# UC San Diego

## UC San Diego Electronic Theses and Dissertations

### Title

Circulation and selection of zoonotic viral epidemics

### Permalink

https://escholarship.org/uc/item/6x2232ks

### Author

Havens, Jennifer L

### Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

# Circulation and selection of zoonotic viral epidemics

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

in

Bioinformatics & Systems Biology

by

Jennifer L. Havens

Committee in charge:

Professor Joel O. Wertheim, Chair
Professor Matt Daugherty, Co-Chair
Professor Vineet Bafna
Professor Niema Moshiri
Professor Tetyana Vasylyeva

2024

Copyright

Jennifer L. Havens, 2024

All rights reserved.

The Dissertation of Jennifer L. Havens is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

# DEDICATION

This dissertation is dedicated to those without whom this work would not be possible:

To my teachers and mentors

To my friends

To my family, by blood and by choice

To healthcare workers, and the individuals who have supported them

Thank you.

TABLE OF CONTENTS

Dissertation Approval Page ..... iii

Dedication ..... iv

Table of Contents ..... v

List of Figures ..... viii

List of Tables ..... x

Acknowledgements ..... xi

Vita ..... xiv

Abstract of the Dissertation ..... xvi

Introduction ..... 1

Chapter 1 The role of socio-economic disparities in the relative success and persistence of SARS-CoV-2 variants in New York City in early 2021 ..... 6
1.1 Introduction ..... 7
1.2 Materials and methods ..... 8
1.2.1 Ethical considerations ..... 8
1.2.2 Genetic data ..... 8
1.2.3 Seroprevalence level and epidemiological data ..... 9
1.2.4 Statistical analyses ..... 10
1.2.5 BEAST analyses ..... 10
1.2.6 Persistence analysis ..... 11
1.3 Results ..... 12
1.3.1 Geographical patterns in seropositivity ..... 12
1.3.2 Correlation of variant prevalence and seropositivity ..... 13
1.3.3 Correlation between seropositivity and socio-demographic data ..... 15
1.3.4 Phylogeographic analysis findings ..... 16
1.3.5 Analysis of persisting lineages ..... 17
1.4 Discussion ..... 19
1.5 Acknowledgments ..... 21
1.5.1 Financial Disclosure statement ..... 21
1.5.2 Author contributions ..... 22

Chapter 2 Synonymous substitution rates preceding the emergence of SARS-CoV-2 variants and during persistent infections ..... 23
2.1 Introduction ..... 23
2.2 Results ..... 26
2.2.1 Identification of persistent infections ..... 26
2.2.2 Variant stem lineages and deer clades ..... 27
2.2.3 Rate of synonymous substitutions ..... 28

2.2.4 Rate of nonsynonymous substitutions . . . 30
2.2.5 Molecular clock and dN/dS . . . 30
2.3 Discussion . . . 32
2.4 Methods . . . 34
2.4.1 Non-variant context . . . 34
2.4.2 Variant stem lineages . . . 34
2.4.3 Persistent intra-host infections . . . 35
2.4.4 Deer associated evolution . . . 35
2.4.5 Bayesian local clock inference and robust counting . . . 35
2.4.6 Inference of dN/dS . . . 36
2.5 Acknowledgments . . . 37

Chapter 3 Dynamics of natural selection preceding human viral epidemics and pandemics 38
3.1 Introduction . . . 38
3.2 Results . . . 40
3.2.1 Selection regime preceding and during epidemics with zoonotic origin . . . 40
3.2.2 Selection regimes in Betacoronaviruses prior to spillovers . . . 44
3.2.3 The reemergence of H1N1 in 1977 . . . 46
3.2.4 Laboratory passage of H1N1 influenza A viruses . . . 48
3.2.5 Selection on attenuated vaccine viruses . . . 48
3.2.6 Serially-passaged coronavirus . . . 49
3.2.7 Artificial selection on influenza A viruses . . . 49
3.3 Discussion . . . 50
3.4 Methods . . . 52
3.4.1 Phylogenic selection analysis . . . 52
3.4.2 Data sets . . . 53
3.4.3 Selection regime preceding and during the H1N1 2009 pandemic . . . 53
3.4.4 Selection regime preceding and during Ebolavirus (EBOV) 2013-2015 West Africa outbreak . . . 54
3.4.5 Selection regime preceding and during Marburg virus disease Angola outbreak . . . 54
3.4.6 Selection regimes in SARS-CoV-like virus . . . 54
3.4.7 Selection regimes in SARS-CoV-2-like virus . . . 55
3.4.8 Selection of H1N1 influenza A viruses . . . 55
3.4.9 Selection of H1N1 influenza A viruses putatively from lake ice . . . 56
3.4.10 Selection on attenuated vaccine measles virus (MeV) . . . 56
3.4.11 Selection on attenuated vaccine mumps virus (MuV) . . . 57
3.4.12 Serially-passaged coronavirus MHV . . . 57
3.4.13 Artificial selection on influenza A virus A/H2N2 . . . 57
3.4.14 Artificial selection on influenza A virus A/H5N1 . . . 57
3.5 Acknowledgments . . . 58

Chapter 4 Phylogeographic analysis reveals an ancient East African origin of human herpes simplexvirus 2 dispersal out-of-Africa . . . 59
4.1 Introduction . . . 59
4.2 Results . . . 61
4.2.1 HSV-2 genomes. . . . 61

4.2.2 Maximum likelihood phylogenetic inference. . . . . 61
4.2.3 Clock rate inference of empirical sequences. . . . . 63
4.2.4 Likelihood of empirical clock rates. . . . . 63
4.2.5 Estimating clock rate of simulated data. . . . . 64
4.2.6 Timing the ancestor of HSV-2. . . . . 66
4.2.7 Phylogeographic analysis indicates East African origin of worldwide lineage. 70
4.3 Discussion . . . . . 71
4.4 Methods . . . . . 74
4.4.1 HSV-2 data set and sequence alignment. . . . . 74
4.4.2 Phylogenetic and initial temporal analysis. . . . . 75
4.4.3 Likelihood of fixed clock rates with empirical data. . . . . 76
4.4.4 Clock rate estimates of simulated data. . . . . 76
4.4.5 Phylogeographic ancestral state reconstruction. . . . . 77
4.4.6 Data Availability . . . . . 78
4.4.7 Code Availability . . . . . 78
4.4.8 Author Contributions Statement . . . . . 78
4.4.9 Competing Interests Statement . . . . . 78
4.5 Acknowledgments . . . . . 79

Epilogue . . . . . 80

Appendix A Supplemental Material for Chapter 1 . . . . . 84
A.1 Supplementary Table . . . . . 84
A.2 Supplementary Figures . . . . . 85
A.3 Supplementary Data Captions . . . . . 88

Appendix B Supplemental Material for Chapter 2 . . . . . 89
B.1 Supplementary data . . . . . 89
B.2 Supplementary Tables . . . . . 89

Appendix C Supplemental Material for Chapter 3 . . . . . 91
C.1 Supplementary Figures . . . . . 91
C.2 Supplementary Tables . . . . . 99

Appendix D Supplemental Material for Chapter 4 . . . . . 101
D.1 Supplementary Figures . . . . . 101
D.2 Supplementary Data Captions . . . . . 104
D.3 Supplementary Tables . . . . . 104

Bibliography . . . . . 106

LIST OF FIGURES


Figure 1.1. Seroprevalence of SARS-CoV-2 antibodies across NYC in November 2020. . 13

Figure 1.2. SARS-CoV-2 variants and demographics across ZIP codes in NYC. ..... 15

Figure 1.3. Seroprevalence quantiles across NYC boroughs. ..... 17

Figure 1.4. Persistence of variant lineages over time in NYC stratified by seroprevalence level. ..... 18

Figure 2.1. Phylogenetic trees of persistent infections. ..... 26

Figure 2.2. Phylogenetic trees of variant stem lineages. ..... 27

Figure 2.3. Rate of nonsynonymous and synonymous substitutions and molecular clock. 29

Figure 3.1. Quantifying selection regimes of key epidemics with zoonotic origin. ..... 41

Figure 3.2. Quantifying selection regime in SARS-CoV. ..... 43

Figure 3.3. Quantifying selection regimes in SARS-CoV-2. ..... 45

Figure 3.4. Quantifying selection regimes in H1N1. ..... 47

Figure 3.5. Quantifying selection regimes of laboratory passaged viruses ..... 49

Figure 4.1. Maximum likelihood phylogenetic analysis of empirical HSV-2 genomes. ... 62

Figure 4.2. Shape of likelihood surface for HSV-2 clock rate on empirical. ..... 64

Figure 4.3. Performance of clock rate inference on simulated data. ..... 65

Figure 4.4. Phylogeography of HSV-2 supports East African origin of out-of-Africa dispersal. ..... 69

Figure 4.5. Phylogeography of HSV-2 supports East African origin of out-of-Africa dispersal. ..... 71

Figure A.1. Proportion of sequenced samples by ZIP code. ..... 85

Figure A.2. ZMaximum clade credibility phylogenetic trees with branches annotated by location seroprevalence inferred in phylogeographic analysis in BEAST. ... 86

Figure A.3. Number of introductions from and to each location in phylogeographic analysis where locations were defined based on seropositivity only as High, Medium (Mid), and Low. ..... 87

Figure C.1. Quantifying selection regimes in SARS-CoV-2 with human outbreak. ..... 91

Figure C.2. Phylogenetic tree of H1N1 HA segment, branch color indicates partition. . . 92

Figure C.3. Phylogenetic tree of H1N1 HA segment, branch color indicates partition. . . 93

Figure C.4. Phylogenetic tree of MeV genome, branch color indicates partition. . . . . 94

Figure C.5. Phylogenetic tree of MuV genome, branch color indicates partition. . . . . 95

Figure C.6. Phylogenetic tree of MHV. . . . . 95

Figure C.7. Phylogenetic tree of H2N2 PB1 segment. . . . . 96

Figure C.8. Phylogenetic tree of H5N1 PB1 segment. . . . . 97

Figure C.9. Diagram of host during evolutionary history. . . . . 98

Figure D.1. Phylogenetic analysis of subset of HSV-2 worldwide lineage. . . . . 101

Figure D.2. Shape of likelihood surface for clock rate of ebolavirus. . . . . 102

Figure D.3. Breakpoint analysis of clock rate inference on simulated data. . . . . 103

Figure D.4. Phylogeography of HSV-2 supports East Africa as origin of out-of-Africa dispersal. . . . . 103

Figure D.5. Internal nodes in backbone of tree around out of Africa migration (within the worldwide lineage). . . . . 104

LIST OF TABLES

Table 2.1. Estimate of dN/dS . . . . . 31

Table 4.1. Age of root and out-of-Africa dispersal across clock rates. . . . . 68

Table A.1. GLM model coefficients . . . . . 84

Table B.1. Molecular clock and rate of nonsynonymous and synonymous substitutions . 90

Table B.2. Synonymous substitutions of variant stem lineages . . . . . 90

Table C.1. Inferred substitutions of 1977 reemergence . . . . . 99

Table C.2. Results of RELAX test . . . . . 100

Table D.1. Age of root and out-of-Africa dispersal across clock rates for robustness analysis of geographic partitioning . . . . . 105

ACKNOWLEDGEMENTS

I want to thank all of the mentors that I have had that offered their time and experience. First and foremost I want to thank Joel Wertheim. Joel has been a constant source of knowledge, advice and support throughout graduate school. He inspired my interest in viruses and introduced me to a new way of doing science. I am also thankful to the rest of my committee, Vineet Bafna, Matt Daugherty, Niema Moshiri, and Tanya Vasylyeva. Specifically, I want to thank Tanya who was always open for conversations and giving advice. And to Vineet, whose commitment to graduate students is an inspiration I will carry with me. I would also like to thank Nicole Scharping for her guidance and shared discussions about careers in research and the university environment.

I had the privilege of wonderful research mentors prior to starting graduate school. Thank you to professors Daniel Martínez and Catherine McFadden, my first research advisers who opened the door to research for me. To Danae Schulz, who guided my first major research project and scientific write up. I would also like to thank Paul Spellman, who was endlessly supportive of my upcoming graduate school career, and was the one to suggest I apply to UCSD BISB.

I would like to thank friends and collaborators who supported both my research and myself, and helped to make an exciting and positive research community. In particular Jonathan Pekar, and Tanya who made my lab experience. As well as Jade Wang and Jordan Zehr.

Teaching shaped my own graduate school education, and I am thankful to all of the other teachers I have worked with and learned from, in particular Pablo Dodero, Haowei Li, Niema, Alex Phan, and Lisa Trahan.

Student (and post-doc) engagement and advocacy were important in my time at UCSD. I would like to thank all those that I worked with, many who became friends, I could accomplish nothing alone - but together so much was achieved. I would also like to thank the people who came before and after, you set us up for success and give me hope for the future. There are so many people involved and I cannot mention everyone but in particular I am thankful to Lauryn Bruce, Michael Cuoco, Lia Gale, Caitlin Guccione, Julia Hardy, Hsuan-lin Her, Pratibha Jagannatha, Alex Jambor, Adam Klie, Hema Kopalle, Douglas Meyer, Hannah Mummey, Gege Qian, and Alexander Wenzel. Thank you to my fiends and peers from Bioinformatics & Systems

Biology, George Armstrong, Owen Chapman, Michelle Franc Ragsac, Adam Jussila, Emily Kobayashi, Clarence Mah, Arya Massarat, Jacklyn Newsome, Jennifer Nguyen, Adam Officer, Gibraan Rahman, Jesse Solvason, and Nancy Yuan, who helped make my time at UCSD filled with good times and supported me during the hard times. Thank you to the whole BISB 2018 cohort as well as all my other friends from USCD who I haven't had the chance to name here. I would like to thank my friends from Mudd, Quentin Barth, Stetson Bost, Julia Cosma, Tiffany Fong, Juliet Forman, Sarah Hale, Marina Knittel, Ben Leghman, Liam Lyod, Geneva Miller, Jake Palanker, Lindsay Popowski, Paige Rinnert, Jacob Roth, Mikayla Stabile, Zoe Tucker, Lewis Viornery, and the whole timesuck crew. And thank you to my friends from home Christiane Adcock, and Alyssa Blue.

I would like to thank the staff, in particular Elyse Witt, and other volunteers (plus the animals) I worked with at the San Diego Society that got me out of my apartment and offered a friendly face every week; the constant interest and support meant so much to me.

I would also like to thank my family, so many people who are cheering me on. In particular Chris, Jeff, and Candy Havens, who have offered their support not just during graduate school, but also all the support that helped me get here.

Finally I am grateful to Sonja Huč, an amazing friend, biologist, and teacher. Sonja provided constant support and discussions about research, teaching, and life. I look forward to writing sessions every week because of her, and without this support this document might just be blank.

Chapter 1, in full, is a reformatted reprint, of the material as it appears in "The role of socio-economic disparities in the relative success and persistence of SARS-CoV-2 variants in New York City in early 2021." Tetyana I. Vasylyeva, Jennifer L. Havens, Jade C. Wang, Elizabeth Luoma, Gabriel W. Hassler, Helly Amin, Steve Di Lonardo, Faten Taki, Enoma Omoregie, Scott Hughes, and Joel O. Wertheim, *Plos Pathogens*, 2024. The dissertation author was the co-primary investigator and co-first author of this paper.

Chapter 2, in full, is a reformatted reprint, of the material as it appears in "Synonymous substitution rates preceding the emergence of SARS-CoV-2 variants and during persistent infections." Jennifer L. Havens, Karthik Gangavarapu, Jade C. Wang, Elizabeth Luoma, Jonathan

E. Pekar, Helly Amin, Steve Di Lonardo, Faten Taki, Enoma Omoregie, Scott Hughes, Kristian G. Andersen, Tetyana I. Vasylyeva, Marc Suchard, and Joel O. Wertheim. *In Preparation*, 2024. The dissertation author was the primary investigator and first author of this text.

Chapter 3, in full, is a reformatted reprint, of the material as it appears in "Dynamics of natural selection preceding human viral epidemics and pandemics." Jennifer L. Havens, Sergei L. Kosakovsky Pond, Jordan D. Zehr, Jonathan E. Pekar, Edyth Parker, Stephen A. Goldstein, Michael Worobey, Kristian G. Andersen, and Joel O. Wertheim. *In Preparation*, 2024. The dissertation author was the primary investigator and first author of this text.

Chapter 4, in full, is a reformatted reprint, of the material as it appears in "Phylogeographic analysis reveals an ancient East African origin of human herpes simplex virus 2 dispersal out-of-Africa." Jennifer L. Havens, Sébastien Calvignac-Spencer, Kevin Merkel, Sonia Burrel, David Boutolleau, and Joel O. Wertheim. *Nature communications*, 2022. The dissertation author was the primary investigator of this paper and first author of this text.

VITA

2018–2024 Ph.D. in Bioinformatics & Systems Biology
University of California San Diego

2014–2018 B.S. in Mathematical and Computational Biology
Harvey Mudd College

PUBLICATIONS

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

**Jennifer L. Havens**, Sergei L. Kosakovsky Pond, Jordan D. Zehr, Jonathan E. Pekar, Edyth Parker, Stephen A. Goldstein, Michael Worobey, Kristian G. Andersen, and Joel O. Wertheim. 2024. "Dynamics of natural selection preceding human viral epidemics and pandemics". *In preparation* △

**Jennifer L. Havens**, Karthik Gangavarapu, Jade C. Wang, Elizabeth Luoma, Jonathan E. Pekar, Helly Amin, Steve Di Lonardo, Faten Taki, Enoma Omoregie, Scott Hughes, Kristian G. Andersen, Tetyana I. Vasylyeva, Marc Suchard, Joel O. Wertheim. 2024. "Synonymous substitution rates preceding the emergence of SARS-CoV-2 variants and during persistent infections." *In preparation* △

Tetyana I. Vasylyeva†, **Jennifer L. Havens**† , Jade C. Wang, Elizabeth Luoma, Gabriel W. Hassler, Helly Amin, Steve Di Lonardo, et al. 2023. "The role of socio-economic disparities in the relative success and persistence of SARS-CoV-2 variants in New York City in early 2021." *Plos Pathogens.* △

Jonathan E. Pekar, Spyros Lytras, Mahan Ghafari, Andrew F. Magee, Edyth Parker, **Jennifer L. Havens**, Aris Katzourakis, et al. 2023. "The Recency and Geographical Origins of the Bat Viruses Ancestral to SARS-CoV and SARS-CoV-2." *bioRxiv.*

Ethan C. Ashby, **Jennifer L. Havens**, Lindsey M. Rollosson, Johanna Hardin, and Danae Schulz. 2023. "Chemical Inhibition of Bromodomain Proteins in Insect-Stage African Trypanosomes Perturbs Silencing of the Variant Surface Glycoprotein Repertoire and Results in Widespread Changes in the Transcriptome." *Microbiology Spectrum* 11 (3): e00147-23.

Tetyana I. Vasylyeva, Courtney E. Fang, Michelle Su, **Jennifer L. Havens**, Edyth Parker, Jade C. Wang, Mark Zeller, et al. 2022. "Introduction and Establishment of SARS-CoV-2 Gamma Variant in New York City in Early 2021." *The Journal of Infectious Diseases* 226 (12): 2142–49. https://doi.org/10.1093/infdis/jiac265.

Jonathan E. Pekar, Andrew Magee, Edyth Parker, Niema Moshiri, Katherine Izhikevich, **Jennifer L. Havens**, Karthik Gangavarapu, et al. 2022. "The Molecular Epidemiology of Multiple

Zoonotic Origins of SARS-CoV-2." *Science* 377 (6609): 960–66.

Edyth Parker, Catelyn Anderson, Mark Zeller, Ahmad Tibi, **Jennifer L. Havens**, Geneviève Laroche, Mehdi Benlarbi, et al. 2022. "Regional Connectivity Drove Bidirectional Transmission of SARS-CoV-2 in the Middle East during Travel Restrictions." *Nature Communications* 13 (1): 4784. https://doi.org/10.1038/s41467-022-32536-1.

Joel O. Wertheim, Jade C. Wang, Mindy Leelawong, Darren P. Martin, **Jennifer L. Havens**, Moinuddin A. Chowdhury, Jonathan E. Pekar, et al. 2022. "Detection of SARS-CoV-2 Intra-Host Recombination during Superinfection with Alpha and Epsilon Variants in New York City." *Nature Communications* 13 (1): 3645.

**Jennifer L. Havens**, Sébastien Calvignac-Spencer, Kevin Merkel, Sonia Burrel, David Boutolleau, and Joel O. Wertheim. 2022. "Phylogeographic Analysis Reveals an Ancient East African Origin of Human Herpes Simplex Virus 2 Dispersal Out-of-Africa." *Nature Communications* 13 (1): 5477. △

Anthony P. West, Joel O. Wertheim, Jade C. Wang, Tetyana I. Vasylyeva, **Jennifer L. Havens**, Moinuddin A. Chowdhury, Edimarlyn Gonzalez, et al. 2021. "Detection and Characterization of the SARS-CoV-2 Lineage B.1.526 in New York." *Nature Communications* 12 (1): 4886. https://doi.org/10.1038/s41467-021-25168-4.

ABSTRACT OF THE DISSERTATION

**Circulation and selection of zoonotic viral epidemics**

by

Jennifer L. Havens

Doctor of Philosophy in Bioinformatics & Systems Biology

University of California San Diego, 2024

Professor Joel O. Wertheim, Chair
Professor Matt Daugherty, Co-Chair

Human history has been marked by viral epidemics of zoonotic origin which have evolved and spread through the human population with devastating effect. In this dissertation, I explore the origins of zoonosis, adaptation, and spread of viruses in heterogeneous human populations. To understand how viral epidemics can circulate in the human population, I first investigate how the first wave of SARS-CoV-2 in New York City impacted different areas and with this context I explore the dynamics of how different variants circulate in these areas during the second wave. Then, I examine the context in which SARS-CoV-2 variants arose, using a synonymous molecular clock to characterize evolution in different environments. I expand on understanding selection, by examining the selection regime in viral host reservoirs, selection preceding human

outbreaks, and selection after zoonosis during human outbreaks for viruses including Ebola virus, Marburg virus, influenza A virus, SARS-CoV, and SARS-CoV-2. Finally, I explore how herpes simplex virus 2 migrated globally and characterize the uncertainty by developing a simulation framework. By understanding the path and evolutionary history of viruses, I hope to contribute to the fundamental understanding of epidemics which is needed to make informed and effective plans to mitigate harm, and perhaps even prevent viral outbreaks in the future.

# Introduction

Viruses from animals regularly jump into humans, but in many cases they do not spread past the primary case. However, sometimes viruses begin to transmit between people, ultimately impacting many more. It is the later situation that I present on in this dissertation: how viruses spread and evolve in the human population.

SARS-CoV-2 jumped into the human population from a non-human animal host, a zoonotic jump, in late 2019 in China[1]. First circulating within the region, SARS-CoV-2 quickly began spreading to other areas and around the world. Within three years there were more than 670 million cases and 6.8 million deaths from SARS-CoV-2[2]. SARS-CoV-2 is not unique, many other zoonotic viruses have entered the human population and continued to circulate.

In 2002 SARS-CoV first entered the human population, close contact between humans and infected palm civets likely facilitated repeated spillover. While it primarily impacted China and Hong Kong, SARS-CoV was considered a global threat and more than 750 people died[3]. In April 2009, the first human cases of a reassortment of H1N1 influenza A virus (H1N1pdm2009) that had been circulating in swine were found in North America. From there, H1N1pdm2009 spread globally and was considered a pandemic by June 2009. From April 2009 to August 2010 there were 18,500 laboratory confirmed deaths, and modeling suggests there were over 280,000 deaths associated with H1N1pdm2009 infection[4]. In 2013 Ebola virus emerged in West Africa, causing the largest human outbreak of Ebola up until that time. The size of the epidemic was attributed to Ebola virus entering dense urban populations[5]. The epidemic resulted in over 11,300 deaths[6].

These recent zoonotic epidemics are a significant health burden on the global population, and they will likely not be the last. With continued interaction at the human-animal interface along with environmental changes it is likely that we will continue be exposed to novel zoonotic

viruses, and with every exposure there is a chance to spark a new epidemic[7;8]. Understanding the dynamics of how viruses have evolved and spread in the past is critical to making an informed and effective response in the future.

Viruses evolve by acquiring substitutions, changes in the genome sequence, that can impact viral traits such as transmissibility and immune evasiveness. Genetic drift and selection pressure can impact what substitutions are acquired and maintained in virus populations. By looking at extant virus genomes we can infer a history of acquiring substitutions using phylogenetics, and in the context of this history we can probe the environment that the virus evolved under. Information available today tells us a story about the past.

In a phylogenetic approach to investigating viruses, we infer a (often bifurcating) tree where the tips are observed virus genomes and internal nodes represent the most recent common ancestor (MRCA) of the descendant branches. This tree inference is based on the understanding that substitutions in the genome are passed down and we expect viruses with similar genomes to be more closely related (have a more recent MRCA) than viruses with more dissimilar genomes. We can associate substitutions to the branches of the phylogenetic tree to understand when substitutions were acquired overtime, and insight into the importance of these changes.

Some parts of the virus genome, the non-coding regions, do not encode for virus protein and it is typically thought to be important for regulatory and structural reasons. In contrast, the coding region encodes virus proteins; it is these regions that I will focus on. Substitutions in the coding region can be divided into two classes: synonymous substitutions, which do not change the encoded amino acid, and nonsynonymous substitutions, which do. Nonsynonymous substitutions tend to be more impacted by selection because changes in the amino acid sequence can have a large impact on virus fitness while synonymous substitutions do not have such downstream impacts. Synonymous substitutions are typically thought to occur at a relatively constant rate[9–11]. Thus, the ratio of nonsynonymous to synonymous substitutions per site (dN/dS) can give information about the strength of selection.

Under neutral selection, when there is little selective pressure, dN/dS is expected to be 1. dN/dS greater than 1 represents a strong selective pressure to change, this is positive or diversifying selection. Positive selection at specific sites of interest, may be indicative of

adaptation and has been of great interest in research[12–14]. In viruses it is common for a few key sites to be under positive selection which can, for example, help the virus evade immune detection[14;15]. In contrast, dN/dS less than 1 represents purifying selection which eliminates changes and is important in maintaining optimum functionality in an environment. The majority of sites in virus genomes are primarily are under purifying selection[9;16;17]. Viruses are constantly under selective pressure, but as the environment changes the strength of selection can change. By characterizing selection in the past we can gain information on the history of what environment and pressures viruses have experienced.

Genome sequence and selection is not the only thing that can be associated with the evolutionary history; the geographic history and circulation dynamics can be reconstructed in the context of a phylogenetic tree. By looking at the geographic locations of virus samples at the tips, we can walk back up the tree to the base and infer the most likely geographic state (i.e. location) of the virus over time. This phylogeographic inference can tell the story of how a virus may have spread across space. There are maximum likelihood[18] and Bayesian[19] phylogeographic approaches for this inference.

In Chapter 1, I focus on SARS-CoV-2 in New York City (NYC) and investigate how SARS-CoV-2 variants impacted different areas within the heterogeneous city. SARS-CoV-2 evolution is primarily characterized by transmission of closely related viruses with identical or nearly identical genomes interspersed with jumps in evolution which precedes the emergence of variants. SARS-CoV-2 variants are often more transmissible or immune evasive than the strain they descend from and once they emerge, they can spread rapidly[20]. Alpha, the first named variant, was characterized by high transmissibility relative to the original strain[21]. While Iota, which arose in NYC, was characterized by immune-evasive properties[21;22]. I present the amount of circulation of these two variants in different neighborhoods of NYC. While direct observation of transmission chains was impossible due to asymptomatic cases and the large number of cases, by using phylogeographic reconstruction I infer how SARS-CoV-2 variants circulated and persisted within and between areas. I then demonstrate how relative success of variants with different traits are associated with socioeconomic factors.

Given the importance of variants for the spread of SARS-CoV-2, in Chapter 2, I investigate

the selection dynamics that precede the emergence of variants. In particular, I quantify selection as measured by dN/dS and the rate of synonymous substitutions on the stem lineages preceding variants, to understand how these variants are evolving. I then compare and contrast the evolution preceding variants with evolution during persistent infection and within non-human animal hosts, which are two proposed sources of variants[23;24]. In this chapter I present on how SARS-CoV-2 evolves once it began to circulate in the human population; how SARS-CoV-2 first arose in humans is also of interest.

The traditional framework to understand zoonosis posits that zoonotic viruses must evolve the traits necessary to sustain human-to-human transmission outside of the natural host reservoir, but before seeding a sustained human outbreak, such as during a cycle of jumps between non-human animal to human and back[25;26]. In Chapter 3. I present on selection around the time of zoonosis. The initial stages of zoonosis are rarely directly observed, however using phylogenetic inference we can reconstruct some of the history of early outbreaks. Specifically, I characterize selection in the natural host reservoir: the population of non-human animals where the virus circulates, selection preceding zoonosis, and selection after zoonosis during human outbreaks. By contrasting selection in these different contexts, I demonstrate where adaptation to humans occurs: in non-human animals or in humans. Understanding where recent zoonotic viruses evolve contextualizes how viruses can enter and spread in the heterogeneous human population.

Finally, in Chapter 4, I present the global spread of herpes simplex virus 2 (HSV-2). HSV-2 is an ancient zoonotic virus[27]. HSV-2 emerged and spread before our modern understanding of viruses as agents of disease and so direct observation of HSV-2 spread is impossible. To understand how it spread we must use tools which can give us insight into the past with data available today. Using a maximum likelihood phylogeographic approach, and developing a simulation framework to evaluate the uncertainty, I infer the path and timing of HSV-2 worldwide migration. In this chapter I will describe the most likely scenario of the migration of HSV-2, but also the bounds of what is possible to quantify given the limited information available.

Altogether, in this dissertation I investigate how zoonotic viruses enter the human population, the selection dynamics in which they evolve, and how viruses spread in the human population. Using a phylogenetic and phylogeographic approach, as well as incorporating

epidemiological information, I reconstruct an unobserved history of viruses using information available today. Insights provided by this dissertation will be valuable in making public health decisions and will assist in understanding and addressing the evolution of SARS-CoV-2.

# Chapter 1

# The role of socio-economic disparities in the relative success and persistence of SARS-CoV-2 variants in New York City in early 2021

Socio-economic disparities were associated with disproportionate viral incidence between neighborhoods of New York City (NYC) during the first wave of SARS-CoV-2. We investigated how these disparities affected the co-circulation of SARS-CoV-2 variants during the second wave in NYC. We tested for correlation between the prevalence, in late 2020/early 2021, of Alpha, Iota, Iota with E484K mutation (Iota-E484K), and B.1-like genomes and pre-existing immunity (seropositivity) in NYC neighborhoods. In the context of varying seroprevalence we described socio-economic profiles of neighborhoods and performed migration and lineage persistence analyses using a Bayesian phylogeographical framework. Seropositivity was greater in areas with high poverty and a larger proportion of Black and Hispanic or Latino residents. Seropositivity was positively correlated with the proportion of Iota-E484K and Iota genomes, and negatively correlated with the proportion of Alpha and B.1-like genomes. The proportion of persisting Alpha lineages declined over time in locations with high seroprevalence, whereas the proportion of persisting Iota-E484K lineages remained the same in high seroprevalence areas. During the second wave, the geographic variation of standing immunity, due to disproportionate disease burden during the first wave of SARS-CoV-2 in NYC, allowed for the immune evasive Iota-E484K variant, but not the more transmissible Alpha variant, to circulate in locations with high pre-existing immunity.

## 1.1 Introduction

New York City (NYC), a city of 8.5 million people, was one of the locations most heavily affected by the first wave of SARS-CoV-2 infections in North America and Europe[28;29]. In March through May 2020, >200,000 laboratory-confirmed coronavirus disease 2019 (COVID-19) cases were reported in NYC[29]. However, in late March 2020, the percentage of positive PCR tests reached 65%[29], indicating that testing was insufficient to capture the full extent of the outbreak. Most prominently, the epidemiology of SARS-CoV-2 during this first wave reflected socio-economic disparities within NYC, as well as the rest of the US[30], with communities of color and those living in areas with lower average socioeconomic status experiencing the highest burden of SARS-CoV-2 cases, hospitalizations, and deaths[31;32]. The higher proportion of SARS-CoV-2 antibody prevalence in NYC neighborhoods with predominantly Black and Hispanic or Latino residents further highlighted inequalities in access to testing[33]. Consequently, after the first wave, NYC neighborhoods were left with drastically different levels of standing immunity, with seropositivity ranging from 11 to 47% in November 2020 at the beginning of the second wave of SARS-CoV-2 infections.

The second wave in NYC, lasting from November 2020 through June 2021, started with an increase in B.1 cases that peaked in early January 2021, followed by a long plateau of high daily number of cases instead of the usual decrease after the peak[22]. The plateau was caused by the introduction and co-circulation of a number of variants of concern (VOC) and interest (VOI) that emerged at the time, including the emergence of the B.1.526 lineage, a "local" NYC strain later designated as Iota VOI, characterized by a number of spike mutations that were previously hypothesized to allow SARS-CoV-2 immune-evasive properties[22;31]. Specifically, E484K, D253G, and S477N mutations in the spike protein have been shown to reduce both vaccine-induced and natural immunity. A sister lineage to the Iota variant, currently designated as B.1.637, shared a number of defining mutations (namely L5F, D253G, and T95I), but also included S477N and Q957R. Although prevalent in NYC in early 2021, B.1.637 is not recognized as a VOC or VOI, by the World Health Organization.

Coinciding with the emergence of the Iota variant, the Alpha VOC, which emerged in

the United Kingdom in November 2020[34], was introduced to NYC and started increasing in prevalence. The Alpha variant was characterized by high transmissibility [40-50% higher than the wild-type Wuhan strain[21]] associated with the N501Y mutation that improves its ACE2 receptor binding ability[35]. Our previous analysis showed that in January-March 2021, both Alpha and Iota variants grew rapidly in NYC, but a clade of Iota with the E484K mutation (Iota-E484K) outpaced Alpha, B.1.637, and the rest of the Iota lineages in growth[22]. At the same time, in the neighboring state of Connecticut and the rest of New York State, regions which were not as heavily impacted by early outbreaks, the Alpha variant was more prevalent compared with Iota[33;36], as would be expected given the higher transmissibility of Alpha.

Here, we examine how the immune-evasive Iota variant was able to compete and co-circulate with the more transmissible Alpha variant in NYC in early 2021. By focusing on the time that precedes the emergence of other variants and the steep rise in vaccination levels, we examine how differences in seropositivity (primarily from naturally acquired immunity) between different NYC neighborhoods created a niche for successful competition and co-circulation of the two SARS-CoV-2 variants. We examine geographical patterns and socio-demographic characteristics of areas more severely affected during the first SARS-CoV-2 wave in NYC. This analysis sheds light on our understanding of the succession of SARS-CoV-2 variants and the ripple effect of socio-economic disparities on epidemic waves.

## 1.2 Materials and methods

### 1.2.1 Ethical considerations

The institutional review board of the University of California San Diego granted ethical approval of this work as human subjects exempt.

### 1.2.2 Genetic data

Public health surveillance by NYC Public Health Laboratory (PHL) and NYC Pandemic Response Laboratory (PRL) sequenced positive SARS-CoV-2 samples from across NYC (S1 Fig , see S1 Table for GISAID and GenBank accession numbers)[22;31]. We analyzed 18,332 SARS-CoV-2 genomes, which had information on ZIP code location and were sampled from 2

November 2020 through 18 June 2021. Pangolin lineages (v4.0.6) were assigned using UshER, and the Scorpio assignments were used to identify VOC/VOIs[37].

We aligned all Alpha, Iota, B.1.637, and other B.1-like (genomes that were not assigned as any of the other VOC/VOIs or B.1.637) genomes using MAFFT v.7.453[38]. We have previously shown that within the Iota variant there was a large clade defined by the E484K mutation[22]; to identify genomes that belong to that clade, we inferred maximum likelihood trees using IQTREE2[18], using the GTR+F+I substitution model, retaining polytomies, and enforcing a minimum branch length of $1x10^{-9}$ substitutions/site. We analyzed genomes in this Iota-E484K sub-variant cluster separately from the remaining Iota analyses.

### 1.2.3 Seroprevalence level and epidemiological data

Publicly available data collected by the NYC Department of Health and Mental Hygiene and available at https://github.com/nychealth/ were used to incorporate information on SARS-CoV-2 seropositivity, and number of cases, by modified ZIP Code Tabulation Area (MODZCTA), subsequently referred to as ZIP code. We used the seropositivity by ZIP code in NYC as of November 2020 based on the cumulative proportion of positive SARS-CoV-2 antibody tests, available at https://github.com/nychealth/coronavirus-data. This testing method is representative of both natural or vaccine-induced immunity, though vaccination was not broadly available at the start of the second wave in NYC.

Vaccination level by ZIP code was assessed as of May 2021 as the proportion of residents who had received at least one dose of COVID-19 vaccination administered by NYC vaccinating facilities and reported to the Citywide Immunization Registry. Neighborhood poverty level was defined based on the proportion of residents in a ZIP code with household incomes less than the federal poverty level, per the American Community Survey 2014–2018. We defined poverty levels as "Low" (<10% of residents in a ZIP code with household income lower than the federal poverty level), "Medium" (10-19.9%), "High" (20-29.9%), and "Very High" (≥30%). We considered five racial/ethnic groups: Asian or Pacific Islander, Black or African American, Hispanic or Latino, Other (including mixed race, American Indian or Alaska Native, unknown, and other), and White. ZIP code, poverty level, and race/ethnicity data were obtained from the NYC DOHMH

population estimates[39].

### 1.2.4 Statistical analyses

We performed a linear model regression between the seropositivity in a ZIP code and the proportion of genomes of each studied lineage (e.g., number of Alpha genomes/number of all SARS-CoV-2 genomes reported for that ZIP code) as reported by PHL and PRL. ZIP codes with less than 10 reported genomes were not included in the analysis. Then, we tested the association between seropositivity and poverty at ZIP code level using analysis of variance (ANOVA). We used the Pearson correlation coefficient to test the correlation of the proportion of the population of a certain race/ethnicity in a ZIP code and seropositivity in the ZIP code. We then calculated the proportion of adjacent ZIP codes (ZIP codes that share a land border) with the similar seropositivity relative to the total number of adjacent ZIP codes and performed a permutation analysis across the NYC ZIP codes, randomly permuting seropositivity 1000 times, to obtain the expected proportion of adjacent ZIP codes with similar seropositivity.

### 1.2.5 BEAST analyses

We conducted Bayesian phylogeographic analyses to estimate rates of the variants migration between locations with different standing immunity. We first used BEAST v1.10.4[40] to reconstruct time-scaled phylogenies for four lineages: Alpha, Iota, Iota-E484K, and B.1.637. We used the GTR+G4 nucleotide substitution model, assuming a strict molecular clock model (fixed to the value of $8x10^{-4}$ substitutions/site/year), and a Skygrid coalescent tree prior with 20 dimensions to reconstruct past population dynamics[41], following previous work on NYC SARS-CoV-2 data[28]. We provided a starting maximum likelihood tree (obtained as described above) to help with convergence. We ran Markov Chain Monte Carlo (MCMC) for 108 generations and investigated the chain convergence using Tracer v1.7.1[42] to confirm that the effective sample size (ESS) of all parameters was >200.

For phylogeographic analysis, we used LogCombiner[40] to resample the obtained tree distributions at a lower frequency. The resulting tree distributions for each lineage included 2000 trees; these tree distributions were used for all consequent phylogeographic analyses. We assigned geographic trait value to sequences in the phylogeographic analysis based on ZIP code

level seroprevalence (grouped into 'High', 'Medium', or 'Low') and borough. We further refer to parts of boroughs with different seroprevalence levels as locations, i.e all ZIP codes with the same seroprevalence level (i.e. 'High') were grouped together within a borough (i.e. 'High') into a location (i.e. 'Bronx_High') in the phylogeographic analysis. We applied Markov jumps model to calculate the number of transitions from one location to another[42] along with the estimated migration rates between locations, assuming asymmetric virus migration flows between locations. We ran MCMC for 108 generations to estimate pairwise migration rates between different locations in this analysis. Xml files for the phylogeographic analyses of each variant are provided as Supplementary materials.

To assess whether certain variants were preferentially migrating from one location to another based on the seropositivity, we calculated the pairwise difference in seropositivity between locations as they were defined in phylogeographical analysis and estimated whether it was correlated with the log-transformed viral lineage migration rate between the locations. We then re-ran this analysis of association separately for each variant directly in the Bayesian framework using previously described GLM model[43]. We also ran a GLM of association in a Bayesian framework using vaccination level in each location as of May 2021 as a covariate for the migration rates.

### 1.2.6 Persistence analysis

We used PersistenceSummarizer, as implemented in BEAST[19], to examine the effect of seropositivity on the persistence of specific variants in a location. PersistenceSummarizer identifies lineages that have been in a location from the beginning to the end of an observation period. For this analysis, we partitioned the observation period into biweekly intervals going backward in time from the most recent genome in each of the datasets. We report the mean abundance of persistent lineages in the posterior tree distribution of each variant by seroprevalence.

## 1.3 Results

### 1.3.1 Geographical patterns in seropositivity

To test whether there was geographical structure in the levels of preexisting immunity to SARS-CoV-2 across NYC as of November 2020, we analyzed whether areas with similar seroprevalence levels were more likely to be geographically adjacent to each other. First, we assigned every ZIP code across NYC into one of three quantiles ("High", "Medium", and "Low") based on their reported seroprevalence levels (Fig 1.1). We found the proportion of adjacent ZIP codes (i.e., ZIP codes sharing a land border) with the same seroprevalence level was higher than expected and had a strong geographical pattern. Specifically, 20.6% of ZIP codes adjacent to ZIP codes with "High" seroprevalence level also had "High" seroprevalence, the expected proportion was 7.2% (95% range 4.2–10.5%). For "Medium" seroprevalence level ZIP codes, the observed proportion of adjacent ZIP codes in the same category was 45.9%, while the proportion expected by chance was 29.1% (95% range 24.1–34.1%). For "Low" seroprevalence level ZIP codes, the observed proportion of adjacent ZIP codes with "Low" seroprevalence level was 50.7%, while the expected was 26% (95% range 20.8–30.7%). For each seroprevalence level the average proportion of adjacent ZIP codes that match the seroprevalence level was higher than expected, indicating that seroprevalence level in NYC has geographic structure.

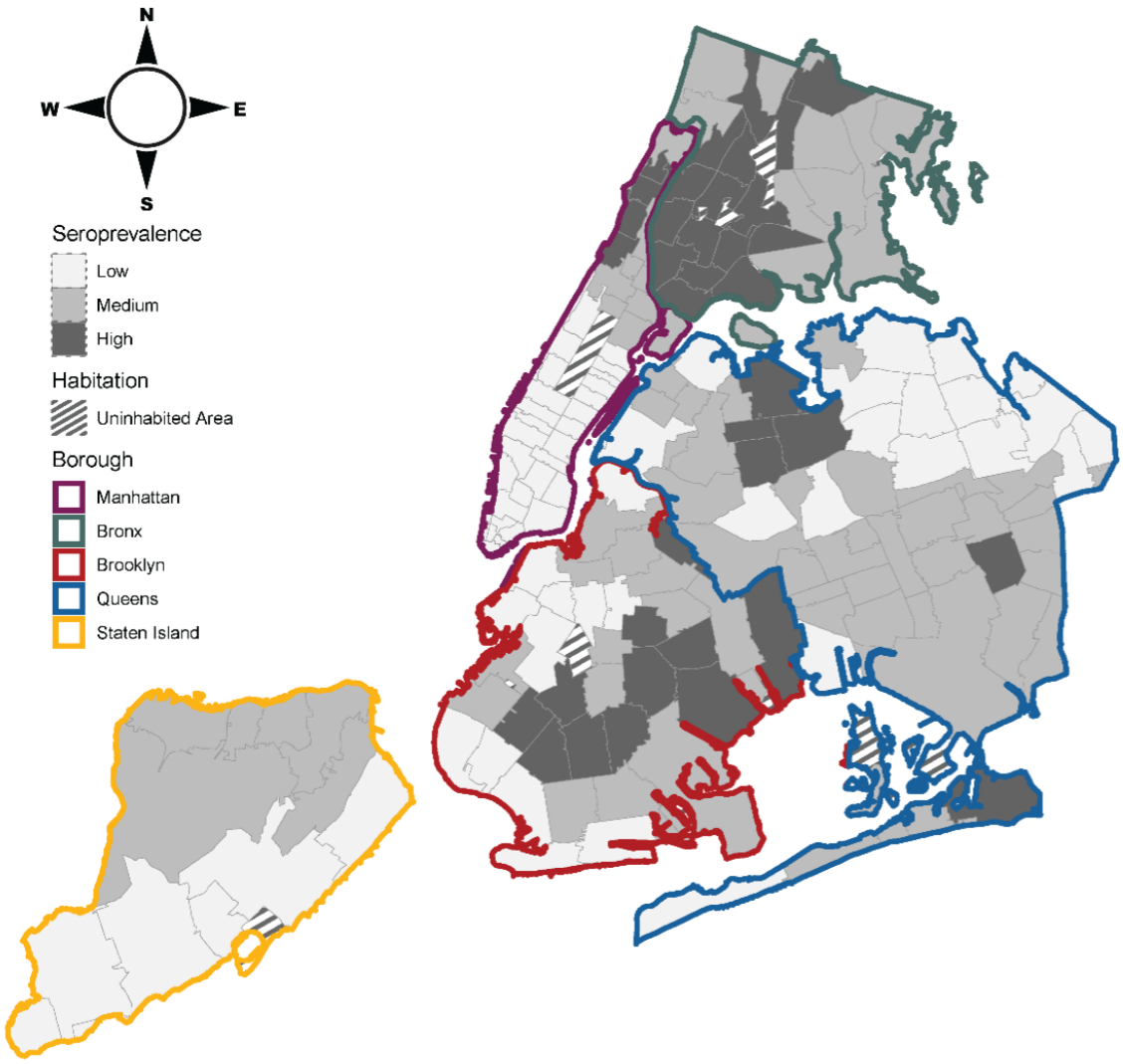



**Figure 1.1. Seroprevalence of SARS-CoV-2 antibodies across NYC in November 2020.** Color borders indicate boroughs. Gray borders denote ZIP codes and shading indicates seroprevalence. The base layer of the map obtained from: Modified-Zip-Code-Tabulation-Areas-MODZCTA-Map , distributed openly under the following terms and conditions: Terms of Use.

### 1.3.2 Correlation of variant prevalence and seropositivity

We estimated whether prevalence of SARS-CoV-2 lineages during the second wave was correlated with estimated seropositivity at the start of the second wave, by ZIP code. We focused on Alpha, Iota, Iota-E484K, and B.1.637 lineages and compared them to the co-circulating B.1-like viruses. In locations with higher pre-existing seropositivity, we observed a higher proportion of

Iota and Iota-E484K genomes, but not other variants or B.1-like lineages (Fig 1.2A). Specifically, we found a positive correlation between the higher pre-existing seropositivity and the proportion of Iota-E484K genomes (R=0.32, p<0.001) and Iota genomes (R=0.19, $p$= 0.014). In contrast, we found a negative correlation with pre-existing seropositivity and the proportion of Alpha (R=-0.16, $p$= 0.040) and B.1-like genomes (R=-0.29, p<0.001); no correlation was found for the B.1.637 lineages (R=0.07, $p$= 0.376) (Fig 1.2B). Although the correlations of seropositivity and some variants are significant, seropositivity does not explain most of the variance in lineage proportion. We note that ZIP codes with low numbers of sampled genomes will have high variance in estimated lineage proportion and high stochasticity in population dynamics which are other major factors that impact estimated lineage proportion in a ZIP code, along with relative susceptibility of a population to different variants due to standing immunity levels.

During the beginning of the second wave, genomes were primarily B.1-like, across ZIP codes with varying levels of seroprevalence (Fig 1.2C). During the peak of the second wave in April 2021, the majority of cases in "Low" seroprevalence level ZIP codes were from the Alpha variant, but Alpha was not the majority in "High" seroprevalence level ZIP codes. In contrast, Iota and Iota-E484K accounted for a greater proportion of cases in "High" seroprevalence level ZIP codes compared to the "Low" seroprevalence level ZIP codes.

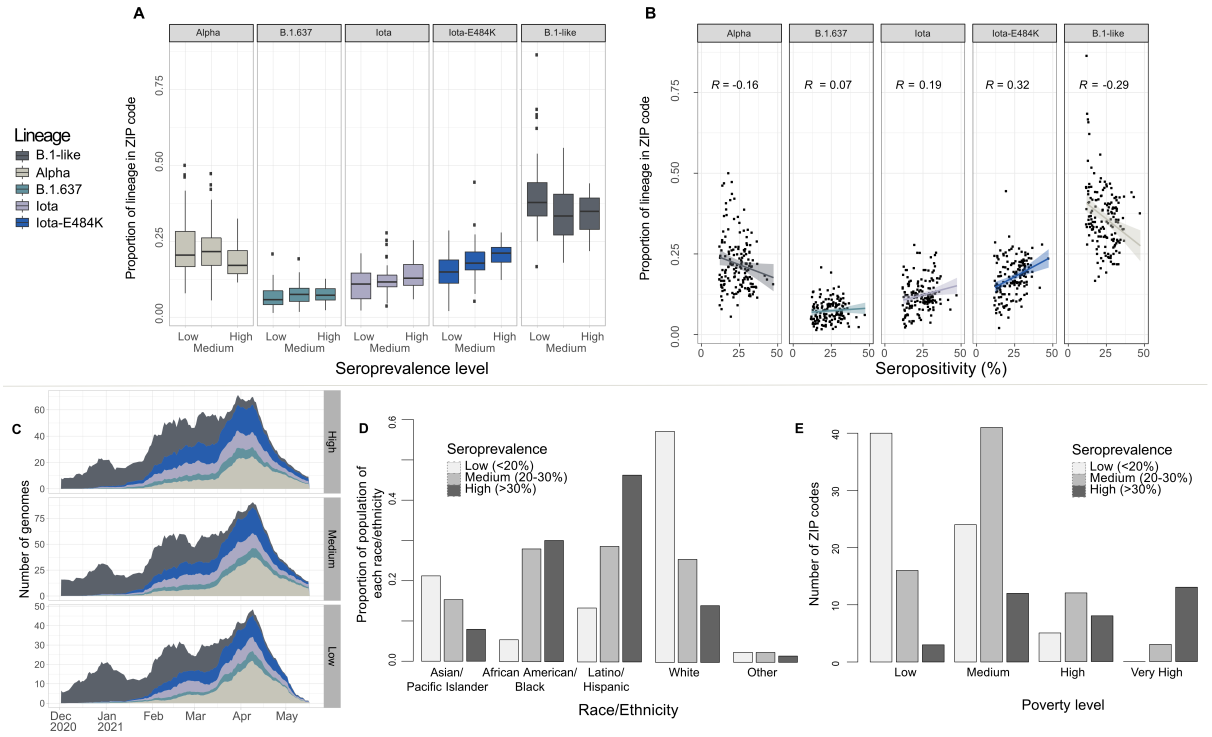



**Figure 1.2. SARS-CoV-2 variants and demographics across ZIP codes in NYC.** Correlation analysis between (A) seroprevalence level and proportion of each lineage, (B) seropositivity and proportion of each lineage; (C) 14 day rolling mean number of genomes of each lineage by ZIP code seroprevalence level; (D) proportion of population of a certain race/ethnicity by seroprevalence level by ZIP code, and (E) number of ZIP codes of different seroprevalence level by poverty level.

### 1.3.3 Correlation between seropositivity and socio-demographic data

We then investigated the relationship between ZIP code seropositivity and sociodemographic composition (Fig 1.2D). We found a negative correlation between the seropositivity and the proportion of residents who identified as White, Asian or Pacific Islander, or Other races (Pearson correlation coefficients -0.73 ($p$<0.001), -0.31 ($p$<0.001), and -0.23 ($p$= 0.002), respectively). Conversely, seropositivity was positively correlated with the proportion of Black or African American and Hispanic or Latino residents (Pearson correlation coefficients 0.49 ($p$<0.001), and 0.63 ($p$<0.001), respectively). Similarly, we found that areas with "High" and "Very High" poverty levels were more likely to have high seropositivity (ANOVA $p$<0.0001; Fig 1.2E).

### 1.3.4 Phylogeographic analysis findings

We performed phylogeographic analysis to infer viral lineage migration rates between regions of seroprevalence levels. To make our approach computationally tractable, we grouped ZIP codes with the same borough and seroprevalence level together ("High", "Medium", or "Low"; Fig 1) (Fig 1.3). Since there was only one "Low" seroprevalence ZIP code in the Bronx, it was included in the Bronx "Medium" seroprevalence group. Staten Island did not have any ZIP codes with "High" seroprevalence level, resulting in a total of 13 total locations for phylogeographic inference (see Fig A.2 for molecular clock trees with locations annotated along the tree branches). We utilized geography as a proxy for prior immunity, rather than as a measure of physical space. We found no significant relationship between the log-transformed between-location viral migration rates and the pairwise seropositivity difference for any of the studied lineages: correlation coefficients 0.04 ($p$= 0.65), 0.02 ($p$= 0.78), 0.17 ($p$= 0.84), and -0.06 ($p$= 0.48) for Alpha, B.1.637, Iota, and Iota-E484K, respectively. There is no significant pattern of migration between locations that is explained by the difference in seroprevalence of those locations.

Inclusion of seropositivity as a generalized linear model covariate directly in the phylogeographic model also showed no statistically significant association with the migration rates (Table A.1). The same result was observed when all locations with the same seropositivity level were merged irrespective of the borough. Additionally, to test whether different levels of vaccination throughout NYC could have affected SARS-CoV-2 migration rates between the different locations as estimated in the phylogeographic analysis described above, we included vaccination level (as of May 2021) as a covariate in our analyses, but similarly, no association was found (Table A.1).

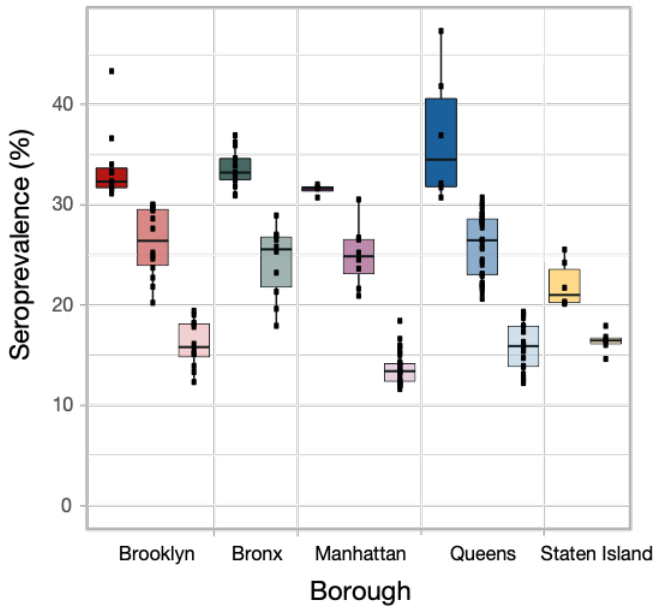



**Figure 1.3. Seroprevalence quantiles across NYC boroughs** Seroprevalence level for ZIP codes (black square); categorized by location groups (high, medium, and low seroprevalence within each borough). Location group summarized by boxplot where shading indicates seroprevalence level, and color is associated with borough.

### 1.3.5 Analysis of persisting lineages

We further ran lineage persistence analyses that allow inference, whether a variant was likely to persist in a location (i.e., if a variant observed in a location is still present in that same location at the end of the two-week observation period) or if it was more likely to migrate to another location or go extinct (i.e., not observed after two weeks). We ran the persistence analysis separately for each of the variants to describe patterns of their circulation within locations of "High", "Medium", and "Low" seroprevalence level. We found that for the Alpha variant, the proportion of lineages persisting in "Low" seroprevalence level locations grew over time, from 19% in mid-January to 81% in early June. Concurrently, the proportion of Alpha lineages was gradually declining in "High" seroprevalence level locations, from 16% to 2% in the same observation period (Fig 1.4A). This finding indicates that Alpha was able to consistently persist for longer than 2 weeks in locations with "Low" seroprevalence level, but repeatedly and rapidly went extinct in "High" seroprevalence locations. Given that seropositivity was not a significant predictor for migration rates between locations with different seroprevalence levels (Table A.1),

and the number of introductions into High and Low seroprevalence regions was similar (Fig 1.4), it is likely that the Alpha lineages repeatedly died out in areas with "High" seroprevalence.

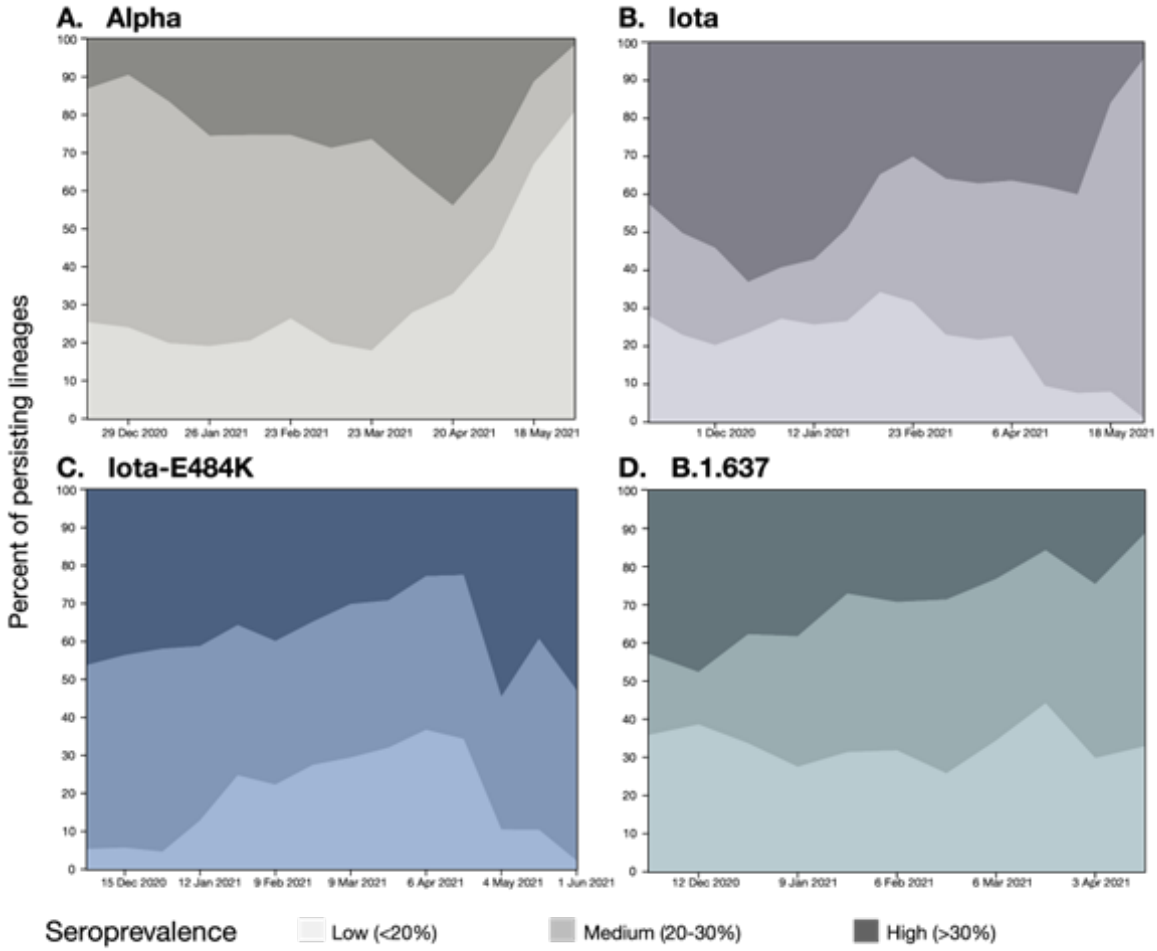



**Figure 1.4. Persistence of variant lineages over time in NYC stratified by seroprevalence level.** Percent of persisting lineages indicates the proportion of lineages present in a location at the observation time that were already present in the same location 2 weeks prior. Shading indicates seroprevalence level.

In contrast, the proportion of persisting Iota-E484K lineages remained relatively stable in all locations in the corresponding time period. Between mid-January and mid-May, the proportion of Iota-E484K lineages that persisted in "High" seroprevalence level locations for at least 2 weeks changed from 41% to 39%, and in "Low" seroprevalence level locations from 13% to 10% (Fig 1.4B). Importantly, Iota-E484K was not able to establish itself at high prevalence in "Low" seroprevalence level locations, likely in part because they were being outcompeted by

infections with the more transmissible Alpha variant.

Other Iota lineages circulated mostly in locations with "Medium" seroprevalence level: the proportion of Iota lineages persisting in both "High" and "Low" seroprevalence level locations was gradually decreasing, while the proportions of Iota lineages persisting in the "Medium" seroprevalence level locations grew from 26% to 96% between mid-January and early June (Fig 1.4C). Between mid-January and mid-April, the proportion of B.1.637 persisting lineages remained stable in the "Low" seroprevalence level locations (28% and 33%), grew in the "Medium" seroprevalence level locations from 34% to 55%, and declined in the "High" seroprevalence level locations from 38% to 11% (Fig 1.4D).

## 1.4 Discussion

Infectious disease thrives on disparity and inequality[30]. Here we demonstrate how socio-economic disparities in New York City created unequal predispositions for successful spread of emerging lineages during the second wave of SARS-CoV-2. We show that those areas that had high seropositivity after the first wave of SARS-CoV-2 and were predominantly high poverty areas with majority Black and Hispanic or Latino residents, were less likely to sustain the highly transmissible Alpha variant. In contrast, these locales were more likely to sustain the Iota-E484K variant; areas with high levels of preexisting seropositivity provided a successful niche for the emergence and spread of an immune evasive variant (in comparison to a highly transmissible but less immune evasive variant) likely due to re-infection of individuals who had acquired SARS-CoV-2 during the first wave.

Our findings confirm that areas with high seropositivity were more likely to have SARS-CoV-2 infections of Iota-E484K than Alpha, as hypothesized earlier[36]; and that once Iota-E484K reached those areas of high seropositivity, it was able to persist there, while Alpha was not. While persistence could be incorrectly inferred in the case that a lineage leaves NYC and returns to the same seroprevalence and borough, this likely would form a very small contribution to the total persistence as we expect the majority of cases to be from local transmission, particularly because Iota variant emerged in NYC[28].

Interestingly, we did not find a correlation between virus migration rates and difference

in seropositivity between locations; this lack of association is not unexpected given the ease of transportation throughout the city. This finding suggests that when the Alpha variant reached "High" seroprevalence areas, it likely went extinct there, unable to re-infect individuals, but did not successfully migrate to other areas. In contrast, in "Low" seroprevalence locations, Alpha prevalence was able to grow and thus maintain its dominant position in the city.

In May 2021, after the second wave, the proportion of partially vaccinated people in NYC ranged from 22-82% and only 5-17% were fully vaccinated. Vaccination level at that point in time was not a main contributing factor to variant spread during the second wave that took place in November 2020 through June 2021, as evidenced by our phylogeographic analysis. Less than 1% of people were vaccinated at the start of the second wave.

In NYC, the introduction of the Alpha variant around the time of Iota emergence[44] likely prevented Iota from establishing at a level that would make it difficult for Alpha to compete. Unlike previous analyses that focused on competition at a country level[45;46], we were able to decipher competition dynamics between variants at the neighborhood level owing to the granularity of seropositivity data and SARS-CoV-2 genome data. Given the higher transmissibility of the Alpha variant, it might have completely eliminated Iota from NYC if not for the pockets of higher seropositivity left in the wake of the first wave of SARS-CoV-2. For example, when the Gamma variant, another lineage with immune-evasive properties, reached NYC in January 2021[47], it was not able to successfully establish in the city and only circulated at low frequencies, likely because its niche in people with pre-existing immunity was taken by Iota-E484K that had been circulating in NYC for several months. In contrast, Iota lineages that did not have the E484K mutation dominated the "Medium" seroprevalence locations, as this might have been the pockets available to them after "Low" seroprevalence locations were occupied by Alpha and "High" by Iota-E484K. Our analysis of importation events also showed that "Medium" seroprevalence areas had more Iota introductions, which could have been by chance but could have also contributed to the higher persistence of Iota in those locations.

The Gamma variant dominated the Brazilian SARS-CoV-2 epidemic in late 2020/early 2021 due to the high levels of pre-existing immunity in Brazilian population[48;49], but in mid-2021 it was quickly replaced by Delta due to the Delta variant's higher transmissibility[50]. At the

same time, the Gamma variant was not able to successfully establish itself in the neighboring Colombia or Peru. This failure is likely due to a founder effect as competing variants, Mu in Colombia and Lambda in Peru, which also have immune-evasive properties, already dominated the Colombian and Peruvian epidemic by the time Gamma reached the country[45;46]. Thus, the time of VOCs emergence and introductions, and mechanisms for increased fitness, is crucial for their success in a location in the presence of other variants.

Our study was able to elucidate how various factors, including geographic variability in seropositivity, a symptom of inequities associated with socio-demographic factors in 2020, affected the competition dynamic between two co-existing viral lineages. Such analyses would not be possible in the post-Omicron era when multiple consecutive waves are dominated by variants with similar properties and where the majority of the population have a mixture of acquired natural and/or vaccine-induced immunity[51]. By focusing on the time preceding the emergence of the Delta and Omicron variants, we show that the consequences of socio-economic disparities in an outbreak can have a ripple effect that can last into subsequent outbreaks, providing an important lesson for future epidemic preparedness and mitigation efforts.

## 1.5 Acknowledgments

We thank NYC DOHMH: Public Health Laboratory: Virology Section, WGS Unit; Bureau of Communicable Diseases for COVID-19 seroprevalence data; Division Management Systems Coordination for data integration; Bureau of Immunization for COVID-19 vaccine data; NYC DOHMH Bureau of Epidemiology Services for NYC MODZCTA data and shapes; Pandemic Response Laboratory for sequencing virus genomes. We thank Volodymyr Minin of the University of California Irvine for his help with Markov Jumps analysis interpretations. J.O.W receives funding through CDC contracts through his institution unrelated to this research and provided compensated expert testimony on SARS-CoV-2 and the COVID-19 pandemic.

### 1.5.1 Financial Disclosure statement

T.I.V. is funded by a Branco Weiss Fellowship. G.W.H. acknowledges funding from the National Institutes of Health (AI154824). This work was supported (in part) by the

Epidemiology and Laboratory Capacity (ELC) for Infectious Diseases Cooperative Agreement (Grant Number: ELC DETECT (6NU50CK000517-01-07) funded by the Centers for Disease Control and Prevention (CDC). The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. Use of trade names is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### 1.5.2 Author contributions

T.I.V., J.L.H., S.H., J.O.W. - conceptualization. J.C.W., E.L., G.W.H., H.A., S.D.L. - data curation. T.I.V., J.L.H., J.C.W. - formal analysis and investigation. T.I.V., J.L.H., J.O.W. - writing - original draft preparation, T.I.V., J.L.H., J.C.W., E.L., G.W.H., H.A., S.D.L., F.T., E.O., S.H., J.O.W. - writing - review & editing.

Chapter 1, in full, is a reformatted reprint, permission pending, of the material as it appears in "The role of socio-economic disparities in the relative success and persistence of SARS-CoV-2 variants in New York City in early 2021." Tetyana I. Vasylyeva, Jennifer L. Havens, Jade C. Wang, Elizabeth Luoma, Gabriel W. Hassler, Helly Amin, Steve Di Lonardo, Faten Taki, Enoma Omoregie, Scott Hughes, and Joel O. Wertheim, *Plos Pathogens*, 2024. The dissertation author was the co-primary investigator and co-first author of this paper.

# Chapter 2

# Synonymous substitution rates preceding the emergence of SARS-CoV-2 variants and during persistent infections

The emergence of variants has shaped the SARS-CoV-2 pandemic. Hypotheses for the origination of variants include emergence from persistent infections, transmission in non-human animal populations after reverse-zoonosis, or cryptic transmission in the human population. We investigated the molecular clock and rate of nonsynonymous and synonymous substitutions in these contexts and along the stem lineages of SARS-CoV-2 variants. Persistent infections were identified by viral genomes from individuals with repeat samples, taken more than 30 days apart, that cluster together on a phylogenetic tree. We find that some persistent infections (4 out of 7 in our analysis) and some variants (eg. Beta, Delta, Iota, Mu, BA.1) have a slower synonymous substitution rate, while other variants (eg. Alpha, Epsilon) have an increase in synonymous substitution rate relative to typical transmission of acute infections. A slower rate of synonymous substitutions along variant stem lineages suggests that viral replication rate is lower, which could occur in isolated body compartments during persistent infection, suggesting that to be the most plausible scenario for the emergence of variants.

## 2.1 Introduction

SARS-CoV-2 emerged in 2019 and quickly spread across the world. The early pandemic was primarily characterized by transmission of closely related viruses with identical or nearly identical genomes of pangolin lineage B.1, or a close descendant of B.1, including B.1.1 (B.1-

like)[20]. Later jumps in evolution interspersed these closely related viruses; in a phylogenetic tree these jumps are represented by a single long branch or a few internal branches with limited intermediate sequences (stem lineage).This evolutionary dynamics precedes the emergence of variants of interest or concern that reinforced the spread of SARS-CoV-2. Monitoring the spread of variants is of paramount epidemiological importance since they are often more transmissible or immune evasive than the strain they descend from and once they emerge they can spread rapidly[20].

One example of dominance of SARS-CoV-2 variants is Alpha, the first named variant, which was characterized by high transmissibility relative to B.1-like virus. The stem lineage preceding Alpha had more nonsynonymous substitutions than expected given the length of time it evolved and was shown to be under positive selection[21;52]. Starting in late 2020 many different variants, with greater transmissibility or immune evasiveness relative to B.1-like virus, arose including Beta, Gamma, Epsilon, Iota, B.1.637 and Mu[22;49;53–55]. These named and unnamed variants were all characterized by many nonsynonymous substitutions on the stem lineage. Although first detected in late 2020, alongside many other variants, Delta spread globally as the predominant lineage late 2021. After the Delta wave, Omicron, in particular the BA.1, BA.2, BA.4, and BA.5 lineages, swept around the world causing a new wave of cases. The stem lineages of Delta and Omicron variants were under positive selection and had more nonsynonymous substitutions than expected; both variants were highly successful and more immune evasive than B.1-like virus[55–59].

While the positive selection and nonsynonymous substitutions on the stem lineage of variants has been repeatedly noted, the behavior with regards to synonymous and non-coding substitutions has not been as well explored[60]. Synonymous substitutions can be impacted by selection, even if it is to a lesser extent than expected for nonsynonymous substitutions, as well as the rate of viral replication and fidelity in sequence replication. Synonymous substitution rate can be affected by host species[61], cell tropism[62] and virus sequestration[63]; a change in synonymous substitution rate is typically associated with a change in viral replication rate[10;11;64]. Ebola virus sequestered in an immunoprivileged compartment with slow rates of replication was the likely source of onward transmission virus years after the initial infection[63;65]. In HIV it has

been found that the synonymous substitution rate varies between infections and impacts the dynamic of the infection[66].

Understanding the dynamics of the stem preceding variants is of great public health concern because it can give insight into how variants arise and hopefully how to prevent their spread in the future. Proposed sources for where evolutionary jumps that give rise to the variants occur include 1) persistent infections 2) transmission in non-human animal populations after reverse-zoonosis and 3) cryptic transmission in the human population[67].

SARS-CoV-2 is typically a short acute infection. However, persistent infections lasting longer than 30 days have been detected, sometimes lasting longer than a year[23]. Some persistent infections are characterized by accelerated evolution resulting in many nonsynonymous mutations that allowed them to develop immune evasive properties[23]. The stem lineage of some variants have the same pattern of accelerated evolution and many nonsynonymous substitution, and so persistent infections have been proposed as a possible source of variants. Evolution in nonhuman animals could also be the source of variants; undersampling of animal reservoirs can result in evolution within non-human hosts to not be observed[24;68]. However, we note that SARS-CoV-2 in nonhuman populations typically acquire host characteristic substitutions, which are not associated with SARS-CoV-2 variant stem lineages[69;70]. It is also possible that variants evolve during typical transmission in the human population, and the long stem lineage is due to undersampling[55;71]. However, this does not explain the accelerated rate of non-synonymous substitutions, or increased positive selection associated with variants.

In this study we contrast the evolutionary dynamics in variant stems, persistent infections, and non-human animal hosts with evolution during transmission between humans. We consider the stem lineages of Alpha, Beta, Gamma, Delta, Epsilon, Iota, B.1.637, and Omicron (we consider the stem of BA.1 and separately the stem of BA.2, BA.4, and BA.5). Using a Bayesian framework with local clocks and robust counting, we infer the molecular clock rate, and rate of nonsynonymous and synonymous substitutions of SARS-CoV-2 in different contexts.

## 2.2 Results

### 2.2.1 Identification of persistent infections

To identify persistent infections we inferred a phylogenetic tree with SARS-CoV-2 genomes from individuals with two or more genomes sampled more than 30 days apart, from New York City SARS-CoV-2 genomic surveillance (n=373). We defined persistent infections, as opposed to repeat infections, as individuals with multiple genomes that clustered on the tree. We found 15 individuals, represented by 46 genomes, that were persistent infections. Of these persistent infections 3 are B.1-like, 2 are Alpha, 2 are Delta, 1 is Iota, 2 are B.1.637, and 5 are BA.1 (Omicron).

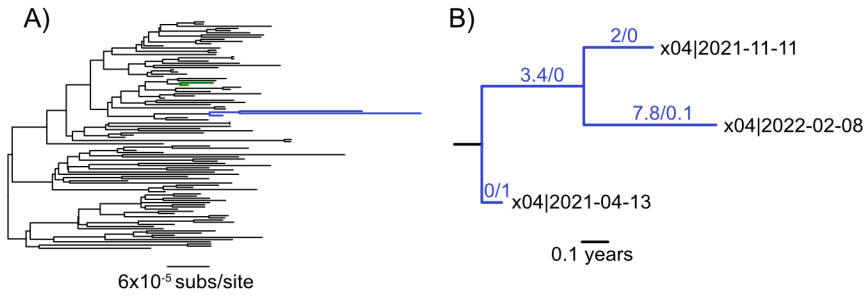



**Figure 2.1. Phylogenetic trees of persistent infections.** (A) Phylogenetic tree of Alpha sequences, persistent infection clades are highlighted in blue (Alpha_x04) and green. (B) Phylogenetic tree of persistent infection Alpha_x04 with time scaled branches annotated with robust counts of nonsynonymous / synonymous substitutions.

Using all sequences of the same variant or a subset of 150 B.1-like genomes for context, we inferred local molecular clocks for each persistent infection clade (Fig 2.1). Local molecular clocks were inferred as the product of the background rate and an increment factor with a log normal prior centered on 1. Of the 15 persistent infections 7 had more than 3 months of evolution over the sum of all branches within the persistent infection clade (person months), and were used for subsequent analysis.

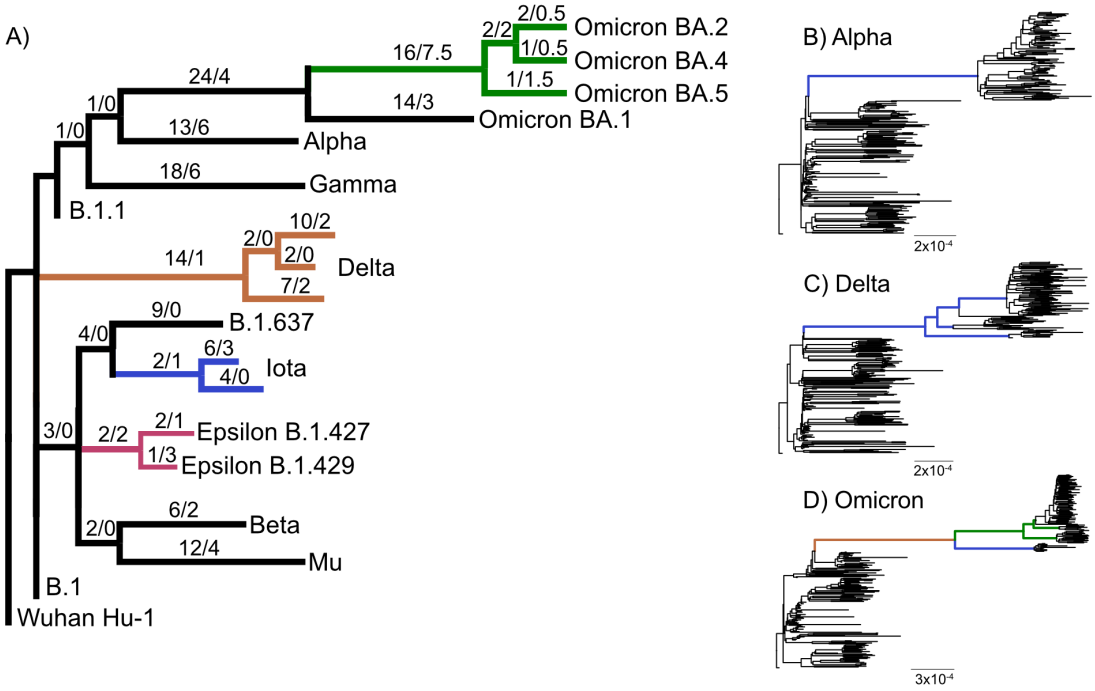



**Figure 2.2. Phylogenetic trees of stem lineages.** (A) Cartoon tree showing SARS-CoV-2 backbone and the relationship between variants, annotated with substitutions formatted as the number of nonsynonymous/synonymous substitutions inferred by robust counting in BEAST. (B) Phylogenetic tree of B.1-like and Alpha sequences, the variant stem branch is highlighted in blue. (C) Phylogenetic tree of B.1-like and Delta sequences, the variant stem lineage is highlighted in blue. (D) Phylogenetic tree of B.1-like and Omicron sequences, BA.1 stem is highlighted in blue, the stem lineage leading to BA.2, BA.4, and BA.5 (BA.2/BA.4/BA.5) is highlighted in green, the base of Omicron, which included in analysis the lineage for BA.1 and BA.2/BA.4/BA.5, is highlighted in brown.

### 2.2.2 Variant stem lineages and deer clades

We inferred a maximum likelihood phylogenetic tree with 10,572 high quality SARS-CoV-2 genomes to define variant stem lineages (Fig 2.2A) as the path from B.1 (Beta, Delta, Epsilon, Mu, Iota, and B.1.637) or B.1.1 (Alpha, Gamma, Omicron) to the base of the variant expansion (Fig 2.2B-D). For subsequent analysis, B.1-like virus and each variant were represented by all available genomes or subsampled down to 150 genomes. Using BEAST we inferred local molecular clocks on variant stem lineages as characterized by sequences of the variant of interest and B.1-like virus.

To assess SARS-CoV-2 evolution in non-human hosts after reverse zoonosis we used 22

sequences from 5 clades of deer associated virus[70]. All deer samples were associated with Alpha and Delta variants, Alpha and Delta sequences (respectively) from NYC human surveillance were used for background context. Local clocks for each deer clade were inferred.

### 2.2.3 Rate of synonymous substitutions

Using BEAST[40] with robust counting[72] we inferred the synonymous and nonsynonymous substitutions. The rate of substitution was inferred as the sum of substitutions divided by the sum of the length of branches of interest (persistent infection clade, variant stem lineages, deer clade, or the background of all other branches). The background rate of synonymous substitutions, for internal branches, in coding regions of B.1-like virus was 6.8 subs/year (HPD 6.1-8.2 subs/year; Supp table B.1).

We then considered the substitutions rate on variant stem lineages. There is a consistent pattern across Beta, Delta, Iota, B.1.637, Mu, and BA.1 where the rate of synonymous substitutions is slower than the estimated background rate (Fig 2.3). The stem lineage of Delta is the combination of branches leading to multiple major genetic expansions, given the length of the Delta stem and background synonymous rate we expect around 11 (HPD 9.45, 12.76) synonymous substitutions, but actually observe 5 substitutions, and synonymous rate of 3.2 subs/year (HPD 2.8-3.8 subs/year), much less than expected (Supp table B.2). Similarly for the Iota variant, we would expect approximately 7 (HPD 6.02, 9.24) synonymous substitutions given the length of the stem lineage, but we only observe 4 substitutions (Supp table B.2), and a synonymous substitution rate of 3.78 subs/year (HPD 2.9-5.1 subs/year), much lower than expected

When we consider persistent infections, 4 of the 7 persistent infections had a slower rate of synonymous substitutions compared to the background (Supp table B.1). For example in the case of Alpha_x04 the synonymous rate is 0.83 subs/year (HPD 0.59-1.8 subs/year) compared to 6.8 synonymous subs/year in the context of typical transmission.

In the Gamma and the BA.2/BA.4/BA.5 stem lineage the rate of synonymous substitutions is not significantly different from the background (Fig 2.3). Similarly, in 3 of the 7 persistent infections, the rate of synonymous substitutions is not significantly different from the

background.

When we consider evolution in clades of deer genomes, we found that there was an increased rate of synonymous substitutions. We also found an increased rate of synonymous substitutions in the Alpha and Epsilon stem lineages, which are outliers compared to the other variants in that respect (Fig 2.3).

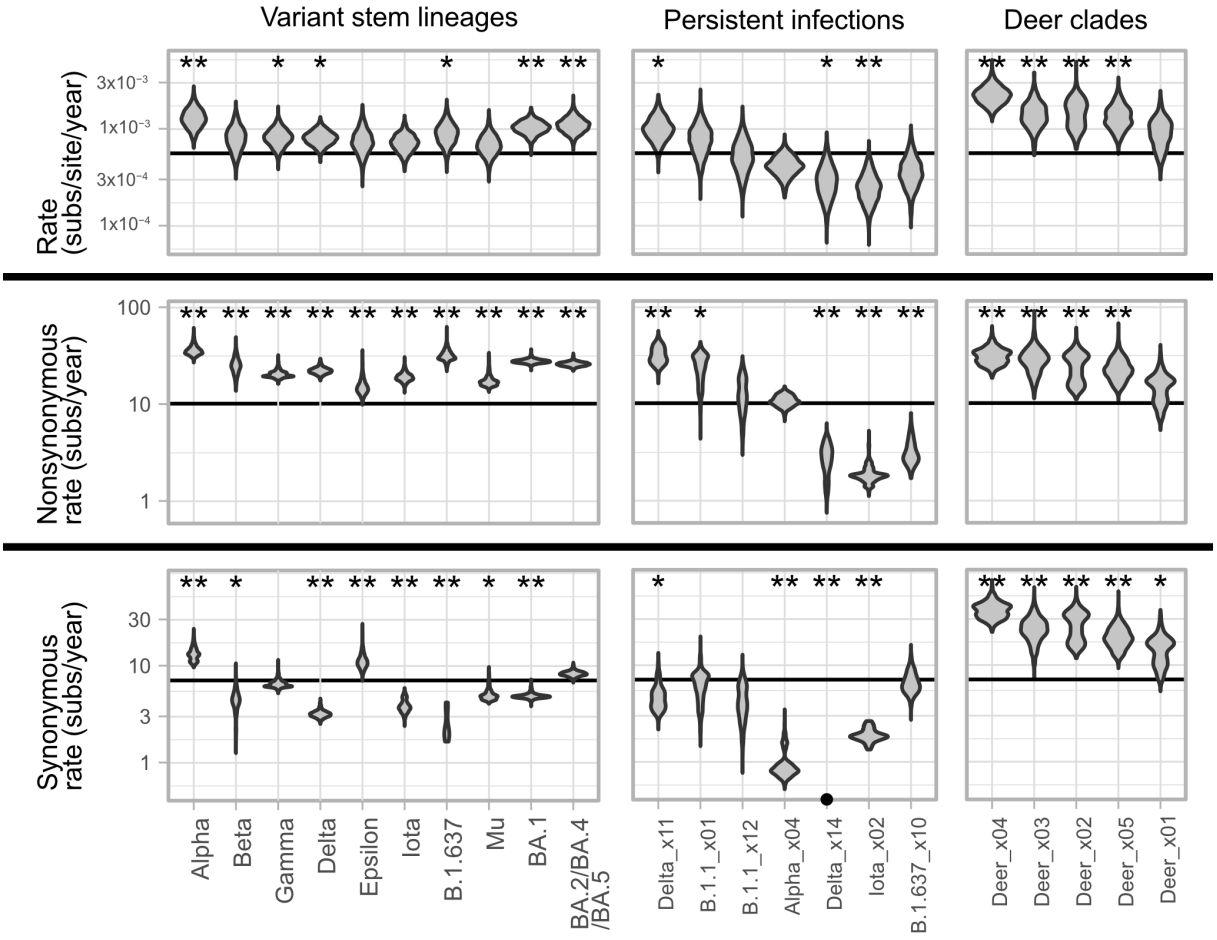



**Figure 2.3. Rate of nonsynonymous and synonymous substitutions and molecular clock.** First row: Summary of molecular clock rate for each variant stem lineage, persistent infection, and deer clade with posterior probability density of difference from background (P > 0.9 *, P > 0.99 **). Background molecular clock median ($5.6x10^{-4}$ substitutions/site/year) indicated by black horizontal line. Second row: Summary of nonsynonymous substitution rate with posterior probability of difference from background (P > 0.9 *, P > 0.99 **). Background nonsynonymous rate median (10.1 substitutions/year) indicated by black horizontal line. Third row) Summary of synonymous substitution rate with posterior probability of difference from background ( P > 0.9 *, P > 0.99 **). Background synonymous rate median (6.8 substitutions/year) indicated by black horizontal line.

### 2.2.4 Rate of nonsynonymous substitutions

There is a consistent pattern across all the variants that we considered, that all the variant stem lineages had an elevated rate of nonsynonymous substitutions relative to the background (Fig 2.3; Supp Table B.1). This elevated rate preceding variants has been previously described. In deer clades we found an increased rate of nonsynonymous substitutions, which is also consistent with previous findings[70]. In persistent infections, 2 of 7 had an increased rate of nonsynonymous substitutions consistent with variant stem lineages. However, in contrast to the other contexts, some persistent infections (n=3) had slower rates, while others had a similar rate of nonsynonymous substitutions compared to the background (n=2)

### 2.2.5 Molecular clock and dN/dS

The overall molecular clock and dN/dS are both impacted by the rates of nonsynonymous and synonymous substitutions. Looking at these measures combined is still a valuable tool, for example dN/dS can provide important information about the strength of selection.

The Alpha stem lineage (Fig 2.2B) had an increased rate of evolution in coding regions ($1.3\text{x}10^{-3}$ HPD: $7.8\text{x}10^{-4}$ - $1.9\text{x}10^{-3}$ subs/site/year) compared to the background coding region clock rate ($5.5\text{x}10^{-4}$ HPD: $5.1\text{x}10^{-4}$ - $6.2\text{x}10^{-4}$ subs/site/year; Supp Table B.1). While both the nonsynonymous and synonymous substitution rate increase, the increase in nonsynonymous was more impactful and dN/dS is increased in the Alpha stem relative to the background (Table 2.1). We see a similar pattern in Gamma and the BA.2/BA.4/BA.5 stem lineage where dN/dS is elevated as is the overall molecular clock rate, due to the increase in nonsynonymous substitution rate.

In the Epsilon stem lineage there is a slight increase in synonymous and nonsynonymous substitution rates, which are in proportion with each other such that the dN/dS is not significantly different from the background rate. Similarly in deer we see that both the nonsynonymous and synonymous rates are increased relative to the background, thus dN/dS is not different from the background. This is consistent with a lack of gene wise positive selection previously described[70]. Unlike Epsilon, in 3 out of 4 deer clades there is an accelerated molecular clock, (Fig 2.3; Supp table B.1), which is consistent with previous findings[70].

**Table 2.1.** Estimate of dN/dS

| | Length (person years) | dN/dS |
|---|---|---|
| Alpha | 1.20 | 0.43 |
| Beta | 2.34 | 0.45 |
| Gamma | 1.32 | 0.94 |
| Delta | 2.84 | 1.58 |
| Epsilon | 0.56 | 0.54 |
| Iota | 2.04 | 1.07 |
| B.1.637 | 29.59 | 0.51† |
| Mu | 1.44 | 0.82 |
| BA.1 | 2.43 | 1.45 |
| BA.2/BA.4/BA.5 | 1.37 | 1.91 |
| Alpha_x04 | 1.24 | 5.35 |
| B11_x01 | 0.32 | 1.64 |
| B11_x12 | 0.49 | 1.24 |
| B1637_x10 | 0.32 | 0† |
| Delta_x14 | 0.36 | 27.62†† |
| Delta_x11 | 0.5 | 3.04 |
| Iota_x02 | 0.55 | 30.28†† |
| Deer_x04 | 1.23 | 0.39 |
| Deer_x05 | 0.91 | 0.54 |
| Deer_x01 | 1.3 | 0.43 |
| Deer_x02 | 0.82 | 0.39 |
| Deer_x03 | 0.45 | 0.48 |
| | | |
| Background | | 0.63 |

† 95% HPD includes 0 nonsynonymous substitutions
†† 95% HPD includes 0 synonymous substitutions

For the variant stem lineages of Beta, Delta, Iota, B.1.637, Mu, and BA.1 dN/dS is elevated relative to the background, as a consequence of the increased nonsynonymous rate and decreased synonymous rate (Table 2.1). The exception to this pattern is the dN/dS of B.1.637, which is inferred to be lower than the background, however because the 95% HPD of the stem lineage includes zero synonymous substitutions, the maximum likelihood based approach is heavily biased. For 2 persistent infections, as with B.1.637 the 95% HPD includes zero synonymous substitutions and 1 includes zero nonsynonymous substitutions, which makes inference of dN/dS difficult (Table 2.1). For Beta, Iota, and Mu there is no significant change in the overall rate of evolution (Fig 2.3). The increased rate of nonsynonymous substitutions is not reflected in an increase in the overall rate, as seen in Alpha, because the slower rate of

synonymous substitutions obfuscates it. We also see this pattern of increased dN/dS and no change in overall rate in 3 of 7 persistent infections (Table 2.1; Fig 2.3). Similarly for Delta and 1 persistent infection there is a slower rate of synonymous substitutions, increase in dN/dS and the overall molecular clock. Although the magnitude of the change in molecular clock in the Delta stem is small in comparison with Alpha, because of the contribution of the synonymous substitution rate.

## 2.3 Discussion

We investigated molecular clocks and selection on the stem lineage that preceded different variants. We found the synonymous substitution rate of variants could be divided into two primary categories where the synonymous substitution rate was higher than expected (eg. Alpha, Epsilon) and where the synonymous substitution rate was lower than expected (eg. Beta, Delta, Iota, Mu, BA.1). The slower rate of synonymous substitutions can obfuscate the increase in nonsynonymous substitutions when looking at the overall molecular clock which has led to the suggestion that Delta could have emerged from typical transmission[52]. However the rate of synonymous substitutions has broadly remained unchanged during typical transmission throughout the pandemic[60], and so typical transmission is not consistent with the history of the stem lineage of Delta and other variants.

We note that we characterize the typical transmission based on surveillance sampling because most evolution on this tree is associated with evolution during transmission between acute infections, however we have not removed any atypical evolution that could be associated with onward transmission from a persistent infection, or other distinct mechanisms, as long as it was not considered a variant. The typical transmission we characterize likely includes evolution from persistent infections that were unidentified, and so our comparisons between this behavior and persistent infections is a conservative estimate. We suggest that persistent infections can have a range of evolutionary behaviors that deviate from typical acute transmission (increased or decreased rate of nonsynonymous and synonymous substitutions), and that variants arise when virus in a persistent infection find a fitness peak and transmit.

In contrast to most variants, SARS-CoV-2 in deer, after reverse zoonosis, showed an

increased rate of nonsynonymous and synonymous substitutions. This increase is consistent with some variants (Alpha). However, transmission in deer SARS-CoV-2 tends to acquire characteristic substitutions which are not seen in variant stem lineages[69;70], suggesting reverse zoonosis is not the primary source of variants.

We explored the dynamics of SARS-CoV-2 persistence infections identified in surveillance sampling. Overall we find that some persistent infections have slower rates of synonymous substitutions, as was seen in some variant stem lineages, which suggests intra-host evolution during persistent infection is a possible source of variants. It has been suggested that the slower rate of evolution could be due to lingering non-replicating and non-viable RNA[73], however we note that there are less synonymous substitutions than expected on the internal branch of the Alpha_x04 clade and the descendants still accumulate mutations (Fig 2.1B) indicating continued replication of virus with slower synonymous substitutions.

SARS-CoV-2 is typically present in the upper respiratory tract, but can move to other areas of the body, where it shows distinct selection patterns[74–77]. It is possible that different body compartments and cell tropism could cause a change in replication rate which would affect the synonymous substitutions rate. One strategy in antiviral pharmaceuticals is to inhibit viral replication, for example Remdesivir inhibits SARS-CoV-2 replication[78]. Slowly replicating virus as a major component in persistence could reduce the effectiveness of such a strategy to treat persistent infections. To understand how SARS-CoV-2 evolves in different areas of the body, we still need more samples from the diverse body compartments. Based on wastewater samples which may represent viruses that are prevalent in the gut, rather than the respiratory tract, there is evidence of different patterns of evolution, however samples from other tissues are limited in comparison to samples from the upper respiratory tract[23].

Ebola virus has been found persisting in immunoprivileged sites where it experienced significantly slower rates of replication[65]. Transmission after persistence likely seeded an outbreak of Ebola virus that was preceded by a stem lineage that had less synonymous substitutions than expected given the amount of time of evolution[63]. Similarly, in SARS-CoV-2 we observe replicating virus in different body compartments[77], persistent infections with slower rates of synonymous substitutions, and stem lineages with molecular clocks different from the background;

together this suggests that SARS-CoV-2 variants could be from persistent infections where virus replicates in different body compartments under conditions different from most acute infections. Persistent infection as a source of variants explains the observation that new variants are not always derived from the current most widely circulating virus, but can be descended from any past virus; which has important implications for vaccine design, in that we cannot assume that new variants will be antigenically similar to the current most prominent strain but can be descended from any virus that has circulated in the past.

## 2.4 Methods

### 2.4.1 Non-variant context

We analyzed SARS-CoV-2 full genome sequences (<1% Ns) from New York City Public Health Laboratory (PHL) collected from February 7th, 2020 to June 1st, 2022 (n=10,572). Pangolin lineages (v4.1.2) were assigned using UShER[79]. Genomes were aligned with MAFFT[80] to the reference genome Wuhan Hu-1, and coding regions were extracted for further analyses.

We used a random subset of 150 genomes from all genomes not assigned to a Pangolin variant to represent the early background context (B.1-like) (Supp data). Using BEAST (v1.10.5)[40] we inferred a molecular clock rate for these early sequences using only the coding regions of $5.4\text{x}10^{-4}$ substitutions/site/year (HPD: $4.9\text{x}10^{-4}$ - $5.9\text{x}10^{-4}$ substitutions/site/year).

### 2.4.2 Variant stem lineages

For each variant we used the B.1-like background context and all variant sequences to infer a tree with IQTREE (v1.6.12) (Supp data)[81]. Using these trees we defined a variant stem lineage as the path between B.1 or B.1.1 to the base of the variant polytomy expansion. For Alpha, Beta, Gamma, and Mu the variant stem was a single branch. The Epsilon variant stem is the branch leading to B.1.427, the branch leading to B.1.429 and their parent branch. Delta, Iota, and B.1.637 variant stem lineages are the set of branches preceding the polytomy expansion that is characteristic of variants. We subset Omicron lineage into two sets for defining the variant stems (BA.1 alone and BA.2 BA.4 and BA.5 combined), but included all Omicron sequences in one alignment for inference. The stem lineage of BA.1 is the branch leading to the most recent

common ancestor (MRCA) of BA.1, BA.2, BA.4, and BA.5 and the branch leading to the MRCA of BA.1. The stem lineage of BA.2, BA.4 and BA.5 combined was multiple branches leading from the MRCA with BA.1 to the MRCA of each lineage (BA.2, BA.4, BA.5).

We note that because each variant stem lineage is defined without other variant sequences there is some shared evolutionary history. There is some shared history from the base of the B.1 polytomy to Beta, Epsilon, Iota, Mu, and B.1.637. And there is some shared history from the bases of B.1.1 to Alpha, Gamma, and Omicron. However, less than 20% of any stem lineage is shared with another variant lineage, with exception of B.1.637 and Iota sharing more. Delta does not have shared evolutionary history with any other stem lineage.

### 2.4.3 Persistent intra-host infections

To identify persistent infections we used all genomes from individuals with more than 2 SARS-CoV-2 genomes sampled more than 30 days apart by New York City Public Health Laboratory, NYC Pandemic Response Laboratory, or Department of Health Wadsworth Center (n=373). We inferred a phylogenetic tree using IQTREE using the GTR+F+G substitution model, retaining polytomies, and enforcing a minimum branch length of 10-9, and identified persistent infections as individuals that had multiple genomes that formed a cluster (samples could be connected without passing through a node representing a sample from another individual). We found a total of 15 individuals with persistent infections, characterized by 42 sequences (Supp data). For each case we inferred a local clock with BEAST, using sequences of the same variant, or the B.1-like subset, as context.

### 2.4.4 Deer associated evolution

To investigate SARS-CoV-2 evolution associated with a deer host we considered 5 clades, described by 22 genomes[70]. Clades were chosen as a random subset of the defined clades that had more than 0.33 years of evolution inferred in previous analysis.

### 2.4.5 Bayesian local clock inference and robust counting

We inferred a local clock for each variant stem lineage, persistent infection, and deer clade using BEAST. For each variant stem lineage we used the B.1-like background context

and all variant sequences. For persistent infections and deer clades, all sequences of the same variant from PHL assigned were used for context. Context sequences (variant and B.1-like) had a tight molecular clock prior ($5.4 \times 10^{-4}$ substitutions/site/year, standard deviation: $2.3 \times 10^{-5}$), inferred previously. We inferred a local clock on the stem of the variant clades, and within the persistent infection and deer clades. For Omicron, the base branch (parent branch of BA.1 and BA.2 BA.4 and BA.5) shared a local clock with the BA.1 stem, while the stem lineage of BA.2 BA.4 and BA.5 and another local clock. We included the base branch for both stem lineages in the post-BEAST analysis. The local clock was set as the product of the background rate and an increment factor with a log normal prior centered on 1. For each analysis we ran 2 chains for $5 \times 10^8$ to $10^9$ states and investigated the chain convergence using Tracer[42], then used LogCombiner[40] to resample and combine the obtained tree distributions at a lower frequency. We used robust counting[72] to infer the amount of synonymous and nonsynonymous substitutions on branches. We did not consider insertions and deletions or noncoding regions.

For each set of branches of interest we summarized the person years (sum of the length of branches of interest), clock rate, synonymous substitution rate, and nonsynonymous substitution rate as compared to the background context. The posterior probability of the difference of the rate relative to the background rate is calculated by counting the number of states in which the local rate is greater than the background for each tree in the posterior, and taking the maximum of the proportion of states that are greater, or one minus that proportion (eg the minimum posterior is 0.5). Moving forward with the results we considered only variant stem lineages, persistent infections, and deer clades with more than 0.25 person years. The variant under monitoring B.1.1.519 and 5 persistent infections were excluded due to this criteria.

### 2.4.6 Inference of dN/dS

Using BUSTED (v2.5.42)[82], we inferred the dN/dS of the branches of interest described above.

## 2.5 Acknowledgments

Chapter 2, in full, is a reformatted reprint, permission pending, of the material as it appears in "Synonymous substitution rates preceding the emergence of SARS-CoV-2 variants and during persistent infections." Jennifer L. Havens, Karthik Gangavarapu, Jade C. Wang, Elizabeth Luoma, Jonathan E. Pekar, Helly Amin, Steve Di Lonardo, Faten Taki, Enoma Omoregie, Scott Hughes, Kristian G. Andersen, Tetyana I. Vasylyeva, Marc Suchard, and Joel O. Wertheim. *In Preparation*, 2024. The dissertation author was the primary investigator and first author of this text.

# Chapter 3

# Dynamics of natural selection preceding human viral epidemics and pandemics

Viral epidemics of zoonotic origin are an important factor in global human health. It has been proposed that zoonotic viruses need to adapt to humans before successfully spreading. To investigate this hypothesis, we characterized the change in strength of positive and purifying selection preceding and after the zoonotic emergence, relative to selection in the natural host, of Ebola virus, Marburg virus, influenza A H1N1 2009 pandemic virus, SARS-CoV, and SARS-CoV-2. We found that there is no evidence of change in selection preceding zoonotic emergence. In contrast, in the early stages of outbreaks after a host switch there is a change in selection. That we consistently see adaptation only after zoonosis suggests that zoonotic emergence occurs when a virus of sufficient transmissibility, that exists in the natural reservoir, comes into contact with humans, jumps, and then spreads and adapts. We also investigated the reemergence of H1N1 in 1977 and found a relaxation of selection preceding the reemergence, which is consistent with laboratory passage; supporting the theory the H1N1 reemergent virus was in a laboratory setting prior to 1977.

## 3.1 Introduction

Human history has been marked by viral epidemics of zoonotic origin—viruses transmitted from a vertebrate reservoir to humans—recently including Ebola virus (EBOV), Marburg virus (MARV), HIV-1, influenza A virus (IAV), SARS-CoV, MERS-CoV, SARS-CoV-2, and mpox virus[83–90]. The processes by which a zoonotic spillover starts an epidemic are unclear. The

traditional framework posits that zoonotic viruses cross the species barrier without the ability to sustain human-to-human transmission and must evolve the necessary traits to spread and persist in humans[25;26]. This adaptation is predicted to occur before human outbreaks, either in the natural host reservoir, in an intermediate host, or during initial stuttering transmission in the human population. Alternatively, the viruses that establish zoonotic epidemics may have the ability to transmit between humans prior to emergence, and once in the human population there is adaptation to become even more successful at transmitting and spreading.

Adaptation to increase transmissibility has been documented to occur shortly after a novel zoonotic virus establishes in humans. During the SARS-CoV 2002–2003 outbreak in humans the virus acquired increased binding efficiency to human ACE2, which is an important receptor for successful infection of human cells[91–93]. During the 2013-2015 Ebola epidemic in West Africa, the virus accumulated mutations that increased tropism for human cells[94]. SARS-CoV-2 accumulated advantageous mutations within the first year of widespread circulation, including D614G which increased transmissibility[95;96] and N501Y together with a cluster of other mutations that arose in multiple variants of concern[15].

As viruses switch between host species or experience other changes in their environments, the intensity, location, and temporal dynamics of selective forces on their genome changes as well. Comparative phylogenetic methods which estimate and compare $\omega$ distributions, where $\omega$ is a parameter estimation of dN/dS, have found broad use in understanding how viruses evolve adaptations[16;17;97–99]. We consider $\omega$ greater than 1 indicative of diversifying or positive selection and $\omega$ less than 1 indicative of purifying selection, the selective regime is a combination of these pressures, which are both impacted by the environment. For example, when viruses are passaged without artificial selection there can be a relaxation of selection[100;101], which we attribute to removing host specific selective pressures, such as competition with the immune system which drives positive selection, or maintaining optimized host specific receptor binding driving purifying selection. Here, we use the RELAX[102] modeling framework to conduct a statistical comparison between sets of branches in viral phylogenies, specifically suited to determine whether viral genomes experience intensified or relaxed selection (both purifying and diversifying selection combined) when they transition to different evolutionary environments. Because recombination

and reassortment in the evolutionary history of zoonotic viruses can confound selection inference, we have extended the RELAX framework to account for multiple evolutionary histories in different genomic segments.

We characterized selection regimes for zoonotic viruses (i) in their natural hosts, (ii) on the stem lineage of the phylogenetic tree immediately preceding a zoonotic viral emergence, (iii) and in the early stages of outbreaks in humans. The stem lineage encompasses evolution associated with early less successful transmission after zoonosis before the most recent common ancestor of extant lineages, as well as any evolution associated with unsampled intermediate hosts. If adaptation facilitating human-to-human transmission occurs prior to the zoonotic jump or during stuttering transmission, we would expect a change in the selection regime on the stem lineage preceding the human outbreak. Alternatively, if the ability to transmit between humans is not the result of selection in the natural host, intermediate host, or during stuttering transmission, but rather a property of the virus in its natural host reservoir, then we would expect the selection regime will not change until the virus is transmitting during a successful outbreak. We explore patterns of changes in selection regime in recent human zoonotic viruses: Orthomyxoviridae, Filoviridae, and Coronaviridae.

## 3.2 Results

### 3.2.1 Selection regime preceding and during epidemics with zoonotic origin

We first explored selective dynamics preceding zoonotic emergences of the 2009 H1N1 influenza A virus pandemic, the West African Ebola virus epidemic, and the Angolan Marburg virus epidemic (Figure 3.1A-C). We compared the selection regimes along viral lineages preceding and during these zoonotic epidemics with selection regimes within the viral host reservoir using the RELAX framework[102]. The RELAX testing framework summarizes diversifying and purifying selection which characterizes the selection regime, and relates the selection regime of test and reference partitions by a selection intensity parameter. Selection intensity (K) greater than 1 indicates intensification, where positive and purifying selection both increase in strength, and K less than 1 indicates relaxation, where positive and purifying selection decrease in strength.

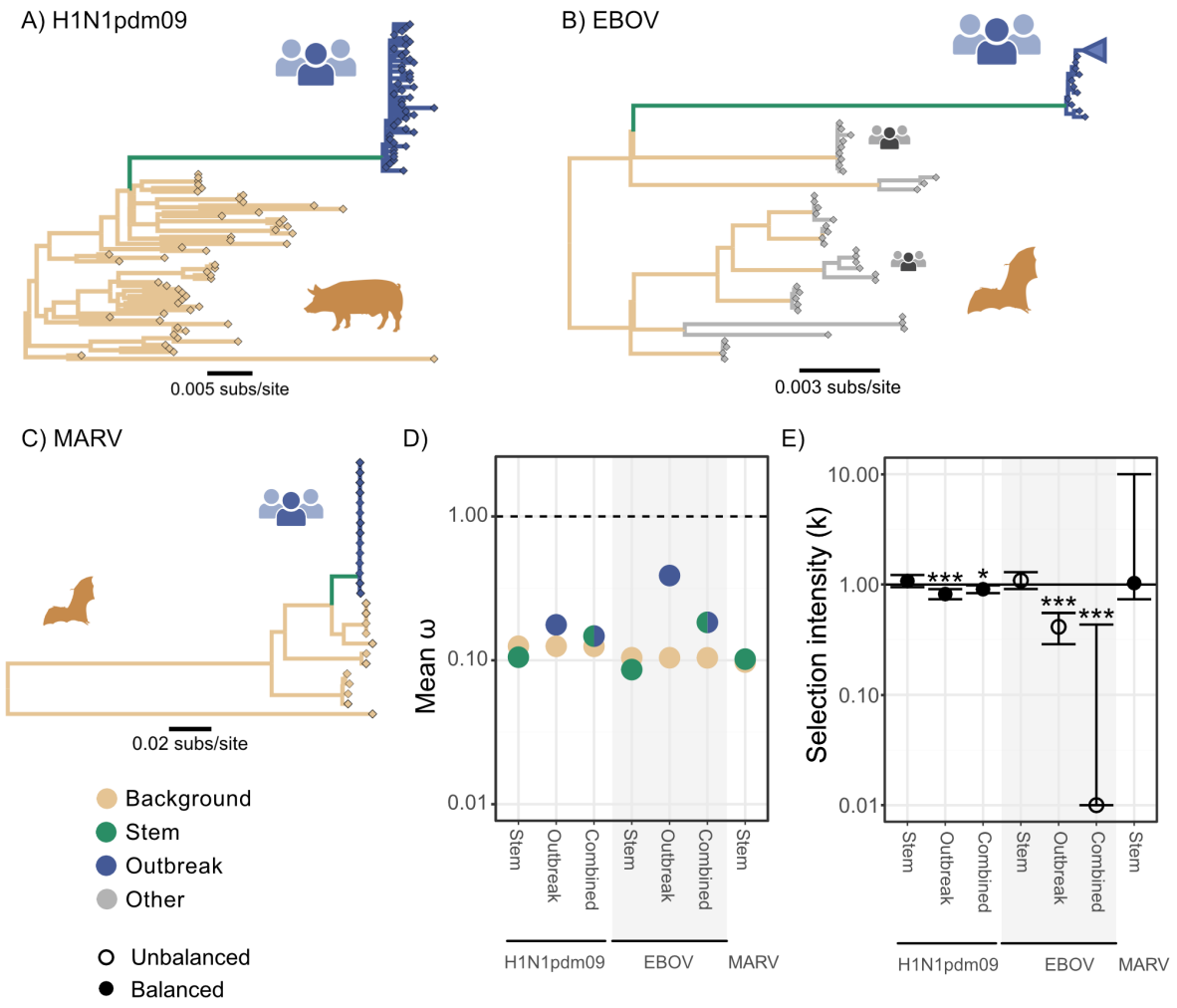



**Figure 3.1. Quantifying selection regimes of key epidemics with zoonotic origin.** Phylogenetic tree of (A) H1N1 2009 pandemic (H1N1pdm09 ) (B) 2013-2015 Ebola virus outbreak (EBOV) (C) the Angola 2004-2005 outbreak of Marburg virus. (D) Mean estimated $\omega$ for each branch partition of background and the test sets including "Combined" which collates the stem and the epidemic clade in test branches (color coded to match the tree) (E) Change in selection intensity comparing selection regimes between background and test branches . Filled circles indicate a balanced model where directionality is identifiable, open circles indicate an unbalanced model and the direction above or below 1 is not identifiable. Significance indicated with * $p$<0.05; ** $p$<0.01; and *** $p$<0.001. Numerical values in Supp Table C.2. Icons created with BioRender.

The 2009 H1N1 influenza A virus (H1N1pdm09) pandemic arose from a novel reassortment of influenza genomic segments from swine and avian viruses. The stem lineage preceding this pandemic encompasses the 9-12 years of evolution separating the most recent common ancestor (MRCA) of H1N1pdm09 in humans and their closest swine virus relatives[103]. We found no

evidence for intensification or relaxation of selection (K=1.1; $p$=0.21) on the stem branch leading to the human pandemic viruses compared with the branches associated with swine viruses (reference partition; Fig 3.1A). We find that the majority of evolution on the lineage preceding the emergence of H1N1pdm09 is not distinct from typical swine influenza virus.

In contrast, we inferred relaxation of selection for the human pandemic within a year following emergence, compared with selection in swine hosts. During these early months of the 2009 H1N1 pandemic, the mean $\omega$ of the virus in humans was greater than the mean $\omega$ associated with viral evolution in the swine host (Sup Table C.2; Fig 3.1D). This increase in $\omega$ was not due to an increase in positive selection but a decrease in purifying selection; we inferred that the human virus experienced a relaxation of selection compared with the selection regime in swine viruses (K=0.82, $p$<0.01). H1N1 during the human outbreak is in a different evolutionary environment than when in swine hosts. When we include the stem lineage as well as pandemic clade, the relaxation of selection compared with background swine virus selection is detectable, though less pronounced (K=0.91, $p$=0.01). That we can detect a change in the selection regime even when the stem lineage is induced with the outbreak, suggests that if there is extensive adaptation along a stem lineage we would be able to detect it even if the host switch is unsampled.

The 2013-2015 West African Ebola epidemic was the largest outbreak to date of Ebola virus (EBOV) disease[83], occurring thousands of kilometers from previously documented human cases of EBOV infection. We tested for a change in the selection regime along the stem lineage leading to the West African EBOV outbreak relative to viruses in the (presumed) bat reservoir along phylogenetic branches between human outbreaks[104] (Fig 3.1B). We found no support for intensification or relaxation preceding the West African epidemic (K=1.09, $p$=0.34), indicating bat virus-like evolutionary dynamics up until the MRCA of the human epidemic clade.

We did find evidence for a significant change in the selection regime during the first 18 months of human-to-human transmission during the West African EBOV epidemic compared with selection on viruses in the host reservoir (K<0.01, $p$<0.01); this signal was also detected when we considered both the stem lineage and viral evolution within the epidemic clade (K=0.41, $p$<0.01). The inferred selection regime for EBOV had a zero weight component for positive selection, confounding our ability to attribute changes in $\omega$ to relaxation or intensification. We

describe this behavior as an "unbalanced model". As with H1N1pdm09, there was an increase in mean $\omega$ in the epidemic clade compared with the background (Sup Table C.2). An increase in mean $\omega$ can be explained by relaxation of purifying selection or intensification of positive selection, the increase does not inform the direction of the change in selection.

Marburg virus (MARV) is associated with hemorrhagic fever, and the natural host reservoir is thought to be bats. One of the largest known MARV outbreaks occurred in Angola 2004-2005[90]. We characterized selection along the stem lineage preceding the epidemic MRCA and compared it with the selection regime on branches between MARV sampled in bats (Fig 3.1C) and found no evidence of intensification or relaxation (K=1.03, $p$=0.90). We were unable to explore selection regimes within the human MARV outbreak due to insufficient viral diversity in the too few sampled genomes.

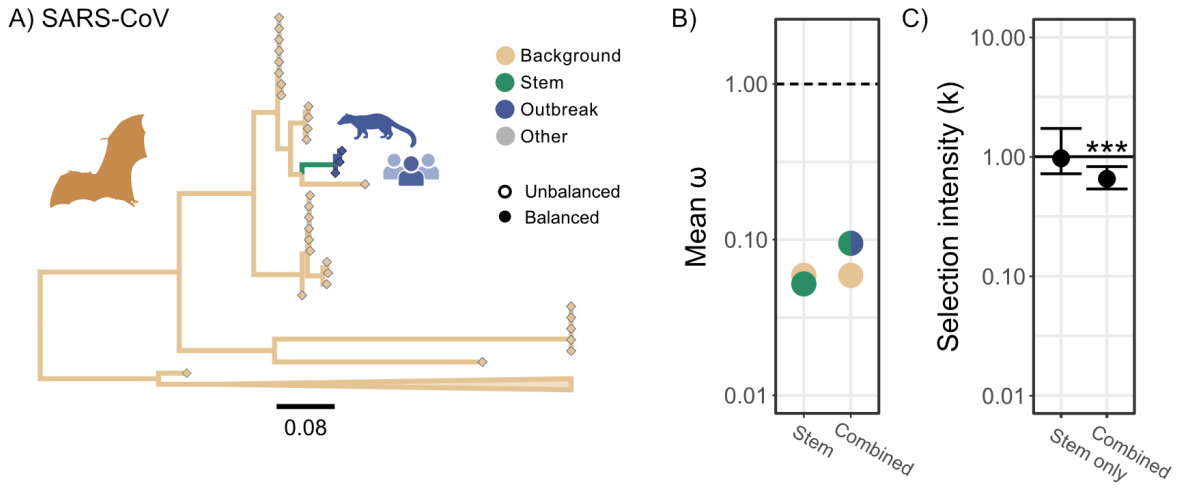



**Figure 3.2. Quantifying selection regime in SARS-CoV.** (A) Phylogenetic tree of SARS-CoV like sarbecoviruses non-recombinant region 22, branch color indicate partition. The SARS-CoV outbreak includes 2 human sequences and 1 civet sequence, which are monophyletic across all regions. (B) Mean $\omega$ for each branch partition of branch sets, color coded to match the tree (C) Change in selection intensity between background and test sets. Filled circles indicate a balanced model where directionality is identifiable, open circles indicate an unbalanced model and the direction above or below 1 is not identifiable. Numerical values in Supp Table C.2. Significance indicated with* $p$<0.05; ** $p$<0.01; and *** $p$<0.001. Icons created with BioRender.

### 3.2.2 Selection regimes in Betacoronaviruses prior to spillovers

Betacoronaviruses are an ancient viral group circulating in bats that have repeatedly experienced zoonotic jumps into humans. The 2002-2003 SARS-CoV pandemic was traced back to palm civets sold for human consumption in live-animal markets in China[105]. The close contact between humans and these infected intermediate host animals likely facilitated repeated spillover. We analyzed 139 SARS-CoV-like sarbecovirus genomes sampled in bats, two human SARS-CoV genomes representing distinct zoonotic introductions, and one SARS-CoV genome sampled from the palm civet intermediate host (Fig 3.2A). Using coding sequences from 31 putatively non-recombinant genomic regions[106], we compared selection along the stem lineage leading from bat SARS-CoV-like viruses to the MRCA of human and palm civet SARS-CoV. We found no significant intensification or relaxation (K=0.97, $p$=0.91) along the stem lineage. The lack of change indicates that, prior to infecting palm civets, the bat SARS-CoV-like virus had not experienced extensive evolution in an evolutionary environment different from the environment in the bat host. However, when we include the lineages encompassing evolution in the intermediate palm civet host preceding emergence in humans, we find that the SARS-CoV selection regime was significantly relaxed compared with related bat viruses (K=0.65, $p$<0.01).This change in selection is evidence that after the host switch the virus is in a new evolutionary environment, and suggests that it is adapting to the new host.

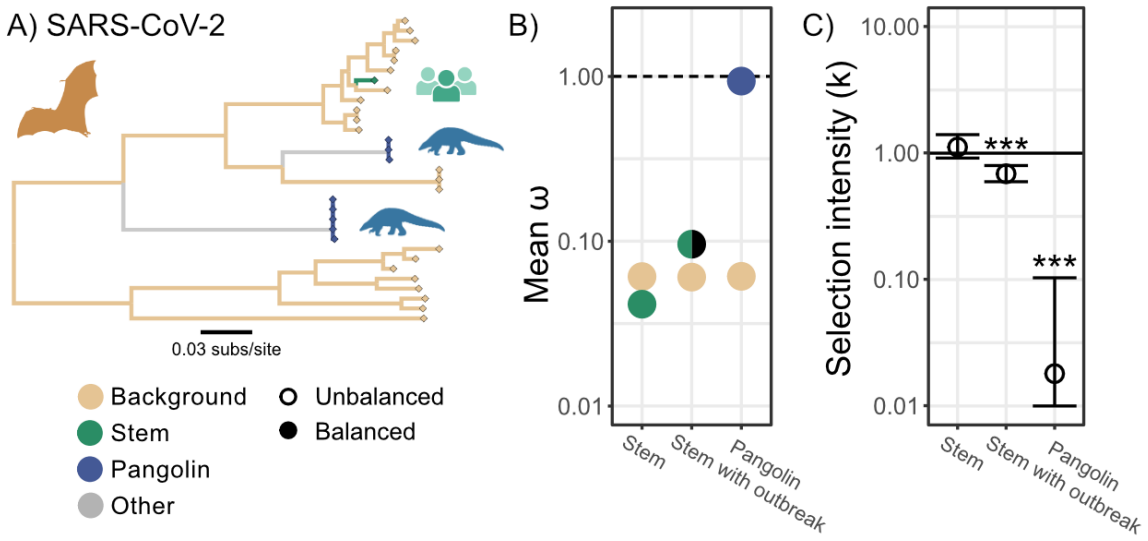



**Figure 3.3. Quantifying selection regimes in SARS-CoV-2.** (A) Phylogenetic tree of SARS-CoV-2 like sarbecoviruses non recombinant region 08, branch color indicate partition. (B) Mean $\omega$ for each branch partition of background and the test sets, color coded to match the tree (Supp Fig C.1) and includes both the stem and outbreak partitions in the test (C) Change in selection intensity comparing selection associated with background to test set. Filled circles indicate a balanced model where directionality is identifiable, open circles indicate an unbalanced model and the direction above or below 1 is not identifiable. Numerical values in Supp Table C.2. Significance indicated with* $p<0.05$; ** $p<0.01$; and *** $p<0.001$. Icons created with BioRender.

In late 2019, SARS-CoV-2 began spreading in central China, thousands of kilometers away from its closest sampled bat virus relatives in southern China and Laos[107]. Across the coding sequences from 15 putatively non-recombinant regions[106], we found no evidence of intensification or relaxation between the selection regime in the bat host reservoir and the stem lineage leading to SARS-CoV-2 emergence (K=1.1, $p$=0.23; Fig 3). Hence, SARS-CoV-2 did not experience extensive evolution in an environment different from related bat viruses or substantial adaptation prior to its emergence in humans. This result does not change if we use a different approach to identifying non-recombinant regions (K=1.02, $p$=0.82; Supp Fig 1), and is consistent with previous analysis[97].

When we consider evolution along the stem lineage in combination with viral evolution during the first 3 months of the outbreak in China, we find evidence for a significant shift in selection regime, consistent with rapid adaptation to humans (K=0.69 , $p<0.01$). As with EBOV we cannot confidently infer the directionality of this change because the model is unbalanced.

The change in selection regime is detectable when the clade of the first world-wide wave, up to October 2020, is included with the stem compared with bat virus background (K=0.56, $p$<0.01; Supp Fig C.1).

In addition to humans, SARS-CoV-2-like sarbecoviruses have also been sampled from pangolins confiscated during investigations of the illegal wildlife trade[108]. We compared selection within the pangolin virus clades with the bat host SARS-CoV-2-like sarbecoviruses. We find a change in the selection regime in pangolin SARS-CoV-2-like virus compared with the viruses in the bat host reservoir (K=0.02, $p$<0.01), suggesting that pangolins can maintain circulation of SARS-CoV-2-like viruses long enough to allow for host specific evolution. However, like EBOV the model is unbalanced and we cannot identify the directionality of the change. Although we do note that the mean $\omega$ associated with pangolin host virus evolution is much greater than the background.

### 3.2.3 The reemergence of H1N1 in 1977

In 1977, H1N1 influenza A virus reemerged in humans after going extinct 20 years prior. This new virus was closely related to a strain that had been circulating in the 1950s[109;110]. It has been proposed that H1N1 reemerged in 1977 without genetic change because it was frozen in a laboratory before being accidentally released as part of a live attenuated vaccine trial or challenge[111]. We infer only 48 substitutions across the entire genome along the stem lineage preceding the 1977 reemergence, with multiple genomic segments experiencing only 2 substitutions (Supp Table C.1). Hence, the 1977 strain is too genetically similar to 1950s H1N1 virus to support a natural reemergence, however the provenance of this virus prior to reemergence remains poorly understood.

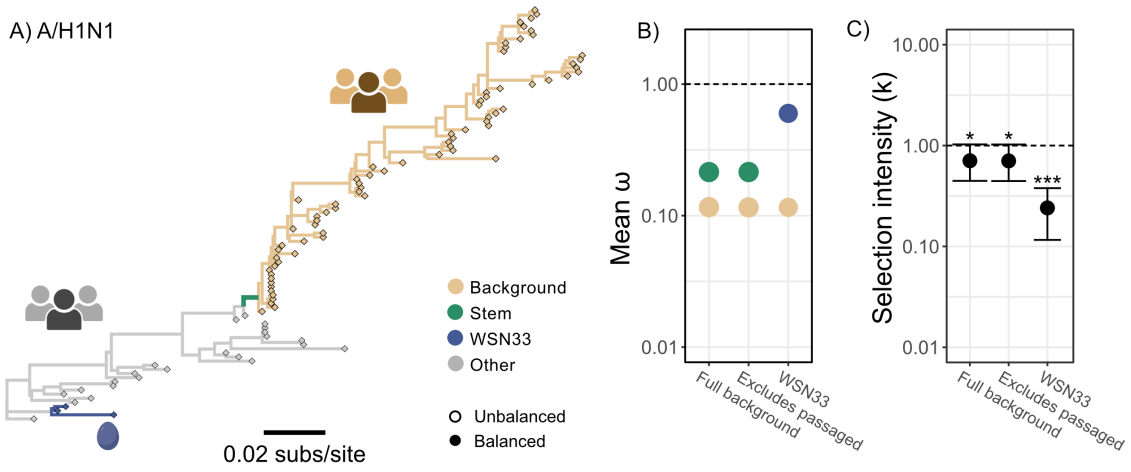



**Figure 3.4. Quantifying selection regimes in H1N1.** (A) Phylogenetic tree of H1N1 HA segment, branch color indicates partition. The stem branch connects 1977 clade and basal ancestors, and is foreground for 1977 flu test. WSN33 (blue) branches are clade associated with laboratory passage. Other branches are excluded from analysis (treated as a nuisance group). (B) Mean $\omega$ for each branch partition of background and the test sets. "Full background" compares selection on stem to displayed background. "Excluded passaged" compares selection on Stem to background excluding tips which have history of passaging (49/74; supp fig xx) "WSN33" compares selection of WSN33 branches to the full background. (C) Change in selection intensity comparing selection associated with background to test set. Filled circles indicate a balanced model where directionality is identifiable, open circles indicate an unbalanced model and the direction above or below 1 is not identifiable. Numerical values in Supp Table C.2. Significance indicated with* $p$<0.05; ** $p$<0.01; and *** $p$<0.001. Icons created with BioRender.

We compared the selection regime of the coding regions across all H1N1 influenza A virus segments on the stem lineage preceding the 1977 pandemic with the evolution post-1977 (Fig 3.4). We found evidence for significant relaxation of selection along the stem lineage (K=0.71, $p$=0.043), in addition to an increase in mean $\omega$. This change in selection regime suggests that in addition to being frozen between the 1950s and 1977, the precursor to the 1977 H1N1 pandemic experienced a limited amount of evolution in an environment distinct from human-to-human transmission (e.g. laboratory passage for attenuation or under artificial selection). This result remains supported even if we exclude the tips associated with influenza A virus isolates that have a known history of laboratory passaging (Supp Fig C.2; K=0.70, $p$=0.042).

### 3.2.4 Laboratory passage of H1N1 influenza A viruses

To better contextualize the atypical evolutionary dynamics preceding the 1977 H1N1 pandemic, we explored whether the signature of laboratory passage would be detected as a change in the selection regime characterized by RELAX. The Wilson Smith 1933 H1N1 (WSN33) isolate is the ancestor of many laboratory strains used in research. We compared the selection regime of 3 early laboratory isolates derived from WSN33 with the selection regime associated with human-to-human H1N1 transmission across all 8 genomic segments (Fig 3.5). There was significant evidence of relaxation of selection in the WSN33 clade (K=0.24, $p<0.01$; Fig 3.5). This relaxation of selection may be due to the removal of host specific constraints, such as competition with the immune system driving positive selection, and maintaining host optimized receptor binding driving purifying selection.

Zhang et al., 2006[112] claimed that influenza A virus persisted in lake ice based on HA gene sequences. However, these putative lake ice viruses were monophyletic and descended from the WSN33-derived positive control (Supp Fig. C.3), suggesting that these samples were instead serially passaged laboratory contaminants[112;113]. Selection on putative lake ice viruses was significantly relaxed compared with that on human H1N1 virus (K=0.1, $p<0.01$; Fig 3.5, which is consistent with laboratory passaging in the absence of selective drivers.

### 3.2.5 Selection on attenuated vaccine viruses

Measles virus (MeV) and mumps virus (MuV) passaged in cell culture to create attenuated viruses for vaccines evolved with increased mean $\omega$, consistent with a relaxation of purifying selection[100;101]. We compared the selection regime of MeV and MuV viruses passaged for vaccine attenuation with selection on branches of viruses transmitted between humans (Supp Fig C.4, Supp Fig C.5). In agreement with previous results, the vaccine attenuated MeV and MuV both have an increased $\omega$ relative to the background branches (Fig 3.5A). Furthermore, we were able to detect significant relaxation of selection in the vaccine attenuated viruses relative to selection in the human host for both MeV (K=0.03, $p<0.01$) and MuV (K=0.84, $p<0.01$) (Fig 3.5).

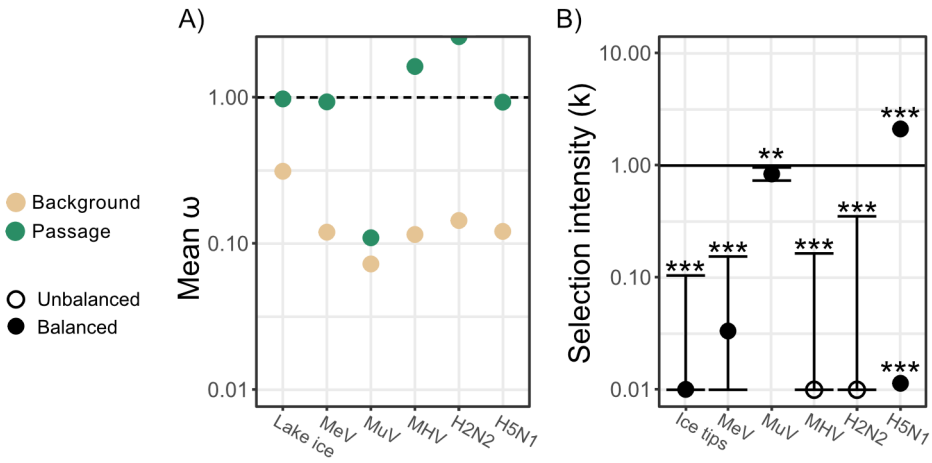



**Figure 3.5. Quantifying selection regimes of laboratory passaged viruses**(A) Mean $\omega$ for each branch partition of background and the test sets. (B) Change in selection intensity, comparing selection associated with transmission between humans (except H5N1 which is compared with selection in avian host reservoir, and MHV to selection associated with murine host) to (i) positive control of H1N1 (Lake ice) (ii) passaging in cell lines of Measles virus (MeV) (iii) Mumps virus (MuV) and (iv) Mouse hepatitis virus (MHV) (v) artificial selection for cold adaptation (H2N2) and (vi) gain of function passaging in ferrets to introduce airborne transmission (H5N1). Filled circles indicate a balanced model where directionality is identifiable, open circles indicate an unbalanced model and the direction above or below 1 is not identifiable. Numerical values in Supp Table C.2. Significance indicated with* $p$<0.05; ** $p$<0.01; and *** $p$<0.001. Icons created with BioRender.

### 3.2.6 Serially-passaged coronavirus

Murine hepatitis virus (MHV) is a betacoronavirus naturally found in mice. Numerous isolates of MHV have been propagated in laboratory cell culture. We compared the selection regime of MHV propagated in cell culture with the selection regime of naturally transmitting MHV (Supp Fig C.6), across 11 putatively non-recombinant regions. We inferred an increased mean $\omega$ in the passaged virus relative to the background. Further, we found evidence of a change in selection in the passaged viruses (K=0, $p$<0.01; Fig 3.5); however, as with Ebola virus, we are unable to identify the direction of change due to an unbalanced model.

### 3.2.7 Artificial selection on influenza A viruses

Artificial selection, e.g., passaging of viruses under selective constraints to produce novel phenotypes, is expected to alter the selection regime. An H2N2 influenza A virus was passaged

at colder temperatures to produce the non-surface proteins for a reassortant live attenuated vaccine that had attenuated replication at human body temperature[114]. We compared the selection regime associated with this cold adapted virus, with selection associated with H2N2 virus transmitted between humans (Supp Fig C.7). We found evidence of a change in selection of the non-surface proteins resulting from artificial selection (K=0, $p<0.01$; Fig 3.5), however the model is unbalanced and we cannot identify the direction of selection intensity change. Notably, we also inferred an increase in the mean $\omega$ in the passaged viruses relative to the virus in humans, consistent with adaptation due to laboratory passage.

In an artificial selection experiment, H5N1 influenza A virus was passaged between ferrets to gain airborne transmissibility[115]. We inferred a significant change in the selection regime due to artificial selection in the new host compared with virus in the host avian reservoir (Fig 3.5; Supp Fig C.8; $p<0.01$). However the inferred direction of change in selection intensity was unstable; relaxation (K=0.01) or intensification (K=2.12) did not have significantly different likelihoods in our model; this is partly due to the relatively low diversity of available sequence.

## 3.3 Discussion

To better understand the evolutionary pressures that lead to epidemics of zoonotic viruses in humans, we characterized the selective regime of these viruses in their natural hosts, immediately preceding zoonotic spillover, and within the epidemic itself. We find that for recent zoonotic epidemics-H1N1pdm2009, EBOV, MARV, SARS-CoV, and SARS-CoV-2-these viruses continued to evolve in a manner consistent with transmission in their host reservoir up until spillover into humans or an intermediate host. These findings are inconsistent with the traditional framework where zoonotic viruses cross the species barrier without the ability to sustain human-to-human transmission and must adapt before sustained human-to-human transmission can be achieved[25;26]. Instead, our findings support a framework whereby zoonotic epidemics result from viruses that are sufficiently transmissible in the human population, and then these viruses adapt over the course of an already established outbreak.

Previous work suggests that SARS-CoV-2 spent some time in an intermediate host before jumping into humans, although there was very little diversity in the viruses that jumped into

the human population[1]. We found that there was no significant change in selection between branches associated with bat hosts and the stem leading to SARS-CoV-2, which is consistent with a previous analysis[97]. This lack of change in selection is not inconsistent with an intermediate host. While we expect an intermediate host to result in a change in selection, a short time in the intermediate host with little accumulation of mutations (scenario depicted in Supp Fig C.9A) would not result in a detectable change in selection. Rather we only expect to detect a change in selection if there is extensive evolution associated with an intermediate host or initial stuttering transmission in humans on the stem lineage (Supp Fig C.9 B,C). In other situations where there is a detectable change in selection it may be attributable to passage in a laboratory context.

Influenza A/H1N1 reemerged in 1977 after a few decades out of circulation, but very little divergence occurred over this time, the conditions preceding reemergence is unclear. Viruses passaged without artificial selection showed relaxation of selection (WSN33, Ice tips, MeV, MuV), or a change in selection without an identifiable direction (MHV). Viruses passaged with artificial selection (H2N2, H5N1) showed a change in selection, although we could not determine the direction of this change. The change in selection regime of H1N1 preceding the 1977 reemergence suggests that it evolved in a context other than typical transmission. The change in selective regime is consistent with passaging at colder temperatures to attenuate replication at human body temperature, as seen with H2N2, or passaging in cell culture without artificial selection, as seen in WSN33 laboratory isolates. This is consistent with the theory that H1N1 was passaged in a laboratory context preceding reemergence, although we cannot distinguish the context[111].

It is not uncommon to focus on a single segment, or address each segment separately, when considering the history of influenza[102;116;117]. To address the challenge of reassortment we extended RELAX to account for different evolutionary histories by applying a single model to multiple sequence alignments with their associated phylogenetic trees. We also applied this multi-region analysis to combine non-recombinant regions for viruses where recombination was likely a large contributing factor (e.g. positive sense RNA viruses). By using the multi-region model we increased the power relative to what would have been possible with a single region, and have detected a significant relaxation of selection on the stem preceding the H1N1 1977 reemergence.

Although H1N1 1977 reemergence is not attributable to a natural reservoir, humans are constantly exposed to animal viruses[118;119]. However most of these exposures do not result in large outbreaks of human to human transmission. Small outbreaks of zoonotic viruses would provide the opportunity to adapt to humans. However this adaptation does not typically get introduced back into the host reservoir, and so does not contribute to the transmissibility of later outbreaks. Recent zoonotic epidemics have been sparked when a virus from the zoonotic host reservoir that is already capable of transmitting between humans, enters the human population and then begins to adapt.

## 3.4 Methods

### 3.4.1 Phylogenic selection analysis

For all influenza datasets coding regions were extracted from each genomic segment, then each region was aligned using MAFFT (v7.409)[80], alignments were validated manually. For all other data sets, full genomes were aligned to the reference genome and split into non-recombinant regions. For MHV, a positive sense RNA virus, putative non-recombinant regions were detected using GARD[120]. SARS-CoV, and SARS-CoV-2, predefined putative non-recombinant regions were used. For all data sets, coding regions were extracted based on the reference genome annotation. For each region a maximum likelihood phylogeny was inferred with IQtree2 (v2.0.4)[81] using the GTR+F+$\Gamma$4 model.

Change in intensity of selection was tested across all genomic regions with the multi-region (joint) RELAX model implemented in HyPhy (v2.5.53)[102]. In the RELAX testing framework we partition branches in a phylogenetic tree into "test" and "reference". Evolution along each tree branch is described using a continuous time discrete space Markov model of codon evolution; the ratio of non-synonymous to synonymous substitution rates ($\omega$) is used to model selection. Along each branch, the evolutionary process is modeled as an independent draw from discrete distributions on $\omega$, with one rate allocated to positive selection ($\omega \geq 1$) and one or two rates allocated to negative selection or neutral evolution ($\omega \geq 1$). Phylogenetic likelihood is computed by integrating over the assignments of $\omega$ to branches, and all model parameters, including $\omega$ distributions, are estimated by maximum likelihood. The selection regimes ($\omega$ distributions) of

the test and reference branches are related by a selection intensity parameter K, via $\omega$(test) = $\omega$(background)K. Selection intensity (K) greater than 1 indicates intensification, where positive and purifying increase in strength (move away from $\omega = 1$), and K less than 1 indicates relaxation, where positive and purifying selection decrease in strength (move closer to $\omega = 1$). We use the likelihood ratio test to determine whether K is significantly different from 1 (K=1 when the selection regime in test and reference are equal).

The joint RELAX test is an extension of RELAX in which multiple evolutionary histories (trees) share the inferred selection regime ($\omega$ distributions) and selection intensity test parameter (K). We ran joint RELAX for each context with every combination of 1) with and without multiple hits, which allows instantaneous multi-nucleotide substitutions, 2) with and without synonymous rate variation, which accommodates site-to-site variation of synonymous substitution rates, and 3) with 2 or 3 rate partitions, which has 1 or 2 partitions for purifying or neutral selection respectively. From these 8 options the best model was chosen with AIC.

### 3.4.2 Data sets

Acessension numbers and sequence names for all data are available upon request. Alignments and annotated trees are available upon request.

### 3.4.3 Selection regime preceding and during the H1N1 2009 pandemic

To characterize selection regime preceding H1N1pdm09, we used previously curated virus genomes[121], with up to 58 stains for each of the 8 influenza genome segments. Additionally, to characterize the outbreak, we considered 39 stains selected from Influenza Virus Resource search with parameters of pandemic H1N1 virus that are full length with all segments sampled from a human host in North America from the FLU project collected from 2009-04-01 to 2010-03-01. We used branches associated with transmission between swine, to represent selection associated with the host reservoir (reference partition). The stem branch leading to the MRCA of H1N1pdm09 was used as the test branch of interest. Additionally, selection on the stem branch preceding H1N1pdm09 and branches within the outbreak clade, was tested against the selection regime of the same reference partition.

### 3.4.4 Selection regime preceding and during Ebolavirus (EBOV) 2013-2015 West Africa outbreak

We analyzed 101 EBOV genomes; of these 50 were selected to represent known human outbreaks including 10 from the 2013-2015 outbreak additionally there were and 50 genomes subsampled from NCBI virus with search parameters of EBOV genomes with complete nucleotides from Guinea, Liberia or Sierra Leone, and human host with sample collected 1 January 2014 to 1 June 2015. EBOV is a negative sense single-stranded RNA virus, as such recombination is not expected to be a major factor in its evolutionary history, so we used all coding regions combined in a single genome wide analysis.

We used branches between outbreaks of EBOV in humans, to represent selection associated with the natural host reservoir (reference partition), which is thought to be bats[122]. Branches that were associated within human outbreak clades were excluded. The long branch leading to the MRCA of the 2013-2015 West African outbreak was used as the test branch of interest. Additionally, selection on the stem branch preceding the 2013-2015 outbreak and branches within the outbreak clade, was tested against the selection regime of the same reference partition.

### 3.4.5 Selection regime preceding and during Marburg virus disease Angola outbreak

We analyzed all MARV genomes sampled from bats available on NCBI (n=12) and 14 sampled from the 2005 outbreak in Angola[90] which is the largest known human outbreak of Marburg virus disease. Similarly to EBOV, MARV is a negative sense single-stranded RNA virus and we used all coding regions combined in a single genome wide analysis. The branches associated with bat samples were used to represent selection in the natural host reservoir, which is thought to be bats. The stem branch leading to the human outbreak was used as the test branch of interest.

### 3.4.6 Selection regimes in SARS-CoV-like virus

In 2002-2003 SARS-CoV spilled over into humans, likely from palm civets sold for human consumption in live-animal markets in China[105]. We used a total of 142 sequences, including 3 SARS-CoV sequences 2 from humans and 1 from a palm civet, using 31 predefined putative

non-recombinant regions[123]. We considered selection preceding the SARS-CoV outbreak by comparing the stem branch leading to the MRCA of SARS-CoV civet and human samples to the branches associated with a bat host. We also compared the early outbreak, using all branches in the SARS-CoV clade with the branch preceding it to selection in the host, using the same reference partition.

### 3.4.7 Selection regimes in SARS-CoV-2-like virus

We considered the selection preceding SARS-CoV-2 pandemic and compared it to selection of SARS-CoV-2-like sarbecoviruses associated with the natural bat host reservoir. We used a total 31 sequences[1] with 15 pre-defined putatively non-recombinant regions[107]. The robustness of changes in selection to non-recombinant regions was tested by repeating the RELAX analyses with non-recombinant regions inferred by 3seq[124], which defined 21 regions. We considered selection preceding the SARS-CoV-2 outbreak by comparing the stem branch leading to SARS-CoV-2, represented by the reference sequence Hu-1 to branches associated with the natural host reservoir of bats.

We also considered selection early in the SARS-CoV-2 pandemic using 2 different contexts. We used a subsample of 50 sequences representing the early outbreak in China, over 3 months, from a previous dataset[1]. Additionally we used a random sample of 39 sequences from NCBI Virus with complete genomes sampled from September 15th to October 1st, 2020 to represent early worldwide transmission. For each of these sequence sets representing the early outbreak we compared selection of the host reference partition to the outbreak clade and the stem leading to SARS-CoV-2.

Of the SARS-CoV-2-like sarbecoviruses, 8 were sampled from pangolins (the rest from bats), which form 2 pangolin clades[108;125]. The stem leading to the pangolin clades, and the pangolin clades were excluded from the reservoir host background. We compared selection within the pangolin clades to the background partition.

### 3.4.8 Selection of H1N1 influenza A viruses

In the 1950s A/H1N1 was replaced by another strain of influenza A virus. In 1977 an H1N1 strain nearly identical to that which circulated in the 1950s reemerged in the human population[109].

We used 100 genomes sampled from before 1950 and after 1977 to build a phylogenetic tree and define the branch leading to the 1977 pandemic clade as the stem lineage. To characterize typical human-to-human transmission we considered the branches of the 1977 pandemic clade. We excluded branches associated with samples from before 1977 from the background due to concerns of passaging, erroneous inclusion of lab strains, and sequence quality. To test for robustness we excluded 49 tips associated with isolates that have a known history of laboratory passaging. We tested if laboratory passage of virus isolates changed selection regime by comparing the background with the clade that includes 3 early laboratory isolates derived from WSN33[126]: WSN/33, and two derivatives A/United Kingdom/1/1933 and A/Wilson-Smith/1933(H1N1), all of which are known to be passaged before sequencing.

We inferred the synonymous and nonsynonymous substitutions preceding the 1977 reemergence using TreeTime (v0.8.6)[127] ancestral state reconstruction.

### 3.4.9 Selection of H1N1 influenza A viruses putatively from lake ice

Zhang et al. purportedly isolated H1N1 from Serbian lake ice, and suggested these samples were frozen in the ice from a local outbreak. We considered the 19 HA sequences purportedly isolated from lake ice, 1 positive control[112], and 19 HA genes from the H1N1 dataset from the 1930s to 1950s clade, excluding tips associated with sequences that have a history of lab passaging[113] for a total of 38 sequences. Branches associated with transmission between humans were used for the reference partition, and branches in the 'lake ice' clade were used as the test partition.

### 3.4.10 Selection on attenuated vaccine measles virus (MeV)

We analyzed 28 measles virus full genomes, that were curated for use to calibrate divergence[128]. This set of genomes is characteristic of MeV evolution with a human host, and branches associated with these samples were used to characterize background selection. We used 13 genomes from vaccine associated strains of MeV, previously curated[101], which were passaged in cell culture to produce an attenuated virus intended for use as a vaccine. Branches associated with the vaccine strains were used as the test branches.

### 3.4.11 Selection on attenuated vaccine mumps virus (MuV)

We analyzed 59 MuV genomes that were selected by subsetting from Nextstrain NA Real-time tracking of mumps virus evolution (nextstrain), and collecting the associated full genomes from NCBI. Branches associated with these samples were used to characterize background selection. We characterized MuV passaged in the lab for attenuation, with 10 MuV genomes associated with vaccine strains, which were previously curated[101].

### 3.4.12 Serially-passaged coronavirus MHV

MHV is a betacoronavirus naturally found in mice. Graepel et al passged MHV 250 times in cell culture[129], we used the passaging control that did not include experimental modifications to characterize selection associated with laboratory passaging, using branches leading to the passaged virus from the input wild type strain. Background selection was characterized with branches between 5 MHV lineage defining strains. 11 non-recombinant regions were inferred using GARD.

### 3.4.13 Artificial selection on influenza A virus A/H2N2

An H2N2 influenza A virus (wt A/Leningrad/134/57) was passaged at colder temperatures to produce the non-surface proteins for a reassortant live attenuated vaccine that had attenuated replication at human body temperature[114]. We characterized the selection regime of passaging with selection as the branches that were descendant from the wildtype leading to isolates sequenced at passages 17 and 47, and the input wild type sequence. The branch leading to the input wild type was excluded. The background context for comparing to cold adaptation of an H2N2 stain was characterized by a random subsample of 50 full H2N2 genomes from Influenza Virus Resource, with all proteins available, excluding vaccine strains and sampled from a human host.

### 3.4.14 Artificial selection on influenza A virus A/H5N1

Herfst et al., passaged virus derived from A/Indonesia/5/2005 10 times in ferrets, which was passaged twice more in two parallel sets of transmission[115]. We considered all branches in the clade of all available passaged sequences (n=6) and the branch leading from the initial

genome (A/Indonesia/5/2005) to this clade as the test partition. The background context was characterized by a random subsample of 94 whole H5N1 genomes from Influenza Virus Resource, with all proteins available, sampled from avian sources from 1980 to 2005. We compared selection dynamics along with branches associated with the H5N1 samples from the avian population (reference set). The branch leading to the human A/Indonesia/5/2005 from the avian samples was excluded from main analysis.

## 3.5 Acknowledgments

We gratefully acknowledge the authors and contributors who generated, or contributed to, the viral genomic sequences on which this research is based. We thank Mark Zeller, Tetyana I. Vasylyeva, and Marc Suchard for their insightful discussion.

Chapter 3, in full, is a reformatted reprint, permission pending, of the material as it appears in "Dynamics of natural selection preceding human viral epidemics and pandemics." Jennifer L. Havens, Sergei L. Kosakovsky Pond, Jordan D. Zehr, Jonathan E. Pekar, Edyth Parker, Stephen A. Goldstein, Michael Worobey, Kristian G. Andersen, and Joel O. Wertheim. *In Preparation*, 2024. The dissertation author was the primary investigator and first author of this text.

# Chapter 4

# Phylogeographic analysis reveals an ancient East African origin of human herpes simplexvirus 2 dispersal out-of-Africa

Human herpes simplex virus 2 (HSV-2) is a ubiquitous, slowly evolving DNA virus. HSV-2 has two primary lineages, one found in West and Central Africa and the other found worldwide. Competing hypotheses have been proposed to explain how HSV-2 migrated out-of-Africa (i)HSV-2 followed human migration out-of-Africa 50-100 thousand years ago, or (ii)HSV-2 migrated via the trans-Atlantic slave trade 150-500 years ago. Limited geographic sampling and lack of molecular clock signal has precluded robust comparison. Here, we analyze newly sequenced HSV-2 genomes from Africa to resolve geography and timing of divergence events within HSV-2. Phylogeographic analysis consistently places the ancestor of worldwide dispersal in East Africa, though molecular clock is too slow to be detected using available data. Rates $4.2\text{x}10^{-8}$-$5.6\text{x}10^{-8}$ substitutions/site/year, consistent with previous age estimates, suggest a worldwide dispersal 22-29 thousand years ago. Thus, HSV-2 likely migrated with humans from East Africa and dispersed after the Last Glacial Maximum.

## 4.1 Introduction

Human herpes simplex virus 1 and 2 (HSV-1, HSV-2) are globally ubiquitous human pathogens. HSV-1 and HSV-2 are slow evolving double-stranded DNA viruses that can infect via oral and genital routes, though HSV-1 is predominantly associated with oral herpes and HSV-2 with genital herpes[130]. Although genital herpes infection is mostly mild, HSV-2 infection can be

fatal in neonates and it is associated with an increased risk of HIV acquisition[131;132]. In 2016, worldwide the prevalence of HSV-2 was 13%; the highest prevalence is in Africa with around 44% of females and 25% of males living with HSV-2[133]. HSV-2 has ancient zoonotic origins, whereby proto-humans in Africa were infected by an ape simplexvirus[27;134;135].

Evolutionarily, HSV-2 is comprised of two major phylogenetic lineages separated by a long internal branch. One of these lineages is predominantly found in African populations (which we refer to here as lineage A), and the other lineage which includes members sampled across the world (which we refer to here as lineage B)[136]. The relationship between these lineages suggests HSV-2 diversified in the human population in Africa before dispersing with modern humans out-of-Africa[136], sometime between 50 and 100 kya[137;138]. Alternatively, it has been hypothesized that HSV-2 left Africa far more recently, sometime between 1671 and 1792 CE, via the trans-Atlantic slave trade[139] which was active between 1501–1867 CE[140]. Phylogeographic analysis of other viruses suggests that the slave trade may have been a conduit for viral introduction into the Americas[141].

The inability to distinguish between these two competing hypotheses operating on vastly different time-scales arises due to a lack of strong temporal signal in the HSV-2 molecular clock when calibrated using viral sampling dates (i.e. tip dating calibration;[142]). There is no detectable relationship between the year of HSV-2 sampling and the distance from the root of the phylogeny[139;143]. The human migration out-of-Africa hypothesis for HSV-2 dispersal arises from molecular clock estimates calibrated based on inference of codivergence events deeper in the simplexvirus tree that suggest a genome-wide clock rate on the order of $10^{-8}$ substitutions/site/year[27;134;136;144]. In contrast, the trans-Atlantic slave trade hypothesis is based upon a power-law rate decay of the molecular clock, whereby more recent branches have a faster clock rate than older branches[145]. The absence of a molecular clock calibrated exclusively within HSV-2 leaves these competing hypotheses unresolved.

Importantly, each of these origin hypotheses are associated with different geographic sources within Africa. If HSV-2 accompanied the earliest human migration out-of-Africa, then the most recent common ancestor (MRCA) of the worldwide diversity in lineage B would emanate from East, Central, or Southern Africa, where the first modern human migration began[138].

However, if HSV-2 left Africa by the trans-Atlantic slave trade, then this lineage would originate from West or Central Africa, where most Africans enslaved via the trans-Atlantic slave trade originated. Currently, most available African HSV-2 genomes are from East Africa[136;139;146], limiting our ability to formally test these hypotheses.

Here, we clarify the location of the emergence of HSV-2 out-of-Africa by sequencing 65 HSV-2 genomes from 18 previously unsampled countries, including a majority of viral samples from West and Central Africa. We further narrow down the timing of the worldwide dispersal through simulation and empirical molecular clock analysis to determine which origin scenarios are compatible with a lack of clock-like signal in HSV-2. We show that HSV-2 lineage B followed human migration from East Africa, but likely postdated the initial out-of-Africa human migration event.

## 4.2 Results

### 4.2.1 HSV-2 genomes.

We sequenced the enriched DNA of 65 new HSV-2 isolates, including 47 with a plausible link to 13 African countries, permitting us to determine sequence across a mean of 77% of the HSV-2 genome (range: 45-91%). These 65 newly sequenced African HSV-2 genomes were combined with previously published HSV-2 genomes resulting in an alignment of 395 taxa, sampled over 47 years from 37 countries (Supplementary Data 1; Figure 4.1A) . The total alignment length was 105,520 nt and recombinant fragments were masked.

### 4.2.2 Maximum likelihood phylogenetic inference.

We inferred a maximum likelihood phylogeny using IQTree2[81], which supported the existence of two distinct HSV-2 lineages (Figure 4.1B). These lineages are separated by a relatively long internal branch with length of 0.009 substitutions/site. The nodes defining lineage A and lineage B both had strong topological support (aBayes = 1.0; Supplementary data 2). Both midpoint rooting and outgroup rooting using the chimpanzee herpesvirus (ChHV) supports rooting the HSV-2 tree on this long branch, consistent with previous studies[136;139]. Lineage A comprised sequences from West Africa (n=9) and Central Africa (n=4).

Lineage B contains 382 sequences from all sampled geographic regions, including from West and Central Africa. The basal sequences of the this lineage are primarily East African. The basal structure of lineage B has strong phylogenetic support (aBayes 0.95-1.00 for the first 8 nodes; Figure 4.1). The crown of lineage B, representing the worldwide diversity, has little observable geographic structure (Figure 4.1B) due to the oversampling of genomes from North America and limited sampling of other regions[139;147;148].

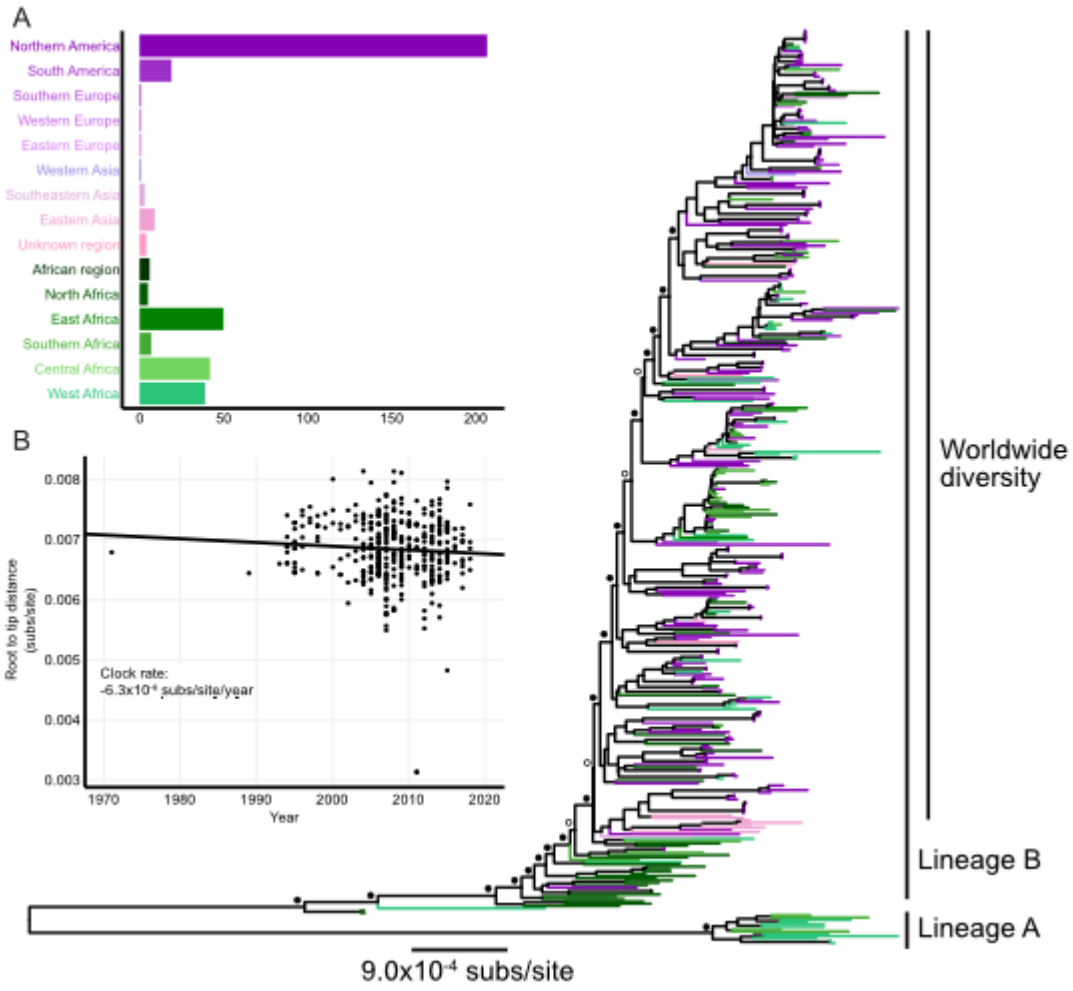



**Figure 4.1. Maximum likelihood phylogenetic analysis of empirical HSV-2 genomes.** Co-expression of Zic and ETS is shown in purple and occurs in the notochord, a6.5 lineage, which gives rise to the anterior sensory vesicle and palps, and four mesenchyme cells shown in light purple. A schematic of the tailbud embryo shows the notochord and a6.5 cell types later in development. Dark coloring represents a6.5 and notochord lineages, and light coloring represents other tissues with expression of Zic and/or ETS.

### 4.2.3 Clock rate inference of empirical sequences.

In an attempt to estimate the timing of divergence events within HSV-2, we performed molecular clock analysis using TreeTime[127]. However, despite the inclusion of HSV-2 genomes sampled over a 47-year timespan, we inferred a negative clock rate (Figure 4.1C) and a coefficient of determination ($R^2$) of the regression between year of sampling and root-to-tip distance of 0.00. These findings indicate a lack of clock-like signal, consistent with previous investigations of the HSV-2 molecular clock using a tip-dating approach[139;143]. We first explored the possibility that this lack of signal was due either to the phylogenetically recessed branches basal to the worldwide diversity in lineage B or the long branch separating the two lineages. However, after excluding the recessed basal taxa, we still failed to detect clock-like signal. Similarly, we failed to find of clock-like signal including only the taxa associated with worldwide diversity within lineage B (Supplementary Fig. D.1) or only lineage A. These sensitivity analyses all produced coefficient of determination for the molecular clock of $R^2 < 0.1$.

### 4.2.4 Likelihood of empirical clock rates.

To understand the plausible range of evolutionary rates that would be compatible with the lack of clock-like signal in the empirical HSV-2 data, we explored the shape of likelihood surface across a series of fixed substitution rates ranging between $10^{-9}$ and $10^{-3}$ substitutions/site/year to determine which molecular clock rates applied to the input ML tree have the highest likelihood and what rates are substantially worse. Given that the maximum likelihood rate estimate is negative, we expected that the slowest substitution rates (i.e., closest to a negative rate) should have the best likelihood. We restricted this analysis to rates between $10^{-9}$ and $10^{-3}$ substitutions/site/year, as this range encompasses previously estimated rates for a variety of viruses. As expected, the slowest rates have the best likelihood scores and the fastest rates have the worst (Figure 4.2). Clock rates from $10^{-9}$ through $4.2 \times 10^{-7}$ substitutions/site/year are within the 10 log-likelihood points of the best likelihood. Substitution rates faster than $4.2 \times 10^{-7}$ substitutions/site/year are a substantially worse fit to the observed data.

We note when we performed the same procedure using a virus with strong molecular clock signal, Ebolavirus from the 2014-2015 West African epidemic[149], the peak of this log-likelihood

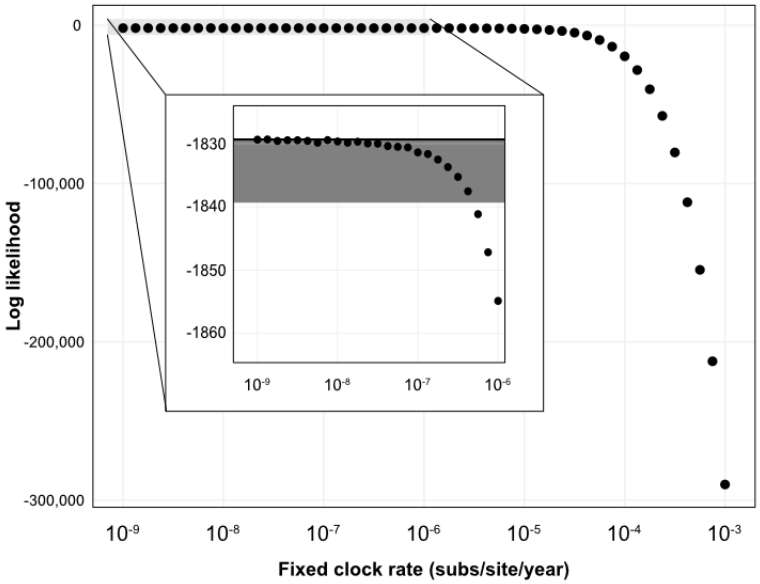



**Figure 4.2. Shape of likelihood surface for HSV-2 clock rate on empirical.** Log-likelihood of HSV-2 chronogram with a range of fixed clock rates estimated using the maximum likelihood topology. The black line is best estimated log likelihood. The grey shaded box is the area shown in insert. Within the insert, the black line is maximum estimated likelihood at rate $1x10^{-9}$ substitutions/site/year, and the dark grey box highlights the range 10 points below the maximum estimated log likelihood and encompasses the rates $4.2x10^{-7}$ substitutions/site/year and slower.

curve reasonably approximated the inferred maximum likelihood rate (Supplementary Fig. D.2).

### 4.2.5 Estimating clock rate of simulated data.

To explore at what substitution rates temporal signal should theoretically be detectable in our HSV-2 dataset, given empirical sampling dates and phylogenetic structure, we simulated HSV-2 genome sequences across the maximum likelihood tree under a series of substitution rates, ranging between $10^{-9}$ and $10^{-3}$ substitutions/site/year. We then inferred maximum likelihood tree branch lengths from these simulated data and re-inferred the clock rate. A simulated rate that had detectable temporal signal using this approach would be inconsistent with the empirical data.

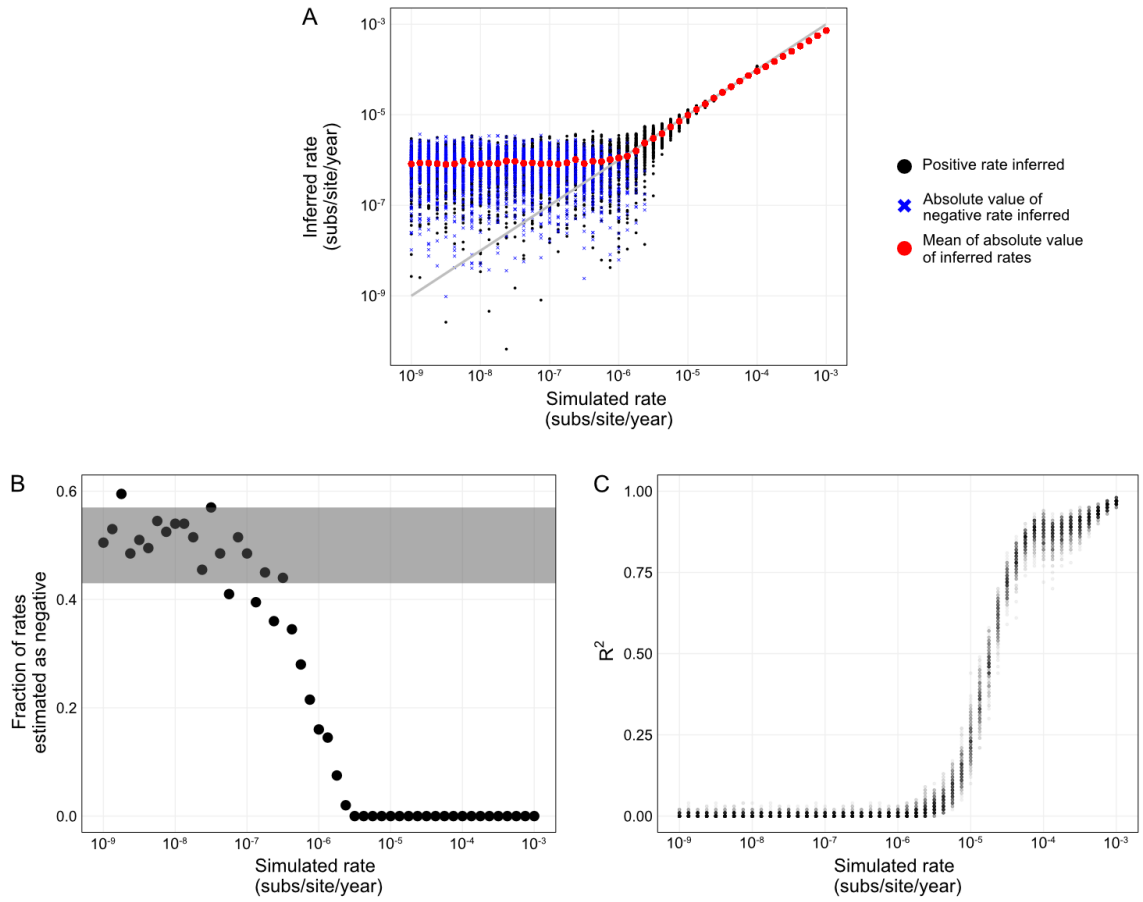



**Figure 4.3. Performance of clock rate inference on simulated data.** (A) Inference of clock rate for a fixed input tree with simulated sequences and empirical dates. Black dot represents estimates of positive rates, a blue x represents the absolute value of estimates of negative rates, and red dots indicate the mean of absolute value for each fixed rate. Grey line indicates parity of simulated and estimated rate. (B) Negative clock rate analysis of the fraction of total simulations for a fixed rate that have a rate estimated as negative. Grey shading represents range for which a negative estimate is equally probable as positive (two-sided binomial test $p<0.05$). (C) $R^2$ coefficient of determination of root-to-tip regression, a transparent dot represents one replicate. 200 replicates of simulated sequences were generated under GTR+F+Γ4 model then clock rate estimated on simulated sequences.

We were able to reliably infer clock rates from sequences simulated at $1.0\text{x}10^{-6}$ substitutions/site/year to $10^{-4}$ substitutions/site/year (Figure 4.3A), indicating the true underlying HSV-2 substitution rate must be slower than these rates. We note that rates faster than $10^{-4}$ substitutions/site/year were underestimated as the age of the root approached the earliest sample date. Based on a breakpoint analysis, for simulated rates slower than $1.0\text{x}10^{-6}$ substi-

tutions/site/year, the mean of the absolute value of inferred clock rates did not capture the simulated rate, unlike faster rates (segmented regression; Supplementary Fig. D.3), suggesting these slower rates are consistent with the empirical data.

As in the empirical analysis, the estimated clock rates simulated under slower rates are often negative (Figure 4.3B). Thus, sequences simulated under a clock rate which results in an estimated negative rate indicates that the true data could have been produced under that rate. The fastest rate at which more than 5% of trials infer a negative rate on the simulated data is $1.8\text{x}10^{-6}$ substitutions/site/year (Figure 4.3B). The fastest rate at which it is equally probable for a negative rate to be inferred as positive rate is $3.2\text{x}10^{-7}$ substitutions/site/year (binomial test; $p<0.05$).

The $R^2$ of the root-to-tip by sample date regression decreases with slower simulated clock rates (Figure 4.3C), because as the clock rate (slope) approaches zero less of the divergence is explained by sampling date. More than half of the $R^2$ estimates were zero in simulations with a rate of $1.8\text{x}10^{-6}$ substitutions/site/year or slower (Figure 4.3C). The simulations of slower rates are consistent with the root-to-tip regression of the empirical HSV-2 dataset, which had an $R^2$ of zero.

### 4.2.6 Timing the ancestor of HSV-2.

The range of substitution rates we explored is broad, accounting for scenarios in which the time of most recent common ancestor (tMRCA), or root age, for HSV-2 is older than modern humans, 6 million years ago at $1.0\text{x}10^{-9}$ substitutions/site/year, to the similarly unrealistic 49 years ago at $1.0\text{x}10^{-3}$ substitutions/site/year ( 4.1, Supplementary Data 3). The youngest tMRCA not rejected by any of our clock-rate simulation was 19 kya, assuming $3.2\text{x}10^{-7}$ substitutions/site/year (Fig. 4B). Using the slowest substitution rate that could be detected in our simulations, $2.4\text{x}10^{-6}$ substitutions/site/year, we infer root age of 2.6 kya (556 BCE). Therefore, the HSV-2 tMRCA must be older than this date in order to be consistent with the lack of clock-like signal in the empirical data (Supplementary Fig. D.4). HSV-2 tMRCAs between 109-146 kya, which are consistent with a previously characterized East-West split in anatomically modern humans in Africa around 110-160 kya[138], were obtained assuming a substitution rate between $4.2\text{x}10^{-8}$

and $5.6x10^{-8}$ substitutions/site/year, respectively (Fig 4A). These tMRCAs are also consistent with previously published HSV-2 root age 90-120 kya, based on externally calibrated molecular clocks[27;136].

**Table 4.1.** Age of root and out-of-Africa dispersal across clock rates.

| Substitution rate[a] | Root age[b] | Out-of-Africa age[b] | Support for East African source[c] | Molecular clock analyses[d] |
|---|---|---|---|---|
| $1.0x10^{-9}$ | 6.1 mya | 1.2 mya | 0.977 | Not rejected |
| $1.3x10^{-9}$ | 4.6 mya | 922kya | 0.977 | Not rejected |
| $1.8x10^{-9}$ | 3.4 mya | 693kya | 0.977 | Not rejected |
| $2.4x10^{-9}$ | 2.6 mya | 519kya | 0.977 | Not rejected |
| $3.2x10^{-9}$ | 1.9 mya | 387kya | 0.977 | Not rejected |
| $4.2x10^{-9}$ | 1.4 mya | 290kya | 0.977 | Not rejected |
| $5.6x10^{-9}$ | 1.1 mya | 219kya | 0.977 | Not rejected |
| $7.5x10^{-9}$ | 809kya | 163kya | 0.977 | Not rejected |
| $1.0x10^{-8}$ | 616kya | 124kya | 0.976 | Not rejected |
| $1.3x10^{-8}$ | 455kya | 93kya | 0.977 | Not rejected |
| $1.8x10^{-8}$ | 341kya | 70kya | 0.976 | Not rejected |
| $2.4x10^{-8}$ | 256kya | 52kya | 0.977 | Not rejected |
| $3.2x10^{-8}$ | 192kya | 39kya | 0.977 | Not rejected |
| $4.2x10^{-8}$ | 146kya | 29kya | 0.977 | Not rejected |
| $5.6x10^{-8}$ | 109kya | 22kya | 0.977 | Not rejected |
| $7.5x10^{-8}$ | 81kya | 16kya | 0.977 | Not rejected |
| $1.0x10^{-7}$ | 61kya | 12kya | 0.977 | Not rejected |
| $1.3x10^{-7}$ | 45kya | 9.3kya | 0.977 | Not rejected |
| $1.8x10^{-7}$ | 34kya | 6.9kya | 0.977 | Not rejected |
| $2.4x10^{-7}$ | 26kya | 5.2kya | 0.978 | Not rejected |
| $3.2x10^{-7}$ | 19kya | 3.9kya | 0.977 | Not rejected |
| $4.2x10^{-7}$ | 14kya | 2.9kya | 0.977 | Rejected by some |
| $5.6x10^{-7}$ | 11kya | 2.2kya | 0.976 | Rejected by some |
| $7.5x10^{-7}$ | 8.1kya | 1.7kya | 0.977 | Rejected by some |
| $1.0x10^{-6}$ | 6.1kya | 1.2kya | 0.976 | Rejected by some |
| $1.3x10^{-6}$ | 4.6kya | 935 ya | 0.976 | Rejected by some |
| $1.8x10^{-6}$ | 3.4kya | 702 ya | 0.976 | Rejected by some |
| $2.4x10^{-6}$ | 2.6kya | 528 ya | 0.976 | Rejected by all |
| $3.2x10^{-6}$ | 1.9kya | 401 ya | 0.975 | Rejected by all |
| $4.2x10^{-6}$ | 1.4kya | 303 ya | 0.976 | Rejected by all |
| $5.6x10^{-6}$ | 1.1kya | 230 ya | 0.977 | Rejected by all |
| $7.5x10^{-6}$ | 826 ya | 175 ya | 0.977 | Rejected by all |
| $1.0x10^{-5}$ | 616 ya | 133 ya | 0.980 | Rejected by all |
| $1.3x10^{-5}$ | 463 ya | 104 ya | 0.982 | Rejected by all |
| $1.8x10^{-5}$ | 353 ya | 80 ya | 0.985 | Rejected by all |
| $2.4x10^{-5}$ | 267 ya | 64 ya | 0.989 | Rejected by all |
| $3.2x10^{-5}$ | 205 ya | 55 ya | 0.990 | Rejected by all |
| $4.2x10^{-5}$ | 159 ya | 52 ya | 0.980 | Rejected by all |
| $5.6x10^{-5}$ | 125 ya | 50 ya | 0.944 | Rejected by all |
| $7.5x10^{-5}$ | 101 ya | 49 ya | 0.827 | Rejected by all |
| $1.0x10^{-4}$ | 84 ya | 49 ya | 0.540 | Rejected by all |
| $1.3x10^{-4}$ | 72 ya | 48 ya | 0.274 | Rejected by all |
| $1.8x10^{-4}$ | 64 ya | 48 ya | 0.129 | Rejected by all |
| $2.4x10^{-4}$ | 59 ya | 48 ya | 0.062 | Rejected by all |
| $3.2x10^{-4}$ | 55 ya | 48 ya | 0.039 | Rejected by all |
| $4.2x10^{-4}$ | 53 ya | 47 ya | 0.026 | Rejected by all |
| $5.6x10^{-4}$ | 51 ya | 47 ya | 0.020 | Rejected by all |
| $7.5x10^{-4}$ | 50 ya | 47 ya | 0.013 | Rejected by all |
| $1.0x10^{-3}$ | 49 ya | 47 ya | 0.010 | Rejected by all |

[a] Substitutions/site/year [b] Prior to most recently sampled virus, 2018 [c] Maximum inferred probability for East African ancestor prior to out-of-Africa dispersal [d] Empirical and simulation analyses described in Figure 4.5

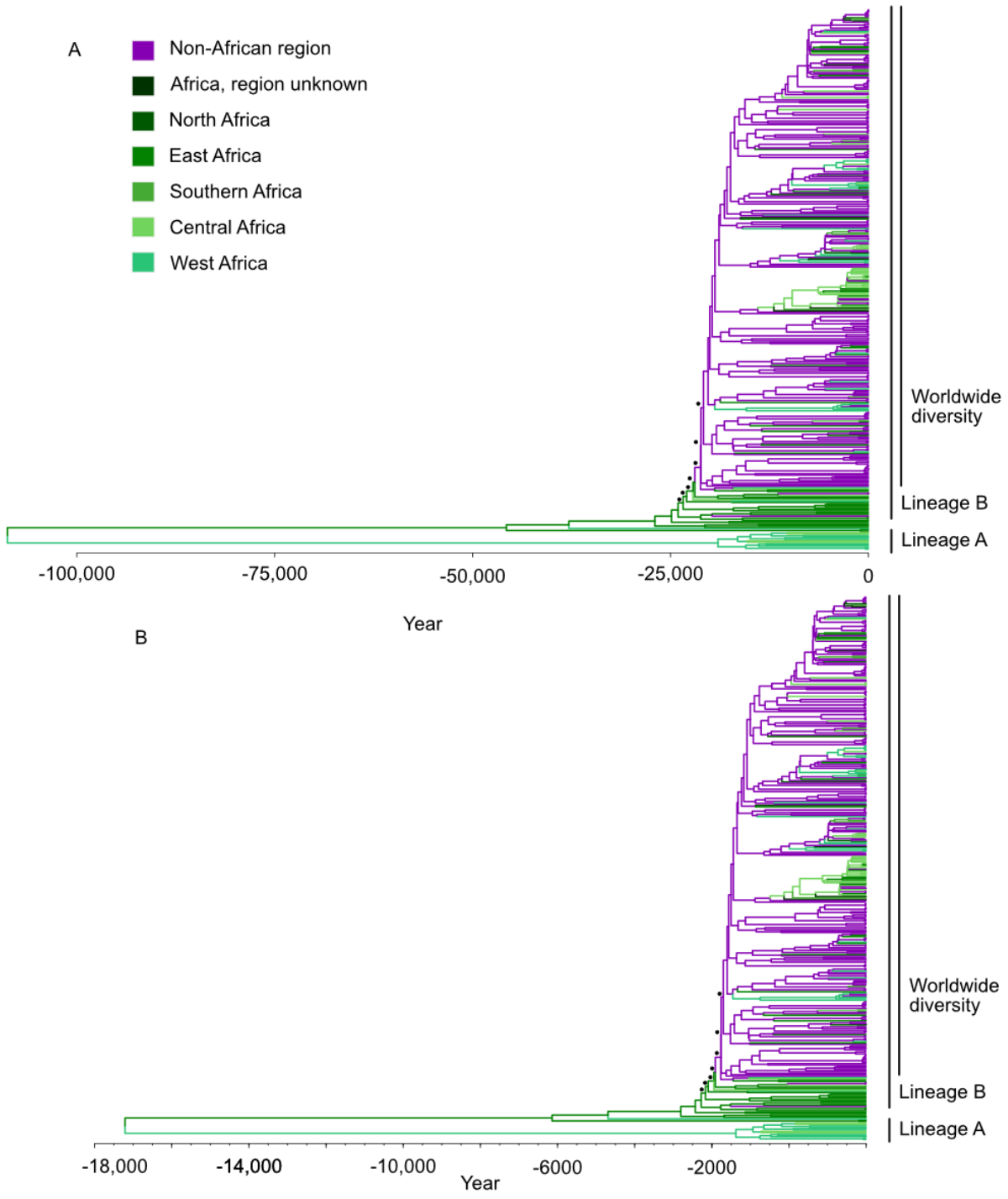



**Figure 4.4. Phylogeography of HSV-2 supports East African origin of out-of-Africa dispersal.** (A) Time tree inferred using the evolutionary rate of 5.6x10-8 substitutions/site/year. (B) Time tree inferred with the fastest rate supported by all clock analyses: 3.2x10-7 substitutions/site/year. Maximum likelihood tree topology scaled to time tree, colored by plurality of support in mugration analysis for each region. Time scale represents the calendar year; negative years are BCE. Black dots indicate nodes used in out-of-Africa weighted age calculation (see Supplementary Fig. D.1).

### 4.2.7 Phylogeographic analysis indicates East African origin of worldwide lineage.

We performed phylogeographic reconstruction in TreeTime on all empirical time trees inferred with fixed clock rates between $10^{-9}$ and $10^{-3}$ substitutions/site/year to determine the ancestral location of HSV-2 and the source of out-of-Africa migration. Using the substitution rates with HSV-2 tMRCA that is consistent with human divergence and published HSV-2 ages ($4.2\text{x}10^{-8}$ and $5.6\text{x}10^{-8}$ substitutions/site/year), we estimate that the ancestor of lineage A was in West Africa and that the out-of-Africa dispersal event within lineage B occurred in East Africa between 29.3 to 21.9 kya (Figure 4.4; Table 1).

The phylogeographic placement of East Africa as the source of the HSV-2 out-of-Africa migration is robust across substitution rates (Table 1). Support for East Africa as the ancestral location of lineage B increases with slower substitution rates, exceeding 0.95 for all evolutionary rates that were not rejected by the molecular clock analyses (i.e., $< 2.4\text{x}10^{-6}$ substitutions/site/year). Even the youngest possible estimate for the timing of HSV-2 out-of-Africa migration that is consistent the lack of clock-like signal, is 3.9 kya (1883 BCE) and supports East Africa as the source of the out-of-Africa migration (Table 4.1). The out-of-Africa age estimate that is consistent our most permissive rejection strategy was 702 years ago (1316 CE), still predating the trans-Atlantic slave trade by 200 years. Nonetheless, even at this younger age, the out-of-Africa dispersal was still inferred to be in East Africa. Furthermore, the inference of East African origin for the worldwide diversity is robust to geographic partitioning schemes of African and non-African regions (Sup Table D.1).

## 4.3 Discussion

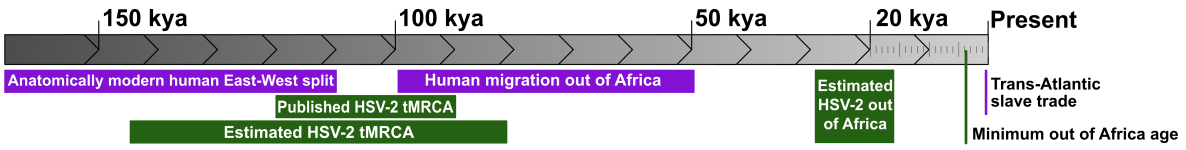



**Figure 4.5. Timeline of inferred HSV-2 diversification events with context of human activity.** Purple segments denote events associated with human activity. Green segments denote viral evolutionary events. Timing of deepest anatomically modern human split inferred to be 110-160 kya[138]. Timing of first human migrations out of Africa estimated to be 50-100 kya[138]. Timing of trans-Atlantic slave trade ca. 1501–1867 CE[150]. Previously published age of HSV-2 tMRCA estimated to be 90-120 kya[27;136]. Age of HSV-2 tMRCA at fixed rates $4.2\text{x}10^{-8}$ – $7.5\text{x}10^{-8}$ substitutions/site/year. Age of extant out-of-Africa dispersal estimated with fixed rates $4.2\text{x}10^{-8}$ – $7.5\text{x}10^{-8}$ substitutions/site/year, is 16-29kya. Most recent age of out-of-Africa dispersal estimated under clock rate that is consistent with molecular clock analysis is 3.9kya ($3.2\text{x}10^{-7}$ substitutions/site/year).

Understanding the geographic spread of human pathogens, both ancient and recent, provides insight into what conditions lead to viral establishment or extinction, informing the potential dynamics of future epidemics[151–153]. To understand the geography and timing of HSV-2 dispersal, we expanded the geographic breadth of HSV-2 genomic sampling and established the sensitivity of tip dated molecular clock inference. Clock rates between 4.2x10-8-5.6x10-8 substitutions/site/year, are consistent with our molecular clock analysis of empirical data and, under these rates we infer an HSV-2 tMRCA of 110-146 kya which is consistent with the previously published timing between 90-120 kya[27;136]. Based on these rates, we conclude that HSV-2 most likely accompanied humans migrating from East Africa and dispersed worldwide between 22 and 29 kya (Figure 4.5). The tMRCA of modern worldwide diversity, between 22 and 29 kya, is consistent with an HSV-2 strain dispersing after the population bottleneck associated with the Last Glacial Maximum[154]. This strain could have been the first HSV-2 to disperse worldwide, or it could have replaced HSV-2 that accompanied humans migrations and later went extinct (as has been reported in HSV-1)[155].

HSV-2 has ancient zoonotic origins, whereby proto-humans in Africa were infected by an ape simplexvirus[156–158]. While the tMRCA of HSV-2, the split of lineage A and B,

may represent the age of HSV-2 in humans this divergence may have occurred sometime after establishing in humans. The divergence within HSV-2 may reflect human migration within Africa, however as HSV-2 likely infected humans the tMRCA almost certainly postdates the origin of modern humans. Therefore it is unlikely that the true substitution rate is slower than $2.4 \times 10^{-8}$ substitutions/site/year, which produce a of 250 kya which is older than evidence of anatomically modern humans[138]. Thus HSV-2 migrating with modern humans during the original out-of-Africa migration between 50 and 100 kya[138], is not consistent with the age of HSV-2 worldwide diversity.

We conclude that a scenario in which HSV-2 left Africa via the trans-Atlantic slave trade between 1500–1850 CE is inconsistent with both molecular clock and phylogeographic inference. Out-of-Africa migrations events more recently than ca. 1300 CE would necessitate a viral substitution rate that would produce some clock-like signal using tip-date calibrated molecular clock (i.e., faster than $1.8 \times 10^{-6}$ substitutions/site/year). Our clock-signal breakpoint analysis (Supplementary Fig. D.3) indicates that an out-of-Africa event post-dating 1083 CE (i.e., substitution rate $1.3 \times 10^{-6}$ substitutions/site/year) would necessitate a molecular clock that could be inferred from the dataset. Further, any internally consistent molecular clock rate places the origin of this out-of-Africa migration in East Africa, on the opposite side of the continent from where the majority of Africans enslaved as part of the trans-Atlantic slave trade originated. We acknowledge that it is possible that HSV-2 was subsequently imported into the Americas via the trans-Atlantic slave trade long after the worldwide dispersal, given that there is evidence of other viral pathogens being transported to the Americas from Africa during this time[141]. Nonetheless, we find that both the timing and geographical source of the HSV-2 worldwide dispersal are incompatible with the slave trade hypothesis.

A possible explanation for the previous inference of a young age of the HSV-2 out-of-Africa dispersal[139] is the misapplication of substitution rate correction for purifying selection. As viruses evolve for long periods of time, purifying selection can lead to underestimation of branch lengths due to substitution saturation[64;159]. In RNA viruses, this effect is most prominent on internal branches >0.1 substitutions/site[134;159], which is an order of magnitude longer than the total tree height of HSV-2 (0.0083 substitutions/site). Bayesian approaches have been developed

to explore deeper divergence events in virus phylogenies[134;160]. Additionally, rescaling branch lengths by a power law decay rate can be used to correct for the underestimation[145]. We note that a refinement of the formulation for the power law decay rate model[161] now describes rate variation through time with a sigmoid curve, acknowledging a more biologically plausible constant evolutionary rate near the tips of viral evolutionary trees with rate decay occurring only deeper in the tree. In other words, although rate decay is almost certainly relevant to deeper divergence events within the simplexvirus phylogeny that show evidence of saturation[134], HSV-2 is young enough that traditional maximum likelihood phylogenetic inference would not underestimate branch lengths, and applying a power law decay rate would inappropriately underestimate the evolutionary time along recent branches.

Previous analysis of HSV-2 geographic structure has primarily partitioned samples into continent based regions[139;147;148]. Using our novel genome sequences, we were able in increase the resolution of geographic analysis, and show that East Africa is the geographic source of HSV-2 worldwide diversity. However, we note the exact location and path the HSV-2 took upon leaving Africa is still unclear because of (i) the lack samples from the Horn of Africa which is the likely source of human out-of-Africa migration, (ii) the sampling bias for viral genomes from North America, and (iii) human migration in the modern era which obscures ancient signal. Increased sampling across global regions could improve geographic signal of HSV-2 migration.

HSV-2 genome samples on the order of thousands of years should provide sufficient temporal signal to estimate the true clock rate of HSV-2. In our simulations, genomes sampled over a 47-year timespan can accurately calibrate a clock rate of $1.3 \times 10^{-6}$ substitutions/site/year and faster. Hence HSV-2 genomes thousands of years old should have sufficient accumulation of signal to acutely infer a clock rate based on tip dates with a rate of $10^{-8}$ substitutions/site/year. The identification of ancient herpes virus sequences (including HSV-1, as in[155]) in mummified archeological specimens[162] suggests the possibility of reconstructing ancient HSV-2 genomes that could be used to inform a tip-dated calibration for this virus and to reveal the ancient paths of HSV-2 transmission.

## 4.4 Methods

### 4.4.1 HSV-2 data set and sequence alignment.

Clinical isolates were recovered from patients diagnosed with genital herpes at La Pitié Salpêtrière – Charles Foix University Hospital (Paris, France) who self-reported overseas origin. HSV-2 isolates were used because of shortages of primary clinical sample material. Isolates were obtained by propagation in subconfluent Vero cell monolayers[163]. A limited number of passages was necessary for the generation of viral stocks.

We sequenced 65 HSV-2 samples collected 2012-2018 primarily from West and Central Africa. In brief, we extracted DNA from supernatant using the QIAmp Viral RNA Mini Kit (Qiagen, Hilden, Germany) and measured DNA content with a Qubit (Thermo Fischer Scientific, Waltham, MA, USA). HSV-2 copy numbers were determined using a quantitative PCR (qPCR) assay[136]. We then prepared dual indexed-libraries using 1ng DNA extract using the Nextera XT Library Preparation Kit and the Nextera XT Index Kit (Illumina, San Diego, CA, USA), which we pooled so individual libraries would all bring in about the same number of HSV-2 genome copies (according to the qPCR results). The pool was subjected to a double round hybridization capture with a MYbaits Custom Target Enrichment Kit designed to cover the whole genomes of HSV-1 (NC_001806), HSV-2 (NC_001798) and varicella zoster virus (VZV, NC_001348; Human herpesvirus 3) (Daicel Arbor Biosciences, Ann Arbor, MI, USA). For both rounds, we followed manufacturer's instructions with the exception that only one fourth of the recommended bait quantity was used. We sequenced the capture product on MiSeq and NextSeq 550 platforms (Illumina) using the MiSeq Reagent Kit v3 (600 cycles; Illumina) and NextSeq Mid Output Kit v2.5 (300 cycles; Illumina), generating a total of 61,926,628 reads.

Low quality bases and adapter sequences were removed from raw reads using Trimmomatic v0.36 (35) and overlapping reads were merged using ClipAndMerge v1.7.8[164]. We then mapped trimmed reads to the HSV-2 RefSeq genome (NC_001798) from which we had removed the TRL and TRS regions[165]. We finally sorted mapping files and removed duplicate reads using the SortSam and MarkDuplicates tools from Picard v2.7.1 (http://broadinstitute.github.io/picard/). We generated consensus sequences form the maps using Geneious Prime v2021.2.2[166], calling

bases at positions covered by at least 20 unique reads and for which more than 95% of the reads were in agreement.

Additionally, we downloaded all available HSV-2 sequences of length greater than 100kbp from GenBank in November 2019. Potentially engineered (n=5) and dubious sequences (n=10) were removed. HSV-2 sequences along with, ChHV (NC_023677) and HSV-1 (NC_001806) were aligned with MAFFT v7.307[80]. Regions with more than 10% of sequences containing gaps were removed with Geneious. Invariant sites were stripped from the alignment to produce a dataset that was computationally tractable for recombination analysis, resulting in an alignment of 25,809 variant sites of ChHV, HSV-1, and HSV-2 sequences. Recombinant fragments were inferred with RDP4[167], using five recombination detection methods (RDP, GENECONV, MaxChi, BootScan and SiScan) and validating recombination events identified by at least two methods. In sequences found to contain recombination, the recombinant fragments within these sequences were masked (with Ns) in the complete alignment (including variant and invariant sites) for subsequent phylogenetic analysis.

### 4.4.2 Phylogenetic and initial temporal analysis.

Subsequent analysis was done on only HSV-2 sequences with the year of sample collection resulting in a final HSV-2 dataset of 395 sequences, of 105,420 sites of which 5,648 sites had variation. Sequences were sampled globally across 47 years, from 1971 to 2018 (Supplementary Data 1).

A maximum likelihood phylogeny for the HSV-2 dataset was made using IQtree2 v2.0.4[81], with a GTR+F+R4 substitution model selected by model testing using ModelTesting with cAIC on the initial alignment[168]. To test uncertainty in the topology we used aBayes to estimate branch support[169]. We used TreeTime v0.7.6[127] to perform molecular clock inference on HSV-2 using tip date calibrations. Default TreeTime parameters were used, except for fixing the root, using input branch lengths, and clock filter was set to 0 (no filtering). The root was fixed, as this rooting is supported by including ChHV as an outgroup and is consistent with previous studies[136]. Input topology and branch lengths were used to because joint optimization is not stable without this approach[170]. With the lack of clock like signal it is was not appropriate to

filter based on clock variation.

### 4.4.3 Likelihood of fixed clock rates with empirical data.

The log likelihood was calculated for time trees with a range of fixed substitution rates using TreeTime. We used a fixed input tree, empirical sample dates, and no clock filter. The rates used were $1x10^{-9}$ substitutions/site/year to $1x10^{-3}$ substitutions/site/year, with a resolution of 8 rates distributed evenly across the log space across each order of magnitude. This distribution exceeds the range of previous clock rate estimates for simplexvirus and HSV-1 [139;144].

To validate the interpretation of the likelihood surface, we performed a similar process with a previously published dataset of 1,610 ebolavirus sequences from the 2014-2015 West African Ebola epidemic [149], which has detectable temporal signal using tip-date calibrations. The Ebola dataset was analyzed with rates from $1x10^{-1}$ substitutions/site/year to $1x10^{-5}$ substitutions/site/year, with a resolution of 3 rates for each order of magnitude spread evenly across log space (Supplementary Fig. D.2).

### 4.4.4 Clock rate estimates of simulated data.

In order to understand what clock rate would result in sufficient temporal signal to detect the molecular clock on the empirical HSV-2 dataset (i.e., range of sampling dates, sequence length, and tree topology), we simulated sequence data across a range of substitution rates, $10^{-9}$ to $10^{-3}$ substitutions/site/year with 8 rates across each order of magnitude. Simulation was done with the nucleotide substitution model parameters inferred by IQtree2, under a GTR+F+$\Gamma$4 model on the HSV-2 dataset. We first inferred a time tree with a fixed clock rate based on our topology using TreeTime and then scaled this tree to substitutions/site with the clock rate. Sequences were simulated on the rescaled tree with Seq-Gen v1.3.4 [171]. We performed 200 replicates of sequence evolution and subsequent analysis.

Branch lengths were estimated on the fixed empirical topology using the simulated sequences with IQtree2 using GTR+F+$\Gamma$4 model. Although the empirical tree inference was performed with free rate variation, we used gamma rate variation in the simulations and subsequent tree inference to maintain consistency between simulation, tree and clock inference. The clock rate and $R^2$ of the root-tip-regression of the simulated data was then estimated with

TreeTime, using previous parameters.

The detection of temporal signal was assessed though different methods using R[172]. The rate at which there was a change in the relationship between the mean absolute value of estimated rate and the simulated rate was inferred by a linear regression break point analysis (R segmented v. 1.2-0).

The point at which the estimated rate reflected the empirical data, in that the estimated clock rate was negative, was determined by two cutoffs. The first cutoff method is the fastest rate at which more than 5% of estimated rates on simulated data were negative. The second cutoff is the fastest rate at which a negative estimate is as likely as a positive estimate, determined by binomial test ($p<0.05$).

### 4.4.5 Phylogeographic ancestral state reconstruction.

We performed a mugration analysis, which is an inference of transition between discrete geographic regions along branches of the phylogenetic tree, to reconstruct the ancestral geographic state. We used TreeTime mugration with default parameters, empirical time trees built with different fixed clock rates and the region of sampling for each tip. Samples with country histories were partitioned into regions, based on the United Nations M49 Standard, into North Africa, West Africa, East Africa, Central Africa, Southern Africa, and countries from all other regions were grouped into non-African region. Six samples had no country of origin, but are known to be sampled from Africa, and were partitioned into a broad Africa, region unknown for phylogeography analysis. Additionally, 4 samples were from an unknown location and were indicated as missing in the mugration analysis.

Robustness analysis of the phylogeographic reconstruction was done by repeating previous mugration analysis with a merged West, Central, and Southern Africa region. Additional robustness analysis split the non-African region into worldwide subregions, and reconstruction was done with North Africa, West Africa, East Africa, Central Africa, Southern Africa, Africa region unknown, Eastern Europe, Southern Europe, Western Europe, Northern America, South America, Southeastern Asia, Western Asia, and missing.

The points of particular interest in the trees are first the age of the root, which represents

the age of HSV-2 diversification, and second the age of the most recent common ancestor (MRCA) of the worldwide diversity within the worldwide lineage, which represents the maximum age for the dispersal of HSV-2 out-of-Africa, assuming an African origin for the lineage B. On the fixed topology, ancestral geographic reconstruction along the basal backbone of the worldwide lineage had majority of support for the African region that is the source of HSV-2 worldwide diversity which decreases until the majority of support is for the non-African region (Supplementary Fig. D.5). We defined the-out of-Africa age as the mean age of the nodes that define the backbone at the base of lineage B, weighted by a factor of one minus the absolute value of the difference of the support for source African region and support for non-African region. This factor accounts for the uncertainty which point on the maximum likelihood phylogeny that HSV-2 migrated out-of-Africa.

Age is reported in number of years ago, relative to 2018, the most resent sample date.

### 4.4.6 Data Availability

All sequence data generated in this study are made available in BioProject under accession code PRJEB48656 (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB48656/). All accession numbers for data used in this manuscript can be found in Supplementary Data 1.

### 4.4.7 Code Availability

Code for running simulation is available at https://github.com/jennifer-bio/HSV2 .

### 4.4.8 Author Contributions Statement

Conceptualization: J.L.H., S.C.S., J.O.W. Data collection: S.B., D.B. Data generation: S.C.S., K.M. Methodology: J.L.H., S.C.S., K.M., J.O.W. Formal analysis: J.L.H., S.C.S., J.O.W. Writing - original draft preparation: J.L.H., J.O.W. Writing - review and editing: All Authors Supervision: S.C.S., J.O.W.

### 4.4.9 Competing Interests Statement

The authors declare no competing interests.

## 4.5 Acknowledgments

We thank Richard Neher for his guidance on reporting and interpreting likelihoods in TreeTime. This research was supported by National Institute of Health R01 (AI135992, J.O.W.).

Chapter 4, in full, is a reformatted reprint, permission pending, of the material as it appears in "Phylogeographic analysis reveals an ancient East African origin of human herpes simplexvirus 2 dispersal out-of-Africa." Jennifer L. Havens, Sébastien Calvignac-Spencer, Kevin Merkel, Sonia Burrel, David Boutolleau, and Joel O. Wertheim. Nature communications, 2022. The dissertation author was the primary investigator of this paper.

# Epilogue

In this dissertation I have presented multiple cases of human virus epidemics where virus traits and human behavior have impacted circulation. With an understanding of the importance of virus traits, I presented evidence that zoonotic viruses can adapt to their novel host once they have jumped into the human population and a specific investigation into how SARS-CoV-2 evolves.

In Chapter 1, I sought to understand how virus variants with different traits circulated in a heterogeneous population. I investigated NYC during the second wave of SARS-CoV-2 and found that Iota, an immune evasive variant, circulated more in areas with high rates of poverty and larger proportion of Black and Hispanic or Latino residents. In contrast I found that Alpha, a variant with high transmissibility, succeeds in areas with low rates of poverty and a larger proportion of White and Asian or Pacific Islander residents. I presented this dynamic in the context of the immune landscape. Areas where immune evasive variants succeeded had higher disease prevalence during the first wave thus higher rates of standing immunity during the second wave. While the same factors associated with socioeconomic status that caused heterogeneous disease burden during the first wave likely impacted transmission during the second wave, the resulting heterogeneous standing immunity was also an important factor. In this chapter, I present evidence that virus traits impacted their relative circulation and spread in a heterogeneous population.

I also contributed research, not expanded upon here, investigating the dynamics of SARS-CoV-2 transmission in Jordan[173] and the strength of the second wave of SARS-CoV-2 in NYC[44;47]. In Jordan, human behavior, specifically regional connectivity and land travel were drivers of transmission. In NYC, we found that variants, with increased transmissibility and immune escape properties, were important drivers in an extended second wave. Understanding

the traits of different variants, as well as human behavior, was important to understand how the virus persisted; as such I investigated how SARS-CoV-2 evolves in the human population.

In general, SARS-CoV-2 tends to accumulate little change between individual transmission events; however, interspersing this are jumps with a large amount of evolution all at once, which define stem lineages of variants. Variants have different traits from their ancestor, such as immune evasion and transmissibility, which appears to be acquired during these jumps in evolution. In Chapter 2, I explore the molecular clock, rates of synonymous and nonsynonymous substitutions, and dN/dS in the stem lineages of variants. I find that the rate of synonymous substitutions is slower in many variant stem lineages than during typical transmission. I found this observation of reduced synonymous substitutions to be particularly interesting to investigate because it involved looking for signal that was missing, rather than looking for an excess. This is a contrast to, for example, finding an excess of nonsynonymous substitutions or an increase in dN/dS, which has been repeatedly noted by many different groups[20;52].

When I examine persistent infections, identified in NYC SARS-CoV-2 surveillance data, there are some cases which show the same pattern of reduced rate of synonymous substitutions and increase in dN/dS. This consistency supports the hypothesis that persistent infections are the source of variants. Variation in the synonymous molecular clock, which is less impacted by selection than the nonsynonymous molecular clock, also supports the hypothesis that persistent infections experience an environment which is not typically seen during acute infection. One possibility is that during persistent infection virus enters different body compartments, and cell tropism or other factors impact replication rate. Understanding how virus persists is vital to developing effective treatment, and potentially reducing the impact of future variants.

In Chapter 2, I present that SARS-CoV-2 variants evolve within the human population rather than in a zoonotic reservoir. While reverse-zoonosis and subsequent reintroduction into humans is possible as has been seen with deer and mink[70;174], it does not appear to be a significant factor in the evolution of SARS-CoV-2 circulating in the human population. In Chapter 3, I explore the context that adaptation to humans occurs in, in a variety of viruses, with a focus on recent zoonosis. I present the comparison of selection in a zoonotic host reservoir with selection in the stem lineage proceeding zoonosis and during the epidemic following zoonosis. I find that

selection on the stem lineage of recent zoonoses are consistent with selection in the host reservoir. Again, I found this finding intriguing to investigate because, rather than focusing on a notable difference, I focus on results that do not show a change, a negative result.

In contrast, the stem lineage leading the reemergance of H1N1 in 1977 is not characteristic of transmission in either humans or a non-human animal reservoir. Rather, selection and molecular clocks are consistent with freezing and passage in a laboratory context. This is an outlier to other recent pandemics that are of zoonotic origin.

In contrast to the lack of change in selection preceding zoonosis relative to the host reservoir, after zoonosis during early human outbreaks the selection regime is different than in the host reservoir. This change is consistent with the known pattern of adaptation after host switch[15;91;93;94]. In this dissertation I present evidence which is consistent with the scenario that some viruses in zoonotic reservoirs are sufficiently fit to transmit in the human population and when a fit virus has the opportunity to infect humans it can make the jump. Once in the human population, a circulating virus can evolve, which impacts it's spread, giving the opportunity to evolve more.

I have examined viruses which are sufficiently fit to circulate in the human population; however, this is not the only interaction humans have with viruses that we can explore. Based on seropositivity data, humans are routinely exposed to zoonotic viruses, including viruses from the Coronaviridae and Filoviridae families[118;122;175], which do not go on to circulate and transmit extensively in the human population. Genomic and epidemiological data on viruses that infect individuals without onward transmission could provide valuable information to understand the requirements for successful zoonosis (zoonotic emergence and subsequent circulation in human population). However, this data is limited, because it is difficult to capture such isolated events. Understanding the relative importance of viral traits and human behavior to successful virus circulation is valuable in determining strategies to prevent and respond to zoonosis.

In Chapter 4, I examined the circulation of HSV-2, which, in contrast to the recent zoonotic viruses I examined earlier in my dissertation, is one of the earliest extant zoonotic pathogens. The limited number of samples from different geographic regions limits the granularity of the phylogeographic reconstruction, so we focused on continent level reconstruction. This

is in stark contrast with the scope of data available for SARS-CoV-2, which can at times be investigated at the level of neighborhoods (Chapter 1). However, even with this limitation, we see the same pattern: human behavior has a fundamental impact on the dynamics and spread of a zoonotic virus. With more information it may be possible to investigate the impact of virus traits on that spread. With respect to timing, we find that HSV-2 likely spread with the global migration of humans. However, given the limited time range of collected samples there is insufficient information to infer clock-like behavior and so there is high uncertainty in the molecular clock inference timing. To address the absence of clock-like behavior we developed a robust simulation framework to characterizes this lack.

As with the lack of change proceeding zoonosis and lack of synonymous substitutions on SARS-CoV-2 variant stem lineages, the lack of clock-like signal in HSV-2 could have been considered a negative result that is not worth proceeding with. However, in this dissertation I have presented multiple situations where the lack of observable information is, in and of itself, information. Using careful contextualization and comparison, null results and missing information can be fundamentally important to developing novel insights that may otherwise be missed.

The work presented in this dissertation has provided novel insight into the dynamics of evolution and circulation of viruses. Specifically, I present the spread and circulation of HSV-2 and SARS-CoV-2, and how human behavior and virus traits impact this dynamic. To understand the context in which virus traits evolve, I present how selection dynamics change in different environments between hosts and contrasting acute and persistent infections. Understanding the dynamics of evolution and spread is important in developing an effective response to novel zoonotic viruses, as well as the challenges we face from zoonotic viruses which are circulating endemically.

# Appendix A

# Supplemental Material for Chapter 1

## A.1 Supplementary Table

**Table A.1.** GLM model coefficients

| Variant | Coefficient (Seroprevalence by Borough) | 95%HPD | Coefficient (Vaccination) | 95%HPD |
|---|---|---|---|---|
| Iota | 0.95 | 0.003-2.5 | 0.985 | 0.001 - 2.95 |
| Iota_E484K | 0.89 | 0.0006-2.8 | 0.966 | 0.001 - 2.87 |
| Alpha | 0.8 | 0.00002-2.4 | 4.92 | 0.02-9.44 |
| B1637 | 0.99 | 0.0001-3.05 | 0.964 | 0.001 - 2.88 |

## A.2 Supplementary Figures

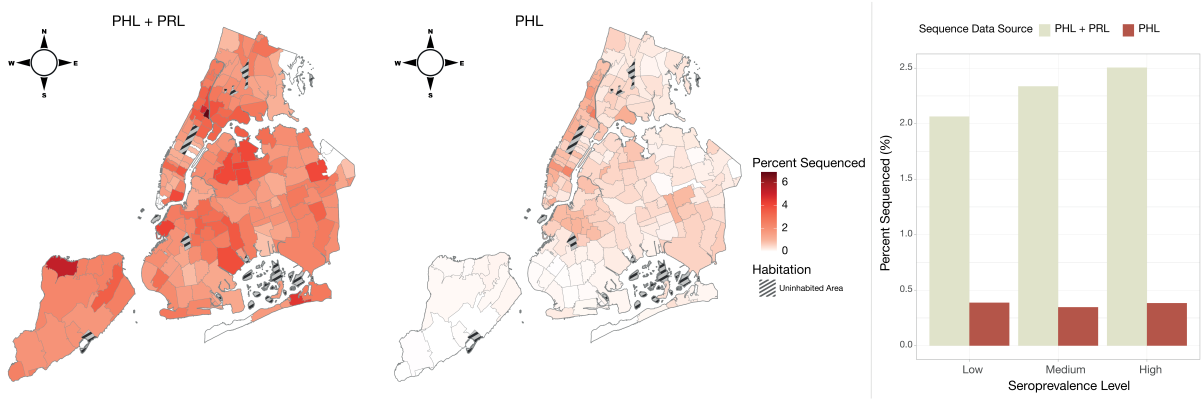



**Figure A.1. Proportion of sequenced samples by ZIP code.** Left panel includes PHL and PRL data, and middle panel is PHL data only. Right panel shows the percent of sequenced genomes for each location used in a phylogeographical analysis (areas of High, Medium, and Low seroprevalence). The base layer of the map obtained from: https://data.cityofnewyork.us/Health/Modified-Zip-Code-Tabulation-Areas-MODZCTA-Map/5fzm-kpwv, distributed openly under the following terms and conditions: https://www.nyc.gov/home/terms-of-use.page.

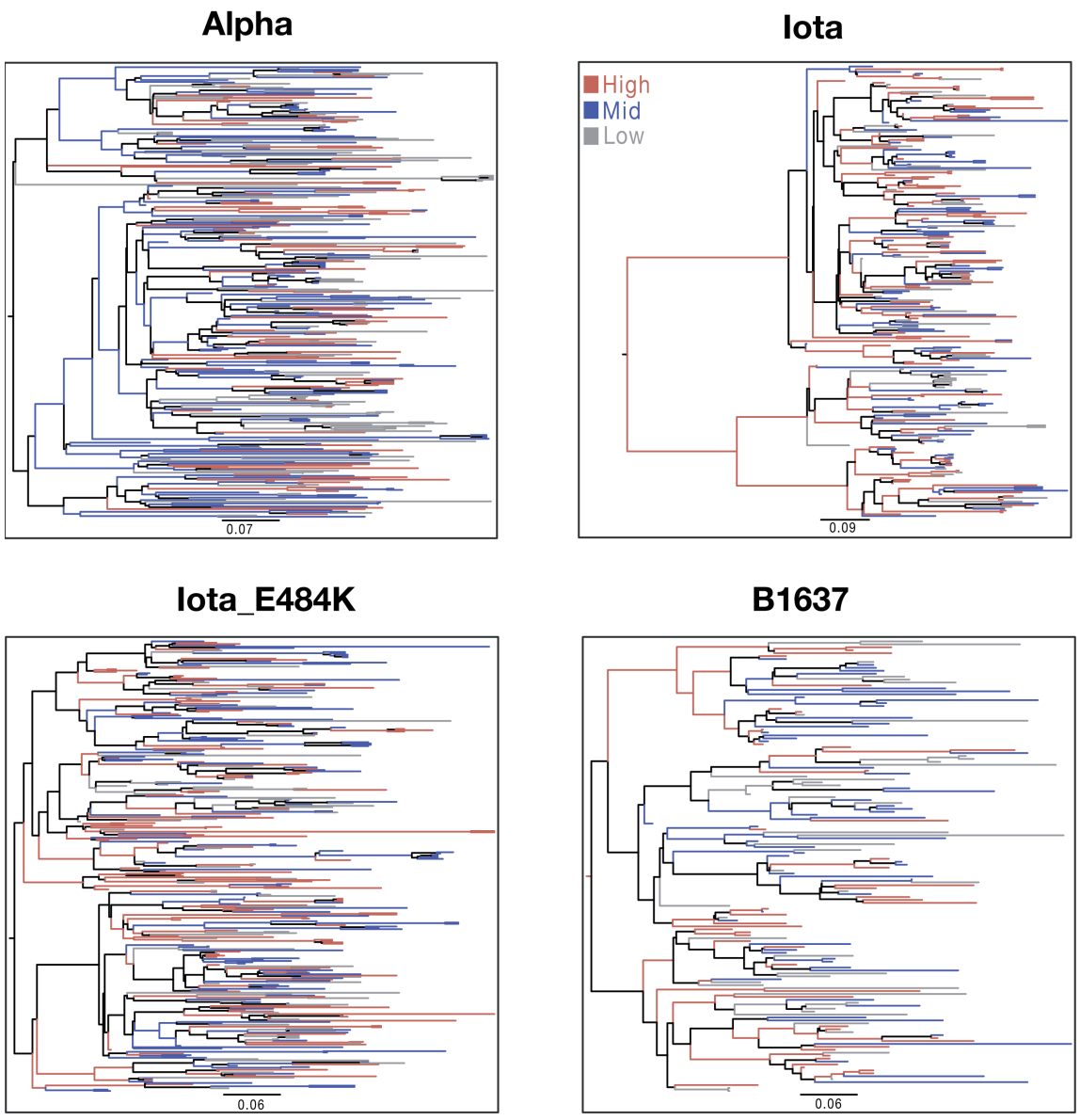



**Figure A.2. Maximum clade credibility phylogenetic trees with branches annotated by location seroprevalence inferred in phylogeographic analysis in BEAST.** Colors denote seroprevalence. Only branches with location posterior probability >70% were annotated.

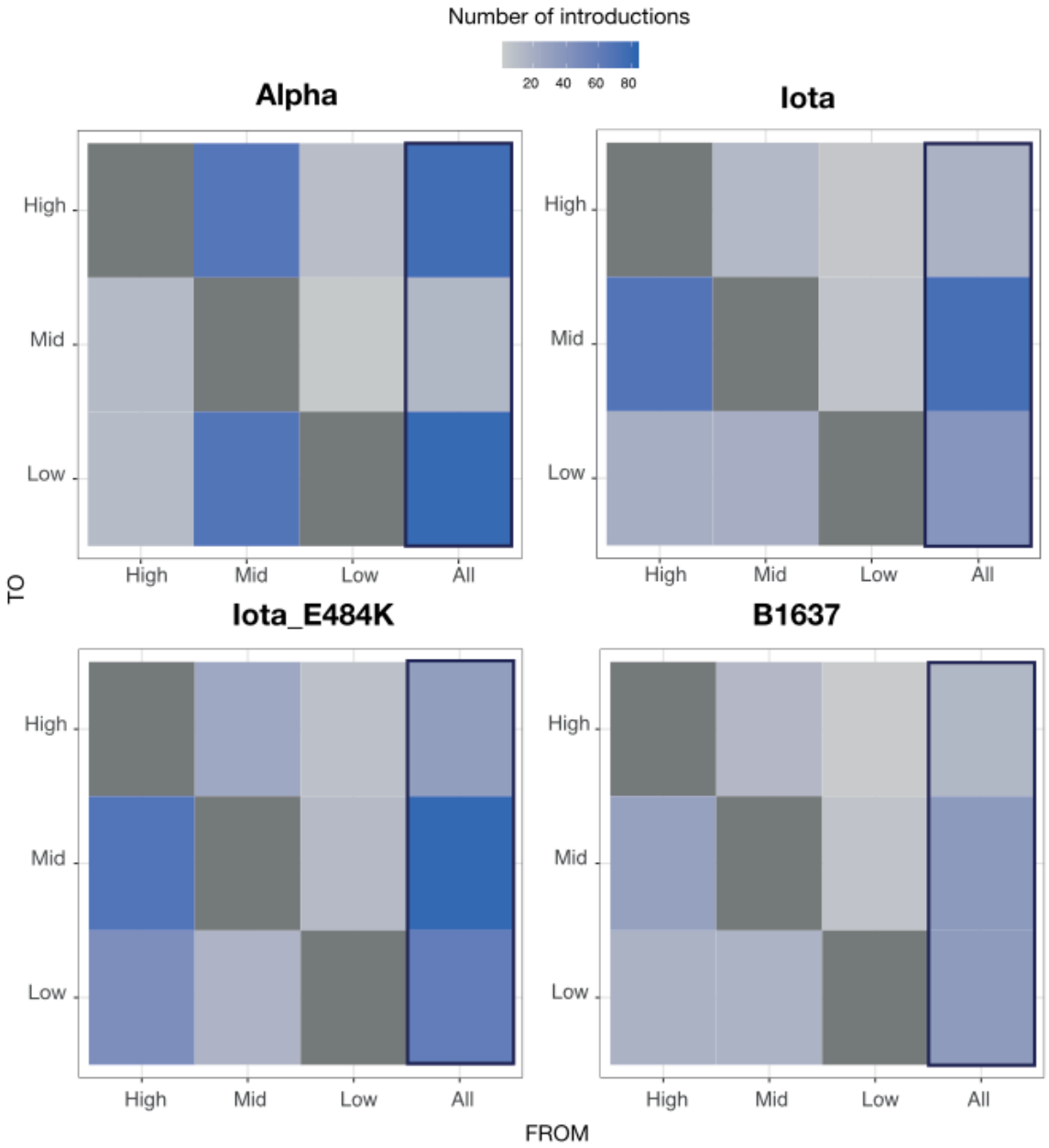



**Figure A.3. Number of introductions from and to each location in phylogeographic analysis where locations were defined based on seropositivity only as High, Medium (Mid), and Low.** Total number of introductions into each location is highlighted with a black square box.

## A.3 Supplementary Data Captions

**Supplementary Data 1** XML for phylogeographic analysis (Alpha). Accessible here.

**Supplementary Data 2** XML for phylogeographic analysis (Iota). Accessible here.

**Supplementary Data 3** XML for phylogeographic analysis (E484K). Accessible here.

**Supplementary Data 4** XML for phylogeographic analysis (B.1.637). Accessible here.

**Supplementary Data 5** Sequence accession numbers. Accessible here.

# Appendix B

# Supplemental Material for Chapter 2

## B.1 Supplementary data

Supp Data 1. Ascension numbers of context sequences is available upon request.

Supp Data 2. Ascension numbers of persistent infection sequences available upon request.

## B.2 Supplementary Tables

**Table B.1.** Molecular clock and rate of nonsynonymous and synonymous substitutions

| | Context | Length (years) | dN/dS | Rate (subs/site/year) | Synonymous subs | Synonymous rate (subs/year) | P (Synonymous rate) | Nonsynonymous subs | Nonsynonymous rate (subs/year) | P (Nonsynonymous rate) |
|---|---|---|---|---|---|---|---|---|---|---|
| Alpha | Variant | 0.43 | 1.2 | 1.29E-03 | 6 | 13.14 | 0.99 | 15 | 35.47 | 0.99 |
| Beta | Variant | 0.45 | 2.34 | 8.00E-04 | 2 | 4.45 | 0.95 | 11 | 25.06 | 0.99 |
| Gamma | Variant | 0.94 | 1.32 | 8.18E-04 | 6 | 6.41 | 0.53 | 19 | 20.11 | 0.99 |
| Delta | Variant | 1.58 | 2.84 | 8.07E-04 | 5 | 3.17 | 0.99 | 35 | 22.2 | 0.99 |
| Epsilon | Variant | 0.54 | 0.56 | 7.38E-04 | 6 | 10.97 | 0.99 | 8 | 14.77 | 0.99 |
| Iota | Variant | 1.07 | 2.04 | 7.20E-04 | 4 | 3.78 | 0.99 | 22 | 18.98 | 0.99 |
| B.1.637 | Variant | 0.51 | 29.59 | 8.96E-04 | 0 | 0 | 0.99 | 16 | 31.4 | 0.99 |
| Mu | Variant | 0.82 | 1.44 | 6.76E-04 | 4 | 4.89 | 0.93 | 14 | 16.97 | 0.99 |
| BA.1 | Variant | 1.45 | 2.43 | 1.04E-03 | 7 | 4.85 | 0.99 | 40 | 27.84 | 0.99 |
| BA.2/BA.4/BA.5 | Variant | 1.91 | 1.37 | 1.09E-03 | 16 | 8.3 | 0.77 | 50 | 26.08 | 0.99 |
| Alpha_x04 | Intrahost | 1.24 | 5.35 | 4.15E-04 | 1 | 0.84 | 0.99 | 13 | 10.67 | 0.84 |
| B11_x01 | Intrahost | 0.32 | 1.64 | 7.75E-04 | 2 | 6.44 | 0.52 | 8 | 25.49 | 0.98 |
| B11_x12 | Intrahost | 0.49 | 1.24 | 5.17E-04 | 2 | 4.1 | 0.83 | 6 | 12.24 | 0.66 |
| B1637_x10 | Intrahost | 0.32 | 0 | 3.49E-04 | 2 | 6.38 | 0.55 | 0 | 0 | 0.99 |
| Delta_x14 | Intrahost | 0.36 | 27.62 | 2.86E-04 | 0 | 0 | 0.99 | 1 | 2.78 | 0.99 |
| Delta_x11 | Intrahost | 0.5 | 3.04 | 9.97E-04 | 2 | 4.51 | 0.9 | 15 | 31.19 | 0.99 |
| Iota_x02 | Intrahost | 0.55 | 30.28 | 2.44E-04 | 0 | 0 | 0.99 | 1 | 1.82 | 0.99 |
| Deer_x04 | Deer | 1.23 | 0.39 | 2.23E-03 | 45 | 36.88 | 0.99 | 38 | 31.06 | 0.99 |
| Deer_x05 | Deer | 0.91 | 0.54 | 1.31E-03 | 17 | 19.07 | 0.99 | 20 | 22.42 | 0.99 |
| Deer_x01 | Deer | 1.3 | 0.43 | 9.18E-04 | 20 | 15.21 | 0.97 | 20 | 15.54 | 0.79 |
| Deer_x02 | Deer | 0.82 | 0.39 | 1.46E-03 | 21 | 26.25 | 0.99 | 20 | 24.83 | 0.99 |
| Deer_x03 | Deer | 0.45 | 0.48 | 1.39E-03 | 11 | 23.45 | 0.99 | 13 | 28.67 | 0.99 |
| All background | Background | | 0.63 | 5.67E-04 | | 6.8 | | | 10.1 | |
| Internal branch | Background | | 0.65 | 5.66E-04 | | 7.02 | | | 10.3 | |



**Table B.2.** Synonymous substitutions of variant stem lineages

| | Rate (subs/site /year) | dN/dS | Length (years) | Synonymous subs | Expected synonymous subs | Synonymous rate (subs/year) | P (Synonymous rate) |
|---|---|---|---|---|---|---|---|
| Alpha | 1.29E-03 | 1.20 | 0.43 | 6 | 3.04 (2.31, 3.5) | 13.14 | 0.99 |
| Beta | 8.00E-04 | 2.34 | 0.45 | 2 | 3.13 (2.04, 4.49) | 4.45 | 0.95 |
| Gamma | 8.18E-04 | 1.32 | 0.94 | 6 | 6.58 (5.36, 7.45) | 6.41 | 0.53 |
| Delta | 8.07E-04 | 2.84 | 1.58 | 5 | 11.08 (9.45, 12.76) | 3.17 | 0.99 |
| Epsilon | 7.38E-04 | 0.56 | 0.54 | 6 | 3.81 (2.62, 4.99) | 10.97 | 0.99 |
| Iota | 7.20E-04 | 2.04 | 1.07 | 4 | 7.53 (6.02, 9.24) | 3.78 | 0.99 |
| B.1.637 | 8.96E-04 | 29.59 | 0.51 | 0 | 3.58 (2.65, 4.33) | 0 | 0.99 |
| Mu | 6.76E-04 | 1.44 | 0.82 | 4 | 5.75 (4.19, 6.63) | 4.89 | 0.93 |
| BA.1 | 1.04E-03 | 2.43 | 1.45 | 7 | 10.17 (9.21, 11.2) | 4.85 | 0.99 |
| BA.2/BA.4/BA.5 | 1.09E-03 | 1.37 | 1.91 | 16 | 13.39 (12.06, 14.68) | 8.3 | 0.77 |

# Appendix C

# Supplemental Material for Chapter 3

## C.1 Supplementary Figures

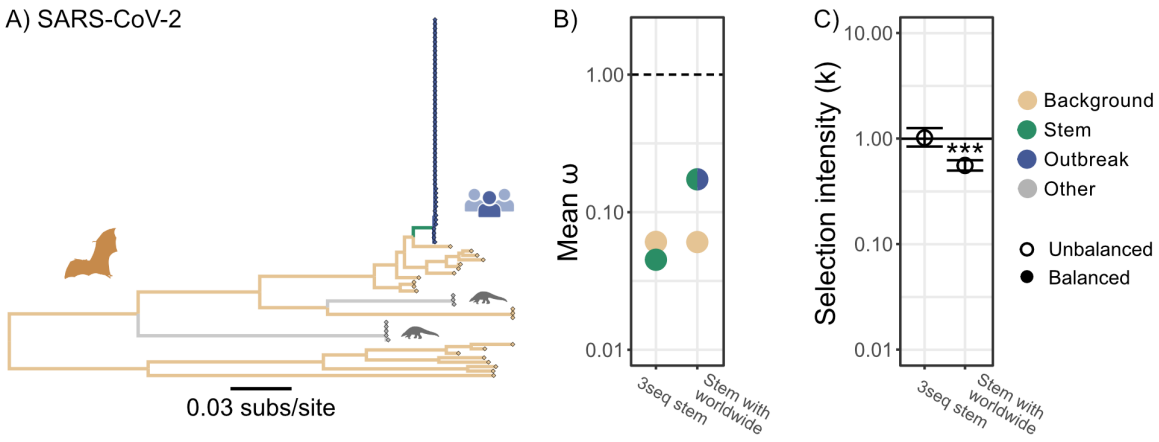



**Figure C.1. Quantifying selection regimes in SARS-CoV-2 with human outbreak.** (A) Phylogenetic tree of SARS-CoV-2 like sarbecoviruses non recombinant region 08, branch color indicate partition. (B) Mean $\omega$ for each branch partition of background and the test sets, including 3seq stem, which compares the stem branch with background, based on alignment of non-recombinant regions inferred with 3seq (no tree displayed), and "combined with worldwide" which compares stem and worldwide outbreak (n , from time ) to background (C) Change in selection intensity comparing selection associated with background to test set. Filled circles indicate a balanced model where directionality is identifiable, open circles indicate an unbalanced model and the direction above or below 1 is not identifiable. Numerical values in Supp Table C.2. Significance indicated with* $p$<0.05; ** $p$<0.01; and *** $p$<0.001. Icons created with BioRender.

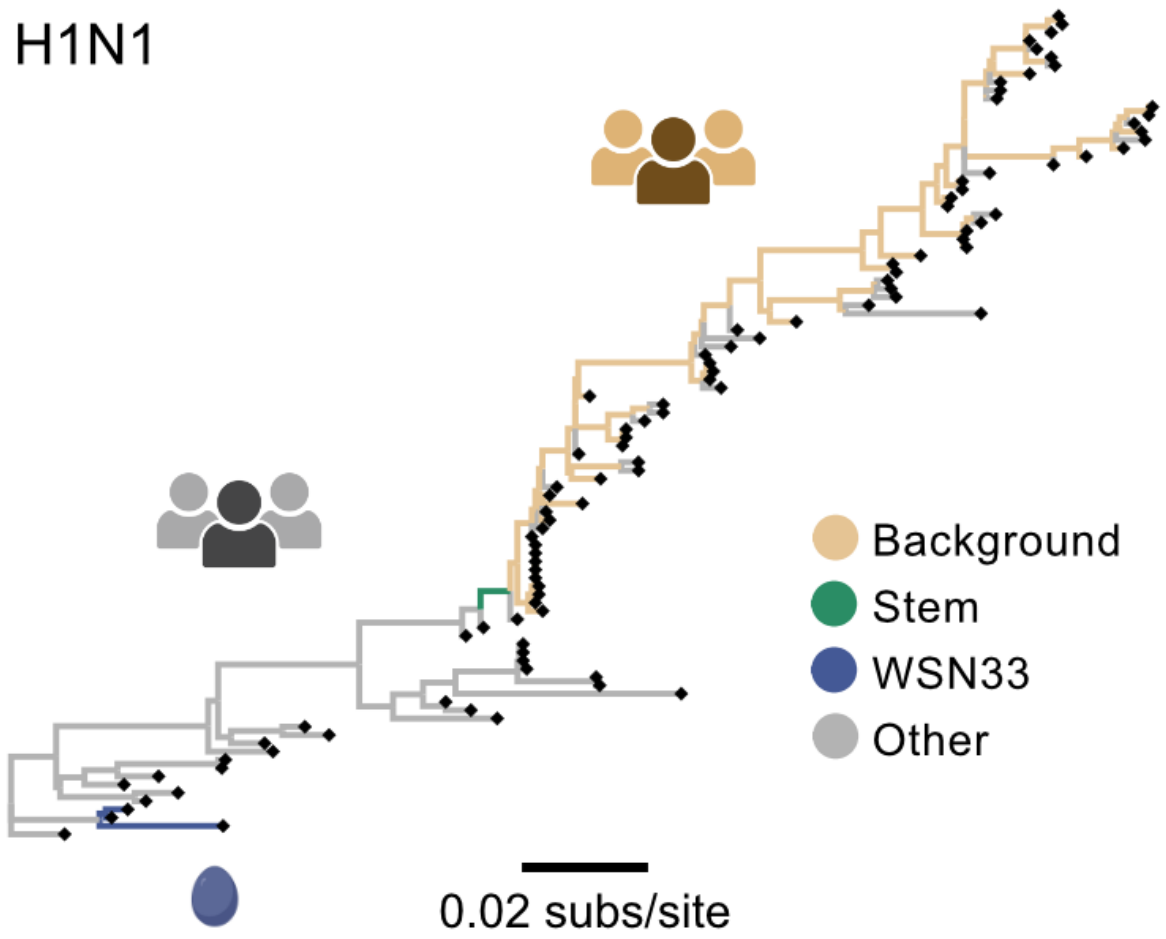



**Figure C.2. Phylogenetic tree of H1N1 HA segment, branch color indicates partition.** The stem branch connects 1977 pandemic reemergence and basal ancestors, and is foreground for 1977 flu test. WSN33 (blue) branches are clade associated with laboratory passaged positive control. Post 1977 branches associated human to human transmission (brown) used as background. Other branches that are pre-1977 or tips that have a history of passaging (grey) are excluded from analysis. Icons created with BioRender.

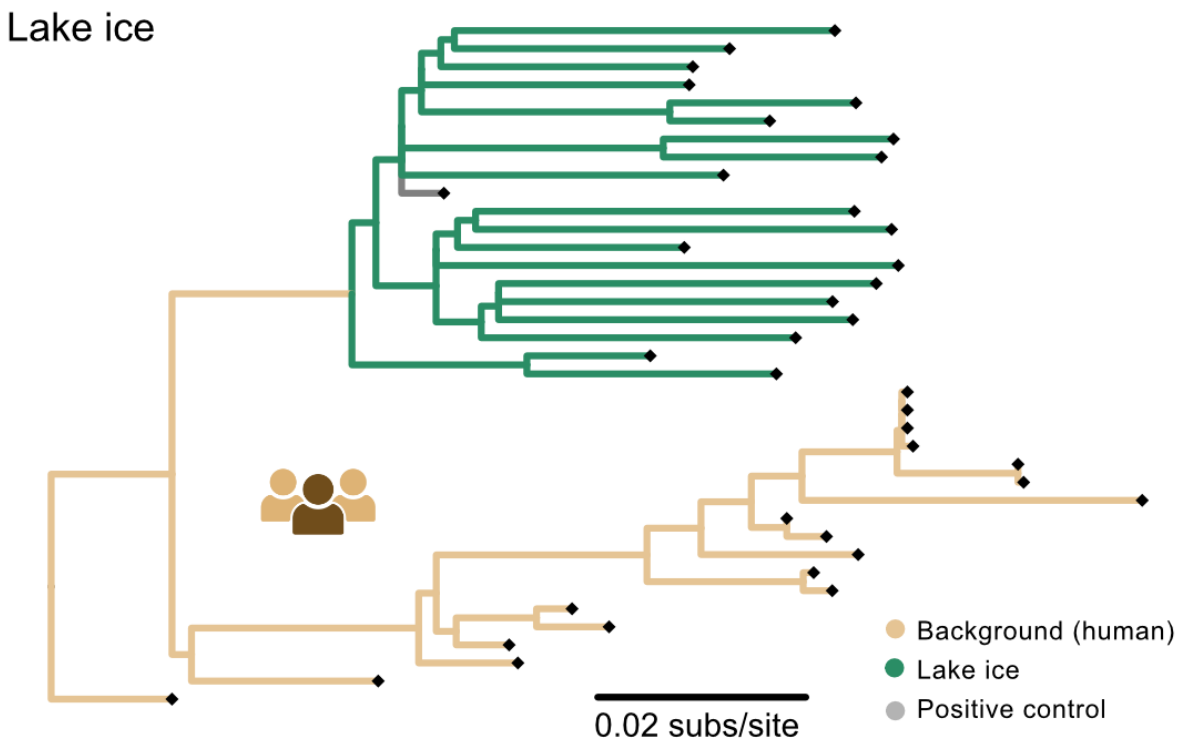



**Figure C.3. Phylogenetic tree of H1N1 HA segment, branch color indicates partition.** Lake ice clade (green) is the foreground and branches associated with human to human transmission (brown) are background of Lake ice test. The study positive control (grey) is excluded. Icons created with BioRender.

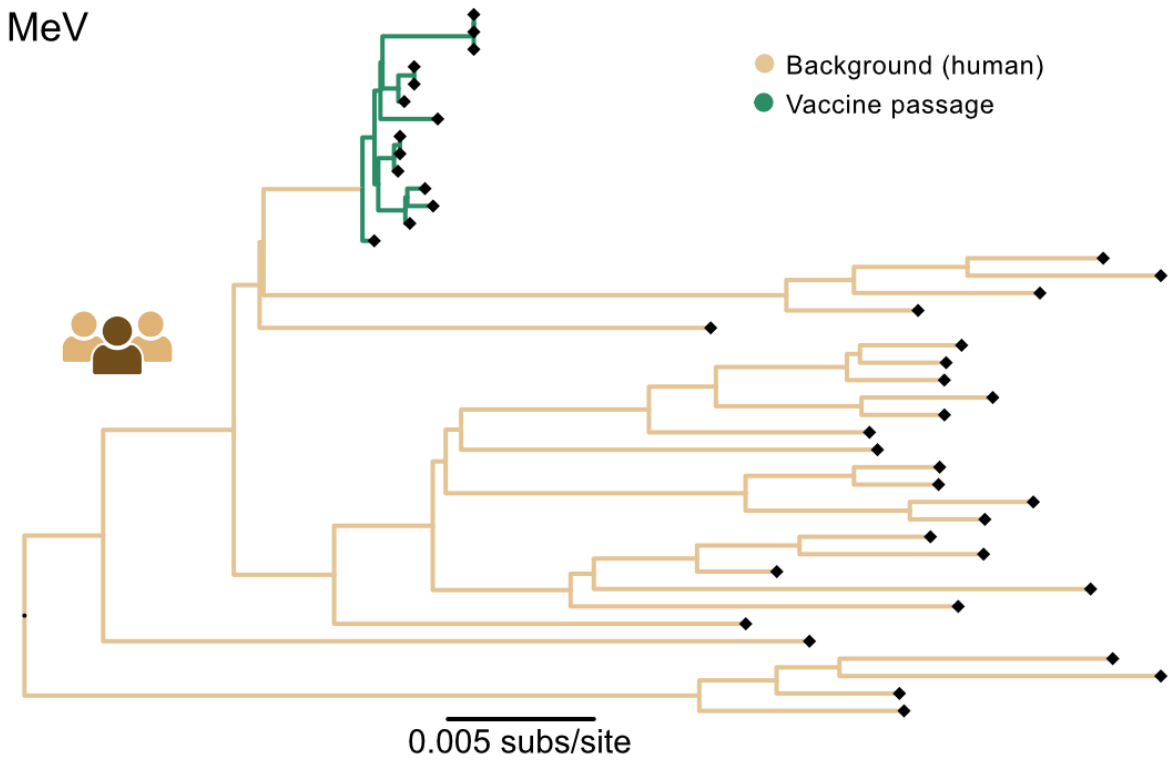



**Figure C.4. Phylogenetic tree of MeV genome, branch color indicates partition.** Vaccine attenuated by passage (green) is the foreground and branches associated with human to human transmission. Icons created with BioRender.

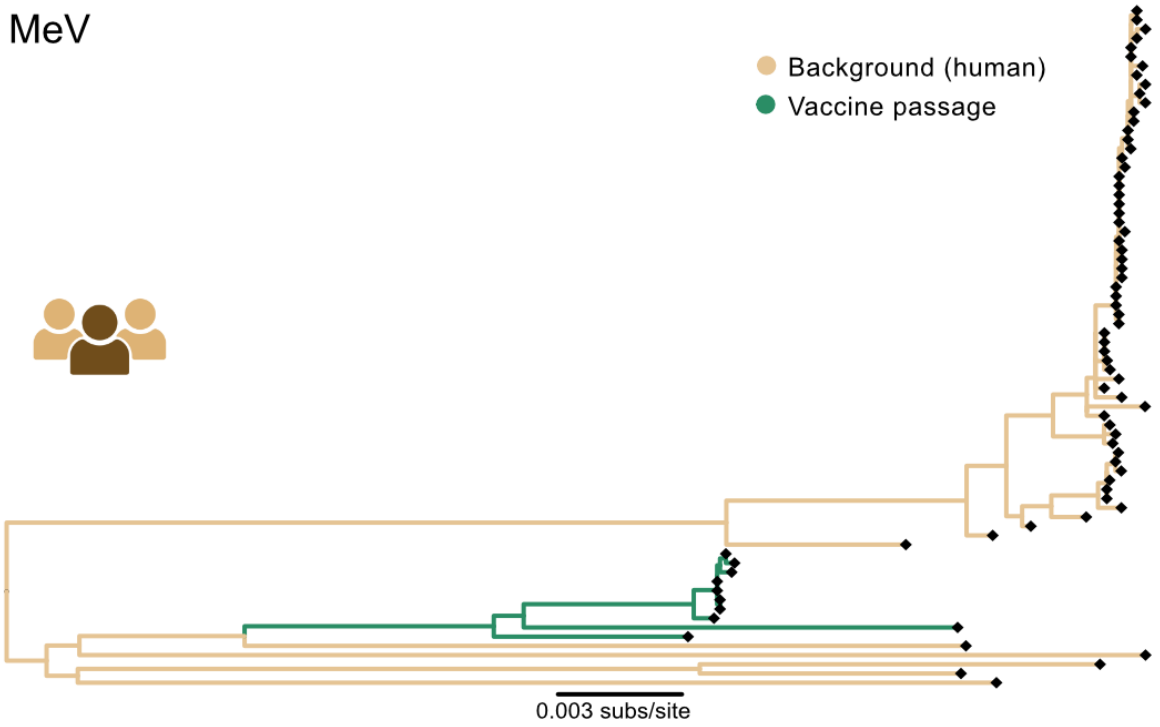



**Figure C.5. Phylogenetic tree of MuV genome, branch color indicates partition.** Vaccine attenuated by passage (green) is the foreground and branches associated with human to human transmission. Icons created with BioRender.

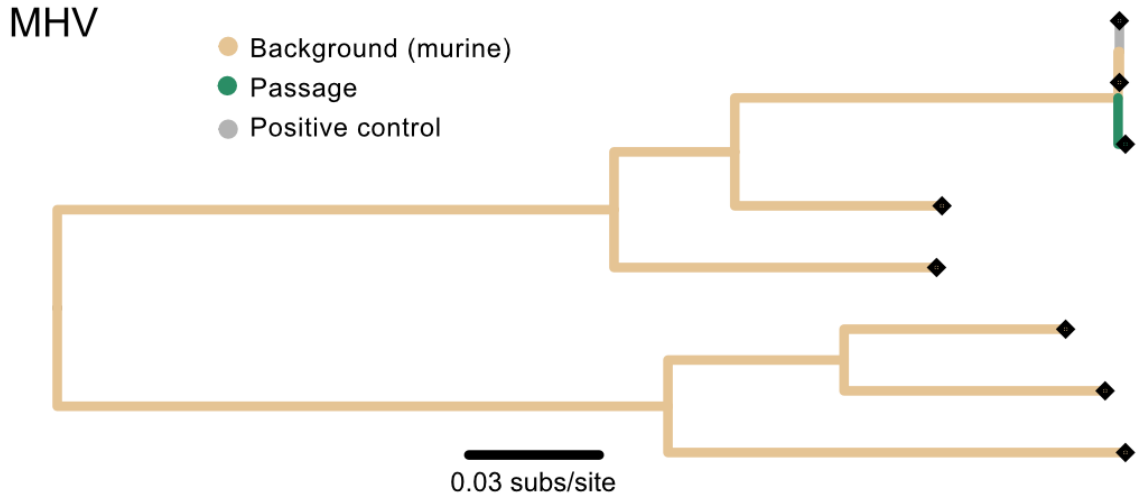



**Figure C.6. Phylogenetic tree of MHV.** Non-recombinant region 07, defined by GARD, is displayed, branch color indicates partition. Laboratory passaged virus (green) is the foreground and branches between murine strains (brown) are background of Lake ice test. The isolate used at the start of passaging (grey) is excluded.

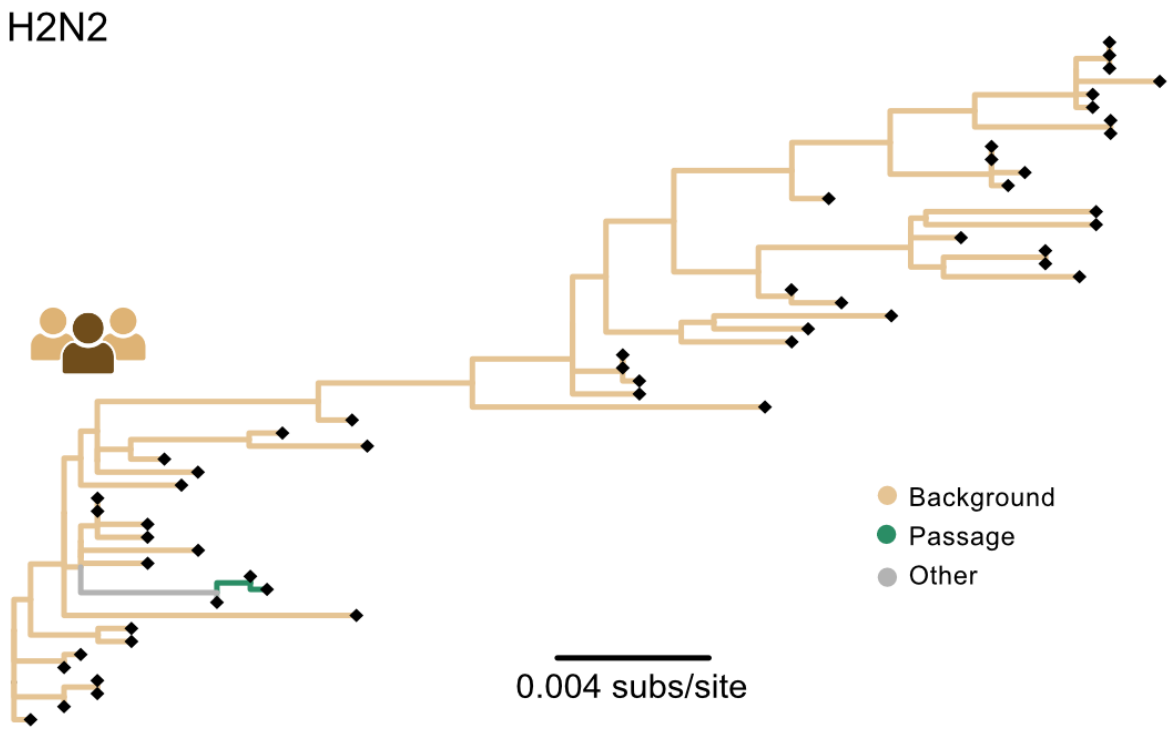



**Figure C.7. Phylogenetic tree of H2N2 PB1 segment.** Branch color indicates partition. Cold adaptation passage (green) is the foreground and branches associated with human to human transmission (brown) are background. The stem to isolate used at the start of passaging (grey) is excluded. Icons created with BioRender.

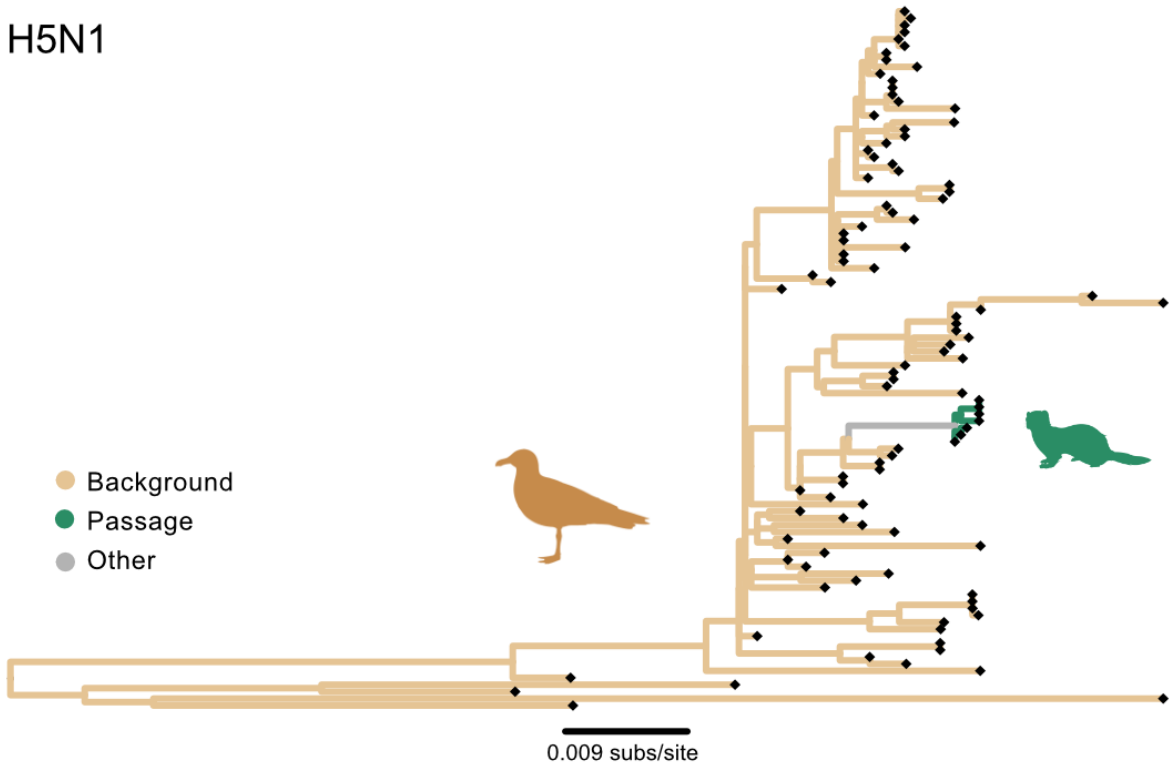



**Figure C.8. Phylogenetic tree of H5N1 PB1 segment.** Branch color indicates partition. Airborne ferret adaptation passage (green) is the foreground and branches associated with avian host reservoir (brown) are background. The stem to isolate used at the start of passaging (grey) is excluded. Icons created with BioRender.

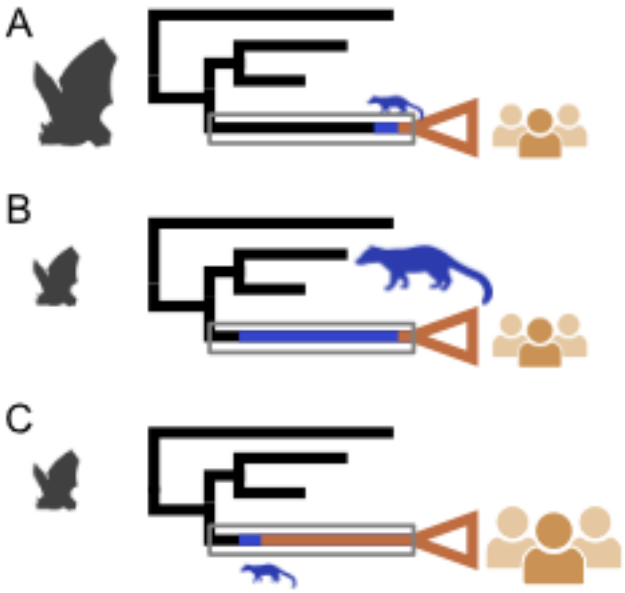



**Figure C.9. Diagram of host during evolutionary history.** Branches are colored by the host of the virus, the natural host reservoir (black), intermediate host (blue) or human (orange) (A) Evolution on the stem lineage leading the MRCA of human outbreak is primarily associated with the natural host reservoir. (B) Evolution on the stem lineage leading the MRCA of human outbreak is primarily associated with (unsampled) intermediate host. (C) Evolution on the stem lineage leading the MRCA of human outbreak is primarily associated with initial stuttering transmission in the human population, not captured by observed samples. Icons created with BioRender

## C.2 Supplementary Tables

**Table C.1.** Results of RELAX test

| type | test | k | k-unstable | p | significance | bestModel | LB | UB | Balanced | Avg. omega reference | Avg. omega - test |
|---|---|---|---|---|---|---|---|---|---|---|---|
| zoonotic | EBOV stem | 1.089 | | 0.34 | | SRV.noMH.3 | 0.911 | 1.295 | FALSE | 0.104 | 0.086 |
| zoonotic | EBOV combined | 6.34E-06 | | 1.49E-04 | *** | SRV.noMH.3 | 6.68E-07 | 0.434 | FALSE | 0.104 | 0.184 |
| zoonotic | EBOV outbreak | 0.413 | | 3.56E-07 | *** | SRV.noMH.3 | 0.288 | 0.553 | FALSE | 0.104 | 0.388 |
| zoonotic | H1N1pdm09 stem | 1.077 | | 0.212 | | SRV.noMH.3 | 0.947 | 1.221 | TRUE | 0.125 | 0.105 |
| zoonotic | H1N1pdm09 combined | 0.907 | | 0.01 | * | SRV.noMH.3 | 0.836 | 0.982 | TRUE | 0.125 | 0.147 |
| zoonotic | H1N1pdm09 outbreak | 0.82 | | 2.23E-05 | *** | SRV.noMH.3 | 0.736 | 0.909 | TRUE | 0.125 | 0.176 |
| zoonotic | MARV | 1.031 | | 0.904 | | SRV.noMH.2 | 0.736 | 50 | TRUE | 0.097 | 0.102 |
| SARS-1 | SARS-1 stem | 0.969 | | 0.907 | | SRV.MH.3 | 0.719 | 1.723 | TRUE | 0.059 | 0.052 |
| SARS-1 | SARS-1 stem with outbreak | 0.653 | | 1.72E-04 | *** | SRV.MH.3 | 0.537 | 0.826 | TRUE | 0.059 | 0.095 |
| SARS-2 | SARS-CoV-2 stem | 1.117 | | 0.236 | | SRV.MH.3 | 0.915 | 1.401 | FALSE | 0.061 | 0.041 |
| SARS-2 | SARS-CoV-2 stem 3seq-NRR | 1.019 | | 0.823 | | SRV.MH.3 | 0.843 | 1.261 | FALSE | 0.061 | 0.045 |
| SARS-2 | SARS-CoV-2 stem with outbreak China | 0.687 | | 1.08E-07 | *** | SRV.MH.3 | 0.595 | 0.798 | FALSE | 0.061 | 0.096 |
| SARS-2 | SARS-CoV-2 stem with outbreak worldwide | 0.558 | | 0 | *** | SRV.MH.3 | 0.499 | 0.625 | FALSE | 0.061 | 0.173 |
| SARS-2 | SARS-CoV-2-like pangolin | 0.018 | | 0 | *** | SRV.MH.3 | 0 | 0.103 | FALSE | 0.061 | 0.936 |
| flu1977 | Flu 1977 full background | 0.707 | | 0.043 | * | SRV.noMH.3 | 0.447 | 1.027 | TRUE | 0.116 | 0.214 |
| flu1977 | Flu 1977 background excludes passaged | 0.705 | | 0.042 | * | SRV.noMH.3 | 0.445 | 1.025 | TRUE | 0.115 | 0.215 |
| passage | MeV | 0.033 | | 0 | *** | SRV.MH.3 | 0 | 0.154 | TRUE | 0.12 | 0.936 |
| passage | MuV | 0.841 | | 0.006 | ** | SRV.MH.2 | 0.734 | 0.961 | TRUE | 0.073 | 0.11 |
| passage | MHV | 0 | | 9.64E-06 | *** | SRV.MH.3 | 0 | 0.165 | FALSE | 0.116 | 1.635 |
| passage | WSN33 | 0.241 | | 0 | *** | SRV.noMH.3 | 0.116 | 0.378 | TRUE | 0.116 | 0.6 |
| passage | Ice tips | 0.01 | | 1.98E-14 | *** | SRV.noMH.2 | 0 | 0.105 | TRUE | 0.314 | 0.981 |
| passage | H2N2 | 0 | | 7.84E-06 | *** | SRV.noMH.2 | 0 | 0.352 | FALSE | 0.144 | 4.518 |
| passage | H5N1 | 0.011 | 2.12 | 2.04E-05 | *** | SRV.noMH.3 | 0 | 0.187 | TRUE | 0.121 | 0.929 |

**Table C.2.** Inferred substitutions of 1977 reemergence

| | Nonsynonymous substitutions | Synonymous substitutions |
|---|---|---|
| HA | 3 | 5 |
| M | 2 | 0 |
| NA | 3 | 2 |
| NP | 4 | 2 |
| NS | 1 | 1 |
| PA | 0 | 2 |
| PB1 | 1 | 12 |
| PB2 | 3 | 7 |

# Appendix D

# Supplemental Material for Chapter 4

## D.1 Supplementary Figures

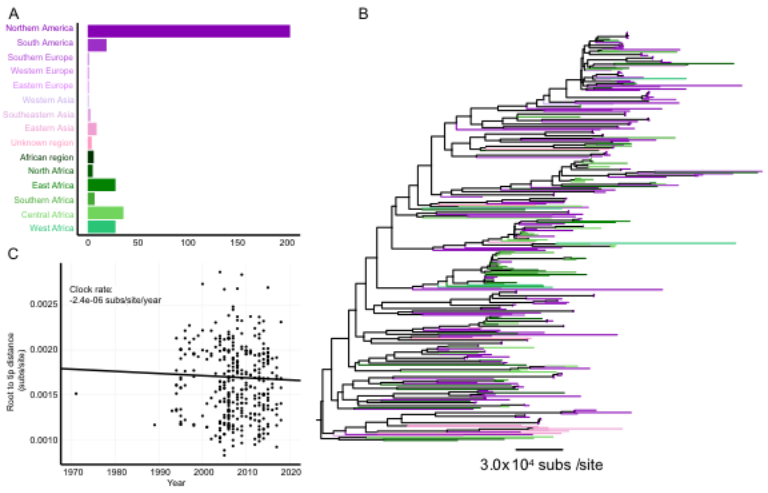



**Figure D.1. Phylogenetic analysis of subset of HSV-2 worldwide lineage.** (A) HSV-2 dataset of 353 sequences in out of Africa clade of worldwide lineage, by number of samples from each region. (B) ML phylogenetic tree inferred with GTR+F+R4 model of sequences described in panel A. Color indicates region of sampling for tip. (C) Root to tip distance of ML tree verse sampling year, with regression estimate of clock rate of -2.4x10$^{-6}$ substitutions/site/year.

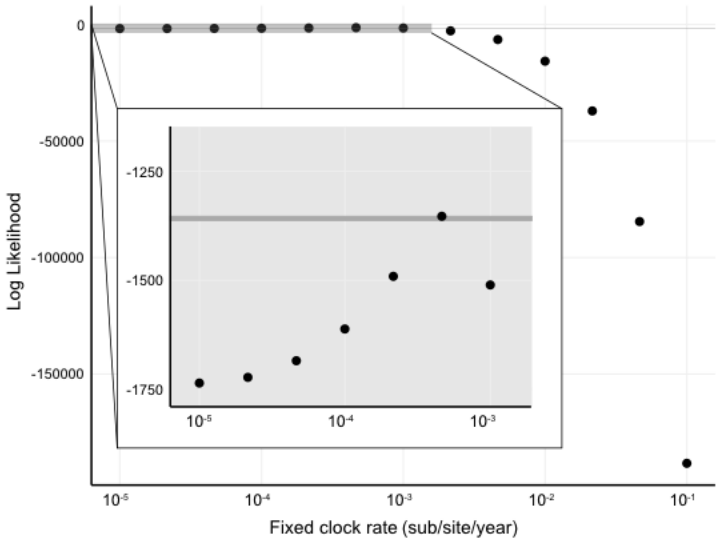



**Figure D.2. Shape of likelihood surface for clock rate of ebolavirus.** Log likelihood of ebolavirus time tree at fixed clock rates estimated using input ML tree (1) of 1610 Ebola genomes, black line is maximum estimated likelihood, grey shaded box is the area shown in insert. (Insert) Black line is maximum estimated likelihood at rate $4.6 \times 10^{-4}$ substitutions/site/year, dark grey box is range from maximum estimated likelihood to 10 points below.

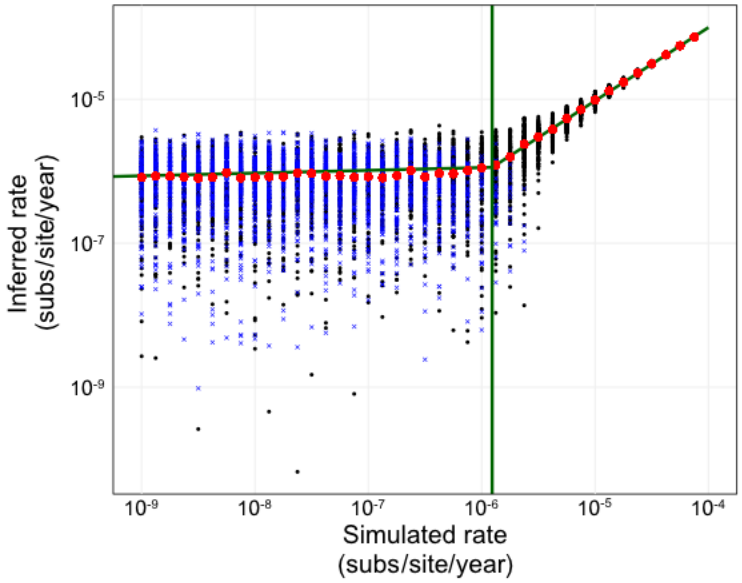



**Figure D.3. Breakpoint analysis of clock rate inference on simulated data.** Estimates of rate for a fixed input tree with simulated sequences and empirical dates. Black dot represents estimates of positive rates, while blue 'x' represent the absolute value of estimates of negative rates, and the mean of absolute value for each fixed rate is indicated by red dot. Sloped green segments represent break point regression of absolute value of estimated rates. Vertical green line represents break point at $1.2\text{x}10^{-6}$ substitutions/site/year (R segmented). 200 replicates were generated under GTR+Γ4 model with fixed rates $1\text{x}10^{-9}$ to $1\text{x}10^{-4}$ substitutions/site/year.

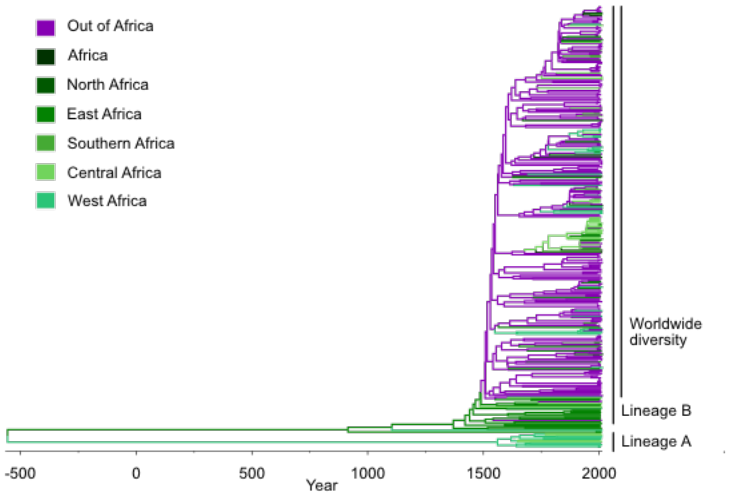



**Figure D.4. Phylogeography of HSV-2 supports East Africa as origin of out-of-Africa dispersal.** Time tree inferred with the fastest rate supported by any clock analysis, $2.4\text{x}10^{-6}$ substitutions/site/year. ML tree topology scaled to time tree, colored by plurality of support in mugration analysis for each region. Time scale represents the calendar year; negative years are BCE.

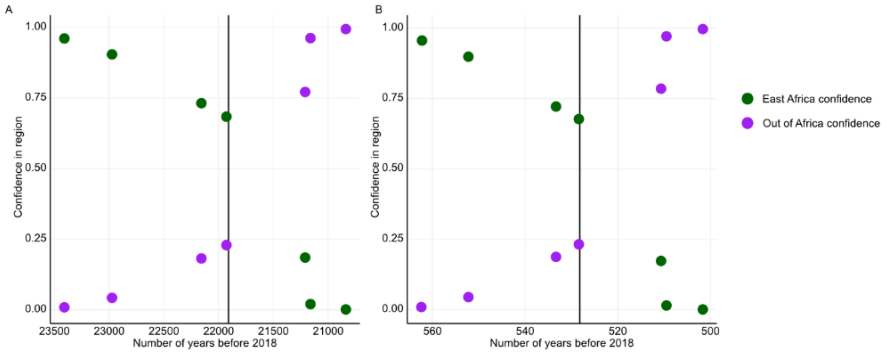



**Figure D.5. Internal nodes in backbone of tree around out of Africa migration (within the worldwide lineage).** Confidence in East Africa placement (green) and out of Africa placement (purple), by number of years before 2018, for time trees with rate $5.8 \times 10^{-8}$ (A) and $2.4 \times 10^{-6}$ (B) substitutions/site/year. Vertical bar indicates average age of nodes with ages weighted by *1-absolute value(East Africa confidence - out of Africa confidence)*.

## D.2 Supplementary Data Captions

**Supplementary Data 1** Sample metadata with sequence information. Available here

**Supplementary Data 2** Description: Nexus formatted treefile of empirical HSV-2 dataset, visualized in Fig. 1. ML tree of 395 HSV-2 sequences estimated with GTR+F+R4. The tree displayed is midpoint rooted, though the identical rooting orientation is recovered using the closely related chimpanzee simplexvirus. Node labels indicate aBayes support. Available here

**Supplementary Data 3** Description: Nexus formatted time trees inferred using with fixed evolutionary rates. Associated summary of ages and phylogeographic source in Table 1. Available here

## D.3 Supplementary Tables

**Table D.1.** Age of root and out-of-Africa dispersal across clock rates for robustness analysis of geographic partitioning

| Partitioning scheme | Substitution rate | Root age[1] | Maximum support for East Africa migration | Rate support in simulations |
|---|---|---|---|---|
| All subregions | 1.0E-08 | 122447 | 0.976 | Not rejected |
| All subregions | 1.3E-08 | 91470 | 0.977 | Not rejected |
| All subregions | 1.8E-08 | 68746 | 0.976 | Not rejected |
| All subregions | 2.4E-08 | 51445 | 0.977 | Not rejected |
| All subregions | 3.2E-08 | 38414 | 0.977 | Not rejected |
| All subregions | 4.2E-08 | 28910 | 0.978 | Not rejected |
| All subregions | 5.6E-08 | 21641 | 0.977 | Not rejected |
| All subregions | 7.5E-08 | 16279 | 0.977 | Not rejected |
| All subregions | 1.0E-07 | 12163 | 0.977 | Not rejected |
| All subregions | 1.3E-07 | 9135 | 0.977 | Not rejected |
| All subregions | 1.8E-07 | 6838 | 0.978 | Not rejected |
| All subregions | 2.4E-07 | 5130 | 0.978 | Not rejected |
| All subregions | 3.2E-07 | 3851 | 0.978 | Not rejected |
| All subregions | 4.2E-07 | 2891 | 0.978 | Intermediate |
| All subregions | 5.6E-07 | 2181 | 0.977 | Intermediate |
| All subregions | 7.5E-07 | 1630 | 0.979 | Intermediate |
| All subregions | 1.0E-06 | 1226 | 0.979 | Intermediate |
| All subregions | 1.3E-06 | 923 | 0.981 | Intermediate |
| All subregions | 1.8E-06 | 693 | 0.982 | Intermediate |
| All subregions | 2.4E-06 | 521 | 0.983 | Intermediate |
| All subregions | 3.2E-06 | 396 | 0.984 | Rejected by all |
| All subregions | 4.2E-06 | 299 | 0.986 | Rejected by all |
| All subregions | 5.6E-06 | 226 | 0.988 | Rejected by all |
| All subregions | 7.5E-06 | 173 | 0.989 | Rejected by all |
| All subregions | 1.0E-05 | 132 | 0.992 | Rejected by all |
| All subregions | 1.3E-05 | 102 | 0.994 | Rejected by all |
| All subregions | 1.8E-05 | 80 | 0.995 | Rejected by all |
| All subregions | 2.4E-05 | 64 | 0.997 | Rejected by all |
| All subregions | 3.2E-05 | 54 | 0.997 | Rejected by all |
| All subregions | 4.2E-05 | 52 | 0.993 | Rejected by all |
| All subregions | 5.6E-05 | 50 | 0.967 | Rejected by all |
| All subregions | 7.5E-05 | 49 | 0.824 | Rejected by all |
| All subregions | 1.0E-04 | 49 | 0.437 | Rejected by all |
| Out of Africa and African merged | 1.0E-08 | 124370 | 0.985 | Not rejected |
| Out of Africa and African merged | 1.3E-08 | 92838 | 0.986 | Not rejected |
| Out of Africa and African merged | 1.8E-08 | 69776 | 0.985 | Not rejected |
| Out of Africa and African merged | 2.4E-08 | 52233 | 0.986 | Not rejected |
| Out of Africa and African merged | 3.2E-08 | 38987 | 0.985 | Not rejected |
| Out of Africa and African merged | 4.2E-08 | 29342 | 0.986 | Not rejected |
| Out of Africa and African merged | 5.6E-08 | 21961 | 0.986 | Not rejected |
| Out of Africa and African merged | 7.5E-08 | 16528 | 0.985 | Not rejected |
| Out of Africa and African merged | 1.0E-07 | 12350 | 0.985 | Not rejected |
| Out of Africa and African merged | 1.3E-07 | 9271 | 0.985 | Not rejected |
| Out of Africa and African merged | 1.8E-07 | 6943 | 0.986 | Not rejected |
| Out of Africa and African merged | 2.4E-07 | 5206 | 0.986 | Not rejected |
| Out of Africa and African merged | 3.2E-07 | 3911 | 0.986 | Not rejected |
| Out of Africa and African merged | 4.2E-07 | 2937 | 0.986 | Intermediate |
| Out of Africa and African merged | 5.6E-07 | 2214 | 0.985 | Intermediate |
| Out of Africa and African merged | 7.5E-07 | 1655 | 0.985 | Intermediate |
| Out of Africa and African merged | 1.0E-06 | 1245 | 0.985 | Intermediate |
| Out of Africa and African merged | 1.3E-06 | 937 | 0.985 | Intermediate |
| Out of Africa and African merged | 1.8E-06 | 704 | 0.985 | Intermediate |
| Out of Africa and African merged | 2.4E-06 | 530 | 0.984 | Intermediate |
| Out of Africa and African merged | 3.2E-06 | 403 | 0.984 | Rejected by all |
| Out of Africa and African merged | 4.2E-06 | 304 | 0.984 | Rejected by all |
| Out of Africa and African merged | 5.6E-06 | 230 | 0.984 | Rejected by all |
| Out of Africa and African merged | 7.5E-06 | 176 | 0.983 | Rejected by all |
| Out of Africa and African merged | 1.0E-05 | 134 | 0.984 | Rejected by all |
| Out of Africa and African merged | 1.3E-05 | 104 | 0.984 | Rejected by all |
| Out of Africa and African merged | 1.8E-05 | 81 | 0.986 | Rejected by all |
| Out of Africa and African merged | 2.4E-05 | 65 | 0.988 | Rejected by all |
| Out of Africa and African merged | 3.2E-05 | 55 | 0.989 | Rejected by all |
| Out of Africa and African merged | 4.2E-05 | 52 | 0.977 | Rejected by all |
| Out of Africa and African merged | 5.6E-05 | 50 | 0.939 | Rejected by all |
| Out of Africa and African merged | 7.5E-05 | 49 | 0.827 | Rejected by all |
| Out of Africa and African merged | 1.0E-04 | 49 | 0.551 | Rejected by all |

# Bibliography


[1] 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, and others. The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. 377(6609):960–966, . Publisher: American Association for the Advancement of Science.

[2] John hopkins university of medicine. coronavirus resource center. 2023. accessed june 14, 2024. https://coronavirus.jhu.edu/map.html, .

[3] Wk Lam, Ns Zhong, and Wc Tan. Overview on SARS in asia and the world. 8:S2–S5, . ISSN 1440-1843. doi: 10.1046/j.1440-1843.2003.00516.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1440-1843.2003.00516.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1440-1843.2003.00516.x.

[4] Fatimah S Dawood, A Danielle Iuliano, Carrie Reed, Martin I Meltzer, David K Shay, Po-Yung Cheng, Don Bandaranayake, Robert F Breiman, W Abdullah Brooks, Philippe Buchy, Daniel R Feikin, Karen B Fowler, Aubree Gordon, Nguyen Tran Hien, Peter Horby, Q Sue Huang, Mark A Katz, Anand Krishnan, Renu Lal, Joel M Montgomery, Kåre Mølbak, Richard Pebody, Anne M Presanis, Hugo Razuri, Anneke Steens, Yeny O Tinoco, Jacco Wallinga, Hongjie Yu, Sirenda Vong, Joseph Bresee, and Marc-Alain Widdowson. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza a h1n1 virus circulation: a modelling study. 12(9):687–695. ISSN 1473-3099. doi: 10.1016/S1473-3099(12)70121-4. URL https://www.sciencedirect.com/science/article/pii/S1473309912701214.

[5] Simon Dellicour, Guy Baele, Gytis Dudas, Nuno R. Faria, Oliver G. Pybus, Marc A. Suchard, Andrew Rambaut, and Philippe Lemey. Phylodynamic assessment of intervention strategies for the west african ebola virus outbreak. 9(1):2222, . ISSN 2041-1723. doi: 10.1038/s41467-018-03763-2. URL https://www.nature.com/articles/s41467-018-03763-2. Publisher: Nature Publishing Group.

[6] Ebola vaccine purchasing commitment from gavi to prepare for future outbreaks. . URL https://www.gavi.org/news/media-room/ebola-vaccine-purchasing-commitment-gavi-prepare-future-outbreaks.

[7] Ioannis Magouras, Victoria J. Brookes, Ferran Jori, Angela Martin, Dirk Udo Pfeiffer, and Salome Dürr. Emerging zoonotic diseases: Should we rethink the animal–human interface? 7. ISSN 2297-1769. doi: 10.3389/fvets.2020.582743. URL https://www.frontiersin.org/

articles/10.3389/fvets.2020.582743. Publisher: Frontiers.

[8] Agustín Estrada-Peña, Richard S. Ostfeld, A. Townsend Peterson, Robert Poulin, and José de la Fuente. Effects of environmental change on zoonotic disease risk: an ecological primer. 30(4):205–214. ISSN 1471-4922, 1471-5007. doi: 10.1016/j.pt.2014.02.003. URL https://www.cell.com/trends/parasitology/abstract/S1471-4922(14)00032-4. Publisher: Elsevier.

[9] Edward C. Holmes. Molecular clocks and the puzzle of RNA virus origins. 77(7):3893–3897. doi: 10.1128/JVI.77.7.3893-3897.2003. URL https://journals.asm.org/doi/full/10.1128/JVI.77.7.3893-3897.2003. Publisher: American Society for Microbiology.

[10] Gareth M. Jenkins, Andrew Rambaut, Oliver G. Pybus, and Edward C. Holmes. Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. 54(2):156–165. ISSN 1432-1432. doi: 10.1007/s00239-001-0064-3. URL https://doi.org/10.1007/s00239-001-0064-3.

[11] Kousuke Hanada, Yoshiyuki Suzuki, and Takashi Gojobori. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. 21(6):1074–1080. ISSN 0737-4038. doi: 10.1093/molbev/msh109. URL https://doi.org/10.1093/molbev/msh109.

[12] Martin D. Smith, Joel O. Wertheim, Steven Weaver, Ben Murrell, Konrad Scheffler, and Sergei L. Kosakovsky Pond. Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. 32(5):1342–1353, . ISSN 0737-4038. doi: 10.1093/molbev/msv022. URL https://doi.org/10.1093/molbev/msv022.

[13] Erik Volz, Swapnil Mishra, Meera Chand, Jeffrey C. Barrett, Robert Johnson, Lily Geidelberg, Wes R. Hinsley, Daniel J. Laydon, Gavin Dabrera, Áine O'Toole, Roberto Amato, Manon Ragonnet-Cronin, Ian Harrison, Ben Jackson, Cristina V. Ariani, Olivia Boyd, Nicholas J. Loman, John T. McCrone, Sónia Gonçalves, David Jorgensen, Richard Myers, Verity Hill, David K. Jackson, Katy Gaythorpe, Natalie Groves, John Sillitoe, Dominic P. Kwiatkowski, The COVID-19 Genomics UK (COG-UK) Consortium, Seth Flaxman, Oliver Ratmann, Samir Bhatt, Susan Hopkins, Axel Gandy, Andrew Rambaut, and Neil M. Ferguson. Transmission of SARS-CoV-2 lineage b.1.1.7 in england: Insights from linking epidemiological and genetic data. page 2020.12.30.20249034, . doi: 10.1101/2020.12.30.20249034. URL https://www.medrxiv.org/content/10.1101/2020.12.30.20249034v2. Publisher: Cold Spring Harbor Laboratory Press.

[14] Yuri I. Wolf, Cecile Viboud, Edward C. Holmes, Eugene V. Koonin, and David J. Lipman. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza a virus. 1(1):34. ISSN 1745-6150. doi: 10.1186/1745-6150-1-34. URL https://doi.org/10.1186/1745-6150-1-34.

[15] Darren P. Martin, Steven Weaver, Houryiah Tegally, Emmanuel James San, Stephen D. Shank, Eduan Wilkinson, Jennifer Giandhari, Sureshnee Naidoo, Yeshnee Pillay, Lavanya

Singh, Richard J. Lessells, NGS-SA, COVID-19 Genomics Uk (cog Uk), Ravindra K. Gupta, Joel O. Wertheim, Anton Nekturenko, Ben Murrell, Gordon W. Harkins, Philippe Lemey, Oscar A. MacLean, David L. Robertson, Tulio de Oliveira, and Sergei L. Kosakovsky Pond. The emergence and ongoing convergent evolution of the n501y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. page 2021.02.23.21252268. doi: 10.1101/2021.02.23.21252268. URL https://www.medrxiv.org/content/10.1101/2021.02.23.21252268v2. Publisher: Cold Spring Harbor Laboratory Press.

[16] Simon D W Frost, Brittany Rife Magalis, and Sergei L Kosakovsky Pond. Neutral theory and rapidly evolving viral pathogens. 35(6):1348–1354. ISSN 0737-4038. doi: 10.1093/molbev/msy088. URL https://doi.org/10.1093/molbev/msy088.

[17] Oliver G. Pybus, Andrew Rambaut, Robert Belshaw, Robert P. Freckleton, Alexei J. Drummond, and Edward C. Holmes. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. 24(3):845–852. ISSN 0737-4038. doi: 10.1093/molbev/msm001. URL https://doi.org/10.1093/molbev/msm001.

[18] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. 37(5):1530–1534. ISSN 0737-4038. doi: 10.1093/molbev/msaa015. URL https://doi.org/10.1093/molbev/msaa015.

[19] Philippe Lemey, Nick Ruktanonchai, Samuel L. Hong, Vittoria Colizza, Chiara Poletto, Frederik Van den Broeck, Mandev S. Gill, Xiang Ji, Anthony Levasseur, Bas B. Oude Munnink, Marion Koopmans, Adam Sadilek, Shengjie Lai, Andrew J. Tatem, Guy Baele, Marc A. Suchard, and Simon Dellicour. Untangling introductions and persistence in COVID-19 resurgence in europe. 595(7869):713–717, . ISSN 1476-4687. doi: 10.1038/s41586-021-03754-2. URL https://www.nature.com/articles/s41586-021-03754-2. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7869 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Epidemiology;Phylogenetics;SARS-CoV-2 Subject_term_id: epidemiology;phylogenetics;sars-cov-2.

[20] Erik Volz. Fitness, growth and transmissibility of SARS-CoV-2 genetic variants. pages 1–11. ISSN 1471-0064. doi: 10.1038/s41576-023-00610-z. URL https://www.nature.com/articles/s41576-023-00610-z. Publisher: Nature Publishing Group.

[21] Nicole L. Washington, Karthik Gangavarapu, Mark Zeller, Alexandre Bolze, Elizabeth T. Cirulli, Kelly M. Schiabor Barrett, Brendan B. Larsen, Catelyn Anderson, Simon White, Tyler Cassens, Sharoni Jacobs, Geraint Levan, Jason Nguyen, Jimmy M. Ramirez, Charlotte Rivera-Garcia, Efren Sandoval, Xueqing Wang, David Wong, Emily Spencer, Refugio Robles-Sikisaka, Ezra Kurzban, Laura D. Hughes, Xianding Deng, Candace Wang, Venice Servellita, Holly Valentine, Peter De Hoff, Phoebe Seaver, Shashank Sathe, Kimberly Gietzen, Brad Sickler, Jay Antico, Kelly Hoon, Jingtao Liu, Aaron Harding, Omid Bakhtar, Tracy Basler, Brett Austin, Duncan MacCannell, Magnus Isaksson, Phillip G. Febbo, David Becker, Marc Laurent, Eric McDonald, Gene W. Yeo, Rob Knight, Louise C. Laurent, Eileen de Feo, Michael Worobey, Charles Y. Chiu, Marc A. Suchard, James T. Lu, William

Lee, and Kristian G. Andersen. Emergence and rapid transmission of SARS-CoV-2 b.1.1.7 in the united states. 184(10):2587–2594.e7. ISSN 0092-8674. doi: 10.1016/j.cell.2021.03.052. URL https://www.sciencedirect.com/science/article/pii/S0092867421003834.

[22] Anthony P. West, Joel O. Wertheim, Jade C. Wang, Tetyana I. Vasylyeva, Jennifer L. Havens, Moinuddin A. Chowdhury, Edimarlyn Gonzalez, Courtney E. Fang, Steve S. Di Lonardo, Scott Hughes, Jennifer L. Rakeman, Henry H. Lee, Christopher O. Barnes, Priyanthi N. P. Gnanapragasam, Zhi Yang, Christian Gaebler, Marina Caskey, Michel C. Nussenzweig, Jennifer R. Keeffe, and Pamela J. Bjorkman. Detection and characterization of the SARS-CoV-2 lineage b.1.526 in new york. 12(1):4886, . ISSN 2041-1723. doi: 10.1038/s41467-021-25168-4. URL https://www.nature.com/articles/s41467-021-25168-4. Publisher: Nature Publishing Group.

[23] Heather M. Machkovech, Anne M. Hahn, Jacqueline Garonzik Wang, Nathan D. Grubaugh, Peter J. Halfmann, Marc C. Johnson, Jacob E. Lemieux, David H. O'Connor, Anne Piantadosi, Wanting Wei, and Thomas C. Friedrich. Persistent SARS-CoV-2 infection: significance and implications. 0(0). ISSN 1473-3099, 1474-4457. doi: 10.1016/S1473-3099(23)00815-0. URL https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(23)00815-0/fulltext. Publisher: Elsevier.

[24] McKenzie N. Sparrer, Natasha F. Hodges, Tyler Sherman, Susan VandeWoude, Angela M. Bosco-Lauth, and Christie E. Mayo. Role of spillover and spillback in SARS-CoV-2 transmission and the importance of one health in understanding the dynamics of the COVID-19 pandemic. 61(7):e01610–22. doi: 10.1128/jcm.01610-22. URL https://journals.asm.org/doi/full/10.1128/jcm.01610-22. Publisher: American Society for Microbiology.

[25] James O. Lloyd-Smith, Dylan George, Kim M. Pepin, Virginia E. Pitzer, Juliet R. C. Pulliam, Andrew P. Dobson, Peter J. Hudson, and Bryan T. Grenfell. Epidemic dynamics at the human-animal interface. 326(5958):1362–1367. doi: 10.1126/science.1177345. URL https://www.science.org/doi/10.1126/science.1177345. Publisher: American Association for the Advancement of Science.

[26] Nathan D. Wolfe, Claire Panosian Dunavan, and Jared Diamond. Origins of major human infectious diseases. 447(7142):279–283. ISSN 1476-4687. doi: 10.1038/nature05775. URL https://www.nature.com/articles/nature05775. Publisher: Nature Publishing Group.

[27] Joel O. Wertheim, Reilly Hostager, Diane Ryu, Kevin Merkel, Samuel Angedakin, Mimi Arandjelovic, Emmanuel Ayuk Ayimisin, Fred Babweteera, Mattia Bessone, Kathryn J. Brun-Jeffery, Paula Dieguez, Winnie Eckardt, Barbara Fruth, Ilka Herbinger, Sorrel Jones, Hjalmar Kuehl, Kevin E. Langergraber, Kevin Lee, Nadege F. Madinda, Sonja Metzger, Lucy Jayne Ormsby, Martha M. Robbins, Volker Sommer, Tara Stoinski, Erin G. Wessling, Roman M. Wittig, Yisa Ginath Yuh, Fabian H. Leendertz, and Sébastien Calvignac-Spencer. Discovery of novel herpes simplexviruses in wild gorillas, bonobos, and chimpanzees supports zoonotic origin of hsv-2. *Molecular Biology and Evolution*, 38(7):2818–2830, 03 2021. ISSN 1537-1719. doi: 10.1093/molbev/msab072. URL https://doi.org/10.1093/molbev/msab072.

[28] Ana S. Gonzalez-Reiche, Matthew M. Hernandez, Mitchell J. Sullivan, Brianne Ciferri, Hala Alshammary, Ajay Obla, Shelcie Fabre, Giulio Kleiner, Jose Polanco, Zenab Khan, Bremy Alburquerque, Adriana van de Guchte, Jayeeta Dutta, Nancy Francoeur, Betsaida Salom Melo, Irina Oussenko, Gintaras Deikus, Juan Soto, Shwetha Hara Sridhar, Ying-Chih Wang, Kathryn Twyman, Andrew Kasarskis, Deena R. Altman, Melissa Smith, Robert Sebra, Judith Aberg, Florian Krammer, Adolfo García-Sastre, Marta Luksza, Gopi Patel, Alberto Paniz-Mondolfi, Melissa Gitman, Emilia Mia Sordillo, Viviana Simon, and Harm van Bakel. Introductions and early spread of SARS-CoV-2 in the new york city area. 369 (6501):297–301. doi: 10.1126/science.abc1917. URL https://www.science.org/doi/full/10.1126/science.abc1917. Publisher: American Association for the Advancement of Science.

[29] Corinne N. Thompson. COVID-19 outbreak — new york city, february 29–june 1, 2020. 69, . ISSN 0149-21951545-861X. doi: 10.15585/mmwr.mm6946a2. URL https://www.cdc.gov/mmwr/volumes/69/wr/mm6946a2.htm.

[30] William Mude, Victor M Oguoma, Tafadzwa Nyanhanda, Lillian Mwanri, and Carolyne Njue. Racial disparities in COVID-19 pandemic cases, hospitalisations, and deaths: A systematic review and meta-analysis. 11:05015. ISSN 2047-2978. doi: 10.7189/jogh.11.05015. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8248751/.

[31] Corinne N. Thompson. Rapid emergence and epidemiologic characteristics of the SARS-CoV-2 b.1.526 variant — new york city, new york, january 1–april 5, 2021. 70, . ISSN 0149-21951545-861X. doi: 10.15585/mmwr.mm7019e1. URL https://www.cdc.gov/mmwr/volumes/70/wr/mm7019e1.htm.

[32] Elizabeth Luoma. Notes from the field: Epidemiologic characteristics of SARS-CoV-2 recombinant variant XBB.1.5 — new york city, november 1, 2022–january 4, 2023. 72. ISSN 0149-21951545-861X. doi: 10.15585/mmwr.mm7208a4. URL https://www.cdc.gov/mmwr/volumes/72/wr/mm7208a4.htm.

[33] Mary E. Petrone, Jessica E. Rothman, Mallery I. Breban, Isabel M. Ott, Alexis Russell, Erica Lasek-Nesselquist, Hamada Badr, Kevin Kelly, Greg Omerza, Nicholas Renzette, Anne E. Watkins, Chaney C. Kalinich, Tara Alpert, Anderson F. Brito, Rebecca Earnest, Irina R. Tikhonova, Christopher Castaldi, John P. Kelly, Matthew Shudt, Jonathan Plitnick, Erasmus Schneider, Steven Murphy, Caleb Neal, Eva Laszlo, Ahmad Altajar, Claire Pearson, Anthony Muyombwe, Randy Downing, Jafar Razeq, Linda Niccolai, Madeline S. Wilson, Margaret L. Anderson, Jianhui Wang, Chen Liu, Pei Hui, Shrikant Mane, Bradford P. Taylor, William P. Hanage, Marie L. Landry, David R. Peaper, Kaya Bilguvar, Joseph R. Fauver, Chantal B. F. Vogels, Lauren M. Gardner, Virginia E. Pitzer, Kirsten St. George, Mark D. Adams, and Nathan D. Grubaugh. Combining genomic and epidemiological data to compare the transmissibility of SARS-CoV-2 variants alpha and iota. 5(1):1–10. ISSN 2399-3642. doi: 10.1038/s42003-022-03347-3. URL https://www.nature.com/articles/s42003-022-03347-3. Publisher: Nature Publishing Group.

[34] Walker A. Sarah, Vihta Karina-Doris, Gethings Owen, Pritchard Emma, Jones Joel, House Thomas, Bell Iain, Bell John I., Newton John N., Farrar Jeremy, Diamond Ian, Studley Ruth,

Rourke Emma, Hay Jodie, Hopkins Susan, Crook Derrick, Peto Tim, Matthews Philippa C., Eyre David W., Stoesser Nicole, and Pouwels Koen B. Tracking the emergence of SARS-CoV-2 alpha variant in the united kingdom. 385(27):2582–2585. doi: 10.1056/NEJMc2103227. URL https://www.nejm.org/doi/full/10.1056/NEJMc2103227. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMc2103227.

[35] Fang Tian, Bei Tong, Liang Sun, Shengchao Shi, Bin Zheng, Zibin Wang, Xianchi Dong, and Peng Zheng. N501y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2. 10:e69091. ISSN 2050-084X. doi: 10.7554/eLife.69091. URL https://doi.org/10.7554/eLife.69091. Publisher: eLife Sciences Publications, Ltd.

[36] Alexis Russell, Collin O'Connor, Erica Lasek-Nesselquist, Jonathan Plitnick, John P. Kelly, Daryl M. Lamson, and Kirsten St. George. Spatiotemporal analyses of 2 co-circulating SARS-CoV-2 variants, new york state, USA. 28(3):650–659. ISSN 1080-6040, 1080-6059. doi: 10.3201/eid2803.211972. URL https://wwwnc.cdc.gov/eid/article/28/3/21-1972_article.htm.

[37] Yatish Turakhia, Bryan Thornlow, Angie S. Hinrichs, Nicola De Maio, Landen Gozashti, Robert Lanfear, David Haussler, and Russell Corbett-Detig. Ultrafast sample placement on existing trees (UShER) empowers real-time phylogenetics for the SARS-CoV-2 pandemic. page 2020.09.26.314971. doi: 10.1101/2020.09.26.314971. URL https://www.biorxiv.org/content/10.1101/2020.09.26.314971v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[38] Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. 33(2):511–518. ISSN 0305-1048. doi: 10.1093/nar/gki198. URL https://doi.org/10.1093/nar/gki198.

[39] NYC DOHMH population estimates, modified from US census bureau interpolated intercensal population estimates, 2000-2021. updated september 2022., .

[40] Marc A Suchard, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond, and Andrew Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. 4(1):vey016. ISSN 2057-1577. doi: 10.1093/ve/vey016. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6007674/.

[41] Verity Hill and Guy Baele. Bayesian estimation of past population dynamics in BEAST 1.10 using the skygrid coalescent model. 36(11):2620–2628. ISSN 0737-4038. doi: 10.1093/molbev/msz172. URL https://doi.org/10.1093/molbev/msz172.

[42] Andrew Rambaut, Alexei J Drummond, Dong Xie, Guy Baele, and Marc A Suchard. Posterior summarization in bayesian phylogenetics using tracer 1.7. 67(5):901–904. ISSN 1063-5157. doi: 10.1093/sysbio/syy032. URL https://doi.org/10.1093/sysbio/syy032.

[43] Philippe Lemey, Samuel L. Hong, Verity Hill, Guy Baele, Chiara Poletto, Vittoria Colizza, Áine O'Toole, John T. McCrone, Kristian G. Andersen, Michael Worobey, Martha I. Nelson,

Andrew Rambaut, and Marc A. Suchard. Accommodating individual travel history and unsampled diversity in bayesian phylogeographic inference of SARS-CoV-2. 11(1):5110, . ISSN 2041-1723. doi: 10.1038/s41467-020-18877-9. URL https://www.nature.com/articles/s41467-020-18877-9. Publisher: Nature Publishing Group.

[44] Anthony P. West, Christopher O. Barnes, Zhi Yang, and Pamela J. Bjorkman. SARS-CoV-2 lineage b.1.526 emerging in the new york region detected by software utility created to query the spike mutational landscape. page 2021.02.14.431043, . doi: 10.1101/2021.02.14.431043. URL https://www.biorxiv.org/content/10.1101/2021.02.14.431043v2. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[45] Gregory S. Orf, Lester J. Pérez, Karl Ciuoderis, Andrés Cardona, Simón Villegas, Juan P. Hernández-Ortiz, Guy Baele, Aurash Mohaimani, Jorge E. Osorio, Michael G. Berg, and Gavin A. Cloherty. The principles of SARS-CoV-2 intervariant competition are exemplified in the pre-omicron era of the colombian epidemic. 11(3):e05346–22. doi: 10.1128/spectrum.05346-22. URL https://journals.asm.org/doi/10.1128/spectrum.05346-22. Publisher: American Society for Microbiology.

[46] Natalia Vargas-Herrera, Roger V Araujo-Castillo, Orson Mestanza, Marco Galarza, Nancy Rojas-Serrano, and Lely Solari-Zerpa. SARS-CoV-2 lambda and gamma variants competition in peru, a country with high seroprevalence. 6:100112. ISSN 2667-193X. doi: 10.1016/j.lana.2021.100112. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8600335/.

[47] Tetyana I Vasylyeva, Courtney E Fang, Michelle Su, Jennifer L Havens, Edyth Parker, Jade C Wang, Mark Zeller, Anna Yakovleva, Gabriel W Hassler, Moinuddin A Chowdhury, Kristian G Andersen, Scott Hughes, and Joel O Wertheim. Introduction and establishment of SARS-CoV-2 gamma variant in new york city in early 2021. 226(12):2142–2149. ISSN 0022-1899. doi: 10.1093/infdis/jiac265. URL https://doi.org/10.1093/infdis/jiac265.

[48] Felipe Gomes Naveca, Valdinete Nascimento, Victor Costa de Souza, André de Lima Corado, Fernanda Nascimento, George Silva, Ágatha Costa, Débora Duarte, Karina Pessoa, Matilde Mejía, Maria Júlia Brandão, Michele Jesus, Luciana Gonçalves, Cristiano Fernandes da Costa, Vanderson Sampaio, Daniel Barros, Marineide Silva, Tirza Mattos, Gemilson Pontes, Ligia Abdalla, João Hugo Santos, Ighor Arantes, Filipe Zimmer Dezordi, Marilda Mendonça Siqueira, Gabriel Luz Wallau, Paola Cristina Resende, Edson Delatorre, Tiago Gräf, and Gonzalo Bello. COVID-19 in amazonas, brazil, was driven by the persistence of endemic lineages and p.1 emergence. 27(7):1230–1238. ISSN 1546-170X. doi: 10.1038/s41591-021-01378-7. URL https://www.nature.com/articles/s41591-021-01378-7. Publisher: Nature Publishing Group.

[49] Nuno R. Faria, Thomas A. Mellan, Charles Whittaker, Ingra M. Claro, Darlan da S. Candido, Swapnil Mishra, Myuki A. E. Crispim, Flavia C. S. Sales, Iwona Hawryluk, John T. McCrone, Ruben J. G. Hulswit, Lucas A. M. Franco, Mariana S. Ramundo, Jaqueline G. de Jesus, Pamela S. Andrade, Thais M. Coletti, Giulia M. Ferreira, Camila A. M. Silva, Erika R. Manuli, Rafael H. M. Pereira, Pedro S. Peixoto, Moritz U. G. Kraemer, Nelson Gaburo, Cecilia da C. Camilo, Henrique Hoeltgebaum, William M. Souza, Esmenia C.

Rocha, Leandro M. de Souza, Mariana C. de Pinho, Leonardo J. T. Araujo, Frederico S. V. Malta, Aline B. de Lima, Joice do P. Silva, Danielle A. G. Zauli, Alessandro C. de S. Ferreira, Ricardo P. Schnekenberg, Daniel J. Laydon, Patrick G. T. Walker, Hannah M. Schlüter, Ana L. P. dos Santos, Maria S. Vidal, Valentina S. Del Caro, Rosinaldo M. F. Filho, Helem M. dos Santos, Renato S. Aguiar, José L. Proença-Modena, Bruce Nelson, James A. Hay, Mélodie Monod, Xenia Miscouridou, Helen Coupland, Raphael Sonabend, Michaela Vollmer, Axel Gandy, Carlos A. Prete, Vitor H. Nascimento, Marc A. Suchard, Thomas A. Bowden, Sergei L. K. Pond, Chieh-Hsi Wu, Oliver Ratmann, Neil M. Ferguson, Christopher Dye, Nick J. Loman, Philippe Lemey, Andrew Rambaut, Nelson A. Fraiji, Maria do P. S. S. Carvalho, Oliver G. Pybus, Seth Flaxman, Samir Bhatt, and Ester C. Sabino. Genomics and epidemiology of the p.1 SARS-CoV-2 lineage in manaus, brazil. 372(6544):815–821. doi: 10.1126/science.abh2644. URL https://www.science.org/doi/10.1126/science.abh2644. Publisher: American Association for the Advancement of Science.

[50] Marta Giovanetti, Vagner Fonseca, Eduan Wilkinson, Houriiyah Tegally, Emmanuel James San, Christian L Althaus, Joilson Xavier, Svetoslav Nanev Slavov, Vincent Louis Viala, Alex Ranieri Jerônimo Lima, Gabriela Ribeiro, Jayme A Souza-Neto, Heidge Fukumasu, Luiz Lehmann Coutinho, Rivaldo Venancio da Cunha, Carla Freitas, Carlos F Campelo de A e Melo, Wildo Navegantes de Araújo, Rodrigo Fabiano Do Carmo Said, Maria Almiron, Tulio de Oliveira, Sandra Coccuzzo Sampaio, Maria Carolina Elias, Dimas Tadeu Covas, Edward C Holmes, José Lourenço, Simone Kashima, and Luiz Carlos Junior de Alcantara. Replacement of the gamma by the delta variant in brazil: Impact of lineage displacement on the ongoing pandemic. 8(1):veac024. ISSN 2057-1577. doi: 10.1093/ve/veac024. URL https://doi.org/10.1093/ve/veac024.

[51] Simon Dellicour, Samuel L. Hong, Verity Hill, Dacia Dimartino, Christian Marier, Paul Zappile, Gordon W. Harkins, Philippe Lemey, Guy Baele, Ralf Duerr, and Adriana Heguy. Variant-specific introduction and dispersal dynamics of SARS-CoV-2 in new york city – from alpha to omicron. 19(4):e1011348, . ISSN 1553-7374. doi: 10.1371/journal.ppat.1011348. URL https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1011348. Publisher: Public Library of Science.

[52] Verity Hill, Louis Du Plessis, Thomas P. Peacock, Dinesh Aggarwal, Rachel Colquhoun, Alesandro M. Carabelli, Nicholas Ellaby, Eileen Gallagher, Natalie Groves, Ben Jackson, J. T. McCrone, Áine O'Toole, Anna Price, Theo Sanderson, Emily Scher, Joel Southgate, Erik Volz, The COVID-19 genomics UK (COG-UK) Consortium, Wendy S. Barclay, Jeffrey C. Barrett, Meera Chand, Thomas Connor, Ian Goodfellow, Ravindra K. Gupta, Ewan M. Harrison, Nicholas Loman, Richard Myers, David L. Robertson, Oliver G. Pybus, and Andrew Rambaut. The origins and molecular evolution of SARS-CoV-2 lineage b.1.1.7 in the UK. URL https://www.biorxiv.org/content/10.1101/2022.03.08.481609v1. Pages: 2022.03.08.481609 Section: New Results.

[53] Xianding Deng, Miguel A. Garcia-Knight, Mir M. Khalid, Venice Servellita, Candace Wang, Mary Kate Morris, Alicia Sotomayor-González, Dustin R. Glasner, Kevin R. Reyes, Amelia S. Gliwa, Nikitha P. Reddy, Claudia Sanchez San Martin, Scot Federman, Jing Cheng, Joanna Balcerek, Jordan Taylor, Jessica A. Streithorst, Steve Miller, Bharath

Sreekumar, Pei-Yi Chen, Ursula Schulze-Gahmen, Taha Y. Taha, Jennifer M. Hayashi, Camille R. Simoneau, G. Renuka Kumar, Sarah McMahon, Peter V. Lidsky, Yinghong Xiao, Peera Hemarajata, Nicole M. Green, Alex Espinosa, Chantha Kath, Monica Haw, John Bell, Jill K. Hacker, Carl Hanson, Debra A. Wadford, Carlos Anaya, Donna Ferguson, Phillip A. Frankino, Haridha Shivram, Liana F. Lareau, Stacia K. Wyman, Melanie Ott, Raul Andino, and Charles Y. Chiu. Transmission, infectivity, and neutralization of a spike l452r SARS-CoV-2 variant. 184(13):3426–3437.e8. ISSN 0092-8674. doi: 10.1016/j.cell.2021.04.025. URL https://www.sciencedirect.com/science/article/pii/S0092867421005055.

[54] Katherine Laiton-Donato, Carlos Franco-Muñoz, Diego A. Álvarez Díaz, Hector Alejandro Ruiz-Moreno, José A. Usme-Ciro, Diego Andrés Prada, Jhonnatan Reales-González, Sheryll Corchuelo, María T. Herrera-Sepúlveda, Julian Naizaque, Gerardo Santamaría, Jorge Rivera, Paola Rojas, Juan Hernández Ortiz, Andrés Cardona, Diana Malo, Franklin Prieto-Alvarado, Fernando Ruiz Gómez, Magdalena Wiesner, Martha Lucia Ospina Martínez, and Marcela Mercado-Reyes. Characterization of the emerging b.1.621 variant of interest of SARS-CoV-2. 95:105038. ISSN 1567-1348. doi: 10.1016/j.meegid.2021.105038. URL https://www.sciencedirect.com/science/article/pii/S1567134821003361.

[55] Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh, Vagner Fonseca, Jennifer Giandhari, Deelan Doolabh, Sureshnee Pillay, Emmanuel James San, Nokukhanya Msomi, Koleka Mlisana, Anne von Gottberg, Sibongile Walaza, Mushal Allam, Arshad Ismail, Thabo Mohale, Allison J. Glass, Susan Engelbrecht, Gert Van Zyl, Wolfgang Preiser, Francesco Petruccione, Alex Sigal, Diana Hardie, Gert Marais, Nei-yuan Hsiao, Stephen Korsman, Mary-Ann Davies, Lynn Tyers, Innocent Mudau, Denis York, Caroline Maslo, Dominique Goedhals, Shareef Abrahams, Oluwakemi Laguda-Akingba, Arghavan Alisoltani-Dehkordi, Adam Godzik, Constantinos Kurt Wibmer, Bryan Trevor Sewell, José Lourenço, Luiz Carlos Junior Alcantara, Sergei L. Kosakovsky Pond, Steven Weaver, Darren Martin, Richard J. Lessells, Jinal N. Bhiman, Carolyn Williamson, and Tulio de Oliveira. Detection of a SARS-CoV-2 variant of concern in south africa. 592(7854):438–443. ISSN 1476-4687. doi: 10.1038/s41586-021-03402-9. URL https://www.nature.com/articles/s41586-021-03402-9. Number: 7854 Publisher: Nature Publishing Group.

[56] John T. McCrone, Verity Hill, Sumali Bajaj, Rosario Evans Pena, Ben C. Lambert, Rhys Inward, Samir Bhatt, Erik Volz, Christopher Ruis, Simon Dellicour, Guy Baele, Alexander E. Zarebski, Adam Sadilek, Neo Wu, Aaron Schneider, Xiang Ji, Jayna Raghwani, Ben Jackson, Rachel Colquhoun, Áine O'Toole, Thomas P. Peacock, Kate Twohig, Simon Thelwall, Gavin Dabrera, Richard Myers, Nuno R. Faria, Carmen Huber, Isaac I. Bogoch, Kamran Khan, Louis du Plessis, Jeffrey C. Barrett, David M. Aanensen, Wendy S. Barclay, Meera Chand, Thomas Connor, Nicholas J. Loman, Marc A. Suchard, Oliver G. Pybus, Andrew Rambaut, and Moritz U. G. Kraemer. Context-specific emergence and growth of the SARS-CoV-2 delta variant. 610(7930):154–160. ISSN 1476-4687. doi: 10.1038/s41586-022-05200-3. URL https://www.nature.com/articles/s41586-022-05200-3. Number: 7930 Publisher: Nature Publishing Group.

[57] Paul Elliott, David Haw, Haowei Wang, Oliver Eales, Caroline E. Walters, Kylie E. C. Ainslie, Christina Atchison, Claudio Fronterre, Peter J. Diggle, Andrew J. Page, Alexander J.

Trotter, Sophie J. Prosolek, The COVID-19 Genomics UK (COG-UK) Consortium, Deborah Ashby, Christl A. Donnelly, Wendy Barclay, Graham Taylor, Graham Cooke, Helen Ward, Ara Darzi, and Steven Riley. Exponential growth, high prevalence of SARS-CoV-2, and vaccine effectiveness associated with the delta variant. 374(6574):eabl9551. doi: 10.1126/science.abl9551. URL https://www.science.org/doi/10.1126/science.abl9551. Publisher: American Association for the Advancement of Science.

[58] Carolina Lucas, Chantal B. F. Vogels, Inci Yildirim, Jessica E. Rothman, Peiwen Lu, Valter Monteiro, Jeff R. Gehlhausen, Melissa Campbell, Julio Silva, Alexandra Tabachnikova, Mario A. Peña-Hernandez, M. Catherine Muenker, Mallery I. Breban, Joseph R. Fauver, Subhasis Mohanty, Jiefang Huang, Albert C. Shaw, Albert I. Ko, Saad B. Omer, Nathan D. Grubaugh, and Akiko Iwasaki. Impact of circulating SARS-CoV-2 variants on mRNA vaccine-induced immunity. 600(7889):523–529. ISSN 1476-4687. doi: 10.1038/s41586-021-04085-y. URL https://www.nature.com/articles/s41586-021-04085-y. Publisher: Nature Publishing Group.

[59] Chee Wah Tan, Wan Ni Chia, Feng Zhu, Barnaby E. Young, Napaporn Chantasrisawad, Shi-Hsia Hwa, Aileen Ying-Yan Yeoh, Beng Lee Lim, Wee Chee Yap, Surinder Kaur M. S. Pada, Seow Yen Tan, Watsamon Jantarabenjakul, Lim Kai Toh, Shiwei Chen, Jinyan Zhang, Yun Yan Mah, Vivian Chih-Wei Chen, Mark I.-C. Chen, Supaporn Wacharapluesadee, Alex Sigal, Opass Putcharoen, David Chien Lye, and Lin-Fa Wang. SARS-CoV-2 omicron variant emerged under immune selection. pages 1–6, . ISSN 2058-5276. doi: 10.1038/s41564-022-01246-1. URL https://www.nature.com/articles/s41564-022-01246-1. Publisher: Nature Publishing Group.

[60] Richard A Neher. Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. 8(2):veac113. ISSN 2057-1577. doi: 10.1093/ve/veac113. URL https://doi.org/10.1093/ve/veac113.

[61] Gwanghun Kim, Hyun Mu Shin, Hang-Rae Kim, and Yuseob Kim. Effects of host and pathogenicity on mutation rates in avian influenza a viruses. 8(1):veac013. ISSN 2057-1577. doi: 10.1093/ve/veac013. URL https://doi.org/10.1093/ve/veac013.

[62] Allison L. Hicks and Siobain Duffy. Cell tropism predicts long-term nucleotide substitution rates of mammalian RNA viruses. 10(1):e1003838. ISSN 1553-7374. doi: 10.1371/journal.ppat.1003838. URL https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1003838. Publisher: Public Library of Science.

[63] Alpha Kabinet Keita, Fara R. Koundouno, Martin Faye, Ariane Düx, Julia Hinzmann, Haby Diallo, Ahidjo Ayouba, Frederic Le Marcis, Barré Soropogui, Kékoura Ifono, Moussa M. Diagne, Mamadou S. Sow, Joseph A. Bore, Sebastien Calvignac-Spencer, Nicole Vidal, Jacob Camara, Mamadou B. Keita, Annick Renevey, Amadou Diallo, Abdoul K. Soumah, Saa L. Millimono, Almudena Mari-Saez, Mamadou Diop, Ahmadou Doré, Fodé Y. Soumah, Kaka Kourouma, Nathalie J. Vielle, Cheikh Loucoubar, Ibrahima Camara, Karifa Kourouma, Giuditta Annibaldis, Assaïtou Bah, Anke Thielebein, Meike Pahlmann, Steven T. Pullan, Miles W. Carroll, Joshua Quick, Pierre Formenty, Anais Legand, Karla Pietro, Michael R.

Wiley, Noel Tordo, Christophe Peyrefitte, John T. McCrone, Andrew Rambaut, Youssouf Sidibé, Mamadou D. Barry, Madeleine Kourouma, Cé D. Saouromou, Mamadou Condé, Moussa Baldé, Moriba Povogui, Sakoba Keita, Mandiou Diakite, Mamadou S. Bah, Amadou Sidibe, Dembo Diakite, Fodé B. Sako, Fodé A. Traore, Georges A. Ki-Zerbo, Philippe Lemey, Stephan Günther, Liana E. Kafetzopoulou, Amadou A. Sall, Eric Delaporte, Sophie Duraffour, Ousmane Faye, Fabian H. Leendertz, Martine Peeters, Abdoulaye Toure, and N' Faly Magassouba. Resurgence of ebola virus in 2021 in guinea suggests a new paradigm for outbreaks. 597(7877):539–543. ISSN 1476-4687. doi: 10.1038/s41586-021-03901-9. URL https://www.nature.com/articles/s41586-021-03901-9. Number: 7877 Publisher: Nature Publishing Group.

[64] Edward C Holmes. Molecular clocks and the puzzle of rna virus origins. *Journal of virology*, 77(7):3893–3897, 2003.

[65] Shannon L. M. Whitmer, Jason T. Ladner, Michael R. Wiley, Ketan Patel, Gytis Dudas, Andrew Rambaut, Foday Sahr, Karla Prieto, Samuel S. Shepard, Ellie Carmody, Barbara Knust, Dhamari Naidoo, Gibrilla Deen, Pierre Formenty, Stuart T. Nichol, Gustavo Palacios, and Ute Ströher. Active ebola virus replication and heterogeneous evolutionary rates in EVD survivors. 22(5):1159–1168. ISSN 2211-1247. doi: 10.1016/j.celrep.2018.01.008. URL https://www.sciencedirect.com/science/article/pii/S2211124718300081.

[66] Philippe Lemey, Sergei L. Kosakovsky Pond, Alexei J. Drummond, Oliver G. Pybus, Beth Shapiro, Helena Barroso, Nuno Taveira, and Andrew Rambaut. Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. 3 (2):e29, . ISSN 1553-7358. doi: 10.1371/journal.pcbi.0030029. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0030029. Publisher: Public Library of Science.

[67] Peter V. Markov, Mahan Ghafari, Martin Beer, Katrina Lythgoe, Peter Simmonds, Nikolaos I. Stilianakis, and Aris Katzourakis. The evolution of SARS-CoV-2. 21(6):361–379. ISSN 1740-1534. doi: 10.1038/s41579-023-00878-2. URL https://www.nature.com/articles/s41579-023-00878-2. Publisher: Nature Publishing Group.

[68] Changshuo Wei, Ke-Jia Shan, Weiguang Wang, Shuya Zhang, Qing Huan, and Wenfeng Qian. Evidence for a mouse origin of the SARS-CoV-2 omicron variant. 48(12):1111–1121. ISSN 1673-8527. doi: 10.1016/j.jgg.2021.12.003. URL https://www.sciencedirect.com/science/article/pii/S1673852721003738.

[69] Vanessa L. Hale, Patricia M. Dennis, Dillon S. McBride, Jacqueline M. Nolting, Christopher Madden, Devra Huey, Margot Ehrlich, Jennifer Grieser, Jenessa Winston, Dusty Lombardi, Stormy Gibson, Linda Saif, Mary L. Killian, Kristina Lantz, Rachel M. Tell, Mia Torchetti, Suelee Robbe-Austerman, Martha I. Nelson, Seth A. Faith, and Andrew S. Bowman. SARS-CoV-2 infection in free-ranging white-tailed deer. 602(7897):481–486. ISSN 1476-4687. doi: 10.1038/s41586-021-04353-x. URL https://www.nature.com/articles/s41586-021-04353-x. Number: 7897 Publisher: Nature Publishing Group.

[70] Dillon S. McBride, Sofya K. Garushyants, John Franks, Andrew F. Magee, Steven H. Overend, Devra Huey, Amanda M. Williams, Seth A. Faith, Ahmed Kandeil, Sanja Trifkovic, Lance Miller, Trushar Jeevan, Anami Patel, Jacqueline M. Nolting, Michael J. Tonkovich, J. Tyler Genders, Andrew J. Montoney, Kevin Kasnyik, Timothy J. Linder, Sarah N. Bevins, Julianna B. Lenoch, Jeffrey C. Chandler, Thomas J. DeLiberto, Eugene V. Koonin, Marc A. Suchard, Philippe Lemey, Richard J. Webby, Martha I. Nelson, and Andrew S. Bowman. Accelerated evolution of SARS-CoV-2 in free-ranging white-tailed deer. 14(1):5105. ISSN 2041-1723. doi: 10.1038/s41467-023-40706-y. URL https://www.nature.com/articles/s41467-023-40706-y. Number: 1 Publisher: Nature Publishing Group.

[71] Lok Bahadur Shrestha, Charles Foster, William Rawlinson, Nicodemus Tedla, and Rowena A. Bull. Evolution of the SARS-CoV-2 omicron variants BA.1 to BA.5: Implications for immune escape and transmission. 32(5):e2381. ISSN 1099-1654. doi: 10.1002/rmv.2381. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rmv.2381. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rmv.2381.

[72] John D. O'Brien, Vladimir N. Minin, and Marc A. Suchard. Learning to count: Robust estimates for labeled distances between molecular sequences. 26(4):801–814. ISSN 0737-4038. doi: 10.1093/molbev/msp003. URL https://doi.org/10.1093/molbev/msp003.

[73] Mahan Ghafari, Matthew Hall, Tanya Golubchik, Daniel Ayoubkhani, Thomas House, George MacIntyre-Cockett, Helen R. Fryer, Laura Thomson, Anel Nurtay, Steven A. Kemp, Luca Ferretti, David Buck, Angie Green, Amy Trebes, Paolo Piazza, Lorne J. Lonie, Ruth Studley, Emma Rourke, Darren L. Smith, Matthew Bashton, Andrew Nelson, Matthew Crown, Clare McCann, Gregory R. Young, Rui Andre Nunes dos Santos, Zack Richards, Mohammad Adnan Tariq, Roberto Cahuantzi, Jeff Barrett, Christophe Fraser, David Bonsall, Ann Sarah Walker, and Katrina Lythgoe. Prevalence of persistent SARS-CoV-2 in a large community surveillance study. 626(8001):1094–1101. ISSN 1476-4687. doi: 10.1038/s41586-024-07029-4. URL https://www.nature.com/articles/s41586-024-07029-4. Publisher: Nature Publishing Group.

[74] Aude Jary, Valentin Leducq, Isabelle Malet, Stéphane Marot, Elise Klement-Frutos, Elisa Teyssou, Cathia Soulié, Basma Abdi, Marc Wirden, Valérie Pourcher, Eric Caumes, Vincent Calvez, Sonia Burrel, Anne-Geneviève Marcelin, and David Boutolleau. Evolution of viral quasispecies during SARS-CoV-2 infection. 26(11):1560.e1–1560.e4. ISSN 1198-743X. doi: 10.1016/j.cmi.2020.07.032. URL https://www.sciencedirect.com/science/article/pii/S1198743X20304407.

[75] Majdouline El Moussaoui, Sebastien Bontems, Cecile Meex, Marie-Pierre Hayette, Marie Lejeune, Samuel L. Hong, Simon Dellicour, Michel Moutschen, Nadine Cambisano, Nathalie Renotte, Vincent Bours, Gilles Darcis, Maria Artesi, and Keith Durkin. Intrahost evolution leading to distinct lineages in the upper and lower respiratory tracts during SARS-CoV-2 prolonged infection. URL https://www.medrxiv.org/content/10.1101/2024.03.03.24303157v1. Pages: 2024.03.03.24303157.

[76] Martina Rueca, Barbara Bartolini, Cesare Ernesto Maria Gruber, Antonio Piralla, Fausto

Baldanti, Emanuela Giombini, Francesco Messina, Luisa Marchioni, Giuseppe Ippolito, Antonino Di Caro, and Maria Rosaria Capobianchi. Compartmentalized replication of SARS-cov-2 in upper vs. lower respiratory tract assessed by whole genome quasispecies analysis. 8(9):1302. ISSN 2076-2607. doi: 10.3390/microorganisms8091302. URL https://www.mdpi.com/2076-2607/8/9/1302. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[77] Jolien Van Cleemput, Willem van Snippenberg, Laurens Lambrechts, Amélie Dendooven, Valentino D'Onofrio, Liesbeth Couck, Wim Trypsteen, Jan Vanrusselt, Sebastiaan Theuns, Nick Vereecke, Thierry P. P. van den Bosch, Martin Lammens, Ann Driessen, Ruth Achten, Ken R. Bracke, Wim Van den Broeck, Jan Von der Thüsen, Hans Nauwynck, Jo Van Dorpe, Sarah Gerlo, Piet Maes, Janneke Cox, and Linos Vandekerckhove. Organ-specific genome diversity of replication-competent SARS-CoV-2. 12:6612. ISSN 2041-1723. doi: 10.1038/s41467-021-26884-7. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8595628/.

[78] Manli Wang, Ruiyuan Cao, Leike Zhang, Xinglou Yang, Jia Liu, Mingyue Xu, Zhengli Shi, Zhihong Hu, Wu Zhong, and Gengfu Xiao. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. 30(3):269–271, . ISSN 1748-7838. doi: 10.1038/s41422-020-0282-0. URL https://www.nature.com/articles/s41422-020-0282-0. Publisher: Nature Publishing Group.

[79] Adriano de Bernardi Schneider, Michelle Su, Angie S Hinrichs, Jade Wang, Helly Amin, John Bell, Debra A Wadford, Áine O'Toole, Emily Scher, Marc D Perry, Yatish Turakhia, Nicola De Maio, Scott Hughes, and Russ Corbett-Detig. SARS-CoV-2 lineage assignments using phylogenetic placement/UShER are superior to pangoLEARN machine-learning method. 10(1):vead085. ISSN 2057-1577. doi: 10.1093/ve/vead085. URL https://doi.org/10.1093/ve/vead085.

[80] Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research*, 33(2): 511–518, 2005.

[81] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37 (5):1530–1534, 2020.

[82] Ben Murrell, Steven Weaver, Martin D. Smith, Joel O. Wertheim, Sasha Murrell, Anthony Aylward, Kemal Eren, Tristan Pollner, Darren P. Martin, Davey M. Smith, Konrad Scheffler, and Sergei L. Kosakovsky Pond. Gene-wide identification of episodic selection. 32(5): 1365–1371. ISSN 0737-4038. doi: 10.1093/molbev/msv035. URL https://doi.org/10.1093/molbev/msv035.

[83] Sylvain Baize, Delphine Pannetier, Lisa Oestereich, Toni Rieger, Lamine Koivogui, N'Faly Magassouba, Barrè Soropogui, Mamadou Saliou Sow, Sakoba Keïta, Hilde De Clerck, Amanda Tiffany, Gemma Dominguez, Mathieu Loua, Alexis Traoré, Moussa Kolié,

Emmanuel Roland Malano, Emmanuel Heleze, Anne Bocquin, Stephane Mély, Hervé Raoul, Valérie Caro, Dániel Cadar, Martin Gabriel, Meike Pahlmann, Dennis Tappe, Jonas Schmidt-Chanasit, Benido Impouma, Abdoul Karim Diallo, Pierre Formenty, Michel Van Herp, and Stephan Günther. Emergence of zaire ebola virus disease in guinea. 371(15):1418–1425. ISSN 0028-4793. doi: 10.1056/NEJMoa1404505. URL https://www.nejm.org/doi/10.1056/NEJMoa1404505. Publisher: Massachusetts Medical Society.

[84] Simon Cauchemez, Pierre Nouvellet, Anne Cori, Thibaut Jombart, Tini Garske, Hannah Clapham, Sean Moore, Harriet Linden Mills, Henrik Salje, Caitlin Collins, Isabel Rodriquez-Barraquer, Steven Riley, Shaun Truelove, Homoud Algarni, Rafat Alhakeem, Khalid AlHarbi, Abdulhafiz Turkistani, Ricardo J. Aguas, Derek A. T. Cummings, Maria D. Van Kerkhove, Christl A. Donnelly, Justin Lessler, Christophe Fraser, Ali Al-Barrak, and Neil M. Ferguson. Unraveling the drivers of MERS-CoV transmission. 113(32):9081–9086. doi: 10.1073/pnas.1519235113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1519235113. Publisher: Proceedings of the National Academy of Sciences.

[85] Feng Gao, Elizabeth Bailes, David L. Robertson, Yalu Chen, Cynthia M. Rodenburg, Scott F. Michael, Larry B. Cummins, Larry O. Arthur, Martine Peeters, George M. Shaw, Paul M. Sharp, and Beatrice H. Hahn. Origin of HIV-1 in the chimpanzee pan troglodytes troglodytes. 397(6718):436–441. ISSN 1476-4687. doi: 10.1038/17130. URL https://www.nature.com/articles/17130. Publisher: Nature Publishing Group.

[86] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, Zhenshun Cheng, Ting Yu, Jiaan Xia, Yuan Wei, Wenjuan Wu, Xuelei Xie, Wen Yin, Hui Li, Min Liu, Yan Xiao, Hong Gao, Li Guo, Jungang Xie, Guangfa Wang, Rongmeng Jiang, Zhancheng Gao, Qi Jin, Jianwei Wang, and Bin Cao. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. 395(10223):497–506. ISSN 0140-6736. doi: 10.1016/S0140-6736(20)30183-5. URL https://www.sciencedirect.com/science/article/pii/S0140673620301835.

[87] Joana Isidro, Vítor Borges, Miguel Pinto, Daniel Sobral, João Dourado Santos, Alexandra Nunes, Verónica Mixão, Rita Ferreira, Daniela Santos, Silvia Duarte, Luís Vieira, Maria José Borrego, Sofia Núncio, Isabel Lopes de Carvalho, Ana Pelerito, Rita Cordeiro, and João Paulo Gomes. Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus. 28(8):1569–1572. ISSN 1546-170X. doi: 10.1038/s41591-022-01907-y. URL https://www.nature.com/articles/s41591-022-01907-y. Publisher: Nature Publishing Group.

[88] J. S. M. Peiris, S. T. Lai, L. L. M. Poon, Y. Guan, L. Y. C. Yam, W. Lim, J. Nicholls, W. K. S. Yee, W. W. Yan, M. T. Cheung, V. C. C. Cheng, K. H. Chan, D. N. C. Tsang, R. W. H. Yung, T. K. Ng, and K. Y. Yuen. Coronavirus as a possible cause of severe acute respiratory syndrome. 361(9366):1319–1325. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(03)13077-2. URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(03)13077-2/fulltext#%20. Publisher: Elsevier.

[89] Eefje JA Schrauwen and Ron AM Fouchier. Host adaptation and transmission of influenza a viruses in mammals. 3(1):1–10. ISSN null. doi: 10.1038/emi.2014.9. URL https://doi.org/10.1038/emi.2014.9. Publisher: Taylor & Francis _eprint: https://doi.org/10.1038/emi.2014.9.

[90] Jonathan S. Towner, Marina L. Khristova, Tara K. Sealy, Martin J. Vincent, Bobbie R. Erickson, Darcy A. Bawiec, Amy L. Hartman, James A. Comer, Sherif R. Zaki, Ute Ströher, Filomena Gomes da Silva, Fernando del Castillo, Pierre E. Rollin, Thomas G. Ksiazek, and Stuart T. Nichol. Marburgvirus genomics and association with a large hemorrhagic fever outbreak in angola. 80(13):6497–6516. ISSN 0022-538X. doi: 10.1128/JVI.00069-06. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1488971/.

[91] Wenhui Li, Chengsheng Zhang, Jianhua Sui, Jens H Kuhn, Michael J Moore, Shiwen Luo, Swee-Kee Wong, I-Chueh Huang, Keming Xu, Natalya Vasilieva, Akikazu Murakami, Yaqing He, Wayne A Marasco, Yi Guan, Hyeryun Choe, and Michael Farzan. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. 24(8):1634–1643. ISSN 0261-4189. doi: 10.1038/sj.emboj.7600640. URL https://www.embopress.org/doi/full/10.1038/sj.emboj.7600640. Num Pages: 1643 Publisher: John Wiley & Sons, Ltd.

[92] Xiu-Xia Qu, Pei Hao, Xi-Jun Song, Si-Ming Jiang, Yan-Xia Liu, Pei-Gang Wang, Xi Rao, Huai-Dong Song, Sheng-Yue Wang, Yu Zuo, Ai-Hua Zheng, Min Luo, Hua-Lin Wang, Fei Deng, Han-Zhong Wang, Zhi-Hong Hu, Ming-Xiao Ding, Guo-Ping Zhao, and Hong-Kui Deng. Identification of two critical amino acid residues of the severe acute respiratory syndrome coronavirus spike protein for its variation in zoonotic tropism transition via a double substitution strategy *. 280(33):29588–29595. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M500662200. URL https://www.jbc.org/article/S0021-9258(20)56384-9/abstract. Publisher: Elsevier.

[93] The Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in china. 303(5664):1666–1669. doi: 10.1126/science.1092002. URL https://www.science.org/doi/full/10.1126/science.1092002. Publisher: American Association for the Advancement of Science.

[94] Richard A. Urbanowicz, C. Patrick McClure, Anavaj Sakuntabhai, Amadou A. Sall, Gary Kobinger, Marcel A. Müller, Edward C. Holmes, Félix A. Rey, Etienne Simon-Loriere, and Jonathan K. Ball. Human adaptation of ebola virus during the west african outbreak. 167(4):1079–1087.e5. ISSN 0092-8674. doi: 10.1016/j.cell.2016.10.013. URL https://www.sciencedirect.com/science/article/pii/S0092867416313964.

[95] Bette Korber, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, Elena E. Giorgi, Tanmoy Bhattacharya, Brian Foley, Kathryn M. Hastie, Matthew D. Parker, David G. Partridge, Cariad M. Evans, Timothy M. Freeman, Thushan I. de Silva, Adrienne Angyal, Rebecca L. Brown, Laura Carrilero, Luke R. Green, Danielle C. Groves, Katie J. Johnson, Alexander J. Keeley, Benjamin B. Lindsey, Paul J. Parsons, Mohammad Raza, Sarah Rowland-Jones, Nikki Smith, Rachel M. Tucker, Dennis Wang, Matthew D. Wyles, Charlene McDanal, Lautaro G. Perez, Haili Tang, Alex Moon-Walker, Sean P. Whelan, Celia C. LaBranche, Erica O. Saphire, and

David C. Montefiori. Tracking changes in SARS-CoV-2 spike: Evidence that d614g increases infectivity of the COVID-19 virus. 182(4):812–827.e19. ISSN 0092-8674. doi: 10.1016/j.cell.2020.06.043. URL https://www.sciencedirect.com/science/article/pii/S0092867420308205.

[96] Erik Volz, Verity Hill, John T. McCrone, Anna Price, David Jorgensen, Áine O'Toole, Joel Southgate, Robert Johnson, Ben Jackson, Fabricia F. Nascimento, Sara M. Rey, Samuel M. Nicholls, Rachel M. Colquhoun, Ana da Silva Filipe, James Shepherd, David J. Pascall, Rajiv Shah, Natasha Jesudason, Kathy Li, Ruth Jarrett, Nicole Pacchiarini, Matthew Bull, Lily Geidelberg, Igor Siveroni, Cherian Koshy, Emma Wise, Nick Cortes, Jessica Lynch, Stephen Kidd, Matilde Mori, Derek J. Fairley, Tanya Curran, James P. McKenna, Helen Adams, Christophe Fraser, Tanya Golubchik, David Bonsall, Catrin Moore, Sarah L. Caddy, Fahad A. Khokhar, Michelle Wantoch, Nicola Reynolds, Ben Warne, Joshua Maksimovic, Karla Spellman, Kathryn McCluggage, Michaela John, Robert Beer, Safiah Afifi, Sian Morgan, Angela Marchbank, Anna Price, Christine Kitchen, Huw Gulliver, Ian Merrick, Joel Southgate, Martyn Guest, Robert Munn, Trudy Workman, Thomas R. Connor, William Fuller, Catherine Bresner, Luke B. Snell, Themoula Charalampous, Gaia Nebbia, Rahul Batra, Jonathan Edgeworth, Samuel C. Robson, Angela Beckett, Katie F. Loveson, David M. Aanensen, Anthony P. Underwood, Corin A. Yeats, Khalil Abudahab, Ben E. W. Taylor, Mirko Menegazzo, Gemma Clark, Wendy Smith, Manjinder Khakh, Vicki M. Fleming, Michelle M. Lister, Hannah C. Howson-Wells, Louise Berry, Tim Boswell, Amelia Joseph, Iona Willingham, Paul Bird, Thomas Helmer, Karlie Fallon, Christopher Holmes, Julian Tang, Veena Raviprakash, Sharon Campbell, Nicola Sheriff, Matthew W. Loose, Nadine Holmes, Christopher Moore, Matthew Carlile, Victoria Wright, Fei Sang, Johnny Debebe, Francesc Coll, Adrian W. Signell, Gilberto Betancor, Harry D. Wilson, Theresa Feltwell, Charlotte J. Houldcroft, Sahar Eldirdiri, Anita Kenyon, Thomas Davis, Oliver Pybus, Louis du Plessis, Alex Zarebski, Jayna Raghwani, Moritz Kraemer, Sarah Francois, Stephen Attwood, Tetyana Vasylyeva, M. Estee Torok, William L. Hamilton, Ian G. Goodfellow, Grant Hall, Aminu S. Jahun, Yasmin Chaudhry, Myra Hosmillo, Malte L. Pinckert, Iliana Georgana, Anna Yakovleva, Luke W. Meredith, Samuel Moses, Hannah Lowe, Felicity Ryan, Chloe L. Fisher, Ali R. Awan, John Boyes, Judith Breuer, Kathryn Ann Harris, Julianne Rose Brown, Divya Shah, Laura Atkinson, Jack C. D. Lee, Adela Alcolea-Medina, Nathan Moore, Nicholas Cortes, Rebecca Williams, Michael R. Chapman, Lisa J. Levett, Judith Heaney, Darren L. Smith, Matthew Bashton, Gregory R. Young, John Allan, Joshua Loh, Paul A. Randell, Alison Cox, Pinglawathee Madona, Alison Holmes, Frances Bolt, James Price, Siddharth Mookerjee, Aileen Rowan, Graham P. Taylor, Manon Ragonnet-Cronin, Fabricia F. Nascimento, David Jorgensen, Igor Siveroni, Rob Johnson, Olivia Boyd, Lily Geidelberg, Erik M. Volz, Kirstyn Brunker, Katherine L. Smollett, Nicholas J. Loman, Joshua Quick, Claire McMurray, Joanne Stockton, Sam Nicholls, Will Rowe, Radoslaw Poplawski, Rocio T. Martinez-Nunez, Jenifer Mason, Trevor I. Robinson, Elaine O'Toole, Joanne Watts, Cassie Breen, Angela Cowell, Catherine Ludden, Graciela Sluga, Nicholas W. Machin, Shazaad S. Y. Ahmad, Ryan P. George, Fenella Halstead, Venkat Sivaprakasam, Emma C. Thomson, James G. Shepherd, Patawee Asamaphan, Marc O. Niebel, Kathy K. Li, Rajiv N. Shah, Natasha G. Jesudason, Yasmin A. Parr, Lily Tong, Alice Broos, Daniel Mair, Jenna Nichols, Stephen N. Carmichael, Kyriaki Nomikou, Elihu Aranday-Cortes, Natasha Johnson, Igor Starinskij, Ana da Silva Filipe, David L. Robertson, Richard J. Orton, Joseph Hughes, Sreenu Vattipally, Joshua B. Singer,

Antony D. Hale, Louissa R. Macfarlane-Smith, Katherine L. Harper, Yusri Taha, Brendan A. I. Payne, Shirelle Burton-Fanning, Sheila Waugh, Jennifer Collins, Gary Eltringham, Kate E. Templeton, Martin P. McHugh, Rebecca Dewar, Elizabeth Wastenge, Samir Dervisevic, Rachael Stanley, Reenesh Prakash, Claire Stuart, Ngozi Elumogo, Dheeraj K. Sethi, Emma J. Meader, Lindsay J. Coupland, Will Potter, Clive Graham, Edward Barton, Debra Padgett, Garren Scott, Emma Swindells, Jane Greenaway, Andrew Nelson, Wen C. Yew, Paola C. Resende Silva, Monique Andersson, Robert Shaw, Timothy Peto, Anita Justice, David Eyre, Derrick Crooke, Sarah Hoosdally, Tim J. Sloan, Nichola Duckworth, Sarah Walsh, Anoop J. Chauhan, Sharon Glaysher, Kelly Bicknell, Sarah Wyllie, Ethan Butcher, Scott Elliott, Allyson Lloyd, Robert Impey, Nick Levene, Lynn Monaghan, Declan T. Bradley, Elias Allara, Clare Pearson, Peter Muir, Ian B. Vipond, Richard Hopes, Hannah M. Pymont, Stephanie Hutchings, Martin D. Curran, Surendra Parmar, Angie Lackenby, Tamyo Mbisa, Steven Platt, Shahjahan Miah, David Bibby, Carmen Manso, Jonathan Hubb, Meera Chand, Gavin Dabrera, Mary Ramsay, Daniel Bradshaw, Alicia Thornton, Richard Myers, Ulf Schaefer, Natalie Groves, Eileen Gallagher, David Lee, David Williams, Nicholas Ellaby, Ian Harrison, Hassan Hartman, Nikos Manesis, Vineet Patel, Chloe Bishop, Vicki Chalker, Husam Osman, Andrew Bosworth, Esther Robinson, Matthew T. G. . Evaluating the effects of SARS-CoV-2 spike mutation d614g on transmissibility and pathogenicity. 184(1):64–75.e11, . ISSN 0092-8674. doi: 10.1016/j.cell.2020.11.020. URL https://www.sciencedirect.com/science/article/pii/S0092867420315373.

[97] Oscar A. MacLean, Spyros Lytras, Steven Weaver, Joshua B. Singer, Maciej F. Boni, Philippe Lemey, Sergei L. Kosakovsky Pond, and David L. Robertson. Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. 19(3). ISSN 1544-9173. doi: 10.1371/journal.pbio.3001115. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7990310/.

[98] Cedric C. S. Tan, Lucy Van Dorp, and Francois Balloux. The evolutionary drivers and correlates of viral host jumps. . ISSN 2397-334X. doi: 10.1038/s41559-024-02353-4. URL https://www.nature.com/articles/s41559-024-02353-4.

[99] Michael Worobey, Guan-Zhu Han, and Andrew Rambaut. Genesis and pathogenesis of the 1918 pandemic h1n1 influenza a virus. 111(22):8107–8112. doi: 10.1073/pnas.1324197111. URL https://www.pnas.org/doi/abs/10.1073/pnas.1324197111. Publisher: Proceedings of the National Academy of Sciences.

[100] Austin L. Hughes. Relaxation of purifying selection on the SAD lineage of live attenuated oral vaccines for rabies virus. 9(5):827–831, . ISSN 1567-1348. doi: 10.1016/j.meegid.2009.04.016. URL https://www.sciencedirect.com/science/article/pii/S1567134809000884.

[101] Austin L. Hughes. Relaxation of purifying selection on live attenuated vaccine strains of the family paxamyxoviridae. 27(11):1685–1690, . ISSN 0264-410X. doi: 10.1016/j.vaccine.2009.01.036. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2707849/.

[102] Joel O. Wertheim, Ben Murrell, Martin D. Smith, Sergei L. Kosakovsky Pond, and Konrad Scheffler. RELAX: Detecting relaxed selection in a phylogenetic framework. 32(3):820–832,

. ISSN 0737-4038. doi: 10.1093/molbev/msu400. URL https://doi.org/10.1093/molbev/msu400.

[103] Gavin J. D. Smith, Dhanasekaran Vijaykrishna, Justin Bahl, Samantha J. Lycett, Michael Worobey, Oliver G. Pybus, Siu Kit Ma, Chung Lam Cheung, Jayna Raghwani, Samir Bhatt, J. S. Malik Peiris, Yi Guan, and Andrew Rambaut. Origins and evolutionary genomics of the 2009 swine-origin h1n1 influenza a epidemic. 459(7250):1122–1125, . ISSN 1476-4687. doi: 10.1038/nature08182. URL https://www.nature.com/articles/nature08182. Number: 7250 Publisher: Nature Publishing Group.

[104] Eric M. Leroy, Brice Kumulungui, Xavier Pourrut, Pierre Rouquet, Alexandre Hassanin, Philippe Yaba, André Délicat, Janusz T. Paweska, Jean-Paul Gonzalez, and Robert Swanepoel. Fruit bats as reservoirs of ebola virus. 438(7068):575–576. ISSN 1476-4687. doi: 10.1038/438575a. URL https://www.nature.com/articles/438575a. Publisher: Nature Publishing Group.

[105] Zhengli Shi and Zhihong Hu. A review of studies on animal reservoirs of the SARS coronavirus. 133(1):74–87. ISSN 0168-1702. doi: 10.1016/j.virusres.2007.03.012. URL https://www.sciencedirect.com/science/article/pii/S0168170207001050.

[106] Jonathan E. Pekar, Andrew Magee, Edyth Parker, Niema Moshiri, Katherine Izhikevich, Jennifer L. Havens, Karthik Gangavarapu, Lorena M. 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. SARS-CoV-2 emergence very likely resulted from at least two zoonotic events. . doi: 10.5281/zenodo.6291628. URL https://zenodo.org/record/6291628.

[107] Sarah Temmam, Khamsing Vongphayloth, Eduard Baquero, Sandie Munier, Massimiliano Bonomi, Béatrice Regnault, Bounsavane Douangboubpha, Yasaman Karami, Delphine Chrétien, Daosavanh Sanamxay, Vilakhan Xayaphet, Phetphoumin Paphaphanh, Vincent Lacoste, Somphavanh Somlor, Khaithong Lakeomany, Nothasin Phommavanh, Philippe Pérot, Océane Dehan, Faustine Amara, Flora Donati, Thomas Bigot, Michael Nilges, Félix A. Rey, Sylvie van der Werf, Paul T. Brey, and Marc Eloit. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. 604(7905):330–336. ISSN 1476-4687. doi: 10.1038/s41586-022-04532-4. URL https://www.nature.com/articles/s41586-022-04532-4. Number: 7905 Publisher: Nature Publishing Group.

[108] Tommy Tsan-Yuk Lam, Na Jia, Ya-Wei Zhang, Marcus Ho-Hin Shum, Jia-Fu Jiang, Hua-Chen Zhu, Yi-Gang Tong, Yong-Xia Shi, Xue-Bing Ni, Yun-Shi Liao, Wen-Juan Li, Bao-Gui Jiang, Wei Wei, Ting-Ting Yuan, Kui Zheng, Xiao-Ming Cui, Jie Li, Guang-Qian Pei, Xin Qiang, William Yiu-Man Cheung, Lian-Feng Li, Fang-Fang Sun, Si Qin, Ji-Cheng Huang, Gabriel M. Leung, Edward C. Holmes, Yan-Ling Hu, Yi Guan, and Wu-Chun Cao. Identifying SARS-CoV-2-related coronaviruses in malayan pangolins. 583(7815):282–285, .

ISSN 1476-4687. doi: 10.1038/s41586-020-2169-0. URL https://www.nature.com/articles/s41586-020-2169-0. Number: 7815 Publisher: Nature Publishing Group.

[109] Katsuhisa Nakajima, Ulrich Desselberger, and Peter Palese. Recent human influenza a (h1n1) viruses are closely related genetically to strains isolated in 1950. 274(5669):334–339. ISSN 1476-4687. doi: 10.1038/274334a0. URL https://www.nature.com/articles/274334a0. Publisher: Nature Publishing Group.

[110] Joel O. Wertheim. The re-emergence of h1n1 influenza virus in 1977: A cautionary tale for estimating divergence times using biologically unrealistic sampling dates. 5(6):e11184. ISSN 1932-6203. doi: 10.1371/journal.pone.0011184. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0011184. Publisher: Public Library of Science.

[111] Michelle Rozo and Gigi Kwik Gronvall. The reemergent 1977 h1n1 strain and the gain-of-function debate. 6(4):e01013–15. ISSN 2150-7511. doi: 10.1128/mBio.01013-15. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4542197/.

[112] Gang Zhang, Dany Shoham, David Gilichinsky, Sergei Davydov, John D. Castello, and Scott O. Rogers. Evidence of influenza a virus RNA in siberian lake ice. 80(24):12229–12235. doi: 10.1128/JVI.00986-06. URL https://journals.asm.org/doi/full/10.1128/JVI.00986-06. Publisher: American Society for Microbiology.

[113] Michael Worobey. Phylogenetic evidence against evolutionary stasis and natural abiotic reservoirs of influenza a virus. 82(7):3769–3774. doi: 10.1128/JVI.02207-07. URL https://journals.asm.org/doi/full/10.1128/JVI.02207-07. Publisher: American Society for Microbiology.

[114] A. I. Klimov, N. J. Cox, W. V. Yotov, E. Rocha, G. I. Alexandrova, and A. P. Kendal. Sequence changes in the live attenuated, cold-adapted variants of influenza a/leningrad/134/57 (h2n2) virus. 186(2):795–797. ISSN 0042-6822. doi: 10.1016/0042-6822(92)90050-y.

[115] Sander Herfst, Eefje J. A. Schrauwen, Martin Linster, Salin Chutinimitkul, Emmie de Wit, Vincent J. Munster, Erin M. Sorrell, Theo M. Bestebroer, David F. Burke, Derek J. Smith, Guus F. Rimmelzwaan, Albert D. M. E. Osterhaus, and Ron A. M. Fouchier. Airborne transmission of influenza a/h5n1 virus between ferrets. 336(6088):1534–1541. ISSN 1095-9203. doi: 10.1126/science.1213362.

[116] Robin M. Bush, Catherine A. Bender, Kanta Subbarao, Nancy J. Cox, and Walter M. Fitch. Predicting the evolution of human influenza a. 286(5446):1921–1925. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.286.5446.1921. URL https://science.sciencemag.org/content/286/5446/1921. Publisher: American Association for the Advancement of Science Section: Research Article.

[117] Yuki Furuse, Kozue Shimabukuro, Takashi Odagiri, Rumi Sawayama, Takashi Okada, Irona Khandaker, Akira Suzuki, and Hitoshi Oshitani. Comparison of selection pressures on the HA gene of pandemic (2009) and seasonal human and swine influenza a h1 subtype

viruses. 405(2):314–321. ISSN 0042-6822. doi: 10.1016/j.virol.2010.06.018. URL https://www.sciencedirect.com/science/article/pii/S0042682210004009.

[118] Joseph Akoi Boré, Joseph W. S. Timothy, Tom Tipton, Ifono Kekoura, Yper Hall, Grace Hood, Stephanie Longet, Kimberly Fornace, Millimono S. Lucien, Sarah Katarina Fehling, Beatrice K. Koivogui, Si'Ana A. Coggins, Eric D. Laing, Christopher C. Broder, N' Faly Magassouba, Thomas Strecker, Jeremy Rossman, Kader Konde, and Miles W. Carroll. Serological evidence of zoonotic filovirus exposure among bushmeat hunters in guinea. 15 (1):4171. ISSN 2041-1723. doi: 10.1038/s41467-024-48587-5. URL https://www.nature.com/articles/s41467-024-48587-5. Publisher: Nature Publishing Group.

[119] Cecilia A. Sánchez, Hongying Li, Kendra L. Phelps, Carlos Zambrana-Torrelio, Lin-Fa Wang, Peng Zhou, Zheng-Li Shi, Kevin J. Olival, and Peter Daszak. A strategy to assess spillover risk of bat SARS-related coronaviruses in southeast asia. 13(1):4380. ISSN 2041-1723. doi: 10.1038/s41467-022-31860-w. URL https://www.nature.com/articles/s41467-022-31860-w. Publisher: Nature Publishing Group.

[120] Sergei L. Kosakovsky Pond, David Posada, Michael B. Gravenor, Christopher H. Woelk, and Simon D. W. Frost. Automated phylogenetic detection of recombination using a genetic algorithm. 23(10):1891–1901. ISSN 0737-4038. doi: 10.1093/molbev/msl051. URL https://doi.org/10.1093/molbev/msl051.

[121] Adrian J. Gibbs, John S. Armstrong, and Jean C. Downie. From where did the 2009 'swine-origin' influenza a virus (h1n1) emerge? 6(1):207. ISSN 1743-422X. doi: 10.1186/1743-422X-6-207. URL https://doi.org/10.1186/1743-422X-6-207.

[122] Stephanie N. Seifert, Robert J. Fischer, Eeva Kuisma, Cynthia Badzi Nkoua, Gerard Bounga, Marc-Joël Akongo, Jonathan E. Schulz, Beatriz Escudero-Pérez, Beal-Junior Akoundzie, Vishnou Reize Bani Ampiri, Ankara Dieudonne, Ghislain Dzeret Indolo, Serge D. Kaba, Igor Louzolo, Lucette Nathalie Macosso, Yanne Mavoungou, Valchy Bel-bebi Miegakanda, Rock Aimé Nina, Kevin Tolovou Samabide, Alain I. Ondzie, Francine Ntoumi, César Muñoz-Fontela, Jean-Vivien Mombouli, Sarah H. Olson, Chris Walzer, Fabien Roch Niama, and Vincent J. Munster. Zaire ebolavirus surveillance near the bikoro region of the democratic republic of the congo during the 2018 outbreak reveals presence of seropositive bats. 16(6):e0010504. ISSN 1935-2735. doi: 10.1371/journal.pntd.0010504. URL https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0010504. Publisher: Public Library of Science.

[123] 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, . URL https://www.biorxiv.org/content/10.1101/2023.07.12.548617v1. Pages: 2023.07.12.548617 Section: New Results.

[124] Ha Minh Lam, Oliver Ratmann, and Maciej F Boni. Improved algorithmic complexity

for the 3seq recombination detection algorithm. 35(1):247–251, . ISSN 0737-4038. doi: 10.1093/molbev/msx263. URL https://doi.org/10.1093/molbev/msx263.

[125] Kangpeng Xiao, Junqiong Zhai, Yaoyu Feng, Niu Zhou, Xu Zhang, Jie-Jian Zou, Na Li, Yaqiong Guo, Xiaobing Li, Xuejuan Shen, Zhipeng Zhang, Fanfan Shu, Wanyi Huang, Yu Li, Ziding Zhang, Rui-Ai Chen, Ya-Jiang Wu, Shi-Ming Peng, Mian Huang, Wei-Jun Xie, Qin-Hui Cai, Fang-Hui Hou, Wu Chen, Lihua Xiao, and Yongyi Shen. Isolation of SARS-CoV-2-related coronavirus from malayan pangolins. 583(7815):286–289. ISSN 1476-4687. doi: 10.1038/s41586-020-2313-x. URL https://www.nature.com/articles/s41586-020-2313-x. Publisher: Nature Publishing Group.

[126] Wilson Smith, C. H. Andrewes, and P. P. Laidlaw. A VIRUS OBTAINED FROM INFLUENZA PATIENTS. 222(5732):66–68, . ISSN 0140-6736. doi: 10.1016/S0140-6736(00)78541-2. URL https://www.sciencedirect.com/science/article/pii/S0140673600785412.

[127] Pavel Sagulenko, Vadim Puller, and Richard A Neher. Treetime: Maximum-likelihood phylodynamic analysis. *Virus evolution*, 4(1):vex042, 2018.

[128] Ariane Düx, Sebastian Lequime, Livia Victoria Patrono, Bram Vrancken, Sengül Boral, Jan F. Gogarten, Antonia Hilbig, David Horst, Kevin Merkel, Baptiste Prepoint, Sabine Santibanez, Jasmin Schlotterbeck, Marc A. Suchard, Markus Ulrich, Navena Widulin, Annette Mankertz, Fabian H. Leendertz, Kyle Harper, Thomas Schnalke, Philippe Lemey, and Sébastien Calvignac-Spencer. Measles virus and rinderpest virus divergence dated to the sixth century BCE. 368(6497):1367–1370. doi: 10.1126/science.aba9411. URL https://www.science.org/doi/10.1126/science.aba9411. Publisher: American Association for the Advancement of Science.

[129] Kevin W. Graepel, Xiaotao Lu, James Brett Case, Nicole R. Sexton, Everett Clinton Smith, and Mark R. Denison. Proofreading-deficient coronaviruses adapt for increased fitness over long-term passage without reversion of exoribonuclease-inactivating mutations. 8(6): e01503–17. ISSN 2150-7511. doi: 10.1128/mBio.01503-17. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5676041/.

[130] David I. Bernstein, Abbie R. Bellamy, III Hook, Edward W., Myron J. Levin, Anna Wald, Marian G. Ewell, Peter A. Wolff, Carolyn D. Deal, Thomas C. Heineman, Gary Dubin, and Robert B. Belshe. Epidemiology, clinical presentation, and antibody response to primary infection with herpes simplex virus type 1 and type 2 in young women. *Clinical infectious diseases*, 56(3):344–351, 2013.

[131] Katharine J Looker, Jocelyn AR Elmes, Sami L Gottlieb, Joshua T Schiffer, Peter Vickerman, Katherine ME Turner, and Marie-Claude Boily. Effect of hsv-2 infection on subsequent hiv acquisition: an updated systematic review and meta-analysis. *The Lancet infectious diseases*, 17(12):1303–1316, 2017.

[132] Aleksandra Knezevic, Jelena Martic, Maja Stanojevic, Sasa Jankovic, Jasminka Nedeljkovic, Ljubica Nikolic, Srdjan Pasic, Borisav Jankovic, and Tanja Jovanovic. Disseminated

neonatal herpes caused by herpes simplex virus types 1 and 2. *Emerging infectious diseases*, 13(2):302, 2007.

[133] Charlotte James, Manale Harfouche, Nicky J Welton, Katherine ME Turner, Laith J Abu-Raddad, Sami L Gottlieb, and Katharine J Looker. Global and regional estimates of herpes simplex virus infection prevalence and incidence in 2016. *Bulletin of the World Health Organization*, 98:315–329, 2020.

[134] Joel O Wertheim, Martin D Smith, Davey M Smith, Konrad Scheffler, and Sergei L Kosakovsky Pond. Evolutionary origins of human herpes simplex viruses 1 and 2. *Molecular biology and evolution*, 31(9):2356–2364, 2014.

[135] Simon J Underdown, Krishna Kumar, and Charlotte Houldcroft. Network analysis of the hominin origin of herpes simplex virus 2 from fossil data. *Virus evolution*, 3(2):vex026, 2017.

[136] Sonia Burrel, David Boutolleau, Diane Ryu, Henri Agut, Kevin Merkel, Fabian H Leendertz, and Sébastien Calvignac-Spencer. Ancient recombination events between human herpes simplex viruses. *Molecular biology and evolution*, 34(7):1713–1721, 2017.

[137] Cosimo Posth, Gabriel Renaud, Alissa Mittnik, Dorothée G. Drucker, Hélène Rougier, Christophe Cupillard, Frédérique Valentin, Corinne Thevenet, Anja Furtwängler, Christoph Wißing, Michael Francken, Maria Malina, Michael Bolus, Martina Lari, Elena Gigli, Giulia Capecchi, Isabelle Crevecoeur, Cédric Beauval, Damien Flas, Mietje Germonpré, Johannes van der Plicht, Richard Cottiaux, Bernard Gély, Annamaria Ronchitelli, Kurt Wehrberger, Dan Grigorescu, Jiří Svoboda, Patrick Semal, David Caramelli, Hervé Bocherens, Katerina Harvati, Nicholas J. Conard, Wolfgang Haak, Adam Powell, and Johannes Krause. Pleistocene mitochondrial genomes suggest a single major dispersal of non-africans and a late glacial population turnover in europe. *Current Biology*, 26(6):827–833, 2016.

[138] Rasmus Nielsen, Joshua M Akey, Mattias Jakobsson, Jonathan K Pritchard, Sarah Tishkoff, and Eske Willerslev. Tracing the peopling of the world through genomics. *Nature*, 541 (7637):302–310, 2017.

[139] Diego Forni, Chiara Pontremoli, Mario Clerici, Uberto Pozzoli, Rachele Cagliani, and Manuela Sironi. Recent out-of-africa migration of human herpes simplex viruses. *Molecular biology and evolution*, 37(5):1259–1271, 2020.

[140] David Eltis and David Richardson. *Atlas of the transatlantic slave trade*. Yale University Press, 1995.

[141] Axel A Guzmán-Solís, Viridiana Villa-Islas, Miriam J Bravo-López, Marcela Sandoval-Velasco, Julie K Wesp, Jorge A Gómez-Valdés, María de la Luz Moreno-Cabrera, Alejandro Meraz, Gabriela Solís-Pichardo, Peter Schaaf, Benjamin R TenOever, Daniel Blanco-Melo, and María C Ávila Arcos. Ancient viral genomes reveal introduction of human pathogenic viruses into mexico during the transatlantic slave trade. 10:e68612. ISSN 2050-084X. doi:

10.7554/eLife.68612. URL https://doi.org/10.7554/eLife.68612. Publisher: eLife Sciences Publications, Ltd.

[142] Alexei J Drummond, Oliver G Pybus, Andrew Rambaut, Roald Forsberg, and Allen G Rodrigo. Measurably evolving populations. *Trends in ecology & evolution*, 18(9):481–488, 2003.

[143] Cadhla Firth, Andrew Kitchen, Beth Shapiro, Marc A Suchard, Edward C Holmes, and Andrew Rambaut. Using time-structured data to estimate evolutionary rates of double-stranded dna viruses. *Molecular biology and evolution*, 27(9):2038–2051, 2010.

[144] Peter Norberg, Shaun Tyler, Alberto Severini, Rich Whitley, Jan-Åke Liljeqvist, and Tomas Bergström. A genome-wide comparative evolutionary analysis of herpes simplex virus type 1 and varicella zoster virus. *PloS one*, 6(7):e22527, 2011.

[145] Pakorn Aiewsakun and Aris Katzourakis. Time-dependent rate phenomenon in viruses. *Journal of virology*, 90(16):7184–7195, 2016.

[146] David M. Koelle, Peter Norberg, Matthew P. Fitzgibbon, Ronnie M. Russell, Alex L. Greninger, Meei-Li Huang, Larry Stensland, Lichen Jing, Amalia S. Magaret, Kurt Diem, Stacy Selke, Hong Xie, Connie Celum, Jairam R. Lingappa, Keith R. Jerome, Anna Wald, and Christine Johnston. Worldwide circulation of hsv-2× hsv-1 recombinant strains. *Scientific reports*, 7(1):1–12, 2017.

[147] Aaron W Kolb, Inna V Larsen, Jacqueline A Cuellar, and Curtis R Brandt. Genomic, phylogenetic, and recombinational characterization of herpes simplex virus 2 strains. *Journal of virology*, 89(12):6427–6434, 2015.

[148] Ruchi M. Newman, Susanna L. Lamers, Brian Weiner, Stuart C. Ray, Robert C. Colgrove, Fernando Diaz, Lichen Jing, Kening Wang, Sakina Saif, Sarah Young, Matthew Henn, Oliver Laeyendecker, Aaron A. R. Tobian, Jeffrey I. Cohen, David M. Koelle, Thomas C. Quinn, and David M. Knipe. Genome sequencing and analysis of geographically diverse clinical isolates of herpes simplex virus 2. *Journal of virology*, 89(16):8219–8232, 2015.

[149] Gytis Dudas, Luiz Max Carvalho, Trevor Bedford, Andrew J. Tatem, Guy Baele, Nuno R. Faria, Daniel J. Park, Jason T. Ladner, Armando Arias, Danny Asogun, Filip Bielejec, Sarah L. Caddy, Matthew Cotten, Jonathan D'Ambrozio, Simon Dellicour, Antonino Di Caro, Joseph W. Diclaro, Sophie Duraffour, Michael J. Elmore, Lawrence S. Fakoli, Ousmane Faye, Merle L. Gilbert, Sahr M. Gevao, Stephen Gire, Adrianne Gladden-Young, Andreas Gnirke, Augustine Goba, Donald S. Grant, Bart L. Haagmans, Julian A. Hiscox, Umaru Jah, Jeffrey R. Kugelman, Di Liu, Jia Lu, Christine M. Malboeuf, Suzanne Mate, David A. Matthews, Christian B. Matranga, Luke W. Meredith, James Qu, Joshua Quick, Suzan D. Pas, My V. T. Phan, Georgios Pollakis, Chantal B. Reusken, Mariano Sanchez-Lockhart, Stephen F. Schaffner, John S. Schieffelin, Rachel S. Sealfon, Etienne Simon-Loriere, Saskia L. Smits, Kilian Stoecker, Lucy Thorne, Ekaete Alice Tobin, Mohamed A. Vandi, Simon J. Watson, Kendra West, Shannon Whitmer, Michael R. Wiley, Sarah M.

Winnicki, Shirlee Wohl, Roman Wölfel, Nathan L. Yozwiak, Kristian G. Andersen, Sylvia O. Blyden, Fatorma Bolay, Miles W. Carroll, Bernice Dahn, Boubacar Diallo, Pierre Formenty, Christophe Fraser, George F. Gao, Robert F. Garry, Ian Goodfellow, Stephan Günther, Christian T. Happi, Edward C. Holmes, Brima Kargbo, Sakoba Keïta, Paul Kellam, Marion P. G. Koopmans, Jens H. Kuhn, Nicholas J. Loman, N'Faly Magassouba, Dhamari Naidoo, Stuart T. Nichol, Tolbert Nyenswah, Gustavo Palacios, Oliver G. Pybus, Pardis C. Sabeti, Amadou Sall, Ute Ströher, Isatta Wurie, Marc A. Suchard, Philippe Lemey, and Andrew Rambaut. Virus genomes reveal factors that spread and sustained the ebola epidemic. *Nature*, 544(7650):309–315, 2017.

[150] David Eltis and David Richardson. *Atlas of the Transatlantic Slave Trade*. Yale University Press. ISBN 978-0-300-18529-4. doi: 10.12987/9780300185294. URL https://www.degruyter.com/document/doi/10.12987/9780300185294/html. Publication Title: Atlas of the Transatlantic Slave Trade.

[151] Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James LN Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *science*, 303(5656):327–332, 2004.

[152] Sebastián Duchêne, Simon YW Ho, Ann G Carmichael, Edward C Holmes, and Hendrik Poinar. The recovery, interpretation and use of ancient pathogen genomes. *Current Biology*, 30(19):R1215–R1231, 2020.

[153] Yassine Souilmi, M Elise Lauterbur, Ray Tobler, Christian D Huber, Angad S Johar, Shayli Varasteh Moradi, Wayne A Johnston, Nevan J Krogan, Kirill Alexandrov, and David Enard. An ancient viral epidemic involving host coronavirus interacting genes more than 20,000 years ago in east asia. *Current Biology*, 31(16):3504–3514, 2021.

[154] Nina Marchi, Laura Winkelbach, Ilektra Schulz, Maxime Brami, Zuzana Hofmanová, Jens Blöcher, Carlos S. Reyna-Blanco, Yoan Diekmann, Alexandre Thiéry, Adamandia Kapopoulou, Vivian Link, Valérie Piuz, Susanne Kreutzer, Sylwia M. Figarska, Elissavet Ganiatsou, Albert Pukaj, Travis J. Struck, Ryan N. Gutenkunst, Necmi Karul, Fokke Gerritsen, Joachim Pechtl, Joris Peters, Andrea Zeeb-Lanz, Eva Lenneis, Maria Teschler-Nicola, Sevasti Triantaphyllou, Sofija Stefanović, Christina Papageorgopoulou, Daniel Wegmann, Joachim Burger, and Laurent Excoffier. The genomic origins of the world's first farmers. 185(11):1842–1859.e18. ISSN 0092-8674. doi: 10.1016/j.cell.2022.04.008. URL https://www.sciencedirect.com/science/article/pii/S009286742200455X.

[155] Meriam Guellil, Lucy van Dorp, Sarah A. Inskip, Jenna M. Dittmar, Lehti Saag, Kristiina Tambets, Ruoyun Hui, Alice Rose, Eugenia D'Atanasio, Aivar Kriiska, Liivi Varul, A. M. H. C. Koekkelkoren, Rimma D. Goldina, Craig Cessford, Anu Solnik, Mait Metspalu, Johannes Krause, Alexander Herbig, John E. Robb, Charlotte J. Houldcroft, and Christiana L. Scheib. Ancient herpes simplex 1 genomes reveal recent viral structure in eurasia. 8(30):eabo4435. doi: 10.1126/sciadv.abo4435. URL https://www.science.org/doi/10.1126/sciadv.abo4435. Publisher: American Association for the Advancement of Science.

[156] Joel O. Wertheim, Martin D. Smith, Davey M. Smith, Konrad Scheffler, and Sergei L. Kosakovsky Pond. Evolutionary origins of human herpes simplex viruses 1 and 2. 31(9): 2356–2364, . ISSN 0737-4038. doi: 10.1093/molbev/msu185. URL https://doi.org/10.1093/molbev/msu185.

[157] Simon J. Underdown, Krishna Kumar, and Charlotte Houldcroft. Network analysis of the hominin origin of herpes simplex virus 2 from fossil data. 3(2). ISSN 2057-1577. doi: 10.1093/ve/vex026. URL https://doi.org/10.1093/ve/vex026.

[158] Joel O Wertheim, Reilly Hostager, Diane Ryu, Kevin Merkel, Samuel Angedakin, Mimi Arandjelovic, Emmanuel Ayuk Ayimisin, Fred Babweteera, Mattia Bessone, Kathryn J Brun-Jeffery, Paula Dieguez, Winnie Eckardt, Barbara Fruth, Ilka Herbinger, Sorrel Jones, Hjalmar Kuehl, Kevin E Langergraber, Kevin Lee, Nadege F Madinda, Sonja Metzger, Lucy Jayne Ormsby, Martha M Robbins, Volker Sommer, Tara Stoinski, Erin G Wessling, Roman M Wittig, Yisa Ginath Yuh, Fabian H Leendertz, and Sébastien Calvignac-Spencer. Discovery of novel herpes simplexviruses in wild gorillas, bonobos, and chimpanzees supports zoonotic origin of HSV-2. . ISSN 0737-4038. doi: 10.1093/molbev/msab072. URL https://doi.org/10.1093/molbev/msab072.

[159] Joel O Wertheim and Sergei L Kosakovsky Pond. Purifying selection can obscure the ancient age of viral lineages. *Molecular biology and evolution*, 28(12):3355–3365, 2011.

[160] Jade Vincent Membrebe, Marc A Suchard, Andrew Rambaut, Guy Baele, and Philippe Lemey. Bayesian inference of evolutionary histories under time-dependent substitution rates. *Molecular biology and evolution*, 36(8):1793–1803, 2019.

[161] Mahan Ghafari, Peter Simmonds, Oliver G Pybus, and Aris Katzourakis. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. *Current Biology*, in Press, 2021.

[162] Sofie Holtsmark Nielsen, Lucy van Dorp, Charlotte J. Houldcroft, Anders G. Pedersen, Morten E. Allentoft, Lasse Vinner, Ashot Margaryan, Elena Pavlova, Vyacheslav Chasnyk, Pavel Nikolskiy, Vladimir Pitulko, Ville N. Pimenoff, François Balloux, and Martin Sikora. 31,600-year-old human virus genomes support a pleistocene origin for common childhood infections. page 2021.06.28.450199. doi: 10.1101/2021.06.28.450199. URL https://www.biorxiv.org/content/10.1101/2021.06.28.450199v1. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

[163] Sonia Burrel, Claire Deback, Henri Agut, and David Boutolleau. Genotypic characterization of ul23 thymidine kinase and ul30 dna polymerase of clinical isolates of herpes simplex virus: natural polymorphism and mutations associated with resistance to antivirals. *Antimicrobial agents and chemotherapy*, 54(11):4833–4842, 2010.

[164] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[165] Alexander Peltzer, Günter Jäger, Alexander Herbig, Alexander Seitz, Christian Kniep, Johannes Krause, and Kay Nieselt. Eager: efficient ancient genome reconstruction. *Genome biology*, 17(1):1–14, 2016.

[166] Matthew Kearse, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, Alex Cooper, Sidney Markowitz, Chris Duran, Tobias Thierer, Bruce Ashton, Peter Meintjes, and Alexei Drummond. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012.

[167] Darren P Martin, Ben Murrell, Michael Golden, Arjun Khoosal, and Brejnev Muhire. Rdp4: Detection and analysis of recombination patterns in virus genomes. *Virus evolution*, 1(1), 2015.

[168] Subha Kalyaanamoorthy, Bui Quang Minh, Thomas KF Wong, Arndt Von Haeseler, and Lars S Jermiin. Modelfinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, 14(6):587–589, 2017.

[169] Maria Anisimova, Manuel Gil, Jean-François Dufayard, Christophe Dessimoz, and Olivier Gascuel. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. 60(5):685–699. ISSN 1063-5157. doi: 10.1093/sysbio/syr041. URL https://doi.org/10.1093/sysbio/syr041.

[170] David A Shaw, Vu C Dinh, and Frederick A Matsen IV. Joint maximum likelihood of phylogeny and ancestral states is not consistent. *Molecular biology and evolution*, 36(10): 2352–2357, 2019.

[171] Andrew Rambaut and Nicholas C Grass. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3): 235–238, 1997.

[172] R Core Team. R: A language and environment for statistical computing, 2013.

[173] Edyth Parker, Catelyn Anderson, Mark Zeller, Ahmad Tibi, Jennifer L. Havens, Geneviève Laroche, Mehdi Benlarbi, Ardeshir Ariana, Refugio Robles-Sikisaka, Alaa Abdel Latif, Alexander Watts, Abdalla Awidi, Saied A. Jaradat, Karthik Gangavarapu, Karthik Ramesh, Ezra Kurzban, Nathaniel L. Matteson, Alvin X. Han, Laura D. Hughes, Michelle McGraw, Emily Spencer, Laura Nicholson, Kamran Khan, Marc A. Suchard, Joel O. Wertheim, Shirlee Wohl, Marceline Côté, Amid Abdelnour, Kristian G. Andersen, and Issa Abu-Dayyeh. Regional connectivity drove bidirectional transmission of SARS-CoV-2 in the middle east during travel restrictions. 13(1):4784. ISSN 2041-1723. doi: 10.1038/s41467-022-32536-1. URL https://www.nature.com/articles/s41467-022-32536-1. Publisher: Nature Publishing Group.

[174] Anne Sofie Hammer, Michelle Lauge Quaade, Thomas Bruun Rasmussen, Jannik Fonager, Morten Rasmussen, Karin Mundbjerg, Louise Lohse, Bertel Strandbygaard, Char-

lotte Sværke Jørgensen, Alonzo Alfaro-Núñez, Maiken Worsøe Rosenstierne, Anette Boklund, Tariq Halasa, Anders Fomsgaard, Graham J. Belsham, and Anette Bøtner. SARS-CoV-2 transmission between mink (neovison vison) and humans, denmark. 27(2):547–551. ISSN 1080-6040. doi: 10.3201/eid2702.203794. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7853580/.

[175] 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, Emily Hagan, Ji-Hua Zhou, Yun-Zhi Zhang, Lin-Fa Wang, Peter Daszak, and Zheng-Li Shi. Serological evidence of bat SARS-related coronavirus infection in humans, china. 33(1):104–107, . ISSN 1674-0769. doi: 10.1007/s12250-018-0012-7. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6178078/.