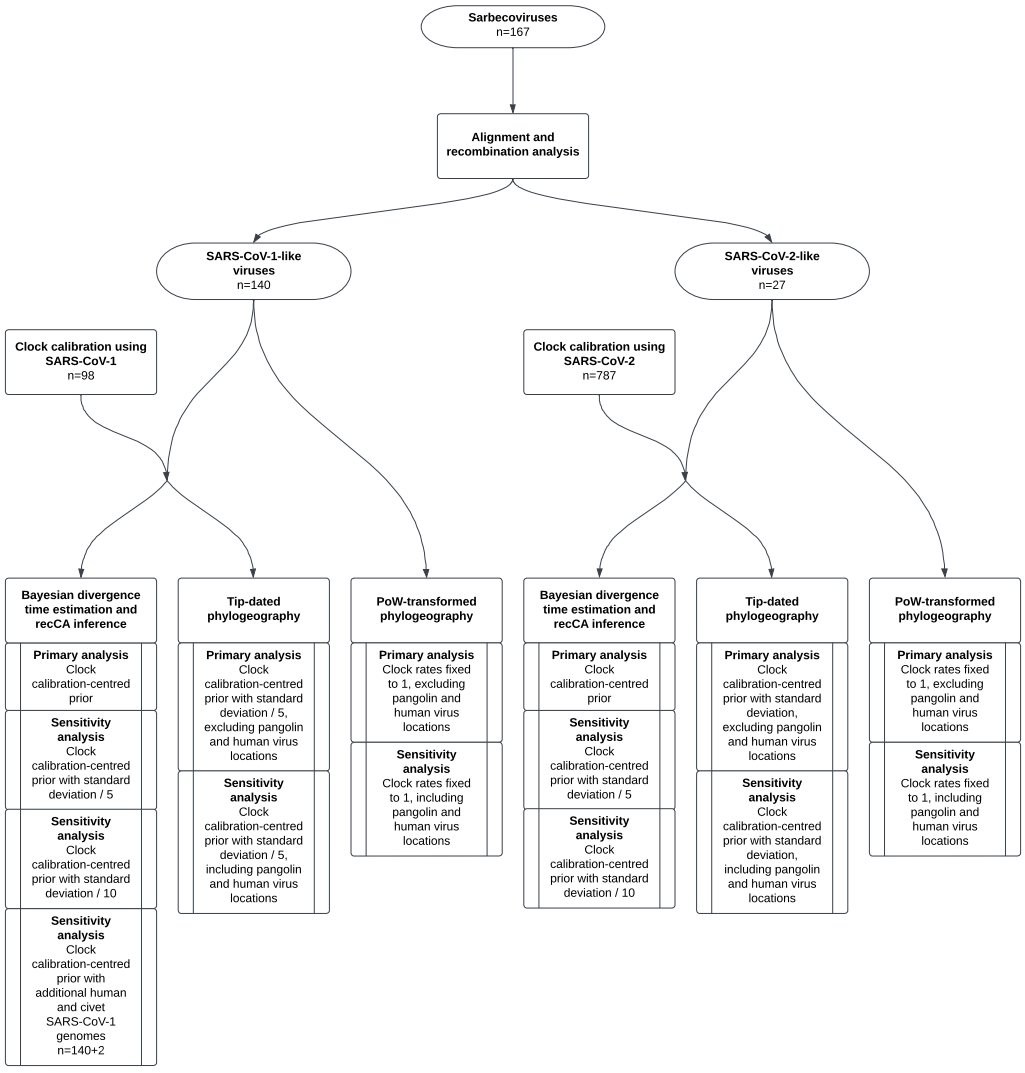


Figure C.2. Workflow to perform phylogenetic and phylogeographic analyses. Starting with 167 sarbecovirus genomes, we align and perform recombination analysis to determine the number of SARS-CoV-1-like and SARS-CoV-2-like viruses and their corresponding NRRs for subsequent analyses. We then calibrate the molecular clock for SARS-CoV-1-like viruses and SARS-CoV-2-like viruses using SARS-CoV-1 genomes and SARS-CoV-2 genomes (these datasets are separate from the SARS-CoV-1-like viruses and SARS-CoV-2-like viruses), respectively. Finally, we perform Bayesian divergence time estimation, recCA inference, and phylogeography analyses, each of which comprised both primary and sensitivity analyses. Ovals indicate the number of sarbecovirus genomes included in subsequent steps. Rectangles without inner vertical lines indicate analyses performed; rectangles with inner vertical lines indicate primary versus sensitivity analyses, with non-bolded text describing the details of the analysis performed.

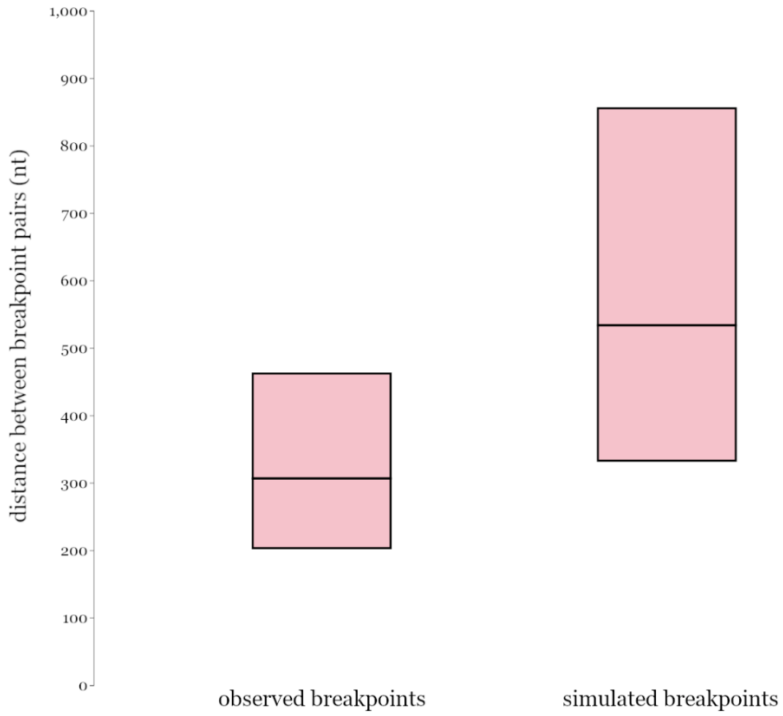


Figure C.3. Mean and 90% confidence intervals of the log-normal breakpoint pair distance distributions for the observed SARS-CoV-1-like and SARS-CoV-2-like estimated breakpoints and the simulated randomly sampled breakpoint sets.

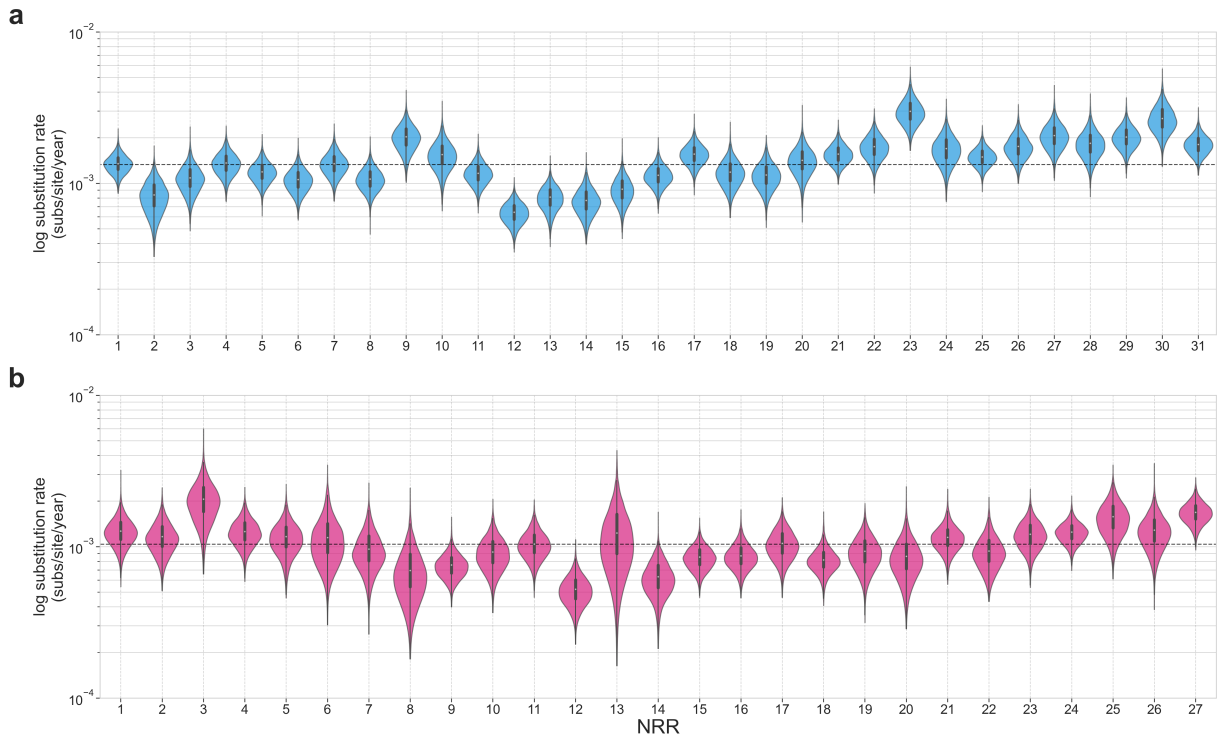


Figure C.4. Substitution rates for the SARS-CoV-1-like and SARS-CoV-2-like viruses. Substitution rates (substitutions/site/year) across the (a) 31 NRRs for the SARS-CoV-1-like viruses and (b) 27 NRRs for the SARS-CoV-2-like viruses. The dashed line in each panel is the respective median substitution rate. The dots and thick lines within the violins indicate the median and interquartile range.

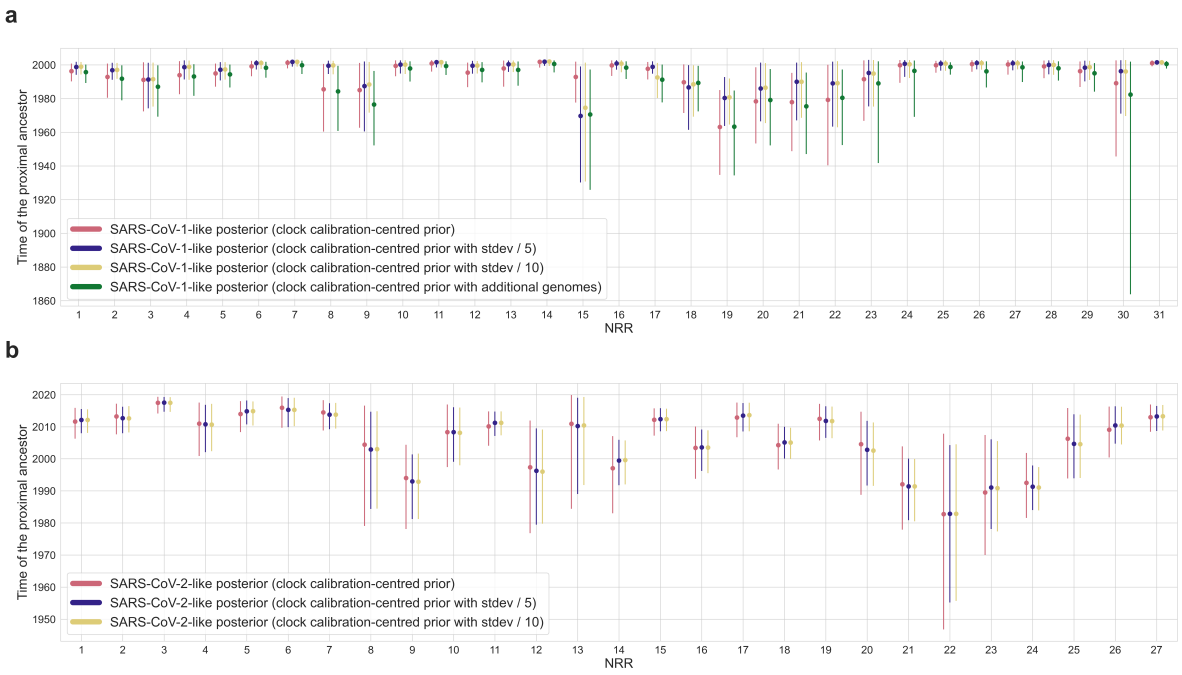


Figure C.5. Time of the closest-inferred ancestor sensitivity analysis. a, Time of the closest-inferred ancestor of SARS-CoV-1 across the 31 NRRs of the SARS-CoV-1-like viruses with the calibrated molecular clock from the SARS-CoV-1 analysis, the calibrated molecular clock but the standard deviation of the clock rate divided by 5, the calibrated molecular clock but the standard deviation of the clock rate divided by 10, and when adding in one SARS-CoV-1 genome from a civet and an additional SARS-CoV-1 genome from a human. b, Time of the closest-inferred ancestor of SARS-CoV-2 across the 27 NRRs of the SARS-CoV-2-like viruses with the calibrated molecular clock from the SARS-CoV-2 analysis, the calibrated molecular clock but the standard deviation of the clock rate divided by 5, and the calibrated molecular clock but the standard deviation of the clock rate divided by 10. The dots indicate the median and the lines are the 95% HPD. Refer to Supplementary Tables 3 and 4 for numerical values.

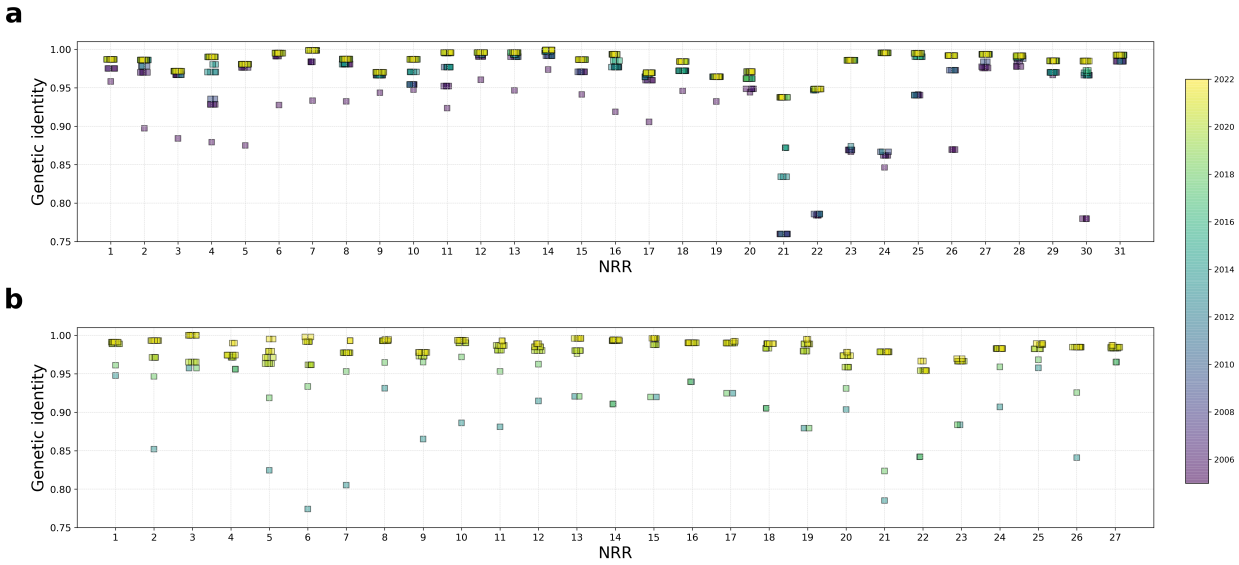


Figure C.6. Genetic similarity of the NRRs of the recCA to each SARS-CoV over time. a, The similarity of each NRR of the recCA of SARS-CoV-1 to SARS-CoV-1 as a function of time. b, The similarity of each NRR of the recCA of SARS-CoV-2 to SARS-CoV-2 as a function of time. Each square represents the similarity of the NRR of the recCA to its relative SARS-CoV at one point in time and is colored by date of the most recent non-human sarbecovirus in the reconstruction. See Fig. 3.2 for the similarity of the complete recCA of each SARS-CoV to its relative SARS-CoV over time.

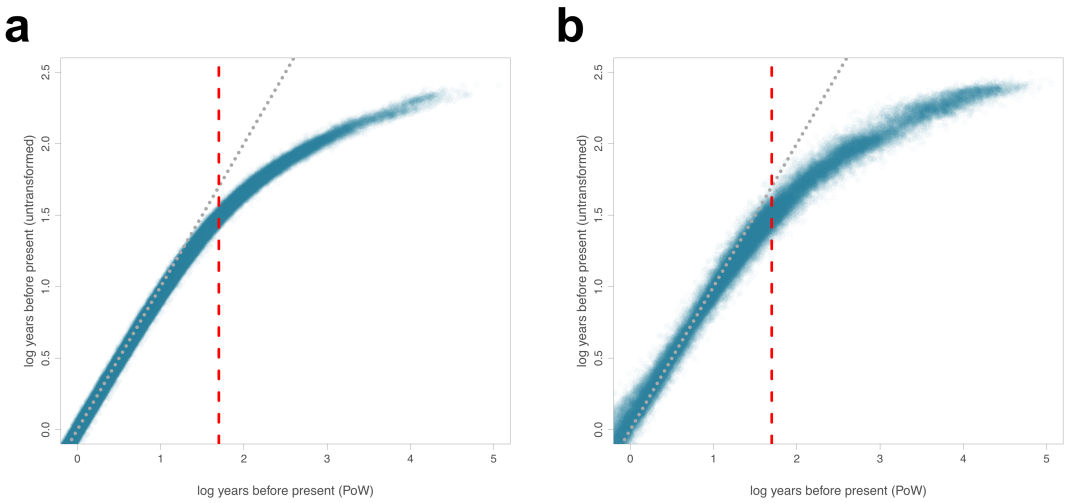


Figure C.7. Prisoner of War (PoW) transformation comparison. The log number of years of each internal node before the most recently sampled taxon compared to the log number of years of the same node after applying a PoW transformation for (a) 100 SARS-CoV-1-like posterior phylogenies for each NRR and (b) 100 SARS-CoV-2-like posterior phylogenies for each NRR. Red dashed line at 50 years before the most recently sampled taxon indicates a threshold beyond which saturation effects substantially influence tMRCA estimates. Diagonal dotted line depicts a line of equality, with a one-to-one correspondence between the two inferences on the two axes.

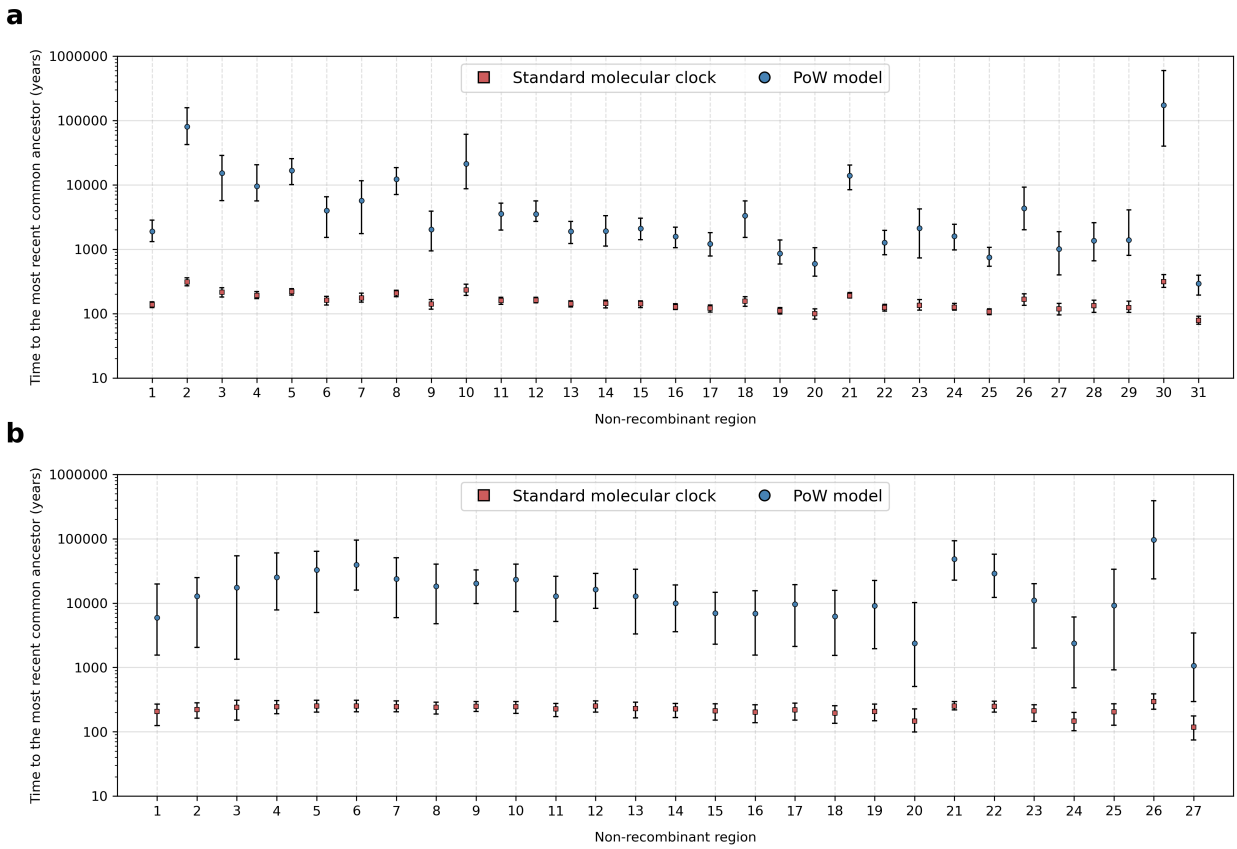


Figure C.8. Time to the most recent common ancestor (tMRCA) after PoW transformation. Estimates of time to the most recent common ancestor for the (a) SARS-CoV-1-like viruses and (b) SARS-CoV-2-like viruses for each NRR when using a standard molecular clock or the PoW model. Each square indicates the median tMRCA and the lines indicate the 95% HPD.

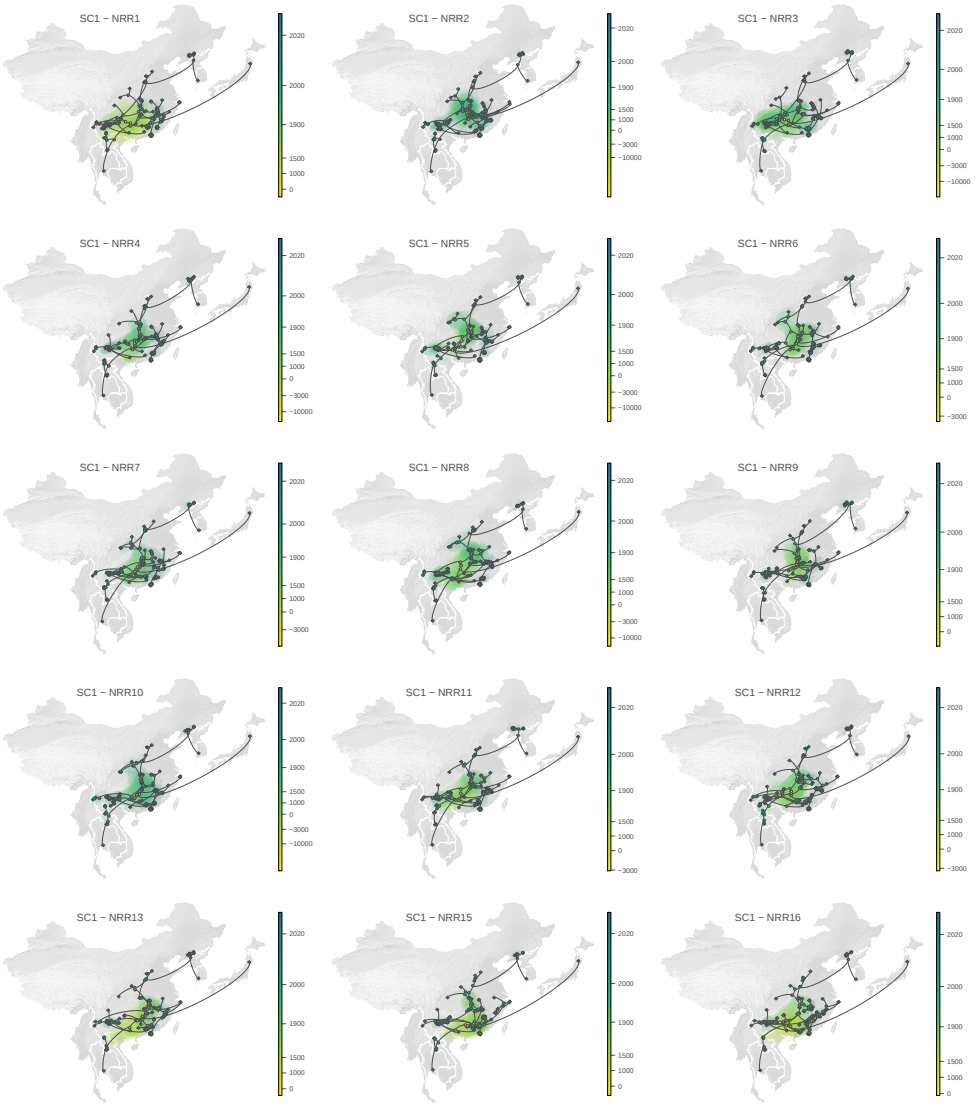


Figure C.9. Continuous phylogeographic reconstruction of the dispersal history of viral lineages for the SARS-CoV-1-like (SC1) viruses. We report the reconstructions obtained for NRRs 1-13 and 15-16 (NRR14 being reported in the main Fig. 3.3g; see Fig. C.10 for NRRs 17-31). For each NRR, we map the maximum clade credibility (MCC) tree whose nodes are coloured from yellow (the time of the most recent common ancestor) to blue (most recent collection date) and are superimposed on 80% HPD polygons coloured according to the same log-transformed time scale and reflecting the uncertainty of the Bayesian phylogeographic inference. HPD polygons were only computed and reported for the last 1,000 years. As in Fig. 3.3g-h, MCC tree tip nodes are displayed as squares (as opposed to dots for internal ones) and dispersal direction (anti-clockwise) of viral lineages is indicated by the edge curvature.

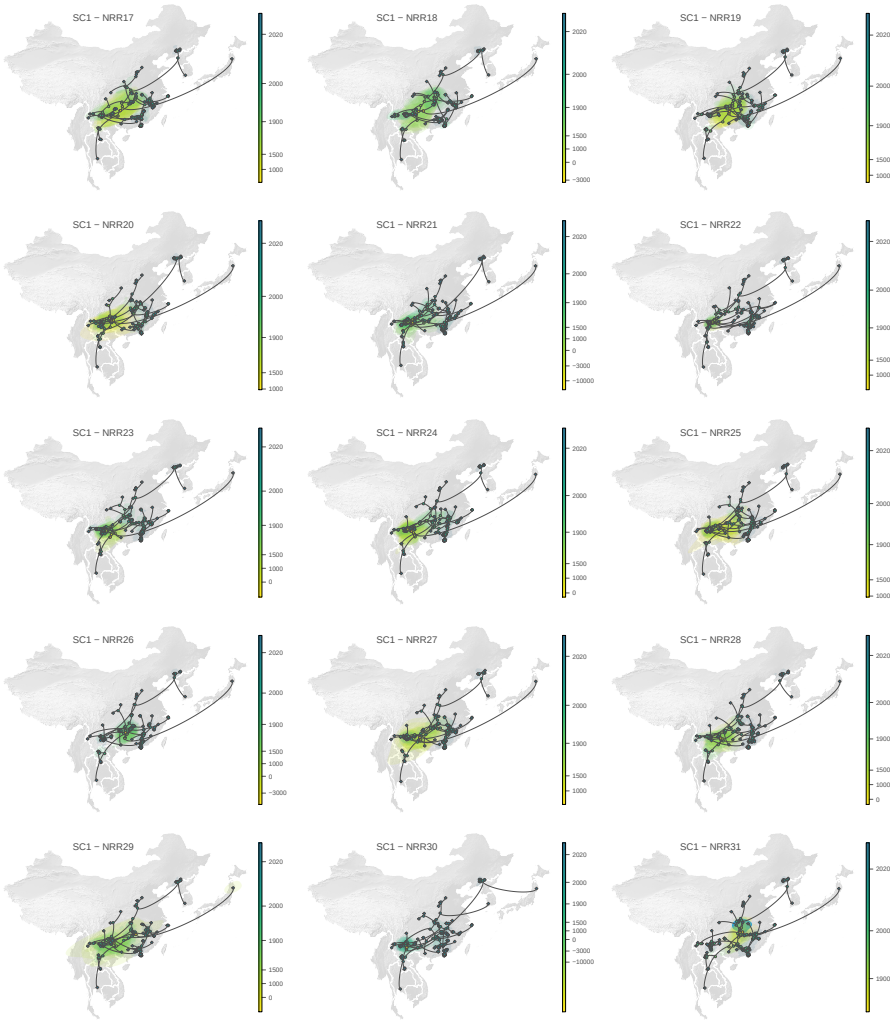


Figure C.10. Continuous phylogeographic reconstruction of the dispersal history of viral lineages for the SARS-CoV-1-like (SC1) viruses. We report the reconstructions obtained for NRRs 17-31 (see Fig. C.9 for NRRs 1-13 and 15-17, NRR14 being reported in the main Fig. 3.3g). For each NRR, we map the maximum clade credibility (MCC) tree whose nodes are coloured from yellow (the time of the most recent common ancestor) to blue (most recent collection date) and are superimposed on 80% HPD polygons coloured according to the same log-transformed time scale and reflecting the uncertainty of the Bayesian phylogeographic inference. HPD polygons were only computed and reported for the last 1,000 years. As in Fig. 3.3g-h, MCC tree tip nodes are displayed as squares (as opposed to dots for internal ones) and dispersal direction (anti-clockwise) of viral lineages is indicated by the edge curvature.

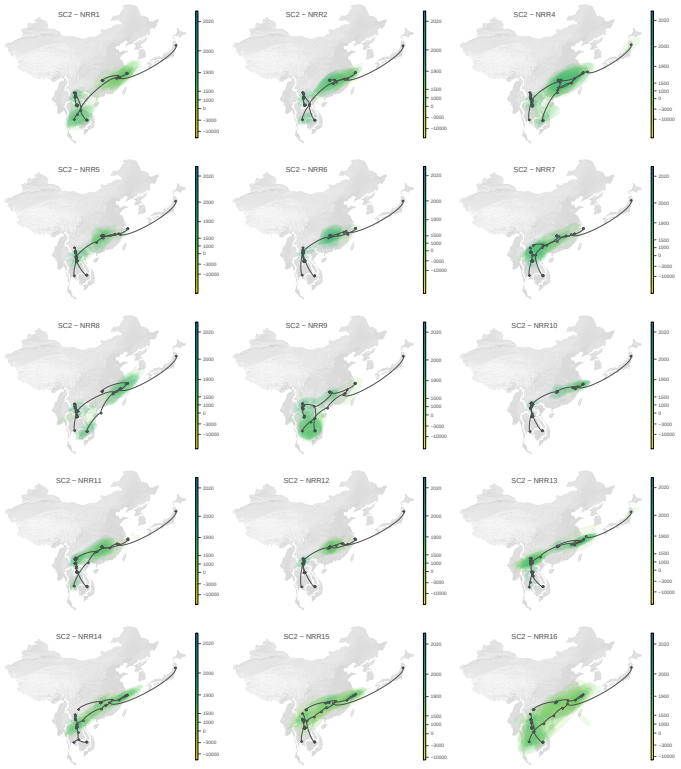


Figure C.11. Continuous phylogeographic reconstruction of the dispersal history of viral lineages for the SARS-CoV-2-like (SC2) viruses. We report the reconstructions obtained for NRRs 1-2 and 4-16 (NRR3 being reported in the main Fig. 3.3h; see Fig. C.12 for NRRs 17-27). For each NRR, we map the maximum clade credibility (MCC) tree whose nodes are coloured from yellow (the time of the most recent common ancestor) to blue (most recent collection date) and are superimposed on 80% HPD polygons coloured according to the same log-transformed time scale and reflecting the uncertainty of the Bayesian phylogeographic inference. HPD polygons were only computed and reported for the last 1,000 years. As in Fig. 3.3g-h, MCC tree tip nodes are displayed as squares (as opposed to dots for internal ones) and dispersal direction (anti-clockwise) of viral lineages is indicated by the edge curvature.

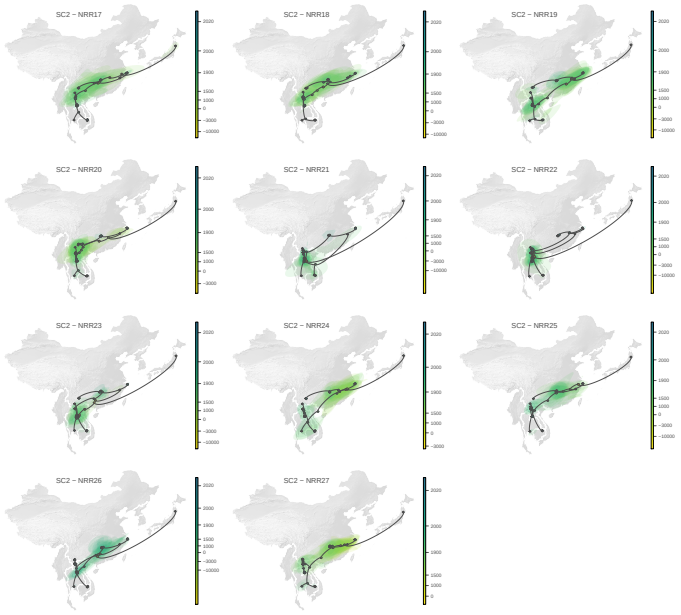


Figure C.12. Continuous phylogeographic reconstruction of the dispersal history of viral lineages for the SARS-CoV-2-like (SC2) viruses. We report the reconstructions obtained for NRRs 17-27 (see Fig. C.11 for NRRs 1-2 and 4-16, NRR3 being reported in the main Fig. 3.3h). For each NRR, we map the maximum clade credibility (MCC) tree whose nodes are coloured from yellow (the time of the most recent common ancestor) to blue (most recent collection date) and are superimposed on 80% HPD polygons coloured according to the same log-transformed time scale and reflecting the uncertainty of the Bayesian phylogeographic inference. HPD polygons were only computed and reported for the last 1,000 years. As in Fig. 3.3g-h, MCC tree tip nodes are displayed as squares (as opposed to dots for internal ones) and dispersal direction (anti-clockwise) of viral lineages is indicated by the edge curvature.



Figure C.13. Distribution maps for species of the *Rhinolophus* genus, part 1. See Fig. C.14-C.15 for remaining bat distribution maps.

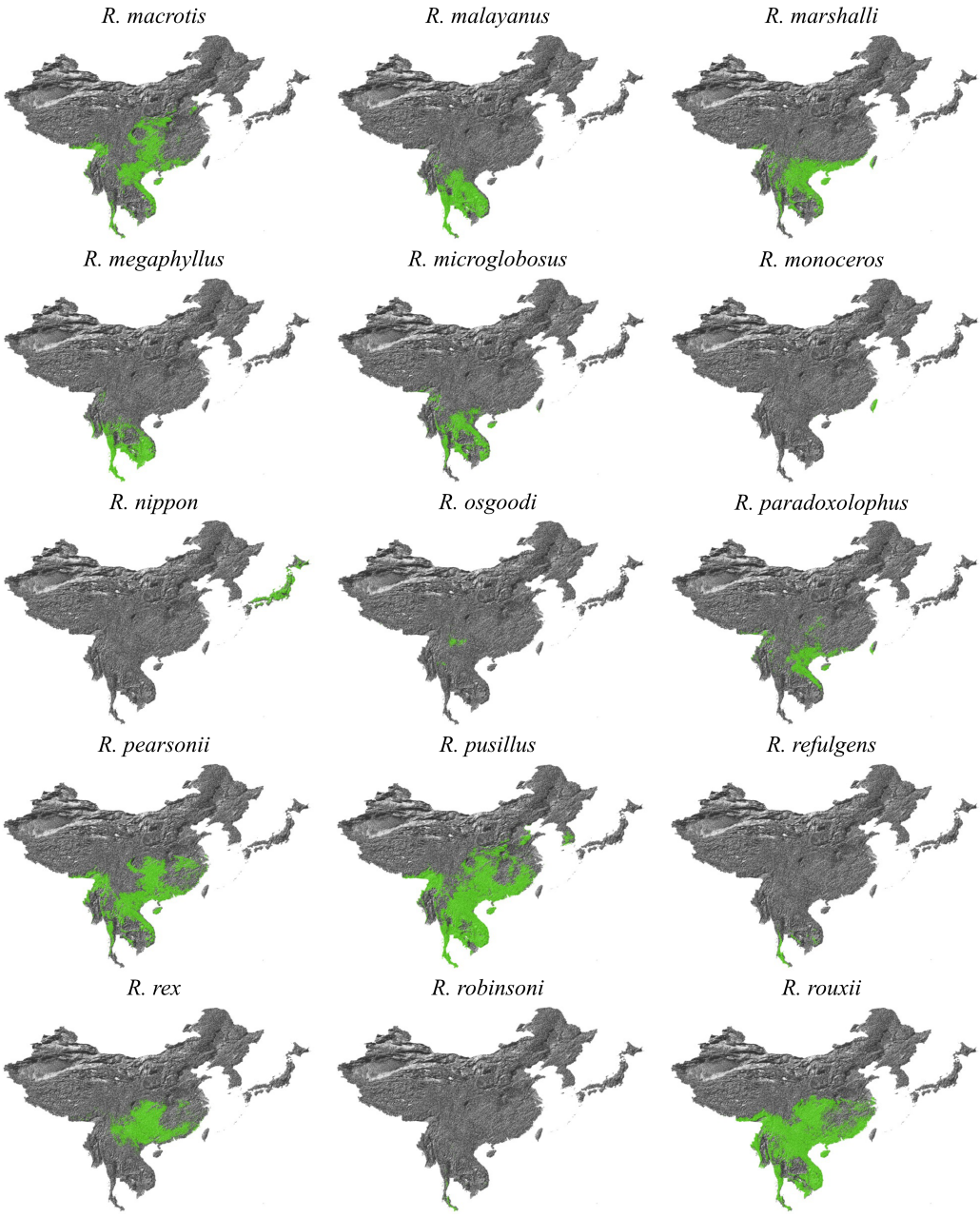


Figure C.14. Distribution maps for species of the *Rhinolophus* genus, part 2. See Fig. C.13 and C.15 for remaining bat distribution maps.

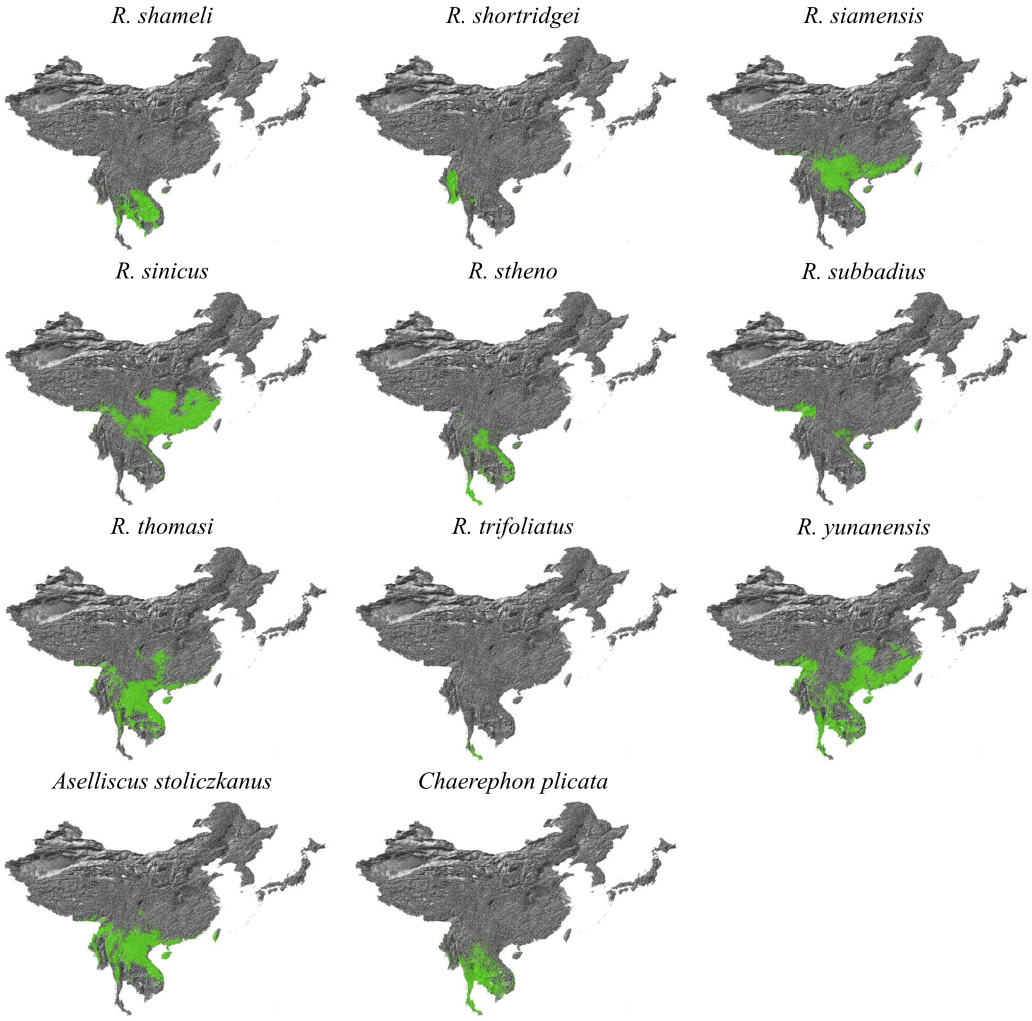


Figure C.15. Distribution maps for species of the *Rhinolophus*, *Ascelliscus*, and *Chaerephon* genera, part 3. Species from the *Ascelliscus* and *Chaerephon* genera are limited to those that were hosts to sarbecoviruses in our analyses. See Fig. C.13–C.14 for remaining bat distribution maps.

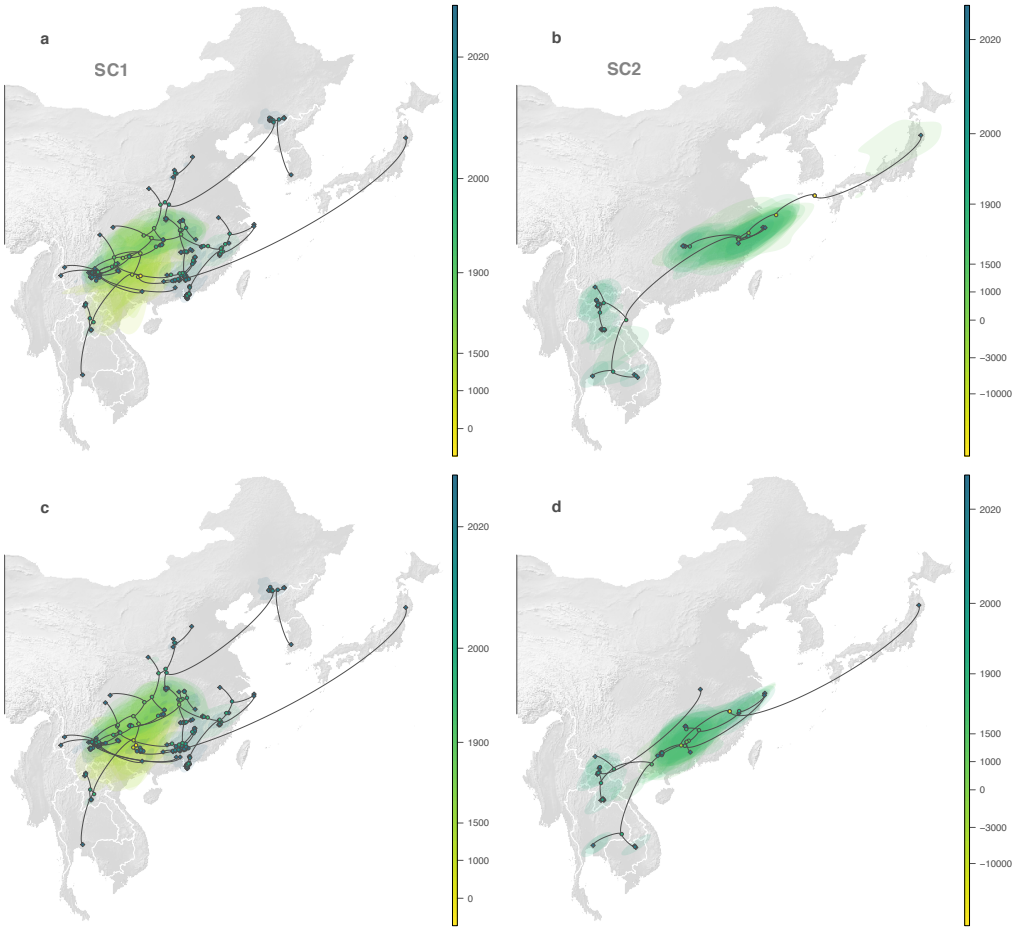


Figure C.16. Comparison of the continuous phylogeographic reconstructions for two selected NRRSs (NRR14 for SARS-CoV-1-like viruses [SC1] on the left, and NRR3 for SARS-CoV-2-like viruses [SC2] on the right) when not considering (a-b) or when considering (c-d) the human and pangolin sequences in the analyses. Comparison of the continuous phylogeographic reconstructions for two selected NRRSs (NRR14 for SARS-CoV-1-like viruses [SC1] on the left, and NRR3 for SARS-CoV-2-like viruses [SC2] on the right) when not considering (a-b) or when considering (c-d) the human and pangolin sequences in the analyses. For each reconstruction, we map the corresponding MCC tree whose nodes are coloured from yellow (the time of the most recent common ancestor) to blue (most recent collection date) and are superimposed on 80% HPD polygons coloured according to a log-transformed time scale and reflecting the uncertainty of the Bayesian phylogeographic inference. HPD polygons were only computed and reported for the last 1,000 years. As in Fig. 3.3g-h, MCC tree tip nodes are displayed as squares (as opposed to dots for internal ones) and dispersal direction (anti-clockwise) of viral lineages is indicated by the edge curvature.

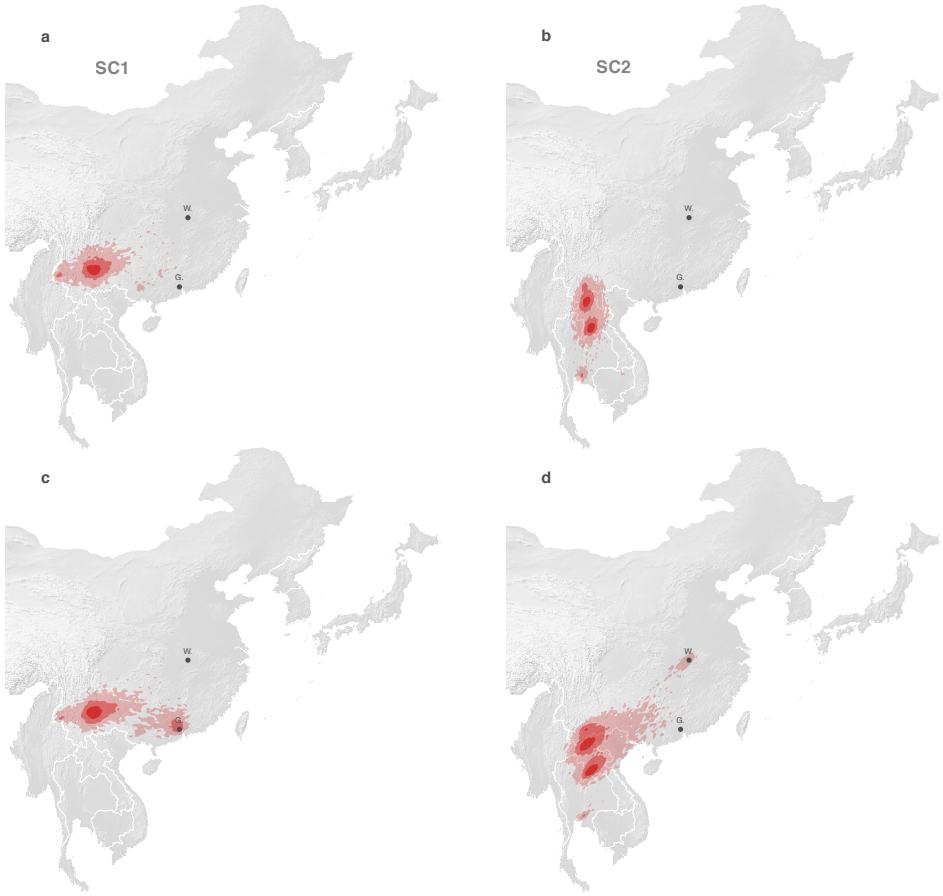


Figure C.17. Comparison of the estimated position of the recCA according to continuous phylogeographic reconstructions that did not consider (a-b) or did consider (c-d) the human and pangolin sequences in the analyses. As in Figure 4g-h, the estimated position of the recCA is based on the continuous phylogeographic reconstruction of all NRRs. We here display the 95%, 75% and 50% HPD regions with an increasing red color darkness. "G." and "W." indicate the position of the cities of Guangzhou and Wuhan, respectively.

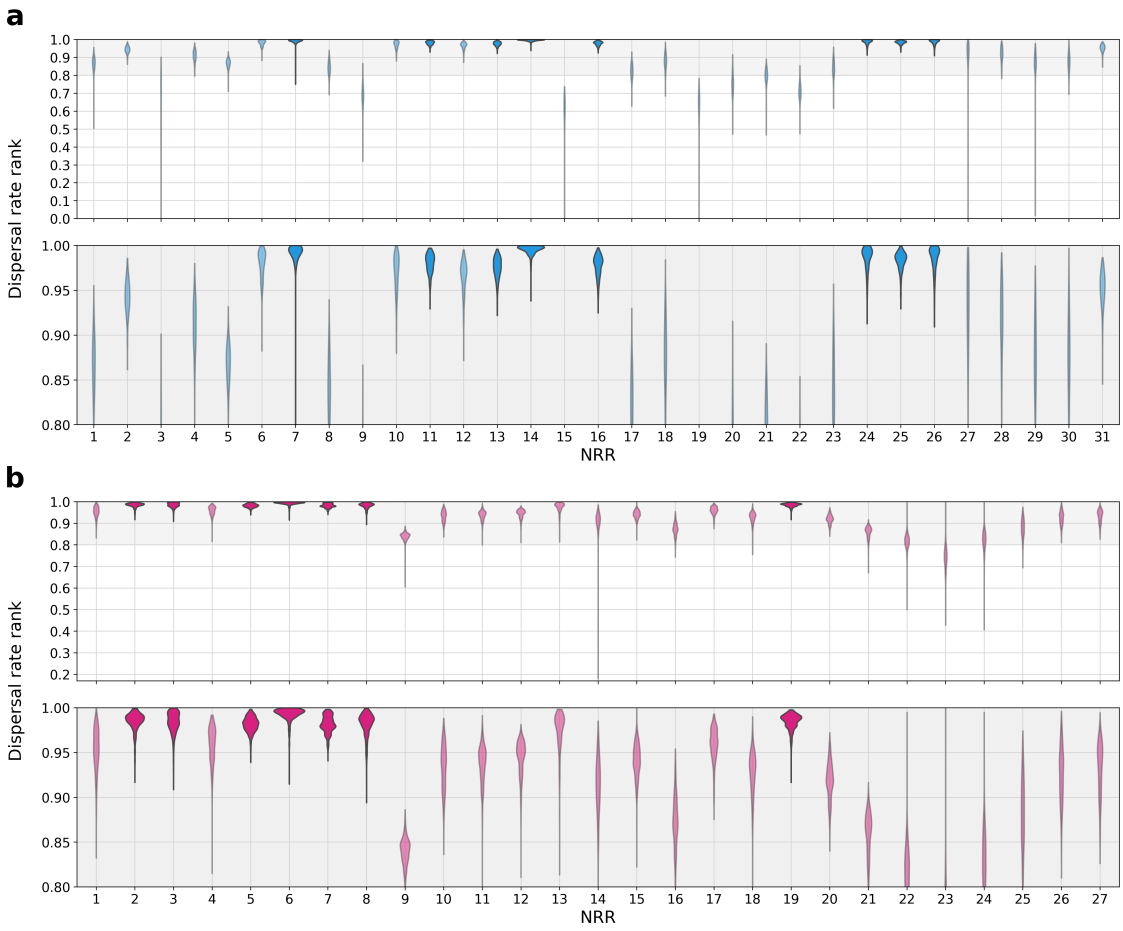


Figure C.18. Ranking of dispersal velocity for the closest-inferred bat virus ancestor to reach the province of emergence. a, The rank for the dispersal velocity necessary to travel from the location of the closest-inferred bat virus ancestor of SARS-CoV-1 from each NRR to the nearest point in Guangdong Province, given the time between emergence and this ancestor and the bat virus dispersal velocities for each NRR. b, The rank for the dispersal velocity necessary to travel from the location of the closest-inferred bat virus ancestor of SARS-CoV-2 from each NRR to the nearest point in Hubei Province, given the time between emergence and this ancestor and the bat virus dispersal velocities for each NRR. The top 20% of the dispersal rate ranks are zoomed-in upon in the lower panels of a and b. Violins are darkened if at least 95% of the dispersal velocity rank posterior is greater than 0.95.

Bibliography

[1] N Zhu et al. A novel coronavirus from patients with pneumonia in china, 2019. *N. Engl. J. Med.*, 382:727–733, 2020.

[2] Q Li et al. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *N. Engl. J. Med.*, 382:1199–1207, 2020.

[3] Livia V Patrono, Bram Vrancken, Matthias Budt, Ariane D"ux, Sebastian Lequime, Seng"ul Boral, M Thomas P Gilbert, Jan F Gogarten, Luisa Hoffmann, David Horst, et al. Archival influenza virus genomes from europe reveal genomic variability during the 1918 pandemic. *Nature communications*, 13(1):2314, 2022.

[4] Michael Worobey, Guan-Zhu Han, and Andrew Rambaut. Genesis and pathogenesis of the 1918 pandemic h1n1 influenza a virus. *Proceedings of the National Academy of Sciences*, 111(22):8107–8112, 2014.

[5] Michael Worobey, Guan-Zhu Han, and Andrew Rambaut. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature*, 508(7495):254–257, 2014.

[6] NR Faria et al. The early spread and epidemic ignition of hiv-1 in human populations. *Science*, 346:56–61, 2014.

[7] M Worobey et al. Direct evidence of extensive diversity of hiv-1 in kinshasa by 1960. *Nature*, 455:661–664, 2008.

[8] Sophie Gryseels, Thomas D Watts, Jean-Marie Kabongo Mpolesha, Brendan B Larsen, Philippe Lemey, Jean-Jacques Muyembe-Tamfum, Dirk E Teuwen, and Michael Worobey. A near full-length hiv-1 genome from 1966 recovered from formalin-fixed paraffin-embedded tissue. *Proceedings of the National Academy of Sciences*, 117(22):12222–12229, 2020.

[9] Michael Worobey, Marlea Gemmel, Dirk E Teuwen, Tamara Haselkorn, Kevin Kunstman, Michael Bunce, Jean-Jacques Muyembe, Jean-Marie M Kabongo, Raphaël M Kalengayi, Eric Van Marck, et al. Direct evidence of extensive diversity of hiv-1 in kinshasa by 1960. *Nature*, 455(7213):661–664, 2008.

[10] Yi Guan, BJ Zheng, YQ He, XL Liu, ZX Zhuang, CL Cheung, SW Luo, Philip H Li, LJ Zhang, YJ Guan, et al. Isolation and characterization of viruses related to the sars coronavirus from animals in southern china. *Science*, 302(5643):276–278, 2003.

[11] Paul M Sharp and Beatrice H Hahn. Origins of hiv and the aids pandemic. *Cold Spring Harbor Perspectives in Medicine:*, 1(1), 2011.

[12] Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the sars coronavirus during the course of the sars epidemic in china. *Science*, 303(5664):1666–1669, 2004.

[13] Edward C Holmes and Andrew Rambaut. Viral evolution and the emergence of sars coronavirus. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1447):1059–1065, 2004.

[14] Michael Worobey, Jim Cox, and Douglas Gill. The origins of the great pandemic. *Evolution, Medicine, and Public Health*, 2019(1):18–25, 2019.

[15] Chi Wai Yip, Chung Chau Hon, Mang Shi, Tommy Tsan-Yuk Lam, Ken Yan-Ching Chow, Fanya Zeng, and Frederick Chi-Ching Leung. Phylogenetic perspectives on the epidemiology and origins of sars and sars-like coronaviruses. *Infection, Genetics and Evolution*, 9(6): 1185–1196, 2009.

[16] KG Andersen et al. The proximal origin of sars-cov-2. *Nat. Med.*, 26:450–452, 2020.

[17] L du Plessis et al. Further musings on the tmrca. *Virological*, 2020. https://virological.org/t/further-musings-on-the-tmrca/340.

[18] Erik M Volz, Katia Koelle, and Trevor Bedford. Viral phylodynamics. *PLoS computational biology*, 9(3):e1002947, 2013.

[19] Matt J. Keeling and Pejman Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton, October 28 2007.

[20] X Hao et al. Reconstruction of the full transmission dynamics of covid-19 in wuhan. *Nature*, 584:420–424, 2020.

[21] J. Giesecke. Primary and index cases. *Lancet*, 384:2024, 2014. doi: 10.1016/S0140-6736(14)62331-X.

[22] M. Worobey. Dissecting the early covid-19 cases in wuhan. *Science*, 374:1202–1204, 2021. doi: 10.1126/science.abm4454.

[23] F. Li, Y.-Y. Li, M.-J. Liu, et al. Household transmission of sars-cov-2 and risk factors for susceptibility and infectivity in wuhan: A retrospective observational study. *Lancet Infect. Dis.*, 21:617–628, 2021. doi: 10.1016/S1473-3099(20)30981-6.

[24] X. Yang, Y. Yu, J. Xu, et al. Clinical course and outcomes of critically ill patients with sars-cov-2 pneumonia in wuhan, china: A single-centered, retrospective, observational study. *Lancet Respir. Med.*, 8:475–481, 2020. doi: 10.1016/S2213-2600(20)30079-5.

[25] C. Huang, Y. Wang, X. Li, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *Lancet*, 395:497–506, 2020. doi: 10.1016/S0140-6736(20)30183-5.

[26] WHO Headquarters. Who-convened global study of origins of sars-cov-2: China part, 2021. URL https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part.

[27] Wendong Li, Zhengli Shi, Meng Yu, Wuze Ren, Craig Smith, Jonathan H Epstein, Hanzhong Wang, Gary Crameri, Zhihong Hu, Huajun Zhang, et al. Bats are natural reservoirs of sars-like coronaviruses. *Science*, 310(5748):676–679, 2005.

[28] Xing-Yi Ge, Jia-Lu Li, Xing-Lou Yang, Aleksei A Chmura, Guangjian Zhu, Jonathan H Epstein, Jonna K Mazet, Ben Hu, Wei Zhang, Cheng Peng, et al. Isolation and characterization of a bat sars-like coronavirus that uses the ace2 receptor. *Nature*, 503(7477):535–538, 2013.

[29] Livia V Patrono, Kamilla Pléh, Liran Samuni, Markus Ulrich, Caroline Röthemeier, Andreas Sachse, Silvia Muschter, Andreas Nitsche, Emmanuel Couacy-Hymann, Christophe Boesch, et al. Monkeypox virus emergence in wild chimpanzees reveals distinct clinical outcomes and viral diversity. *Nature Microbiology*, 5(7):955–965, 2020.

[30] Mirela D'arc, Ahidjo Ayouba, Amandine Esteban, Gerald H Learn, Vanina Boué, Florian Liegeois, Lucie Etienne, Nikki Tagg, Fabian H Leendertz, Christophe Boesch, et al. Origin of the hiv-1 group o epidemic in western lowland gorillas. *Proceedings of the National Academy of Sciences*, 112(11):E1343–E1352, 2015.

[31] Michael Worobey, Paul Telfer, Sandrine Souqui'ere, Meredith Hunter, Clint A Coleman, Michael J Metzger, Patricia Reed, Maria Makuwa, Gail Hearn, Shaya Honarvar, et al. Island biogeography reveals the deep history of siv. *Science*, 329(5998):1487–1487, 2010.

[32] L.-F. Wang, Z. Shi, S. Zhang, H. Field, P. Daszak, and B. T. Eaton. Review of bats and sars. *Emerg. Infect. Dis.*, 12:1834–1840, 2006.

[33] W. Ren, W. Li, M. Yu, P. Hao, Y. Zhang, P. Zhou, S. Zhang, G. Zhao, Y. Zhong, S. Wang, L.-F. Wang, and Z. Shi. Full-length genome sequences of two sars-like coronaviruses in horseshoe bats and genetic variation analysis. *J. Gen. Virol.*, 87:3355–3359, 2006.

[34] S. K. P. Lau, P. C. Y. Woo, K. S. M. Li, Y. Huang, H.-W. Tsoi, B. H. L. Wong, S. S. Y. Wong, S.-Y. Leung, K.-H. Chan, and K.-Y. Yuen. Severe acute respiratory syndrome coronavirus-like virus in chinese horseshoe bats. *Proc. Natl. Acad. Sci. U. S. A.*, 102:14040–14045, 2005.

[35] S. Lytras, J. Hughes, D. Martin, P. Swanepoel, A. de Klerk, R. Lourens, S. L. Kosakovsky Pond, W. Xia, X. Jiang, and D. L. Robertson. Exploring the natural origins of sars-cov-2 in the light of recombination. *Genome Biol. Evol.*, 2022. doi:

10.1093/gbe/evac018.

[36] Report of the who-china joint mission on coronavirus disease 2019 (covid-19). 2020. https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19).

[37] H Fei et al. The impact of the covid-19 epidemic on tuberculosis control in china. *Lancet Reg. Health - West. Pac.*, 3:100032, 2020.

[38] Who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020. 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

[39] C Huang et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet*, 395:497–506, 2020.

[40] X Zhang et al. Viral and host factors related to the clinical outcome of covid-19. *Nature*, 583:437–440, 2020.

[41] F Wu et al. A new coronavirus associated with human respiratory disease in china. *Nature*, 579:265–269, 2020.

[42] R Lu et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet Lond. Engl.*, 395:565–574, 2020.

[43] J Ma. Coronavirus: China's first confirmed covid-19 case traced back to november 17. *South China Morning Post*, 2020. https://www.scmp.com/news/china/society/article/3074991/coronavirus-chinas-first-confirmed-covid-19-case-traced-back.

[44] A Rambaut. Phylodynamic analysis — 176 genomes — 6 mar 2020. *Virological*, 2020. https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356.

[45] S Duchene et al. Temporal signal and the phylodynamic threshold of sars-cov-2. *Virus Evol.*, 6:veaa061, 2020.

[46] BF Keele et al. Chimpanzee reservoirs of pandemic and nonpandemic hiv-1. *Science*, 313: 523–526, 2006.

[47] T Bedford et al. Cryptic transmission of sars-cov-2 in washington state. *Science*, 370: 571–575, 2020.

[48] Q Bi et al. Epidemiology and transmission of covid-19 in 391 cases and 1286 of their close contacts in shenzhen, china: a retrospective cohort study. *Lancet Infect. Dis.*, 20:911–919, 2020.

[49] MA Suchard et al. Bayesian phylogenetic and phylodynamic data integration using beast

1.10. *Virus Evol.*, 4:vey016, 2018.

[50] L Pipes et al. Assessing uncertainty in the rooting of the sars-cov-2 phylogeny. *Mol. Biol. Evol.*, 2020. 10.1101/2020.06.19.160630.

[51] N Moshiri et al. Favites: simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics*, 35:1852–1861, 2019.

[52] R Laxminarayan et al. Epidemiology and transmission dynamics of covid-19 in two indian states. *Science*, 370:691–697, 2020.

[53] A Endo et al. Estimating the overdispersion in covid-19 transmission using outbreak sizes outside china. *Wellcome Open Res.*, 5:67, 2020.

[54] XK Xu et al. Reconstruction of transmission pairs for novel coronavirus disease 2019 (covid-19) in mainland china: estimation of superspreading events, serial interval, and hazard of infection. *Clin. Infect. Dis.*, 71:3163–3167, 2020.

[55] H Li et al. Clustering and superspreading potential of sars-cov-2 infections in hong kong. *Nat. Med.*, 26:1714–1719, 2020.

[56] H Li et al. Human-animal interactions and bat coronavirus spillover potential among rural residents in southern china. *Biosaf. Health*, 1:84–90, 2019.

[57] M Worobey et al. The emergence of sars-cov-2 in europe and north america. *Science*, 370: 564–570, 2020.

[58] L du Plessis et al. Establishment and lineage dynamics of the sars-cov-2 epidemic in the uk. *Science*, 371:708–712, 2021.

[59] A Deslandes et al. Sars-cov-2 was already spreading in france in late december 2019. *Int. J. Antimicrob. Agents*, 55:106006, 2020.

[60] X Deng et al. Genomic surveillance reveals multiple introductions of sars-cov-2 into northern california. *Science*, 369:582–587, 2020.

[61] CDC COVID-19 Response Team et al. Evidence for limited early spread of covid-19 within the united states, january–february 2020. *MMWR Morb. Mortal. Wkly. Rep.*, 69(22): 680–684, 2020.

[62] G Chavarria-Miró et al. Title. *http://medrxiv.org/lookup/doi/10.1101/2020.06.13.20129627*, 2020.

[63] G Fongaro et al. Title. *https://www.medrxiv.org/content/10.1101/2020.06.26.20140731v1*, 2020.

[64] G La Rosa, P Mancini, G Bonanno Ferraro, C Veneri, M Iaconelli, L Bonadonna, L Lucentini, and E Suffredini. Sars-cov-2 has been circulating in northern italy since december 2019: Evidence from environmental monitoring. *Sci. Total Environ.*, 750:141711, 2021.

[65] A Amendola, S Bianchi, M Gori, D Colzani, M Canuti, E Borghi, MC Raviglione, GV Zuccotti, and E Tanzi. Evidence of sars-cov-2 rna in an oropharyngeal swab specimen, milan, italy, early december 2019. *Emerg. Infect. Dis.*, 27:648–650, 2021.

[66] EO Nsoesie, B Rader, YL Barnoon, L Goodwin, and J Brownstein. Analysis of hospital traffic and search engine data in wuhan china indicates early disease activity in the fall of 2019. *Available at https://dash.harvard.edu/handle/1/42669767*, 2020.

[67] J Peccia, A Zulli, DE Brackney, ND Grubaugh, EH Kaplan, A Casanovas-Massana, AI Ko, AA Malik, D Wang, M Wang, et al. Measurement of sars-cov-2 rna in wastewater tracks community infection dynamics. *Nat. Biotechnol.*, pages 1–4, 2020.

[68] E Volz, S Mishra, M Chand, JC Barrett, R Johnson, L Geidelberg, WR Hinsley, DJ Laydon, G Dabrera, Á O'Toole, et al. Title. *medRxiv*, 2020. in press, doi:10.1101/2020.12.30.20249034.

[69] B Korber et al. Tracking changes in sars-cov-2 spike: Evidence that d614g increases infectivity of the covid-19 virus. *Cell*, 182:812–827.e19, 2020.

[70] S Lytras et al. Title. *https://www.biorxiv.org/content/10.1101/2021.01.22.427830v2*, 2021.

[71] MF Boni, P Lemey, X Jiang, TTY Lam, BW Perry, TA Castoe, A Rambaut, and DL Robertson. Evolutionary origins of the sars-cov-2 sarbecovirus lineage responsible for the covid-19 pandemic. *Nat. Microbiol.*, pages 1–10, 2020.

[72] J Pekar and J Wertheim. Data for "timing the sars-cov-2 index case in hubei province". *Data Dryad*, 2021.

[73] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *Lancet Infect. Dis.*, 20:533–534, 2020. doi: 10.1016/S1473-3099(20)30120-1.

[74] L.-L. Ren, Y.-M. Wang, Z.-Q. Wu, Z.-C. Xiang, L. Guo, T. Xu, Y.-Z. Jiang, Y. Xiong, Y.-J. Li, X.-W. Li, H. Li, G.-H. Fan, X.-Y. Gu, Y. Xiao, H. Gao, J.-Y. Xu, F. Yang, X.-M. Wang, C. Wu, L. Chen, Y.-W. Liu, B. Liu, J. Yang, X.-R. Wang, J. Dong, L. Li, C.-L. Huang, J.-P. Zhao, Y. Hu, Z.-S. Cheng, L.-L. Liu, Z.-H. Qian, C. Qin, Q. Jin, B. Cao, and J.-W. Wang. Identification of a novel coronavirus causing severe pneumonia in human: A descriptive study. *Chin. Med. J. (Engl.)*, 133:1015–1024, 2020. doi: 10.1097/CM9.0000000000000722.

[75] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, S. Beltekian, and X. Roser. Coronavirus pandemic (covid-19), 2022. URL https://ourworldindata.org/covid-deaths.

[76] A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, and 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.

[77] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, and Y.-Z. Zhang. A new coronavirus associated with human respiratory disease in china. *Nature*, 579:265–269, 2020. doi: 10.1038/s41586-020-2008-3.

[78] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W. J. Liu, D. Wang, W. Xu, E. C. Holmes, G. F. Gao, G. Wu, W. Chen, W. Shi, and W. Tan. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet*, 395:565–574, 2020. doi: 10.1016/S0140-6736(20)30251-8.

[79] S. Lytras, J. Hughes, D. Martin, P. Swanepoel, A. de Klerk, R. Lourens, S. L. Kosakovsky Pond, W. Xia, X. Jiang, and D. L. Robertson. Exploring the natural origins of sars-cov-2 in the light of recombination. *Genome Biol. Evol.*, 14:evac018, 2022. doi: 10.1093/gbe/evac018.

[80] R. F. Garry. Early appearance of two distinct genomic lineages of sars-cov-2 in different wuhan wildlife markets suggests sars-cov-2 has a natural origin, 2021. URL https://virological.org/t/early-appearance-of-two-distinct-genomic-lineages-of-sars-cov-2-in-different-wuhan-wildlife-markets-suggest 691.

[81] N. De Maio, C. Walker, R. Borges, L. Weilguny, G. Slodkowicz, and N. Goldman. Issues with sars-cov-2 sequencing data, 2020. URL https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.

[82] M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, and P. Lemey. The emergence of sars-cov-2 in europe and north america. *Science*, 370:564–570, 2020. doi: 10.1126/science.abc8169.

[83] J. O. Wertheim, M. Steel, and M. J. Sanderson. Accuracy in near-perfect virus phylogenies. *Syst. Biol.*, 71:426–438, 2022. doi: 10.1093/sysbio/syab069.

[84] S. Temmam, K. Vongphayloth, E. Baquero, S. Munier, M. Bonomi, B. Regnault, B. Douangboubpha, Y. Karami, D. Chrétien, D. Sanamxay, V. Xayaphet, P. Paphaphanh, V. Lacoste, S. Somlor, K. Lakeomany, N. Phommavanh, P. Pérot, O. Dehan, F. Amara, F. Donati, T. Bigot, M. Nilges, F. A. Rey, S. van der Werf, P. T. Brey, and M. Eloit. Bat coronaviruses related to sars-cov-2 and infectious for human cells. *Nature*, 604:330–336, 2022. doi: 10.1038/s41586-022-04532-4.

[85] J. B. Pease and M. W. Hahn. More accurate phylogenies inferred from low-recombination

regions in the presence of incomplete lineage sorting. *Evolution*, 67:2376–2384, 2013. doi: 10.1111/evo.12118.

[86] J. Ratcliff and P. Simmonds. Potential apobec-mediated rna editing of the genomes of sars-cov-2 and other coronaviruses and its impact on their longer term evolution. *Virology*, 556:62–72, 2021. doi: 10.1016/j.virol.2020.12.018.

[87] P. Simmonds. Rampant c→u hypermutation in the genomes of sars-cov-2 and other coronaviruses: Causes and consequences for their short- and long-term evolutionary trajectories. *MSphere*, 5:e00408–20, 2020. doi: 10.1128/mSphere.00408-20.

[88] P. Simmonds and M. A. Ansari. Extensive c->u transition biases in the genomes of a wide range of mammalian rna viruses; potential associations with transcriptional mutations, damage- or host-mediated editing of viral rna. *PLOS Pathog.*, 17:e1009596, 2021. doi: 10.1371/journal.ppat.1009596.

[89] P. Forster, L. Forster, C. Renfrew, and M. Forster. Phylogenetic network analysis of sars-cov-2 genomes. *Proc. Natl. Acad. Sci. U.S.A.*, 117:9241–9243, 2020. doi: 10.1073/pnas.2004999117.

[90] J. D. Bloom. Recovery of deleted deep sequencing data sheds more light on the early wuhan sars-cov-2 epidemic. *Mol. Biol. Evol.*, 38:5211–5224, 2021. doi: 10.1093/molbev/msab246.

[91] M. A. Caraballo-Ortiz, S. Miura, M. Sanderford, T. Dolker, Q. Tao, S. Weaver, S. L. K. Pond, and S. Kumar. Tophap: Rapid inference of key phylogenetic structures from common haplotypes in large genome collections with limited diversity. *Bioinformatics*, 38:2719–2726, 2022. doi: 10.1093/bioinformatics/btac186.

[92] S. Kumar, Q. Tao, S. Weaver, M. Sanderford, M. A. Caraballo-Ortiz, S. Sharma, S. L. K. Pond, and S. Miura. An evolutionary portrait of the progenitor sars-cov-2 and its dominant offshoots in covid-19 pandemic. *Mol. Biol. Evol.*, 38:3046–3059, 2021. doi: 10.1093/molbev/msab118.

[93] N. Moshiri, M. Ragonnet-Cronin, J. O. Wertheim, and S. Mirarab. Favites: Simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics*, 35:1852–1861, 2019. doi: 10.1093/bioinformatics/bty921.

[94] J. Pekar, M. Worobey, N. Moshiri, K. Scheffler, and J. O. Wertheim. Timing the sars-cov-2 index case in hubei province. *Science*, 372:412–417, 2021. doi: 10.1126/science.abf8003.

[95] S. Hsiang, D. Allen, S. Annan-Phan, K. Bell, I. Bolliger, T. Chong, H. Druckenmiller, L. Y. Huang, A. Hultgren, E. Krasovich, P. Lau, J. Lee, E. Rolf, J. Tseng, and T. Wu. The effect of large-scale anti-contagion policies on the covid-19 pandemic. *Nature*, 584:262–267, 2020. doi: 10.1038/s41586-020-2404-8.

[96] A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge. The challenges of

modeling and forecasting the spread of covid-19. *Proceedings of the National Academy of Sciences*, 117:16732–16738, 2020. doi: 10.1073/pnas.2006520117.

[97] S. Sanche, Y. T. Lin, C. Xu, E. Romero-Severson, N. Hengartner, and R. Ke. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerging Infectious Diseases*, 26:1470–1477, 2020. doi: 10.3201/eid2607.200282.

[98] T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, M.-L. Huang, A. Nalla, G. Pepper, A. Reinhardt, H. Xie, L. Shrestha, T. N. Nguyen, A. Adler, E. Brandstetter, S. Cho, D. Giroux, P. D. Han, K. Fay, C. D. Frazar, M. Ilcisin, K. Lacombe, J. Lee, A. Kiavand, M. Richardson, T. R. Sibley, M. Truong, C. R. Wolf, D. A. Nickerson, M. J. Rieder, J. A. Englund, J. Hadfield, E. B. Hodcroft, J. Huddleston, L. H. Moncla, N. F. Müller, R. A. Neher, X. Deng, W. Gu, S. Federman, C. Chiu, J. S. Duchin, R. Gautom, G. Melly, B. Hiatt, P. Dykema, S. Lindquist, K. Queen, Y. Tao, A. Uehara, S. Tong, D. MacCannell, G. L. Armstrong, G. S. Baird, H. Y. Chu, J. Shendure, K. R. Jerome, H. Y. Chu, M. Boeckh, J. A. Englund, M. Famulare, B. R. Lutz, D. A. Nickerson, L. M. Starita, M. Thompson, J. Shendure, T. Bedford, A. Adler, E. Brandstetter, S. Cho, C. D. Frazar, D. Giroux, P. D. Han, J. Hadfield, S. Huang, M. L. Jackson, A. Kiavand, L. E. Kimball, K. Lacombe, J. Logue, V. Lyon, K. L. Newman, M. Richardson, T. R. Sibley, M. L. Zigman Suchsland, M. Truong, and C. R. Wolf. Cryptic transmission of sars-cov-2 in washington state. *Science*, 370:571–575, 2020. doi: 10.1126/science.abc0523.

[99] M. Zeller, K. Gangavarapu, C. Anderson, A. R. Smither, J. A. Vanchiere, R. Rose, D. J. Snyder, G. Dudas, A. Watts, N. L. Matteson, R. Robles-Sikisaka, M. Marshall, A. K. Feehan, G. Sabino-Santos Jr., A. R. Bell-Kareem, L. D. Hughes, M. Alkuzweny, P. Snarski, J. Garcia-Diaz, R. S. Scott, L. I. Melnik, R. Klitting, M. McGraw, P. Belda-Ferre, P. DeHoff, S. Sathe, C. Marotz, N. D. Grubaugh, D. J. Nolan, A. C. Drouin, K. J. Genemaras, K. Chao, S. Topol, E. Spencer, L. Nicholson, S. Aigner, G. W. Yeo, L. Farnaes, C. A. Hobbs, L. C. Laurent, R. Knight, E. B. Hodcroft, K. Khan, D. N. Fusco, V. S. Cooper, and P. Lemey. Emergence of an early sars-cov-2 epidemic in the united states. *Cell*, 184:4939–4952.e15, 2021. doi: 10.1016/j.cell.2021.07.030.

[100] C. Alteri, V. Cento, A. Piralla, V. Costabile, M. Tallarita, L. Colagrossi, S. Renica, F. Giardina, F. Novazzi, S. Gaiarsa, E. Matarazzo, M. Antonello, C. Vismara, R. Fumagalli, O. M. Epis, M. Puoti, C. F. Perno, and F. Baldanti. Genomic epidemiology of sars-cov-2 reveals multiple lineages and early spread of sars-cov-2 infections in lombardy, italy. *Nature Communications*, 12:434, 2021. doi: 10.1038/s41467-020-20688-x.

[101] L. du Plessis and O. Pybus. Further musings on the tmrca. *Virological*, 2020. URL https://virological.org/t/further-musings-on-the-tmrca/340.

[102] Centers for Disease Control and Prevention (CDC). Prevalence of igg antibody to sars-associated coronavirus in animal traders—guangdong province, china, 2003. *MMWR Morb. Mortal. Wkly. Rep.*, 52:986–987, 2003.

[103] A. Marí Saéz, S. Weiss, K. Nowak, et al. Investigating the zoonotic origin of the west

african ebola epidemic. *EMBO Mol. Med.*, 7:17–23, 2015. doi: 10.15252/emmm.201404792.

[104] X. Zhang, Y. Tan, Y. Ling, et al. Viral and host factors related to the clinical outcome of covid-19. *Nature*, 583:437–440, 2020. doi: 10.1038/s41586-020-2355-0.

[105] E. O. Nsoesie, B. Rader, Y. L. Barnoon, et al. Analysis of hospital traffic and search engine data in wuhan china indicates early disease activity in the fall of 2019. *Dig. Acc. Scholar. Harv.*, 2, 2019.

[106] L. Chang, L. Zhao, Y. Xiao, et al. Serosurvey for sars-cov-2 among blood donors in wuhan, china from september to december 2019. *Protein Cell*, 2022. URL https://doi.org/10.1093/procel/pwac013.

[107] E. C. Holmes, S. A. Goldstein, A. L. Rasmussen, et al. The origins of sars-cov-2: A critical review. *Cell*, 184:4848–4856, 2021. doi: 10.1016/j.cell.2021.08.017.

[108] M. Worobey, J. I. Levy, L. M. Serrano, et al. The huanan seafood wholesale market in wuhan was the early epicenter of the covid-19 pandemic. *Science*, 377:951–959, 2022. doi: 10.1126/science.abp8715.

[109] X. Xiao, C. Newman, C. D. Buesching, et al. Animal sales from wuhan wet markets immediately prior to the covid-19 pandemic. *Sci. Rep.*, 11:11898, 2021. doi: 10.1038/s41598-021-91470-2.

[110] C. M. Freuling, A. Breithaupt, T. Müller, et al. Susceptibility of raccoon dogs for experimental sars-cov-2 infection. *Emerg. Infect. Dis.*, 26:2982–2985, 2020. doi: 10.3201/eid2612.203733.

[111] S. M. Porter, A. E. Hartwig, H. Bielefeldt-Ohmann, A. M. Bosco-Lauth, and J. Root. Susceptibility of wild canids to severe acute respiratory syndrome coronavirus 2 (sars-cov-2). *bioRxiv*, 2022. URL https://doi.org/10.1101/2022.01.27.478082. 478082 [Preprint].

[112] G. Gao, W. Liu, P. Liu, et al. Surveillance of sars-cov-2 in the environment and animal samples of the huanan seafood market. *Research Square*, 2022. https://doi.org/10.21203/rs.3.rs-1370392/v1.

[113] L. du Plessis, J. T. McCrone, A. E. Zarebski, et al. Establishment and lineage dynamics of the sars-cov-2 epidemic in the uk. *Science*, 371:708–712, 2021. doi: 10.1126/science.abf2946.

[114] Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the sars coronavirus during the course of the sars epidemic in china. *Science*, 303:1666–1669, 2004. doi: 10.1126/science.1092002.

[115] G. Dudas, L. M. Carvalho, A. Rambaut, et al. Mers-cov spillover at the camel-human interface. *eLife*, 7:e31257, 2018. doi: 10.7554/eLife.31257.

[116] J. A. Lednicky, M. S. Tagliamonte, S. K. White, et al. Independent infections of porcine deltacoronavirus among haitian children. *Nature*, 600:133–137, 2021. doi: 10.1038/s41586-021-04111-z.

[117] B. Kan, M. Wang, H. Jing, et al. Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms. *J. Virol.*, 79:11892–11900, 2005. doi: 10.1128/JVI.79.18.11892-11900.2005.

[118] K. G. Andersen, A. Rambaut, W. I. Lipkin, et al. The proximal origin of sars-cov-2. *Nat. Med.*, 26:450–452, 2020. doi: 10.1038/s41591-020-0820-9.

[119] V. L. Hale, P. M. Dennis, D. S. McBride, et al. Sars-cov-2 infection in free-ranging white-tailed deer. *Nature*, 602:481–486, 2022. doi: 10.1038/s41586-021-04353-x.

[120] J. C. Chandler, S. N. Bevins, J. W. Ellis, et al. Sars-cov-2 exposure in wild white-tailed deer (*odocoileus virginianus*). *Proc. Natl. Acad. Sci. U.S.A.*, 118:e2114828118, 2021. doi: 10.1073/pnas.2114828118.

[121] L. Lu, R. S. Sikkema, F. C. Velkers, et al. Adaptation, spread and transmission of sars-cov-2 in farmed minks and associated humans in the netherlands. *Nat. Commun.*, 12:6802, 2021. doi: 10.1038/s41467-021-27096-9.

[122] B. B. Oude Munnink, R. S. Sikkema, D. F. Nieuwenhuijse, et al. Transmission of sars-cov-2 on mink farms between humans and mink and back to humans. *Science*, 371:172–177, 2021. doi: 10.1126/science.abe5901.

[123] S. V. Kuchipudi, M. Surendran-Nair, R. M. Ruden, et al. Multiple spillovers from humans and onward transmission of sars-cov-2 in white-tailed deer. *Proc. Natl. Acad. Sci. U.S.A.*, 119:e2121644119, 2022. doi: 10.1073/pnas.2121644119.

[124] H.-L. Yen, T. H. C. Sit, C. J. Brackman, et al. Transmission of sars-cov-2 delta variant (ay.127) from pet hamsters to humans, leading to onward human-to-human transmission: A case study. *Lancet*, 399:1070–1078, 2022. doi: 10.1016/S0140-6736(22)00326-9.

[125] Y. Shu and J. McCauley. Gisaid: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.*, 22:30494, 2017. doi: 10.2807/1560-7917.ES.2017.22.13.30494.

[126] K. Katoh and D. M. Standley. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, 30:772–780, 2013. doi: 10.1093/molbev/mst010.

[127] N. De Maio, C. Walker, R. Borges, et al. Masking strategies for sars-cov-2 alignments, 2020. URL https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480.

[128] B. Q. Minh, H. A. Schmidt, O. Chernomor, et al. Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, 37:1530–1534,

2020. doi: 10.1093/molbev/msaa015.

[129] P. Sagulenko, V. Puller, and R. A. Neher. Treetime: Maximum-likelihood phylodynamic analysis. *Virus Evol.*, 4:vex042, 2018. doi: 10.1093/ve/vex042.

[130] M. A. Suchard, P. Lemey, G. Baele, et al. Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus Evol.*, 4:vey016, 2018. doi: 10.1093/ve/vey016.

[131] N. Moshiri. Favites-covid-lite: A simplified (and much faster) simulation pipeline specifically for covid-19 contact + transmission + phylogeny + sequence simulation, 2022. URL https://github.com/niemasd/FAVITES-COVID-Lite.

[132] X. Hao, S. Cheng, D. Wu, et al. Reconstruction of the full transmission dynamics of covid-19 in wuhan. *Nature*, 584:420–424, 2020. doi: 10.1038/s41586-020-2554-8.

[133] J. E. Pekar, A. Magee, E. Parker, et al. The molecular epidemiology of multiple zoonotic origins of sars-cov-2, 2022.

[134] J. E. Pekar, A. Magee, E. Parker, et al. The molecular epidemiology of multiple zoonotic origins of sars-cov-2, 2022.

[135] Xian-Dan Lin, Wen Wang, Zong-Yu Hao, Zhao-Xiao Wang, Wen-Ping Guo, Xiao-Qing Guan, Miao-Ruo Wang, Hong-Wei Wang, Run-Hong Zhou, Ming-Hui Li, Guang-Peng Tang, Jun Wu, Edward C. Holmes, and Yong-Zhen Zhang. Extensive diversity of coronaviruses in bats from china. *Virology*, 507:1–10, 2017.

[136] M. F. Boni, P. Lemey, X. Jiang, T. T.-Y. Lam, B. W. Perry, T. A. Castoe, A. Rambaut, and D. L. Robertson. Evolutionary origins of the sars-cov-2 sarbecovirus lineage responsible for the covid-19 pandemic. *Nat Microbiol.*, 5:1408–1417, 2020.

[137] J. O. Wertheim, D. K. W. Chu, J. S. M. Peiris, S. L. Kosakovsky Pond, and L. L. M. Poon. A case for the ancient origin of coronaviruses. *J. Virol.*, 87:7039–7045, 2013.

[138] M. Ghafari, P. Simmonds, O. G. Pybus, and A. Katzourakis. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. *Curr. Biol.*, 31: 4689–4696.e5, 2021.

[139] Y. Souilmi, M. E. Lauterbur, R. Tobler, C. D. Huber, A. S. Johar, S. V. Moradi, W. A. Johnston, N. J. Krogan, K. Alexandrov, and D. Enard. An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in east asia. *Curr. Biol.*, 31: 3504–3514.e9, 2021.

[140] M. Worobey, P. Telfer, S. Souquière, M. Hunter, C. A. Coleman, M. J. Metzger, P. Reed, M. Makuwa, G. Hearn, S. Honarvar, P. Roques, C. Apetrei, M. Kazanji, and P. A. Marx. Island biogeography reveals the deep history of siv. *Science*, 329, 2010.

[141] J. O. Wertheim and S. L. Kosakovsky Pond. Purifying selection can obscure the ancient age of viral lineages. *Mol. Biol. Evol.*, 28:3355–3365, 2011.

[142] P. Simmonds, P. Aiewsakun, and A. Katzourakis. Prisoners of war - host adaptation and its constraints on virus evolution. *Nat. Rev. Microbiol.*, 17:321–328, 2019.

[143] S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. W. Frost. Gard: a genetic algorithm for recombination detection. *Bioinformatics*, 22:3096–3098, 2006.

[144] N. F. Müller, K. E. Kistler, and T. Bedford. A bayesian approach to infer recombination patterns in coronaviruses. *Nat. Commun.*, 13, 2022.

[145] J. E. Pekar, A. Magee, E. Parker, N. Moshiri, K. Izhikevich, J. L. Havens, K. Gangavarapu, L. M. Malpica Serrano, A. Crits-Christoph, N. L. Matteson, M. Zeller, J. I. Levy, J. C. Wang, S. Hughes, J. Lee, H. Park, M.-S. Park, K. Ching Zi Yan, R. T. P. Lin, M. N. Mat Isa, Y. M. Noor, T. I. Vasylyeva, R. F. Garry, E. C. Holmes, A. Rambaut, M. A. Suchard, K. G. Andersen, M. Worobey, and J. O. Wertheim. The molecular epidemiology of multiple zoonotic origins of sars-cov-2. *Science*, 377:960–966, 2022.

[146] Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the sars coronavirus during the course of the sars epidemic in china. *Science*, 303:1666–1669, 2004.

[147] K. Xiao, J. Zhai, Y. Feng, N. Zhou, X. Zhang, J.-J. Zou, N. Li, Y. Guo, X. Li, X. Shen, Z. Zhang, F. Shu, W. Huang, Y. Li, Z. Zhang, R.-A. Chen, Y.-J. Wu, S.-M. Peng, M. Huang, W.-J. Xie, Q.-H. Cai, F.-H. Hou, W. Chen, L. Xiao, and Y. Shen. Isolation of sars-cov-2-related coronavirus from malayan pangolins. *Nature*, 583:286–289, 2020.

[148] T. T.-Y. Lam, N. Jia, Y.-W. Zhang, M. H.-H. Shum, J.-F. Jiang, H.-C. Zhu, Y.-G. Tong, Y.-X. Shi, X.-B. Ni, Y.-S. Liao, W.-J. Li, B.-G. Jiang, W. Wei, T.-T. Yuan, K. Zheng, X.-M. Cui, J. Li, G.-Q. Pei, X. Qiang, W. Y.-M. Cheung, L.-F. Li, F.-F. Sun, S. Qin, J.-C. Huang, G. M. Leung, E. C. Holmes, Y.-L. Hu, Y. Guan, and W.-C. Cao. Identifying sars-cov-2-related coronaviruses in malayan pangolins. *Nature*, 583:282–285, 2020.

[149] P. Liu, W. Chen, and J.-P. Chen. Viral metagenomics revealed sendai virus and coronavirus infection of malayan pangolins (manis javanica). *Viruses*, 2019. doi: 10.3390/v11110979.

[150] S. Duchêne, E. C. Holmes, and S. Y. W. Ho. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. Biol. Sci.*, 281, 2014. doi: 10.1098/rspb.2014.0732.

[151] P. Aiewsakun and A. Katzourakis. Time-dependent rate phenomenon in viruses. *J. Virol.*, 90:7184–7195, 2016.

[152] N. S. Trovão, M. A. Suchard, G. Baele, M. Gilbert, and P. Lemey. Bayesian inference reveals host-specific contributions to the epidemic expansion of influenza a h5n1. *Mol. Biol.*

Evol., 32:3264–3275, 2015.

[153] R. Clements, N. S. Sodhi, M. Schilthuizen, and P. K. L. Ng. Limestone karsts of southeast asia: Imperiled arks of biodiversity. *Bioscience*, 56:733–742, 2006.

[154] A. Hassanin, V. T. Tu, M. Curaudeau, and G. Csorba. Inferring the ecological niche of bat viruses closely related to sars-cov-2 using phylogeographic analyses of rhinolophus species. *Sci. Rep.*, 11, 2021.

[155] K. Funakoshi and F. Maeda. Foraging activity and night-roost usage in the japanese greater horseshoe bat, rhinolophus ferrumequinum nippon. *Mammal Study*, 28:1–10, 2003.

[156] F. Bontadina, H. Schofield, and B. Naef-Daenzer. Radio-tracking reveals that lesser horseshoe bats (rhinolophus hipposideros) forage in woodland. *J. Zool.*, 258:281–290, 2002.

[157] U. Goiti, J. R. Aihartza, D. Almenar, and E. Salsamendi. Seasonal foraging by rhinolophus euryale (rhinolophidae) in an atlantic rural landscape in northern iberian peninsula. *Acta Chiropt.*, 8:141–155, 2006.

[158] S. J. Rossiter, G. Jones, R. D. Ransome, and E. M. Barratt. Parentage, reproductive success and breeding behaviour in the greater horseshoe bat (rhinolophus ferrumequinum). *Proc. Biol. Sci.*, 267:545–551, 2000.

[159] L. Henley, O. Jones, F. Mathews, and T. E. Woolley. Coupling bayesian theory and static acoustic detector data to model bat motion and locate roosts. *arXiv [q-bio.QM]*, 2021. (available at http://arxiv.org/abs/2106.11969).

[160] A. de Klerk, P. Swanepoel, R. Lourens, M. Zondo, I. Abodunran, S. Lytras, O. A. MacLean, D. Robertson, S. L. Kosakovsky Pond, J. D. Zehr, V. Kumar, M. J. Stanhope, G. Harkins, B. Murrell, and D. P. Martin. Conserved recombination patterns across coronavirus subgenera. *Virus Evol.*, 8:veac054, 2022.

[161] Who-convened global study of origins of sars-cov-2: China part. 2021. URL https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part.

[162] D. Russo, G. Jones, and A. Migliozzi. Habitat selection by the mediterranean horseshoe bat, rhinolophus euryale (chiroptera: Rhinolophidae) in a rural area of southern italy and implications for conservation. *Biol. Conserv.*, 107:71–81, 2002.

[163] S. K. P. Lau, K. S. M. Li, Y. Huang, C.-T. Shek, H. Tse, M. Wang, G. K. Y. Choi, H. Xu, C. S. F. Lam, R. Guo, K.-H. Chan, B.-J. Zheng, P. C. Y. Woo, and K.-Y. Yuen. Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related rhinolophus bat coronavirus in china reveal bats as a reservoir for acute, self-limiting infection that allows recombination events. *J. Virol.*, 84:2808–2819, 2010.

[164] Z. Bihari. Characteristics of the northernmost population of rhinolophus ferrumequinum in the carpathian basin. *Acta Theriol.*, 46:13–21, 2001.

[165] Y. Guan, B. J. Zheng, Y. Q. He, X. L. Liu, Z. X. Zhuang, C. L. Cheung, S. W. Luo, P. H. Li, L. J. Zhang, Y. J. Guan, K. M. Butt, K. L. Wong, K. W. Chan, W. Lim, K. F. Shortridge, K. Y. Yuen, J. S. M. Peiris, and L. L. M. Poon. Isolation and characterization of viruses related to the sars coronavirus from animals in southern china. *Science*, 302: 276–278, 2003.

[166] M. Worobey. Dissecting the early covid-19 cases in wuhan. *Science*, 374:1202–1204, 2021.

[167] M. Worobey, J. I. Levy, L. M. Serrano, A. Crits-Christoph, J. E. Pekar, S. A. Goldstein, A. L. Rasmussen, M. U. G. Kraemer, C. Newman, M. P. G. Koopmans, M. A. Suchard, J. O. Wertheim, P. Lemey, D. L. Robertson, R. F. Garry, E. C. Holmes, A. Rambaut, and K. G. Andersen. The huanan seafood wholesale market in wuhan was the early epicenter of the covid-19 pandemic. *Science*, abp8715, 2022.

[168] H. L. Ward, R. D. Ransome, G. Jones, and S. J. Rossiter. Determinants and patterns of reproductive success in the greater horseshoe bat during a population recovery. *PLoS One*, 9:e87199, 2014.

[169] Species structure of the bat community hibernating in muierilor cave (southern carpathians, romania). URL https://adatbank.ro/denever/cikkek/Dragu_2009.pdf.

[170] M. López-Roig, H. Bourhy, R. Lavenir, and J. Serra-Cobo. Seroprevalence dynamics of european bat lyssavirus type 1 in a multispecies bat colony. *Viruses*, 6:3386–3399, 2014.

[171] D. T. S. Hayman, P. M. Cryan, P. D. Fricker, and N. G. Dannemiller. Long-term video surveillance and automated analyses reveal arousal patterns in groups of hibernating bats. *Methods Ecol. Evol.*, 8:1813–1821, 2017.

[172] C. Stawski, C. K. R. Willis, and F. Geiser. The importance of temporal heterothermy in bats. *J. Zool.*, 2014. URL https://zslpublications.onlinelibrary.wiley.com/doi/abs/10.1111/jzo.12105.

[173] S. Murakami, T. Kitamura, H. Matsugo, H. Kamiki, K. Oyabu, W. Sekine, A. Takenaka-Uema, Y. Sakai-Tagawa, Y. Kawaoka, and T. Horimoto. Isolation of bat sarbecoviruses, japan. *Emerg. Infect. Dis.*, 28:2500–2503, 2022.

[174] I. Ahlén and H. J. Baagøe. Behavior of scandinavian bats during migration and foraging at sea. *J. Mammal.*, 90:1318–1323, 2009.

[175] D. G. Constantine. Geographic translocation of bats: known and potential problems. *Emerg. Infect. Dis.*, 9:17–21, 2003.

[176] D. I. Solick and C. M. Newman. Oceanic records of north american bats and implications

for offshore wind energy development in the united states. *Ecol. Evol.*, 11:14433–14447, 2021.

[177] A. C. Breed, H. E. Field, C. S. Smith, J. Edmonston, and J. Meers. Bats without borders: long-distance movements and implications for disease risk management. *Ecohealth*, 7: 204–212, 2010.

[178] R. C. Anderson, R. Chakravarty, and S. Raman. Bats in the maldives: a review of historical data and first record of a vagrant long-winged tomb bat (emballonuridae: Taphozous longimanus). *Journal of Bat Research & Conservation*, 15(1):78–83, 2022. URL https://www.researchgate.net/profile/Rohit-Chakravarty/publication/367136026_Bats_in_the_Maldives_a_review_of_historical_data_and_first_record_of_a_vagrant_Long-winged_Tomb_Bat_Emballonuridae_Taphozous-longimanus.pdf.

[179] Y. Nakazawa and C. J. Bae. Quaternary paleoenvironmental variation and its impact on initial human dispersals into the japanese archipelago. *Palaeogeogr. Palaeoclimatol. Palaeoecol.*, 512:145–155, 2018.

[180] M. Dobson. Patterns of distribution in japanese land mammals. *Mamm. Rev.*, 24:91–111, 1994.

[181] J. J. Shi and D. L. Rabosky. Speciation dynamics during the global radiation of extant bats. *Evolution*, 69:1528–1545, 2015.

[182] H. L. Wells, M. Letko, G. Lasso, B. Ssebide, J. Nziza, D. K. Byarugaba, I. Navarrete-Macias, E. Liang, M. Cranfield, B. A. Han, M. W. Tingley, M. Diuk-Wasser, T. Goldstein, C. K. Johnson, J. A. K. Mazet, K. Chandran, V. J. Munster, K. Gilardi, and S. J. Anthony. The evolutionary history of ace2 usage within the coronavirus subgenus sarbecovirus. *Virus Evol.*, 7:veab007, 2021.

[183] Z. Wu, Y. Han, Y. Wang, B. Liu, L. Zhao, J. Zhang, H. Su, W. Zhao, L. Liu, S. Bai, J. Dong, L. Sun, Y. Zhu, S. Zhou, Y. Song, H. Sui, J. Yang, J. Wang, S. Zhang, Z. Qian, and Q. Jin. A comprehensive survey of bat sarbecoviruses across china in relation to the origins of sars-cov and sars-cov-2. *Natl Sci Rev*, 2022.

[184] A. Latinne, B. Hu, K. J. Olival, G. Zhu, L. Zhang, H. Li, A. A. Chmura, H. E. Field, C. Zambrana-Torrelio, J. H. Epstein, B. Li, W. Zhang, L.-F. Wang, Z.-L. Shi, and P. Daszak. Origin and cross-species transmission of bat coronaviruses in china. *Nat. Commun.*, 11: 4235, 2020.

[185] J. Suzuki, R. Sato, T. Kobayashi, T. Aoi, and R. Harasawa. Group b betacoronavirus in rhinolophid bats, japan. *J. Vet. Med. Sci.*, 76:1267–1269, 2014.

[186] S. M. Kasibhatla, M. Kinikar, S. Limaye, M. M. Kale, and U. Kulkarni-Kale. Understanding evolution of sars-cov-2: A perspective from analysis of genetic diversity of rdrp gene. *J. Med. Virol.*, 92:1932–1937, 2020.

[187] H. K. H. Luk, X. Li, J. Fung, S. K. P. Lau, and P. C. Y. Woo. Molecular epidemiology, evolution and phylogeny of sars coronavirus. *Infect. Genet. Evol.*, 71:21–30, 2019.

[188] Alexander Crits-Christoph, Joshua I Levy, Jonathan Pekar, Stephen A Goldstein, Reema Singh, Zach Hensel, Karthik Gangavarapu, Matthew B Rogers, Niema Moshiri, Robert F Garry, et al. Genetic tracing of market wildlife and viruses at the epicenter of the covid-19 pandemic. *bioRxiv*, pages 2023–09, 2023.

[189] Bradley Pickering, Oliver Lung, Finlay Maguire, Peter Kruczkiewicz, Jonathon D Kotwa, Tore Buchanan, Marianne Gagnier, Jennifer L Guthrie, Claire M Jardine, Alex Marchand-Austin, et al. Divergent sars-cov-2 variant emerges in white-tailed deer with deer-to-human transmission. *Nature Microbiology*, 7(12):2011–2024, 2022.

[190] Ning Wang, Shi-Yue Li, Xing-Lou Yang, Hui-Min Huang, Yu-Ji Zhang, Hua Guo, Chu-Ming Luo, Maureen Miller, Guangjian Zhu, Aleksei A Chmura, et al. Serological evidence of bat sars-related coronavirus infection in humans, china. *Virologica Sinica*, 33:104–107, 2018.

[191] Manuel Ruiz-Aravena, Clifton McKee, Amandine Gamble, Tamika Lunn, Aaron Morris, Celine E Snedden, Claude Kwe Yinda, Julia R Port, David W Buchholz, Yao Yu Yeo, et al. Ecology, evolution and spillover of coronaviruses from bats. *Nature Reviews Microbiology*, 20(5):299–314, 2022.

[192] Wan-Ting He, Xin Hou, Jin Zhao, Jiumeng Sun, Haijian He, Wei Si, Jing Wang, Zhiwen Jiang, Ziqing Yan, Gang Xing, et al. Virome characterization of game animals in china reveals a spectrum of emerging pathogens. *Cell*, 185(7):1117–1129, 2022.

[193] Liang-Jun Chen, Xian-Dan Lin, Wen-Ping Guo, Jun-Hua Tian, Wen Wang, Xu-Hua Ying, Miao-Ruo Wang, Bin Yu, Zhan-Qiu Yang, Mang Shi, et al. Diversity and evolution of avian influenza viruses in live poultry markets, free-range poultry and wild wetland birds in china. *Journal of General Virology*, 97(4):844–854, 2016.

[194] Emma E Glennon, Freya L Jephcott, Olivier Restif, and James LN Wood. Estimating undetected ebola spillovers. *PLoS neglected tropical diseases*, 13(6):e0007428, 2019.

[195] Y Shu and J McCauley. Gisaid: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.*, 22(13), 2017.

[196] BQ Minh, HA Schmidt, O Chernomor, D Schrempf, MD Woodhams, A von Haeseler, and R Lanfear. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, 37:1530–1534, 2020.

[197] A Rambaut et al. Exploring the temporal structure of heterochronous sequences using tempest (formerly path-o-gen). *Virus Evol.*, 2, 2016. doi:10.1093/ve/vew007.

[198] A Rambaut, AJ Drummond, D Xie, G Baele, and MA Suchard. Posterior summarization in bayesian phylogenetics using tracer 1.7. *Syst. Biol.*, 67:901–904, 2018.

[199] MS Gill, P Lemey, NR Faria, A Rambaut, B Shapiro, and MA Suchard. Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.*, 30:713–724, 2013.

[200] AL Barabási and R Albert. Emergence of scaling in random networks. *Science*, 286: 509–512, 1999.

[201] J Mossong, N Hens, M Jit, P Beutels, K Auranen, R Mikolajczyk, M Massari, S Salmaso, GS Tomba, J Wallinga, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLOS Med.*, 5:e74, 2008.

[202] FD Sahneh, A Vajdi, H Shakeri, F Fan, and C Scoglio. Gemfsim: A stochastic simulator for the generalized epidemic modeling framework. *J. Comput. Sci.*, 22:36–44, 2017.

[203] O Ratmann, EB Hodcroft, M Pickles, A Cori, M Hall, S Lycett, C Colijn, B Dearlove, X Didelot, S Frost, et al. Phylogenetic tools for generalized hiv-1 epidemics: findings from the pangea-hiv methods comparison. *Mol. Biol. Evol.*, 34:185–203, 2017.

[204] N Moshiri. Treeswift: A massively scalable python tree package. *SoftwareX*, 11:100436, 2020.

[205] J. Hadfield, C. Megill, S. M. Bell, et al. Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, 34:4121–4123, 2018. doi: 10.1093/bioinformatics/bty407.

[206] A. Rambaut. figtree, 2018. URL https://github.com/rambaut/figtree/releases.

[207] H. Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34:3094–3100, 2018. doi: 10.1093/bioinformatics/bty191.

[208] H. Li, B. Handsaker, A. Wysoker, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25:2078–2079, 2009. doi: 10.1093/bioinformatics/btp352.

[209] N. D. Grubaugh, K. Gangavarapu, J. Quick, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using primalseq and ivar. *Genome Biol.*, 20:8, 2019. doi: 10.1186/s13059-018-1618-7.

[210] virus evolution. gofasta, 2022. URL https://github.com/virus-evolution/gofasta.

[211] G. Dudas. baltic: baltic - backronymed adaptable lightweight tree import code for molecular phylogeny manipulation, analysis and visualization, 2021. URL https://github.com/evogytis/baltic.

[212] S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, et al. Gard: A genetic algorithm for recombination detection. *Bioinformatics*, 22:3096–3098, 2006. doi: 10.1093/bioinformatics/btl474.

[213] D. P. Martin, B. Murrell, M. Golden, et al. Rdp4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.*, 1:vev003, 2015. doi: 10.1093/ve/vev003.

[214] H. M. Lam, O. Ratmann, and M. F. Boni. Improved algorithmic complexity for the 3seq recombination detection algorithm. *Mol. Biol. Evol.*, 35:247–251, 2018. doi: 10.1093/molbev/msx263.

[215] M. F. Boni, P. Lemey, X. Jiang, et al. Evolutionary origins of the sars-cov-2 sarbecovirus lineage responsible for the covid-19 pandemic. *Nat. Microbiol.*, 5:1408–1417, 2020. doi: 10.1038/s41564-020-0771-4.

[216] A. Rambaut, T. T. Lam, L. Max Carvalho, and O. G. Pybus. Exploring the temporal structure of heterochronous sequences using tempest (formerly path-o-gen). *Virus Evol.*, 2:vew007, 2016. doi: 10.1093/ve/vew007.

[217] A. Rambaut, A. J. Drummond, D. Xie, et al. Posterior summarization in bayesian phylogenetics using tracer 1.7. *Syst. Biol.*, 67:901–904, 2018. doi: 10.1093/sysbio/syy032.

[218] Epinow2: Estimate realtime case counts and time-varying epidemiological parameters, 2020. URL https://github.com/epiforecasts/EpiNow2.

[219] N. Moshiri. Niemagraphgen: A memory-efficient global-scale contact network simulation toolkit, 2022.

[220] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999. doi: 10.1126/science.286.5439.509.

[221] J. Mossong, N. Hens, M. Jit, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLOS Med.*, 5:e74, 2008. doi: 10.1371/journal.pmed.0050074.

[222] F. D. Sahneh, A. Vajdi, H. Shakeri, et al. Gemfsim: A stochastic simulator for the generalized epidemic modeling framework. *J. Comput. Sci.*, 22:36–44, 2017. doi: 10.1016/j.jocs.2017.08.014.

[223] F. Zhou, T. Yu, R. Du, et al. Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: A retrospective cohort study. *Lancet*, 395:1054–1062, 2020. doi: 10.1016/S0140-6736(20)30566-3.

[224] J. Yang, X. Chen, X. Deng, et al. Disease burden and clinical severity of the first pandemic wave of covid-19 in wuhan, china. *Nat. Commun.*, 11:5411, 2020. doi: 10.1038/s41467-020-19238-2.

[225] N. Moshiri. Treeswift: A massively scalable python tree package. *SoftwareX*, 11:100436, 2020. doi: 10.1016/j.softx.2020.100436.

[226] First chinese coronavirus cases may have been infected in october 2019, says new

research, 2021. URL https://www.scmp.com/news/china/science/article/3126499/first-chinese-covid-19-cases-may-have-been-infected-october-2019.

[227] Clock and tmrca based on 27 genomes, 2020. URL https://virological.org/t/clock-and-tmrca-based-on-27-genomes/347/6.

[228] L. Pipes, H. Wang, J. P. Huelsenbeck, and R. Nielsen. Assessing uncertainty in the rooting of the sars-cov-2 phylogeny. *Mol. Biol. Evol.*, 38:1537–1543, 2021. doi: 10.1093/molbev/msaa316.

[229] T. Murata, A. Sakurai, M. Suzuki, et al. Shedding of viable virus in asymptomatic sars-cov-2 carriers. *MSphere*, 6:e00019–21, 2021. doi: 10.1128/mSphere.00019-21.

[230] T. Sekizuka, K. Itokawa, T. Kageyama, et al. Haplotype networks of sars-cov-2 infections in the diamond princess cruise ship outbreak. *Proc. Natl. Acad. Sci. U.S.A.*, 117:20198–20201, 2020. doi: 10.1073/pnas.2006824117.

[231] Y. Turakhia, B. Thornlow, A. S. Hinrichs, et al. 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.

[232] P. Zhou, X.-L. Yang, X.-G. Wang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579:270–273, 2020. doi: 10.1038/s41586-020-2012-7.

[233] M. Ghafari, L. du Plessis, J. Raghwani, et al. Purifying selection determines the short-term time dependency of evolutionary rates in sars-cov-2 and ph1n1 influenza. *Mol. Biol. Evol.*, 39:msac009, 2022. doi: 10.1093/molbev/msac009.

[234] S. Duchêne, E. C. Holmes, and S. Y. W. Ho. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. Biol. Sci.*, 281:20140732, 2014. doi: 10.1098/rspb.2014.0732.

[235] J. T. Wu, K. Leung, M. Bushman, et al. Estimating clinical severity of covid-19 from the transmission dynamics in wuhan, china. *Nat. Med.*, 26:506–510, 2020. doi: 10.1038/s41591-020-0822-7.

[236] R. Ke, E. Romero-Severson, S. Sanche, and N. Hengartner. Estimating the reproductive number r0 of sars-cov-2 in the united states and eight european countries and implications for vaccination. *J. Theor. Biol.*, 517:110621, 2021. doi: 10.1016/j.jtbi.2021.110621.

[237] L. Pellis, F. Scarabel, H. B. Stage, et al. Challenges in control of covid-19: Short doubling time and long delay to effect of interventions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 376:20200264, 2021. doi: 10.1098/rstb.2020.0264.

[238] Q. Li, X. Guan, P. Wu, et al. Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.*, 382:1199–1207, 2020. doi: 10.1056/

NEJMoa2001316.

[239] M. Chinazzi, J. T. Davis, M. Ajelli, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (covid-19) outbreak. *Science*, 368:395–400, 2020. doi: 10.1126/science.aba9757.

[240] R. Li, S. Pei, B. Chen, et al. Substantial undocumentedinfection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, 368:489–493, 2020. doi: 10.1126/science.abb3221.

[241] N. Moshiri. Coatran: Coalescent tree simulation along a transmission network, 2020. URL https://doi.org/10.1101/2020.11.10.377499.

[242] K. M. Braun, G. K. Moreno, C. Wagner, et al. Acute sars-cov-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLOS Pathog.*, 17: e1009849, 2021. doi: 10.1371/journal.ppat.1009849.

[243] J. Ma. Coronavirus: China's first confirmed covid-19 case traced back to november 17.

[244] K. Katoh and D. M. Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30:772–780, 2013.

[245] T. A. Hall. Bioedit a user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. *Nucleic Acids Symposium Series*, 41:95–98, 1999.

[246] S. L. Kosakovsky Pond, A. F. Y. Poon, R. Velazquez, S. Weaver, N. L. Hepler, B. Murrell, S. D. Shank, B. R. Magalis, D. Bouvier, A. Nekrutenko, S. Wisotsky, S. J. Spielman, S. D. W. Frost, and S. V. Muse. Hyphy 2.5-a customizable platform for evolutionary hypothesis testing using phylogenies. *Mol. Biol. Evol.*, 37:295–299, 2020.

[247] A. Di Franco, R. Poujol, D. Baurain, and H. Philippe. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.*, 19:21, 2019.

[248] D. P. Martin, S. Lytras, A. G. Lucaci, W. Maier, B. Grüning, S. D. Shank, S. Weaver, O. A. MacLean, R. J. Orton, P. Lemey, M. F. Boni, H. Tegally, G. W. Harkins, C. Scheepers, J. N. Bhiman, J. Everatt, D. G. Amoako, J. E. San, J. Giandhari, A. Sigal, and et al. Selection analysis identifies clusters of unusual mutational changes in omicron lineage ba.1 that likely impact spike function. *Mol. Biol. Evol.*, 39, 2022. URL doi:10.1093/molbev/msac061.

[249] L.-T. Nguyen, H. A. Schmidt, and A. von Haeseler. Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32:268–274, 2015.

[250] B. Q. Minh, M. A. T. Nguyen, and A. von Haeseler. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.*, 30:1188–1195, 2013.

[251] J. Huerta-Cepas, F. Serra, and P. Bork. Ete 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.*, 33:1635–1638, 2016.

[252] X. H. Zhou and S. Gao. Confidence intervals for the log-normal mean. *Stat. Med.*, 16: 783–790, 1997.

[253] M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut. Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus Evol.*, 4: vey016, 2018.

[254] M. A. R. Ferreira and M. A. Suchard. Bayesian analysis of elapsed times in continuous-time markov chains. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 36:355–368, 2008.

[255] A. Rambaut, A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard. Posterior summarization in bayesian phylogenetics using tracer 1.7. *Syst. Biol.*, 67:901–904, 2018.

[256] M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, and M. A. Suchard. Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.*, 30:713–724, 2013.

[257] X. Ji, A. A. Fisher, S. Su, J. L. Thorne, B. Potter, P. Lemey, G. Baele, and M. A. Suchard. Scalable bayesian divergence time estimation with ratio transformations. *arXiv [q-bio.PE]*, 2021. URL http://arxiv.org/abs/2110.13298.

[258] X. Ji, Z. Zhang, A. Holbrook, A. Nishimura, G. Baele, A. Rambaut, P. Lemey, and M. A. Suchard. Gradients do grow on trees: A linear-time o(n)-dimensional gradient for statistical phylogenetics. *Mol. Biol. Evol.*, 37:3047–3060, 2020.

[259] D. L. Ayres, M. P. Cummings, G. Baele, A. E. Darling, P. O. Lewis, D. L. Swofford, J. P. Huelsenbeck, P. Lemey, A. Rambaut, and M. A. Suchard. Beagle 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst. Biol.*, 68:1052–1061, 2019.

[260] A. A. Fisher, X. Ji, Z. Zhang, P. Lemey, and M. A. Suchard. Relaxed random walks at scale. *Syst. Biol.*, 70:258–267, 2021.

[261] S. Dellicour, R. Rose, N. R. Faria, P. Lemey, and O. G. Pybus. Seraphim: studying environmental rasters and phylogenetically informed movements. *Bioinformatics*, 32: 3204–3206, 2016.

[262] H. Zhou, J. Ji, X. Chen, Y. Bi, J. Li, Q. Wang, T. Hu, H. Song, R. Zhao, Y. Chen, M. Cui, Y. Zhang, A. C. Hughes, E. C. Holmes, and W. Shi. Identification of novel bat coronaviruses sheds light on the evolutionary origins of sars-cov-2 and related viruses. *Cell*, 184:4380–4391.e14, 2021.

[263] GBIF.org. Gbif occurrence download, 2023. URL https://doi.org/10.15468/dl.xc2jvq.

[264] GBIF.org. Gbif occurrence download, 2022. URL https://doi.org/10.15468/dl.82sgat.

[265] K. C. Tanalgo, J. A. G. Tabora, H. F. M. de Oliveira, D. Haelewaters, C. T. Beranek, A. Otálora-Ardila, E. Bernard, F. Gonçalves, A. Eriksson, M. Donnelly, J. M. González, H. F. Ramos, A. C. Rivas, P. W. Webala, S. Deleva, R. Dalhoumi, J. Maula, D. Lizarro, L. F. Aguirre, M. N. R. M. Quibod, J. Barros, M. A. Turcios-Casco, M. Martínez, D. I. Ordoñez-Mazier, J. A. S. Orellana, E. J. Ordoñez-Trejo, D. Ordoñez, A. Chornelia, J. M. Lu, C. Xing, S. Baniya, R. L. Muylaert, L. H. Dias-Silva, N. Ruadreo, and A. C. Hughes. Darkcides 1.0, a global database for bats in karsts and caves. *Scientific Data*, 9:155, 2022.

[266] J. Alroy. Latitudinal gradients in the ecology of new world bats. *Glob. Ecol. Biogeogr.*, 28: 784–792, 2019.

[267] D. J. Becker, D. E. Crowley, A. D. Washburne, and R. K. Plowright. Temporal and spatial limitations in global surveillance for bat filoviruses and henipaviruses. *Biol. Lett.*, 15: 20190423, 2019.

[268] C. Di Gregorio, M. Iannella, and M. Biondi. Revealing the role of past and current climate in shaping the distribution of two parapatric european bats, myotis daubentonii and m. capaccinii. *The European Zoological Journal*, 88:669–683, 2021.

[269] A. Jargalsaikhan, B. Ariunbold, M. Dalannast, E. Purevee, W. K. Paek, K. Tsagaan, and O. Ganbold. Molecular phylogenetic analysis of bats in the family vespertilionidae in mongolia. *Journal of Asia-Pacific Biodiversity*, 15:329–335, 2022.

[270] S. V. Kruskop, I. V. Artyushin, A. P. Yuzefovich, E. Undrakhbayar, A. S. Speranskya, A. A. Lisenkova, A. A. Bannikova, and V. S. Lebedev. Genetic diversity of mongolian long-eared bats (plecotus; vespertilionidae; chiroptera). *acta.*, 22:243–255, 2020.

[271] H.-C. Kuo, S.-F. Chen, Y.-P. Fang, J. Flanders, and S. J. Rossiter. Comparative rangewide phylogeography of four endemic taiwanese bat species. *Mol. Ecol.*, 23:3566–3586, 2014.

[272] Y. Mokrani, F. Mimeche, Y. Nouidjem, and M. Saheb. Rapid assessment of cave-dwelling bat diversity in the chebket es-sellaoua mountains (eastern algeria). *Arx. Misc. Zool.*, 16: 112–120, 2018.

[273] S. Raman, T. T. Shameer, U. Pooja, and A. C. Hughes. Identifying priority areas for bat conservation in the western ghats mountain range, peninsular india. *J. Mammal.*, 104: 49–61, 2022.

[274] D. Scherrer, P. Christe, and A. Guisan. Modelling bat distributions and diversity in a mountain landscape using focal predictors in ensemble of small models. *Divers. Distrib.*, 25:770–782, 2019.

[275] A. C. Hughes, C. Satasook, P. J. J. Bates, S. Bumrungsri, and G. Jones. The projected effects of climatic and vegetation changes on the distribution and diversity of southeast asian bats. *Glob. Chang. Biol.*, 18:1854–1865, 2012.

[276] A. C. Hughes. Mapping priorities for conservation in southeast asia. *Biol. Conserv.*, 209: 395–405, 2017.

[277] P. Brun, N. E. Zimmermann, C. Hari, L. Pellissier, and D. Karger. Chelsa-bioclim+ a novel set of global climate-related predictors at kilometre-resolution. EnviDat. https://doi.org/10.16904/envidat.332, 2022.

[278] A. Trabucco and R. J. Zomer. Global high-resolution soil-water balance. figshare. Dataset. Https://doi.org/10.6084/m9.figshare.7707605.v3, 2019.

[279] N. Lang, W. Jetz, K. Schindler, and J. D. Wegner. A high-resolution canopy height model of the earth. Available at https://langnico.github.io/globalcanopyheight/, 2019.

[280] P. Potapov, X. Li, A. Hernandez-Serna, A. Tyukavina, M. C. Hansen, A. Kommareddy, A. Pickens, S. Turubanova, H. Tang, C. E. Silva, J. Armston, R. Dubayah, J. B. Blair, and M. Hofton. Mapping global forest canopy height through integration of gedi and landsat data. *Remote Sens. Environ.*, 253:112165, 2021.

[281] T. W. Crowther, H. B. Glick, K. R. Covey, C. Bettigole, D. S. Maynard, S. M. Thomas, J. R. Smith, G. Hintler, M. C. Duguid, G. Amatulli, M.-N. Tuanmu, W. Jetz, C. Salas, C. Stam, D. Piotto, R. Tavani, S. Green, G. Bruce, S. K. Williams, S. K. Wiser, M. O. Huber, G. M. Hengeveld, G.-J. Nabuurs, E. Tikhonova, P. Borchardt, C.-F. Li, L. W. Powrie, M. Fischer, A. Hemp, J. Homeier, P. Cho, A. C. Vibrans, P. M. Umunay, S. L. Piao, C. W. Rowe, M. S. Ashton, P. R. Crane, and M. A. Bradford. Mapping tree density at a global scale. *Nature*, 525:201–205, 2015.

[282] T. Hengl, J. Mendes de Jesus, G. B. M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N. H. Batjes, J. G. B. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel, and B. Kempen. Soilgrids250m: Global gridded soil information based on machine learning. *PLoS One*, 12:e0169748, 2017.

[283] S. J. Phillips, M. Dudík, and R. E. Schapire. A maximum entropy approach to species distribution modeling. page 83, 2004.