Improving sampling design and surveillance strategies for inferring the spatiotemporal dynamics of emerging infectious diseases

Joseph Tsui Lok Hei

New College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

Abstract

Emerging infectious diseases represent a continuing threat to global public health. Our ability to respond effectively requires accurate inference of where and when pathogens first emerged, and how they subsequently spread through populations. Despite recent progress, there remains substantial gaps in disease surveillance systems, resulting in incomplete or biased sampling and delayed situational awareness. The primary goal of this thesis is to better understand the implications of these limitations for both downstream inferences and outbreak response, and to develop more effective sampling design and surveillance strategies.

In Chapter 2, I apply phylogeographic methods to the introduction and subsequent local spread of SARS-CoV-2 Omicron BA.1 in the UK, demonstrating how human mobility shaped its dissemination across multiple spatial scales. I also show that travel restrictions implemented at the time were largely ineffective, partly due to the delayed detection of local transmission in international travel hubs. This finding motivates Chapter 3, where I investigate how limited testing resources can be optimally allocated across a mobility network for more accurate inference of the underlying disease distribution during an epidemic. Chapter 4 examines how undersampling of local infections leads to underestimation of viral importation – a limitation in phylogeographic inference highlighted in Chapter 2 – by developing a theoretical framework that characterises the underlying sampling process. Building on this, Chapter 5 explores the broader impacts of heterogeneous sampling on phylogeographic inference and how different sampling strategies can mitigate them, using a simulation-based evaluation framework developed in this work.

Together, these studies provide insights into how limitations in disease surveillance affect our ability to infer the spatiotemporal dynamics of pathogen spread, while offering practical approaches for mitigating these biases. I conclude by discussing how methodologies developed in this thesis can be generalised to other questions in epidemiology and public health, particularly considering recent advances in artificial intelligence.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Prof. Oliver Pybus and Prof. Moritz Kraemer. Without their steady support and guidance, this DPhil would not have been possible. I am particularly grateful for the freedom they have given me to explore my own ideas, follow my curiosity, and forge my own path. Thank you for always believing that I would "figure it out", even at times when I wasn't sure myself.

I am especially indebted to Oliver for taking me under his supervision and helping me find my place in Oxford. Your insights and guidance - whether it is through the manuscript editing or our many walking meetings in University Park - have been instrumental in all my work and growth during my DPhil. It has been a true privilege to learn from you. Equally, I owe special thanks to Moritz, whose mentorship, trust, and enthusiasm have been critical to every step of my DPhil journey. Your generosity - both in providing me with exciting research opportunities and entrusting me with valuable data and resources - has been invaluable in shaping my development as a researcher. Above all, the scientific discoveries I have shared with both of you have been both a privilege and joy. I genuinely believe in the value and significance of the work that we have done and plan to do - and that belief that has carried me through many long nights and tired mornings. I look forward to many more shared adventures in science in the years to come.

I am sincerely grateful to New College for their support during my time in Oxford, particularly through the Yeotown Scholarship, which made this DPhil journey was possible.

I would also like to thank the many collaborators I had the pleasure of working with during my DPhil. To Houriiyah Tegally, Monika Moir, Eduan Wilkinson, Jenicca Poongavanan, and the wider team at CERI – thank you for all your inspiring work, and for allowing me to contribute to some of these efforts. A special thanks to Houriiyah and Jenicca for our road trip to Cape Point, it was certainly an experience unforgettable in more ways than one! To Mengyan Zhang and Elizaveta Semenova in the Computer Science Department - it has been a joy working with you both, and I can't wait to tackle the long list of extensions we have planned. To Louis du Plessis at ETH - despite our brief overlap in Oxford, our conversation during the first week of my DPhil left a lasting

impression. Thank you for generously sharing your experience and always helpful insights, both then and during our subsequent collaboration, and I look forward to many more to come. Finally, my sincere gratitude also to Philippe Lemey, Marc Suchard, Simon Dellicour, Simon Cauchemez, Andrew Rambaut, Ben Lambert, Seth Flaxman, Samir Bhatt, John McCrone, Jayna Raghwani, and Verity Hill for their invaluable guidance and mentorship throughout various projects - learning from each of you has truly been a privilege.

A huge thank you to all my lab mates, both current and former: Prathyush, Bernardo, Yalda, Yan, Tiggy, John, Abhishek, Pip, Simon, and Alex. I have gained immensely from all the work we have done together, and I look forward to many more exciting projects ahead. More importantly, thank you for the countless conversations about science, life, and everything in between - and for always making sure that I never missed out on free cakes. To Sumali - the *chosen* one who is always ahead - both in DPhil timeline, life, and, most importantly, age. I will always look back fondly at our time working together in the office, and place the blame firmly on you for my relative absence ever since your departure. You have been sorely missed. To Rhys - thank you for being a reliable travel companion over the past few years, and a constant source of chatter and unprompted musical performance. I wouldn't have it any other way. Finally, a special thank you to Rosario - from talking through your Master's project during my first year to late-night work sessions both as fellow DPhil students, it's remarkable how far we have come. Thank you for all your support, encouragement, laughter, and exceptional patience with my ever-growing pizza debt.

I am deeply grateful to my family for their unwavering love and support. To my parents especially - who have always cared more about my happiness and well-being than any academic milestones or accolades. I recognise that witnessing the many challenges and hardships of the DPhil journey from afar must not have been easy, and I am thankful beyond words for your unyielding patience and love. Thank you also for giving me the courage to follow my own path, for walking with me on this winding and bumpy road, and for reminding me that there's always backup when I need it - whether it is a phone call or the ever-open option of medical school.

This acknowledgement would not be complete without mentioning Alice - my girlfriend and partner in crime. Thank you for coming into my life and bringing much-needed spontaneity and just the right amount of chaos - both big and small. Thank you for keeping me grounded when the road is smooth, and for standing by me when it isn't.

Thank you for reading all my work and for caring about it even more than you let on. I am sorry my DPhil journey hasn't brought you as many crazy adventures as you had hoped, but we have a lifetime ahead of us to make up for it.

Finally, I would like to dedicate this thesis to all those who continue working tirelessly for what they believe in, despite the odds. I moved to the UK to begin this DPhil during a particularly tumultuous time in Hong Kong, when young people brighter and more qualified than I were stripped of their opportunities to fight for a better future, simply because they were brave enough to try. Much has changed since then, but their stories - and those of countless others around the world - is a constant reminder of what a tremendous privilege it is to be in a position where I can contribute, however modestly, to a better world. And that when given the chance, no matter how difficult, I must always try.

"*If you work really hard and you're kind, amazing things will happen."*

- Conan O'Brien

Statement of contributions and publications

The work presented in this thesis was carried out between October 2021 and May 2025, during my time as a DPhil student at the Department of Biology, University of Oxford, under the supervision of Prof. Oliver Pybus and Prof. Moritz Kraemer. It represents my own original research, except where otherwise acknowledged. Parts of the work presented in this thesis have been published or are under peer-review, as outlined below. Where the research was conducted in collaboration with others, I have detailed the nature and extent of my contributions. Related studies to which I have contributed during my DPhil studies, but are not presented as part of this thesis, are listed at the end of this statement.

Chapter 2

This chapter presents a collaborative study that I led jointly with M.U.G.K and O.G.P. I was responsible for all aspects of this work, including project conception and planning, the design and implementation of a pipeline for performing large-scale phylodynamic and phylogeographic inference (with J.T.M.), curation of genomic and epidemiological data (with R.P.D.I., R.E.P., and V.H.), and subsequent data analysis and interpretation of results (with S.B., B.L., V.H., J.T.M., P.B., R.E.P., J.R., P.L., and S.D.). P.B. and S.C. were responsible for performing the branching process simulation, the results of which I analysed and interpreted. I wrote the initial draft of the manuscript (with supervision from M.U.G.K and O.G.P.), which all authors edited and reviewed.

> **Tsui, J.L.-H.**, McCrone, J.T., Lambert, B., Bajaj, S., Inward, R.P.D., Bosetti, P., Pena, R.E., Tegally, H., Hill, V., Zarebski, A.E., Peacock, T.P., Liu, L., Wu, N., Davis, M., Bogoch, I.I., Khan, K., Kall, M., Abdul Aziz, N.I.B., Colquhoun, R., O'Toole, Á., Jackson, B., Dasgupta, A., Wilkinson, E., de Oliveira, T., COVID-19 Genomics UK (COG-UK) consortium¶, Connor, T.R., Loman, N.J., Colizza,

V., Fraser, C., Volz, E., Ji, X., Gutierrez, B., Chand, M., Dellicour, S., Cauchemez, S., Raghwani, J., Suchard, M.A., Lemey, P., Rambaut, A., Pybus, O.G. and Kraemer, M.U.G. (2023) 'Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1', *Science*, 381(6655), pp. 336–343.

Chapter 3

This chapter presents a collaborative study that I led jointly with M.Z. I was responsible for all aspects of this work, including project conception and design (with M.Z., S.F., E.S., and M.U.G.K.), the development of a computational framework for performing simulation experiments, and subsequent data analysis and interpretation of results. I wrote the initial draft of the manuscript (with critical input from M.Z., P.S., M.A.S., O.G.P., S.F., E.S., and M.U.G.K.), which all authors edited and reviewed.

Tsui, J.L.-H.*, Zhang, M.*, Sambaturu, P., Busch-Moreno, S., Suchard, M.A., Pybus, O.G., Flaxman, S., Semenova, E. and Kraemer, M.U.G. (2024) 'Toward optimal disease surveillance with graph-based active learning', *Proceedings of the National Academy of Sciences of the United States of America*, 121(52), p. e2412424121.

Chapter 4

This chapter presents a collaborative study that I led. I was responsible for all aspects of this work, including project conception and design, derivation of key analytical results, curation of relevant literature (with R.E.P.), development of simulation models, and subsequent data analysis and interpretation of results. L.d.P., P.S., L.T., M.U.G.K., O.G.P. advised on methodologies. I wrote the initial draft of the manuscript, which all authors edited and reviewed.

Tsui, J.L.-H., Sambaturu, P., Pena, R.E., Too, L., Gutierrez, B., Inward, R., Kraemer, M.U.G., du Plessis, L. and Pybus, O.G. (2025) 'Transmission lineage dynamics and the detection of viral importation in emerging epidemics', *medRxiv*. Available at: https://doi.org/10.1101/2025.03.05.25323408.

Chapter 5

This chapter is based on a manuscript currently in preparation. I was responsible for all aspects of this work, including project conception and planning, the design and implementation of a new simulation-based evaluation framework for optimising genomic sampling for phylogeographic inference, and subsequent data analysis and interpretation of results in two applications of the framework. I also wrote the initial draft of the manuscript, with critical input from M.U.G.K. and O.G.P.

> **Tsui, J.L.-H.**, Inward, R., Kraemer, M.U.G. and Pybus, O.G. (in preparation) '*SOPHI: Sandbox for Optimising genomic sampling for PHylogeographic Inference*'.

Related publications (* indicates joint-first authorship)

1. Chen, S., Creswell, R., Hounsell, R., Cantrell, L., Bajaj, S., Dahal, P., **Hei, J.T.L.**, Kolade, O., Amswych, M. 'ayan, Naidoo, R., Fowler, T., Hopkins, S., Stepniewska, K., Voysey, M., White, L., Shretta, R. and Lambert, B. (2025) 'Mass testing for discovery and control of COVID-19 outbreaks in adult social care: an observational study and cost-effectiveness analysis of 14 805 care homes in England', *BMJ Public Health*, 3(1).

2. Kraemer, M.U.G.*, **Tsui, J.L.-H.***, Chang, S.Y.*, Lytras, S., Khurana, M.P., Vanderslott, S., Bajaj, S., Scheidwasser, N., Curran-Sebastian, J.L., Semenova, E., Zhang, M., Unwin, H.J.T., Watson, O.J., Mills, C., Dasgupta, A., Ferretti, L., Scarpino, S.V., Koua, E., Morgan, O., Tegally, H., Paquet, U., Moutsianas, L., Fraser, C., Ferguson, N.M., Topol, E.J., Duchêne, D.A., Stadler, T., Kingori, P., Parker, M.J., Dominici, F., Shadbolt, N., Suchard, M.A., Ratmann, O., Flaxman, S., Holmes, E.C., Gomez-Rodriguez, M., Schölkopf, B., Donnelly, C.A., Pybus, O.G., Cauchemez, S. and Bhatt, S. (2025) 'Artificial intelligence for modelling infectious disease epidemics', *Nature*, 638(8051), pp. 623–635.

3. Chen, Z., **Tsui, J.L.-H.**, Cai, J., Su, S., Viboud, C., du Plessis, L., Lemey, P., Kraemer, M.U.G. and Yu, H. (2025) 'Disruption of seasonal influenza circulation and evolution during the 2009 H1N1 and COVID-19 pandemics in Southeastern Asia', *Nature communications*, 16(1), p. 475.

4. Tamayo Cuartero, C., Carnegie, A.C., Cucunuba, Z.M., Cori, A., Hollis, S.M., Van Gaalen, R.D., Baidjoe, A.Y., Spina, A.F., Lees, J.A., Cauchemez, S., Santos, M., Umaña, J.D., Chen, C., Gruson, H., Gupte, P., **Tsui, J.**, Shah, A.A., Millan, G.G.,

Quevedo, D.S., Batra, N., Torneri, A. and Kucharski, A.J. (2025) 'From the 100 Day Mission to 100 lines of software development: how to improve early outbreak analytics', *The Lancet. Digital health*, 7(2), pp. e161–e166.

5. Poongavanan, J., Lourenço, J., **Tsui, J.L.**, Colizza, V., Ramphal, Y., Baxter, C., Kraemer, M.U.G., Dunaiski, M., de Oliveira, T. and Tegally, H. (2024) 'Dengue virus importation risks in Africa: a modelling study', *The Lancet. Planetary health*, 8(12).

6. Chen, Z., **Tsui, J.L.-H.**, Gutierrez, B., Busch Moreno, S., du Plessis, L., Deng, X., Cai, J., Bajaj, S., Suchard, M.A., Pybus, O.G., Lemey, P., Kraemer, M.U.G. and Yu, H. (2024) 'COVID-19 pandemic interventions reshaped the global dispersal of seasonal influenza viruses', *Science*, 386(6722), p. eadq3003.

7. Gutierrez, B.*, **Tsui, J.L.-H.***, Pullano, G.*, Mazzoli, M.*, Gangavarapu, K.*, Inward, R.P.D., Bajaj, S., Evans Pena, R., Busch-Moreno, S., Suchard, M.A., Pybus, O.G., Dunner, A., Puentes, R., Ayala, S., Fernandez, J., Araos, R., Ferres, L., Colizza, V. and Kraemer, M.U.G. (2024) 'Routes of importation and spatial dynamics of SARS-CoV-2 variants during localized interventions in Chile', *PNAS nexus*, 3(11), p. gae483.

8. Bajaj, S.*, Chen, S.*, Creswell, R.*, Naidoo, R.*, **Tsui, J.L.***, Kolade, O., Nicholson, G., Lehmann, B., Hay, J.A., Kraemer, M.U.G., Aguas, R., Donnelly, C.A., Fowler, T., Hopkins, S., Cantrell, L., Dahal, P., White, L.J., Stepniewska, K., Voysey, M. and Lambert, B. (2024) 'COVID-19 testing and reporting behaviours in England across different sociodemographic groups: a population-based study using testing data and data from community prevalence surveillance surveys', *The Lancet. Digital health*, 6(11).

9. **Tsui, J.L.-H.***, Pena, R.E.*, Moir, M.*, Inward, R.P.D., Wilkinson, E., San, J.E., Poongavanan, J., Bajaj, S., Gutierrez, B., Dasgupta, A., de Oliveira, T., Kraemer, M.U.G., Tegally, H. and Sambaturu, P. (2024) 'Impacts of climate change-related human migration on infectious diseases', *Nature Climate Change*, 14(8), pp. 793–802.

10. Chen, S.*, Hounsell, R.*, Cantrell, L.*, **Tsui, L.H.***, Naidoo, R.*, Dahal, P., Creswell, R., Bajaj, S., Flegg, J.A., Fowler, T., Hopkins, S., Lambert, B., Voysey, M., White, L.J., EY-Oxford Health Analytics Consortium, Stepniewska, K. and Shretta, R. (2024) 'Assessing the impact of testing for COVID-19 using lateral flow devices in NHS acute trusts in England', *medRxiv*. Available at: https://doi.org/10.1101/2024.06.06.24308561.

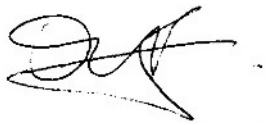
11. Yang, Q., Wang, B., Lemey, P., Dong, L., Mu, T., Wiebe, R.A., Guo, F., Trovão, N.S., Park, S.W., Lewis, N., **Tsui, J.L.-H.**, Bajaj, S., Cheng, Y., Yang, L., Haba, Y., Li, B., Zhang, G., Pybus, O.G., Tian, H. and Grenfell, B. (2024) 'Synchrony of Bird Migration with Global Dispersal of Avian Influenza Reveals Exposed Bird Orders', *Nature communications*, 15(1), p. 1126.

12. Tegally, H.*, Wilkinson, E.*, **Tsui, J.L.-H.***, Moir, M.*, Martin, D., Brito, A.F., Giovanetti, M., Khan, K., Huber, C., Bogoch, I.I., San, J.E., Poongavanan, J., Xavier, J.S., Candido, D. da S., Romero, F., Baxter, C., Pybus, O.G., Lessells, R.J., Faria, N.R., Kraemer, M.U.G. and de Oliveira, T. (2023) 'Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern', *Cell*, 186(15), pp. 3277–3290.e16.

13. Robert, A., **Tsui Lok Hei, J.**, Watson, C.H., Gsell, P.-S., Hall, Y., Rambaut, A., Longini, I.M., Jr, Sakoba, K., Kucharski, A.J., Touré, A., Danmadji Nadlaou, S., Saidou Barry, M., Fofana, T.O., Lansana Kaba, I., Sylla, L., Diaby, M.L., Soumah, O., Diallo, A., Niare, A., Diallo, A., Eggo, R.M., Caroll, M.W., Henao-Restrepo, A.M., Edmunds, W.J. and Hué, S. (2023) 'Quantifying the value of viral genomics when inferring who infected whom in the 2014-16 Ebola virus outbreak in Guinea', *Virus evolution*, 9(1), p. vead007.

14. Viana, R., Moyo, S., Amoako, D.G., Tegally, H., Scheepers, C., Althaus, C.L., Anyaneji, U.J., Bester, P.A., Boni, M.F., Chand, M., Choga, W.T., Colquhoun, R., Davids, M., Deforche, K., Doolabh, D., du Plessis, L., Engelbrecht, S., Everatt, J., Giandhari, J., Giovanetti, M., Hardie, D., Hill, V., Hsiao, N.-Y., Iranzadeh, A., Ismail, A., Joseph, C., Joseph, R., Koopile, L., Kosakovsky Pond, S.L., Kraemer, M.U.G., Kuate-Lere, L., Laguda-Akingba, O., Lesetedi-Mafoko, O., Lessells, R.J., Lockman, S., Lucaci, A.G., Maharaj, A., Mahlangu, B., Maponga, T., Mahlakwane, K., Makatini, Z., Marais, G., Maruapula, D., Masupu, K., Matshaba, M., Mayaphi, S., Mbhele, N., Mbulawa, M.B., Mendes, A., Mlisana, K., Mnguni, A., Mohale, T., Moir, M., Moruisi, K., Mosepele, M., Motsatsi, G., Motswaledi, M.S., Mphoyakgosi, T., Msomi, N., Mwangi, P.N., Naidoo, Y., Ntuli, N., Nyaga, M., Olubayo, L., Pillay, S., Radibe, B., Ramphal, Y., Ramphal, U., San, J.E., Scott, L., Shapiro, R., Singh, L., Smith-Lawrence, P., Stevens, W., Strydom, A., Subramoney, K., Tebeila, N., Tshiabuila, D., **Tsui, J.**, van Wyk, S., Weaver, S., Wibmer, C.K., Wilkinson, E., Wolter, N., Zarebski, A.E., Zuze, B., Goedhals, D., Preiser, W., Treurnicht, F., Venter, M., Williamson, C., Pybus, O.G., Bhiman, J., Glass, A., Martin, D.P., Rambaut, A., Gaseitsiwe, S., von Gottberg, A. and de Oliveira, T. (2022) 'Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa', *Nature*, 603(7902), pp. 679–686.

As Joseph's supervisors, I confirm that the information presented here is representative of his contributions to the work listed in this thesis.



Prof. Oliver G. Pybus
Professor of Evolution & Infectious Disease
Department of Biology
University of Oxford

Prof. Moritz U. G. Kraemer
Professor of Epidemiology & Data Science
Department of Biology
University of Oxford

Table of Contents

List of Figures

1 Introduction

1.1 Global burden of emerging infectious diseases

Throughout human history, few forces have shaped the course of civilisations as profoundly as the spread of infectious diseases. From the Athenian Plague in 430 BC to the recent COVID-19 pandemic, the emergence of novel pathogens capable of widespread human-to-human transmission has not only led to the loss of millions of lives, but also fundamental changes in population dynamics, socio-political structures, as well as governance and warfare (1-3).

A prominent example of the impact of infectious diseases is the bubonic plague, often known as the Black Death, which arrived in Europe in the mid-14th century through trade routes connecting Europe, Asia, and the Middle East (4-6). The spread of the causative pathogen Yersinia pestis - a zoonotic bacteria found in small mammals such as rodents and their fleas - killed an estimated 25-50 million people, or 30-60% of the total population in Europe at the time (4, 7, 8). The impact of this unprecedented mortality went far beyond its immediate death toll: labour shortage as a result of the epidemic led surviving workers to demand higher wages and better living conditions, followed with attempts by the ruling class to preserve feudal norms which were met with social unrest and revolts; meanwhile, authority of medical experts and religious institutions were challenged as they failed to contain the disease spread (9, 10). Crucially, the lack of a fundamental understanding of the disease's aetiology and transmission mechanism not only led to widespread fear and superstitious beliefs, but also control measures that were

largely ineffective and likely further prolonged the outbreak (11-13). Unfortunately, similar patterns of societal upheavals were observed in subsequent large-scale outbreaks of the bubonic plague as well as other pathogens, including the 1520-1521 smallpox epidemic in Mexico which led to at least 5 million deaths (14-16), the seven major cholera pandemics between 1817-1947 with more than 23 million deaths in India alone (17-19), and the 1918 influenza pandemic, or Spanish flu, claiming an estimated 50-100 million lives globally amid the aftermath of World War I (20).

With the advent of modern medicine and germ theory in the 19th and 20th centuries however, the mortality resulting from these once-lethal illnesses has been greatly reduced or even eliminated in some parts of the world (e.g., smallpox, polio, measles) (21-23). A better understanding of their underlying transmission dynamics has also enabled the design of more effective public health and social measures (also often referred to as non-pharmaceutical interventions (NPIs)) - such as mask mandates, isolation and quarantine measures, and social-distancing policies - which have helped mitigate the scale and impact of many outbreaks. Nevertheless, the threat of emerging infectious diseases remains ever present, with vulnerable populations in low- and middle-income countries (LMICs) being disproportionately affected - due to not only heightened susceptibility and risk of exposure (e.g., from inadequate sanitation and overcrowding), but also greater burden during outbreaks as a result of limited healthcare access and reduced capacity to respond effectively. For example, in the 2014-2016 West Africa Ebola epidemic which predominantly struck Guinea, Liberia, and Sierra Leone, fragile healthcare infrastructure and limited diagnostic capacity, together with unsafe burial practices and mistrust in government institutions, allowed the virus to spread widely before containment measures took effect (24, 25). Beyond the 11,325 deaths (26), the epidemic led to further weakening of the already strained healthcare system and disrupted

routine immunisation programmes for other preventable diseases, such as measles and polio (27, 28).

At the same time, globalisation and the resulting increased connectivity between human populations mean that high-income countries are by no means insulated from the threat of emerging infectious diseases. Frequent air travel and porous national borders allow local outbreaks to rapidly escalate into global public health emergencies in a matter of weeks, as demonstrated by a number of recent outbreaks - including the 2003 SARS epidemic (29, 30), 2009 H1N1 influenza pandemic (31, 32), the Zika virus epidemic in 2015-2016 (33, 34), and the COVID-19 pandemic (35, 36). In particular for COVID-19, the ability of a pathogen to spread rapidly through the air traffic network was showcased not only during the initial emergence of the SARS-CoV-2 virus, but also in subsequent waves driven by variants with increased transmissibility (e.g., Alpha, Delta, Mu) or immune escape mutations (e.g., Beta, Gamma, Omicron), each originated in different parts of the world and rapidly displaced previously dominant variants in most affected countries. This phenomenon, combined with the virus's high reproduction number and long incubation period that enables undetected and sustained asymptomatic spread (37, 38), contributed to the unprecedented scale and impact of the COVID-19 pandemic, despite its relatively moderate case fatality rate (39-41). As of March 2025, over 700 million cases and 7 million confirmed deaths have been reported globally (42), with excess mortality estimates suggesting the true death toll may be between two to four times higher (43, 44). In addition to a reduction in global life expectancy (45, 46), ongoing research has also highlighted various indirect negative effects as a result of the pandemic, including an increase in the global prevalence of depression (47, 48), rising levels of public distrust of governments and public health authorities (49, 50), and higher rate of preventable illnesses especially in developing countries due to disrupted immunisation

schedules (51-53). Meanwhile, the economic impact is estimated to be in the trillions of dollars (54), with long-term consequences on global trade, labour markets, and social mobility that are likely to persist for years to come (55-57).

As the world continues to grapple with the repercussions of the COVID-19 pandemic, there is growing concern among public health communities about the risk of another pandemic (58). Climate change and land use changes are forcing animals to migrate, potentially creating new opportunities for zoonotic pathogens to spill over into humans (59); mass human migration due to regional conflicts and extreme weather conditions is reshaping the global distribution of endemic diseases, potentially introducing these pathogens into previously unaffected, immunologically naive populations (60, 61); antimicrobial resistance, driven by overuse and misuse of antibiotics in humans and animals, threatens to render many of our current treatments ineffective (62, 63); finally, the spread of misinformation and rising distrust in public health authorities continues to hamper our efforts to prevent and manage infectious disease outbreaks. At the time of writing, the ongoing m-pox outbreak remains a public health emergency of international concern (PHEIC), as declared by the World Health Organisation (WHO) in May 2022 (64), with increasing case numbers reported in Uganda and Zambia (65); meanwhile, the US is experiencing a resurgence of measles cases due to declining vaccination rates (66-68). Although it is impossible to predict if and when the next pandemic will occur, it is clear that the global burden of emerging infectious diseases remains an ever-present risk to human society, and that sustained investment in research and pandemic preparedness will continue to be a global priority for years to come.

1.2 Spatiotemporal dynamics of emerging infectious diseases

Understanding the spatiotemporal dynamics of emerging infectious diseases is vital to our ability to anticipate, prepare for, and manage future outbreaks. For emerging infectious diseases in particular, this includes identifying where and when they first emerged, and understanding how they subsequently propagate through human populations. One of the earliest records of the spatiotemporal distribution of disease occurrence can be traced back to Hippocrates (460-370 BC), whose treatise "Airs, Waters, and Places" highlighted how environmental factors such as climate, geography, and seasonality shaped the prevalence of illnesses in different populations; centuries later, John Snow's (1813-1858) spatial analysis of cholera cases during the 1854 London epidemic famously traced the source of outbreak to a contaminated water pump (69). These early observations led to a key insight central to modern epidemiology - that the distribution of disease occurrence is neither static nor random, but follows distinct spatial and temporal patterns that can be explained by the complex interactions between socio-demographics, human behaviour, and environmental conditions.

Among these different factors, one that received particular attention during the early development of modern epidemiology is the effect of population distribution and socio-economic disparities. Densely populated cities where millions live, travel, and work in close proximity create environments where frequent contacts facilitate sustained disease transmission between humans. In low-income countries especially, population density is often associated with lower socio-economic status, with poorer communities living in overcrowded housing that lacks adequate sanitation and has limited access to clean water (70-72). For instance, studies have shown that tuberculosis (TB) disproportionately affects impoverished communities in densely populated urban neighbourhoods, where cramped housing and poor ventilation allows effective and rapid

airborne transmission; meanwhile, malnutrition and hesitancy to seek medical care further heightens disease susceptibility and burden (73, 74). In addition to human-to-human transmission, these living conditions also make ideal breeding grounds for various disease vectors including mosquitoes and rodents, enabling the emergence and spread of vector-borne and water-borne diseases. As a result, diseases such as cholera, typhoid fever, and Chagas disease often exhibit distinct spatial patterns that can be linked to the underlying socio-economic conditions of the affected populations, where higher rates of transmission and mortality are associated with lower income and limited access to healthcare facilities (75-77).

While these socio-demographic factors represent important predictors of transmission intensity for many infectious diseases, this by no means implies that disease risk is confined to specific regions or communities - as some early investigations of these patterns suggested - often resulting in policies of social segregation and forced isolation (78, 79). This misconception was also partly the result of the relatively slow rate of spread observed for most historical outbreaks, following patterns of gradual spatial dissemination (also known as contagious diffusion) driven by local interactions and spatial proximity. Not unlike the diffusion of air particles in continuous space, a defining pattern of contagious diffusion is the concentric radial expansion of infected regions, typically centred around densely populated human settlements and along major trade routes (80-82). More importantly, the continuous nature of contagious diffusion means that disease spread is often disrupted by different types of barriers, which could be physical (e.g., mountains, rivers, and deserts), social (e.g., administrative boundaries between countries), and epidemiological (e.g., communities with higher levels of immunity due to vaccination and therefore lower risk of sustained transmission) (83). Indeed, previous studies have shown that the spread of many historical outbreaks, such

as the Black Death, exhibited wave-like spreading patterns that can be well explained by contagious diffusion (84-86). A more recent example of such dynamics can be found in the 2014-2016 Ebola outbreak in Western Africa, where the early spread showed patterns of radial expansion centred around epicentres in districts bordering Guinea and Liberia (87, 88). Geographic features and administrative borders between districts likely slowed its encroachment in certain directions, while heterogeneous distribution of population density and difference in local customs between districts further shaped the path of transmission (87, 89, 90).

However, the spatiotemporal patterns of recent large-scale outbreaks have challenged this paradigm. Since the rise of globalisation, growing connectivity among human populations - driven by the expansion of global trade and air travel - has enabled infectious diseases to spread far more rapidly and across greater distances, often bypassing social and geographic barriers that once constrained their reach. Importantly, the hierarchical structure of human mobility networks means that an emerging pathogen can rapidly disseminate over large geographic scales through pathways mediated by densely connected travel hubs (e.g., major international airports), followed by transmission across short distances driven by local mobility patterns or spatial proximity (91, 92). This phenomenon, also known as hierarchical network diffusion, leads to disease spread with evolving spatial scales over the course of an outbreak, as evident in many recent large-scale outbreaks such as SARS, MERS, and COVID-19 (93-96). It should be also noted that this applies to not only the spread of a novel pathogen in an immunologically naive population, but also the dissemination of immune escape variants, as observed during the COVID-19 pandemic with the successive waves of antigenetically distinct variants, each replacing the previously dominant variant in most affected countries within weeks or months of its emergence (97-100).

Beyond human mobility, the dissemination of many infectious diseases is also influenced by environmental conditions, either through changes in pathogen survivability or host susceptibility. For example, for respiratory diseases such as SARS, MERS, and COVID-19, which spread predominantly by airborne transmission, lower humidity has been shown to increase the distance over which virus-containing droplets can travel through air, increasing the probability that they are inhaled by susceptible individuals (101, 102). Meanwhile, lower temperature during winters has been associated with higher transmission rates due to a combination of factors including enhanced virus survival, impaired immune response in humans, and increased indoor crowding (103-106). Whereas for vector-borne diseases, the spatial distribution, abundance, and activity of vectors are known to exhibit periodic fluctuations as a result of changes in temperature, precipitation, and vegetation cover, leading to seasonal variations in transmission intensity. This is particularly relevant for diseases such as malaria and dengue fever, where the distribution of relevant vectors, *Anopheles* mosquitoes and *Aedes aegypti* mosquitoes, respectively, are strongly influenced by environmental conditions (107-110). Over longer timescales, there is also growing evidence of gradual, large-scale shifts in these seasonal patterns driven by climate change. These shifts not only pose further challenges to tracking and predicting the spread of these climate-sensitive diseases, but also influence the frequency and spatial distribution of zoonotic spillover events, introducing additional variability to the spatiotemporal dynamics of emerging infectious diseases.

1.3 Infectious disease surveillance and data collection

It is clear from the discussion so far that the multifaceted nature of infectious disease spread cannot be fully understood without robust and high-quality data across multiple

domains, including demographic, epidemiological, genomic, and mobility data. More importantly, the constant evolution and inherent stochasticity of infectious disease spread means that continuous and systematic data collection is necessary to provide timely insights for understanding the underlying transmission mechanism and informing the design of effective public health interventions.

Traditionally, epidemiologists and public health officials have primarily relied on data collected by two different types of disease surveillance systems: passive surveillance and active surveillance. In passive surveillance, the primary source of data consists of voluntary reporting from healthcare providers, laboratories, and institutions during routine operations. Examples of such data include patient records, death reports, and regular laboratory testing. By identifying cases with specific clinical symptoms (through a process more generally known as syndromic surveillance), it is possible to identify early warning signals for emerging outbreaks and changing disease trends. For example, during the 2009 H1N1 pandemic, data from a large general practitioner database in the UK provided detailed spatial mapping of local cases, enabling public health officials to implement pre-emptive measures such as distribution of antiviral prophylaxis and vaccination campaign targeting populations at risk (111-113). Another notable example can be found in late 2019, when routine hospital reports of pneumonia cases with unknown causes in Wuhan, China, were flagged by the national Pneumonia of Unknown Etiology (PUE) passive surveillance system (114). The resulting investigations led to the identification of SARS-CoV-2 as the causative agent (115, 116), providing critical early signals for local and global responses.

More recently, and especially since the COVID-19 pandemic, many passive surveillance systems have expanded to incorporate novel data sources. Among them, genomic surveillance - the systematic collection and sequencing of pathogen genomes

from infected hosts - has become an essential tool for identifying emerging variants, tracking viral evolution, and informing the design of effective vaccines (115). For example, regular sequencing of positive SARS-CoV-2 samples enabled the early detection of the Beta (118) and Omicron (100) variant in South Africa, prompting rapid national and global outbreak responses; genomic markers such as S-gene target failure (SGTF) in SARS-CoV-2 sequences also allowed the monitoring of variant-specific incidence and prevalence trends in real-time (119, 120). Meanwhile, Next-Generation Sequencing (NGS) has enabled the direct sequencing of environmental or pooled biological samples, without needing to isolate a specific pathogen beforehand as required by traditional PCR (polymerase chain reaction)-based approaches (121-124). This technique underpins wastewater surveillance (though PCR is also used), which has emerged as a promising approach for unbiased, non-invasive monitoring of community-level prevalence and viral diversity (125-127), with potential applications for tracking between-country transmission through sampling at airports or on individual aircrafts (128, 129). In parallel, digital surveillance has gained prominence by leveraging internet search histories, social media trends, and self-reported health data collected via mobile apps to monitor disease outbreaks in near real-time (130). In the UK, for example, the ZOE Health Study app allows users to report their COVID-19 symptoms, test results, and vaccination status - generating one of the largest citizen-participatory datasets to date (131). This dataset has since been used to detect COVID-19 hotspots in the UK (132), and to evaluate changes in symptom profiles (133) and vaccine effectiveness (134). In addition to providing population-wide insights at relatively low costs, these methods are generally less susceptible to the impact of variation in health-seeking behaviour and access to healthcare, which are known to introduce bias and latency to data collected through more traditional passive surveillance approaches.

Meanwhile, active surveillance remains the strategy of choice when complete and high-resolution data are needed. Unlike passive surveillance which relies on routinely collected data, active surveillance adopts a proactive case-finding approach using techniques including patient interviews, household surveys, and targeted testing. In some cases, it may also involve targeted sequencing of samples collected from incoming travellers at border crossings (135, 136), or individuals from infection clusters exhibiting unusual growth - potentially signalling the emergence of a novel variant (100). This approach is particularly useful in settings where underreporting is expected, such as in low-resource areas or during outbreaks of diseases with substantial asymptomatic transmission. More importantly, surveillance efforts such as contact-tracing and individual patient interviews can provide highly detailed metadata critical to understanding the transmission mechanisms of an emerging novel pathogen. For example, during the early phase of the 2014-2016 West Africa Ebola outbreak, contact-tracing efforts identified a single funeral as the source of at least 28 secondary cases and 9 deaths in Sierra Leone, lending further support to the hypothesis of unsafe burial being a key driver of Ebola spread (137). During the COVID-19 pandemic, border screening also became an important approach for detecting and containing imported infections. Travel histories of identified infectious travellers provided valuable insights into the dispersal patterns of SARS-CoV-2 variants and the impact of the global air traffic network, with implications for the design of NPIs intended to delay or prevent the dissemination of an emerging variant (138-142).

1.4 Inferring the spatiotemporal dynamics of infectious disease spread

There is little doubt that disease surveillance has played a critical role in providing data necessary to guide public health response and understand the spatiotemporal dynamics of

disease spread during both historical and modern epidemics. However, the unprecedented scale of the COVID-19 pandemic has highlighted a number of important limitations inherent in existing surveillance infrastructures. For instance, many passive surveillance systems, relying on routine clinical testing and self-reporting, quickly became overwhelmed due to limited testing capacity, resulting in substantial underreporting and bias towards detecting cases with severe symptoms or in high-resource populations (143-145). Active surveillance efforts, such as household surveys and contact-tracing, similarly struggled under logistical constraints and resource limitation given the large volume of cases (146, 147). Meanwhile, delays in laboratory testing and uneven distribution of testing resources introduced further latency and biases in the mapping of virus diversity across geographic locations (148). As a result, epidemiologists and public health officials frequently faced challenges in deriving timely and unbiased insights from data containing substantial gaps and sampling biases across both space and time.

While these challenges are not entirely new, the magnitude and urgency of COVID-19 spurred substantial progress in developing increasingly sophisticated and powerful mathematical methods capable of robust inference of disease dynamics despite limited data, building on approaches developed in response to previous outbreaks (149).

At a basic level, these approaches can be divided into two main categories: statistical modelling and mechanistic modelling. Statistical modelling emphasises the fitting and prediction of patterns in observed data, without necessarily embedding explicit assumptions about the underlying biological process. Examples of its applications include time-series analysis to identify trends and seasonal patterns in observed case numbers over time (e.g., using AutoRegressive Integrated Moving Average (ARIMA) models (150)), and spatial analysis to understand the correlation between socio-demographic factors and the spatial clustering of observed cases, or to infer the underlying disease

distribution across space given the location of observed cases (e.g., using Spatial Autoregressive (SAR) (151) or Conditional Autoregressive (CAR) (152) model). Recent advances focus on applying Bayesian frameworks to simultaneously model spatial and temporal disease given historical case data, with the incorporation of additional data streams such as mobility and environmental data (153-156). Importantly, Bayesian approaches enable the quantification of uncertainty in model predictions as well as the incorporation of prior information - both key advantages compared to traditional frequentist approaches especially in the context of incomplete and biased data. For example, in a recent study published during the COVID-19 pandemic, data from a randomised surveillance study were used to mitigate the effect of ascertainment bias in data from targeted testing efforts in the UK, using a Bayesian framework known as causal debiasing (157); this technique has since been adopted to show how testing behaviours vary across different socio-demographic groups in the UK (158).

Mechanistic modelling, on the other hand, simulates infectious disease spread based on theoretical biological principles or known transmission mechanisms. Examples of mechanistic models include compartmental models, such as the Susceptible-Infectious-Recovered (SIR) model, which partitions the population into different compartments, with individuals moving between them at rates governed by a system of differential equations that describes the underlying transmission process (159). Most recent developments have focused on extending these models to include more compartments (e.g., to account for exposed individuals in an SEIR model) or geographically distinct patches to simulate the effects of human movement on disease spread (also known as a metapopulation model). For instance, during the 2016 Zika outbreak and the COVID-19 pandemic, large-scale metapopulation models such as GLEAM (160) which combines high-resolution demographic and hierarchical mobility

data, have been used to model the spatiotemporal dynamics of early dispersal, revealing the role of the global air traffics in the rapid dissemination of the virus (34, 161). While both statistical and mechanistic models can be used to estimate the value of key epidemiological parameters (e.g., basic reproduction number R_0 and effective reproduction number R(t) or R_t) that are regularly used to describe and predict the temporal dynamics of a disease outbreak, mechanistic models also provide a natural framework for simulating different outbreak scenarios and exploring the potential impact of different NPIs on disease dynamics. More recently, agent-based models, which simulate the behaviour of individual agents and their interactions in an explicit spatial environment, have also gained popularity for their ability to capture the heterogeneity in human behaviour and complex social interactions. Although these models have been applied in real-world settings (162-164), their widespread use remains limited compared to more traditional models, due in part to their resource-intensive computation and the need for detailed data on individual-level interactions.

Over the past decades, and especially during the COVID-19 pandemic, phylodynamics and phylogeography have also emerged as powerful inference approaches for investigating the spatiotemporal dynamics of disease spread. By combining information about the evolutionary relationships between sampled pathogens inferred from their genomes with relevant epidemiological data (e.g., sampling time), these methods enable the estimation of key epidemiological parameters (e.g., birth rate and sampling rate) as well as historical variation in the size of the underlying viral population even before the pathogen was first detected (165-168). With the incorporation of geographic information in phylogeography, it is also possible to reconstruct historical pathogen movement by modelling the spatial dispersal as an evolution process along an estimated phylogenetic tree. For example, by modelling pathogen movement in

continuous space (169, 170), Dellicour et al. measured the rate of spatial expansion of West Nile virus in North America and identified temperature as a key predictor of viral dispersal (171). Whereas for large-scale outbreaks involving transmission events across multiple countries or even continents, a discrete approach is typically employed to model pathogen movement as "jumps" between discrete geographic locations (172-175). This approach was used extensively during the COVID-19 pandemic to infer the source-sink dynamics of SARS-CoV-2 spread (35, 151, 176, 177) and to examine the impact of different environmental and epidemiological factors (e.g., population size, air traffic volume) on dispersal patterns (141, 176, 178, 179). Importantly, the decreasing cost of genomic sequencing and advances in high-performance computing have allowed the analyses of increasingly large viral genomic datasets, often revealing spatiotemporal patterns in disease spread that span multiple spatial scales. Although traditional surveillance efforts such as contact-tracing and border screening can also provide data necessary to measure or infer such dynamics, their resolution and coverage are generally limited by logistical and resource constraints, especially during large-scale and prolonged outbreaks.

It is important to note that, despite substantial progress in developing and advancing these inference methods, there remain important limitations. For example, in metapopulation models, the inherent assumption of homogeneity within subpopulations has been shown to result in biased estimates of key epidemiological parameters, especially those associated with disease dynamics driven by complex social mixing between different demographics, for which data available for model calibration are scarce (180, 181). For phylogeographic analyses, it is well-known that heterogeneous sampling of pathogen genomes (i.e. when the distribution of the sampled genomes is not representative of the underlying viral spread) can lead to inaccurate spatial

reconstructions that are biased towards densely sampled locations (182-184); although some approaches (e.g., structured coalescent models (173, 174, 185)) have been shown to be less susceptible to such biases compared to others (e.g., continuous-time Markov chain models), their application to large-scale outbreaks has so far been limited to due computational inefficiency. Understanding how these biases arise across a diverse range of outbreak contexts and data sampling processes - and therefore how they can be mitigated - remains an active area of research critical for ensuring robust and unbiased inference of the spatiotemporal dynamics of infectious disease spread.

1.5 Resource constraints and decision-making in disease surveillance

The challenge to understand and respond effectively to the COVID-19 pandemic has mobilised substantial resources to improve our ability to perform large-scale disease surveillance and to derive timely and useful insights from the data collected. In parallel, there is also growing interest among the public health community in the optimisation of surveillance systems - ranging from earlier efforts focused on the design of optimal sentinel networks (183-185), to the development of data-driven adaptive strategies (187, 189-191), including approaches that leverage recent advances in machine learning techniques such as active learning and reinforcement learning (192-194). At the core of this growing trend lies the inevitable challenge of allocating increasing, yet finite, resources among competing surveillance priorities.

For example - at the outset of an outbreak involving a novel pathogen, the foremost priority is the rapid detection and isolation of cases to 1) prevent further transmission, and 2) gather critical data to identify the pathogen and its transmission mechanisms. This typically involves deploying surveillance teams to perform active case-finding, contact-tracing, and collecting genomic samples from infected patients for

laboratory testing. However, as the outbreak progresses, surveillance priorities may shift to monitoring overall disease trends to facilitate near-term forecasts or identifying disease hotspots for targeted interventions. Meanwhile, and especially during later phases of the outbreak, epidemiologists might continue to perform retrospective analysis of early case data and environmental samples from where the disease first detected - with the overarching goal of identifying risk factors associated with the emergence of the pathogen and preventing similar outbreaks in the future.

Making informed decisions about how to allocate resources across these evolving priorities and as new data becomes available is far from straightforward. Importantly, as each stage of an outbreak requires different types of information, policymakers face difficult trade-offs between accuracy, timeliness, and cost. For example, shifting resources from active case-finding to monitoring disease trends at a population level is likely to reduce the immediate capacity for case isolation and contact-tracing efforts, while providing essential data for reliable forecasting and long-term planning. Similarly, prioritising rapid diagnostic tests over PCR-based genomic sequencing provides timely data for rapid outbreak response, but reduces the availability of genomic data needed for retrospective phylogeographic reconstruction of disease spread. The increasing diversity and volume of available data also presents epidemiologists and public health officials with the additional challenge of balancing data completeness and inference robustness against the computational and time costs of resource-intensive analyses. Striking the optimal balance in these trade-offs requires an understanding of how each type of data is being collected and potential biases inherent in the collection process, how these biases impact the accuracy and robustness of relevant inferences, and the utility of collecting more data or incorporating additional data streams in addressing specific research

questions or informing specific policy decisions. These are the key questions that motivated the work presented in this thesis.

1.6 Thesis overview

This thesis presents a series of interconnected studies, each addressing a distinct aspect of the challenges outlined above. It begins with an empirical study of SARS-CoV-2 spread in 2021-2022, revealing critical limitations in current disease surveillance systems and illustrating how these limitations can lead to biased inferences and suboptimal outbreak response. Subsequent studies examine these limitations further, with a focus on developing corresponding mitigation strategies using analytical and computational techniques drawn from genomic epidemiology, ecology, and machine learning. In addition to providing new theoretical insights, this thesis introduces novel methodologies and frameworks that contribute to ongoing research efforts to improve sampling design and surveillance strategies for inferring the spatiotemporal dynamics of emerging infectious diseases. The following provides a brief overview of each chapter.

Chapter 2 resents a collaborative study I led as first and co-corresponding author, titled "*Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1*" and published in *Science* (2023). Using >115,000 genomes, I reconstructed the invasion process of SARS-CoV-2 Omicron BA.1 into England through discrete phylogeography, showing that the intensity of viral importation increased exponentially despite travel restrictions targeted at southern African countries where the variant was first detected. I then demonstrated using a combination of air traffic data, reported case numbers, and individual travel histories, that this was due to a rapid increase in local prevalence at highly connected travel hubs such as Spain and the United States, which were not covered by the travel ban and therefore acted as major secondary exporters of the virus. Having

characterised the invasion process, I then considered the spatiotemporal dynamics of local spread in the UK through both continuous and discrete phylogeography. This revealed a distinct two-stage process driven initially by the hierarchical travel network within the country, followed by patterns of local diffusion centred around densely populated urban conurbations. In addition to furthering our understanding of the spatiotemporal dynamics of infectious disease spread, my work in this chapter identified a number of key limitations and challenges in existing surveillance systems and sampling design for phylogeographic inference, which motivated the subsequent work presented in this thesis.

Chapter 3 presents a collaborative study I led as first and co-corresponding author, titled "*Toward optimal disease surveillance with graph-based active learning*" and published in *PNAS* (2024). The ability to track the spatial spread of an emerging pathogen is critical to the design of effective containment strategies, as highlighted in Chapter 2. In low-resource settings, where comprehensive testing is not feasible, accurately inferring the underlying spatial distribution of infections requires careful decisions about how limited testing resources should be allocated to maximise information gain, given prior test results and patterns of human mobility between geographic locations. In this study, I showed that this decision-making process can be modelled as an iterative node classification problem on an undirected and unweighted graph, in which nodes represent geographic locations and edges represent movement of infected individuals. This formulation enabled the application of selection strategies developed in the field of active learning - a subfield of machine learning concerned with the selection of data instances for labelling to optimise model training - which I evaluated across a range of simulated outbreak scenarios on both synthetic and empirical networks. I further proposed a novel sampling policy that outperformed existing ones in most outbreak scenarios, particularly in low-budget settings. This work represents an initial

step towards the design of more cost-effective and adaptive surveillance systems for providing data necessary to estimate the underlying distribution of disease prevalence from partial observations - an essential task for early risk assessment and outbreak preparedness.

Chapter 4 presents a collaborative study I led as first and co-corresponding author, titled "*Transmission lineage dynamics and the detection of viral importation in emerging epidemics*" (currently under peer-review). The introduction of an emerging pathogen via the movement of infectious travellers plays a critical role in shaping the early dynamics of local transmission, as seen in Chapter 2. While phylogeography enables the detection and enumeration of such events through the reconstruction of transmission lineages, it is well recognised that the number of detected introductions often substantially underestimates the true number due to undersampling of infections. However, the mechanism underlying this underestimation remains poorly characterised. In this study, I addressed this gap by developing a theoretical framework to model the coupled dynamics of viral importation and local transmission, showing how these dynamics shape the size distribution of local transmission lineages over time. Using both deterministic and stochastic agent-based simulations, I further demonstrated that the probability of detecting individual importation events depends on their timing of occurrence, sampling proportion, underlying migration rate, and local transmission conditions. These findings have important implications for the interpretation of viral movement estimates from phylogeography, especially in evaluating the effectiveness of containment strategies aimed at limiting spatial spread, and in comparing the relative contribution of viral importation among multiple sources in an emerging epidemic.

Chapter 5 introduces a new simulation-based evaluation framework called SOPHI ("*Sandbox for Optimising genomic sampling for PHylogeographic Inference*"),

which I applied to address two open questions regarding the impact of heterogeneous sampling on discrete phylogeographic inference. In the first application, I demonstrated that the detection of viral importation events depends on both the sampling proportion and the underlying migration rate, confirming that findings from Chapter 3 can be generalised to more realistic outbreak conditions. I also showed that the extent to which independent transmission lineages become aggregated - a known phenomenon that contributes to the underestimation of the number of viral importation events - depends primarily on the number of sampled infections at the source location per detectable lineage. In the second application, I used SOPHI to explore how different sampling schemes affect the source attribution of early viral importation events in a multi-deme mobility network, under two outbreak scenarios with varying degrees of sampling bias. In addition to providing insights into how heterogeneous sampling of pathogen genomes gives rise to biased estimates of viral movement, these applications demonstrate the utility of SOPHI as a practical framework for guiding the systematic exploration of the impact of sampling biases and the design of more robust mitigation strategies. The SOPHI framework is open-source and freely available as a web application at http://www.sophi-oxf.io/.

1.7 References

1. Tulchinsky, T.H. and Varavikova, E.A. (2014) 'A History of Public Health', *The New Public Health*, p. 1.
2. Sampath, S., Khedr, A., Qamar, S., Tekin, A., Singh, R., Green, R. and Kashyap, R. (2021) 'Pandemics Throughout the History', *Cureus*, 13(9), p. e18136.
3. McNeill, W. (2010) *Plagues and Peoples*. Anchor.
4. Barbieri, R., Signoli, M., Chevé, D., Costedoat, C., Tzortzis, S., Aboudharam, G., Raoult, D. and Drancourt, M. (2020) 'Yersinia pestis: the Natural History of Plague', *Clinical Microbiology Reviews*, 34(1), pp. e00044–19.
5. Glatter, K.A. and Finkelman, P. (2020) 'History of the Plague: An Ancient Pandemic for the Age of COVID-19', *The American Journal of Medicine*, 134(2), p. 176.
6. Bramanti, B., Namouchi, A., Schmid, B.V., Dean, K.R. and Stenseth, N.C. (2019) 'Reply to Barbieri et al.: Out of the Land of Darkness: Plague on the fur trade routes', *Proceedings of the National Academy of Sciences*, 116(16), pp. 7622–7623.
7. Yersin, A. (1994) '[Bubonic plague in Hong Kong. 1894]', *Revue medicale de la Suisse romande*, 114(5), pp. 393–395.
8. 'Plague: History and contemporary analysis' (2013) *Journal of Infection*, 66(1), pp. 18–26.
9. Sloan, A.W. (1981) 'The Black Death in England', *South African Medical Journal*, 59(18), pp. 645-650.
10. Bailey, M. (1996) 'Demographic decline in late medieval England: some thoughts on recent research', *The Economic History Review*, 49(1), pp. 1-19.
11. Tognotti, E. (2013) 'Lessons from the History of Quarantine, from Plague to Influenza A', *Emerging Infectious Diseases*, 19(2), p. 254.
12. Bencard, A. (2021) 'Epidemics before microbiology: stories from the plague in 1711 and cholera in 1853 in Copenhagen', *Apmis*, 129(7), p. 372.
13. Acuna-Soto, R., Stahle, D.W., Cleaveland, M.K. and Therrell, M.D. (2002) 'Megadrought and Megadeath in 16th Century Mexico', *Emerging Infectious Diseases*, 8(4), p. 360.
14. Hopkins, D.R. (1983) *Princes and Peasants: Smallpox in History*.
15. Crosby, A.W. (1972) *The Columbian Exchange: Biological and Cultural Consequences of 1492*. Greenwood.
16. Pollitzer, R. (1954) 'Cholera studies: 1. History of the disease', *Bulletin of the World Health Organization*, 10(3), p. 421.
17. Arnold, D. (1986) 'Cholera and colonialism in British India', *Past & present*, (113), pp. 118–151.
18. Azizi, M.H. and Azizi, F. (2010) 'History of Cholera Outbreaks in Iran during the 19th and 20th Centuries', *Middle East Journal of Digestive Diseases*, 2(1), p. 51.
19. Spreeuwenberg, P., Kroneman, M. and Paget, J. (2018) 'Reassessing the Global Mortality Burden of the 1918 Influenza Pandemic', *American Journal of Epidemiology*, 187(12), p. 2561.
20. Johnson, N.P. and Mueller, J. (2002) Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bulletin of the history of medicine*, 76(1), pp. 105–115.
21. Strassburg M.A. (1982) The global eradication of smallpox. *American journal of infection control*, 10(2), pp. 53–59.
22. Cochi, S.L., Freeman, A., Guirguis, S., Jafari, H. and Aylward, B. (2014) 'Global polio eradication initiative: lessons learned and legacy', *The Journal of infectious diseases*, 210 Suppl 1(Suppl 1), pp. S540–6.

23. Minta, A.A., Ferrari, M., Antoni, S., Lambert, B., Sayi, T.S., Hsu, C.H., Steulet, C., Gacic-Dobo, M., Rota, P.A., Mulders, M.N., Wimmer, A., Bose, A.S., O'Connor, P. and Crowcroft, N.S. (2024) 'Progress Toward Measles Elimination — Worldwide, 2000–2023', *Morbidity and Mortality Weekly Report*, 73(45), p. 1036.
24. Coltart, C.E.M., Lindsey, B., Ghinai, I., Johnson, A.M. and Heymann, D.L. (2017) 'The Ebola outbreak, 2013–2016: old lessons for new epidemics', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **372,** 20160297.
25. Onyekuru, N.A., Ihemezie, E.J., Ezea, C.P., Apeh, C.C. and Onyekuru, B.O. (2023) 'Impacts of Ebola disease outbreak in West Africa: Implications for government and public health preparedness and lessons from COVID-19', *Scientific African*, 19, p. e01513.
26. Bell, B.P. (2016) 'Overview, Control Strategies, and Lessons Learned in the CDC Response to the 2014–2016 Ebola Epidemic', *MMWR Supplements*, 65. Available at: https://doi.org/10.15585/mmwr.su6503a2.
27. Takahashi, S., Metcalf, C.J.E., Ferrari, M.J., Moss, W.J., Truelove, S.A., Tatem, A.J., Grenfell, B.T. and Lessler, J. (2015) 'Reduced vaccination and the risk of measles and other childhood infections post-Ebola', *Science*, 347(6227), pp. 1240-1242.
28. Camara, B.S., Delamou, A., Diro, E., El Ayadi, M.A., Béavogui, A.H., Sidibé, S., Grovogui, F.M., Takarinda, K.C., Kolié, D., Sandouno, S.D., Okumura, J., Baldé, M.D., Van Griensven, J. and Zachariah, R. (2017) 'Influence of the 2014–2015 Ebola outbreak on the vaccination of children in a rural district of Guinea', *Public Health Action*, 7(2), p. 161.
29. Colizza, V., Barrat, A., Barthélemy, M. and Vespignani, A. (2007) 'Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study', *BMC Medicine*, 5(1), pp. 1–13.
30. Stadler, K., Masignani, V., Eickmann, M., Becker, S., Abrignani, S., Klenk, H.-D. and Rappuoli, R. (2003) 'SARS — beginning to understand a new virus', *Nature Reviews Microbiology*, 1(3), pp. 209–218.
31. Fraser, C., Donnelly, C.A., Cauchemez, S., Hanage, W.P., Van Kerkhove, M.D., Déirdre Hollingsworth, T., Griffin, J., Baggaley, R.F., Jenkins, H.E., Lyons, E.J., Jombart, T., Hinsley, W.R., Grassly, N.C., Balloux, F., Ghani, A.C., Ferguson, N.M., Rambaut, A., Pybus, O.G., Lopez-Gatell, H., Alpuche-Aranda, C.M., Chapela, I.B., Zavala, E.P., Guevara, D.M.E., Checchi, F., Garcia, E., Hugonnet, S., Roth, C. and The WHO Rapid Pandemic Assessment Collaboration (2009) 'Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings', *Science*, 324(5934), pp. 1557-1561
32. Bajardi, P., Poletto, C., Ramasco, J.J., Tizzoni, M., Colizza, V. and Vespignani, A. (2011) 'Human Mobility Networks, Travel Restrictions, and the Global Spread of 2009 H1N1 Pandemic', *PLoS ONE*, 6(1), p. e16591.
33. Bogoch, I.I., Brady, O.J., Kraemer, M.U.G., German, M., Creatore, M.I., Brent, S., Watts, A.G., Hay, S.I., Kulkarni, M.A., Brownstein, J.S. and Khan, K. (2016) 'Potential for Zika virus introduction and transmission in resource limited countries in Africa and Asia-Pacific: A modeling study', *The Lancet. Infectious diseases*, 16(11), pp. 1237-1245.
34. Zhang, Q., Sun, K., Chinazzi, M., Pastore y Piontti, A., Dean, N.E., Rojas, D.P., Merler, S., Mistry, D., Poletti, P., Rossi, L., Bray, M., Elizabeth Halloran, M., Longini, I.M., Jr and Vespignani, A. (2017) 'Spread of Zika virus in the Americas', *Proceedings of the National Academy of Sciences of the United States of America*, 114(22), p. E4334.
35. Tegally, H., Wilkinson, E., Tsui, J.L., Moir, M., Martin, D., Brito, A.F., Giovanetti, M., Khan, K., Huber, C., Bogoch, I.I., San, J.E., Poongavanan, J., Xavier, J.S.,

Candido, D.D.S., Romero, F., Baxter, C., Pybus, O.G., Lessells, R.J., Faria, N.R., Kraemer, M.U.G. and de Oliveira, T. (2023) 'Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern', *Cell*, 186(15), pp. 3277-3290.e16
36. Li, J., Lai, S., Gao, G.F. and Shi, W. (2021) 'The emergence, genomic diversity and global spread of SARS-CoV-2', *Nature*, 600(7889), pp. 408–418.
37. Petersen, E., Koopmans, M., Go, U., Hamer, D.H., Petrosillo, N., Castelli, F., Storgaard, M., Al Khalili, S. and Simonsen, L. (2020) 'Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics', *The Lancet. Infectious diseases*, 20(9), pp. e238-e244.
38. He, X., Lau, E.H.Y., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y.C., Wong, J.Y., Guan, Y., Tan, X., Mo, X., Chen, Y., Liao, B., Chen, W., Hu, F., Zhang, Q., Zhong, M., Wu, Y., Zhao, L., Zhang, F., Cowling, B.J., Li, F. and Leung, G.M. (2020) 'Temporal dynamics in viral shedding and transmissibility of COVID-19', *Nature Medicine*, 26(5), pp. 672–675.
39. Rajgor, D.D., Lee, M.H., Archuleta, S., Bagdasarian, N. and Quek, S.C. (2020) 'The many estimates of the COVID-19 case fatality rate', *The Lancet. Infectious diseases*, 20(7), pp. 776-777.
40. Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Gavrilov, D., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E. and Roser, M. (2020) 'Mortality Risk of COVID-19', *Our World in Data*. Available at: https://ourworldindata.org/mortality-risk-covid (Accessed: 23 March 2025).
41. Piret, J. and Boivin, G. (2021) 'Pandemics Throughout History', *Frontiers in Microbiology*, 11, p. 631736.
42. *COVID-19 circulation* (no date) *datadot*. Available at: https://data.who.int/dashboards/tuberculosis/tuberculosis-epidemiological-profile/covid-vums-line-chart (Accessed: 1 May 2025).
43. Msemburi, W., Karlinsky, A., Knutson, V., Aleshin-Guendel, S., Chatterji, S. and Wakefield, J. (2022) 'The WHO estimates of excess mortality associated with the COVID-19 pandemic', *Nature*, 613(7942), pp. 130–137.
44. Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Gavrilov, D., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E. and Roser, M. (2020) 'Excess mortality during the Coronavirus pandemic (COVID-19)', *Our World in Data*. Available at: https://ourworldindata.org/excess-mortality-covid (Accessed: 23 March 2025).
45. Guogui, H., Fei, G., Lihua, L., Lucy, T., Zhiming, C., Massimiliano, T., Zimmermann Klaus F., Marika, F. and Silva, S.S.M. (2024) 'Changing impact of COVID-19 on life expectancy 2019–2023 and its decomposition: Findings from 27 countries', *SSM - Population Health*, 25, p. 101568.
46. GBD 2021 Demographics Collaborators (2024) 'Global age-sex-specific mortality, life expectancy, and population estimates in 204 countries and territories and 811 subnational locations, 1950-2021, and the impact of the COVID-19 pandemic: a comprehensive demographic analysis for the Global Burden of Disease Study 2021', *Lancet (London, England)*, 403(10440), pp. 1989-2056.
47. COVID-19 Mental Disorders Collaborators (2021) 'Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic', *Lancet*, 398(10312), pp. 1700-1712.
48. Bueno-Notivol, J., Gracia-García, P., Olaya, B., Lasheras, I., López-Antón, R. and Santabárbara, J. (2021) 'Prevalence of depression during the COVID-19 outbreak: A

meta-analysis of community-based studies', *International journal of clinical and health psychology*, 21(1), 100196.
49. Jørgensen, F., Bor, A., Rasmussen, M.S., Lindholt, M.F. and Petersen, M.B. (2022) 'Pandemic fatigue fueled political discontent during the COVID-19 pandemic', *Proceedings of the National Academy of Sciences*, 119(48), p. e2201266119.
50. Eichengreen, B., Saka, O. and Aksoy, C.G. (2020) *The Political Scar of Epidemics*. w27401. National Bureau of Economic Research. Available at: https://doi.org/10.3386/w27401.
51. Shet, A., Carr, K., Danovaro-Holliday, M.C., Sodha, S.V., Prosperi, C., Wunderlich, J., Wonodi, C., Reynolds, H.W., Mirza, I., Gacic-Dobo, M., O'Brien, K.L. and Lindstrand, A. (2022) 'Impact of the SARS-CoV-2 pandemic on routine immunisation services: evidence of disruption and recovery from 170 countries and territories', *The Lancet. Global health*, 10(2), pp. e186-e194.
52. Moyo, S., Ashok, A., Myers, L., Nyankieya, R., Sharma, S. and Prasad, R. (2024) 'The impact of COVID-19 on routine child immunisation in South Africa', *BMC Public Health*, 24(1), pp. 1–9.
53. Causey, K., Fullman, N., Sorensen, R.J.D., Galles, N.C., Zheng, P., Aravkin, A., Danovaro-Holliday, M.C., Martinez-Piedra, R., Sodha, S.V., Velandia-González, M.P., Gacic-Dobo, M., Castro, E., He, J., Schipp, M., Deen, A., Hay, S.I., Lim, S.S. and Mosser, J.F. (2021) 'Estimating global and regional disruptions to routine childhood vaccine coverage during the COVID-19 pandemic in 2020: a modelling study', *Lancet (London, England)*, 398(10299), pp. 522-534.
54. McKibbin, W. and Fernando, R. (2023) 'The global economic impacts of the COVID-19 pandemic', *Economic Modelling*, 129, p. 106551.
55. Faramarzi, A., Norouzi, S., Dehdarirad, H., Aghlmand, S., Yusefzadeh, H. and Javan-Noughabi, J. (2024) 'The global economic burden of COVID-19 disease: a comprehensive systematic review and meta-analysis', *Systematic Reviews*, 13(1), pp. 1–10.
56. Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M. and Agha, R. (2020) 'The socio-economic implications of the coronavirus pandemic (COVID-19): A review', *International Journal of Surgery (London, England)*, 78, pp. 185-193.
57. Das, K., Behera, R.L. and Paital, B. (2022) 'Socio-economic impact of COVID-19', *COVID-19 in the Environment*. Elsevier, pp. 153–190.
58. Baker, R.E., Mahmud, A.S., Miller, I.F., Rajeev, M., Rasambainarivo, F., Rice, B.L., Takahashi, S., Tatem, A.J., Wagner, C.E., Wang, L.-F., Wesolowski, A. and Metcalf, C.J.E. (2021) 'Infectious disease in an era of global change', *Nature Reviews Microbiology*, 20(4), pp. 193–205.
59. Carlson, C.J., Albery, G.F., Merow, C., Trisos, C.H., Zipfel, C.M., Eskew, E.A., Olival, K.J., Ross, N. and Bansal, S. (2022) 'Climate change increases cross-species viral transmission risk', *Nature*, 607(7919), pp. 555–562.
60. Tsui, J.L.-H., Pena, R.E., Moir, M., Inward, R.P.D., Wilkinson, E., San, J.E., Poongavanan, J., Bajaj, S., Gutierrez, B., Dasgupta, A., de Oliveira, T., Kraemer, M.U.G., Tegally, H. and Sambaturu, P. (2024) 'Impacts of climate change-related human migration on infectious diseases', *Nature Climate Change*, 14(8), pp. 793–802.
61. Findlater, A. and Bogoch, I.I. (2018) 'Human Mobility and the Global Spread of Infectious Diseases: A Focus on Air Travel', *Trends in Parasitology*, 34(9), p. 772.

62. Zhang, Z., Zhang, Q., Wang, T., Xu, N., Lu, T., Hong, W., Penuelas, J., Gillings, M., Wang, M., Gao, W. and Qian, H. (2022) 'Assessment of global health risk of antibiotic resistance genes', *Nature Communications*, 13(1), pp. 1–11.
63. Levy, S.B. and Marshall, B. (2004) 'Antibacterial resistance worldwide: causes, challenges and responses', *Nature Medicine*, 10(12), pp. S122–S129.
64. *Mpox outbreak* (no date). Available at: https://www.who.int/emergencies/situations/mpox-outbreak (Accessed: 1 May 2025).
65. World Health Organization (2025) *Global Mpox Trends*. Available at: https://worldhealthorg.shinyapps.io/mpx_global/#8_Disclaimers (Accessed: 1 May 2025).
66. Nowogrodzki, J. (2025) *Measles is surging in the US: how bad could it get?*, *Nature Publishing Group UK*. Available at: https://doi.org/10.1038/d41586-025-00786-w (Accessed: 1 May 2025).
67. Gambrell, A., Sundaram, M. and Bednarczyk, R.A. (2022) 'Estimating the number of US children susceptible to measles resulting from COVID-19-related vaccination coverage declines', *Vaccine*, 40(32), pp. 4574-4579.
68. CDC (2025) *Measles Cases and Outbreaks*, *Measles (Rubeola)*. Available at: https://www.cdc.gov/measles/data-research/index.html (Accessed: 23 March 2025).
69. Snow, J. (1991) '[On the mode of communication of cholera. 1855]', *Salud publica de Mexico*, 33(2), pp. 194–201.
70. Alirol, E., Getaz, L., Stoll, B., Chappuis, F. and Loutan, L. (2011) 'Urbanisation and infectious diseases in a globalised world', *The Lancet. Infectious diseases*, 11(2), pp. 131-41.
71. Sokolow, S.H., Nova, N., Jones, I.J., Wood, C.L., Lafferty, K.D., Garchitorena, A., Hopkins, S.R., Lund, A.J., MacDonald, A.J., LeBoa, C., Peel, A.J., Mordecai, E.A., Howard, M.E., Buck, J.C., Lopez-Carr, D., Barry, M., Bonds, M.H. and De Leo, G.A. (2022) 'Ecological and socioeconomic factors associated with the human burden of environmentally mediated pathogens: a global analysis', *The Lancet. Planetary health*, 6(11), pp. 870-879.
72. Jedwab, R., Loungani, P. and Yezar, A. (2021) 'Comparing cities in developed and developing countries: Population, land area, building height and crowding', *Regional Science and Urban Economics*, 86, p. 103609.
73. Hargreaves, J.R., Boccia, D., Evans, C.A., Adato, M., Petticrew, M. and Porter, J.D.H. (2011) 'The Social Determinants of Tuberculosis: From Evidence to Action', *American Journal of Public Health*, 101(4), p. 654-62.
74. Duarte, R., Lönnroth, K., Carvalho, C., Lima, F., Carvalho, A.C.C., Muñoz-Torrico, M., and Centis, R. (2018) 'Tuberculosis, social determinants and co-morbidities (including HIV)'. *Pulmonology*, *24*(2), 115–119.
75. Richterman, A., Sainvilien, D.R., Eberly, L. and Ivers, L.C. (2018) 'Individual and Household Risk Factors for Symptomatic Cholera Infection: A Systematic Review and Meta-analysis', *The Journal of infectious diseases*, 218(suppl_3), pp. 154-164.
76. Lee, J.-S., Mogasale, V.V., Mogasale, V. and Lee, K. (2016) 'Geographical distribution of typhoid risk factors in low and middle income countries', *BMC Infectious Diseases*, 16(1), pp. 1–10.
77. del Pilar Fernández, M., Gaspe, M.S. and Gürtler, R.E. (2019) 'Inequalities in the social determinants of health and Chagas disease transmission risk in indigenous and creole households in the Argentine Chaco', *Parasites & Vectors*, 12, p. 184.
78. Newman, K.L.S. (2012) 'Shutt up: bubonic plague and quarantine in early modern England', *Journal of social history*, 45(3), pp. 809–834.

79. Swanson, M.W. (1977) 'The Sanitation Syndrome: Bubonic Plague and Urban Native Policy in the Cape Colony, 1900–19091', *The Journal of African History*, 18(3), pp. 387–410.
80. Yue, R.P.H., Lee, H.F. and Wu, C.Y.H. (2017) 'Trade routes and plague transmission in pre-industrial Europe', *Scientific Reports*, 7(1), pp. 1–10.
81. Cliff, A.D. and Haggett, P. (1989) 'Spatial aspects of epidemic control', *Progress in Human Geography*, 13(3), pp. 315-347.
82. Sigler, T., Mahmuda, S., Kimpton, A., Loginova, J., Wohland, P., Charles-Edwards, E. and Corcoran, J. (2021) 'The socio-spatial determinants of COVID-19 diffusion: the impact of globalisation, settlement characteristics and population', *Globalization and Health*, 17(1), pp. 1–14.
83. Gould, P.R. (1969) 'Spatial Diffusion, Resource Paper No. 4'. Washington D.C.: Commission on College Geography. Available at: http://files.eric.ed.gov/fulltext/ED120029.pdf (Accessed: 1 May 2025).
84. Benedict, C.A. (1996) *Bubonic Plague in Nineteenth-Century China*. Stanford University Press.
85. Marvel, S.A., Martin, T., Doering, C.R., Lusseau, D. and Newman, M.E.J. (2013) 'The small-world effect is a modern phenomenon', *arXiv*. Available at: http://arxiv.org/abs/1310.2636 (Accessed: 23 March 2025).
86. Pyle, G.F. (1969) 'The diffusion of cholera in the United States in the nineteenth century', *Geographical analysis*, 1(1), pp. 59-75.
87. Dudas, G., Carvalho, L.M., Bedford, T., Tatem, A.J., Baele, G., Faria, N.R., Park, D.J., Ladner, J.T., Arias, A., Asogun, D., Bielejec, F., Caddy, S.L., Cotten, M., D'Ambrozio, J., Dellicour, S., Di Caro, A., Diclaro, J.W., Duraffour, S., Elmore, M.J., Fakoli, L.S., Faye, O., Gilbert, M.L., Gevao, S.M., Gire, S., Gladden-Young, A., Gnirke, A., Goba, A., Grant, D.S., Haagmans, B.L., Hiscox, J.A., Jah, U., Kugelman, J.R., Liu, D., Lu, J., Malboeuf, C.M., Mate, S., Matthews, D.A., Matranga, C.B., Meredith, L.W., Qu, J., Quick, J., Pas, S.D., Phan, M.V.T., Pollakis, G., Reusken, C.B., Sanchez-Lockhart, M., Schaffner, S.F., Schieffelin, J.S., Sealfon, R.S., Simon-Loriere, E., Smits, S.L., Stoecker, K., Thorne, L., Tobin, E.A., Vandi, M.A., Watson, S.J., West, K., Whitmer, S., Wiley, M.R., Winnicki, S.M., Wohl, S., Wölfel, R., Yozwiak, N.L., Andersen, K.G., Blyden, S.O., Bolay, F., Carroll, M.W., Dahn, B., Diallo, B., Formenty, P., Fraser, C., Gao, G.F., Garry, R.F., Goodfellow, I., Günther, S., Happi, C.T., Holmes, E.C., Kargbo, B., Keïta, S., Kellam, P., Koopmans, M.P.G., Kuhn, J.H., Loman, N.J., Magassouba, N. 'faly, Naidoo, D., Nichol, S.T., Nyenswah, T., Palacios, G., Pybus, O.G., Sabeti, P.C., Sall, A., Ströher, U., Wurie, I., Suchard, M.A., Lemey, P. and Rambaut, A. (2017) 'Virus genomes reveal factors that spread and sustained the Ebola epidemic', *Nature*, 544(7650), pp. 309–315.
88. Kramer, A.M., Tomlin Pulliam, J., Alexander, L.W., Park, A.W., Rohani, P. and Drake, J.M. (2016) 'Spatial spread of the West Africa Ebola epidemic', *Royal Society Open Science*, 3(8), 160294.
89. Dalziel, B.D., Kissler, S., Gog, J.R., Viboud, C., Bjørnstad, O.N., Metcalf, C.J.E. and Grenfell, B.T. (2018) 'Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities', *Science*, 362(6410), pp. 75-79.
90. Rader, B., Scarpino, S.V., Nande, A., Hill, A.L., Adlam, B., Reiner, R.C., Pigott, D.M., Gutierrez, B., Zarebski, A.E., Shrestha, M., Brownstein, J.S., Castro, M.C., Dye, C., Tian, H., Pybus, O.G. and Kraemer, M.U.G. (2020) 'Crowding and the shape of COVID-19 epidemics', *Nature Medicine*, 26(12), pp. 1829–1834.
91. Pastor-Satorras, R., Van Mieghem Piet, C.C. and Vespignani, A. (2015) 'Epidemic processes in complex networks', *Reviews of Modern Physics*, 87(3), pp. 925–979.

92. Viboud, C., Bjørnstad, O.N., Smith, D.L., Simonsen, L., Miller, M.A. and Grenfell, B.T. (2006) 'Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza', *Science*, 312(5772), pp. 447-451.
93. Brockmann, D. and Helbing, D. (2013) 'The Hidden Geometry of Complex, Network-Driven Contagion Phenomena', *Science*, 342(6164), pp. 1337-1342.
94. Klamser, P.P., Zachariae, A., Maier, B.F., Baranov, O., Jongen, C., Schlosser, F. and Brockmann, D. (2024) 'Inferring country-specific import risk of diseases from the world air transportation network', *PLOS Computational Biology*, 20(1), p. e1011775.
95. Poletto, C., Pelat, C., Levy-Bruhl, D., Yazdanpanah, Y., Boelle, P.Y. and Colizza, V. (2014) 'Assessment of the Middle East respiratory syndrome coronavirus (MERS-CoV) epidemic in the Middle East and risk of international spread using a novel maximum likelihood analysis approach', *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, 19(23), 20824.
96. Watts, D.J., Muhamad, R., Medina, D.C. and Dodds, P.S. (2005) 'Multiscale, resurgent epidemics in a hierarchical metapopulation model', *Proceedings of the National Academy of Sciences*, 102(32), pp. 11157–11162.
97. Dhar, M.S., Marwal, R., Vs, R., Ponnusamy, K., Jolly, B., Bhoyar, R.C., Sardana, V., Naushin, S., Rophina, M., Mellan, T.A., Mishra, S., Whittaker, C., Fatihi, S., Datta, M., Singh, P., Sharma, U., Ujjainiya, R., Bhatheja, N., Divakar, M.K., Singh, M.K., Imran, M., Senthivel, V., Maurya, R., Jha, N., Mehta, P., A, V., Sharma, P., Vr, A., Chaudhary, U., Soni, N., Thukral, L., Flaxman, S., Bhatt, S., Pandey, R., Dash, D., Faruq, M., Lall, H., Gogia, H., Madan, P., Kulkarni, S., Chauhan, H., Sengupta, S., Kabra, S., Indian SARS-CoV-2 Genomics Consortium (INSACOG)‡, Gupta, R.K., Singh, S.K., Agrawal, A., Rakshit, P., Nandicoori, V., Tallapaka, K.B., Sowpati, D.T., Thangaraj, K., Bashyam, M.D., Dalal, A., Sivasubbu, S., Scaria, V., Parida, A., Raghav, S.K., Prasad, P., Sarin, A., Mayor, S., Ramakrishnan, U., Palakodeti, D., Seshasayee, A.S.N., Bhat, M., Shouche, Y., Pillai, A., Dikid, T., Das, S., Maitra, A., Chinnaswamy, S., Biswas, N.K., Desai, A.S., Pattabiraman, C., Manjunatha, M.V., Mani, R.S., Arunachal Udupi, G., Abraham, P., Atul, P.V. and Cherian, S.S. (2021) 'Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India', *Science*, 374(6570), pp. 995–999.
98. Tegally, H., Wilkinson, E., Althaus, C.L., Giovanetti, M., San, J.E., Giandhari, J., Pillay, S., Naidoo, Y., Ramphal, U., Msomi, N., Mlisana, K., Amoako, D.G., Everatt, J., Mohale, T., Nguni, A., Mahlangu, B., Ntuli, N., Khumalo, Z.T., Makatini, Z., Wolter, N., Scheepers, C., Ismail, A., Doolabh, D., Joseph, R., Strydom, A., Mendes, A., Davis, M., Mayaphi, S.H., Ramphal, Y., Maharaj, A., Karim, W.A., Tshiabuila, D., Anyaneji, U.J., Singh, L., Engelbrecht, S., Fonseca, V., Marais, K., Korsman, S., Hardie, D., Hsiao, N., Maponga, T., van Zyl, G., Marais, G., Iranzadeh, A., Martin, D., Alcantara, L.C.J., Bester, P.A., Nyaga, M.M., Subramoney, K., Treurnicht, F.K., Venter, M., Goedhals, D., Preiser, W., Bhiman, J.N., Av, G., Williamson, C., Lessells, R.J. and de Oliveira, T. (2021) 'Rapid replacement of the Beta variant by the Delta variant in South Africa', *medRxiv*. Available at: https://doi.org/10.1101/2021.09.23.21264018.
99. Carabelli, A.M., Peacock, T.P., Thorne, L.G., Harvey, W.T., Hughes, J., Peacock, S.J., Barclay, W.S., de Silva, T.I., Towers, G.J. and Robertson, D.L. (2023) 'SARS-CoV-2 variant biology: immune escape, transmission and fitness', *Nature Reviews Microbiology*, 21(3), pp. 162–177.
100. Viana, R., Moyo, S., Amoako, D.G., Tegally, H., Scheepers, C., Althaus, C.L., Anyaneji, U.J., Bester, P.A., Boni, M.F., Chand, M., Choga, W.T., Colquhoun, R.,

Davids, M., Deforche, K., Doolabh, D., du Plessis, L., Engelbrecht, S., Everatt, J., Giandhari, J., Giovanetti, M., Hardie, D., Hill, V., Hsiao, N.-Y., Iranzadeh, A., Ismail, A., Joseph, C., Joseph, R., Koopile, L., Kosakovsky Pond, S.L., Kraemer, M.U.G., Kuate-Lere, L., Laguda-Akingba, O., Lesetedi-Mafoko, O., Lessells, R.J., Lockman, S., Lucaci, A.G., Maharaj, A., Mahlangu, B., Maponga, T., Mahlakwane, K., Makatini, Z., Marais, G., Maruapula, D., Masupu, K., Matshaba, M., Mayaphi, S., Mbhele, N., Mbulawa, M.B., Mendes, A., Mlisana, K., Mnguni, A., Mohale, T., Moir, M., Moruisi, K., Mosepele, M., Motsatsi, G., Motswaledi, M.S., Mphoyakgosi, T., Msomi, N., Mwangi, P.N., Naidoo, Y., Ntuli, N., Nyaga, M., Olubayo, L., Pillay, S., Radibe, B., Ramphal, Y., Ramphal, U., San, J.E., Scott, L., Shapiro, R., Singh, L., Smith-Lawrence, P., Stevens, W., Strydom, A., Subramoney, K., Tebeila, N., Tshiabuila, D., Tsui, J., van Wyk, S., Weaver, S., Wibmer, C.K., Wilkinson, E., Wolter, N., Zarebski, A.E., Zuze, B., Goedhals, D., Preiser, W., Treurnicht, F., Venter, M., Williamson, C., Pybus, O.G., Bhiman, J., Glass, A., Martin, D.P., Rambaut, A., Gaseitsiwe, S., von Gottberg, A. and de Oliveira, T. (2022) 'Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa', *Nature*, 603(7902), pp. 679–686.

101. Wang, C.C., Prather, K.A., Sznitman, J., Jimenez, J.L., Lakdawala, S.S., Tufekci, Z. and Marr, L.C. (2021) 'Airborne transmission of respiratory viruses', *Science*, 373(6558), p. eabd9149.
102. Yang, W. and Marr, L.C. (2011) 'Dynamics of Airborne Influenza A Viruses Indoors and Dependence on Humidity', *PLOS ONE*, 6(6), p. e21481.
103. Nichols, G.L., Gillingham, E.L., Macintyre, H.L., Vardoulakis, S., Hajat, S., Sarran, C.E., Amankwaah, D. and Phalkey, R. (2021) 'Coronavirus seasonality, respiratory infections and weather', *BMC Infectious Diseases*, 21(1), pp. 1–15.
104. Wang, J., Tang, K., Feng, K., Lin, X., Lv, W., Chen, K. and Wang, F. (2021) 'Impact of temperature and relative humidity on the transmission of COVID-19: a modelling study in China and the United States', *BMJ Open*, 11(2), p. e043863.
105. Paynter, S., Ware, R. S., Sly, P. D., Williams, G. and Weinstein, P. (2015). 'Seasonal immune modulation in humans: observed patterns and potential environmental drivers', *The Journal of infection*, *70*(1), pp. 1–10.
106. Dowell, S.F. and Ho, M.S. (2004) 'Seasonality of infectious diseases and severe acute respiratory syndrome-what we don't know can hurt us', *The Lancet. Infectious diseases*, 4(11), pp. 704-708.
107. Colón-González, F.J., Sewe, M.O., Tompkins, A.M., Sjödin, H., Casallas, A., Rocklöv, J., Caminade, C. and Lowe, R. (2021) 'Projecting the risk of mosquito-borne diseases in a warmer and more populated world: a multi-model, multi-scenario intercomparison modelling study', *The Lancet. Planetary health*, 5(7), pp. e404-e414.
108. Parham, P.E. and Michael, E. (2010) 'Modeling the effects of weather and climate change on malaria transmission', *Environmental health perspectives*, 118(5), pp. 620-626.
109. Chen, Y., Xu, Y., Wang, L., Liang, Y., Li, N., Lourenço, J., Yang, Y., Lin, Q., Wang, L., Zhao, H., Cazelles, B., Song, H., Liu, Z., Wang, Z., Brady, O.J., Cauchemez, S. and Tian, H. (2024) 'Indian Ocean temperature anomalies predict long-term global dengue trends', *Science*, 384(6696), pp. 639-646.
110. Kraemer, M.U.G., Sinka, M.E., Duda, K.A., Mylne, A.Q.N., Shearer, F.M., Barker, C.M., Moore, C.G., Carvalho, R.G., Coelho, G.E., Van Bortel, W., Hendrickx, G., Schaffner, F., Elyazar, I.R.F., Teng, H.-J., Brady, O.J., Messina, J.P., Pigott, D.M., Scott, T.W., Smith, D.L., Wint, G.R.W., Golding, N. and Hay, S.I.

(2015) ‘The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus’, *eLife*, 4, p. e08347.
111. Harcourt, S.E., Smith, G.E., Elliot, A.J., Pebody, R., Charlett, A., Ibbotson, S., Regan, M. and Hippisley-Cox, J. (2012) ‘Use of a large general practice syndromic surveillance system to monitor the progress of the influenza A(H1N1) pandemic 2009 in the UK’, *Epidemiology & Infection*, 140(1), pp. 100–105.
112. *Independent review into the response to the 2009 swine flu pandemic* (2010) *GOV.UK*. Available at: https://www.gov.uk/government/publications/independent-review-into-the-response-to-the-2009-swine-flu-pandemic (Accessed: 23 March 2025).
113. Sethi, M. and Pebody, R. (2010) *Pandemic H1N1 (Swine) Influenza Vaccine Uptake amongst Patient Groups in Primary Care in England 2009/10*. Department of Health. Available at https://assets.publishing.service.gov.uk/media/5a7cad05ed915d7c983bc39c/dh_121014.pdf (Accessed: 30 April 2025).
114. Bogoch, I.I., Watts, A., Thomas-Bachli, A., Huber, C., Kraemer, M.U.G. and Khan, K. (2020) ‘Pneumonia of unknown aetiology in Wuhan, China: potential for international spread via commercial air travel’, *Journal of travel medicine*, 27(2), taaa008.
115. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F. and Tan, W. (2020) ‘A Novel Coronavirus from Patients with Pneumonia in China, 2019’, *The New England journal of medicine*, 382(8), pp. 727-733.
116. Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C. and Zhang, Y.Z. (2020) ‘A new coronavirus associated with human respiratory disease in China’, *Nature*, 579(7798), pp. 265-269.
117. Hill, V., Githinji, G., Vogels, C.B.F., Bento, A.I., Chaguza, C., Carrington, C.V.F. and Grubaugh, N.D. (2023) ‘Toward a global virus genomic surveillance network’, *Cell host & microbe*, 31(6), pp. 861-873.
118. Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E.J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A.J., Engelbrecht, S., Van Zyl, G., Preiser, W., Petruccione, F., Sigal, A., Hardie, D., Marais, G., Hsiao, N.-Y., Korsman, S., Davies, M.-A., Tyers, L., Mudau, I., York, D., Maslo, C., Goedhals, D., Abrahams, S., Laguda-Akingba, O., Alisoltani-Dehkordi, A., Godzik, A., Wibmer, C.K., Sewell, B.T., Lourenço, J., Alcantara, L.C.J., Kosakovsky Pond, S.L., Weaver, S., Martin, D., Lessells, R.J., Bhiman, J.N., Williamson, C. and de Oliveira, T. (2021) ‘Detection of a SARS-CoV-2 variant of concern in South Africa’, *Nature*, 592(7854), pp. 438–443.
119. Volz, E., Mishra, S., Chand, M., Barrett, J.C., Johnson, R., Geidelberg, L., Hinsley, W.R., Laydon, D.J., Dabrera, G., O’Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C.V., Boyd, O., Loman, N.J., McCrone, J.T., Gonçalves, S., Jorgensen, D., Myers, R., Hill, V., Jackson, D.K., Gaythorpe, K., Groves, N., Sillitoe, J., Kwiatkowski, D.P., Flaxman, S., Ratmann, O., Bhatt, S., Hopkins, S., Gandy, A., Rambaut, A. and Ferguson, N.M. (2021) ‘Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England’, *Nature*, 593(7858), pp. 266–269.
120. Washington, N.L., Gangavarapu, K., Zeller, M., Bolze, A., Cirulli, E.T., Km, S.B., Larsen, B.B., Anderson, C., White, S., Cassens, T., Jacobs, S., Levan, G.,

Nguyen, J., Ramirez, J.M., Rivera-Garcia, C., Sandoval, E., Wang, X., Wong, D., Spencer, E., Robles-Sikisaka, R., Kurzban, E., Hughes, L.D., Deng, X., Wang, C., Servellita, V., Valentine, H., De Hoff, P., Seaver, P., Sathe, S., Gietzen, K., Sickler, B., Antico, J., Hoon, K., Liu, J., Harding, A., Bakhtar, O., Basler, T., Austin, B., MacCannell, D., Isaksson, M., Febbo, P.G., Becker, D., Laurent, M., McDonald, E., Yeo, G.W., Knight, R., Laurent, L.C., de Feo, E., Worobey, M., Chiu, C.Y., Suchard, M.A., Lu, J.T., Lee, W. and Andersen, K.G. (2021) 'Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States', *Cell*, 184(10), pp. 2587-2594.e7.

121. Chiu, C.Y. and Miller, S.A. (2019) 'Clinical metagenomics', *Nature Reviews Genetics*, 20(6), pp. 341–355.
122. Gu, W., Deng, X., Lee, M., Sucu, Y.D., Arevalo, S., Stryke, D., Federman, S., Gopez, A., Reyes, K., Zorn, K., Sample, H., Yu, G., Ishpuniani, G., Briggs, B., Chow, E.D., Berger, A., Wilson, M.R., Wang, C., Hsu, E., Miller, S., DeRisi, J.L. and Chiu, C.Y. (2020) 'Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids', *Nature Medicine*, 27(1), pp. 115–124.
123. Ko, K.K.K., Chng, K.R. and Nagarajan, N. (2022) 'Metagenomics-enabled microbial surveillance', *Nature Microbiology*, 7(4), pp. 486–496.
124. Miller, R.R., Montoya, V., Gardy, J.L., Patrick, D.M. and Tang, P. (2013) 'Metagenomics for pathogen detection in public health', *Genome Medicine*, 5(9), pp. 1–14.
125. Morvan, M., Jacomo, A.L., Souque, C., Wade, M.J., Hoffmann, T., Pouwels, K., Lilley, C., Singer, A.C., Porter, J., Evens, N.P., Walker, D.I., Bunce, J.T., Engeli, A., Grimsley, J., O'Reilly, K.M. and Danon, L. (2022) 'An analysis of 45 large-scale wastewater sites in England to estimate SARS-CoV-2 community prevalence', *Nature Communications*, 13(1), pp. 1–9.
126. Amman, F., Markt, R., Endler, L., Hupfauf, S., Agerer, B., Schedl, A., Richter, L., Zechmeister, M., Bicher, M., Heiler, G., Triska, P., Thornton, M., Penz, T., Senekowitsch, M., Laine, J., Keszei, Z., Klimek, P., Nägele, F., Mayr, M., Daleiden, B., Steinlechner, M., Niederstätter, H., Heidinger, P., Rauch, W., Scheffknecht, C., Vogl, G., Weichlinger, G., Wagner, A.O., Slipko, K., Masseron, A., Radu, E., Allerberger, F., Popper, N., Bock, C., Schmid, D., Oberacher, H., Kreuzinger, N., Insam, H. and Bergthaler, A. (2022) 'Viral variant-resolved wastewater surveillance of SARS-CoV-2 at national scale', *Nature Biotechnology*, 40(12), pp. 1814–1822.
127. Diamond, M.B., Keshaviah, A., Bento, A.I., Conroy-Ben, O., Driver, E.M., Ensor, K.B., Halden, R.U., Hopkins, L.P., Kuhn, K.G., Moe, C.L., Rouchka, E.C., Smith, T., Stevenson, B.S., Susswein, Z., Vogel, J.R., Wolfe, M.K., Stadler, L.B. and Scarpino, S.V. (2022) 'Wastewater surveillance of pathogens can inform public health responses', *Nature Medicine*, 28(10), pp. 1992–1995.
128. St-Onge, G., Davis, J.T., Hébert-Dufresne, L., Allard, A., Urbinati, A., Scarpino, S.V., Chinazzi, M. and Vespignani, A. (2025) 'Pandemic monitoring with global aircraft-based wastewater surveillance networks', *Nature Medicine*, 31(3), pp. 788–796.
129. Li, J., Hosegood, I., Powell, D., Tscharke, B., Lawler, J., Thomas, K.V. and Mueller, J.F. (2023) 'A global aircraft-based wastewater genomic surveillance network for early warning of future pandemics', *The Lancet. Global health*, 11(5), pp. 791-795.
130. Tsao, S.F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L. and Butt, Z.A. (2021) 'What social media told us in the time of COVID-19: a scoping review', *The Lancet. Digital health*, 3(3), pp. 175-194.

131. Menni, C., Valdes, A.M., Freidin, M.B., Sudre, C.H., Nguyen, L.H., Drew, D.A., Ganesh, S., Varsavsky, T., Cardoso, M.J., El-Sayed Moustafa, J.S., Visconti, A., Hysi, P., Bowyer, R.C.E., Mangino, M., Falchi, M., Wolf, J., Ourselin, S., Chan, A.T., Steves, C.J. and Spector, T.D. (2020) 'Real-time tracking of self-reported symptoms to predict potential COVID-19', *Nature Medicine*, 26(7), pp. 1037–1040.
132. Varsavsky, T., Graham, M.S., Canas, L.S., Ganesh, S., Capdevila Pujol, J., Sudre, C.H., Murray, B., Modat, M., Jorge Cardoso, M., Astley, C.M., Drew, D.A., Nguyen, L.H., Fall, T., Gomez, M.F., Franks, P.W., Chan, A.T., Davies, R., Wolf, J., Steves, C.J., Spector, T.D. and Ourselin, S. (2021) 'Detecting COVID-19 infection hotspots in England using large-scale self-reported data from a mobile application: a prospective, observational study', *The Lancet. Public health*, 6(1), pp. e21–e29.
133. Menni, C., Valdes, A.M., Polidori, L., Antonelli, M., Penamakuri, S., Nogal, A., Louca, P., May, A., Figueiredo, J.C., Hu, C., Molteni, E., Canas, L., Österdahl, M.F., Modat, M., Sudre, C.H., Fox, B., Hammers, A., Wolf, J., Capdevila, J., Chan, A.T., David, S.P., Steves, C.J., Ourselin, S. and Spector, T.D. (2022) 'Symptom prevalence, duration, and risk of hospital admission in individuals infected with SARS-CoV-2 during periods of omicron and delta variant dominance: a prospective observational study from the ZOE COVID Study', *Lancet (London, England)*, 399(10335), pp. 1618–1624.
134. Antonelli, M., Penfold, R.S., Merino, J., Sudre, C.H., Molteni, E., Berry, S., Canas, L.S., Graham, M.S., Klaser, K., Modat, M., Murray, B., Kerfoot, E., Chen, L., Deng, J., Österdahl, M.F., Cheetham, N.J., Drew, D.A., Nguyen, L.H., Pujol, J.C., Hu, C., Selvachandran, S., Polidori, L., May, A., Wolf, J., Chan, A.T., Hammers, A., Duncan, E.L., Spector, T.D., Ourselin, S. and Steves, C.J. (2022) 'Risk factors and disease profile of post-vaccination SARS-CoV-2 infection in UK users of the COVID Symptom Study app: a prospective, community-based, nested, case-control study', *The Lancet. Infectious diseases*, 22(1), pp. 43-55.
135. Du, P., Ding, N., Li, J., Zhang, F., Wang, Q., Chen, Z., Song, C., Han, K., Xie, W., Liu, J., Wang, L., Wei, L., Ma, S., Hua, M., Yu, F., Wang, L., Wang, W., An, K., Chen, J., Liu, H., Gao, G., Wang, S., Huang, Y., Wu, A.R., Wang, J., Liu, D., Zeng, H. and Chen, C. (2020) 'Genomic surveillance of COVID-19 cases in Beijing', *Nature Communications*, 11(1), pp. 1–9.
136. Williams, G.H., Llewelyn, A., Brandao, R., Chowdhary, K., Hardisty, K.M. and Loddo, M. (2021) 'SARS-CoV-2 testing and sequencing for international arrivals reveals significant cross border transmission of high risk variants into the United Kingdom', *EClinicalMedicine*, 38, 101021.
137. Curran, K.G. (2016) 'Cluster of Ebola Virus Disease Linked to a Single Funeral — Moyamba District, Sierra Leone, 2014', *MMWR. Morbidity and Mortality Weekly Report*, 65(8), pp. 202-205.
138. Böhmer, M.M., Buchholz, U., Corman, V.M., Hoch, M., Katz, K., Marosevic, D.V., Böhm, S., Woudenberg, T., Ackermann, N., Konrad, R., Eberle, U., Treis, B., Dangel, A., Bengs, K., Fingerle, V., Berger, A., Hörmansdorfer, S., Ippisch, S., Wicklein, B., Grahl, A., Pörtner, K., Muller, N., Zeitlmann, N., Boender, T.S., Cai, W., Reich, A., An der Heiden M, Rexroth, U., Hamouda, O., Schneider, J., Veith, T., Mühlemann, B., Wölfel, R., Antwerpen, M., Walter, M., Protzer, U., Liebl, B., Haas, W., Sing, A., Drosten, C. and Zapf, A. (2020) 'Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series', *The Lancet. Infectious diseases*, 20(8), pp. 920-928.
139. de Jesus, J.G., Sacchi, C., da Silva Candido, D., Claro, I.M., Sales, F.C.S., Manuli, E.R., da Silva, D.B.B., de Paiva, T.M., Pinho, M.A.B., de Oliveira Santos, K.C., Hill,

S.C., Aguiar, R.S., Romero, F., dos Santos, F.C.P., Gonçalves, C.R., do Carmo Timenetsky, M., Quick, J., Croda, J.H.R., de Oliveira, W., Rambaut, A., Pybus, O.G., Loman, N.J., Sabino, E.C. and Faria, N.R. (2020) 'Importation and early local transmission of COVID-19 in Brazil, 2020', *Revista do Instituto de Medicina Tropical de São Paulo*, 62, p. e30.

140. Giovanetti, M., Benvenuto, D., Angeletti, S. and Ciccozzi, M. (2020) 'The first two cases of 2019-nCoV in Italy: Where they come from?', *Journal of Medical Virology*, 92(5), p. 518.
141. Lemey, P., Hong, S.L., Hill, V., Baele, G., Poletto, C., Colizza, V., O'Toole, Á., McCrone, J.T., Andersen, K.G., Worobey, M., Nelson, M.I., Rambaut, A. and Suchard, M.A. (2020) 'Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2', *Nature Communications*, 11(1), pp. 1–14.
142. McCrone, J.T., Hill, V., Bajaj, S., Pena, R.E., Lambert, B.C., Inward, R., Bhatt, S., Volz, E., Ruis, C., Dellicour, S., Baele, G., Zarebski, A.E., Sadilek, A., Wu, N., Schneider, A., Ji, X., Raghwani, J., Jackson, B., Colquhoun, R., O'Toole, Á., Peacock, T.P., Twohig, K., Thelwall, S., Dabrera, G., Myers, R., Faria, N.R., Huber, C., Bogoch, I.I., Khan, K., du Plessis, L., Barrett, J.C., Aanensen, D.M., Barclay, W.S., Chand, M., Connor, T., Loman, N.J., Suchard, M.A., Pybus, O.G., Rambaut, A. and Kraemer, M.U.G. (2022) 'Context-specific emergence and growth of the SARS-CoV-2 Delta variant', *Nature*, 610(7930), pp. 154–160.
143. Dean, N. (2022) Tracking COVID-19 infections: time for change, *Nature*, 602(7896), p. 185.
144. Qasmieh, S.A., Robertson, M.M., Teasdale, C.A., Kulkarni, S.G., Jones, H.E., Larsen, D.A., Dennehy, J.J., McNairy, M., Borrell, L.N. and Nash, D. (2023) 'The prevalence of SARS-CoV-2 infection and other public health outcomes during the BA.2/BA.2.12.1 surge, New York City, April–May 2022', *Communications Medicine*, 3(1), pp. 1–12.
145. Qasmieh, S.A., Robertson, M.M., Teasdale, C.A., Kulkarni, S.G. and Nash, D. (2023) 'Estimating the Period Prevalence of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infection During the Omicron (BA.1) Surge in New York City (NYC), 1 January to 16 March 2022', *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 76(3), pp. 499-502.
146. Umeozuru, C.M., Usman, A.B., Olorukooba, A.A., Abdullahi, I.N., John, D.J., Lawal, L.A., Uwazie, C.C. and Balogun, M.S. (2022) 'Performance of COVID-19 case-based surveillance system in FCT, Nigeria, March 2020 –January 2021', *PLOS ONE*, 17(4), p. e0264839.
147. Fawole, O.I., Bello, S., Adebowale, A.S., Bamgboye, E.A., Salawu, M.M., Afolabi, R.F., Dairo, M.D., Namale, A., Kiwanuka, S., Monje, F., Namuhani, N., Kabwama, S., Kizito, S., Ndejjo, R., Seck, I., Diallo, I., Makhtar, M., Leye, M., Ndiaye, Y., Fall, M., Bassoum, O., Mapatano, M.A., Bosonkie, M., Egbende, L., Lazenby, S., Wang, W., Liu, A., Bartlein, R., Sambisa, W. and Wanyenze, R. (2023) 'COVID-19 surveillance in Democratic Republic of Congo, Nigeria, Senegal and Uganda: strengths, weaknesses and key Lessons', *BMC Public Health*, 23(1), pp. 1–15.
148. Brito, A.F., Semenova, E., Dudas, G., Hassler, G.W., Kalinich, C.C., Kraemer, M.U.G., Ho, J., Tegally, H., Githinji, G., Agoti, C.N., Matkin, L.E., Whittaker, C., Howden, B.P., Sintchenko, V., Zuckerman, N.S., Mor, O., Blankenship, H.M., de Oliveira, T., Lin, R.T.P., Siqueira, M.M., Resende, P.C., Vasconcelos, A.T.R., Spilki, F.R., Aguiar, R.S., Alexiev, I., Ivanov, I.N., Philipova, I., Carrington, C.V.F.,

Sahadeo, N.S.D., Branda, B., Gurry, C., Maurer-Stroh, S., Naidoo, D., von Eije, K.J., Perkins, M.D., van Kerkhove, M., Hill, S.C., Sabino, E.C., Pybus, O.G., Dye, C., Bhatt, S., Flaxman, S., Suchard, M.A., Grubaugh, N.D., Baele, G. and Faria, N.R. (2022) 'Global disparities in SARS-CoV-2 genomic surveillance', *Nature Communications*, 13(1), pp. 1–13.

149. Cori, A. and Kucharski, A. (2024) 'Inference of epidemic dynamics in the COVID-19 era and beyond', *Epidemics*, 48, 100784.
150. Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015) *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.
151. Anselin, L. (1988) *Spatial Econometrics: Methods and Models*. Springer Netherlands.
152. Besag, J., York, J. and Mollié, A. (1991) 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the Institute of Statistical Mathematics*, 43(1), pp. 1–20.
153. Alaimo Di Loro, P., Böhning, D. and Sahu, S.K. (2024) 'A Bayesian spatio-temporal Poisson auto-regressive model for the disease infection rate: application to COVID-19 cases in England', *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, p. 24.
154. Cortes-Ramirez, J., Wilches-Vega, J.D., Caicedo-Velasquez, B., Paris-Pineda, O.M. and Sly, P.D. (2024) 'Spatiotemporal hierarchical Bayesian analysis to identify factors associated with COVID-19 in suburban areas in Colombia', *Heliyon*, 10(9), e30182
155. Lowe, R., Lee, S.A., O'Reilly, K.M., Brady, O.J., Bastos, L., Carrasco-Escobar, G., de Castro Catão, R., Colón-González, F.J., Barcellos, C., Carvalho, M.S., Blangiardo, M., Rue, H. and Gasparrini, A. (2021) 'Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study', *The Lancet. Planetary health*, 5(4), pp. 209-219.
156. Ward, T., Morris, M., Gelman, A., Carpenter, B., Ferguson, W., Overton, C. and Fyles, M. (2023) 'Bayesian spatial modelling of localised SARS-CoV-2 transmission through mobility networks across England', *PLoS computational biology*, 19(11), e1011580.
157. Nicholson, G., Lehmann, B., Padellini, T., Pouwels, K.B., Jersakova, R., Lomax, J., King, R.E., Mallon, A.-M., Diggle, P.J., Richardson, S., Blangiardo, M. and Holmes, C. (2021) 'Improving local prevalence estimates of SARS-CoV-2 infections using a causal debiasing framework', *Nature Microbiology*, 7(1), pp. 97–107.
158. Bajaj, S., Chen, S., Creswell, R., Naidoo, R., Tsui, J.L., Kolade, O., Nicholson, G., Lehmann, B., Hay, J. A., Kraemer, M. U. G., Aguas, R., Donnelly, C. A., Fowler, T., Hopkins, S., Cantrell, L., Dahal, P., White, L.J., Stepniewska, K., Voysey, M., Lambert, B. and EY-Oxford Health Analytics Consortium (2024) 'COVID-19 testing and reporting behaviours in England across different sociodemographic groups: a population-based study using testing data and data from community prevalence surveillance surveys', *The Lancet. Digital health*, *6*(11), pp. e778–e790
159. Kermack, W.O. and McKendrick, A.G. (1927) 'A contribution to the mathematical theory of epidemics', *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 115(772), pp. 700-721.*
160. Balcan, D., Gonçalves, B., Hu, H., Ramasco, J.J., Colizza, V. and Vespignani, A. (2010) 'Modeling the spatial spread of infectious diseases: the GLobal Epidemic and Mobility computational model', *Journal of computational science*, 1(3), p. 132.
161. Davis, J.T., Chinazzi, M., Perra, N., Mu, K., Pastore y Piontti, A., Ajelli, M., Dean, N.E., Gioannini, C., Litvinova, M., Merler, S., Rossi, L., Sun, K., Xiong, X.,

Longini, I.M., Halloran, M.E., Viboud, C. and Vespignani, A. (2021) 'Cryptic transmission of SARS-CoV-2 and the first COVID-19 wave', *Nature*, 600(7887), pp. 127–132.
162. Aleta, A., Martín-Corral, D., Bakker, M.A., Pastore y Piontti, A., Ajelli, M., Litvinova, M., Chinazzi, M., Dean, N.E., Halloran, M.E., Longini, I.M., Pentland, A., Vespignani, A., Moreno, Y. and Moro, E. (2022) 'Quantifying the importance and location of SARS-CoV-2 transmission events in large metropolitan areas', *Proceedings of the National Academy of Sciences*, 119(26), p. e2112182119.
163. Kerr, C.C., Stuart, R.M., Mistry, D., Abeysuriya, R.G., Rosenfeld, K., Hart, G.R., Núñez, R.C., Cohen, J.A., Selvaraj, P., Hagedorn, B., George, L., Jastrzębski, M., Izzo, A.S., Fowler, G., Palmer, A., Delport, D., Scott, N., Kelly, S.L., Bennette, C.S., Wagner, B.G., Chang, S.T., Oron, A.P., Wenger, E.A., Panovska-Griffiths, J., Famulare, M. and Klein, D.J. (2021) 'Covasim: An agent-based model of COVID-19 dynamics and interventions', *PLoS Computational Biology*, 17(7), p. e1009149.
164. Hinch, R., Probert, W.J.M., Nurtay, A., Kendall, M., Wymant, C., Hall, M., Lythgoe, K., Cruz, A.B., Zhao, L., Stewart, A., Ferretti, L., Montero, D., Warren, J., Mather, N., Abueg, M., Wu, N., Legat, O., Bentley, K., Mead, T., Van-Vuuren, K., Feldner-Busztin, D., Ristori, T., Finkelstein, A., Bonsall, D.G., Abeler-Dörner, L. and Fraser, C. (2021) 'OpenABM-Covid19—An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing', *PLOS Computational Biology*, 17(7), p. e1009146.
165. Pybus, O.G., Rambaut, A. and Harvey, P.H. (2000) 'An integrated framework for the inference of viral population history from reconstructed genealogies', *Genetics*, 155(3), pp. 1429-1437.
166. Stadler, T. (2009) 'On incomplete sampling under birth–death models and connections to the sampling-based coalescent' (2009) *Journal of Theoretical Biology*, 261(1), pp. 58–66.
167. Stadler, T. (2010) 'Sampling-through-time in birth-death trees', *Journal of theoretical biology*, 267(3), pp. 396-404.
168. Pybus, O.G., Charleston, M.A., Gupta, S., Rambaut, A., Holmes, E.C. and Harvey, P.H. (2001) 'The Epidemic Behavior of the Hepatitis C Virus', *Science*, 292(5525), pp. 2323-2325.
169. Lemey, P., Rambaut, A., Welch, J.J. and Suchard, M.A. (2010) 'Phylogeography takes a relaxed random walk in continuous space and time', *Molecular biology and evolution*, 27(8), pp. 1877-1885.
170. Gill, M.S., Ls, T.H., Baele, G., Lemey, P. and Suchard, M.A. (2017) 'A Relaxed Directional Random Walk Model for Phylogenetic Trait Evolution', *Systematic biology*, 66(3), pp. 299-319.
171. Dellicour, S., Lequime, S., Vrancken, B., Gill, M.S., Bastide, P., Gangavarapu, K., Matteson, N.L., Tan, Y., du Plessis, L., Fisher, A.A., Nelson, M.I., Gilbert, M., Suchard, M.A., Andersen, K.G., Grubaugh, N.D., Pybus, O.G. and Lemey, P. (2020) 'Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework', *Nature Communications*, 11(1), pp. 1–11.
172. Lemey, P., Rambaut, A., Drummond, A.J. and Suchard, M.A. (2009) 'Bayesian Phylogeography Finds Its Roots', *PLOS Computational Biology*, 5(9), p. e1000520.
173. Vaughan, T.G., Kühnert, D., Popinga, A., Welch, D. and Drummond, A.J. (2014) 'Efficient Bayesian inference under the structured coalescent', *Bioinformatics (Oxford, England)*, 30(16), pp. 2272–2279.

174. De Maio, N., Wu, C.-H., O'Reilly, K.M. and Wilson, D. (2015) 'New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation', *PLOS Genetics*, 11(8), p. e1005421.
175. Müller, N.F., Rasmussen, D.A. and Stadler, T. (2017) 'The Structured Coalescent and Its Approximations', *Molecular biology and evolution*, 34(11), pp. 2970-2981.
176. Lemey, P., Ruktanonchai, N., Hong, S.L., Colizza, V., Poletto, C., Van den Broeck, F., Gill, M.S., Ji, X., Levasseur, A., Oude Munnink, B.B., Koopmans, M., Sadilek, A., Lai, S., Tatem, A.J., Baele, G., Suchard, M.A. and Dellicour, S. (2021) 'Untangling introductions and persistence in COVID-19 resurgence in Europe', *Nature*, 595(7869), pp. 713–717.
177. Shibemba, A.L., Obong, A.E., Fadeyi, A., Anas, A., Shabaan, A.E., Adel, A.M., Shoka, A.M.A., AbdelhamidW, Laraqui, A., Laatiris, A., Bellefquih, A.M., Faouzi, A., AbdelmoulahF, Essabbar, A., Bimouhen, A., Hilali, A., AbdoI, Padane, A., Sangaré, A.K., Soumah, A.K., Djimde, A., Toure, A., Kanteh, A., Bashein, A., Salama, A., Lojuan, A., Bayih, A.G., Negash, A.A., Lissom, A., AbidN, Konou, A.A., Shama, A., AbosedeO, AbouelnagaS, Ali, A., Anang, A.K., Tesfaye, A., Khan, A.K., Tayachew, A., Mihret, A., AlEmam, A.A., Hawi, A., AdesegunA, Desta, A.F., Ghassan, A., Traoré, A., Adjiratou, A.A., Sadji, A., Egli, A., Mendes, A., Abera, A., Candé, A., Alaoui, A., Pedro, A., AgbodziB, AgeezA, M., Ayouba, A., Alarif, A., Kayed, A.E., El-Taweel, A., Elsayed, A., Gad, A.F., Fakhfakh, A., Kandeil, A., AhmedM, Mostafa, A., AhmedO, S., Reggad, A., Taha, A., Vida, A., Bensalem, A., Sivro, A., Hachid, A., AjiliF, AjogbasileF, V., Adegnika, A.A., Siliadin, A., Natasha, A.P., Balde, A., Lemtudo, A., Sanaa-amine, A., AlaruusiA, M., Ouro-Medeli, A., Nyunja, A., Rizzo, A., Abdissa, A., Debela, A.T., Mancon, A., Marcello, A., Goredema, A., Greninger, A., Ndjolo, A., Niyomwungere, A., Mari, A., Mayor, A., AliM, A., Zumla, A., Ben Kahla, A., Grad, A., Kabanda, A., Tia, A., Camara, A., Zumla, A., Dieng, A.B., Maiga, A.I., Sall, A.A., Daou, A., Naguib, A., Souiri, A., Zouaki, A., Ghita, A., Mveang-Nzoghe, A., Koné, A., Koné, A.M.M., Ahouidi, A., Benyahia, A., Nagiub, A., AmerK, E., Dorkenoo, A., Barkat, A., Dia, A., Mbaye, A., Mboup, A., Thiam, A.S., Lahlou, A.I., Wadi, A.S., Ehlan, A., Marion, A., Aziza, A., Strydom, A., Abbad, A., Nkeshimana, A., Mulu, A., Maaroufi, A., Luquette, A.E., Shiningavamwe, A., Moreira-Soto, A., Azman, A., Bennett, A.J., Tarupiwa, A., Latifa, A., Badjo, A., Angelov, A., Brisebarre, A., Detweiler, A.M., Ndong, A., Werno, A., NDiaye, A.J.S., Sander, A.-L., AnnajrB, D'Aprile, A., van der Linden, A., van der Eijk, A., Erhart, A., Somboro, A.M., Dagnran, A., Ahumibe, A., Levasseur, A., van der Linden, A., Dara, A., Jegede-Williams, A., AouniM, ArjarquahA, Uwituze, A., Upton, A., Poda, A., Somé, A., Zongo, A., Massinga, A., AsareK, M., Katabazi, A.F., Ferjani, A., Dieng, A., Gaye-Gaye, A., AtigaN, Weldemariam, A.G., Arjaquah, A., AuldA, Ba-Diallo, A., ElMoussi, A., Sena, A., Mohamed, A., Farghaly, A., Ayo-AleB, AyoadeF, Adegnika, A.A., More, A., James, A.B., Nawfel, A., Diergaardt, A., Okwuraiwe, A.P., Taiwo, B.O., BabooS, B., BahadoorB, S., BahnassyA, A., Sanyang, B., BakryU, Touré, B.F., Iwalokun, B., BandaR, BaneS, Johnson, B., Cisse, B., Murwira, B., Munnink, B.O., Batchi-BouyouArmel, L., BatraR, Dhaala, B., Lamiae, B., BelkhirA, B., AyedI, B., Morton, B., Moussa, B., Wulf, B., Ndeboko, B., Rachid, B., Lindsey, B.B., Foulkes, B.H., Hounkpatin, B., Selekon, B., BensaidM, Mpairwe, B., Bagaya, B.S., Vanmechelen, B., Lell, B., Mutai, B., Adnew, B., Makamure, B., BighignoliB, Ndiaye, B.P., Adhikari, B.N., BitrouS, M., Akenji, B.M., Shobayo, B., Zuze, B., Nlavo, B.M., Belfquih, B., Boujemla, B., Yatassaye, B., Aymane, B., Andika, B., Yemi, B.K., Kleinhans, B., Galvao, B., Embaló, B.D.P., Manene, B., Pascal, B., Capel, C., Anu,

C., Tchibozo, C., Madeira, C., Cortes, C., Elisabeth, C., Kifude, C., Tayimetha, C.Y., Arnold, C.E., Okoi, C., Waruhiu, C., Godwe, C., Obiekea, C., Nkenfou, C., ChabukaL, B., Nejjari, C., ChambaroH, ChandaD, ChangulaK, Touzani, C.D., Kayuki, C., Nyagupe, C., Ssuuna, C., Hicham, C., Watters, C.M., ChawechH, Sokhna, C., Mohamed, C., Nwuba, C., ChilufyaC, Chukwu, C., Anyika, C., ChipimoP, J., ChitangaS, ChitenjeM, ChiwaulaM, J., Taha, C., Mansell, C., Butel, C., Devaux, C.A., Drosten, C., Ranaivoson, C., Utpatel, C., Malabat, C., ChtourouA, ChukwuG, Daubenberger, C., Kakai, C.G., Masakwe, C., Rafai, C.D., Chenwi, C., Misita, C.M., Muli, C., GeurtsvanKessel, C.H., Maufrais, C., Mbegnan, C., CrawfordM, Tato, C.M., Yiaba, C., D'AmoreN, Damilari, D., Ramadan, D., DamenaD, Ouso, D., Pisanelli, D., Licastro, D., Hammer, D., Nieuwenhuijse, D., Kateete, D.P., Wilkinson, D., Mouangala, D.L., Alemayehu, D.H., DawoodR, M., SanctisR, D., Mettle, D.N.A., Koita, D., Lodiongo, D.K., Laryea, D., Wandera, D.N., Leta, D., Noureddine, D., Takou, D., Tefera, D.A., Batra, D., Ndongo, D., DiabA, Diagne, D., Mabika, D.A., Bamourou, D., Samaté, D., DiarraB, Raoult, D., Mutangana, D., Nyepetsi, D.E.T., Ada, D.O.N., Salah, D., Cedola, D., Harding, D., Yeboah-Manu, D., Ange, D., Djomsi, D.M., Drinkovic, D., DraperC, Sahr, D.F., Odewale, E., Sedjro, E., Armand, E.-M., Kigozi, E., Koskei, E., Nkwembe, E., de Oliveira Filho, E.F., da Costa, E.M., Kiritu, E., Lim, E.S., Ozer, E.A., EgyirB, Abuelenein, E.A., Ngole, E.M., Aissam, E.A., Zahra, E.A.F., Adil, E.H., Abdelmjid, E.O., El-ShaqnqeryH, El-ZayatM, Abualas, E., Hicham, E., ElargoubiA, Kidane, E., Rizgalla, E., ElhosenyM, M., ElhosienyF, W., Carniel, E., Nyakarungu, E., ElkhateebS, M., Benaissa, E., El Fahime, E., ElouanassM, ElsissyM, H., Mbongo-Nkama, E., Elwaleed M. Elamin Sara A.I Latif, Edu, E., Orsini, E., Skarwan, E., Stefanov, E.K., Nasinghe, E., Nepolo, E., Ogunbayo, E., Munger, E., Permal, E., Gaies, E., EnnibiH, Khalid, E., Kotey, E., Smit, E., Lelo, E.A., Adu, E., Delaporte, E., Katagirya, E., Kezakarayagwa, E., Muthanje, E., Magalhães, E.L.M., Asiedu, E., Livo, E., Umumararungu, E., Chitechi, E., Omuseni, E., Alain, E.M., Quansah, E.B., Bonney, E.Y., EzzelarabM, H., Cassama, F., Niama, F.R., Arena, F., FaggioniG, FahsbenderE, Sigei, F., Abdellah, F., Jouali, F., FarawylaH, Hilali, F., Chgouri, F., El Falaki, F., El ansari, F.Z., Ebied, F., Bssaibis, F., Al Onifade, F., Ousmane, F., Khardine, F.A., Khardine, F., FedorovA, V., Alemu, F., Jamal, F., Sun, F., Doudou, F., FikryA, E., FilloS, Bracchitta, F., Kaboré, F., Fki-berrajah, L., Donati, F., Fenollar, F., Sorie, F.S.S., FolarinoO, Folorunsho, F., Kabatesi, F., Muyembe-Mawete, F., Malagon, F., Kiemdé, F., Asiedu-Bekoe, F.E., Nkongho, F., Lemoine, F., Tei-Maya, FreitasR, H., Fuh-NebaT, GaaloulI, Kabamba, G., Gutema, G., Emna, G., Gadzama, G., Mahmoud, G., Ouangole, G., GargouriS, GarryR, McAuliffe, G., Kimita, G., Manouana, G.P., Tesfaye, G., Awinda, G., Kyei, G.B., Michuki, G., Ondo, G.N., van Rooyen, G., Marais, G., Abichu, G., Beyene, G.T., Tollera, G., Hailu, G., Hamza, G., Kayali, G., El Amin, G., Mhlanga, G., Kibet, G., Hounkanrin, G., GiordaniF, Teka, G., Masete, G., GoldstoneR, GomaaC, Mokhtar, G., Lo, G., GosnellB, I., GottbergA, Belournou, G.A., Oni, G., Vincent, G., Wani, G., Barreh, G.A., Garcia, G., Deng, G., GusevaN, P., Mbembo, G.P.M., Padzys, G.S., GwayiS, Ben Romdhane, H., Naija, H., Diallo, H., HadadA, HafezM, M., Ladhari, H., Elshora, H., Dadi, H., Mohammed, H., Kabbaj, H., Hafez, H., HalafawyA, Tinto, H., Ndiaye, H.D., Assane, H., HamdaniT, N., HamdyM, S., Namagembe, H., HammadM, HammamiA, Gharmaz, H., Andersson, H.S., Dakka, H., El Jebari, H., Soliman, H.K., Houda, H., HarveyR, Benkirane, H., Aguenaou, H., Ihazmad, H., HassanR, HassanW, Ahmed, H., Sogodogo, H., Seth-Smith, H., Razafimanjato, H., Koka, H., Mouhssine, H., Mwebesa, H., Perez, H., Tchoudjin, H.C.P., Elannaz, H., Oumzil, H., Frempong,

H.O., Fidelia, H., Elarab, H.E., Xie, H., Benrahma, H., Arrouchi, H., Guedouar, H., Luo, H., Bassène, H., Si, H., Goodfellow, I., Guindo, I., Halilou, I., Abdillahi, I.S., Lahlou, I.-A., Diawara, I., Egoh, I., IgeF, A., IknaneA, A., Bnouyahia, I., Osman, I.O., Boubaker, I.B.-B., Hassan, I.A., Foda, I., Smyej, I., Kacem, I., Mdini, I., Mkada, I., Mandomando, I., Comas, I., Mdini, I., InglêsL, Mudau, I., Joel, I.Y., Chestakova, I., Cancino, I., Phiri, I., Boussoukou, I.P.M., IsmailA, Osei-Wusu, I., Maman, I., Barilar, I., Asante, I.A., Gerard, I., Heikel, J., Souopgui, J., Marx, J., JalalD, Nourlil, J., El Atar, J., Ben Khelil, J., Rahoui, J., Fekkak, J., Mutisya, J., Ussher, J., Exler, J.F., MaCauley, J., Majanja, J., JannooN, Manneh, J., Scharnberg, J., Schlotterbeck, J., Chimedza, J., Bouzid, J., Niyibigira, J.B., Djontu, J.C., Maritz, J., Cigolo, J.-C.M., Monemou, J.-L., Umuringa, J.D., Delerce, J., Ndaruhutse, J., Nkurunziza, J., Asamoah, J.A., Sherwood, J., Wang, J., Marti-Carerras, J., Mutungi, J.K., Mutungi, J., Siawaya, J.F.D., Koivogui, J., de Ligt, J., Njuguna, J., Rumunu, J., Tembo, J., Waitumbi, J., Adeniji, J.A., Rigby, J., Hellemans, J., Fokam, J., DeRisi, J.L., Makhema, J., Mugisha, J., Shaibu, J.O., Oliver-Commey, J., Bwogi, J., Freeman, J., Isong, J.A., Nyataya, J., Ayoola, J., Appiah-Kubi, J., Hultquist, J.F., Gedeon, J., Sokei, J., Howard, J., Schneider, J., Campbell, J., Elvy, J., JumaaA, B., Lee, J., Lessler, J., O'Grady, J., Kourouma, K., Ghedira, K., KalantarK, Mahlakwane, K., KamounS, Mulonga, K., KapataP, C., KapayaF, Kapin'aM, HakimH, K., KasambaraW, Yassine, K., Tesfaye, K., Subramoney, K., KatyshevA, D., KayeyiN, Barnes, K., Delaney, K., KazorinaE, V., Seru, K., KeitaS, Durkin, K., Jerome, K.R., René, K.K., Swanson, K., Maeka, K.K., Attiku, K.O., Sanders, K., Ella, K.Z., Tuki, K., Elhag, K.A., Gueye, K., Amer, K., Mostafa, K.E., KharatN, KhumaloZ, Stoecker, K., KimL, Bishop-Lilly, K.A., KimothoJ, Lahlou, K.D., Bonney, K., Tegueni, K., Wahab, K.W., KolomoetsE, V., Jain, K., Asmamaw, K., Kossi, K., Jambo, K., Yao, K.T., Dürr, K., Ouffoué, K., KrasnovY, M., Kandaswamy, K.K., Andersen, K., KritskyA, A., David, K., Macheke, K., KutyrevV, V., KwendaS, Maluzi, K., Lee, K., Long, K.A., Grantz, K., Simons, L.M., Serrano, L., Lagos State Government, Elsawy, L., Sbabou, L., Lamboni, L., Holland, L.A., Shrestha, L., Sangaré, L., Anga, L., Jelly, L., Hoornaert, L., Linh, L.T.K., Kooepile, L., MC Intyre, L.-A., Mutesa, L., Okoli, L., Ouedraogo, L., LesegoKuate-lere, LetaD, LetaiefA, LiboroG, Kanjau, L., LinL, Boatemaa, L., Houhamdi, L., Ian, L.W., ListaF, LiweweM, M., Mulenga, L., Voegtly, L.J., Shipingana, L., Micelli, L., Kwasah, L., Allam, L., Lefrançois, L., Salma, L., Amenga-Etego, L.N., Brechar, L., Mewono, L., Estrella, L.A., Mhuulu, L., Newton, L., Ulrich, M., Mogotsi, M.T., Abderrahmane, M., MadW, MadiW, Y., Stange, M., MagdeldinS, Kharrat, M., MahlanguB, Shehata, M., MahrousN, MaidaA, Camara, M., MakoriT, MalamaK, Bestehorn-Willmann, M.S., MalooI, Lo, M.B.C., Keita, M.B., Bah, M.S., Sow, M.S., Djingarey, M.H., Maiga, M., Elsaid, M.H., Zahran, M.H., Diakite, M., Ben Sassi, M., Pambou, M., ManickchundN, Puente, M.T., ManrajS, S., MansourT, Mukhtar, M.M., Tongo, M., Nabila, M., Mills, M., Artesi, M., Patrizio, M.P., Gismondo, M.R., Lipsi, M.R., Sulaiman, M.K., Kujabi, M., Amougou, M., Okomo, M.C., Chabert-Consen, M.M., Vernet, M.-A., Hayette, M.-P., Gdoura, M., Reynders, M., Barbet, M., Koopmans, M., Boter, M., Siedner, M., Abebe, M., Antwerpen, M.H., Melloul, M., Foudi, M.M., Peeters, M., Hsiao, M., DeAlmeida, M., Lalemi, M., Davies, M.-A., MasahiroK, Sambou, M., MathaboM, Parker, M.D., Walter, M.C., MathurH, Blakiston, M., Storey, M., Bates, M., Rogers, M., Pauthner, M., Vanpeene, M., Margaglione, M., Bloomfield, M., Abdelfattah, M., Soliman, M.S., Fall, M., MdlaloseK, Huang, M.-L., MehtaS, Albert, M., Joël Aïssi, M.A., Bizard, M., Mouad, M., Laamarti, M., Douffan, M., MhallaS, Addidle, M., Marks, M., Nagel, M., Deschenes, M.V., Davids, M., Balm, M., Lin, M., Tan, M.,

MihreteA, Buhari, M.O., Cá, M.A., Adusei-Poku, M., Mwangi, M., Kamel, M., MirandaJ, Prince-David, M., Eshun, M., Wayengera, M., Mishra, M., Eloualid, M., Elalaoui, M.A., MnguniA, MnyameniF, Matshaba, M., MohaleT, Elgohary, M.A.-S., Ali, M.A., Ben Moussa, M., Chenaoui, M., El Sayes, M., Elhadidi, M., Seadawy, M.G., Abdoelraheem, M.H., Hassany, M., Aboubaker, M.H., MohamedK, S., Kamal, M., Rhajaoui, M., Seadawy, M., Shamel, M., Shemis, M., MohammedK, S., Elfihri, M.W.C., Mechita, M.B., Gomaa, M., Bitew, M., MomohM, Alkarim, M.O.A., Pacome, M., Akinola, M., MonteA, MonuirG, MonzeM, MookoM, MoralesA, N., Andres, M.-S., Povogui, M., Mosepele, M., Chilufya, M., Joloba, M., Luutu, M., Elouennass, M., Elhoseiny, M., Elnakib, M., Yakout, M., Gardoul, M., Hemlali, M., MatoumbaA, M., Ben Sassi, M., Safer, M., Essabbar, M., Mbow, M., Maloum, M.N., Sakho, M., Odugbemi, M., MtshaliP, MubembaB, Mugabe, M., MufindaM, Faisal, M., Fowora, M.A., Enatha, M., MuleyaW, Ibrahim, M.E., MupetaF, Clarisse, M., Ali, M., Allam, M., Mouallif, M., MuuoS, N., MwangombaW, Seffar, M., MzumaraT, E., N'dilimabakaN, Soara, N., NabliA, El Mrimar, N., Rodrigues, N., Sitoe, N., Leye, N., NaguibA, NalubambaK, S., El Guindy, N.M., Siegfried, N., Hihi, N., Amar, N., NaryshkinaE, A., Endjala, N., Garus-Oas, N., Kapata, N., Ndiaye, N., Frans, N., NdamN, T., Kane, N.C.T., Ndack, N., Fainguem, N., Nnaemeka, N., Pentikainen, N., NdwigaL, Mabunda, N., NeffN, NegashA, A., Stambouli, N., NetoZ, Ngonga, D.M., NgosaW, Otuonye, N.M., Niatou-SingaF, S., Ndam, N.T., Feasey, N., Mwikwabe, N., NicodJ, Vidal, N., Freed, N., Mishra, N., Ben Alaya, N., Savaliya, N., Baker, N., Mbondoukwe, N.P., Mdlalose, N., Nduka, N., Abo Shama, N.M., Jalal, N., Gideon, N., NtuliN, Abilio, N., Yusuf, N.I., Wayoro, O., OchwotoM, Ebong, O., Ofori-BoaduL, Claire, O.A.M., Elroby, O., Olabisi, O., Ojo, O., Popoola, OlfertL, Silander, O., Obafemi, O., Amoo, O.S., Oluwasemowo, O., Akanbi, O.A., Laguda-Akingba, O., Adewumi, O., Askander, O., Elahmer, O., OmarS, OmilabuS, Kutkat, O., Adesuyi, O., Carey, O.F., Lesetedi, O., OngeraE, Bareng, O.T., OnwuamahC, K., Ope-EweO, Mansour, O., Kanjerwa, O., OtengF, Touzani, O., Mohammed, O.A.S., Diop, O., Ouadghiri, O., Gueye, O., Owusu-NyantakyiC, OyefoluA, Ayansola, O., Nthiga, P., PalombaS, PanjaL, Diaw, P.A., ParkD, Manga, P., PatelH, Motshosi, P., PatooM, Amoth, P., Descheemaeker, P., Mavingui, P., Tuyisenge, P., PattooM, Dobi, P., Liberator, P., Essone, P.N., da Costa Jarra Manneh, P.J., Sene, P.Y., Paixão, P.A.C.R., Roychoudhury, P., Uche, P.O., Zhou, P., Diallo, P.M., PereiraA, Akogbeto, P., Bauer, P., Skidmore, P.T., Raimond, P., Mmatshepho, P.-M., Ashton, P., PhilipC, El-Duah, P., Soglo, P.M., Phiri, P.W., Wagner, P., Colson, P., Dussart, P., Meyer, P.L., Ashton, P., Tushabe, P., Fournier, P.-E., Maes, P., Yu, P.A., Manangazira, P., Adewumi, P., Yang, Q., Mohktar, Q., QuansahE, B., QuashieP, Quedraogo, R., Maqsood, R., Githii, R., Abi, R., Benhida, R., El Jaoudi, R., Mentag, R.A., Ahmed, R., Algheriani, R., Simbi, R., Abubeker, R.C., Nari, R., Lorenzo-Redondo, R., Galal, RaoufA, Saizonou, R., Lumembe, R., Mubichi, R., Cer, R.Z., Dawood, R., Kassab, R., Liyai, R., RehnA, Pereira, R.J., Sikkema, R., Abderrazak, R., Armanious, R.M., Daghfous, R., Gouider, R., Adegbola, R., Lako, R.L.L., Molenkamp, R., Webby, R., Yeboah, R., RickS, Tagajdid, R., Nfor, R.Z., Yandoko, R.N., Newton, R., Rutayisire, R., Daniels, R.S., Bikangui, R., Kohoun, R.K., Kamga, R., Nguema, R.M., Shapiro, R., RogersJ, Kamulegeya, R., Laamrti, R., Lontchi, R.A.L., Kiiza, R., Galiwango, R.M., De Nittis, R., RoshdyW, H., Houechenou, R.M.A., Ben Othman, R., Inndi, R., SaadM, A., Amzazi, S., Nsanzimana, S., Diallo, S., SadjiY, A., El Mazouri, S., Elkochri, S., Ghoulame, S., Karidioula, S., Sankhe, S., Gevao, S., SahrP, F., SaibuJ, O., Ouedraogo, S., SaiidS, Ndure, S.L., Keita, S., SalahD, Hussein, S.E., SalahH, SalehA, A., SalehM, Sourakatou, S., Roberts, S., Abid, S., Sayed, S., SalouM, SaluO,

B., Lissauer, S., Bingono, S.O.O., Ndiour, S., Fageer, S.M., Girgis, S.A., SamirM, SamirO, Benkeroum, S., Ibrahim, S.F., Fageer, S.M., Assoumou, S.Z., Saiid, S., Khamadi, S.A., Limbaso, S.K., Armoo, S., Kirimunda, S., Sorie, S., Symekher, S.L., Owaka, S., Ferjani, S., Alaoui-Amine, S., Anna-Lena, S., Safietou, S., Jiménez-Serrano, S., Kamara, S., Chammam, S., Mahmoud, S., Jefferies, S., Rubin, S., Stanley, S., Sathees, S., Abdella, S., Chamman, S., Mwesigwa, S., SawaH, SeadawyM, G., Ellis, S., Bontems, S., Ramaologa, S., Zinyowera, S., Kumordjie, S., Ngomtcho, S.C.H., Awunyo, S., Sadeuh-Mba, S.A., Niane, S.S., Okeyo, S., Altamura, S., Agbenyo, S.B., Bakhash, S.M., Lockman, S., Bedri, S.A., Soliman, S., ShalabyL, Wilson, S., Muttaiyah, S., Abimbola, S., Hsu, S., ShcherbakovaS, A., Osiany, S., Shawky, S., Helmy, S., ShevtsovaA, P., Moustafa, S., Wohl, S., Johane, S., Kagnissode, S.A.M., Opanda, S.M., Mayaphi, S., Bióté, S.T., Monego, S.D., Ruhweza, S.P., Eckstein, S., Reuben, S., SinyangeN, Rachakonda, S., Andriam, S.F., Viegas, S., Sogodogo, S., Ndongo, S., Ajibaye, S., Langat, S., Fwoloshi, S.S.G., Charlene, S.P., Henson, S., Haddad, S.V., Bedié, SosedovaE, A., Kartti, S., Amira, S., Mboup, S., Maïté, S., Dao, S., Salifou, S., Pallerla, S.R., Niemann, S., Borrmann, S., Asiimwe, S., Eggan, S.M., Ochola, S., Wanok, S., Reynolds, S.J., Olorunnimbe, S., Bunga, S., SujeewonC, Babatunde, S., Engelbrecht, S., Morpeth, S., Taylor, S., Handrick, S., Gatare, S., Melingui, S., SymekerS, L., Devatchagni, S., TahaA, G., Chouati, T., Maatoug, T., Bajjou, T., TakadaA, TaloaK, A., Seyoum, T., TanM, Stander, T., Niemann, T., Aanniz, T., Zaher, T.R.E., Takawira, T., TatoC, TeferaD, A., Fred, T.-M., Gelanew, T., Rufael, T., Le Viet, T., Tefelo, T., Kagoné, T., Gnimadi, T.A.C., ThiongoK, Briese, T., ThomasVan, L., Mphoyakgosi, T., de Silva, T.I., Roloff, T., Blackmore, T., Rollo, T., Lutalo, T., Ibrahim, T.A.M., TombolomakoT, B., Adepetun, T., Tomkins-TinchC, Wawina-Bokalanga, T., Sobajo, T., Nadia, T., Essayagh, T., Nwako, T., Traoré, TrikiH, TsiryR, Randriambolamanantsoa, T., Madisa, T., TurkiM, UgwuC, A., Emokpae, U., Carri, V.D., Micheli, V., RooyenG, V., VanaerschotM, Magnussen, V., Mohr, V., Sathyendran, V., Playle, V., VianaR, VickosU, Corman, V.M., Mukonka, V., Ofula, V., Ahyong, V., Appiah, V., Bours, V., M'cormack, V.V., Hope, V., Lieu, V.H.-T., Raharinosy, V., Meshack, W., WadondaN, WadulaJ, Ali, W., Sule, W., Bulimo, W.D., Yiaba, W.C., WaruhiuC, N., Fares, W., WebbyR, WeldemariamA, G., Jo, W.K., WoelfelR, WolhS, WurieI, Crespin, X., Ren, X., Wang, X., Krasnov, Y.M., Dia, Y.A., Ndiaye, Y.D.S., Yusuf, Y.M., Sawadogo, Y., Yadouleton, Y., Maidane, Y.G., Tsegay, Y., Layibo, Y., El Hady, Y., Sekhsokh, Y., Moatasim, Y.A., Sadji, YeboahC, Mohammed, Y., Makki, Y.R., Akhoud, Y., Ushijima, Y., Jimoh, Y., Badou, Y., ZablonJ., M.D., ZablonJ, O., Hamzaoui, Z., Regragui, Z., Wuduri, Z., Souma, Z., Waberi, Z.A., Imam, Z.S., Tarnagda, Z., Faouzia, Z., ZhangJ, Li, Z., Maiga, Z., Kasmy, Z., ZongMinko, O., ZorganiA, Yassine, Z., Issa, Z., ZuluP and Xiang, Z. (2022) 'The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding genomic surveillance', *Science*, 378(6615), p. eabq5358.

178. Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C.A., Smith, D.J., Pybus, O.G., Brockmann, D. and Suchard, M.A. (2014) 'Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2', *PLOS Pathogens*, 10(2), p. e1003932.

179. Vasylyeva, T.I., Havens, J.L., Wang, J.C., Luoma, E., Hassler, G.W., Amin, H., Di Lonardo, S., Taki, F., Omoregie, E., Hughes, S. and Wertheim, J.O. (2024) '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*, 20(6), p. e1012288.

180. Zachreson, C., Chang, S., Harding, N. and Prokopenko, M. (2022) 'The effects of local homogeneity assumptions in metapopulation models of infectious disease', *Royal Society open science*, 9(7), p. 211919.
181. Ball, F., Britton, T., House, T., Isham, V., Mollison, D., Pellis, L. and Scalia Tomba, G. (2015) 'Seven challenges for metapopulation models of epidemics, including households models' (2015) *Epidemics*, 10, pp. 63–67.
182. Layan, M., Müller, N.F., Dellicour, S., De Maio, N., Bourhy, H., Cauchemez, S. and Baele, G. (2023) 'Impact and mitigation of sampling bias to determine viral spread: Evaluating discrete phylogeography through CTMC modeling and structured coalescent model approximations', *Virus evolution*, 9(1), p. vead010.
183. Bedford, T., Riley, S., Barr, I.G., Broor, S., Chadha, M., Cox, N.J., Daniels, R.S., Palani Gunasekaran, C., Hurt, A.C., Kelso, A., Klimov, A., Lewis, N.S., Li, X., McCauley, J.W., Odagiri, T., Potdar, V., Rambaut, A., Shu, Y., Skepner, E., Smith, D.J., Suchard, M.A., Tashiro, M., Wang, D., Xu, X., Lemey, P. and Russell, C.A. (2015) 'Global circulation patterns of seasonal influenza viruses vary with antigenic drift', *Nature*, 523(7559), p. 217.
184. Kalkauskas, A., Perron, U., Sun, Y., Goldman, N., Baele, G., Guindon, S. and De Maio, N. (2021) 'Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk', *PLOS Computational Biology*, 17(1), p. e1008561.
185. Müller, N.F., Rasmussen, D. and Stadler, T. (2018) 'MASCOT: parameter and state inference under the marginal structured coalescent approximation', *Bioinformatics (Oxford, England)*, 34(22), pp. 3843–3848.
186. Bajardi, P., Barrat, A., Savini, L. and Colizza, V. (2012) 'Optimizing surveillance for livestock disease spreading through animal movements', *Journal of the Royal Society, Interface*, 9(76), pp. 2814–2825.
187. Pei, S., Teng, X., Lewis, P. and Shaman, J. (2021) 'Optimizing respiratory virus surveillance networks using uncertainty propagation', *Nature Communications*, 12(1), pp. 1–10.
188. Polgreen, P.M., Chen, Z., Segre, A.M., Harris, M.L., Pentella, M.A. and Rushton, G. (2009) 'Optimizing influenza sentinel surveillance at the state level', *American journal of epidemiology*, 170(10), pp. 1300-1306.
189. Zhang, D., Ge, Y., Wang, J., Liu, H., Zhang, W.B., Wu, X., B M Heuvelink, G., Wu, C., Yang, J., Ruktanonchai, N.W., Qader, S.H., Ruktanonchai, C.W., Cleary, E., Yao, Y., Liu, J., Nnanatu, C.C., Wesolowski, A., Cummings, D.A.T., Tatem, A. J., and Lai, S. (2024) 'Optimizing the detection of emerging infections using mobility-based spatial sampling', *International Journal of Applied Earth Observation and Geoinformation*, 131, p. 103949.
190. Spott, R., Pletz, M.W., Fleischmann-Struzek, C., Kimmig, A., Hadlich, C., Hauert, M., Lohde, M., Jundzill, M., Marquet, M., Dickmann, P., Schüchner, R., Hölzer, M., Kühnert, D. and Brandt, C. (2024) 'Exploring the Spatial Distribution of Persistent SARS-CoV-2 Mutations - Leveraging mobility data for targeted sampling', *eLife*, 13, RP94045.
191. Tan, Q., Zhang, C., Xia, J., Wang, R., Zhou, L., Du, Z., and Shi, B. (2025) 'Information-guided adaptive learning approach for active surveillance of infectious diseases', *Infectious Disease Modelling*, 10(1), pp. 257–267.
192. Bastani, H., Drakopoulos, K., Gupta, V., Vlachogiannis, I., Hadjichristodoulou, C., Lagiou, P., Magiorkinis, G., Paraskevis, D. and Tsiodras, S. (2021) 'Efficient and targeted COVID-19 border testing via reinforcement learning', *Nature*, 599(7883), pp. 108–113.

193. Tsui, J.L.-H., Zhang, M., Sambaturu, P., Busch-Moreno, S., Suchard, M.A., Pybus, O.G., Flaxman, S., Semenova, E. and Kraemer, M.U.G. (2024) 'Toward optimal disease surveillance with graph-based active learning', *Proceedings of the National Academy of Sciences of the United States of America*, 121(52), p. e2412424121.
194. Borges, D.G.F., Coutinho, E.R., Cerqueira-Silva, T., Grave, M., Vasconcelos, A.O., Landau, L., Coutinho, A.L.G.A., Ramos, P.I.P., Barral-Netto, M., Pinho, S.T.R., Barreto, M.E. and Andrade, R.F.S. (2025) 'Combining machine learning and dynamic system techniques to early detection of respiratory outbreaks in routinely collected primary healthcare records', *BMC Medical Research Methodology*, 25(1), pp. 1–20.

2

Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1

I began my DPhil in October 2021 with an initial focus on methodological development to incorporate spatial structures and mobility data in phylodynamic inference. However, just weeks into my first term, the emergence of SARS-CoV-2 B.1.1.529 - later designated a Variant of Concern (VOC) named Omicron - led to a shift in my research direction. Following my involvement in an international collaborative effort to characterise the variant's emergence and early spread in southern Africa (leading to the first publication describing the variant, Viana et al., 2022), my work continued in this direction, focusing specifically on the introduction and subsequent local dissemination of Omicron BA.1 in the UK. Building on earlier efforts to reconstruct SARS-CoV-2 spread through the joint analysis of epidemiological, genomic, and human mobility data, the study presented in this chapter represents one of the largest of its kind utilising over 115,000 viral genomes, revealing how human geography and mobility shaped the variant's spread across multiple spatial scales. Findings from this work also highlighted important limitations in existing disease surveillance systems and sampling design for phylogeographic inference, helping define the research questions pursued in subsequent chapters.

A manuscript describing this work was first made available on MedRxiv as a preprint on 4th January 2023, and later published in *Science* on 20th July 2023, under the title "*Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1*". It is presented here in full, with minor modifications to ensure consistency of formatting and style within this thesis.

Tsui, J.L.-H., McCrone, J.T., Lambert, B., Bajaj, S., Inward, R.P.D., Bosetti, P., Pena, R.E., Tegally, H., Hill, V., Zarebski, A.E., Peacock, T.P., Liu, L., Wu, N., Davis, M., Bogoch, I.I., Khan, K., Kall, M., Abdul Aziz, N.I.B., Colquhoun, R., O'Toole, Á., Jackson, B., Dasgupta, A., Wilkinson, E., de Oliveira, T., COVID-19 Genomics UK (COG-UK) consortium¶, Connor, T.R., Loman, N.J., Colizza, V., Fraser, C., Volz, E., Ji, X., Gutierrez, B., Chand, M., Dellicour, S., Cauchemez, S., Raghwani, J., Suchard, M.A., Lemey, P., Rambaut, A., Pybus, O.G. and Kraemer, M.U.G. (2023) 'Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1', *Science*, 381(6655), pp. 336–343.

2.1 Abstract

SARS-CoV-2 variants of concern (VOCs) now arise in the context of heterogeneous human connectivity and population immunity. Through a large-scale phylodynamic analysis of 115,622 Omicron BA.1 genomes, we identified >6,000 introductions of the antigenically-distinct VOC into England and analysed their local transmission and dispersal history. We find that six of the eight largest English Omicron lineages were already transmitting when Omicron was first reported in southern Africa (22 November 2021). Multiple data sets show importation of Omicron continued despite subsequent restrictions on travel from southern Africa, due to export from well-connected secondary locations. Initiation and dispersal of Omicron transmission lineages in England was a two-stage process that can be explained by models of the country's human geography and hierarchical travel network. Our results enable a comparison of the processes that drive the invasion of Omicron and other VOCs across multiple spatial scales.

2.2 Introduction

Since the emergence of SARS-CoV-2 in late 2019, multiple variants of concern (VOCs) have sequentially dominated the pandemic worldwide. The Omicron VOC (Pango lineage B.1.1.529, later divided into lineages including BA.1 and BA.2) was discovered in late

November 2021 through genomic surveillance in Botswana and South Africa and a traveller from South Africa in Hong Kong (1) and designated a VOC by the World Health Organisation on 26 November (2). An initial surge in Omicron cases in South Africa indicated a higher transmission rate than previous VOCs (3), which studies later attributed to a shorter serial interval, increased immune evasion, and greater intrinsic transmissibility (4-7). The mechanism for greater transmissibility is hypothesised to be altered tropism and higher replication in the upper respiratory tract (8, 9). Together with waning levels of population immunity from previous infections and vaccination (10), local transmission of Omicron BA.1 was reported soon thereafter in travel hubs worldwide, including New York City and London by early December 2021, despite travel restrictions on international flights from multiple southern African countries (11, 12).

Following the first confirmed case of Omicron in England on 27 November 2021 (13), Omicron prevalence increased rapidly across all regions of England, with Greater London prevalence peaking first in mid-December at ~6%, followed by the South East region (14). Other metropolitan areas in North West and North East England saw similar but delayed increases in prevalence with observed peaks between early- and mid-January 2022. By January 2022, Omicron incidence had declined substantially in Greater London and other southern regions, resulting in decreasing prevalence from north to south England (15). Rapid growth in infections during the initial emergence of Omicron in England prompted the UK government to impose interventions including a move to "Plan B" non-pharmaceutical restrictions (mandatory COVID pass for entry into certain venues, face coverings, and work-from-home guidance) on 8 December 2021 (16) and an accelerated program of booster vaccination for all adults by mid-December 2021 (17). SARS-CoV-2 prevalence in England decreased later in January 2022, coincident with a

falling proportion of BA.1 infections as lineage BA.2 became the dominant lineage; BA.2 was itself later replaced by lineages BA.4 and BA.5 (18-20).

Understanding and quantifying the relative contributions of the factors that determined the arrival and spatial dissemination of Omicron BA.1 in England can help inform the design of spatially-targeted interventions against VOCs (21). Here, we analyse the Omicron BA.1 wave in England, using a dataset of 48,748 Omicron BA.1 genomes from England. This dataset represents ~1% of all confirmed Omicron BA.1 cases in England during the study period and is combined with aggregated and anonymized human mobility and epidemiological data from Lower Tier Local Authorities (LTLAs) in England.

2.3 International importation and Omicron BA.1 lineage dynamics

To investigate the timing of virus importations into England and the dynamics of the resulting local transmission lineages, we undertook a large-scale phylodynamic analysis of 115,622 SARS-CoV-2 Omicron genomes, sampled globally between 8 November 2021 and 31 January 2022. About 42% (N=48,748) were sampled from England and sequenced by the COVID-19 Genomics UK (COG-UK) consortium (22). All available genomes (from COG-UK and GISAID (23) on 12 and 9 April 2022 respectively) sampled before 28 November 2021 were included; later genomes were subsampled randomly in proportion to weekly Omicron case incidence while maintaining a ~1:1 ratio between English and non-English samples. To reduce potential bias caused by heterogeneous sequencing coverage, we performed a weighted subsampling of the English genomes using a previously developed procedure that accounts for variation in the number of sequences sampled per reported case at the Upper Tier Local Authority (UTLA) level (24).

We identified at least 6,455 [95% HPD: 6,184 to 6,722] independent importation events. Most imports from outside of England (69.9% [95% HPD: 69.0 to 70.7]) led to singletons (i.e., a single genome sampled in England associated with an importation event, which did not lead to observable local transmission in our dataset). The earliest importation is estimated between 5 and 18 November (approximated as the midpoint between the inferred times of the most recent common ancestor (MRCA) of the transmission lineage and the parent of the MRCA (PMRCA)). Between the first introduction and mid-December 2021, we reconstruct an approximately exponential increase in the daily number of imports, before a plateau in early January 2022 (Fig. 2.1C). Daily importation rate may have risen between 22 November (when Omicron was first reported) and 25 November (when travel restrictions started; Fig. 2.1C). Increased outflows of air passengers before (and possibly in anticipation of) the imposition of travel restrictions have been reported for SARS-CoV-2 elsewhere (25, 26). The importation rate appears to re-accelerate early in December, despite restrictions on incoming international travel from 11 southern African countries; imports then could have originated from BA.1 outbreaks in other countries in late November and early December 2021.

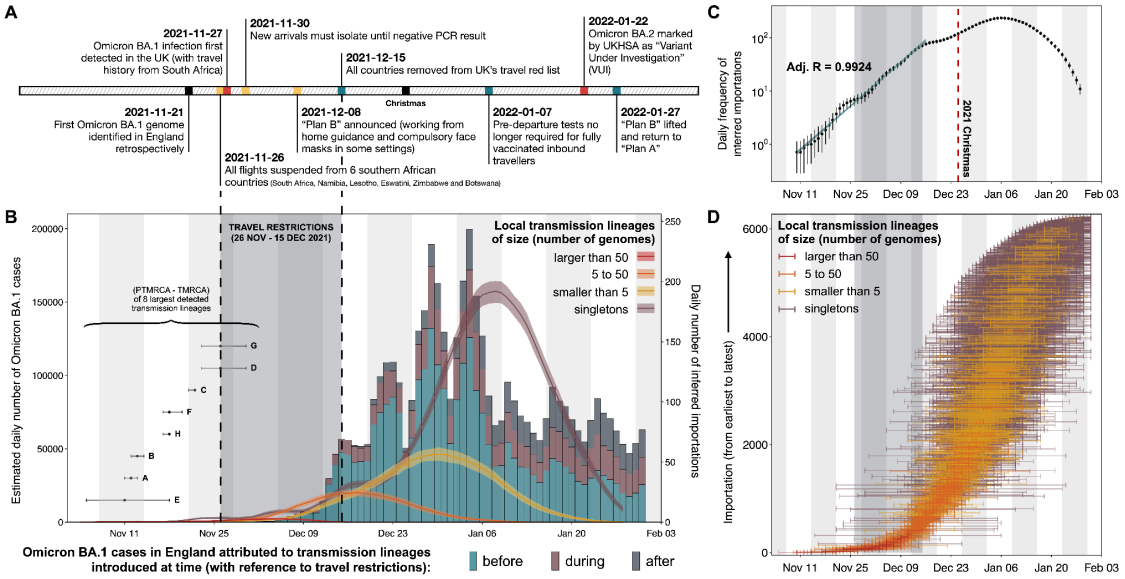


Fig. 2.1. Dynamics of BA.1 transmission lineages in England. (A) Timeline of events during the BA.1 wave in England until February 2022. (B) Histogram of estimated daily number of BA.1 cases, coloured according to the proportion of cases attributable to transmission lineages imported at different times (shaded region shows period of travel restrictions). Curves show the estimated daily frequency of importation (7-day rolling average), coloured according to the size of resulting local transmission lineages; shading denotes the associated 95% HPD. For each of the eight largest detected transmission lineages (A to H), the estimated time of importation, TMRCA (inferred time of most recent common ancestor) and TPMRCA (inferred time of parent of MRCA) (bottom left of the panel). (C) Daily frequency of importation (7-day rolling average; black dots) estimated from phylodynamic analysis, without stratification by size of resulting local transmission lineage; error bars denote the associated 95% HPD. Solid blue line represents an exponential model fitted to the observed 7-day rolling average values. (D) Distribution of TPMRCAs and TMRCAs of all 6,455 detected introductions. Each horizontal line represents a single introduction event that led to a transmission lineage or singleton; the left limit indicates the TPMRCA and the right limit indicates the TMRCA (or genome sample date, for a singleton).

To explore this hypothesis, we calculate the Estimated Importation Intensity (EII) of Omicron BA.1 from countries with the highest air traffic volumes to England, capturing 80% of incoming passengers. For each source location, the EII combines the weekly average COVID-19 test positivity rate, weekly relative prevalence of Omicron BA.1 genomes, and monthly number of observed air passengers travelling to England and thus represents a relative rate of importation (refer to Section 2.6.5 for more details; Figs. A.4-

A.6 in Appendix A). While the earliest imports were inferred to have come mostly from South Africa, we observe a diversification in the inferred sources of BA.1 imports by late November/early December 2021 (Fig. 2.2A), during the period of travel restrictions (mandatory hotel quarantine (27)) on international travel from South Africa. We conclude that the exponential growth of BA.1 importations through mid-December is in part due to introductions from countries other than South Africa (Figs. 2.1B, 2.2), as a result of their growing Omicron epidemics and substantial air travel volumes to England (Fig. A.4 in Appendix A). When travel restrictions on 11 southern African countries were first announced (Fig. 2.1A), BA.1 genome sequences from only four countries had been uploaded to GISAID (23). We note that our work is not designed to quantitatively assess the impact of travel restrictions on infection numbers in England.

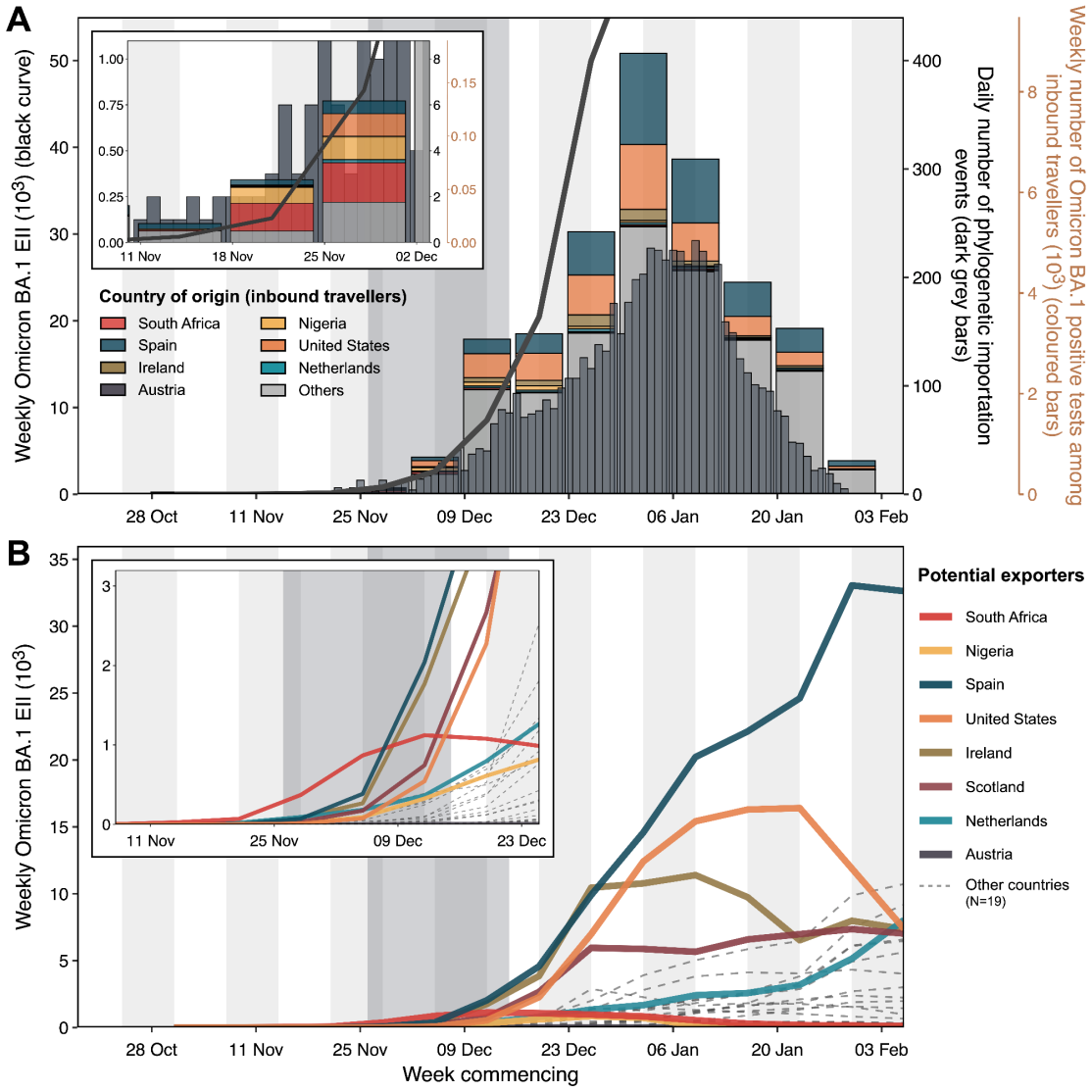


Fig. 2.2. Dynamics of Omicron BA.1 importation into England. (A) Solid curve represents the aggregated EII for 27 countries with the highest air passenger volumes to England between November 2021 and January 2022 (collectively comprising ~80% of air passengers in this period). Coloured bars show the weekly number of inbound travellers who were tested positive for BA.1 following their arrivals in the UK, extracted from travel data compiled by the UKHSA; segments are coloured according to country of origin. Grey bars show the estimated daily number of importation events from phylodynamic analysis. Inset shows a magnified view of early trends. Shaded region indicates the period of travel restrictions on travel from southern African countries. (B) Estimated weekly number of Omicron BA.1 cases arriving in England from 27 countries with the highest air passenger volumes to England between November 2021 and January 2022 (same as those in panel A). Thick solid lines represent weekly EII from eight countries that contribute substantially to overall EII at different times; thin grey lines represent other countries. Inset shows a magnified view of early trends. Shaded region indicates the period of travel restrictions.

To cross-validate the importation dynamics inferred from viral genomes and EIIs using independent data, we collated the travel history of inbound travellers who later tested positive for BA.1 following their arrivals (data generated by the UK Health Security Agency; more details in Section 2.6.3). The early temporal profile of importation from these data is consistent with that inferred from both the EIIs and the phylodynamic analysis (until mid-December; Fig. 2.2A), with the growth of the latter being slightly lagged (Fig. 2.2A). This observation is consistent with previous studies and is likely due to the time-lag between international importation and the first local transmission event observable from genomic data (28). The relative frequency of genomically-identified BA.1 imports among travellers from South Africa and Nigeria declined in mid-December as importation from other countries began to dominate, consistent with the EII results. Observed imports from the phylodynamic analysis also declined in January, likely due to right censoring (the last genome in our dataset was sampled on 31 January).

As with the emergence of previous VOCs in England (28, 29), we find that transmission lineage sizes are overdispersed (Fig. A.2 in Appendix A), with most sampled genomes belonging to a few large transmission lineages. The eight largest lineages (>700 genomes each) together comprise >60% of the English genomes in our dataset (Fig. 2.1B). We infer that six of these eight were imported before restrictions on travel from southern African countries were introduced (26 November), and three could have been introduced before the first epidemiological signal of Omicron (a change in S-gene target failure, SGTF, samples identified by a private lab in South Africa on 15 November; Fig. 2.1B). While aggregation of lineages due to unsampled genetic diversity outside England could have resulted in earlier importation estimates (30), this is unlikely given the enrichment of early genomes and consistency of the observed lineage size distribution with that from simulation (Figs. A.7, A.8 in Appendix A). We observe a strong

association between the size and time of importation of local transmission lineages, with most large transmission lineages attributed to early introductions, before mid-November (Fig. 2.1B). This pattern is recapitulated by a simple mathematical model; if all lineages share the same transmission characteristics, then the date of importation is the main determinant of transmission lineage size when the epidemic in the recipient location is growing exponentially (see Section 2.6.11; Figs. A7, A.8 in Appendix A).

We estimate that ~400 transmission lineages (including the eight largest) resulted from importation before the end of travel restrictions on 15 December (29 lineages were introduced before 26 November). Although these early imports account for only a small proportion (~6%) of the estimated number of introductions, they are responsible collectively for ~80% of estimated BA.1 infections in England by the end of January 2022.

2.4 Human mobility drives spatial expansion and heterogeneity in Omicron BA.1 growth

The rapid increase in Omicron importation in late 2021 led to the establishment of local transmission chains, initially concentrated in Greater London and neighbouring LTLAs in South West and East England. This coincided with early increases in BA.1 prevalence in those regions, as observed from SGTF data and epidemiological prevalence surveys (15). To investigate further the spatiotemporal dynamics of BA.1 in England, we reconstructed the dispersal history of all identified transmission lineages (with >4 genomes) using spatially-explicit phylogeographic techniques. Genomic sample sizes were highly representative of the estimated number of BA.1 cases at the UTLA level in England (Figs. A.9, A.10 in Appendix A).

We observe distinct stages to the spread of BA.1 across England, with the eight largest transmission lineages sharing broadly similar patterns of spatial dispersal. Unlike other VOCs, the first detected BA.1 transmission lineages are more evenly distributed among regions, with ~20% in Greater London, ~15% in the South East, and 13% in the North West (if only introductions before December 2021 are considered, the value for Greater London is 27%). However, most early cases outside Greater London resulted in limited local spatial diffusion (Fig. 2.3, and Figs. A.11, A.12 in Appendix A).

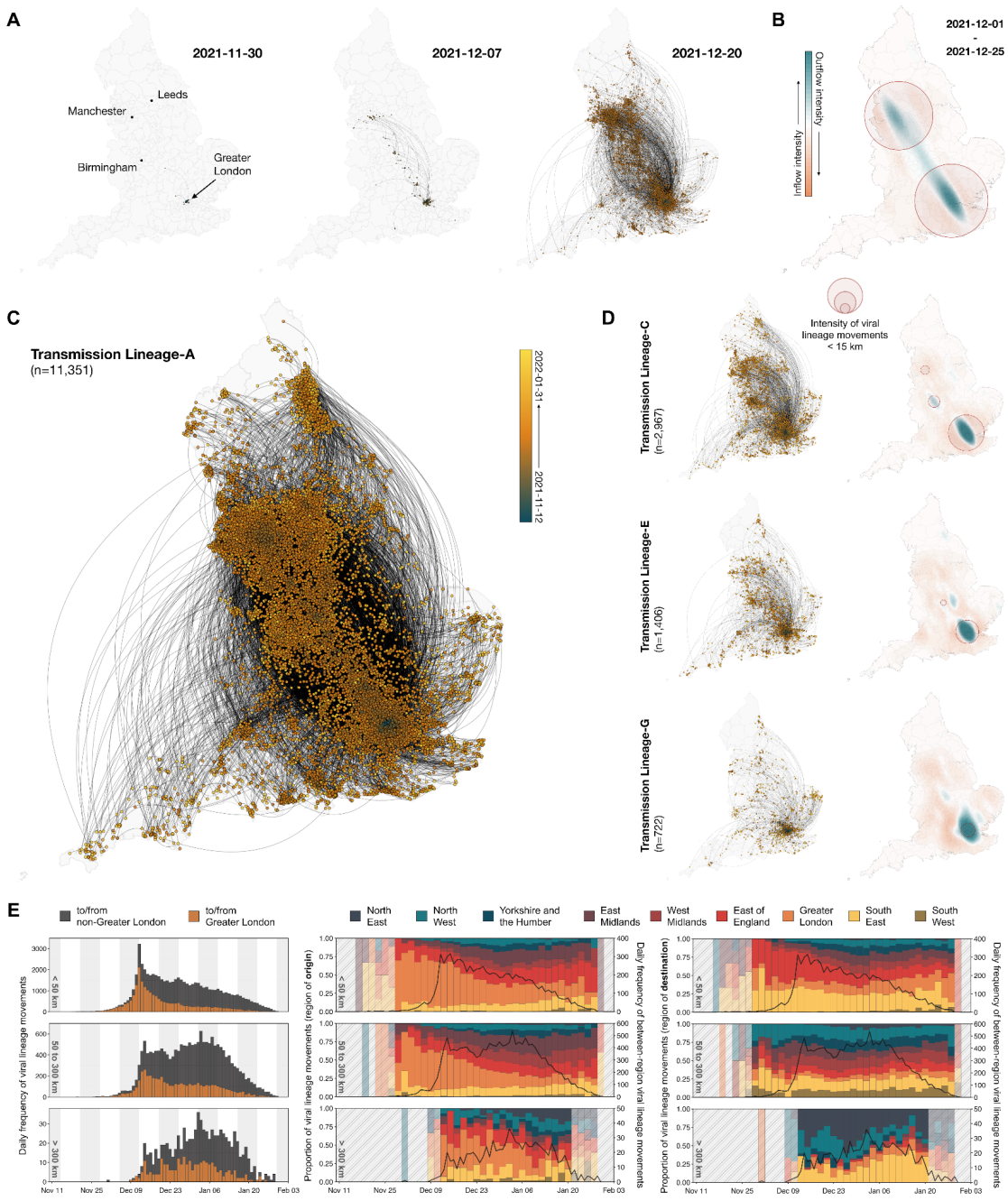


Fig. 2.3. Spatiotemporal dynamics of BA.1 transmission lineages in England. (A and C) Continuous phylogeographic reconstruction of the dispersal history of Transmission Lineage-A, the largest detected BA.1 transmission lineage). Nodes are coloured according to inferred date of occurrence and edge curvature (anti-clockwise) represents the direction of viral lineage movement. Panel A shows the progress of dissemination at three specific times, while panel C shows the complete construction. (B) Geographical distribution of the inflow and outflow of viral lineages within Transmission Lineage-A, from the 1 December to 25 December 2021. Blue colours indicate areas with high intensity of viral lineage outflow; red colours indicate those with high intensity of inflow. Red circles indicate areas with high densities of local viral movements (distances <15 km); circle radii are proportional to that density. (D) Continuous phylogeographic

reconstructions of Transmission Lineages-C, E, and G (as per panel C) with corresponding geographical distributions of viral lineage inflow and outflow (as per panel B). Fig. A.12 in Appendix A provides equivalent figures for Transmission Lineages-B, D, F and H. (E) Plots in each row show viral lineage movements across different spatial scales (top: <50 km, middle: 50 to 300 km, bottom: >300 km). (Left) Histograms show the daily frequency of viral lineage movements; colours indicate whether the origin and/or destination of inferred lineage movements occurred in Greater London. (Middle/Right) Solid black lines represent the daily frequency of among-region viral lineage movements. Vertical bars indicate the proportions of viral lineage movements (aggregated at 2-day intervals); colours indicate origin/destination locations. Shaded grey areas indicate periods when there were <9 inferred viral lineage movements per day.

Initial long-distance viral lineage movements from Greater London repeatedly arrived in multiple urban (as classified in (30)) conurbations in early/mid-December 2021, but local transmission was not established immediately. The fraction of viral lineage movements that were local (within-city) remained between 25%-50% from December 2021 to January 2022 in all areas except Greater London (~90%) and Greater Manchester (~60%). This fraction grew when local mobility levels recovered after the holiday period (31-34), coinciding with the establishment of local transmission across most LTLAs in England (Fig. A.11 in Appendix A). Further, cities other than Greater London acted primarily as sinks throughout the BA.1 wave, with limited backflow of long-distance viral lineages from North West England to Greater London (e.g. Transmission Lineages-A and -B; similar dynamics are seen also for South West England; Fig. 2.3E). We define locations as sinks/sources according to whether there was a net flow of viral lineages into/out of the location over the study period.

Even after the establishment of local transmission in most English LTLAs, Greater London continued to be a source of mid-to-long range viral lineage movements (Fig. 2.3E). This is expected given Greater London's role as a major hub in England's mobility network (similar trends were observed for the Alpha wave in 2020 (26)). The importance of Greater London as a source of short range (<50 km) lineage movements

declined through time (Fig. 2.3E, left-top) and we observe a secondary peak in the frequency of mid-to-long range movements (>50 km) driven predominantly by lineages emanating from the Midlands and southern England (Fig. 2.3E, middle and right). These observations are consistent with epidemiological data showing that most areas outside of southern England experienced a BA.1 incidence peak only in the last week of December 2021 or the first week of January 2022 (Fig. A.13 in Appendix A).

To assess the contribution of demographic, epidemiological, and mobility-related factors to the dissemination of BA.1 in England, we used a phylogeographic generalised linear model (GLM) to test the association of those factors with viral lineage movements among LTLAs, during two distinct periods (before 26 December 2021, and between 26 December 2021 and 31 January 2022; see Section 2.6.13) (32, 33, 35). Using this time-inhomogeneous model we find evidence for a dynamic spatial transmission process, with the estimated effect size and relative importance of most predictors varying over time (Fig. 2.4B; ranking of predictors based on their deviance measure are shown in boxes). During the earlier "expansion" period of lineage dissemination, we observe strong support for the gravity model predictors (a spatial interaction model in which travel intensity between pairs of locations increases with origin and destination population sizes but decreases with distance). Consistent with results from continuous phylogeography (Fig. 2.3), this early period is characterised by directional viral dissemination; lineage movements tend to originate from Greater London (Fig. 2.4B) and this is particularly pronounced for smaller transmission lineages (Fig. 2.3, Fig. A.12 in Appendix A). For LTLAs with earlier times of peak incidence, we also find greater outflow of virus lineages during the expansion period (in three of four analyses) and a lower inflow of viral lineages during the post-expansion period (in four of four analyses; Figs. 2.4, Fig. A.14 in Appendix A). These results reflect the network-driven nature of Omicron's geographic

spread, with variation in the timing of peak incidence reflecting varying degrees of connection to locations where frequent importation seeded early transmission chains (36).

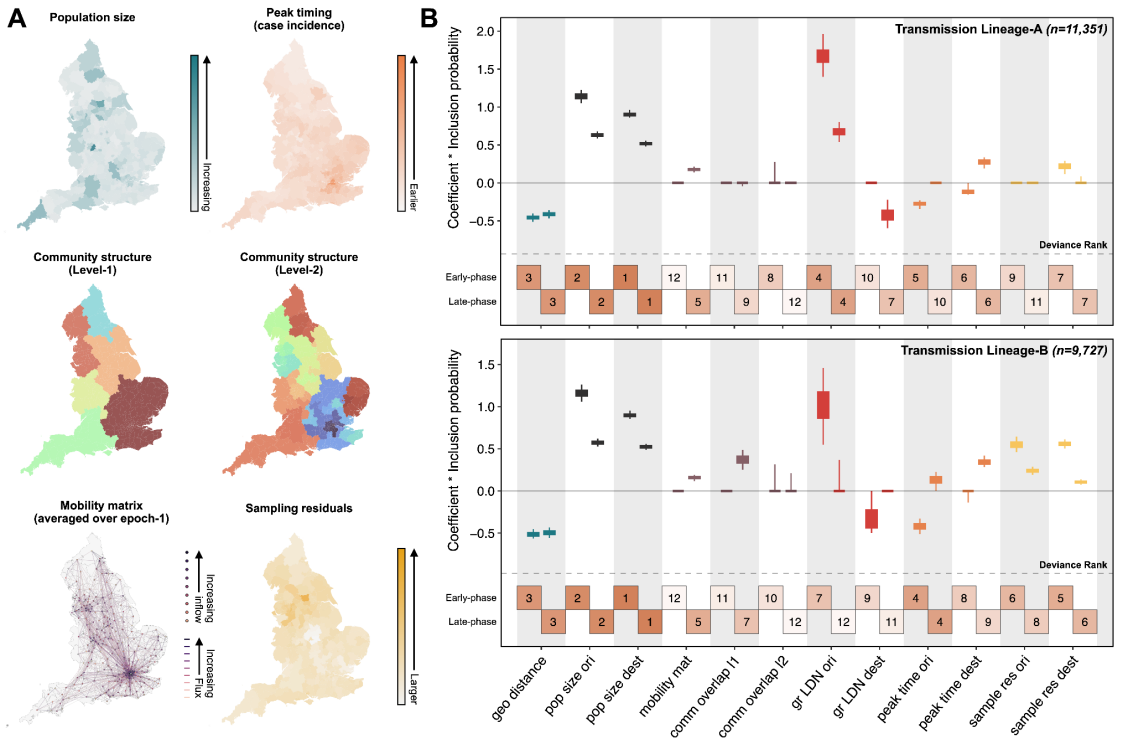


Fig. 2.4. Predictors of BA.1 viral lineage movements in England. (A) Map at LTLA level of model predictors included in the phylogeographic GLM analysis for Transmission Lineage-A. (B) For each predictor, the box and whiskers show the posterior distribution of the product of the log predictor coefficient and the predictor inclusion probability; the left-hand and right-hand values show the estimates for before and after 26 December, respectively. Top and bottom panels show estimates for Transmission Lineages-A and -B respectively. Posterior distributions are coloured according to predictor type: geographic distances (geo distance, dark blue), population sizes at origin and destination (pop size ori/dest, black), aggregated mobility matrix (mobility mat, purple), mobility-based community membership level 1 and level 2 (comm overlap l1 & l2, purple), Greater London origin and destination (gr LDN ori/dest, red), time of peak incidence at origin and destination (peak time ori/dest, orange), the residual of a regression of sample size against case count regression at origin or destination (sample res ori/dest, yellow). Boxes at the bottom of each panel are numbered and shaded to show the ranking of predictors based on their deviance measure (more details in Sectino 2.6.15), with 1 indicating the largest deviance (most important predictor) and 12 indicating the smallest (least important predictor).

The human mobility predictor is supported consistently only in the post-expansion phase (Fig. 2.4B), after local transmission had been established in most LTLAs. This reflects a transition from unidirectional long-distance movements to more homogeneous local dissemination. Conversely, support for the gravity model predictors decreased over time (Fig. 2.4B), consistent with the notion that the gravity model better predicts city-to-city movement and poorly describes diffusion-like mobility over short distances in urban areas (32). Importantly, the phylogeographic GLM results are consistent among the transmission lineages analysed (Fig. 2.4B), and when a simpler time-homogenous model is used (Fig. A.15 in Appendix A). These findings corroborate our continuous phylogeography analyses (Fig. 2.3) and epidemiological studies showing strong local spatial structure of the BA.1 wave (14, 15). We also explored whether booster vaccine uptake (per capita at the LTLA level) is supported as a predictor under a time-inhomogeneous model, but found no significant support (see Section 2.6.14), possibly due to collinearity of this factor with other predictors or limited spatial heterogeneity in vaccine uptake.

2.5 Discussion

We find that most infections during the Omicron BA.1 wave in England can be traced back to a small number of introductions, which likely arrived before or during travel restrictions on incoming passengers from southern Africa. Although the rate of importation continued to increase after mid-December (Fig. 2.1C), the largest English transmission lineages tended to be those introduced earlier (Fig. 2.1D). These results augment previous investigations of VOCs in England and elsewhere (28, 37), highlighting that international travel restrictions can have limited impacts if applied after local exponential growth is established and in the absence of local control measures. Our

analyses indicate that epidemics of BA.1 in multiple locations outside the country where BA.1 was first detected contributed substantially to the growth of BA.1 importation into England in December 2021 (38). The impact of targeted travel restrictions may thus be constrained by the existence of multiple pathways between any two countries in the global aviation network, and such pathways often traverse highly-connected locations with large travel volumes that can act as secondary sources of early importation (36). UK travel restrictions were intended to delay the expansion of BA.1 locally while offering additional vaccination to at-risk individuals. However, Omicron had likely already spread internationally by the time it was detected in late November 2021, allowing the establishment of secondary locations of exportation (38, 39). Therefore, any proposed global systems that aim to rapidly detect and respond to new VOCs (and emerging infectious diseases in general) should be designed around the connection structure of human mobility networks. Despite this, there are likely to be scenarios under which travel restrictions can help control, contain, or delay the spread of emerging infections (40, 41); much further theoretical and empirical work is needed to improve and inform rapid decision-making regarding travel during public health emergencies.

Our two phylogeographic analyses (Figs. 2.3, 2.4) jointly show how Omicron BA.1 disseminated rapidly across England, with Greater London central in its initial dissemination. Early viral movements outside of Greater London were dominated by medium-to-long-distance travel from there; local transmission in recipient locations was observed later, coinciding with an increase in human mobility after the winter holidays (Fig. A.18 in Appendix A). The epidemic is revealed to be a network-driven phenomenon with an initial expansion phase that is well described by a gravity model, followed by a period of sustained local transmission propagated by short-distance movement (36).

With this study, we can now compare the transmission histories of three VOC waves in England (Alpha (26), Delta (29), and Omicron) and contrast factors that influenced their dispersals. First, Omicron and Delta were introduced through international importation, whereas Alpha appeared to have originated in England (42). For both Omicron and Delta, early introductions from their presumed location of origin were followed by growth in importation intensity from secondary locations. While early Delta transmission clusters were observed mainly in North West England, early Omicron infections were found mostly in Greater London (15, 18). Second, different NPIs and restrictions on within-country travel were implemented during the VOC waves. Although Delta arrived when NPIs in England were being relaxed, its initial spread was delayed due to lower mobility levels following a national lockdown (29). In contrast, Omicron was introduced when mobility had largely recovered to pre-pandemic levels (Fig. A.18 in Appendix A). Alpha was observed to rapidly expand from its proposed origin in southeast England, in part due to holiday travels (26) and was subsequently brought under control when local mobility reduced after the introduction of NPIs (26). Third, the dissemination of each VOC is likely to be differentially affected by spatial variation in population immunity. Such variation was likely limited during Delta's emergence due to high population levels of vaccination and previous infection, and also during Omicron's emergence due to the antigenic novelty of BA.1 (9, 43, 44). In contrast, initial growth rates of Alpha in England were found to be affected by local variation in previous attack rates (26). These findings highlight two key questions for future work: how do spatiotemporal interactions between importation and local transmission shape the spread of a VOC, and how can we efficiently evaluate the interplay of factors that drive the dissemination of new VOCs within a country.

We interpret our phylodynamic results in the context of several limitations. First, as discussed previously (28), the inferred number of importation events underestimates the true number of independent introductions due to incomplete sampling and uneven sequencing coverage worldwide (45). Nevertheless, we were able to cross-validate our phylodynamic results using independent epidemiological data (Figs. A7, A.8 in Appendix A). Second, to maintain computational tractability and remove potential sampling bias, we subsampled all available English Omicron genomes, accounting for geographical variations in sequencing coverage and prevalence. However, even after this subsampling, the spatial and temporal sampling was not perfectly representational (Fig. 2.4A, Fig. A.9 in Appendix A). This could be due to spatial variation in case reporting rate or because the maximum sequencing capacity was exceeded in locations with high incidence. Third, our phylogeographic GLM analysis, which explores the association of factors with virus lineage movement, should be interpreted in light of potential biases in the mobility data. For example, mobility in sparsely populated locations may be poorly captured due to censoring to protect user anonymity, and the degree to which smartphone data is representative of the whole population is affected by variation in smartphone use among locations. Work is ongoing to assess how human mobility data can be best applied to the prediction and description of infectious disease invasion dynamics (46, 47).

Omicron BA.1 was replaced by lineage BA.2 in February 2022 and later by lineage BA.5 in June 2022 (18, 19). Although the Public Health Emergency of International Concern has ended (48) and the public health burden of COVID-19 has lessened due to reduced average disease severity and increased population immunity, the continued antigenic evolution of SARS-CoV-2 means that future VOCs of unknown virulence remain possible. One priority in preparing for the next VOC, or novel pathogen emergence, is to develop and implement robust pipelines for large-scale genomic and

epidemiological analyses supported by unified data infrastructures (49, 50) a challenging task that will be realised only through the close coordination of public health efforts worldwide.

2.6 Materials and methods

2.6.1 Genomic data

All SARS-CoV-2 sequences used in this study were downloaded on 12 April 2022. All available international (non-England, including Wales, Scotland and Northern Ireland independently) sequences were downloaded from GISAID (23) while English samples marked as community surveillance (pillar 2) were acquired from COG-UK. Historically, pillar 2 testing sites were instructed to select a number of 96 well plates for sequencing proportional to the fraction of total tests that week. Pillar 2 surveillance is intended to represent a random sample of community cases in the UK, with only 8% of being associated with testing for special reasons, i.e. 'attended-event', 'attended-outbreak-venue', 'confirmatory-test-borders', 'contact-testing-study', 'test-for-contact-self-referral', 'test-for-contact-tracing', 'test-for-contact-tracing-app', 'venue-outbreak'. However, given the changes in testing behaviour and regulations that occurred during the study period we cannot rule out the possibility that there are some biases in the data set. These were partially addressed in the subsampling mentioned below.

Sequences were aligned and filtered as part of the COG-UK datapipe analysis hosted by CLIMB. This analysis removed duplicate and environmental sequences, and flagged samples with improbable collection dates (see https://github.com/COG-UK/datapipe for details). All sequences with impossible or improbable collection dates were removed. To further minimise dating errors caused by retrospective sequencing,

only samples published to COG-UK or GISAID (23) within four weeks of sample collection were included. Sequences were aligned to the reference Wuhan-Hu-1 (genbank accession MN908947.3) with minimap2 and samples with less than 93% coverage were discarded. Sequence coverage weights were calculated for English sequences (24) to ensure they could be subsampled proportional to the number of reported cases in each Upper Tier Local Authority using a two-week sliding window. Scorpio was run as part of Pangolin and sequences identified as BA.1 or BA.2 were selected for further analysis.

2.6.2 Estimated Omicron BA.1 case incidence (from COVID-19 case count and S-gene target failure data)

Daily number of new COVID-19 cases by specimen date in each LTLA were downloaded from https://coronavirus.data.gov.uk/details/download (last accessed on 26 June 2022). S-gene target failure data were provided by UKHSA via a data sharing agreement. The presence of a genetic deletion on the spike protein of the Omicron BA.1 sub-variant produces SGTG in most PCR tests which can be used as a proxy for BA.1 infections. We used daily SGTF PCR-positive tests as a proxy (because these were time- and cost-effective as a test compared to genetic sequencing to ascertain variants) for Omicron BA.1 infection in conjunction with reported case data to estimate daily number of new BA.1 cases. However, small sample sizes in the SGTF dataset could lead to extreme scaling, i.e. zero or 100% of cases could be attributed to BA.1 infections if for example none or all samples were SGTF positive. Hence, we calculated BA.1 cases in a Bayesian framework using uninformative $\mathrm{Beta}(1,1)$ priors and the observed proportion of BA.1 infections (from the SGTF dataset) to estimate the posterior proportion of BA.1 cases which was then scaled up by the number of reported cases from the coronavirus data

download. We can also use the estimated uncertainty from the posterior distribution to get lower and upper bounds in the scaled-up BA.1 case numbers.

2.6.3 Travel history of genomically-identified Omicron BA.1 imports (from UK Health Security Agency)

The travel history data compiled and provided by the UK Health Security Agency (UKHSA) contains the country of origin for all identified inbound travellers arriving in the UK (excluding traveller from GBR and those who travelled to multiple countries) and later tested positive for Omicron BA.1 (confirmed using whole genome sequencing and genotyping) during the weeks from 1 November 2021 to 31 January 2022. This data integrates three sources of travel information: 1) Passenger Locator Forms (PLF) completed at the port of entry by travellers entering the UK from any country, listing all countries visited within the previous 10 days, 2) contact tracing records listing all countries visited by the traveller in the 14 days prior to symptom onset (or a positive test result), and 3) travel information attached to the positive test record including whether the test originated from a managed quarantine facility, whether the diagnostic laboratory was a laboratory specifically testing international arrivals, and if the reason for testing was listed as "isolation-testing".

2.6.4 International passenger flight data arriving in England

We evaluated travel data generated from the International Air Transport Association (IATA) to quantify passenger volumes originating from international airports and arriving in England. IATA data accounts for approximately 90% of passenger travel

itineraries on commercial flights, excluding transportation via unscheduled charter flights (the remainder is modelled using market intelligence).

2.6.5 Estimated importation intensity of Omicron BA.1 from potential exporters

We estimated and compared the weekly importation intensity of SARS-CoV-2 Omicron BA.1 from 27 countries (including Scotland and Northern Ireland independently) with the highest air passenger volumes arriving in England between 7 Nov 2021 and 26 March 2022 (collectively accounting for ~80% of the total air passenger volume during this period). The weekly importation intensity is an estimate of the number of Omicron BA.1 cases arriving in England during a given week from a specified source location, calculated by multiplying together the estimated weekly prevalence of Omicron BA.1 at the source location and the number of air passengers arriving in any England airport from the source location.

We estimated the weekly number of air passengers arriving in England using monthly air traffic data, assuming a uniform daily distribution of passengers throughout the month and aggregating to a weekly level. The weekly prevalence of Omicron BA.1 can be divided into two components, namely, 1) the average non-variant-specific COVID19 prevalence in the week, and 2) the proportion of infections that are Omicron BA.1. The latter was estimated from the proportion of sequenced SARS-CoV-2 genomes that were of Omicron BA.1 (as available from GISAID (23), https://gisaid.org/; last accessed on 2 May 2023). In order to reduce the effects of small numbers, the proportion of infections that are Omicron BA.1 for any given day was first calculated by considering all sequences sampled over the preceding two weeks. This was then further aggregated at the weekly level to calculate the weekly average Omicron BA.1 proportion. To estimate the non-variant-specific COVID19 prevalence and to account for potential biases that

might result from differences in the case reporting rate between countries, we used test positivity as a proxy for the underlying weekly prevalence at the source locations. Daily test positivity rates at the country level were downloaded from OWID (https://ourworldindata.org/; last accessed on 3 April 2023) and their weekly averages were computed. For Scotland and Northern Ireland, daily test positivity rates were calculated using data downloaded from GOV.UK COVID-19 Dashboard (https://coronavirus.data.gov.uk/; last accessed on 23 April 2023), assuming that all reported cases were identified from Pillar 1 and 2 testing. We note that no reliable testing data for Egypt could be found and therefore it was omitted in this EII analysis. Egypt was however included in a subsequent sensitivity analysis where case incidence per capita was used as a proxy for the underlying prevalence (see below), in which it was not observed to be a substantial contributor to BA.1 importation.

In a sensitivity analysis, we further calculated EIIs for Spain and the United States at the autonomous community- and state-level, respectively, to account for any local (within-country) heterogeneities in Omicron BA.1 prevalence and air traffic volume. For Spain, both weekly case incidences and test positivity rates at the autonomous community level were downloaded from the European Centre for Disease Prevention and Control Data Dashboard (https://www.ecdc.europa.eu/en/publications-data/archive-historical-data-testing-volume-covid-19; last accessed on 22 April 2023). For the US, weekly case incidences at the state-level were calculated using data from https://github.com/nytimes/covid-19-data (last accessed on 13 November 2022); weekly test positivity rates were calculated using data from https://github.com/govex/COVID-19/blob/master/data_tables/testing_data/time_series_covid19_US.csv (last accessed on 4 April 2023). We note that three different approaches were used to calculate test positivity for the US states depending on the availability of different test statistics, namely: (A)

positive specimens / total specimens, i.e. the number of positive PCR tests divided by the total number of PCR tests given, (B) positive people / total encounters, i.e. number of people who tested positive (PCR) divided by the total number of PCR tests given, and (C) positive people / total specimens, i.e. the number of people who tested positive (PCR) divided by the total number of PCR tests given. For states where multiple measures of the positivity rates are possible, the optimal approach was applied according to the order (A), (B) and (C), with approach (A) being the optimal approach. For Washington state (WA) in particular, no appropriate measure of the denominator in the calculation of positivity rate is available using any of the approaches, and as a result the national average positivity rates were used instead.

We note that some autonomous communities in Spain (e.g. La Rioja, Valencian Community) and some states in the US (e.g. New Jersey) appeared to have anomalously high (relative to national average) test positivity rates for certain weeks during the study period, potentially due to targeted testing efforts or changes in testing policies. For Spain in particular, there were five autonomous communities (Aragon, Asturias, Melilla, Cantabria, and Region of Murcia) for which the number of sequences sampled per week was consistently below ten over the study period, and as a result, the relative prevalence of Omicron BA.1 could not be reliably estimated. The relative prevalence of BA.1 for these autonomous communities was therefore imputed using the national relative prevalence of BA.1. Despite these data limitations, we find that the total EIIs for Spain and the US (after aggregating over all autonomous communities and states, respectively) are consistent with those calculated using national average positivity rates (Figs. 2.2, Fig. A.5 in Appendix A). To further assess the potential bias that might have resulted from using test positivity rate as a proxy for the underlying prevalence at the source locations, we separately calculated EIIs using weekly COVID-19 case incidence per capita as the

proxy instead (Fig. A.5 in Appendix A). Daily reported case numbers at the country-level were downloaded from the same data sources as in the analysis using test positivity rates (https://ourworldindata.org/; last accessed on 3 April 2023) (with the inclusion of Egypt) and their weekly averages were computed. Similarly, EIIs for Spain and the US were further calculated at the autonomous community and state level using case incidence per capita as a proxy for prevalence. To highlight the potential bias that might have resulted from variations in case reporting between countries, the weekly number of tests taken per capita is calculated for each country, as shown in Fig. A.3 in Appendix A.

We observe that the EII for South Africa did not drop to zero despite the travel restrictions. There are several reasons for this: i) restrictions were imposed between 26 November and 15 December 2021 but our air traffic data is aggregated at a monthly level, and therefore we might not be able to capture any weekly variation in travel intensity (potentially higher right before/after the restrictions and lower during the restrictions); ii) restrictions at the time required travellers to isolate in a government-approved facility for 10 days and take a test on day 2 and 8 during their stay (27). EII therefore does not constitute a measure of the potential for onward local transmission but rather a measure of the relative contribution of Omicron BA.1 infected travellers arriving in England from different countries. According to data from the National Audit Office, approximately 6,000 people were quarantined in managed quarantine hotels between weeks starting on 11 November 2021 and 11 December 2021 (52).

2.6.6 UK population estimates

Mid-year population estimates for England in 2020 at the LTLA level were downloaded from

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/popul

ationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland. Population sizes were used as the denominator in calculating numbers of COVID-19 cases per capita and normalised local mobility in each LTLA.

2.6.7 Vaccination data with age breakdown

Daily vaccination data with age breakdown at the Lower Tier Local Authority (LTLA) level were downloaded from https://coronavirus.data.gov.uk/metrics/doc/vaccinationsAgeDemographics. The dataset consists of daily cumulative number and percentages of people who have received either a 1st dose, 2nd dose, or booster dose (of any type) since the start of the pandemic in each LTLA, with age breakdown by roughly 5-year intervals (5-11, 12-15, 16-17, 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85-89, 90+).

2.6.8 Aggregated and anonymised human mobility data

We used the Google COVID-19 Aggregated Mobility Research Dataset described in detail in (53, 54), which contains anonymized relative mobility flows aggregated over users who have turned on the *Location History* setting, which is turned off by default. This is similar to the data used to show how busy certain types of places are in Google Maps – helping identify when a local business tends to be the most crowded. The mobility flux is aggregated per week, between pairs of approximately 5km^2 cells worldwide, and for the purpose of this study further aggregated for LTLAs in the United Kingdom (https://geoportal.statistics.gov.uk/datasets/lower-tier-local-authority-to-upper-tier-

local-authority-december-2016-lookup-in-england-and-wales/explore) for the time period of November 2019 to January 31st, 2022.

To produce this dataset, machine learning is applied to log data to automatically segment it into semantic trips. To provide strong privacy guarantees (55), all trips were anonymized and aggregated using a differentially private mechanism to aggregate flows over time (see https://policies.google.com/technologies/anonymization). This research is done on the resulting heavily aggregated and differentially private data. No individual user data was ever manually inspected, only heavily aggregated flows of large populations were handled. All anonymized trips are processed in aggregate to extract their origin and destination location and time. For example, if n users travelled from location a to location b within time interval t, the corresponding cell (a, b, t) in the tensor would be n±err, where err is Laplacian noise. The automated Laplace mechanism adds random noise drawn from a zero mean Laplacian distribution and yields (ϵ, δ)-differential privacy guarantee of $\epsilon = 0.66$ and $\delta = 2.1 \times 10{-}29$ per metric. Specifically, for each week W and each location pair (A, B), we compute the number of unique users who took a trip from location A to location B during week W. To each of these metrics, we add Laplace noise from a zero-mean distribution of scale 1/0.66. We then remove all metrics for which the noisy number of users is lower than 100, following the process described in (55), and publish the rest. This yields that each metric we publish satisfies (ε, δ)-differential privacy with values defined above. The parameter ϵ controls the noise intensity in terms of its variance, while δ represents the deviation from pure ϵ-privacy. The closer they are to zero, the stronger the privacy guarantees.

These results should be interpreted in light of several important limitations. First, the Google mobility data is limited to smartphone users who have opted in to Google's *Location History* feature, which is off by default. These data may not be representative

of the population as whole, and furthermore their representativeness may vary by location. Importantly, these limited data are only viewed through the lens of differential privacy algorithms, specifically designed to protect user anonymity and obscure fine detail. Moreover, comparisons across rather than within locations are only descriptive since these regions can differ in substantial ways.

2.6.9 Changes in case reporting rate in the United Kingdom

To assess the degree of changes in case reporting rate in the United Kingdom, we compared the weekly national case incidence downloaded from the GOV.UK COVID-19 Dashboard with that estimated by the UK Office of National Statistics (ONS). Specifically, since we were interested in the relative changes over time rather than the absolute values, a linear regression of the ONS case incidence estimates against case incidence from the GOV.UK COVID-19 Dashboard was performed and the residuals from the model were examined (Fig. A.10 in Appendix A).

As described in further details below, Omicron sequences from England were subsampled with sample weights calculated from the ratio between the cumulative number of reported cases and the cumulative number of sequences collected in the preceding two weeks for any given date. Therefore, to assess the potential bias that might have resulted from changes in case reporting rate in the context of the subsampling of English genomes, a similar linear regression analysis as above was performed, with additional (two-preceding-weeks) smoothing applied to both the ONS and GOV.UK case incidences (Fig. A.10 in Appendix A). We note that is some evidence for an increase in case reporting rate in the first week (starting on 28 November 2021; like due to targeted COVID-19 testing and increases in contact tracing intensity at the beginning of the BA.1 wave) and a decrease in the last week of the study period (likely as a result of a policy

change on 11 January 2022 (56), where people who received positive LFD test results for COVID-19 were no longer required to take a confirmatory PCR test). However, the overall magnitude of the changes in case reporting rate is small and therefore we do not expect any substantial bias in our inferences as a result.

2.6.10 Phylogenetic and importation analysis

We developed a large-scale phylogenetic analysis pipeline following a similar approach as in du Plessis et al. (2021) (28) with additional extensions and modifications to ensure the computational tractability of analyses of up to hundreds of thousands of SARS-CoV-2 sequences (57) (Fig. A.1 in Appendix A).

First, the study period was divided into two phases: i) from 21 November 2021 (sample date of the earliest known genome of the Omicron variant in England, sequenced retrospectively) to 28 November 2021, and ii) from 29 November 2021 to 31 January 2022. The time of division between the two phases was chosen on the basis of an expected change in importation intensity as a result of the implementation of travel restrictions targeted at multiple southern African countries starting on 28 November 2021. With the relatively few genomes available from the first phase and to account for an increased risk of importations prior to the travel restrictions, all 874 available sequences (from both England and non-England locations) were included. Owing to the large number of genome samples collected during the second phase, a downsampling strategy was applied to ensure that the analysis was computationally tractable. To generate a manageable dataset of global sequences, first we computed a crude estimate of the number of new Omicron cases in each country in each epi-week by multiplying the number of reported COVID-19 cases (downloaded from https://github.com/owid; last accessed on 4 May 2022) by the proportion of sampled genomes that were of the Omicron variant PANGO

lineages BA.1 and BA.2, using metadata available from GISAID (23) (https://gisaid.org/; accessed on 12 April 2022). The number of global sequences to be sampled in each epi-week was then allocated in proportion to the estimated total number of Omicron lineages BA.1 and BA.2 cases in the week whilst maintaining a dataset size of ~50,000. In a given epi-week, countries with an estimated number of Omicron cases that accounted for at least 0.5% of the estimated global total were considered as potential exporters. Genome samples were then allocated in proportion to the estimated number of cases among these potential exporters, with the remaining allocation randomly distributed among the non-exporter countries. There was a slight enrichment for samples collected in the early phase of the Omicron wave (early December 2021), where we ensured that a minimum of 4,000 genomes were sampled for each epi-week where available. A similar approach was used to curate a dataset of 21,039 Omicron genomes sampled from Wales, Scotland and Northern Ireland, again using relevant metadata from GISAID (23) and epidemiological data available on (https://api.coronavirus.data.gov.uk/v1/data; last accessed on last accessed on 4 May 2022). This downsampling procedure resulted in a dataset of 59,647 global (non-English) sequences. To generate a dataset of English genomes of roughly the same size, 60,000 sequences were randomly sampled from the COG-UK master alignment whilst accounting for variations in sequencing coverage and prevalence amongst UTLAs over time, using the same method as in Volz et al. (58). This resulted in a combined dataset of 140,686 genomes of which 42.6% were sampled in England with the remaining from non-England locations.

Despite substantial downsampling, estimating a phylogenetic tree for hundreds of thousands of SARS-CoV-2 sequences remains a challenge, with most standard programs only able to handle up to thousands of sequences. To tackle this, we first estimated a maximum likelihood (ML) tree for the 874 sequences collected during the first phase of

the study period using IQTREE (59) with the GTR+G substitution model, rooted with reference genome Wuhan-Hu-1 (GenBank accession MN908947.3) as an outgroup. Five molecular clock outliers were identified and subsequently removed, after examining the root-to-tip regression plot from TreeTime (60). The resulting tree was then used as a starting tree from which a parsimony tree was estimated by inserting individual sequences sequentially and in chronological order according to sample dates, using the recently developed UShER placement tool (61). During each step in the iterative process, all sequences sampled on a given date were considered for placement whilst excluding sequences with 5 or more equally parsimonious placements. Sequences excluded in a previous step were appended to the next batch for reconsideration. The resulting tree was then optimised through 6 iterations of matOptimize (62) with SPR radius of 40 and 100 for the first 5 and final iteration respectively. This iterative tree building process resulted in a phylogeny of 115,634 sequences (with 25,921 (18.3%) sequences excluded due to uncertainty in sample placement). Next we used Chronumental (63) (a recently developed time-tree estimation tool for handling large phylogenies) to estimate a randomly resolved time-calibrated tree, with inferred tip dates that maximise the evidence lower bound under a probabilistic model. By comparing the inferred tip dates with sample dates and examining a root-to-tip plot, 12 molecular clock outliers were further removed, resulting in a final phylogeny of 115,622 sequences.

To further reduce the computational resources and time required, we divided the phylogeny estimated above into smaller tree partitions according to sub-lineage (of Omicron) assignment as defined by the Pango nomenclature (64). Using a custom Python script, subtrees with a high degree of clustering of sequences of the same descendant lineage of Omicron were identified, whilst accounting for some level of ambiguity in lineage assignment (e.g. a tree partition may contain up to 25% of sequences that are of

a minority sub-lineage before it is subdivided into multiple partitions), as would be expected given the high sampling density and variations in sequencing quality. Further merging of these identified subtrees resulted in five final tree partitions, labelled BA.1 (n=38,522), BA.1.1 (n=37,028), BA.1.15 (n=12,229), BA.1.17 (n=21,549), and BA.2 (6,294) according to the sub-lineage represented by the majority of sequences in each partition. Given that the primary focus of this study is the invasion dynamics of Omicron BA.1 in England, the BA.2 partition was omitted in all further downstream analyses.

Having divided the phylogeny into smaller tree partitions of computationally manageable size, we then performed time-calibration of the subtrees using a recently implemented model in BEAST v1.10 (65) which replaces the traditional tree-likelihood with a more efficient likelihood based on a simple Poisson model, thus allowing Bayesian phylogenetic analyses of up to tens of thousands of sequences. In this approach, the tree operators are constrained such that only node heights and polytomy resolutions are sampled, whilst the tree topology is fixed to that of a data tree which we generated using Treetime (60) with a fixed clock rate of 7.5x10E-4 substitutions/site/yr. Using a Skygrid coalescent tree prior (66) with grid points at weekly intervals, we ran between 2 and 6 MCMC chains of 3×10^8 to 2.4×10^9 iterations for each tree partition independently. The first 33% to 40% of each chain was discarded as burn-in and resampled every 1×10^6 to 2.4×10^8 states before merging using LogCombiner, resulting in 1,200 posterior tree samples for each tree partition. Model convergence and mixing was assessed using Tracer (67).

To reconstruct the importation dynamics of Omicron BA.1, we then used a two-state asymmetric discrete trait analysis (DTA) model implemented in BEAST v1.10 (65), using the posterior tree samples estimated above as the empirical tree distributions. For each tree partition, we ran two MCMC chains of 5 million iterations each, resampled

every 9,000 states and with the first 10% discarded as burn-in. TreeAnnotator 1.10 (65) was used to generate a maximum clade credibility (MCC) tree for each subtree, in which each internal node is assigned a posterior probability of representing a transmission event in England. Nodes with a posterior probability of >0.5 were identified as introductions; a small number of nodes with ambiguous location assignment (posterior probability = 0.5) were ignored in downstream analyses. To identify the local transmission lineage resulting from each of the introductions, a depth-first search was performed following the same procedure as in du Plessis et al. (2021) (28), where a path starting from each internal node that corresponds to an introduction is traversed forwards in time until a non-England node is encountered or there are no more nodes to be explored. By convention, introductions that led to only a single sampled English sequence were labelled as singletons; only introductions that led to more than one observed local transmission event were labelled as transmission lineages. The time of importation of each transmission lineage was estimated by taking the mid-point between the internal node corresponding to the introduction and its parent.

Our methodology estimating the time of importation of transmission lineages is likely to result in an apparent "expansion" of the temporal profile of inferred importation intensity (daily number of infected travellers arriving in England) relative to its true underlying distribution. This could be explained by an increase in importation-lag (time elapsed between when a lineage is inferred to have been imported and the first observed local transmission event) over time as shown previously by du Plessis et al. (28), due to transmission lineages from later importation having fewer genomes as they had less time to grow, and are therefore more likely to have estimated time of importation that is later than the true value as a result of mis-identification of the true root node. To verify this effect, we compared the inferred importation intensity from the above phylodynamic

analysis with travel history data (generated by UKHSA) of inbound travellers who were tested positive for BA.1 after their arrivals in the UK. We observed a broadly consistent temporal profile in the importation intensity inferred from the two datasets, with that from the phylodynamic analysis being slightly lagged in time as expected (Fig. 2.2). However, we note that the robustness of this comparison is potentially limited by variations in sampling intensity as a result of rapidly changing testing policies for arriving travellers in the United Kingdom during the later part of January 2022 (68).

2.6.11 Exponential growth of daily frequency of importations

In the absence of any travel restrictions and changes in human mobility as a result of the emergence of a new VOC, the importation intensity during the initial phase of the invasion would be expected to follow a pattern of exponential growth that mirrors the increase in number of infections in the exporting countries. To verify and examine any potential deviation from this pattern, we fitted a simple exponential model to the 7-day rolling average daily number of importations inferred from the phylodynamic analysis. Specifically, we fitted the model using least-squares regression to the inferred daily numbers of importations during the period between the beginning of November 2021 and a range of cut-off dates. The cut-off date that resulted in the highest adjusted R^2 value can be interpreted as an estimate of the time when the growth of importation intensity began to deviate from an exponential trajectory.

2.6.12 Continuous phylogeographic reconstruction of local spread

To reconstruct the spatiotemporal patterns of the Omicron BA.1 wave in England, all local transmission lineages (as identified from the MCC trees generated from the 2-state

discrete trait analysis described above) with five or more sequences were extracted for continuous phylogeographic analyses. Each sequence was assigned a latitude and a longitude randomly from within the postal district (metadata provided by COG-UK) where the sample was collected. For each transmission lineage, we performed the continuous phylogeographic reconstruction on a fixed tree (pruned from the MCC tree) using a relaxed random walk model (69) implemented in BEAST 1.10.4 (65), with a Cauchy distribution to model the among-branch heterogeneity in dispersal velocity. Following a similar approach as in McCrone et al. (29), the eight largest transmission lineages (labelled A to H, from largest to smallest) containing >700 sequences were inferred independently, with the remaining smaller transmission lineages (n=524) inferred in a single joint analysis with a shared diffusion model (i.e. same parameter estimates for likelihood, precision matrix, correlation, etc, but independent estimates for diffusion rate and trait likelihood). Owing to variations in the extent of spatial dispersal among these smaller transmission lineages (with larger lineages being more spatially dispersed in general), 12 were subsequently omitted from downstream analyses and 18 were further inferred independently. Model convergence and mixing was assessed using Tracer v1.7 (67). For the independent analyses of the eight largest transmission lineages, we ran between 2 and 5 MCMC chains of 200 to 300 million iterations, sampling every 10,000 to 80,000 states and removing the first 10% to 33% of each chain as burn-in, resulting in 10,000 to 13,5000 trees sampled from the posterior distribution. For the independent analyses of the 18 smaller transmission lineages with fewer than 700 sequences, we ran 2 MCMC chains each of 200 million iterations which we then merged after resampling every 30,000 states and removing the first 10% as burn-in, giving 12,000 posterior trees per transmission lineage. Finally, in the joint analysis, 8 independent chains of 200 million were run with sampling every 120,000 states. They were combined

after removing the first 10% as burn-in, again resulting in 12,000 posterior tree samples for each transmission lineage. These posterior tree samples were then used to generate an annotated MCC tree for each transmission lineage using TreeAnnotator (65).

To facilitate subsequent analyses of viral lineage movements at the LTLA level, we mapped the inferred location of each internal node in the transmission lineages to its corresponding LTLA by checking whether the inferred coordinates are contained within the associated polygon. In the case where an enclosing polygon could not be found (e.g. a small proportion of internal nodes were inferred to lie in the small spaces between neighbouring polygons), the polygon that is geographically closest to the inferred location was assigned.

2.6.13 Discrete phylogeographic reconstruction of local spread with Generalised Linear Model (GLM) parameterisation

We used the approach of discrete phylogeography with generalised linear model (GLM) to parameterise transition rates between locations and test the association of viral lineage dispersal with a number of geographical, demographic, epidemiological and human mobility-related factors (see Table A.3 in Appendix A for full list of predictors). Specifically, to test the gravity model as a predictor of viral lineage movements, we considered in the GLM analysis the population size at the origin and destination location of each movement and the geographical distance between them. To further capture any heterogeneities in aggregated human mobility at the city-level (which are unlikely to be adequately described by the gravity model), we also included the aggregated mobility matrix and community memberships as predictors. We allowed these mobility-related predictors to vary across different phases in the time-inhomogeneous model to test for temporal variations in aggregated human mobility patterns and also potentially time-

varying effect of mobility on viral dispersal. We observed from both epidemiological data and continuous phylogeography that many LTLAs in Greater London experienced an earlier uptick in Omicron BA.1 cases compared to most LTLAs with other regions of England. To capture this asynchronicity in local epidemic dynamics and investigate its impact on viral dispersal, we considered in the GLM analysis whether each viral movement started or ended in the Greater London region and additionally the time of first peak in Omicron BA.1 case incidence at the origin and destination location. Furthermore, we also tested for the impact of sampling bias by including a predictor based on the residuals from a simple regression of sample size against Omicron BA.1 cases for both the origin and destination location. Due to the small number of sequences collected in some LTLAs especially during later phases of the epidemic, the regression residuals were computed using sample sizes and case counts aggregated over the whole study period in both the time-homogeneous and time-inhomogeneous models.

Unlike continuous phylogeography where each sequence is assigned a unique set of coordinates in continuous space, discrete phylogeography requires that sequences are grouped into discrete geographical units. The level of granularity of these geographical units depends on a number of factors including i) the desired level of resolution at which the dispersal history is to be reconstructed, ii) the amount of heterogeneities present within each geographical unit, and iii) the maximum number of geographical units beyond which the analysis becomes computationally intractable. To capture heterogeneities in viral movements at the city-level and to allow comparisons with results from continuous phylogeography, we allocated sequences to their corresponding LTLAs using a lookup table which provides unique mapping between postal districts and LTLAs.

The current computational architecture and implementation of the discrete phytogeographic GLM model limits the number of discrete units possible to 256, which

is smaller than the number of LTLAs across which sequences were sampled for some of the larger transmission lineages. To tackle this, we aggregated LTLAs where appropriate to reduce the number of geographic units. In order to minimise the resulting information loss, we first considered LTLAs with the fewest sequences and performed a merging operation if an adjacent LTLA with at least one sampled genome could be found. In the case where multiple adjacent LTLAs were available, the LTLA with the largest number of sampled genomes was chosen for the merger. After each merging operation, the list of LTLAs (or geographical units after merging) ranked by the number of sampled genomes was recalculated for the next iterative step (it is therefore possible for an LTLA to be involved with multiple merging operations). This process continued until there were only 253 geographic units in each transmission lineage. For the geographical units consisting of multiple LTLAs, each statistic of interest was averaged over the relevant LTLAs, weighted by population size where appropriate. For transmission lineages with sequences sampled in fewer than 256 LTLAs, no merging was performed.

The discrete phylogeographic GLM model parameterizes the log of between-location transition rates as a log linear function of the predictors. Continuous predictors (geographical distances, population sizes, aggregated mobility matrices, peak timing in case incidence, sampling residuals) were therefore log-transformed and standardised after adding a pseudo-count to each entry where appropriate. Binary variables (community memberships, Greater London/non-Greater London) were encoded as 0 and 1. In the mobility-related predictors, there was missing data for one or two geographical units in some transmission lineages (due to mobility data being unavailable for South Tyneside and City of London), which we labelled as NA and later integrated out in our Bayesian inference. For the aggregated mobility matrix predictor with continuous values in the large-scale transmission analyses, we confronted this using a new Hamiltonian Monte

Carlo (HMC) kernel to jointly sample all missing covariates from their posterior distributions building on similar efforts in the BEAST framework (70, 71). The HMC kernel produces distant proposals with relatively high acceptance rate for the Metropolis algorithm by exploiting numerical solutions to the Hamiltonian dynamics. We performed the analyses using the code available in the hmc-clock branch of the BEAST codebase (available at https://github.com/beast-dev/beast-mcmc/tree/hmc-clock) in conjunction with the BEAGLE code available in the hmc-clock branch of the codebase (available at https://github.com/beagle-dev/beagle-lib/tree/hmc-clock). We ran the analyses on a set of 100 empirical trees for each transmission lineage extracted from the BEAST importation analysis and ran sufficiently long chains sampling every 500 generations, or combined multiple chains (excluding adequate burn-ins), to ensure effective sample sizes (ESSs) > 100 for the continuous parameters as diagnosed using Tracer (67). A custom R script was used to summarise and visualise the posterior coefficient estimates and inclusion probabilities of each predictor.

2.6.14 Discrete phylogeography with GLM: effect of booster uptake

The rollout of booster vaccination in the United Kingdom began in September 2021 (72) and was initially prioritised for those aged 50 and above as they are at a higher risk of severe symptoms and hospitalisation from infection. Eligibility for boosters was extended to those aged 40 and above on 22 November 2021, and subsequently to all adults on 30 November 2021 (73). This resulted in spatial variations in booster uptake that are strongly correlated with the underlying age structure of the population (Fig. A.16, A and C, in Appendix A), which is in turn correlated with Omicron BA.1 prevalences due age-dependent transmission patterns as shown by Elliott et al. (2022) (15) (Fig. A.16, B and D, in Appendix A).

To adjust for age structure as a confounder, we here devise an effective measure of the booster uptake that is independent of the underlying age structure of the population. Using vaccination data (downloaded from the GOV.UK COVID-19 Dashboard) consisting of the percentage of people in different age groups who have received a booster dose, we calculate the overall booster uptake in each Lower Tier Local Authority (LTLA) assuming an age distribution that is the same as the national population-weighted average age distribution (computed from mid-2020 population estimates published by the UK Office of National Statistics). This is equivalent to the overall proportion of the population who would have received a booster dose in an LTLA given its observed age-specific booster uptake (with roughly 5-year grouping), assuming that it has the same age structure as the national average. Similar to other covariates included in the GLM analysis, for the geographical units consisting of multiple LTLAs, the effective booster uptake is averaged over the relevant LTLAs weighted by population size.

2.6.15 Discrete phylogeography with GLM: likelihood-deviance measure

To evaluate the relative importance of the different predictors in the time-inhomogeneous GLM analysis, we have developed and implemented a new phylogeographic model-fit measure which builds upon standard, permutation-based machine learning approaches to assessing variable importance (75).

Starting from the posterior $p(\Theta|Y, x_{11}, \ldots, x_{ke}, \ldots, x_{K2})$ with phylogeographic likelihood $p(Y|\Theta, x_{11}, \ldots, x_{ke}, \ldots, x_{K2})$ where Θ represents all model parameters and x_{ke} represents the vector of covariate values for covariate $k \in \{1, \ldots, K\}$ (K being the total number of predictors in the model) in epoch $e \in \{1,2\}$, we define the deviance for this covariate as $d_{ke} = \log p(Y|\Theta, x_{11}^{\pi}, \ldots, x_{ke}^{\pi}, \ldots, x_{K2}^{\pi}) - \log p(Y|\Theta, x_{11}, \ldots, x_{ke}, \ldots, x_{K2})$ where x_{ke}^{π} is a random permutation of the observed covariate vector x_{ke}. To estimate the

posterior distribution of d_{ke}, we computed values of d_{ke} for each MCMC sample, s, with realised model parameters $\Theta^{(s)}$ by setting $\Theta = \Theta^{(s)}$, and randomly permuting x_{ke} with equal probability for all possible permutations. For each covariate and in each epoch, we estimated the posterior distribution of the likelihood-deviance resulting from a random permutation of the covariate values. We then compared the resulting posterior distributions and ranked the importance of each covariate in predicting the geographic locations Y. The covariate marginal posterior modal (most probable) ranking was then reported, as shown in Fig. 2.4 and Fig. A.14 in Appendix A (with 1 being most important). While this approach averages over all possible marginal permutations and therefore has improved performance over earlier permutation-based measures in machine learning (76), it may nevertheless return limited discrimination among highly correlated covariates. Permute-and-relearn importance methods (77) are able to overcome this limitation but remain computationally impractical given the numbers of tips and the size of the state-spaces considered in this study.

Using the above approach, we find that the predictor rankings do not always reflect differences in absolute effect size and that they help to identify similarities and differences between transmission lineages, as well as between epochs for a given transmission lineage. In the two largest transmission lineages, the gravity model covariates are consistently the most important covariates, in both the early- and late-epoch. For Transmission Lineage-A, the Greater London origin predictor is the next important predictor throughout the study period. The Greater London origin predictor is also more important in Transmission Lineage-A than in Transmission Lineage-B, for which a change in importance of this predictor between the early- and late-epoch is observed. In Transmission Lineage-B, the origin peak time predictor is the next important predictor after the gravity model predictors. For both transmission lineages, we observe

a large and consistent increase in the importance of the mobility matrix predictor between the early- and late-epoch.

We also note that the magnitude of the likelihood-deviance estimates scales with the size of the dataset (and therefore the number of tips in the transmission lineages). As such, the deviance estimates do not provide a relative measure of fit across transmission lineages.

2.6.16 Branching process model and comparison of transmission lineage size distributions

To verify that the time of importation is the key determinant of transmission lineage size, we compared the size distribution of empirically observed transmission lineages with that from a model that simulates the branching process of transmission lineages following importation. Simulated importation dates are set to the dates estimated from the phylodynamic analysis and simulated transmissions occur at the spatially homogeneous growth rates estimated from the daily number of reported COVID-19 cases (from the GOV.UK COVID-19 Dashboard) and SGTF data in England. Due to the low number of Omicron BA.1 cases at the beginning of the epidemic, which can lead to unreliable estimates of the initial growth rate, we performed a series of simulations with a range of different starting growth rates (taken from estimates during early parts of the invasion). We computed the Kullback-Leibler (KL) distance between the size distribution of simulated lineages and that of lineages inferred from phylodynamic analysis (Table A.1 in Appendix A). The growth rate that minimised the KL distance was then used to impute the initial growth rate in the best-fit model. We note that, given the simple nature of the model, we did not take into account any uncertainties associated with the case growth rates but relied only on the central estimates. As a sensitivity analysis for the potential

bias that might have resulted from this and also any changes in case reporting rate during the study period, we repeated the simulations using case incidence estimates from the UK Office of National Statistics (ONS) (see Fig. A.10 in Appendix A for comparisons between case incidence data from UK.GOV COVID-19 Dashboard and ONS estimates). We observed consistent results as those obtained using the case incidence data from the GOV.UK COVID19 Dashboard (Figs. A.7, A.8 in Appendix A).

2.7 References

1. Viana, R., Moyo, S., Amoako, D.G., Tegally, H., Scheepers, C., Althaus, C.L., Anyaneji, U.J., Bester, P.A., Boni, M.F., Chand, M., Choga, W.T., Colquhoun, R., Davids, M., Deforche, K., Doolabh, D., du Plessis, L., Engelbrecht, S., Everatt, J., Giandhari, J., Giovanetti, M., Hardie, D., Hill, V., Hsiao, N.-Y., Iranzadeh, A., Ismail, A., Joseph, C., Joseph, R., Koopile, L., Kosakovsky Pond, S.L., Kraemer, M.U.G., Kuate-Lere, L., Laguda-Akingba, O., Lesetedi-Mafoko, O., Lessells, R.J., Lockman, S., Lucaci, A.G., Maharaj, A., Mahlangu, B., Maponga, T., Mahlakwane, K., Makatini, Z., Marais, G., Maruapula, D., Masupu, K., Matshaba, M., Mayaphi, S., Mbhele, N., Mbulawa, M.B., Mendes, A., Mlisana, K., Mnguni, A., Mohale, T., Moir, M., Moruisi, K., Mosepele, M., Motsatsi, G., Motswaledi, M.S., Mphoyakgosi, T., Msomi, N., Mwangi, P.N., Naidoo, Y., Ntuli, N., Nyaga, M., Olubayo, L., Pillay, S., Radibe, B., Ramphal, Y., Ramphal, U., San, J.E., Scott, L., Shapiro, R., Singh, L., Smith-Lawrence, P., Stevens, W., Strydom, A., Subramoney, K., Tebeila, N., Tshiabuila, D., Tsui, J., van Wyk, S., Weaver, S., Wibmer, C.K., Wilkinson, E., Wolter, N., Zarebski, A.E., Zuze, B., Goedhals, D., Preiser, W., Treurnicht, F., Venter, M., Williamson, C., Pybus, O.G., Bhiman, J., Glass, A., Martin, D.P., Rambaut, A., Gaseitsiwe, S., von Gottberg, A. and de Oliveira, T. (2022) 'Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa', *Nature*, 603(7902), pp. 679–686.
2. World Health Organisation (no date) *Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern*. Available at: https://www.who.int/news-room/statements/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern (Accessed: 1 April 2022).
3. Pulliam, J.R.C., van Schalkwyk, C., Govender, N., von Gottberg, A., Cohen, C., Groome, M.J., Dushoff, J., Mlisana, K. and Moultrie, H. (2022) 'Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa', *Science*, 376(6593), p. eabn4947.
4. Rössler, A., Riepler, L., Bante, D., von Laer, D. and Kimpel, J. (2022) 'SARS-CoV-2 Omicron Variant Neutralization in Serum from Vaccinated and Convalescent Persons', *The New England journal of medicine*, 386(7), pp. 698–700.
5. Lyngse, F.P., Mortensen, L.H., Denwood, M.J., Christiansen, L.E., Møller, C.H., Skov, R.L., Spiess, K., Fomsgaard, A., Lassaunière, R., Rasmussen, M., Stegger, M., Nielsen, C., Sieber, R.N., Cohen, A.S., Møller, F.T., Overvad, M., Mølbak, K., Krause, T.G. and Kirkeby, C.T. (2022) 'Household transmission of the SARS-CoV-2 Omicron variant in Denmark', *Nature communications*, 13(1), p. 5573.
6. Backer, J.A., Eggink, D., Andeweg, S.P., Veldhuijzen, I.K., van Maarseveen, N., Vermaas, K., Vlaemynck, B., Schepers, R., van den Hof, S., Reusken, C.B. and Wallinga, J. (2022) 'Shorter serial intervals in SARS-CoV-2 cases with Omicron BA.1 variant compared with Delta variant, the Netherlands, 13 to 26 December 2021', *Eurosurveillance*, 27(6), 2200042.
7. UK Health Security Agency (2022) *SARS-CoV-2 variants of concern and variants under investigation in England: Technical briefing 36*. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1056487/Technical-Briefing-36-22.02.22.pdf (Accessed: 23 May 2022).
8. Hui, K.P.Y., Ho, J.C.W., Cheung, M.-C., Ng, K.-C., Ching, R.H.H., Lai, K.-L., Kam, T.T., Gu, H., Sit, K.-Y., Hsin, M.K.Y., Au, T.W.K., Poon, L.L.M., Peiris, M.,

Nicholls, J.M. and Chan, M.C.W. (2022) 'SARS-CoV-2 Omicron variant replication in human bronchus and lung ex vivo', *Nature*, 603(7902), pp. 715–720.
9. Peacock, T.P., Brown, J.C., Zhou, J., Thakur, N., Sukhova, K., Newman, J., Kugathasan, R., Yan, A.W.C., Furnon, W., De Lorenzo, G. and Others (2022) 'The altered entry pathway and antigenic distance of the SARS-CoV-2 Omicron variant map to separate domains of spike protein', *biorxiv*, 15, p. e0241955.
10. Cele, S., Jackson, L., Khoury, D.S., Khan, K., Moyo-Gwete, T., Tegally, H., San, J.E., Cromer, D., Scheepers, C., Amoako, D.G., Karim, F., Bernstein, M., Lustig, G., Archary, D., Smith, M., Ganga, Y., Jule, Z., Reedoy, K., Hwa, S.-H., Giandhari, J., Blackburn, J.M., Gosnell, B.I., Abdool Karim, S.S., Hanekom, W., NGS-SA, COMMIT-KZN Team, von Gottberg, A., Bhiman, J.N., Lessells, R.J., Moosa, M.-Y.S., Davenport, M.P., de Oliveira, T., Moore, P.L. and Sigal, A. (2022) 'Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization', *Nature*, 602(7898), pp. 654–656.
11. Mallapaty, S. (2021) 'Omicron-variant border bans ignore the evidence, say scientists', *Nature*, 600(7888), p. 199.
12. Prime Minister's Office (2021) *Prime Minister sets out new measures as Omicron variant identified in UK: 27 November 2021.* GOV.UK. Available at https://www.gov.uk/government/news/prime-minister-sets-out-new-measures-as-omicron-variant-identified-in-uk-27-november-2021 (Accessed: 22 May 2022).
13. UK Health Security Agency (2021) *COVID-19 variants identified in the UK - latest updates.* GOV.UK. Available at https://www.gov.uk/government/news/covid-19-variants-identified-in-the-uk-latest-updates (Accessed: 15 April 2022).
14. Elliott, P., Bodinier, B., Eales, O., Wang, H., Haw, D., Elliott, J., Whitaker, M., Jonnerby, J., Tang, D., Walters, C.E., Atchison, C., Diggle, P.J., Page, A.J., Trotter, A.J., Ashby, D., Barclay, W., Taylor, G., Ward, H., Darzi, A., Cooke, G.S., Chadeau-Hyam, M. and Donnelly, C.A. (2022) 'Rapid increase in Omicron infections in England during December 2021: REACT-1 study', *Science*, 375(6587), pp. 1406–1411.
15. Elliott, P., Eales, O., Steyn, N., Tang, D., Bodinier, B., Wang, H., Elliott, J., Whitaker, M., Atchison, C., Diggle, P.J., Page, A.J., Trotter, A.J., Ashby, D., Barclay, W., Taylor, G., Ward, H., Darzi, A., Cooke, G.S., Donnelly, C.A. and Chadeau-Hyam, M. (2022) 'Twin peaks: The Omicron SARS-CoV-2 BA.1 and BA.2 epidemics in England', *Science*, 376(6600), p. eabq4411.
16. Prime Minister's Office (2021) *Prime Minister confirms move to Plan B in England.* GOV.UK. Available at https://www.gov.uk/government/news/prime-minister-confirms-move-to-plan-b-in-england (Accessed: 2 May 2022).
17. NHS England (2021) NHS sets out next steps to accelerate COVID-19 booster rollout. Available at https://www.england.nhs.uk/2021/12/nhs-sets-out-next-steps-to-accelerate-covid-19-booster-rollout/ (Accessed 14 March 2022).
18. Elliott, P., Eales, O., Bodinier, B., Tang, D., Wang, H., Jonnerby, J., Haw, D., Elliott, J., Whitaker, M., Walters, C.E., Atchison, C., Diggle, P.J., Page, A.J., Trotter, A.J., Ashby, D., Barclay, W., Taylor, G., Ward, H., Darzi, A., Cooke, G.S., Chadeau-Hyam, M. and Donnelly, C.A. (2022) 'Dynamics of a national Omicron SARS-CoV-2 epidemic during January 2022 in England', *Nature Communications*, 13(1), pp. 1–10.
19. Chadeau-Hyam, M., Tang, D., Eales, O., Bodinier, B., Wang, H., Jonnerby, J., Whitaker, M., Elliott, J., Haw, D., Walters, C.E., Atchison, C., Diggle, P.J., Page, A.J., Ashby, D., Barclay, W., Taylor, G., Cooke, G., Ward, H., Darzi, A., Donnelly, C.A. and Elliott, P. (2022) 'Omicron SARS-CoV-2 epidemic in England during

February 2022: A series of cross-sectional community surveys', *The Lancet regional health. Europe*, 21, p. 100462.
20. Roemer, C., Hisner, R., Frohberg, N., Sakaguchi, H., Gueli, F. and Peacock, T.P. (2023) *SARS-CoV-2 evolution, post-Omicron*. Virological.org. Available at https://virological.org/t/sars-cov-2-evolution-post-omicron/911 (Accessed 14 March 2022).
21. Chang, S., Vrabac, D., Leskovec, J. and Ugander, J. (2022) 'Estimating Geographic Spillover Effects of COVID-19 Policies From Large-Scale Mobility Networks', *arXiv [cs.CY]*. Available at: http://arxiv.org/abs/2212.06224. (Accessed: 2 July 2022).
22. COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk (2020) 'An integrated national scale SARS-CoV-2 genomic surveillance network', *Lancet Microbe*, 1(3), pp. e99–e100.
23. Shu, Y. and McCauley, J. (2017) 'GISAID: Global initiative on sharing all influenza data – from vision to reality', *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, 22(13), 30494.
24. robj (2021) *robj411/sequencing_coverage: for B.1.1.7 phylodynamic analysis.* Zenodo. Available at https://zenodo.org/record/4599180 (Accessed: 5 March 2022).
25. Kraemer, M.U.G., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D.M., Open COVID-19 Data Working Group, du Plessis, L., Faria, N.R., Li, R., Hanage, W.P., Brownstein, J.S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O.G. and Scarpino, S.V. (2020) 'The effect of human mobility and control measures on the COVID-19 epidemic in China', *Science*, 368(6490), pp. 493–497.
26. Kraemer, M.U.G., Hill, V., Ruis, C., Dellicour, S., Bajaj, S., McCrone, J.T., Baele, G., Parag, K.V., Battle, A.L., Gutierrez, B., Jackson, B., Colquhoun, R., O'Toole, Á., Klein, B., Vespignani, A., COVID-19 Genomics UK (COG-UK) Consortium, Volz, E., Faria, N.R., Aanensen, D.M., Loman, N.J., du Plessis, L., Cauchemez, S., Rambaut, A., Scarpino, S.V. and Pybus, O.G. (2021) 'Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence', *Science*, 373(6557), pp. 889–895.
27. Department of Health, Social Care (2021) *6 African countries added to red list to protect public health as UK designates new variant under investigation.* GOV.UK. Available at https://www.gov.uk/government/news/six-african-countries-added-to-red-list-to-protect-public-health-as-uk-designates-new-variant-under-investigation (Accessed: 28 March 2022).
28. du Plessis, L., McCrone, J.T., Zarebski, A.E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T.R., Faria, N.R., Jackson, B., Loman, N.J., O'Toole, Á., Nicholls, S.M., Parag, K.V., Scher, E., Vasylyeva, T.I., Volz, E.M., Watts, A., Bogoch, I.I., Khan, K., COVID-19 Genomics UK (COG-UK) Consortium, Aanensen, D.M., Kraemer, M.U.G., Rambaut, A. and Pybus, O.G. (2021) 'Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK', *Science*, 371(6530), pp. 708–712.
29. McCrone, J.T., Hill, V., Bajaj, S., Pena, R.E., Lambert, B.C., Inward, R., Bhatt, S., Volz, E., Ruis, C., Dellicour, S., Baele, G., Zarebski, A.E., Sadilek, A., Wu, N., Schneider, A., Ji, X., Raghwani, J., Jackson, B., Colquhoun, R., O'Toole, Á., Peacock, T.P., Twohig, K., Thelwall, S., Dabrera, G., Myers, R., Faria, N.R., Huber, C., Bogoch, I.I., Khan, K., du Plessis, L., Barrett, J.C., Aanensen, D.M., Barclay, W.S., Chand, M., Connor, T., Loman, N.J., Suchard, M.A., Pybus, O.G., Rambaut, A. and Kraemer, M.U.G. (2022) 'Context-specific emergence and growth of the SARS-CoV-2 Delta variant', *Nature*, 610(7930), pp. 154–160.

30. Office of National Statistics (2016) *2011 rural/urban classification*. Available at https://www.ons.gov.uk/methodology/geography/geographicalproducts/ruralurbanclassifications/2011ruralurbanclassification (Accessed: 4 April 2022).
31. Charu, V., Zeger, S., Gog, J., Bjørnstad, O.N., Kissler, S., Simonsen, L., Grenfell, B.T. and Viboud, C. (2017) 'Human mobility and the spatial transmission of influenza in the United States', *PLoS computational biology*, 13(2), p. e1005382.
32. Kraemer, M.U.G., Faria, N.R., Reiner, R.C., Jr, Golding, N., Nikolay, B., Stasse, S., Johansson, M.A., Salje, H., Faye, O., Wint, G.R.W., Niedrig, M., Shearer, F.M., Hill, S.C., Thompson, R.N., Bisanzio, D., Taveira, N., Nax, H.H., Pradelski, B.S.R., Nsoesie, E.O., Murphy, N.R., Bogoch, I.I., Khan, K., Brownstein, J.S., Tatem, A.J., de Oliveira, T., Smith, D.L., Sall, A.A., Pybus, O.G., Hay, S.I. and Cauchemez, S. (2017) 'Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015-16: a modelling study', *The Lancet. Infectious diseases*, 17(3), pp. 330–338.
33. Kraemer, M.U.G., Golding, N., Bisanzio, D., Bhatt, S., Pigott, D.M., Ray, S.E., Brady, O.J., Brownstein, J.S., Faria, N.R., Cummings, D.A.T., Pybus, O.G., Smith, D.L., Tatem, A.J., Hay, S.I. and Reiner, R.C., Jr (2019) 'Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings', *Scientific reports*, 9(1), p. 5151.
34. Finkenstädt, B. and Grenfell, B. (1998) 'Empirical determinants of measles metapopulation dynamics in England and Wales', *Proceedings. Biological sciences / The Royal Society*, 265(1392), pp. 211–220.
35. Nouvellet, P., Bhatia, S., Cori, A., Ainslie, K.E.C., Baguelin, M., Bhatt, S., Boonyasiri, A., Brazeau, N.F., Cattarino, L., Cooper, L.V., Coupland, H., Cucunuba, Z.M., Cuomo-Dannenburg, G., Dighe, A., Djaafara, B.A., Dorigatti, I., Eales, O.D., van Elsland, S.L., Nascimento, F.F., FitzJohn, R.G., Gaythorpe, K.A.M., Geidelberg, L., Green, W.D., Hamlet, A., Hauck, K., Hinsley, W., Imai, N., Jeffrey, B., Knock, E., Laydon, D.J., Lees, J.A., Mangal, T., Mellan, T.A., Nedjati-Gilani, G., Parag, K.V., Pons-Salort, M., Ragonnet-Cronin, M., Riley, S., Unwin, H.J.T., Verity, R., Vollmer, M.A.C., Volz, E., Walker, P.G.T., Walters, C.E., Wang, H., Watson, O.J., Whittaker, C., Whittles, L.K., Xi, X., Ferguson, N.M. and Donnelly, C.A. (2021) 'Reduction in mobility and COVID-19 transmission', *Nature communications*, 12(1), p. 1090.
36. Brockmann, D. and Helbing, D. (2013) 'The Hidden Geometry of Complex, Network-Driven Contagion Phenomena', *Science*, 342(6164), pp. 1337-1342.
37. Murall, C.L., Fournier, E., Galvez, J.H., N'Guessan, A., Reiling, S.J., Quirion, P.-O., Naderi, S., Roy, A.-M., Chen, S.-H., Stretenowich, P., Bourgey, M., Bujold, D., Gregoire, R., Lepage, P., St-Cyr, J., Willet, P., Dion, R., Charest, H., Lathrop, M., Roger, M., Bourque, G., Ragoussis, J., Shapiro, B.J. and Moreira, S. (2021) 'A small number of early introductions seeded widespread transmission of SARS-CoV-2 in Québec, Canada', *Genome medicine*, 13(1), p. 169.
38. Tegally, H., Wilkinson, E., Tsui, J.L.-H., Moir, M., Martin, D., Brito, A.F., Giovanetti, M., Khan, K., Huber, C., Bogoch, I.I., San, J.E., Poongavanan, J., Xavier, J.S., Candido, D. da S., Romero, F., Baxter, C., Pybus, O.G., Lessells, R.J., Faria, N.R., Kraemer, M.U.G. and de Oliveira, T. (2023) 'Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern', *Cell*, 186(15), pp. 3277–3290.e16.
39. Brett, T.S. and Rohani, P. (2022) 'Containing novel SARS-CoV-2 variants at source is possible with high-intensity sequencing', *PNAS nexus*, 1(4), p. gac159.

40. Ferguson, N.M., Cummings, D.A.T., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S. and Burke, D.S. (2005) 'Strategies for containing an emerging influenza pandemic in Southeast Asia', *Nature*, 437(7056), pp. 209–214.
41. Fraser, C., Riley, S., Anderson, R.M. and Ferguson, N.M. (2004) 'Factors that make an infectious disease outbreak controllable', *Proceedings of the National Academy of Sciences of the United States of America*, 101(16), pp. 6146–6151.
42. Hill, V., Du Plessis, L., Peacock, T.P., Aggarwal, D., Colquhoun, R., Carabelli, A.M., Ellaby, N., Gallagher, E., Groves, N., Jackson, B., McCrone, J.T., O'Toole, Á., Price, A., Sanderson, T., Scher, E., Southgate, J., Volz, E., Barclay, W.S., Barrett, J.C., Chand, M., Connor, T., Goodfellow, I., Gupta, R.K., Harrison, E.M., Loman, N., Myers, R., Robertson, D.L., Pybus, O.G. and Rambaut, A. (2022) 'The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK', *Virus evolution*, 8(2), p. veac080.
43. Willett, B.J., Grove, J., MacLean, O.A., Wilkie, C., De Lorenzo, G., Furnon, W., Cantoni, D., Scott, S., Logan, N., Ashraf, S., Manali, M., Szemiel, A., Cowton, V., Vink, E., Harvey, W.T., Davis, C., Asamaphan, P., Smollett, K., Tong, L., Orton, R., Hughes, J., Holland, P., Silva, V., Pascall, D.J., Puxty, K., da Silva Filipe, A., Yebra, G., Shaaban, S., Holden, M.T.G., Pinto, R.M., Gunson, R., Templeton, K., Murcia, P.R., Patel, A.H., Klenerman, P., Dunachie, S., PITCH Consortium, COVID-19 Genomics UK (COG-UK) Consortium, Haughney, J., Robertson, D.L., Palmarini, M., Ray, S. and Thomson, E.C. (2022) 'SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway', *Nature microbiology*, 7(8), pp. 1161–1179.
44. Stein, C., Nassereldine, H., Sorensen, R.J.D., Amlag, J.O., Bisignano, C., Byrne, S., Castro, E., Coberly, K., Collins, J.K., Dalos, J., Daoud, F., Deen, A., Gakidou, E., Giles, J.R., Hulland, E.N., Huntley, B.M., Kinzel, K.E., Lozano, R., Mokdad, A.H., Pham, T., Pigott, D.M., Reiner, R.C., Jr, Vos, T., Hay, S.I., Murray, C.J.L. and Lim, S.S. (2023) 'Past SARS-CoV-2 infection protection against re-infection: a systematic review and meta-analysis', *The Lancet*, 401(10379), pp. 833–842.
45. Brito, A.F., Semenova, E., Dudas, G., Hassler, G.W., Kalinich, C.C., Kraemer, M.U.G., Ho, J., Tegally, H., Githinji, G., Agoti, C.N., Matkin, L.E., Whittaker, C., Bulgarian SARS-CoV-2 sequencing group, Communicable Diseases Genomics Network (Australia and New Zealand), COVID-19 Impact Project, Danish Covid-19 Genome Consortium, Fiocruz COVID-19 Genomic Surveillance Network, GISAID core curation team, Network for Genomic Surveillance in South Africa (NGS-SA), Swiss SARS-CoV-2 Sequencing Consortium, Howden, B.P., Sintchenko, V., Zuckerman, N.S., Mor, O., Blankenship, H.M., de Oliveira, T., Lin, R.T.P., Siqueira, M.M., Resende, P.C., Vasconcelos, A.T.R., Spilki, F.R., Aguiar, R.S., Alexiev, I., Ivanov, I.N., Philipova, I., Carrington, C.V.F., Sahadeo, N.S.D., Branda, B., Gurry, C., Maurer-Stroh, S., Naidoo, D., von Eije, K.J., Perkins, M.D., van Kerkhove, M., Hill, S.C., Sabino, E.C., Pybus, O.G., Dye, C., Bhatt, S., Flaxman, S., Suchard, M.A., Grubaugh, N.D., Baele, G. and Faria, N.R. (2022) 'Global disparities in SARS-CoV-2 genomic surveillance', *Nature communications*, 13(1), p. 7003.
46. Wardle, J., Bhatia, S., Kraemer, M.U.G., Nouvellet, P. and Cori, A. (2023) 'Gaps in mobility data and implications for modelling epidemic spread: A scoping review and simulation study', *Epidemics*, 42, p. 100666.
47. Citron, D.T., Guerra, C.A., Dolgert, A.J., Wu, S.L., Henry, J.M., Sánchez C, H.M. and Smith, D.L. (2021) 'Comparing metapopulation dynamics of infectious diseases under different models of human movement', *Proceedings of the National Academy of Sciences of the United States of America*, 118(18), p. e2007488118

48. World Health Organisation (2023) *Statement on the fifteenth meeting of the International Health Regulations (2005) Emergency Committee regarding the coronavirus disease (COVID-19) pandemic*. Available at https://www.who.int/news/item/05-05-2023-statement-on-the-fifteenth-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-coronavirus-disease-(covid-19)-pandemic (Accessed: 20 April 2022).
49. Moorthy, V., Morgan, O., Ihekweazu, C. and Swaminathan, S. (2022) 'WHO principles speed up ethical sharing of pathogen genomic data', *Nature*, 611(7936), p. 449.
50. Hill, V., Ruis, C., Bajaj, S., Pybus, O.G. and Kraemer, M.U.G. (2021) 'Progress and challenges in virus genomic epidemiology', *Trends in parasitology*, 37(12), pp. 1038–1049.
51. joetsui (2023) *joetsui1994/omicron-BA.1-invasion-dynamics: 12 may, 2023 - v1.0*. Zenodo. Available at https://zenodo.org/record/7928698 (Accessed: 1 May 2025).
52. National Audit Office (2022) *Managing cross-border travel during the COVID-19 pandemic*. National Audit Office, London, England. Available at https://www.nao.org.uk/reports/managing-cross-border-travel-during-the-covid-19-pandemic/?nab=2 (Accessed: 28 April 2025).
53. Kraemer, M.U.G., Sadilek, A., Zhang, Q., Marchal, N.A., Tuli, G., Cohn, E.L., Hswen, Y., Perkins, T.A., Smith, D.L., Reiner, R.C., Jr and Brownstein, J.S. (2020) 'Mapping global variation in human mobility', *Nat Hum Behav*, 4(8), pp. 800–810.
54. Lemey, P., Ruktanonchai, N., Hong, S.L., Colizza, V., Poletto, C., Van den Broeck, F., Gill, M.S., Ji, X., Levasseur, A., Oude Munnink, B.B., Koopmans, M., Sadilek, A., Lai, S., Tatem, A.J., Baele, G., Suchard, M.A. and Dellicour, S. (2021) 'Untangling introductions and persistence in COVID-19 resurgence in Europe', *Nature*, 595(7869), pp. 713–717.
55. Wilson, R.J., Zhang, C.Y., Lam, W., Desfontaines, D., Simmons-Marengo, D. and Gipson, B. (2019) 'Differentially Private SQL with Bounded User Contribution', *arXiv [cs.CR]*. Available at: http://arxiv.org/abs/1909.01917 (Accessed: 29 April 2025).
56. UK Health Security Agency (2022) *Confirmatory PCR tests to be temporarily suspended for positive lateral flow test results.* GOV.UK. Available at https://www.gov.uk/government/news/confirmatory-pcr-tests-to-be-temporarily-suspended-for-positive-lateral-flow-test-results (Accessed: 2 September 2022).
57. Rambaut, A., Suchard, M., Ji, X., Baele, G., Drummond, A., Bastide, P., gabehassler, Lemey, P., mtolkoff, Zhang, Z., Carvalho, L.M.F., jessiewu, Bedford, T., mandevgill, Walter, McCrone, J.T., Bielejec, F., athos, Karcher, M., Ayres, D.L., Cheung, C., Hall, M., Fourment, M., gcybis, Vallard, T., Bilge, A., Chang, J., Hill, V., Nahata, K. & beast-dev (2021) *beast-mcmc: BEAST v1.10.5 pre-release of ThorneyTreeLikelihood v0.1.2*. Zenodo. Available at: https://zenodo.org/record/5361043 (Accessed: 14 April 2022).
58. Volz, E., Mishra, S., Chand, M., Barrett, J.C., Johnson, R., Geidelberg, L., Hinsley, W.R., Laydon, D.J., Dabrera, G., O'Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C.V., Boyd, O., Loman, N.J., McCrone, J.T., Gonçalves, S., Jorgensen, D., Myers, R., Hill, V., Jackson, D.K., Gaythorpe, K., Groves, N., Sillitoe, J., Kwiatkowski, D.P., Flaxman, S., Ratmann, O., Bhatt, S., Hopkins, S., Gandy, A., Rambaut, A. and Ferguson, N.M. (2021) 'Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England', *Nature*, pp. 1–17.
59. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A. and Lanfear, R. (2020) 'IQ-TREE 2: New Models and Efficient

Methods for Phylogenetic Inference in the Genomic Era', *Molecular biology and evolution*, 37(5), pp. 1530–1534.
60. Sagulenko, P., Puller, V. and Neher, R.A. (2018) 'TreeTime: Maximum-likelihood phylodynamic analysis', *Virus evolution*, 4(1), p. vex042.
61. Turakhia, Y., Thornlow, B., Hinrichs, A.S., De Maio, N., Gozashti, L., Lanfear, R., Haussler, D. and Corbett-Detig, R. (2021) 'Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic', *Nature genetics*, 53(6), pp. 809–816.
62. Ye, C., Thornlow, B., Hinrichs, A., Kramer, A., Mirchandani, C., Torvi, D., Lanfear, R., Corbett-Detig, R. and Turakhia, Y. (2022) 'matOptimize: a parallel tree optimization method enables online phylogenetics for SARS-CoV-2', *Bioinformatics* , 38(15), pp. 3734–3740.
63. Sanderson, T. (2022) 'Chronumental: time tree estimation from very large phylogenies', *bioRxiv*. Available at: https://doi.org/10.1101/2021.10.27.465994 (Accessed 5 April 2022).
64. Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L. and Pybus, O.G. (2020) 'A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology', *Nature microbiology*, 5(11), pp. 1403–1407.
65. Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J. and Rambaut, A. (2018) 'Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10', *Virus evolution*, 4(1), p. vey016.
66. Gill, M.S., Lemey, P., Faria, N.R., Rambaut, A., Shapiro, B. and Suchard, M.A. (2013) 'Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci', *Molecular biology and evolution*, 30(3), pp. 713–724.
67. Rambaut, A., Drummond, A.J., Xie, D., Baele, G. and Suchard, M.A. (2018) 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7', *Systematic biology*, 67(5), pp. 901–904.
68. Department of Transport (2021) *Travel to England from another country – COVID-19 rules.* GOV.UK. Available at https://www.gov.uk/guidance/travel-to-england-from-another-country-during-coronavirus-covid-19 (Accessed: 18 May 2022).
69. Lemey, P., Rambaut, A., Welch, J.J. and Suchard, M.A. (2010) 'Phylogeography takes a relaxed random walk in continuous space and time', *Molecular biology and evolution*, 27(8), pp. 1877–1885.
70. Ji, X., Zhang, Z., Holbrook, A., Nishimura, A., Baele, G., Rambaut, A., Lemey, P. and Suchard, M.A. (2020) 'Gradients Do Grow on Trees: A Linear-Time O(N)-Dimensional Gradient for Statistical Phylogenetics', *Molecular biology and evolution*, 37(10), pp. 3047–3060.
71. Baele, G., Gill, M.S., Lemey, P. and Suchard, M.A. (2020) 'Hamiltonian Monte Carlo sampling to estimate past population dynamics using the skygrid coalescent model in a Bayesian phylogenetics framework', *Wellcome open research*, 5, p. 53.
72. Joint Committee on Vaccination and Immunisation (JCVI) (2021) *JCVI statement regarding a COVID-19 booster vaccine programme for winter 2021 to 2022.* GOV.UK. Available at: https://www.gov.uk/government/publications/jcvi-statement-september-2021-covid-19-booster-vaccine-programme-for-winter-2021-to-2022/jcvi-statement-regarding-a-covid-19-booster-vaccine-programme-for-winter-2021-to-2022 (Accessed: 1 May 2022).
73. Department of Health and Social Care (2021) *People urged to get booster jabs to keep your family protected this Christmas*. GOV.UK. Available at https://www.gov.uk/government/news/people-urged-to-get-booster-jabs-to-keep-

your-family-protected-this-christmas (Accessed: 27 April 2022).
74. Department of Health and Social Care (2021) *All adults to be offered COVID-19 boosters by end of January*. GOV.UK. Available at https://www.gov.uk/government/news/all-adults-to-be-offered-covid-19-boosters-by-end-of-january (Accessed: 27 April 2022).
75. Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32.
76. Fisher, A., Rudin, C. and Dominici, F. (2019) 'All Models are Wrong, but are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously', *Journal of Machine Learning Research, 20(177), pp. 1-81*
77. Mentch, L. and Hooker, G. (2014) 'Quantifying uncertainty in random forests via confidence intervals and hypothesis tests', *arXiv [stat.ML]*. Available at: https://jmlr.org/papers/volume17/14-168/14-168.pdf (Accessed: 14 April 2023).

Appendix A:
Supplementary materials for Chapter 2

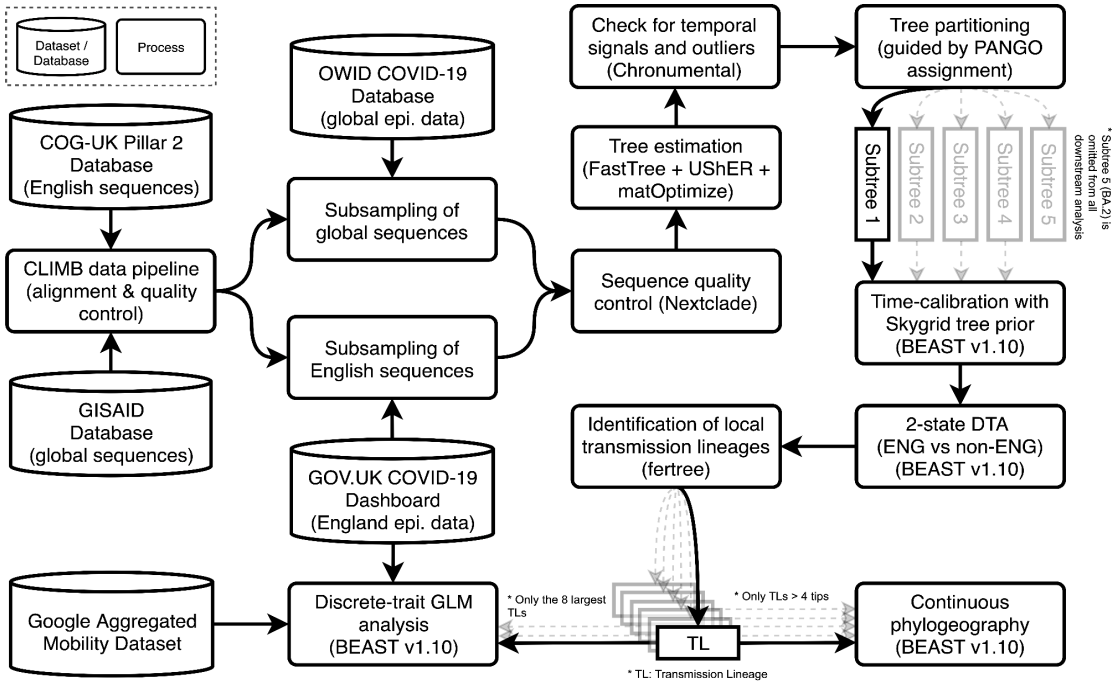


Fig. A.1: Outline of phylodynamic analysis pipeline. A high-level overview of the various processes and phylodynamic analyses performed, as well as any relevant programs and packages for each step. Note that each subtree (except for the subtree containing only Omicron BA.2 sequences which we have omitted from further downstream analysis) from the tree-partitioning procedure is passed onto further downstream analyses independently.

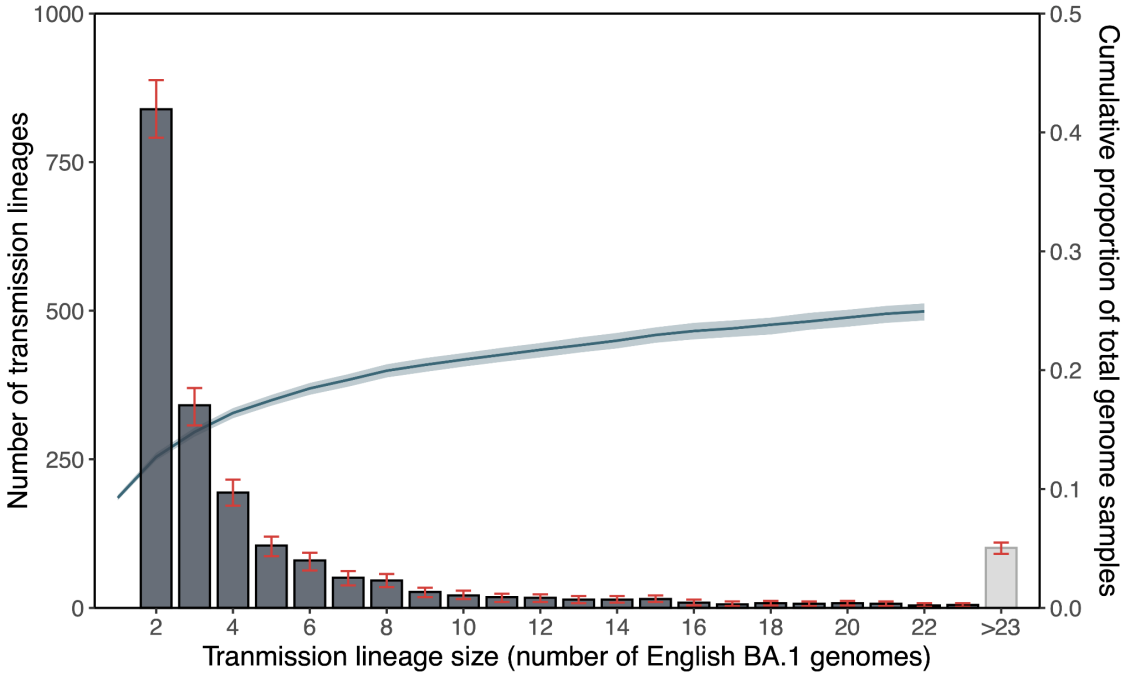


Fig. A.2. Distribution of local transmission lineage sizes from phylodynamic analysis. Grey bars show the number of transmission lineages of different sizes; red error bars denote the 95% HPDs across the posterior tree distribution. Blue solid line represents the cumulative proportion of English Omicron BA.1 genomes in our dataset accounted for by transmission lineages up to a certain size; shading denotes the 95% HPD across the posterior tree distribution.

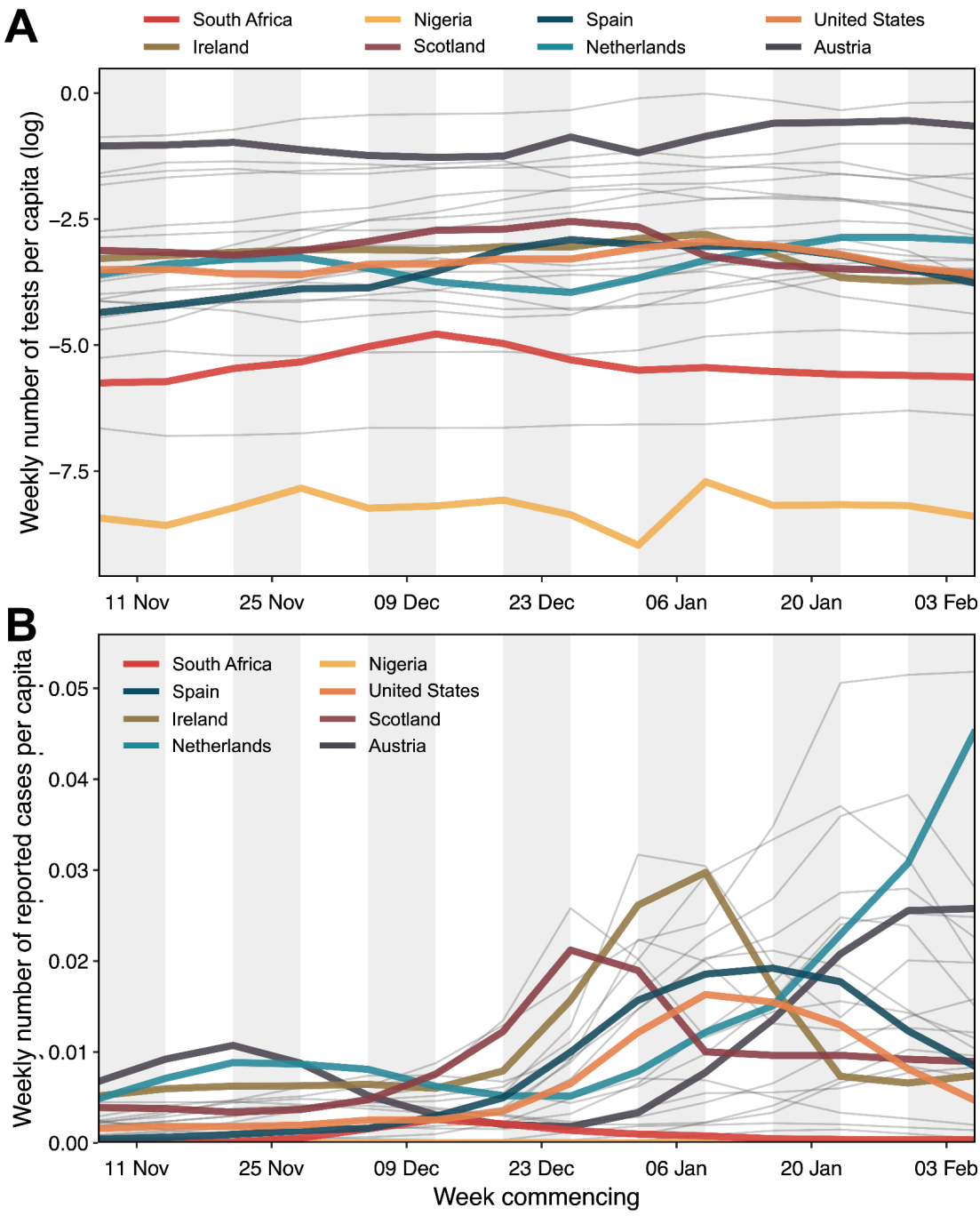


Fig. A.3: Variations in case reporting rates between countries. (A) Weekly number of tests performed per capita (log-transformed) and (B) weekly number of reported cases per capita for 27 countries (including Scotland and Northern Ireland) with the highest air passenger volumes arriving in England between November 2021 and January 2022 (collectively accounting for ~80% of total air passenger volume in this period). Thick solid lines represent a subset of eight selected countries with notable contribution to the overall intensity of Omicron BA.1 importation into England at different points during the study period; thin grey lines represent all other countries.

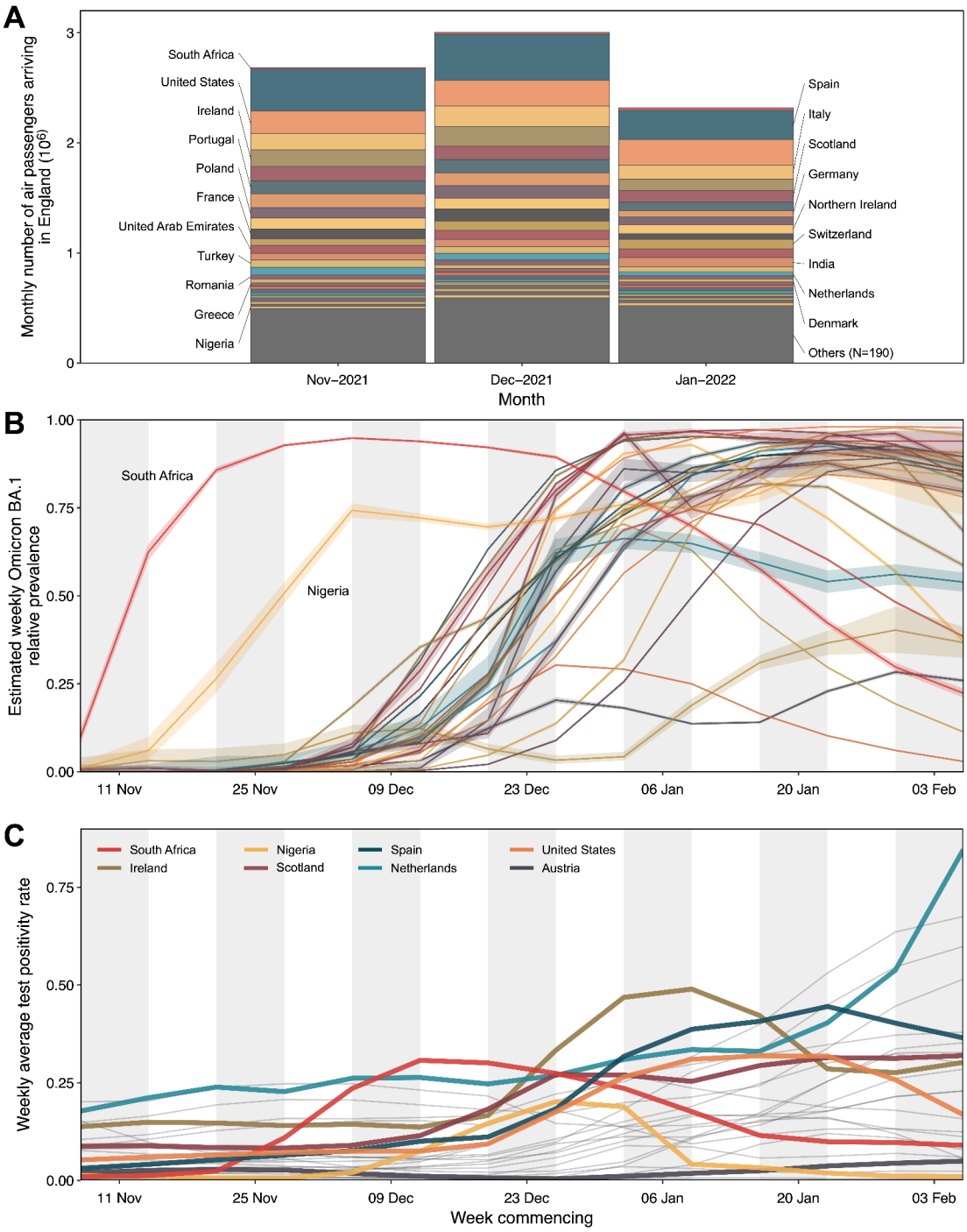


Fig. A.4: Components of BA.1 Estimated Importation Intensity (EII). (A) Monthly number of air passengers arriving in England from all countries (N=217) between November 2021 and January 2022. Area of each coloured block indicates the number of air passengers arriving from a given country (out of the 27 countries for which EIIs are calculated) during a given month; grey blocks at the bottom represent air traffic volume from all other countries (N=190). (B) Estimated weekly relative prevalence of Omicron BA.1 in the 27 selected countries during the study period; shaded region represents the 95% CI. (C) Weekly average test positivity rate in the 27 selected countries during the study period. Thick solid lines represent a subset of eight selected countries with notable

contribution to the overall intensity of Omicron BA.1 importation into England at different points during the study period; thin grey lines represent all other countries.

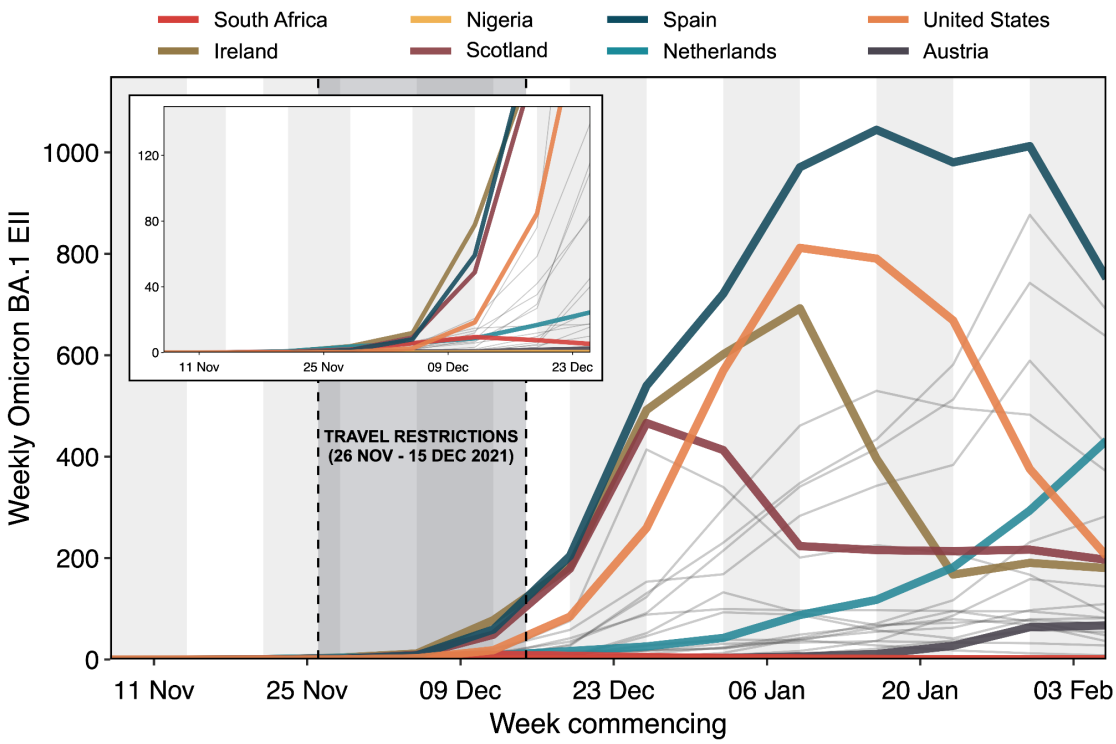


Fig. A.5: Estimated Importation Intensity (EII) of BA.1 from selected potential exporters, using case incidence per capita as proxy for underlying prevalence. Estimated weekly number of Omicron BA.1 cases arriving in England from 27 countries (including Scotland and Northern Ireland) with the highest air passenger volumes arriving in England between November 2021 and January 2022 (collectively accounting for ~80% of total air passenger volume in this period), using weekly number of reported cases as a proxy for trends in the underlying prevalence. Thick solid lines represent EIIs from eight selected countries with notable contribution to the overall intensity of Omicron BA.1 importation into England at different points during the study period; thin grey lines represent all other countries. Inset shows a magnified view of early trends. Grey shaded region represents the period (26 November to 15 December 2021) when travel restrictions on international arrivals from multiple southern African countries were implemented.

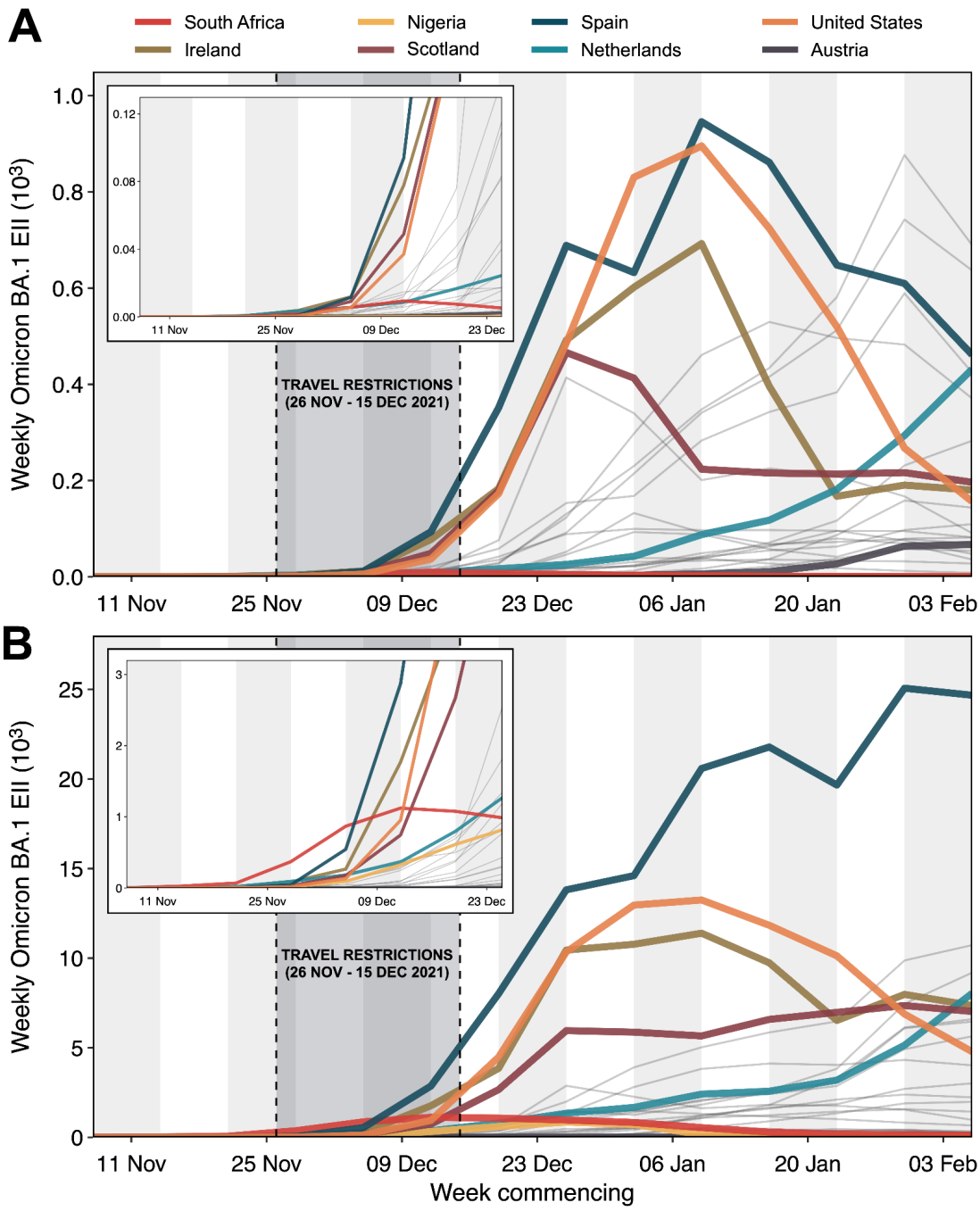


Fig. A.6: Estimated Importation Intensity (EII) of BA.1 from selected potential exporters, with within-country disaggregation for Spain and the United States. Estimated weekly number of Omicron BA.1 cases arriving in England from 27 countries (including Scotland and Northern Ireland) with the highest air passenger volumes arriving in England between November 2021 and January 2022 (collectively accounting for ~80% of total air passenger volume in this period), using (A) weekly number of reported cases and (B) weekly average test positivity rate as a proxy for trends in the underlying prevalence. EII for Spain and the United States were aggregated from multiple EIIs calculated at the autonomous community- and state-level, respectively. Thick solid lines represent EIIs from eight selected countries with notable contribution to the overall

intensity of Omicron BA.1 importation into England at different points during the study period; thin grey lines represent all other countries. Insets show a magnified view of early trends. Grey shaded region represents the period (26 November to 15 December 2021) when travel restrictions on international arrivals from multiple southern African countries were implemented.

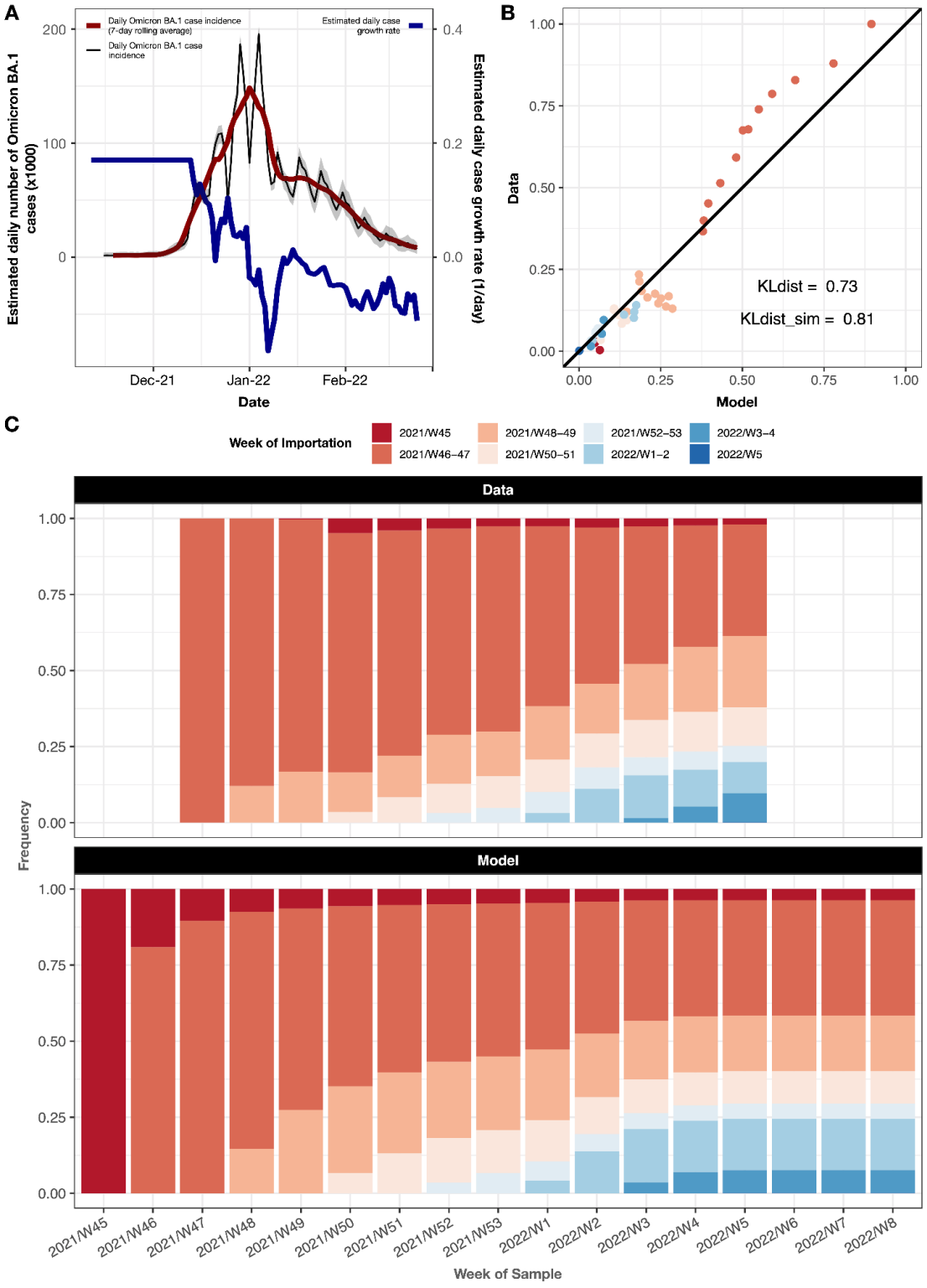


Fig. A.7. Comparison of transmission lineage size distribution from phylodynamic analysis versus simulated results from a branching process model. (A) Black and red solid lines represent the estimated daily and 7-day rolling average daily number of Omicron BA.1 cases in England. Grey shaded region represents the 95% CI associated with the estimated daily number of Omicron BA.1 cases. Blue solid line represents the estimated daily growth rate, with the initial values imputed using an estimate of the

growth rate on 13 December 2021. (B and C) Weekly proportion of local Omicron BA.1 infections resulting from importations at different times throughout the epidemic, with comparison between empirical observations from the phylodynamic analysis (C, top) and predictions from the branching process model (C, bottom).

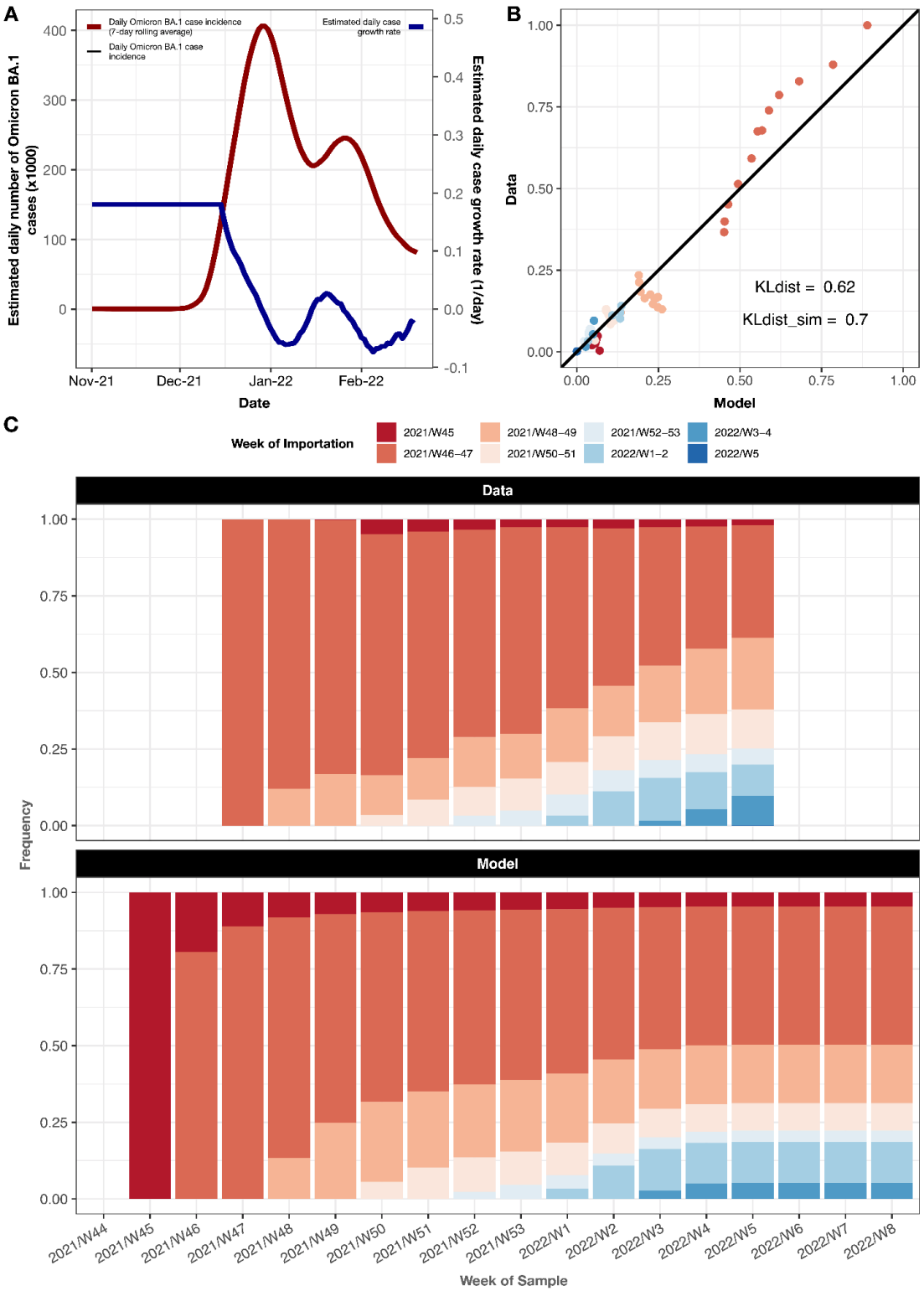


Fig. A.8. Comparison of transmission lineage size distribution from phylodynamic analysis versus simulated results from a branching process model (sensitivity analysis using ONS case incidence estimates). (A) Black and red solid lines represent the estimated daily and 7-day rolling average daily number of Omicron BA.1 cases in England, from the UK Office of National Statistics (ONS). Blue solid line represents the estimated daily growth rate, with the initial values imputed using an estimate of the growth rate on 13 December 2021. (B and C) Weekly proportion of local Omicron BA.1

infections resulting from importations at different times throughout the epidemic, with comparison between empirical observations from the phylodynamic analysis (C, top) and predictions from the branching process model (C, bottom).

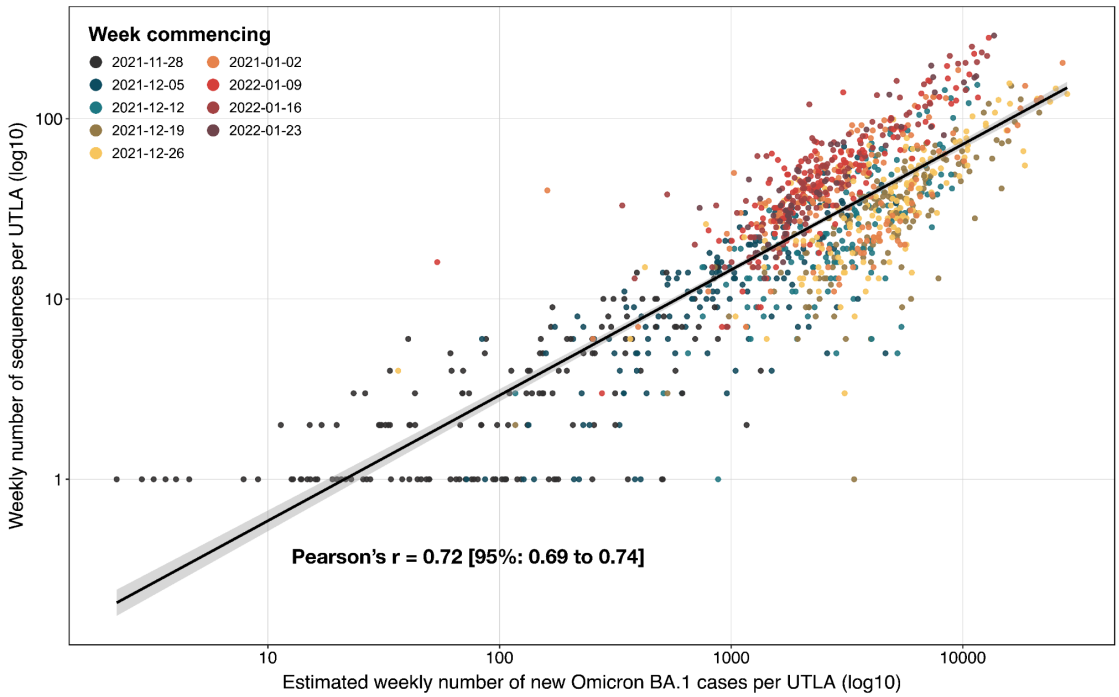


Fig. A.9. Correlation between estimated number of BA.1 cases and number of Omicron BA.1 genomes sampled across UTLAs in England. Circles are coloured by week commencing date. Solid black line represents the line of best-fit; shaded region represents the 95% CI. We note the clustering of circles corresponding to the same week along the line of best-fit, indicating small changes in sequencing coverage across time but not across UTLAs. Only data between week starting on 28 November 2021 and week starting on 23 January 2022 are presented (English genomes were subsampled in proportion to weekly reported case incidence during this period).

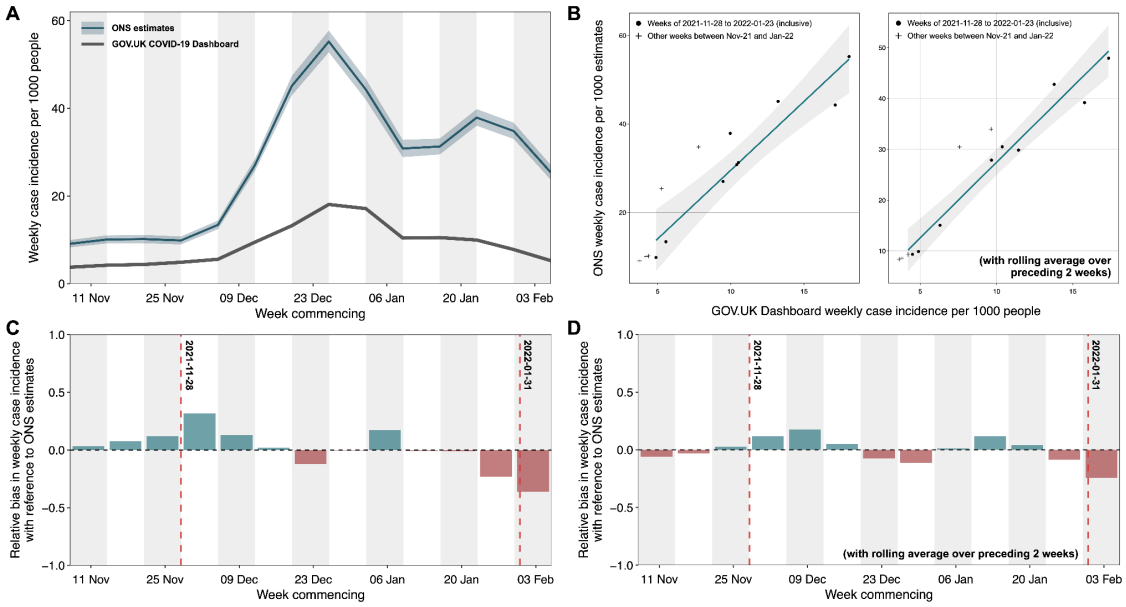


Fig. A.10: Comparison of case incidence from the GOV.UK COVID19 Dashboard against estimates from the UK Office of National Statistics. (A) Weekly number of positive COVID-19 cases in England as reported by the GOV.UK COVID19 Dashboard (https://coronavirus.data.gov.uk/) (solid black line); weekly number of COVID-19 cases in England as estimated from positivity rates by the UK Office of National Statistics (ONS) (solid blue line; shading denotes the associated 95% CI). Vertical red dashed lines indicate the start date and end date of the period during which English genomes were sampled in proportional to the weekly number of reported. (B) Weekly number of COVID-19 cases per 1000 people as estimated by ONS versus that from the GOV.UK COVID19 Dashboard, with (right) and without (left) smoothing over the preceding two weeks for each given date. Blue lines show the least-squares fit and the shading denotes the associated 95% CI. Black dots represent weekly case incidences between 28 November 2021 and 29 January 2022; black crosses represent all other weekly case incidences during the study period. (C, D) Residuals from a linear regression between the weekly number of COVID-19 cases per capita as estimated by ONS versus that from the GOV.UK COVID19 Dashboard (between week starting on 31 October 2021 and week starting on 23 January 2022), with (D) and without (C) smoothing over the preceding two weeks for each given date. Boxes are coloured red (negative) or blue (positive) according to the sign of the relative bias.

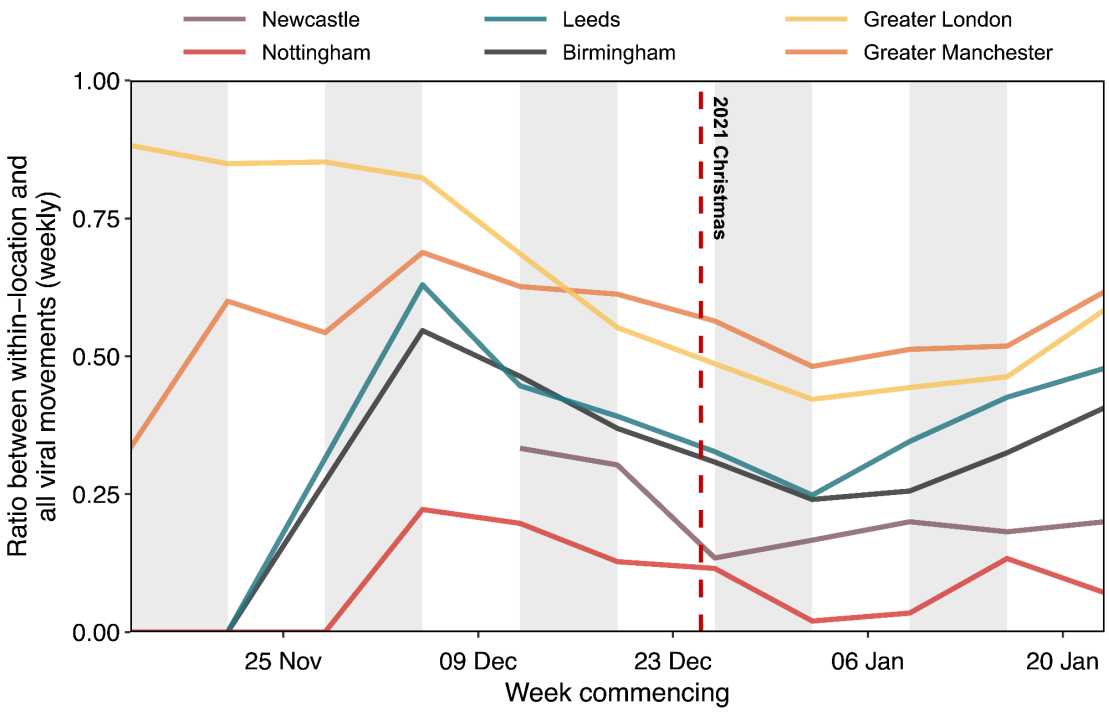


Fig. A.11. Within-location versus all viral lineage movements for major cities in England. Each solid line represents the ratio between the frequency of within-location and all viral lineage movements per week, as inferred from continuous phylogeography for 6 major cities in England. For Greater Manchester and Greater London, viral lineages associated with multiple Lower Tier Local Authorities were aggregated in the calculation of these ratios. The timing of each viral lineage movement was assumed to be half-way between the inferred time of the nodes corresponding to the origin and destination.

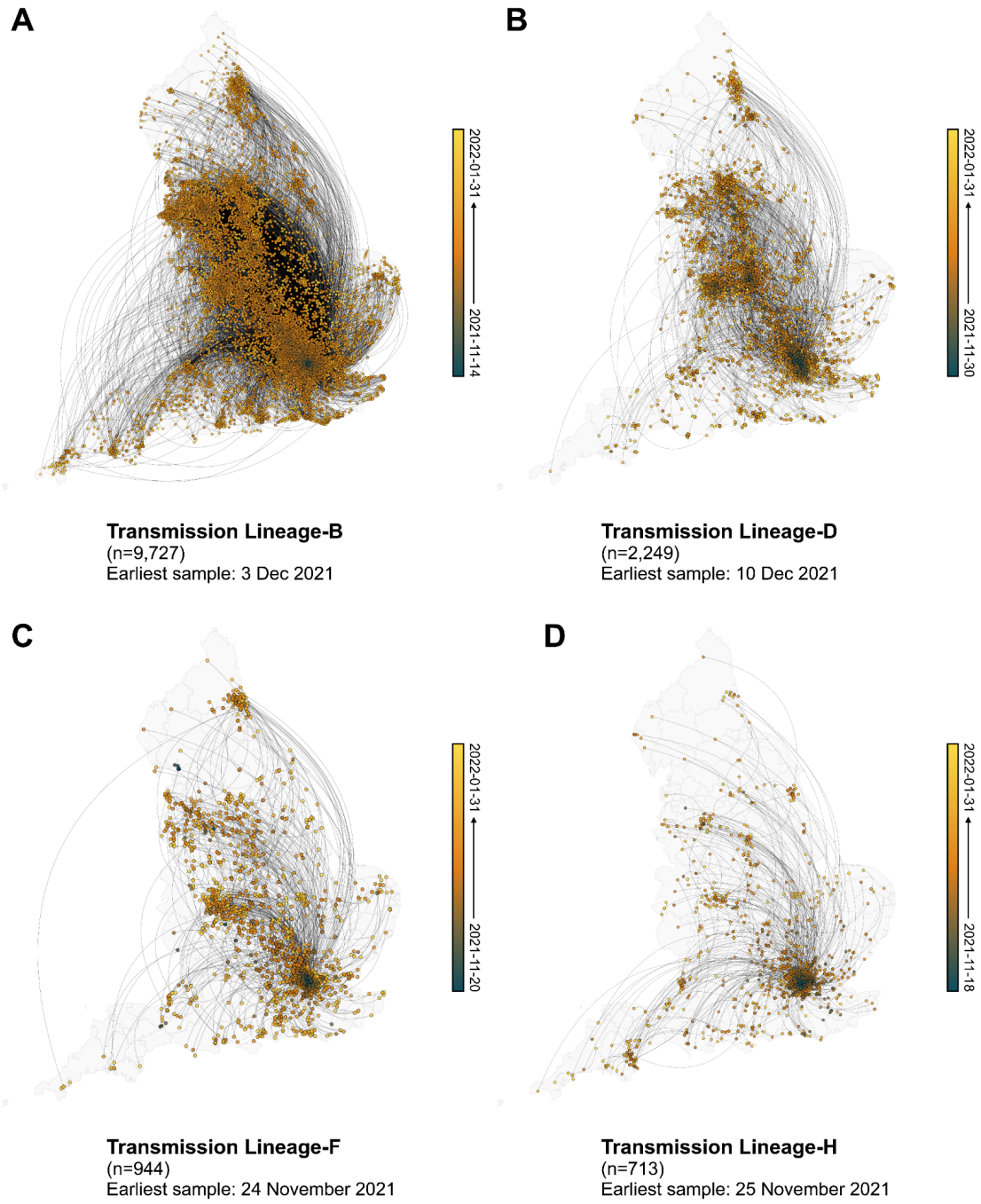


Fig. A.12. Spatiotemporal dynamics of BA.1 transmission lineages in England (Transmission Lineage-B, D, F and H). Maps showing viral lineage movements inferred from continuous phylogeography for (A) Transmission Lineage-B, (B) Transmission Lineage-D, (C) Transmission Lineage-F, and (D) Transmission Lineage-H. Nodes are coloured according to inferred date of occurrence and the direction of viral lineage movement is indicated by edge curvature (anti-clockwise).

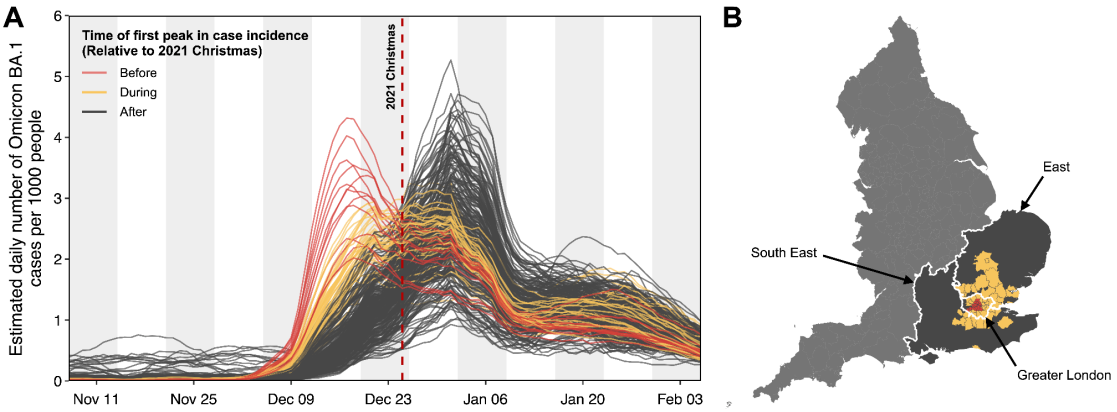


Fig. A.13. Spatial variations in timing of first peak of BA.1 case incidence across England. Estimated daily number of Omicron BA.1 cases per 1000 people at the Lower Tier Local Authority (LTLA) level (7-day rolling average), coloured according to the timing of their first peak relative to Christmas 2021 (specifically, whether the interval during which the daily number of Omicron BA.1 cases exceed 85% of the peak incidence lies entirely before (red), after(dark grey), or encloses (yellow) 25 December 2021 (Christmas). (B) Map showing the spatial distribution of the timing of the first peak in Omicron BA.1 case incidence at the LTLA level, following the same colour scheme as in (A).

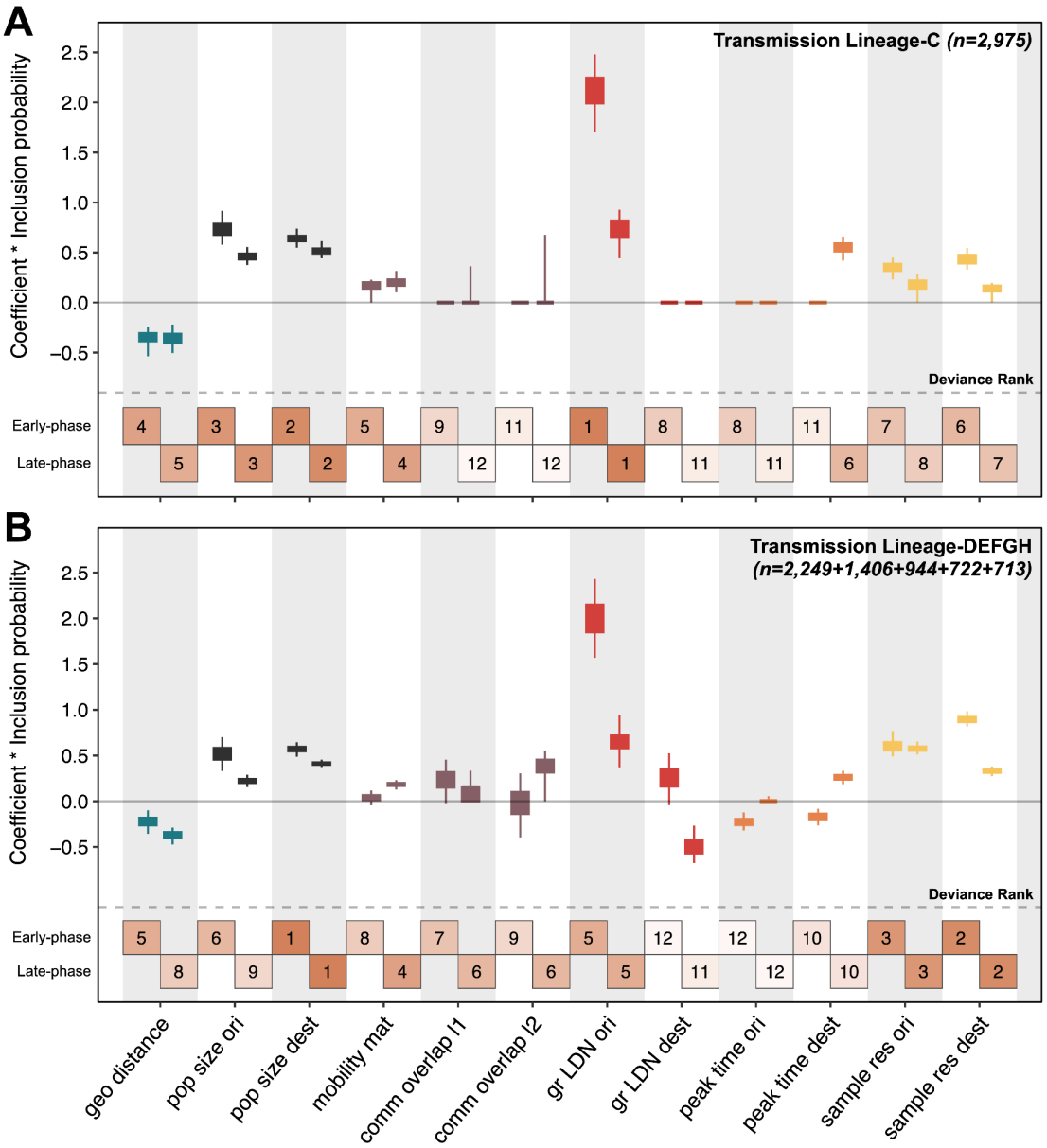


Fig. A.14. Predictors of BA.1 viral lineage movements in England in the time-inhomogeneous discrete-phylogeography with GLM model. For each predictor, the box and whiskers show the posterior distribution of the product of the log predictor coefficient and the predictor inclusion probability; the left-hand and right-hand values show the estimates for before and after 26 December, respectively. Top panel (A) shows estimates for Transmission Lineage-C and bottom panel (B) shows those for Transmission Lineages D, E, F, G, and H analysed in a joint model. Posterior distributions are coloured according to predictor type: geographic distances (geo distance, dark blue), population sizes at origin and destination (pop size ori & pop size dest, black), aggregated mobility (mobility mat, purple), mobility-based community membership level 1 and level 2 (comm overlap l1 & l2, purple), Greater London origin and destination (gr LDN ori & gr LDN dest, red), time of peak incidence at the origin and destination (peak time ori & peak time dest, orange) and the residual of a regression of sample size against case count regression at either the origin and destination (sample res ori & sample res dest, yellow).

Boxes at the bottom of each panel are numbered and shaded to show the ranking of predictors based on their deviance measure (more details in Section 2.6.15), with 1 indicating the largest deviance (most important predictor) and 12 indicating the smallest (least important predictor).

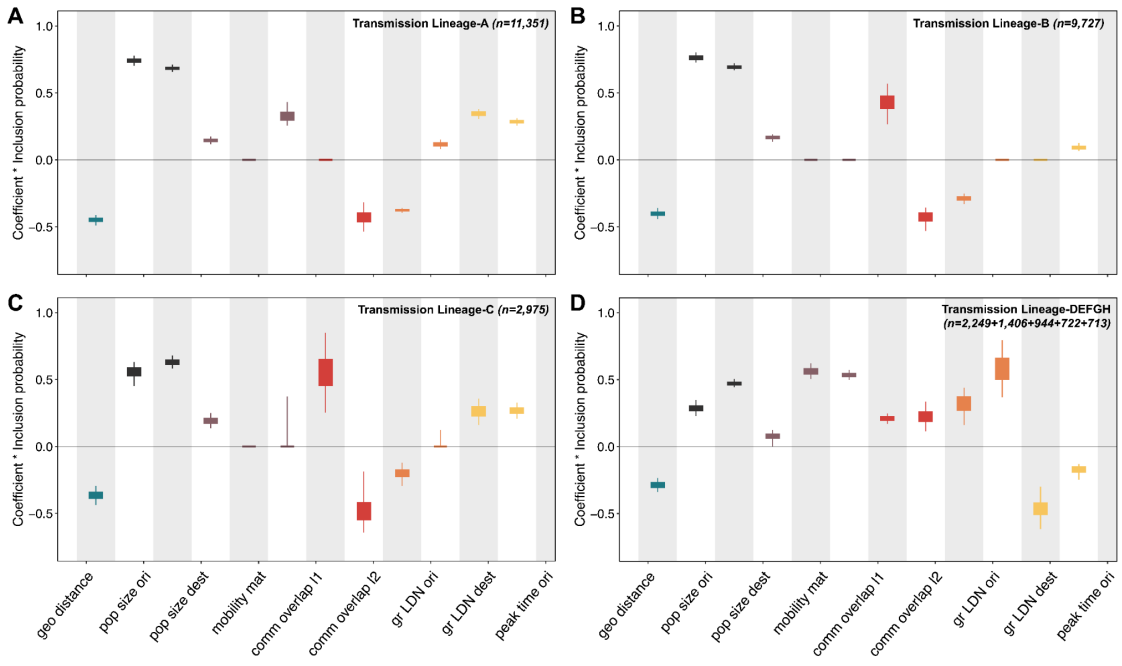


Fig. A.15. Predictors of BA.1 viral lineage movements in England in the time-homogeneous discrete-phylogeography with GLM model. Each panel corresponds to an independent analysis for Transmission Lineage-A (A), Transmission Lineage-B (B), Transmission Lineage-C (C), and Transmission Lineages D, E, F, G and H together in a joint model (D). For each predictor within a panel, the box and whiskers show the posterior distributions of the product of the log predictor coefficient and the predictor inclusion probability. Posterior distributions are coloured according to predictor type: geographic distances (geo distance, dark blue), population sizes at origin and destination (pop size ori & pop size dest, black), aggregated mobility (mobility mat, purple), mobility-based community membership level 1 and level 2 (comm overlap l1 & l2, purple), Greater London origin and destination (gr LDN ori & gr LDN dest, red), time of peak incidence at origin and destination (peak time ori & peak time dest, orange) and the residual of a regression of sample size against case count regression at either the origin and destination (sample res ori & sample res dest, yellow).

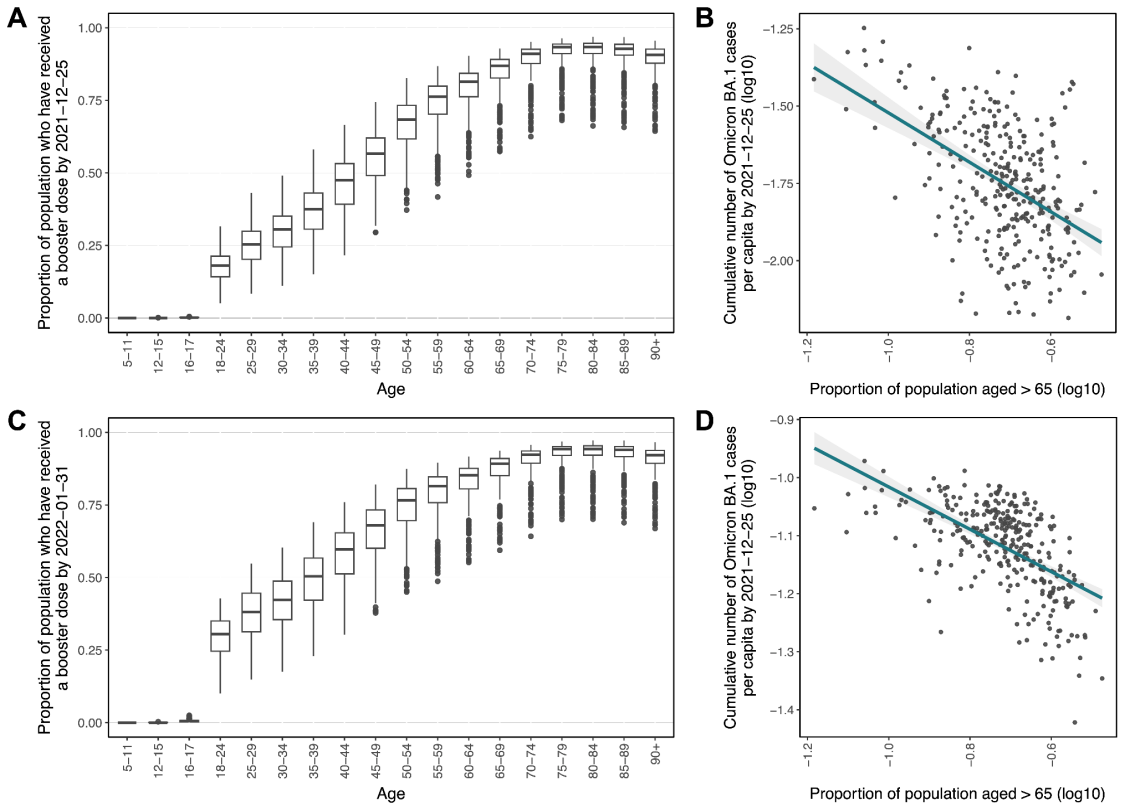


Fig. A.16: Dependency of booster uptakes and cumulative Omicron BA.1 case counts on population age structure. (A, C) Distribution of the proportion of age-specific population in each Lower Tier Local Authority (LTLA) who have received a booster dose by 25 December 2021 and 31 January 2022, respectively. Each box extends from the 25th to 75th percentile of the distribution for the corresponding age group; the midline within each box represents the median; the vertical lines represent the lower and upper limits and the dots denote the outliers. (B, D) Cumulative number of Omicron BA.1 cases per capita (log10-transformed) versus proportion of the population aged above 65 (log10-transformed). Each dot represents an LTLA. Blue lines show the least-squares fit and the shading denotes the associated 95% CI.

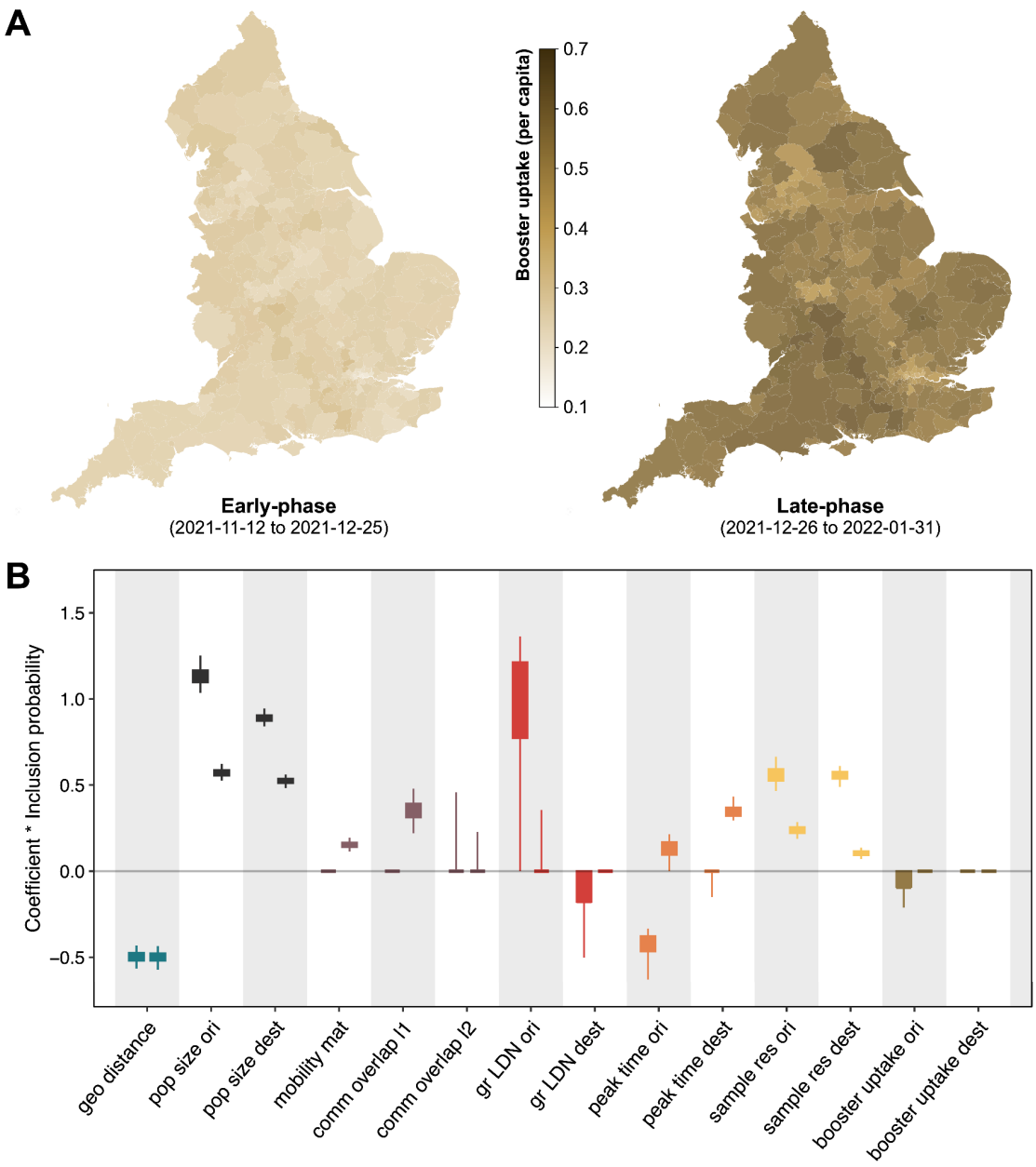


Fig. A.17: Booster uptake as a predictor of BA.1 viral lineage movements in England. (A) Map of age-corrected effective booster uptake at the Lower Tier Local Authority (LTLA) level, averaged over the early-phase (12 November 2021 to 25 December 2021) (left) and the late-phase of the epidemic (26 December 2021 to 31 January 2022) (right). The effective booster uptake is defined as the proportion of the population who would have received a booster dose having accounted for age-specific booster uptakes, assuming the national average population age structure. (B) For each predictor, the box and whiskers show the posterior distribution of the product of the log predictor coefficient and the predictor inclusion probability; the left-hand and right-hand values show the estimates for before and after 26 December, respectively. Posterior distributions are coloured according to predictor type: geographic distances (geo distance, dark blue), population sizes at origin and destination (pop size ori & pop size dest, black), aggregated mobility (mobility mat, purple), mobility-based community membership level

1 and level 2 (comm overlap l1 & l2, purple), Greater London origin and destination (gr LDN ori & gr LDN dest, red), time of peak incidence at origin and destination (peak time ori & peak time dest, orange), the residual of a regression of sample size against case count regression at either origin and destination (sample res ori & sample res dest, yellow), and effective booster uptake at origin and destination (booster uptake ori & booster uptake dest, brown).

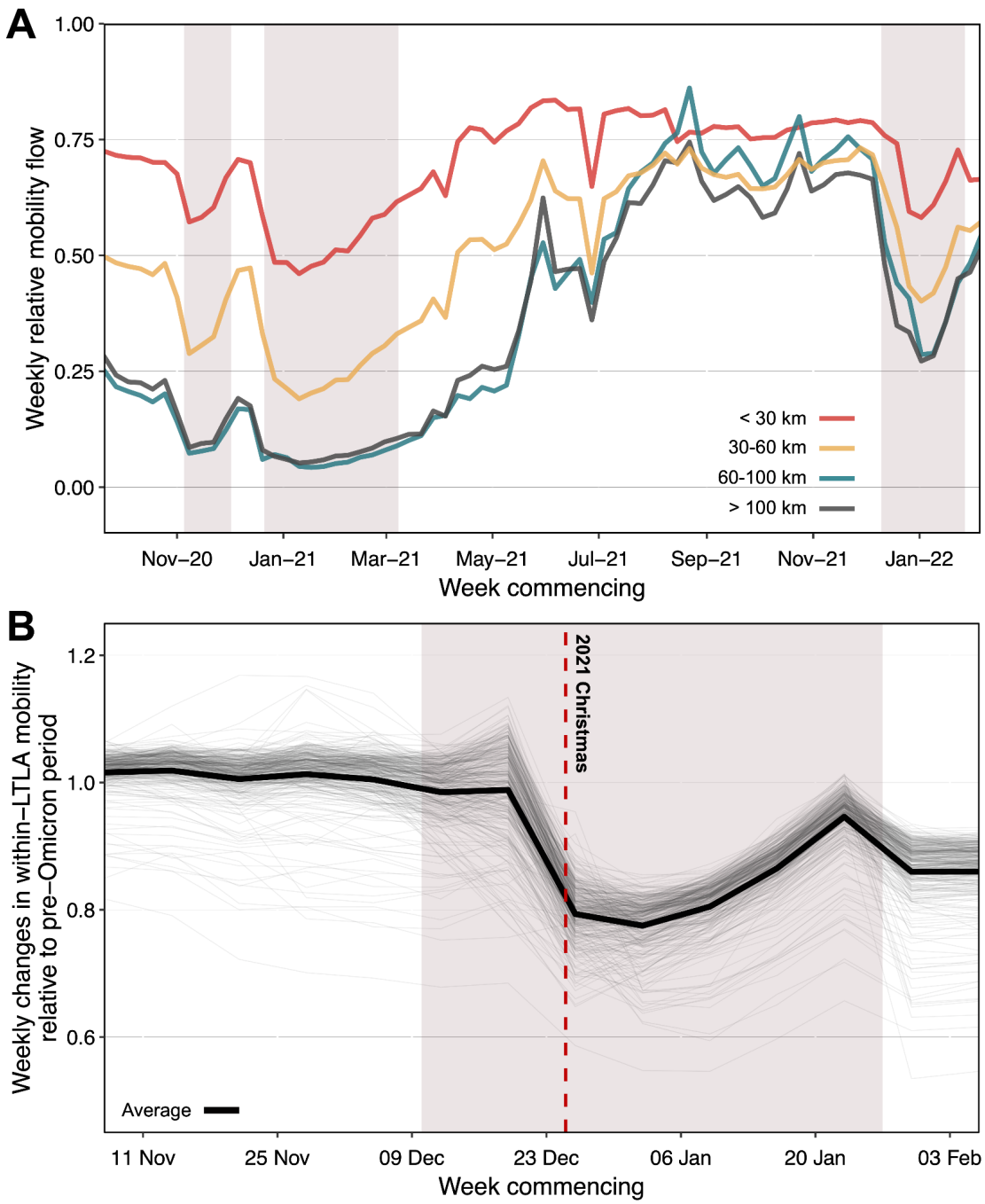


Fig. A.18. Trends in human mobility across England. (A) Weekly human mobility flows relative to pre-pandemic levels (averaged over period from 3 November 2019 to 28 December 2019) across different spatial scales (red: <30 km, yellow: 30-60 km, blue: 60-100 km, dark grey: >100 km). (B) Weekly changes in within-LTLA mobility relative to pre-Omicron levels (averaged over period from 12 September 2021 to 6 November 2021 for each LTLA individually). Thick black line represents the weekly mobility changes averaged over all LTLAs; each thin grey line represents the weekly mobility changes for a single LTLA. Vertical line shows 24th of December 2021.

Table A.1. Imputation of initial growth rate in branching process model. Kullback-Leibler (KL) divergences ($\mathrm{KL1} = D_{KL}(\mathrm{Data} \,||\, \mathrm{Model})$ and $\mathrm{KL2} = D_{KL}(\mathrm{Model} \,||\, \mathrm{Data})$) and the Wasserstein distance between the weekly proportions of local Omicron BA.1 infections attributed to importations that occurred at different times, as (i) inferred from phylodynamic analysis and (ii) predicted by branching process models with different initial growth rates. The initial growth rate in each model was imputed using estimates taken from the early growth phase of the epidemic (2021-12-04 to 2021-12-15). The smallest value for each distance measure is shown in bold.

Initial growth rate taken from	KL1	KL2	Wasserstein
2021-12-04	2.79	2.88	5.48
2021-12-05	2.05	2.02	4.57
2021-12-06	1.38	1.28	3.49
2021-12-07	1.47	1.37	3.66
2021-12-08	2.87	3.03	5.58
2021-12-09	9.10	15.98	9.26
2021-12-10	9.37	16.90	9.35
2021-12-11	6.90	9.71	8.38
2021-12-12	2.11	1.94	4.56
2021-12-13	**0.90**	**0.73**	**2.01**
2021-12-14	2.89	2.63	6.14
2021-12-15	5.26	5.07	8.93

Table A.2. Imputation of initial growth rate in branching process model (sensitivity analysis using incidence estimates from the UK Office of National Statistics). Kullback-Leibler (KL) divergences ($\mathrm{KL1} = D_{KL}(\mathrm{Data}\,||\,\mathrm{Model})$ and $\mathrm{KL2} = D_{KL}(\mathrm{Model}\,||\,\mathrm{Data})$) and the Wasserstein distance between the weekly proportions of local Omicron BA.1 infections attributed to importations that occurred at different times, as (i) inferred from phylodynamic analysis and (ii) predicted by branching process models with different initial growth rates. The initial growth rate in each model was imputed using estimates taken from the early growth phase of the epidemic (2021-11-20 to 2021-12-20). The smallest value for each distance measure is shown in bold. Note that the UK Office of National Statistics (ONS) case incidence (central) estimates are used here for the estimation of daily case growth rates, instead of case incidence data from the GOV.UK COVID-19 Dashboard.

Initial growth rate taken from	KL1	KL2	Wasserstein
2021-11-20	5.20	4.85	5.17
2021-11-21	5.35	4.97	5.24
2021-11-22	5.49	5.09	5.29
2021-11-23	4.82	4.58	5.04
2021-11-24	5.08	4.76	5.13
2021-11-25	5.36	4.97	5.24
2021-11-26	3.66	3.88	4.68
2021-11-27	3.67	3.89	4.68
2021-11-28	2.78	3.59	4.66
2021-11-29	2.66	3.99	5.10
2021-11-30	3.03	4.77	5.68
2021-12-01	5.07	8.19	7.46
2021-12-02	11.08	23.92	9.89
2021-12-03	12.13	28.45	10.19
2021-12-04	11.67	26.32	10.06
2021-12-05	10.05	19.76	9.58
2021-12-06	8.87	15.78	9.18
2021-12-07	7.59	12.20	8.70
2021-12-08	9.51	18.03	9.40
2021-12-09	9.32	17.35	9.34
2021-12-10	9.24	17.08	9.31
2021-12-11	7.15	10.78	8.51
2021-12-12	4.99	6.21	7.36

2021-12-13	2.74	2.80	5.43
2021-12-14	1.58	1.43	3.83
2021-12-15	**0.79**	**0.62**	**1.60**
2021-12-16	1.31	1.12	3.03
2021-12-17	2.72	2.45	5.75
2021-12-18	4.31	4.04	7.90
2021-12-19	6.02	5.89	9.72
2021-12-20	7.21	7.28	10.81

Table A.3. Predictors of viral lineage movements in England. Summary table and descriptions of predictors considered in the discrete phylogeographic GLM analysis. For location-specific predictors (as indicated by an asterisk), we included both an origin and a destination covariate in the GLM model.

Predictor	Domain	Time-varying	Description
Geographical distance between origin and destination	Geography	No	Geographical distance between the origin and destination calculated using the Haversine formula in km
Population size*	Demography	No	Population size at the origin/destination, from population estimates obtained by the Office of National Statistics in mid-year 2020
Aggregated mobility matrix	Human mobility	Yes	Average weekly number of trips taken between origin and destination estimated from Google COVID-19 Aggregated Mobility Research Dataset; given that the mobility flux between two locations does not in general differ from symmetry in a statistically significant manner (i.e. the magnitude of mobility flux in either direction is generally very similar for a given connection), we considered a mobility matrix that was symmetrised
Community memberships from mobility network (level-1/2)*	Human mobility	Yes	A binary variable [0,1] indicating whether the origin and destination belong to the same community at level-1/2; community structures were identified from the human mobility network as described by the aggregated mobility matrix, using the community detection algorithm Infomap (1, 2), with level-1/2 corresponding to the tree-depth at which the communities were extracted; level-1 has a higher level of aggregation (fewer communities) compared to level-2 (more communities)
Greater London / non-Greater London indicator*	Geography/ Epidemiology	No	A binary variable [0,1] indicating whether the origin/destination is in the Greater London region
Timing of peak in Omicron BA.1 case incidence*	Epidemiology	No	Timing of first peak in Omicron BA.1 case incidence at the origin/destination, measured as the number of days from 1 December 2021

Sampling residuals*	Sampling	No	Residuals from a regression of sample size against Omicron BA.1 case count at origin/destination; time-invariant residuals were used due to the small number of samples in some locations during the post-expansion phase

Table A.4. Definitions of Lower Tier Local Authorities (LTLAs). Table of LTLAs that have been deprecated and aggregated into newly defined LTLAs, according to recent definitions used in the report of population estimates for the UK in mid-2020, compiled by the Office of National Statistics, UK.

Most recent LTLA definition	**Deprecated LTLA definition(s)**
E06000058 (Bournemouth, Christchurch and Poole)	E06000028, E06000029, E07000048
E06000060 (Buckinghamshire)	E07000004, E07000005, E07000006, E07000007
E06000059 (Dorset)	E07000049, E07000050, E07000051, E07000052, E07000053
E07000244 (East Suffolk)	E07000205, E07000206
E07000244 (West Suffolk)	E07000201, E07000204
E07000246 (Somerset West and Taunton)	E07000190, E07000191

References

1. Rosvall, M., Axelsson, D. and Bergstrom, C.T. (2009) ‘The map equation’, *The European physical journal. Special topics*, 178(1), pp. 13–23.
2. Rosvall, M. and Bergstrom, C.T. (2011) ‘Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems’, *PloS one*, 6(4), p. e18209.

3

Toward optimal disease surveillance with graph-based active learning

Following my work on Omicron BA.1, my research focus shifted to address a broader challenge of improving our approach to disease surveillance. This shift was motivated by a key finding from Chapter 2 suggesting that the targeted travel restrictions implemented during the Omicron outbreak were largely ineffective. This ineffectiveness stemmed in part from policy decisions based on incomplete risk assessments - specifically, the failure to account for viral importation from major international travel hubs not covered by the travel bans. This highlighted an urgent need for surveillance systems capable of providing more accurate and timely risk assessments to inform containment efforts. The work presented in this chapter represents an initial step towards addressing this need, through the development of an adaptive test deployment framework that enables more accurate inference of the underlying disease distribution across a mobility network under resource constraints.

A manuscript describing this work was first made available on MedRxiv as a preprint on 21st June 2024, and later published in *PNAS* on 19th December 2024, under the title "*Toward optimal disease surveillance with graph-based active learning*". It is presented here in full, with minor modifications to ensure consistency of formatting and style within this thesis.

Tsui, J.L.-H.*, Zhang, M.*, Sambaturu, P., Busch-Moreno, S., Suchard, M.A., Pybus, O.G., Flaxman, S., Semenova, E. and Kraemer, M.U.G. (2024) 'Toward optimal disease

surveillance with graph-based active learning', *Proceedings of the National Academy of Sciences of the United States of America*, 121(52), p. e2412424121.

(* indicates joint-first authorship)

3.1 Abstract

Tracking the spread of emerging pathogens is critical to the design of timely and effective public health responses. Policymakers face the challenge of allocating finite resources for testing and surveillance across locations, with the goal of maximizing the information obtained about the underlying trends in prevalence and incidence. We model this decision-making process as an iterative node classification problem on an undirected and unweighted graph, in which nodes represent locations and edges represent movement of infectious agents among them. To begin, a single node is randomly selected for testing and determined to be either infected or uninfected. Test feedback is then used to update estimates of the probability of unobserved nodes being infected and to inform the selection of nodes for testing at the next iterations, until certain test budget is exhausted. Using this framework, we evaluate and compare the performance of previously developed active learning policies for node selection, including Node Entropy and Bayesian Active Learning by Disagreement. We explore the performance of these policies under different outbreak scenarios using simulated outbreaks on both synthetic and empirical networks. Further, we propose a policy that considers the distance-weighted average entropy of infection predictions among neighbours of each candidate node. Our proposed policy outperforms existing ones in most outbreak scenarios given small test budgets, highlighting the need to consider an exploration–exploitation trade-off in policy design. Our findings could inform the design of cost-effective surveillance strategies for

emerging and endemic pathogens and reduce uncertainties associated with early risk assessments in resource-constrained situations.

3.2 Introduction

Infectious disease surveillance is necessary for managing infectious disease outbreaks, enabling public health authorities to monitor and respond to ongoing disease spread. Notable examples in the past decade include the 2014–2016 West African and 2018–2020 Kivu Ebola virus epidemics, and the COVID-19 pandemic, for which the early detection and continued tracking of the virus' spread helped to inform the design of interventions including targeted vaccination (1-5), case isolation (6-10), and social distancing (11-14). Without timely and accurate surveillance data, the effectiveness of these interventions would likely have been compromised. For example, travel restrictions targeted at countries where new variants of SARS-CoV-2 were first observed were rendered largely ineffective by delays in case detection and insufficient pathogen sequencing (15, 16). Similarly, the lack of baseline testing prior to the 2015–2016 Zika virus epidemic in the Americas likely contributed to the delay in the identification of the scale of disease spread, thereby allowing the virus to disseminate to new locations before a coordinated response was initiated (17, 18).

Well-documented examples of effective disease surveillance have been limited largely to within-country initiatives [e.g., the Real-time Assessment of Community Transmission (REACT) in the United Kingdom (19) and the National Notifiable Diseases Surveillance System (NNDSS) in the United States (20)], while globally coordinated programs remain rare (21). This can lead to disproportionate or inequitable distributions of testing resources within and between regions or countries, with some locations able to conduct large-scale mass testing for sustained periods of time, while others manage only

sparse or sporadic testing (22, 23). One study showed that the intensity of viral genomic sequencing during the COVID-19 pandemic was positively associated with Research & Development expenditures at a country level (24). This likely allowed the virus to continue proliferating undetected in locations with insufficient testing, potentially prolonging local outbreaks.

Previous research on infectious disease surveillance has focused primarily on developing models to identify sentinel sites or subpopulations, with the objective of classifying nodes in networks that could serve as observational units for monitoring disease spread (25-27). Since the COVID-19 pandemic, there has been growing interest in the design of optimal control measures to contain transmission (28), with some studies examining the cost-effectiveness of different strategies for testing and isolation (29-32); one recent study also explored the impact of different air travel regulations on the likelihood of a local epidemic escalating into a global pandemic (33). However, the effectiveness of these interventions depends ultimately on the capacity of local authorities to conduct surveillance and to collectively provide i) timely data of where the disease has been detected (34-36), and ii) an accurate assessment of overall disease distribution (both presence and absence of infections) at any stage of an outbreak – a challenge which, to the best of our knowledge, has received little attention to date.

This study attempts to address this problem; specifically, we consider how testing should be performed across a mobility network, with the objective of providing accurate estimates of where a disease is present, given a fixed budget of testing resources. We hypothesize that the design of an appropriate policy for this task can be formulated as a node classification problem with active learning (AL), where the objective is to select nodes in a partially observed graph for labelling in a manner that maximizes the performance of a model predicting the label of unobserved nodes, while minimizing the

amount of labelled data required (37). Importantly, we note that this differs from a related problem that arises frequently in the context of early case detection and contact-tracing – also referred to as active search in Garnett et al. (38) – where the objective is to find as many infected individuals as possible given a fixed test budget. This motivates the development of an adaptive test deployment framework, which we use to evaluate and compare the performance of previously developed AL policies for infectious disease surveillance. We further propose a policy that takes into consideration graph-based uncertainties, named Selection by Local Entropy (LE), which we show outperforms similar existing policies in most outbreak scenarios and on networks with a diverse range of structural properties, including those commonly found in empirical human mobility networks, especially when test budgets are small.

3.3 Materials and methods

3.3.1 Disease surveillance as a node classification task

We consider the deployment of a disease surveillance program on a mobility network as a node classification task, in which the mobility network is represented as an undirected and unweighted graph $G = (V, E)$, with nodes $v_i \in V$ representing locations, and edges $(v_i, v_j) \in E$ representing the existence of movement of infectious agents between nodes v_i and v_j. Assuming that there is an underlying distribution of infections resulting from an infectious disease outbreak, the goal of a policymaker (or agent) in this classification task is to predict the presence or absence of the disease (or whether disease prevalence is above or below a certain threshold) at any unobserved node, given the knowledge of the infection status of a subset of nodes in the network.

To generate an underlying disease distribution across the mobility network, we simulate an infectious disease outbreak by modelling its spread as a stochastic Susceptible-Infected (SI) process on graphs, such that transmission can occur only between an infected node and an uninfected node if there is an edge between them. We assume that the outbreak originates from a single, randomly selected node and terminates when a certain proportion (10%, 30%, or 50%) of nodes become infected (Fig. 3.1A, red compartment; see also column 3 in Table 3.2 and Section B.1 in Appendix B for further details). Importantly, we assume that the timescale over which transmission occurs is sufficiently longer than the timescale over which testing resources are deployed, such that the resulting disease distribution at the end of the simulated outbreak can be considered as static over the course of the surveillance program (Fig. 3.1A, blue compartment). To indicate the underlying disease distribution, we assign each node v_i in the mobility network a binary label $y_i \in \{0,1\}$ to represent its infection status, where $y_i = 1$ if the node is infected (disease presence) and $y_i = 0$ if uninfected (disease absence).

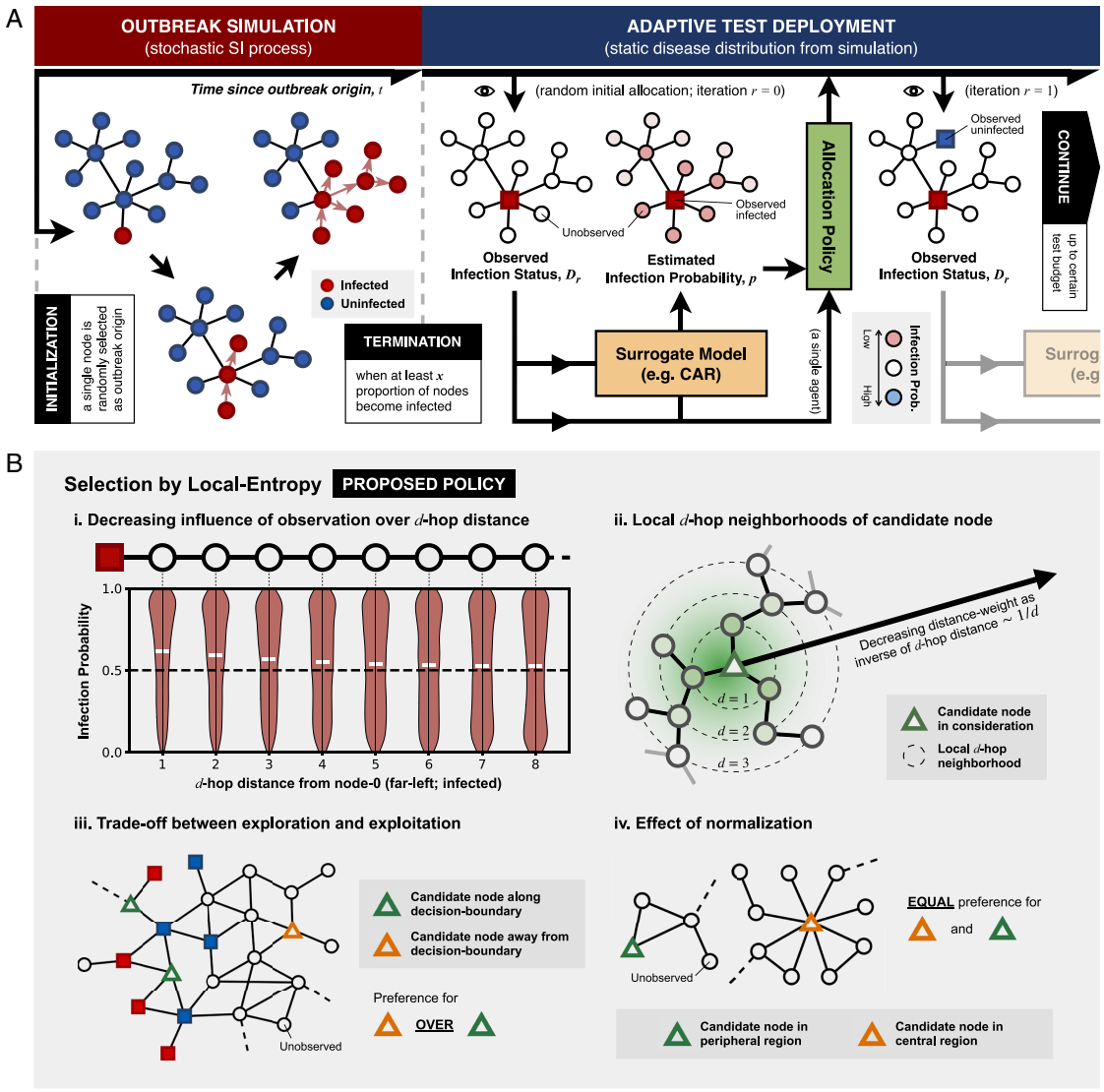


Fig. 3.1. Disease surveillance on a static graph as a node classification task with active learning. (A) A schematic illustration of the simulation of infectious disease spread on an undirected and unweighted graph (Left-hand side, red compartment), followed by the implementation of a disease surveillance program under an adaptive test deployment framework assuming a static disease distribution (Right-hand side, blue compartment). The flow of information/data from one component of the framework to another is represented as arrows. The eye symbol indicates when the underlying disease distribution is queried, thereby revealing the true infection status of a selected node. (B) Key concepts behind our proposed policy named Selection by Local Entropy. i) An example showcasing the decreasing influence of an observed node on the estimated infection probability of remaining unobserved nodes in a graph with a chain-like structure. The violin plot shows the posterior distribution of the infection probabilities for the remaining unobserved nodes at different d-hop distances from the observed node on the far-left (node 0); the posterior mean of the probability of each node being infected is indicated by a white horizontal line. The black dashed line indicates an infection probability of 0.5 (i.e., most uncertain). ii) An

illustration of the concept of local d-hop neighbourhoods, represented by black dashed concentric circles, centred around a candidate node (green triangle). The green shading indicates the distance weight which decreases with increasing d-hop distance from the candidate node following an inverse relationship. iii) An example showcasing the trade-off between exploration and exploitation, with Selection by Local Entropy preferring the selection of the candidate node in the unexplored region (orange triangle) over candidate nodes lying along decision boundaries (green triangles). iv) An example illustrating the effect of normalisation by the sum of distance weights over all d-hop neighbourhoods (see definition of Local Entropy), resulting in an equal preference for candidate nodes that lie in the peripheral (green triangle) and central (orange triangle) region of a graph.

Provided that the infection status of a subset of nodes is observed, the infection status of remaining unobserved nodes can then be inferred probabilistically by considering their connections to the observed nodes; we refer to the model that performs this inference as a surrogate model (orange box in Fig. 3.1A). Here, we adopt an approach known as Conditional Autoregressive (CAR) model (39), which estimates the probability that each node is infected (or its posterior distribution under a Bayesian framework) conditional on the infection status of the observed nodes alone (i.e., there are no external data informing the probability estimates except for the observed infection status; see Section B.2 in Appendix B for a detailed description of the model). To assess the degree to which the surrogate model is able to correctly predict the infection status of remaining unobserved nodes given the observed data, we evaluate the Area Under the Receiver Operating Characteristics Curve (AUC) by comparing the infection probability estimates (posterior mean from the CAR model) with the true underlying infection status, where a higher AUC indicates a better predictive performance.

3.3.2 Test allocation as an active learning task

Given a fixed test budget (i.e., a fixed number of tests to be allocated), the predictive performance of the surrogate model will vary depending on which nodes are selected for testing [a task known as AL (37)] and therefore the observed data that are available for model training. To maximize this performance, we consider a number of existing AL policies with a particular focus on those that are adaptive, i.e., policies that select nodes for testing in an iterative fashion until the test budget is exhausted (37). At each iteration, observed data from previous tests are used as input to retrain the surrogate model and to generate infection probability estimates for remaining unobserved nodes; these estimates are then used to guide the selection process at the next iteration, with selection criterion depending on the policy of choice (Fig. 3.1A, blue compartment).

We consider two adaptive AL policies in this study, namely, Node Entropy (NE) (40) and Bayesian Active Learning by Disagreement (BALD) (41). Both of these policies are uncertainty-based, as they select nodes for testing according to where the surrogate model's predictions are considered to be most uncertain (refer to Table 1 for detailed descriptions of both policies, and Section B.3 in Appendix B for BALD specifically). For comparison, we also consider two nonadaptive, graph-based AL policies, i.e., unobserved nodes are selected for testing by considering only their positions in the network, without using information from previous test iterations (Table 1 for more detailed descriptions).

Table 3.1. Summary of policies considered in this study. Abbreviation for each policy is shown in brackets following the policy name. For all policies, random tie-breaking is performed if and when there are multiple candidate nodes given equal preference according to a selection criterion.

Allocation policy	Policy type	Brief description
Node Entropy (NE) (40)		Select the unlabelled node with the highest entropy in its label prediction according to the surrogate model.
Bayesian Active Learning by Disagreement (BALD) (41)		Select the unlabelled node with the highest mutual information between label prediction and posterior from the surrogate model.
Local Entropy (LE) *(our proposed policy)*		Select the unlabelled node with the highest *Local Entropy*, as defined by Eqs. 1–3, with $\lambda = 0$ (maximal exploration).
Degree Centrality (DC)	- Graph-based - Non-adaptive	Select the unlabelled node with the highest degree centrality (most connections).
PageRank Centrality (PC)		Select the unlabelled node with the highest PageRank centrality (42).
Reactive-Infected (RI)	- Benchmark - Adaptive	Select at random an unlabelled node among immediate neighbours of nodes that are known to be infected from previous observations, if available; otherwise, sample randomly from remaining unlabelled nodes.
Random (RAND)	- Benchmark - Non-adaptive	Select an unlabelled node at random.

Table 3.2. Summary of all experiments conducted in this study.

Experiment	Graph(s)	Outbreak scenario(s)
Preliminary *(only uncertainty-based policies are considered)*	Aperiodic lattice graph (with square tiling)	50 random outbreak realizations, with each outbreak terminating when at least 30% of the nodes become infected ($I/N = 0.3$).

Synthetic graphs	Periodic lattice graph (with square tiling) (*graph-based policies are not considered)*	50 random outbreak realizations for each termination condition ($I/N =$ 0.1, 0.3, 0.5); this amounts to a total of 150 random outbreak realization for each graph.
	A random graph generated by the Barabási–Albert model (46), with each node having a minimum of two connections ($m = 2$)	
	A random graph generated by the stochastic block model (47), with low-modularity settings (Section B.4 in Appendix B)	
	A random graph generated by the stochastic block model (47), with high-modularity settings (Section B.4 in Appendix B)	
Empirical human mobility networks	Graphs derived from aggregated mobility data collected from mobile phone users in Italy at the provincial level during March to May, 2020 (48), with thinning thresholds $T_{thinning} =$ 10%, 15%, 20% (Section B.5 in Appendix B)	
	Graphs derived from global air traffic data collected at the country level during January to March, 2020 (49), with thinning thresholds $T_{thinning} =$ 2.5%, 5%, 7.5% (Section B.6 in Appendix B)	

3.3.3 Our proposed policy: Selection by Local Entropy (LE)

One potential drawback of using selection criteria based on uncertainty-based metrics alone (as seen in NE and BALD) is that they can lead to a bias toward selecting nodes from regions with highly heterogeneous node labels. In the context of infectious disease surveillance, this can be interpreted as a preference for "exploitation" in an exploration–

exploitation trade-off, where exploitation means the selection of nodes that lie along the boundaries between infected and uninfected regions (i.e., decision boundaries) and therefore have highly uncertain infection status predictions despite the availability of data, and "exploration" means the selection of nodes from less observed regions of the graph and therefore with uncertain infection status predictions that are informed by little data [panel (iii) in Fig. 3.1B]. Previous attempts to account for this trade-off have been made, particularly in the context of AL with Graph Neural Network (GNN) models (43), whereby the exploration of less observed regions is encouraged by increasing the probability that a node is selected according to the number of unlabelled neighbours it has (44), or the degree to which the candidate node is representative of its unlabelled neighbours in feature space according to their node attributes (45).

With insights from these previous efforts, we propose here a policy which we refer to as Selection by Local Entropy (LE). This policy evaluates the informativeness of an unlabelled node by taking into account not only the uncertainty in the predicted label of the candidate node itself but also that of surrounding nodes. At a given iteration r, we define the Local Entropy of an unlabelled node v_k as a linear combination of the entropy of the label prediction for node v_k itself, denoted by $\Omega_{k,r}^{self}$, and the distance-weighted average entropy of the label predictions for surrounding nodes, denoted by $\Omega_{k,r}^{surr}$, as follows,

$$\Omega_{k,r} = \lambda\Omega_{k,r}^{self} + (1-\lambda)\Omega_{k,r}^{surr} \quad (1)$$

with $\lambda \in [0,1]$, and

$$\Omega_{k,r}^{self} = \mathrm{H}(v_k|\boldsymbol{D_r}) \quad (2)$$

$$\Omega_{k,r}^{surr} = \frac{\sum_{d=1}^{d_{max}} \sum_{v_i \in V(d,v_k)} \mathrm{H}(v_i|\boldsymbol{D_r})/d}{\sum_{d=1}^{d_{max}} \sum_{v_i \in V(d,v_k)} 1/d} \quad (3)$$

where $\mathrm{H}(v_k|\boldsymbol{D_r})$ is the entropy of the label prediction for node v_k, conditional on the currently observed data $\boldsymbol{D_r} = \{(v_1, y_1), (v_2, y_2), \ldots, (v_n, y_n)\}$.

Key insights that motivate the definition of Local Entropy can be summarized as follows:

1. The information that can be gained from the observation of a node is likely to be greater if it is in close proximity to other unlabelled nodes with highly uncertain label predictions [see panel (i) in Fig. 3.1B].
2. The influence that a new observation has on the label predictions for surrounding nodes decays with increasing hopping distance d. This, together with insight (1), motivates the definition of $\Omega_{k,r}^{surr}$ for a given candidate node v_k, as the sum of the entropies of the label predictions for all surrounding nodes, with the contribution from nodes in each d-hop neighbourhood [denoted by $V(d, v_k)$] weighted by the inverse of their hopping distance, $1/d$ (Eq. 3). This summation extends up to a maximum d-hop distance d_{max}, beyond which the influence of a new observation on the label predictions for unobserved nodes is assumed to be negligible. Altogether, $\Omega_{k,r}^{surr}$ serves as a proxy measure of the total impact that an observation of a candidate node v_k is likely to have on the label predictions for surrounding nodes [see panel (ii) in Fig. 3.1B].
3. This sum, as described in (2), is normalised by the sum of the distance weights $(1/d)$ across all d-hop neighbourhood (up to a hopping distance of d_{max}); this prevents a bias where centrally located nodes would have larger values of $\Omega_{k,r}^{surr}$, simply due to having more connections. As a result of this normalisation, there is an equal preference for nodes in both the peripheral regions (with low centrality) and central regions (with high centrality) of a network, assuming that both regions are equally unexplored [panel (iv) in Fig. 3.1B].

4. Although $\Omega_{k,r}^{self}$ and $\Omega_{k,r}^{surr}$ are generally correlated for a given candidate node k, the strength and direction of the correlation is not always the same. For example, an unobserved node in a relatively unexplored region of the network is likely to have both high $\Omega_{k,r}^{self}$ and $\Omega_{k,r}^{surr}$, whereas an unobserved node lying along a decision boundary would have low $\Omega_{k,r}^{surr}$ (due to surrounding nodes being observed) but high $\Omega_{k,r}^{self}$ (due to conflicting information from nearby observed nodes with opposite infection status). The design choice of our proposed selection criterion as a linear combination of $\Omega_{k,r}^{self}$ and $\Omega_{k,r}^{surr}$ therefore provides a practical heuristic to balance the trade-off between exploration and exploitation through different values of the weighting parameter λ. In the case where $\lambda = 1$, we recover the uncertainty-based policy NE which performs node selection based on node entropy $\Omega_{k,r}^{self}$ alone; see Fig. B.2 in Appendix B for a sensitivity analysis exploring the effect of different values of λ.

Note that we set d_{max} to the graph diameter d_G (i.e., the largest geodesic distance between any pair of nodes) and $\lambda = 0$ (i.e., maximal exploration) in all subsequent considerations of our proposed policy LE. See Figs. B.1 and B.2 in Appendix B for sensitivity analyses exploring the impact of different values of d_{max} and λ, considering outbreak scenarios on an aperiodic lattice graph (with $I/N = 0.5$); future work should investigate how these results generalise to more diverse outbreak settings and network structures.

3.3.4 Policy evaluation under different network structures and outbreak scenarios

We conduct three sets of experiments as summarized in Table 2, with each experiment considering a different graph and outbreak scenario. Specifically, we consider synthetic graphs generated by different generative models (column 2 in Table 2) and therefore with different degree distributions and varying levels of community structure and structural disorder. We also consider two empirical human mobility datasets (row 3 in Table 2), from which we derive two unweighted and undirected graphs following a procedure known as graph thinning, where only mobility flows above a certain thinning threshold are preserved (see Sections B.5, B.6 in Appendix B for details).

To explore the impact of different stages of outbreak progression on policy performance, we simulate outbreaks with different termination conditions, as measured by the proportion of nodes that are infected (column 3 in Table 2). For each random outbreak realization on a given network, 25 different nodes are randomly selected as the initial labelled node, with nodes of either infection status being equally likely to be selected; at the beginning of each experiment, the infection status of the same initial labelled node is made available to all agents (with each agent being assigned one of the policies being considered). This is done to account for any variability in policy performance resulting from different initial observations, as well as stochasticity from the Markov chain Monte Carlo (MCMC) inference process and from random tie-breaking whenever two or more candidate nodes are given equal preference by a policy according to its selection criterion.

3.3.5 Measuring policy performance and test budget specifications

Following the selection of an initial labelled node for a simulated outbreak, as described above, we assess the performance of a given agent (policymaker) at each test iteration by evaluating the AUC, based on a comparison between the current label predictions from the surrogate model (given the available data) and the true infection status of remaining unobserved nodes. The performance of an agent at a given test iteration r can therefore be interpreted as the performance of its designated policy for a given test budget r, assuming no further test deployments.

In each experiment, which considers simulated outbreaks at a specific stage of progression on a given graph, we compare the performance of different policies over a range of test budgets. The maximum test budget is determined by the median number of test iterations required by the Reactive-Infected (RI) policy to identify all infected nodes across all relevant outbreak realizations. Since RI mimics a "contact tracing" approach (Table 1 for a more detailed description of RI), this maximum represents the average minimum number of tests required to identify all infected nodes in a given outbreak scenario. It is therefore only when considering test budgets below this maximum that the objective of accurately predicting the presence or absence of a disease of interest across a mobility network—without complete identification of all infected nodes—may be considered relevant to public health decisions. In all following experiments, we compare the performance of the different policies only at test iterations up to this maximum; full results are presented in Figs. B.3, B.4 in Appendix B.

3.4 Main results

3.4.1 Disease surveillance on an aperiodic regular lattice graph

As a preliminary experiment to illustrate the differences between the uncertainty-based policies considered, we evaluate and compare their performance on an aperiodic regular lattice graph with square tiling. We observe that our proposed policy LE on average performs better than both NE and BALD at small numbers of test iterations ($r < 30$; Fig. 3.2B). LE and NE show similar performance between $r = 30$ and $r = 50$; at $r > 50$, however, NE overtakes LE as the best performing policy with an AUC that rapidly approaches 1, while both LE and BALD struggle to attain a perfect AUC. The difference in performance between LE and NE can be understood in the context of the exploration–exploitation trade-off as described above: at small r, LE encourages an even allocation of tests across the graph (exploration), while NE favours regions with highly heterogeneous disease distributions (exploitation) (see columns 2 and 3 in Fig. 3.2A)—this results in a more rapid increase in model performance for LE as r increases. At large r, however, the greater preference for exploitation by NE means that almost all nodes along the decision boundary are sampled; this results in an AUC that rapidly approaches 1. Although LE also shows a preferential selection of nodes close to the decision boundary at large r, it does so at a much slower rate than does NE.

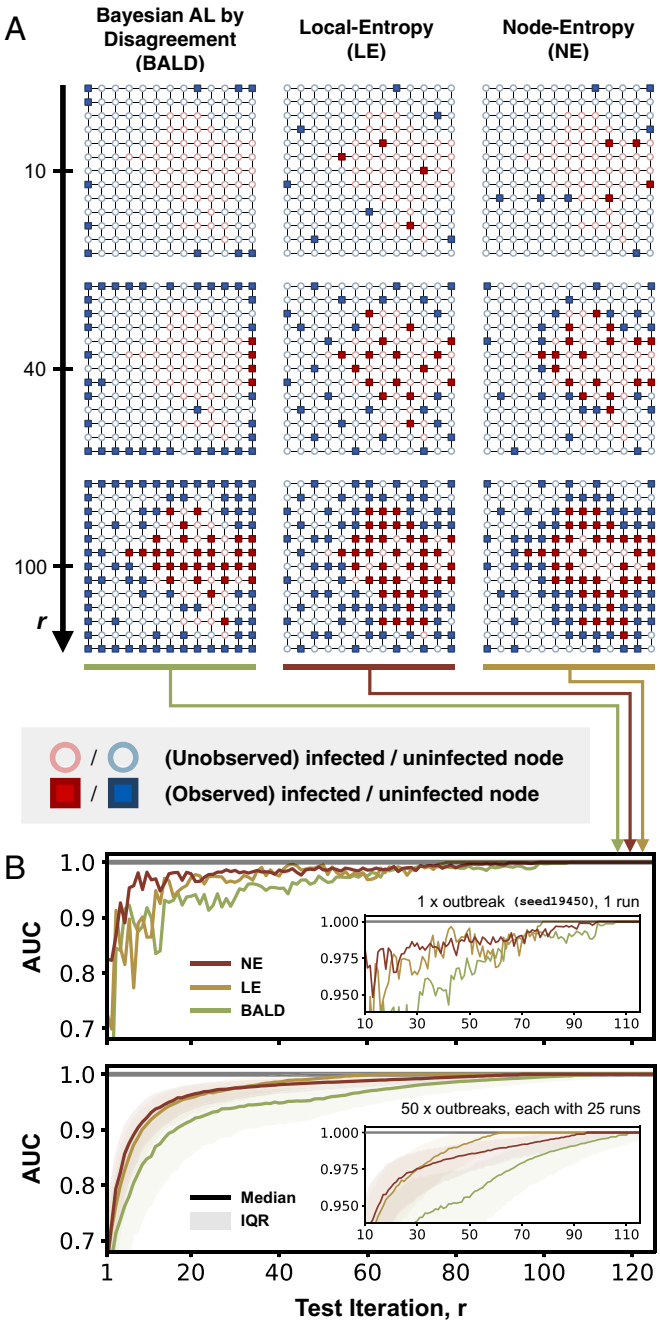


Fig. 3.2. Comparison of Selection by Local Entropy (LE) with existing uncertainty-based policies in the context of simulated outbreaks on an aperiodic lattice graph. (A) Test allocation by three selected agents, with each agent assigned a different policy (LE, NE, or BALD). Each square panel shows the distribution of observed (squares) and unobserved (circles) nodes up to a given test iteration ($r = 10$, 40, 100; as indicated by labels on the Left) for a given agent (as indicated by labels at the Top). Each node is coloured according to its true infection status (red if infected and blue if uninfected, with circles that represent unobserved nodes having a lower opacity). (B) The top plot shows the performance of the three selected agents for a single outbreak realization, as measured by the AUC; higher AUC values indicate better performance. The Bottom plot shows the performance of LE and the two

existing uncertainty-based policies (NE and BALD), each summarized across 1,250 agents (50 outbreak realizations, each with 25 unique initial labelled nodes); the shaded region represents the interquartile range and the solid line represents the median. The Inset in each plot shows the same data in the interval $10 \leq r \leq 115$ on an enlarged scale.

BALD on average performs worse than NE and LE across all test budgets. This is due to its apparent preferential selection of low-degree nodes (either in the corners or along the edges); only at $r > 40$ (at which point no low-degree nodes remain) does BALD exhibit a pattern of test allocation that resembles that of NE. The observed underperformance of BALD is consistent with results from a previous evaluation of existing AL policies for node classification (50), likely explained by the fact that BALD does not consider the graph structure in its formulation (41).

3.4.2 Disease surveillance on synthetic graphs

There are three key observations from our results presented in Fig. 3.3. First, all policies except for BALD and RI outperform random allocation (RAND) across all outbreak scenarios, especially at large r when the performance of random allocation appears to only increase slowly with increasing r. Given the preferential selection of low-degree nodes by BALD, as mentioned, it is not surprising that BALD only shows comparable performance in the periodic lattice graph which has no degree variation. Second, uncertainty-based policies (NE, BALD, and LE) underperform substantially compared to graph-based heuristics (DC and PC) on the synthetic graph generated by the Barabási–Albert model (hereafter referred to as the BA graph), especially when considering outbreaks at early ($I/N = 0.1$) or intermediate ($I/N = 0.3$) stages, with DC and PC together being ranked top greater than 50% of the time, on average, across all test budgets (Fig. B.7 and Table B.4 in Appendix B). This observation can be explained by considering

the infection-assortativity, $r_{\text{infection}}$, which in the context of disease distribution, is a measure of the tendency for two connected nodes to share the same infection status [as has been repeatedly shown in empirical studies that mobility synchronizes epidemics across locations (51); see Section B.7 in Appendix B for definition of infection-assortativity]. Evaluating the average $r_{\text{infection}}$ value across all outbreak realizations on each graph shows that outbreaks on the BA graph have on average the lowest $r_{\text{infection}}$ at 0.20 [compared to 0.64 for the periodic lattice graph, 0.48 and 0.63 for the graphs generated by the stochastic block model (SB graph) with low and high modularity (52), respectively]. A low (but positive) $r_{\text{infection}}$ value indicates a weak tendency for two connected locations to share the same infection status, and therefore a low degree of homophily in the underlying disease distribution. This results in an overall poor predictive performance from the surrogate model, which in turn limits the effectiveness of uncertainty-based policies. In such cases, it may then be advantageous to consider node centrality alone during node selection, especially at small r when there are little data to inform model predictions. Note also that PC tends to perform better than DC—this is not unexpected given that nodes with the most connections are not necessarily the most central in a network.

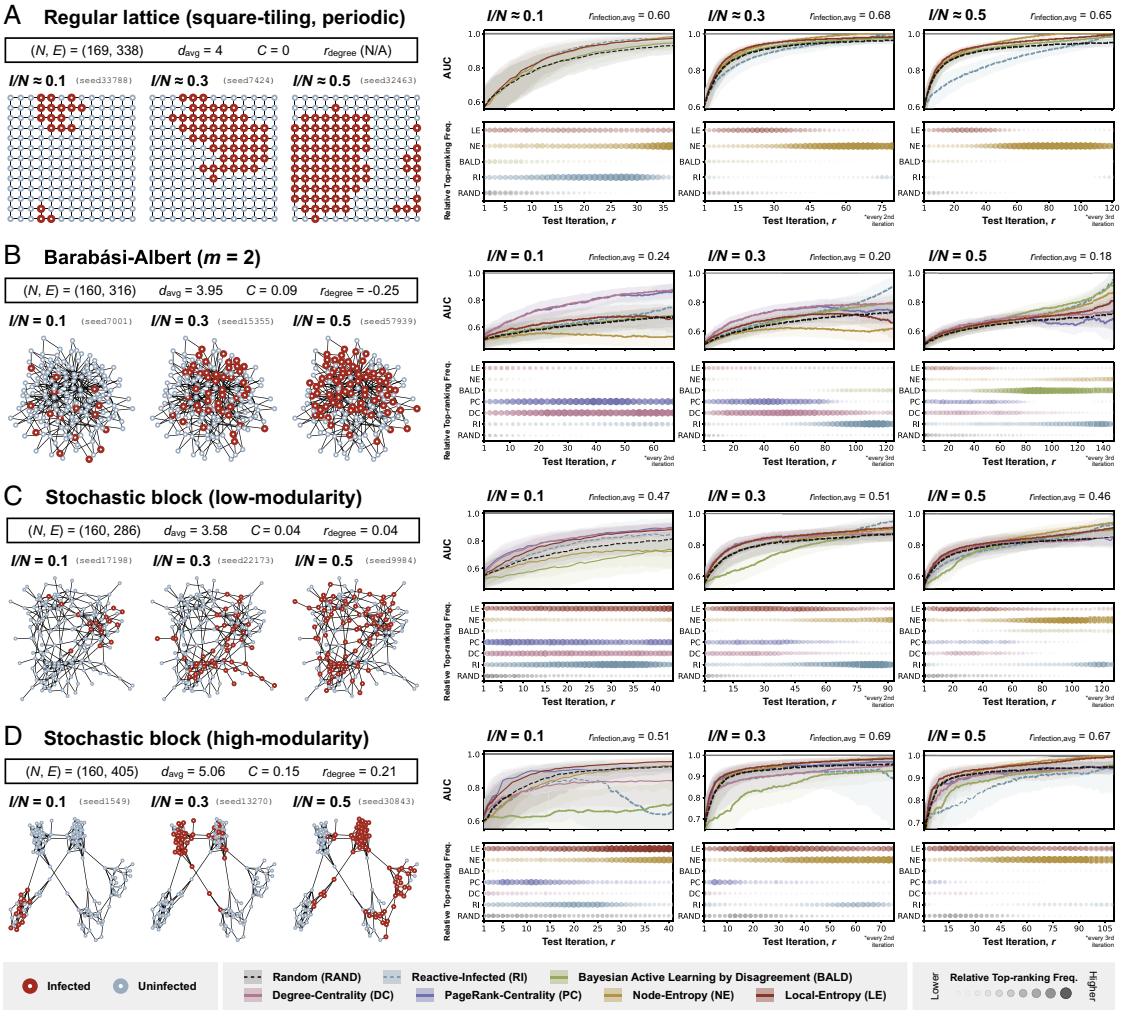


Fig. 3.3. Policy evaluation with simulated outbreaks on synthetic graphs. Each panel (A, B, C, and D) corresponds to results from experiments with simulated outbreaks on a different synthetic graph (panel A: a periodic regular lattice graph with square tiling; panel B: a random graph generated by the Barabási–Albert model, with each node having a minimum of two connections ($m = 2$)); (C): a random graph generated by the stochastic block model at low-modularity settings; (D) a random graph generated by the stochastic block model at high-modularity settings). Summary statistics relevant to the structure of each graph [number of nodes (N), number of edges (E), average node degree (d_{avg}), clustering coefficient (C), and degree assortivity (r_{degree})] are shown in the Top-Left part of each panel. In the Bottom-Left part of each panel are visualizations of three selected disease distributions (with their corresponding seeds shown), each at a different stage of outbreak progression as measured by the proportion of nodes infected ($I/N = 0.1, 0.3, 0.5$); nodes are coloured according to their true infection status (red if infected and blue if uninfected). In the Right part of each panel, each column shows results from experiments considering disease distributions at a different stage of outbreak progression. The Top plot in each column shows the performance of policies considered in the corresponding experiment, as measured by the AUC, with a higher AUC indicating a better performance; the shaded

region represents the interquartile range and the solid line represents the median. The Bottom plot in each column shows the frequency with which each policy is ranked top according to its AUC at each iteration (or every 2nd or 3rd iteration, where indicated), normalised by the difference between the highest and lowest frequencies across different test budgets in the corresponding experiment (refer to Fig. B.7 and Tables B.1-B.4 in Appendix B for the absolute frequencies); a larger circle with a greater opacity indicates a higher frequency of the policy being ranked top at a given test iteration. The performance of each policy is only shown up to the median number of test iterations required for all infected nodes to be observed among agents assigned to Reactive-Infected (RI) (see Fig. B.3 in Appendix B for full results).

Finally, we observe generally favourable performance for LE across most of the outbreak scenarios on graphs with a high degree of structural order (unlike the BA graph, as described), especially at small r. At larger r, however, we again observe superior performance for NE, with AUCs that rapidly approach 1. This can again be explained by the preference for exploitation over exploration by NE, which leads to the complete observation of the decision boundary between infected and uninfected regions given a sufficient number of tests. This is also reflected in the observation that NE is substantially more likely to be ranked top at large r (right-panel in Fig. B.7 and Tables B.1-B.3 in Appendix B), compared to LE at small r (partly because of the limited information available when the number of observed nodes is small and therefore smaller distinction in policy performance).

3.4.3 Disease surveillance on empirical human mobility networks

From Fig. 3.4, it is evident that the two graphs derived from empirical human mobility data have markedly different structural properties. Graph A, generated from aggregated mobility data derived from mobile phone trajectories in Italy at the provincial level (48), shows distinct community structures that closely resemble the SB graphs described in the previous section. In contrast, Graph B, generated from the global air traffic data collected

at the country level (49), displays structural properties similar to those of the BA graph. This is consistent with previous studies showing that the global air traffic network has scale-free properties (53, 54) [e.g., both have a negative degree assortativity (Fig. 3.4B), indicating a hub-and-spoke rather than hub-and-hub structure (55)].

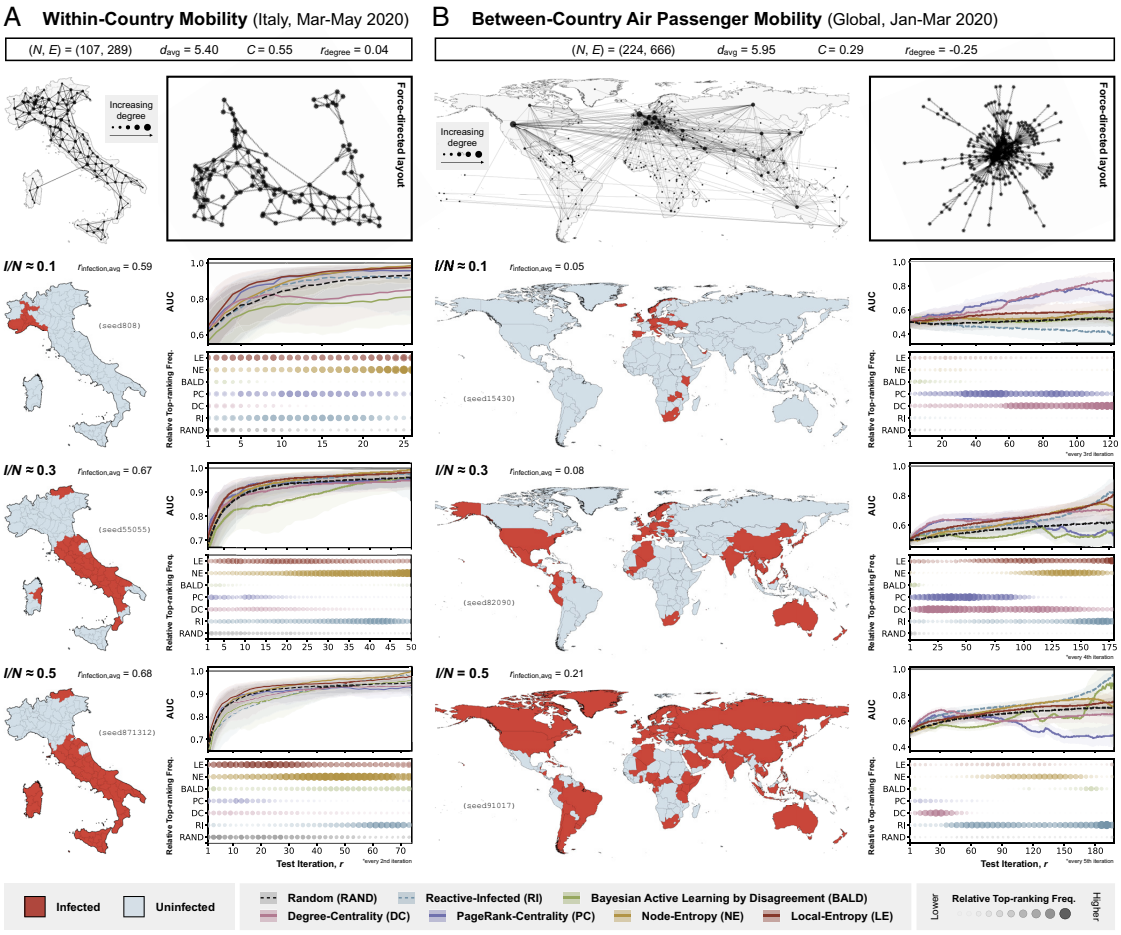


Fig. 3.4. Policy evaluation with simulated outbreaks on graphs derived from empirical human mobility data. Each panel (A and B) corresponds to results from experiments with simulated outbreaks on a graph derived from a different empirical human mobility dataset (panel A: within-country movements from smartphone data collected in Italy at the provincial level between March and May 2020 (48), with thinning threshold $T_{\text{thinning}} = 15\%$; panel B: between-country human movement from air traffic data collected between January and March 2020 (49), with thinning threshold $T_{\text{thinning}} = 5\%$). Summary statistics relevant to the structure of each graph [number of nodes (N), number of edges (E), average node degree (d_{avg}), clustering coefficient (C), and degree assortivity (r_{degree})] are shown at the top of each panel, followed by (left) a visualization of the graph overlaid on a corresponding map (with the size of each node indicating node degree) and (right) a visualization of the same

graph in a force-directed layout. The 2nd to 4th rows of each panel correspond to the different stages of outbreak progression at which the performance of the different policies is evaluated, as measured by the proportion of nodes infected ($I/N = 0.1, 0.3, 0.5$). In the left part of each row is the visualization of a selected disease distribution (from one of 50) on the corresponding map; tiles are coloured according to their true infection status (red if infected and blue if uninfected). In the right part of each panel, each column shows results from experiments considering disease distributions at a different stage of outbreak progression. The Top plot in each column shows the performance of policies considered in the corresponding experiment, as measured by the AUC, with a higher AUC indicating a better performance; the shaded region represents the interquartile range and the solid line represents the median. The Bottom plot in each column shows the frequency with which each policy is ranked top according to its AUC at each iteration (or every nth iteration, where indicated), normalised by the difference between the highest and lowest frequencies across different test budgets in the corresponding experiment (refer to Fig. B.8, and Tables B.5, B.6 in Appendix B for the absolute frequencies); a larger circle with a greater opacity indicates a higher frequency of the policy being ranked top at a given test iteration. The performance of each policy is only shown up to the median number of test iterations required for all infected nodes to be observed among agents assigned to Reactive-Infected (RI) (see Fig. B.4 in Appendix B for full results).

We observe that policy performances on Graphs A and B are similar to those from experiments on the SB graphs and BA graph, respectively. Most notably for Graph A, LE again shows rapid increases in model performance given small numbers of test iterations, only to be surpassed by NE at large r, as expected; this observation is consistent across different stages of outbreak progression (Fig. B.8 in Appendix B). For Graph B, graph-based policies (DC, PC) outperform uncertainty-based policies especially at small r, again consistent with results from experiments on the BA graph. However, the superior performance of these graph-based policies extends only to larger values of r if the outbreak under surveillance is in its early stages (i.e., $I/N = 0.1$); at later stages of outbreak progression, the performance of these policies decreases with further increases in r.

This counterintuitive observation can be explained by considering the changes in the distribution of the decision boundary between the infected and uninfected regions in

the graph during a transmission process. At the beginning of an outbreak, nodes that are centrally located are more likely to be infected early on due to their high degree of connectivity. This implies that most of the decision boundary between infected and uninfected regions can be found close to the central nodes, thus explaining the superior performance of graph-based policies which preferentially select nodes with high degree of centrality. As the outbreak progresses, the decision boundary shifts toward the periphery of the graph with the already infected central nodes acting as secondary hubs of the emerging pathogen. This results in a decrease in the performance of graph-based policies, as the central nodes continue to be targeted while the peripheral regions of the graph (where most heterogeneities in the disease distribution lie) remain largely unexplored. Note that a similar drop in the performance of PC (columns 2 and 3 in Fig. 3.3B) at large r during later stages of outbreak progression ($I/N = 0.3$ and $I/N = 0.5$) can also be observed.

The same reasoning can also potentially explain the unexpected superior performance of BALD at large r during later stages of an outbreak on the BA graph ($I/N = 0.5$) in Fig. 3.3B; see also middle- and right-panels in Fig. B.7 and Table B.4 in Appendix B), with most heterogeneities in disease distribution lying in the peripheral regions that are preferentially sampled by BALD. Whereas during the early stages of an outbreak, most heterogeneities in disease distribution are likely to be found in the central regions of a network, therefore resulting in the superior performance of graph-based policies (DC, PC) which target highly connected nodes and RI at small r ($I/N = 0.1$ in Fig. 3.3C and D; see also left-panel in Fig. B.7, and Tables B.2, B.3 in Appendix B), albeit with only modest top-ranking frequencies given the small number of observed nodes. More generally, provided that the number of infected nodes is sufficiently small and that they are confined to a small, local region of the graph, any policy for which there

is a high probability of selecting an infected node is likely to perform well compared to other policies, especially when given a small test budget.

3.5 Discussion

In this work, we investigated how a finite amount of testing resources should be allocated across a network of locations connected by mobility, in order to maximise the information gained about the underlying distribution of an infectious disease. We formulate this task as a node classification problem with active learning, with the objective of providing accurate assessment of where the disease is likely to be present or absent given a fixed test budget. We proposed a policy that selects nodes for testing according to a measure of the distance-weighted average entropy of the label predictions in the local neighbourhood of a given candidate node. We then evaluated and compared the performance of different policies, including our proposed policy, under a range of different outbreak scenarios and graph structures.

Our results show that in general there is not a single policy that performs optimally across all outbreak scenarios. Instead, the performance of a given policy depends on both the test budget available (relative to the size of the network) and the geometry of the underlying disease distribution, which is in turn determined by network structure and extent of the outbreak. For example, graph-based policies that target central nodes perform better than uncertainty-based policies when the underlying disease spread cannot be modelled with a high degree of accuracy and certainty, as is often the case during early stages of an outbreak when the aetiology is unknown. Conversely, uncertainty-based policies are typically more effective in highly ordered networks with well-defined community structures. In particular, with our proposed policy (Selection by Local Entropy) which considers graph-based uncertainties in its selection criterion, we were

able to show that more frequent exploration results in better performance given a small test budget, while targeting regions in the network with observed heterogeneous disease distribution (exploitation) is more favourable given a large test budget. Finally, we find that following an approach akin to contact tracing (selecting immediate neighbours of infected nodes) generally leads to inferior performance compared to other policies in terms of characterizing the overall disease distribution. A comprehensive assessment of the overall distribution could potentially allow for a more detailed study of the underlying transmission process (e.g., identifying drivers of spread by iteratively refitting prediction models of disease progression on a network), and provide an opportunity to improve the joint modelling of infectious diseases and sampling more generally.

It should be noted that, while we are able to obtain insights into how different policies are likely to behave under different scenarios, a quantitative assessment of their overall performance—and the extent to which one policy should be recommended over another given any outbreak—requires a more detailed and systematic examination across the various parameter spaces considered, which is beyond the scope of this work. Such assessments are particularly important in comparing the costs and benefits of different policies, especially when little is known about the transmission dynamics of the disease or when the underlying mobility network is unknown (56, 57); future studies should focus on developing appropriate evaluation metrics with consideration of relevant public health contexts and under more realistic model assumptions (see below).

Although we observe consistent results across experiments on both synthetic graphs and empirically derived networks, it is important to interpret these findings in the context of the assumptions made, particularly regarding their generalizability to real-world scenarios. A key limitation of our approach is the assumption that the underlying mobility network can be represented as an undirected and unweighted graph. In reality,

mobility networks are highly heterogeneous with mobility fluxes that vary across both regions (e.g., air traffic among European countries versus African countries) and directions (e.g., net inflow of air passengers arriving at tourist destinations during holiday seasons). This limitation is also relevant to infectious diseases with alternative modes of transmission (e.g., sexually transmitted diseases, vector-borne diseases), for which the network capturing the spatial correlation in disease distribution may involve factors other than human movement and cannot be adequately described by an undirected and unweighted graph. For example, in the case of a vector-borne disease, edges in the corresponding network might represent the absence of geographic barriers that prevent vector movement, with edge weights indicating the environmental suitability for vector survival both at the origin/destination location and during transit, which could be time-varying especially for climate-sensitive infectious diseases such as dengue (58–61) and malaria (62–64). Future extensions should consider alternative surrogate models that are able to incorporate these effects when generating label predictions, e.g., GNNs, Gaussian Processes on graphs (65, 66), and spatial mechanistic models that explicitly model the movement of infectious individuals.

Another limitation of this study is the assumption of static disease distributions. This implies that the timescale over which transmission events between locations occur is sufficiently longer than the timescale of test deployment, such that the underlying disease distribution can be treated as static. While this is unlikely to be a realistic assumption for most disease outbreaks—except for some endemic diseases that are more slowly changing in their prevalence, such as HIV/AIDS (67) and Tuberculosis (TB) (68) – it nevertheless allowed us to gain theoretical insights into the various factors one must consider when designing disease surveillance strategies given different network structures and outbreak scenarios. To address this limitation, future work should consider

the correlation in infection status not only between nodes but also across time, given either prior assumptions of the underlying transmission dynamics or information from historical transmission events that are inferred to have occurred given the data. In this dynamic setting, it might also be advisable to consider testing multiple locations at once [similar to batch AL (69)], as opposed to only a single location per iteration as presented in our work. Further, future work should also consider the incorporation of external time-series data (e.g., frequency of patients with specific symptoms, rate of hospitalization) and other data types (e.g., pathogen genomic data, wastewater data) that are independent of surveillance efforts and explore how such data can be used to inform test allocation. Finally, we assume here an idealized implementation of disease surveillance, with i) no observational noise (i.e., the true infection status of a selected node is always revealed upon testing); ii) equal access to testing resources at all nodes (i.e., there are no restrictions on which nodes can be selected for testing); and iii) equal cost per test across all nodes and times. However, in practice, i) the infection status of a location could be misclassified due to measurement error or low prevalence, leading to differences in test informativeness; ii) test deployment at certain locations may be hindered by logistical challenges and limitations in local infrastructures (24); and iii) testing cost may vary across both space and time due to factors such as local disease prevalence, geographical accessibility, staffing and operational expenses (e.g., higher personnel costs in remote or hazardous locations). Future studies should consider more realistic assumptions of how testing resources are deployed and how varying costs impact the design of allocation strategies (e.g., greater resources might be necessary in areas with low disease prevalence, larger populations, or regions with higher logistical and operational costs).

Our findings are relevant to infectious disease surveillance in resource-constrained settings and in situations where practical challenges render the complete

detection of all infected populations unfeasible or cost-inefficient. We propose a flexible and principled approach to evaluating the design and execution of adaptive surveillance strategies with the overall aim of maximizing the information gained from each round of testing. More generally, our adaptive test deployment framework can be extended to consider transmission processes with greater complexities (e.g., SEIR models, spatially explicit semi-mechanistic models, alternative transmission pathways) and more realistic mobility networks (e.g., as directed and weighted graphs, with time-varying edge weights and node attributes) that are derived from empirical data, and with additional constraints to account for imperfect testing (e.g., observational noise and delay in test feedback, presence of nodes that are inaccessible to surveillance efforts). Applications of our model in real-world contexts could provide the opportunity for more cost-effective and rapid identification and monitoring of pathogens while reducing the uncertainties associated with early risk assessments of infectious diseases.

3.6 References

1. Kucharski, A.J., Eggo, R.M., Watson, C.H., Camacho, A., Funk, S. and John Edmunds, W. (2016) 'Effectiveness of Ring Vaccination as Control Strategy for Ebola Virus Disease', *Emerging infectious diseases*, 22(1), p. 105.
2. Wells, C.R., Pandey, A., Parpia, A.S., Fitzpatrick, M.C., Meyers, L.A., Singer, B.H. and Galvani, A.P. (2019) 'Ebola vaccination in the Democratic Republic of the Congo', *Proceedings of the National Academy of Sciences of the United States of America*, 116(20), pp. 10178–10183.
3. Henao-Restrepo, A.M., Camacho, A., Longini, I.M., Watson, C.H., John Edmunds, W., Egger, M., Carroll, M.W., Dean, N.E., Diatta, I., Doumbia, M., Draguez, B., Duraffour, S., Enwere, G., Grais, R., Gunther, S., Gsell, P.-S., Hossmann, S., Watle, S.V., Kondé, M.K., Kéïta, S., Kone, S., Kuisma, E., Levine, M.M., Mandal, S., Mauget, T., Norheim, G., Riveros, X., Soumah, A., Trelle, S., Vicari, A.S., Røttingen, J.-A. and Kieny, M.-P. (2017) 'Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!)', *The Lancet*, 389(10068), pp. 505–518.
4. Wrigley-Field, E., Kiang, M.V., Riley, A.R., Barbieri, M., Chen, Y.-H., Duchowny, K.A., Matthay, E.C., Van Riper, D., Jegathesan, K., Bibbins-Domingo, K. and Leider, J.P. (2021) 'Geographically targeted COVID-19 vaccination is more equitable and averts more deaths than age-based thresholds alone', *Science Advances*, 7(40), p. eabj2099.
5. Keeling, M.J., Moore, S., Penman, B.S. and Hill, E.M. (2023) 'The impacts of SARS-CoV-2 vaccine dose separation and targeting on the COVID-19 epidemic in England', *Nature communications*, 14(1), pp. 1–10.
6. Keita, M., Polonsky, J.A., Ahuka-Mundeke, S., Ilumbulumbu, M.K., Dakissaga, A., Boiro, H., Anoko, J.N., Diassy, L., Ngwama, J.K., Bah, H., Tosalisana, M.K., Kitenge Omasumbu, R., Chérif, I.S., Boland, S.T., Delamou, A., Yam, A., Flahault, A., Dagron, S., Gueye, A.S., Keiser, O. and Fall, I.S. (2023) 'A community-based contact isolation strategy to reduce the spread of Ebola virus disease: an analysis of the 2018-2020 outbreak in the Democratic Republic of the Congo', *BMJ global health*, 8(6), p. e011907.
7. Chan, Y.H. and Nishiura, H. (2020) 'Estimating the protective effect of case isolation with transmission tree reconstruction during the Ebola outbreak in Nigeria, 2014', *Journal of the Royal Society, Interface / the Royal Society*, 17(169), 20200498.
8. Fang, L.-Q., Yang, Y., Jiang, J.-F., Yao, H.-W., Kargbo, D., Li, X.-L., Jiang, B.-G., Kargbo, B., Tong, Y.-G., Wang, Y.-W., Liu, K., Kamara, A., Dafae, F., Kanu, A., Jiang, R.-R., Sun, Y., Sun, R.-X., Chen, W.-J., Ma, M.-J., Dean, N.E., Thomas, H., Longini, I.M., Halloran, M.E. and Cao, W.-C. (2016) 'Transmission dynamics of Ebola virus disease and intervention effectiveness in Sierra Leone', *Proceedings of the National Academy of Sciences*, 113(16), pp. 4488–4493.
9. Auranen, K., Shubin, M., Erra, E., Isosomppi, S., Kontto, J., Leino, T. and Lukkarinen, T. (2023) 'Efficacy and effectiveness of case isolation and quarantine during a growing phase of the COVID-19 epidemic in Finland', *Scientific reports*, 13(1), p. 298.
10. Kucharski, A.J., Klepac, P., Conlan, A.J.K., Kissler, S.M., Tang, M.L., Fry, H., Gog, J.R., Edmunds, W.J. and CMMID COVID-19 working group (2020) 'Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission

of SARS-CoV-2 in different settings: a mathematical modelling study', *The Lancet infectious diseases*, 20(10), pp. 1151–1160.
11. Brauner, J.M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A.B., Leech, G., Altman, G., Mikulik, V., Norman, A.J., Monrad, J.T., Besiroglu, T., Ge, H., Hartwick, M.A., Teh, Y.W., Chindelevitch, L., Gal, Y. and Kulveit, J. (2021) 'Inferring the effectiveness of government interventions against COVID-19', *Science*, 371(6531), p. eabd9338.
12. Flaxman, S., Mishra, S., Gandy, A., Unwin, H.J.T., Mellan, T.A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J.W., Monod, M., Ghani, A.C., Donnelly, C.A., Riley, S., Vollmer, M.A.C., Ferguson, N.M., Okell, L.C. and Bhatt, S. (2020) 'Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe', *Nature*, 584(7820), pp. 257–261.
13. Halvorsen, G.S., Simonsen, L. and Sneppen, K. (2022) 'Spatial model of Ebola outbreaks contained by behavior change', *PloS one*, 17(3), p. e0264425.
14. Funk, S., Ciglenecki, I., Tiffany, A., Gignoux, E., Camacho, A., Eggo, R.M., Kucharski, A.J., John Edmunds, W., Bolongei, J., Azuma, P., Clement, P., Alpha, T.S., Sterk, E., Telfer, B., Engel, G., Parker, L.A., Suzuki, M., Heijenberg, N. and Reeder, B. (2017) 'The impact of control strategies and behavioural changes on the elimination of Ebola from Lofa County, Liberia', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 372(1721), 20160302.
15. Tegally, H., Wilkinson, E., Tsui, J.L., Moir, M., Martin, D., Brito, A.F., Giovanetti, M., Khan, K., Huber, C., Bogoch, I.I., San, J.E., Poongavanan, J., Xavier, J.S., Candido, D.D.S., Romero, F., Baxter, C., Pybus, O.G., Lessells, R.J., Faria, N.R., Kraemer, M.U.G. and de Oliveira, T. (2023) 'Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern', *Cell*, 186(15), pp. 3277-3290.e16
16. Tsui, J.L.-H., McCrone, J.T., Lambert, B., Bajaj, S., Inward, R.P.D., Bosetti, P., Pena, R.E., Tegally, H., Hill, V., Zarebski, A.E., Peacock, T.P., Liu, L., Wu, N., Davis, M., Bogoch, I.I., Khan, K., Kall, M., Abdul Aziz, N.I.B., Colquhoun, R., O'Toole, Á., Jackson, B., Dasgupta, A., Wilkinson, E., de Oliveira, T., COVID-19 Genomics UK (COG-UK) consortium¶, Connor, T.R., Loman, N.J., Colizza, V., Fraser, C., Volz, E., Ji, X., Gutierrez, B., Chand, M., Dellicour, S., Cauchemez, S., Raghwani, J., Suchard, M.A., Lemey, P., Rambaut, A., Pybus, O.G. and Kraemer, M.U.G. (2023) 'Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1', *Science*, 381(6655), pp. 336–343.
17. Grubaugh, N.D., Saraf, S., Gangavarapu, K., Watts, A., Tan, A.L., Oidtman, R.J., Ladner, J.T., Oliveira, G., Matteson, N.L., Kraemer, M.U.G., Vogels, C.B.F., Hentoff, A., Bhatia, D., Stanek, D., Scott, B., Landis, V., Stryker, I., Cone, M.R., Kopp, E.W., IV, Cannons, A.C., Heberlein-Larson, L., White, S., Gillis, L.D., Ricciardi, M.J., Kwal, J., Lichtenberger, P.K., Magnani, D.M., Watkins, D.I., Palacios, G., Hamer, D.H., GeoSentinel Surveillance Network, Gardner, L.M., Alex Perkins, T., Baele, G., Khan, K., Morrison, A., Isern, S., Michael, S.F. and Andersen, K.G. (2019) 'Travel Surveillance and Genomics Uncover a Hidden Zika Outbreak during the Waning Epidemic', *Cell*, 178(5), p. 1057.
18. Hoffman, S.J. and Silverberg, S.L. (2018) 'Delays in Global Disease Outbreak Responses: Lessons from H1N1, Ebola, and Zika', *American journal of public health*, 108(3), pp. 329–333.
19. Riley, S., Atchison, C., Ashby, D., Donnelly, C.A., Barclay, W., Cooke, G.S., Ward, H., Darzi, A., Elliott, P. and REACT study group (2020) 'REal-time Assessment of

Community Transmission (REACT) of SARS-CoV-2 virus: Study protocol', *Wellcome open research*, 5, p. 200.
20. Centers for Disease Control and Prevention (no date) *National notifiable diseases surveillance system (NNDSS)*. Available at https://www.cdc.gov/nndss/index.html (Accessed: 31 May 2024).
21. World Health Organization (2024) *Global influenza surveillance and response system (GISRS).* Available at https://www.who.int/initiatives/global-influenza-surveillance-and-response-system (Accessed 31 May 2024).
22. Zhou, S., Feng, X., Hu, Y., Yang, J., Chen, Y., Bastow, J., Zheng, Z.-J. and Xu, M. (2023) 'Factors associated with the utilization of diagnostic tools among countries with different income levels during the COVID-19 pandemic', *Global health research and policy*, 8(1), p. 45.
23. Han, A.X., Toporowski, A., Sacks, J.A., Perkins, M.D., Briand, S., van Kerkhove, M., Hannay, E., Carmona, S., Rodriguez, B., Parker, E., Nichols, B.E. and Russell, C.A. (2023) 'SARS-CoV-2 diagnostic testing rates determine the sensitivity of genomic surveillance programs', *Nature genetics*, 55(1), pp. 26–33.
24. Brito, A.F., Semenova, E., Dudas, G., Hassler, G.W., Kalinich, C.C., Kraemer, M.U.G., Ho, J., Tegally, H., Githinji, G., Agoti, C.N., Matkin, L.E., Whittaker, C., Howden, B.P., Sintchenko, V., Zuckerman, N.S., Mor, O., Blankenship, H.M., de Oliveira, T., Lin, R.T.P., Siqueira, M.M., Resende, P.C., Vasconcelos, A.T.R., Spilki, F.R., Aguiar, R.S., Alexiev, I., Ivanov, I.N., Philipova, I., Carrington, C.V.F., Sahadeo, N.S.D., Branda, B., Gurry, C., Maurer-Stroh, S., Naidoo, D., von Eije, K.J., Perkins, M.D., van Kerkhove, M., Hill, S.C., Sabino, E.C., Pybus, O.G., Dye, C., Bhatt, S., Flaxman, S., Suchard, M.A., Grubaugh, N.D., Baele, G. and Faria, N.R. (2022) 'Global disparities in SARS-CoV-2 genomic surveillance', *Nature communications*, 13(1), pp. 1–13.
25. Pei, S., Teng, X., Lewis, P. and Shaman, J. (2021) 'Optimizing respiratory virus surveillance networks using uncertainty propagation', *Nature communications*, 12(1), pp. 1–10.
26. Bajardi, P., Barrat, A., Savini, L. and Colizza, V. (2012) 'Optimizing surveillance for livestock disease spreading through animal movements', *Journal of the Royal Society, Interface / the Royal Society*, 9(76), pp. 2814–2825.
27. Ansari, S., Heitzig, J. and Moosavi, M.R. (2023) 'Optimizing testing strategies for early detection of disease outbreaks in animal trade networks via MCMC', *Chaos* , 33(4), 043144.
28. Beregi, S. and Parag, K.V. (2024) 'Optimal algorithms for controlling infectious diseases in real time using noisy infection data', *medRxiv*. Available at: https://doi.org/10.1101/2024.05.24.24307878 (Accessed: 20 June 2024).
29. Xia, M., Bottcher, L. and Chou, T. (2022) 'Controlling epidemics through optimal allocation of test kits and vaccine doses across networks', *IEEE transactions on network science and engineering*, 9(3), pp. 1422–1436.
30. Meirom, E.A., Maron, H., Mannor, S. and Chechik, G. (2020) 'Controlling Graph Dynamics with Reinforcement Learning and Graph Neural Networks', arXiv. Available at: http://arxiv.org/abs/2010.05313 (Accessed: 24 May 2024).
31. Du, Z., Pandey, A., Bai, Y., Fitzpatrick, M.C., Chinazzi, M., Pastore Y Piontti, A., Lachmann, M., Vespignani, A., Cowling, B.J., Galvani, A.P. and Meyers, L.A. (2021) 'Comparative cost-effectiveness of SARS-CoV-2 testing strategies in the USA: a modelling study', *The Lancet. Public health*, 6(3), pp. e184–e191.
32. Grassly, N.C., Pons-Salort, M., Parker, E.P.K., White, P.J., Ferguson, N.M. and Imperial College COVID-19 Response Team (2020) 'Comparison of molecular

testing strategies for COVID-19 control: a mathematical modelling study', *The Lancet infectious diseases*, 20(12), pp. 1381–1389.
33. Stenseth, N.C., Schlatte, R., Liu, X., Pielke, R., Jr, Li, R., Chen, B., Bjørnstad, O.N., Kusnezov, D., Gao, G.F., Fraser, C., Whittington, J.D., Bai, Y., Deng, K., Gong, P., Guan, D., Xiao, Y., Xu, B. and Johnsen, E.B. (2023) 'How to avoid a local epidemic becoming a global pandemic', *Proceedings of the National Academy of Sciences of the United States of America*, 120(10), p. e2220080120.
34. Zhang, D., Ge, Y., Wang, J., Liu, H., Zhang, W.-B., Wu, X., Heuvelink, G.B.M., Wu, C., Yang, J., Ruktanonchai, N.W., Qader, S.H., Ruktanonchai, C.W., Cleary, E., Yao, Y., Liu, J., Nnanatu, C.C., Wesolowski, A., Cummings, D.A.T., Tatem, A.J. and Lai, S. (2024) 'Optimizing the detection of emerging infections using mobility-based spatial sampling', *International Journal of Applied Earth Observation and Geoinformation, 131, 103949.*
35. Spott, R., Pletz, M.W., Fleischmann-Struzek, C., Kimmig, A., Hadlich, C., Hauert, M., Lohde, M., Jundzill, M., Marquet, M., Dickmann, P., Schüchner, R., Hölzer, M., Kühnert, D. and Brandt, C. (2024) 'Exploring the Spatial Distribution of Persistent SARS-CoV-2 Mutations - Leveraging mobility data for targeted sampling', *eLife*, 13, RP94045.
36. Oliveira, J.F., Alencar, A.L., Mcls, C., Vasconcelos, A.O., Cunha, G.G., Miranda, R.B., Fmhs, F., Silva, C., Gustani-Buss, E., Khouri, R., Cerqueira-Silva, T., Landau, L., Barral-Netto, M. and Ramos, P.I.P. (2024) 'Human mobility patterns in Brazil to inform sampling sites for early pathogen detection and routes of spread: a network modelling and validation study', *The Lancet. Digital health*, 6(8), pp. e570-e579.
37. Settles, B. and Craven, M.W. (2008) 'An analysis of active learning strategies for sequence labeling tasks', *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, Honolulu, Hawaii: Association for Computational Linguistics, pp. 1070–1079.
38. Garnett, R., Krishnamurthy, Y., Xiong, X., Schneider, J.G. and Mann, R.P. (2012) 'Bayesian optimal active search and surveying', *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*. Available at: https://arxiv.org/abs/1206.6406 (Accessed: 14 July 2025).
39. Besag, J., York, J. and Mollié, A. (1991) 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the Institute of Statistical Mathematics*, 43(1), pp. 1–20.
40. Lewis, D.D. and Catlett, J. (1994) 'Heterogeneous uncertainty sampling for supervised learning', in Cohen, W.W. and Hirsh, H. (eds.) *Machine Learning Proceedings 1994*. San Francisco, CA: Morgan Kaufmann, pp. 148–156.
41. Houlsby, N., Huszár, F., Ghahramani, Z. and Lengyel, M. (2011) 'Bayesian Active Learning for Classification and Preference Learning', *arXiv*. Available at: https://arxiv.org/pdf/1112.5745.pdf (Accessed: 24 May 2024).
42. Brin, S. and Page, L. (1998) 'The anatomy of a large-scale hypertextual Web search engine' *Computer Networks and ISDN Systems*, 30(1-7), pp. 107–117.
43. Kipf, T.N. and Welling, M. (2017) 'Semi-supervised classification with graph convolutional networks', *arXiv*. Available at: https://doi.org/10.48550/arXiv.1609.02907 (Accessed: 24 May 2024).
44. Madhawa, K. and Murata, T. (2020b) 'MetAL: Active Semi-Supervised Learning on Graphs via Meta-Learning', in *Asian Conference on Machine Learning*. *Asian Conference on Machine Learning*, PMLR, pp. 561–576.
45. Cai, H., Zheng, V. and Chang, K. (2017) 'Active Learning for Graph Embedding', *arXiv*. Available at: https://arxiv.org/pdf/1705.05085.pdf (Accessed: 24 May 2024).

46. Barabási, A.-L. and Albert, R. (1999) ‘Emergence of Scaling in Random Networks’, *Science*, 286(5439), pp. 509-512.
47. Holland, P.W., Laskey, K.B. and Leinhardt, S. (1983) ‘Stochastic blockmodels: First steps’, *Social Networks*, 5(2), pp. 109–137.
48. Pepe, E., Bajardi, P., Gauvin, L., Privitera, F., Lake, B., Cattuto, C. and Tizzoni, M. (2020) ‘COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown’, *Scientific Data*, 7(1), pp. 1–7.
49. Rudolf, S. (2022) *Source for ‘How to avoid a local epidemic becoming a global pandemic?’* Zenodo. Available at: https://doi.org/10.5281/ZENODO.7472836 (Accessed: 3 March 2024).
50. Madhawa, K. and Murata, T. (2020a) ‘Active Learning for Node Classification: An Evaluation’, *Entropy* , 22(10), p. 1164.
51. Viboud, C., Bjørnstad, O.N., Smith, D.L., Simonsen, L., Miller, M.A. and Grenfell, B.T. (2006) ‘Synchrony, waves, and spatial hierarchies in the spread of influenza’, *Science*, 312(5772), pp. 447–451.
52. Newman, M.E.J. (2006) ‘Modularity and community structure in networks’, *Proceedings of the National Academy of Sciences*, 103(23), pp. 8577–8582.
53. Guimerà, R., Mossa, S., Turtschi, A. and Amaral, L.A.N. (2005) ‘The worldwide air transportation network: Anomalous centrality, community structure, and cities’ global roles’, *Proceedings of the National Academy of Sciences of the United States of America*, 102(22), pp. 7794–7799.
54. Diop, I.M., Cherifi, C., Diallo, C. and Cherifi, H. (2021) ‘Revealing the component structure of the world air transportation network’, *Applied network science*, 6(1), p. 92.
55. Sun, X., Wandelt, S. and Zhang, A. (2020) ‘How did COVID-19 impact air transportation? A first peek through the lens of complex networks’, *Journal of Air Transport Management*, 89, p. 101928.
56. Ramiadantsoa, T., Metcalf, C.J.E., Raherinandrasana, A.H., Randrianarisoa, S., Rice, B.L., Wesolowski, A., Randriatsarafara, F.M. and Rasambainarivo, F. (2022) ‘Existing human mobility data sources poorly predicted the spatial spread of SARS-CoV-2 in Madagascar’, *Epidemics*, 38, p. 100534.
57. Kraemer, M.U.G., Golding, N., Bisanzio, D., Bhatt, S., Pigott, D.M., Ray, S.E., Brady, O.J., Brownstein, J.S., Faria, N.R., Cummings, D.A.T., Pybus, O.G., Smith, D.L., Tatem, A.J., Hay, S.I. and Reiner, R.C., Jr (2019) ‘Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings’, *Scientific reports*, 9(1), p. 5151.
58. Xu, L., Stige, L.C., Chan, K.-S., Zhou, J., Yang, J., Sang, S., Wang, M., Yang, Z., Yan, Z., Jiang, T., Lu, L., Yue, Y., Liu, X., Lin, H., Xu, J., Liu, Q. and Stenseth, N.C. (2017) ‘Climate variation drives dengue dynamics’, *Proceedings of the National Academy of Sciences of the United States of America*, 114(1), pp. 113–118.
59. Trejo, I., Barnard, M., Spencer, J.A., Keithley, J., Martinez, K.M., Crooker, I., Hengartner, N., Romero-Severson, E.O. and Manore, C. (2023) ‘Changing temperature profiles and the risk of dengue outbreaks’, *PLOS Climate*, 2(2), p. e0000115.
60. Mordecai, E.A., Cohen, J.M., Evans, M.V., Gudapati, P., Johnson, L.R., Lippi, C.A., Miazgowicz, K., Murdock, C.C., Rohr, J.R., Ryan, S.J., Savage, V., Shocket, M.S., Stewart Ibarra, A., Thomas, M.B. and Weikel, D.P. (2017) ‘Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models’, *PLoS neglected tropical diseases*, 11(4), p. e0005568.

61. Watts, D.M., Burke, D.S., Harrison, B.A., Whitmire, R.E. and Nisalak, A. (1987) 'Effect of temperature on the vector efficiency of Aedes aegypti for dengue 2 virus', *The American journal of tropical medicine and hygiene*, 36(1), pp. 143–152.
62. Arab, A., Jackson, M.C. and Kongoli, C. (2014) 'Modelling the effects of weather and climate on malaria distributions in West Africa', *Malaria journal*, 13, p. 126.
63. Wang, Z., Liu, Y., Li, Y., Wang, G., Lourenço, J., Kraemer, M., He, Q., Cazelles, B., Li, Y., Wang, R., Gao, D., Li, Y., Song, W., Sun, D., Dong, L., Pybus, O.G., Stenseth, N.C. and Tian, H. (2022) 'The relationship between rising temperatures and malaria incidence in Hainan, China, from 1984 to 2010: a longitudinal cohort study', *The Lancet. Planetary health*, 6(4), pp. e350–e358.
64. Santos-Vega, M., Martinez, P.P., Vaishnav, K.G., Kohli, V., Desai, V., Bouma, M.J. and Pascual, M. (2022) 'The neglected role of relative humidity in the interannual variability of urban malaria in Indian cities', *Nature communications*, 13(1), p. 533.
65. Zhi, Y.-C., Ng, Y.C. and Dong, X. (2020) 'Gaussian Processes on Graphs via Spectral Kernel Learning', arXiv. Available at https://doi.org/10.48550/arXiv.2006.07361 (Accessed: 13 October 2024).
66. Borovitskiy, V., Azangulov, I., Terenin, A., Mostowsky, P., Deisenroth, M.P. and Durrande, N. (2021) 'Matern gaussian processes on graphs', *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)* (PMLR, 2020), vol. 130, pp. 2593–2601.
67. Dwyer-Lindgren, L., Cork, M.A., Sligar, A., Steuben, K.M., Wilson, K.F., Provost, N.R., Mayala, B.K., VanderHeide, J.D., Collison, M.L., Hall, J.B., Biehl, M.H., Carter, A., Frank, T., Douwes-Schultz, D., Burstein, R., Casey, D.C., Deshpande, A., Earl, L., El Bcheraoui, C., Farag, T.H., Henry, N.J., Kinyoki, D., Marczak, L.B., Nixon, M.R., Osgood-Zimmerman, A., Pigott, D., Reiner, R.C., Ross, J.M., Schaeffer, L.E., Smith, D.L., Davis Weaver, N., Wiens, K.E., Eaton, J.W., Justman, J.E., Opio, A., Sartorius, B., Tanser, F., Wabiri, N., Piot, P., Murray, C.J.L. and Hay, S.I. (2019) 'Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017', *Nature*, 570(7760), pp. 189–193.
68. Bie, S., Hu, X., Zhang, H., Wang, K. and Dou, Z. (2021) 'Influential factors and spatial–temporal distribution of tuberculosis in mainland China', *Scientific reports*, 11(1), pp. 1–8.
69. Hoi, S.C.H., Jin, R., Zhu, J. and Lyu, M.R. (2006) 'Batch mode active learning and its application to medical image classification', in *Proceedings of the 23rd international conference on Machine learning (ICML '06)*, New York, New York, USA: ACM Press, pp. 417-424.
70. Tsui, J.L.-H. (2024) *joetsui1994/towards-optimal-disease-surveillance-AL: 15 October 2024 - v0.1 Release (v1-release)*. Zenodo. Available at https://doi.org/10.5281/zenodo.13935694. (Deposited: 15 October 2024).

Appendix B:
Supplementary materials for Chapter 3

B.1 Stochastic susceptible-infected process on graphs

Each node is initially in the susceptible (S) state. At $t = 0$, a single node is randomly selected and set to the infected (I) state. At each following time step where $t > 0$, we assume that there is a fixed probability $p = 0.1$ for any infected node to infect their susceptible neighbors, i.e. any nodes that are in the susceptible state and are connected to the infected node by an edge. In this simple SI process, it is assumed that infected nodes do not recover or become immune - once infected, they remain infected indefinitely with the same constant probability of onward transmission, p, throughout the rest of the simulation. This process continues until a certain proportion of nodes are infected, as specified by a target I/N value, where I is the number infected nodes and N is the total number of nodes in the graph. Different values of I/N indicate different stages of outbreak progression at the time of disease surveillance.

An important implication resulting from the assumptions made in the SI process as described, is that a node can only be infected if at least one of its immediate neighbours is also infected (with the exception of the initially infected node), i.e. all infected nodes must be connected in the graph. This implies that there can only be a single infected region, however with potentially multiple uninfected regions and therefore multiple decision-boundaries (lines or surfaces separating infected and uninfected regions). The distribution of these decision-boundaries in a network varies between outbreaks depending on both the network structure and the stage of outbreak progression (proportion of nodes that are infected).

B.2 Conditional Autoregressive Model (CAR) as a surrogate model

The Conditional Autoregressive (CAR) model (1) is widely used in the small area estimation domain, where data typically consist of observations $\boldsymbol{y} = [y_1, y_2, \dots, y_n]$ over a set of n spatial units, which in the context of our study represent the observed infection status of a subset of locations in a mobility network. The CAR model assumes that the value of a variable at one location (e.g., infection status) depends on the values at neighbouring locations, with weights specified by a spatial adjacency matrix $\boldsymbol{A}$. For unweighted models, such as the one we work with in this paper, the adjacency matrix $\boldsymbol{A}$ is binary, capturing the presence or absence of edges between corresponding nodes. The spatial random effect $\boldsymbol{f} = [f_1, f_2, \dots, f_n]$ follows a multivariate normal prior with precision matrix $\boldsymbol{Q}$:

$$f \sim \mathcal{N}(0, \boldsymbol{Q}^{-1}) \tag{1}$$

$$\boldsymbol{Q} = \tau(\boldsymbol{D} - \alpha\boldsymbol{A}) \tag{2}$$

where $\boldsymbol{D}$ is a diagonal matrix with diagonal elements corresponding to the number of neighbours each location has, and τ is the scale parameter with prior $\tau \sim \text{logNormal}(0, 0.1)$. The parameter α captures the amount of spatial correlation between connected locations and can take any value between -1 (perfect heterophily, i.e. neighbouring nodes have opposite labels) and 1 (perfect homophily, i.e. neighbouring nodes have the same labels). In all experiments, we set $\alpha = 0.95$ to reflect our prior assumption of a strong spatial correlation in disease distribution (across a mobility network), consistent with assumptions frequently made in real-world outbreak investigations and observations from numerous empirical studies, notably those examining the spread of respiratory diseases, such as H1N1 and SARS-CoV-2 across the global air traffic network (2–4). We also set $\alpha < 1$ to retain a proper CAR prior ($\alpha = 1$ yields an intrinsic CAR prior with a singular matrix) to prevent over-smoothing of sharp

decision-boundaries, as well as to avoid numerical instability and poor MCMC convergence associated with near-singular precision matrices Q as $\alpha \to 1$. Given the primary aim of this study is to compare the relative performance of different allocation policies under a fixed surrogate model, the exploration of how different assumptions of spatial dependence (e.g., smaller α values or α as a learned parameter) would impact policy performance is left as an important direction for future work.

Much like Gaussian Processes (GPs) (5, 6) are a standard model of choice for continuous space modelling, the CAR model is a default choice for spatial statistics over a discrete areas. Future work should explore a wider range of surrogates, such as label propagation (7), Graph Neural Networks (GNNs) (8), and GPs on graphs when no knowledge about the spread of the disease is available, or spatially explicit mechanistic models like SIR and SEIR when the underlying transmission mechanics are known.

B.3 Bayesian Active Learning by Disagreement (BALD)

Bayesian Active Learning by Disagreement (BALD) (9) is one of the state-of-the-art acquisition policies in active learning. It selects the data instances that maximise the decrease in expected posterior entropy,

$$v_{r+1} = \underset{v \in V}{\operatorname{argmax}}\, I(\theta; y|v, \boldsymbol{D_r}) \tag{3}$$

where θ is the latent parameters and $\boldsymbol{D_r}$ is the set of data instances labelled up to iteration r, and the mutual information, I, is defined as follows:

$$
\begin{aligned}
I(\theta; y|v, \boldsymbol{D_r}) &= \mathrm{H}(\theta|\boldsymbol{D_r}) - \mathbb{E}_{y\sim p(y|v,\boldsymbol{D_r})}\mathrm{H}(\theta|y, v, \boldsymbol{D_r}) \\
&= \mathrm{H}(y|v, \boldsymbol{D_r}) - \mathbb{E}_{\theta\sim p(\theta|\boldsymbol{D_r})}\mathrm{H}(y|\theta, v, \boldsymbol{D_r}) \\
&= \mathrm{H}[\int p(y|v,\theta)p(\theta|\boldsymbol{D_r})d\theta] - \int \mathrm{H}[p(y|v,\theta)]p(\theta|\boldsymbol{D_r})d\theta \\
&\approx \mathrm{H}\left[\frac{1}{n}\sum_{i=1}^{n} p(y|v,\theta_i)\right] - \frac{1}{n}\sum_{i=1}^{n}\mathrm{H}[p(y|v,\theta_i)] \qquad (4)
\end{aligned}
$$

where $\theta_i \sim p(\theta|\boldsymbol{D_r})$.

For Gaussian Process classification tasks, Houlsby et al. (2011) (9) provided approximations of BALD. This formulation highlights that the mutual information can be approximated using posteriors obtained numerically. Hence, one can use surrogates of any complexity as long as their parameters can be estimated in a Bayesian manner.

B.4 Generating random graphs with community structure using the stochastic block model

We used the stochastic block (SB) model (10) to generate random graphs with different levels of community structure. We began by first specifying the number of communities, k, and the size of each community. In this study, we set k to 5 with the size of each community selected at random while keeping the total number of nodes in the graph at 160. To control the level of community structure, we varied the value of parameters p_{intra} and p_{inter}, i.e. the probability of connection within a community and between communities, respectively. For example, a high p_{intra} with a low p_{inter} indicates a strong community structure, with nodes within communities being tightly connected and only sparse connections between communities. To generate a random graph with a high level of community structure, we set the parameters to $(p_{intra}, p_{inter}) = (0.14, 0.001)$; and to generate a random graph with a lower level of community structure, we set $(p_{intra}, p_{inter}) = (0.08, 0.005)$.

One common way to quantify the level of community structure present in a graph is to compute its modularity (11). The modularity of a graph is a measure of the degree to which it can be partitioned into distinct modules or communities; it is defined as the fraction of the edges that fall within communities minus the expected fraction of edges that would fall within communities if edges were distributed randomly. Give a graph with adjacency matrix $\boldsymbol{A}$, its modularity is given by

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{10}$$

where m is the total number of edges in the graph, k_i and k_j are the degrees of node i and j, and $\delta(c_i, c_j)$ is 1 if node i and j are in the same community, and 0 otherwise.

Applying the above formula shows that the random graph generated using the SB model with a high level of community structure has a modularity of 0.72, whereas the random graph with a lower level of community structure has a modularity of 0.62.

B.5 Pre-processing of within-country human mobility data collected at provincial level in Italy

A dataset containing daily aggregated mobility data collected from mobile phone users in Italy at provincial level, covering the period between 18 January and 26 June 2020 (12), was downloaded from https://covid19mm.github.io/data.html on 26 February 2024. The data consists of the daily number of smartphone users moving both within and between 107 provinces, normalized by the number of active users each week, which has been shown to be roughly constant throughout the collection period. Here we focus our analysis on the period between March and May (inclusive), during which a national lockdown (from 9 March to 18 May, 2020) was imposed by the Italian government in response to the emerging COVID-19 outbreak.

To construct a static graph from the mobility data, we first summed the mobility flows over both directions for each pair of provinces to obtain a symmetric matrix for each day, which was then averaged across the 3-month period. The resulting matrix was then converted into an unweighted graph using a procedure known as graph-thinning. In this process, edges representing pairs of provinces were ranked according to their total mobility flow as calculated earlier; edges were then removed one at a time starting from those ranked the lowest while ensuring that the graph remained connected. This iterative process continued until a certain target proportion of the original edges remained; this target proportion is known as the thinning-threshold. Finally, all edge weights are removed.

The choice of this threshold takes into consideration the balance between 1) the need to remove edges with very low mobility flows and are therefore less relevant to the overall structure of the graph, versus 2) the need to retain enough edges in order to preserve important structural properties (e.g., presence of travel hubs and community structure) of the graph. With these in mind, the thinning-thresholds of 10%, 15% and 20% were specified. To ensure robustness, the same experiments were repeated on each of the resulting graphs (see Fig. B.5); however, only results from experiments performed on the graph with a thinning-threshold of 15% are presented in Fig. 3.4.

B.6 Pre-processing of between-country air traffic data collected at country level

A dataset containing monthly air traffic data collected at country level, covering the period between January and March 2020 (13), was downloaded from https://zenodo.org/records/7472836 on 1 March 2024. The data consists of the monthly number of air passengers travelling both within and between countries. To construct an

undirected and unweighted graph from the data, the same procedure as described in Section B.5 was performed. Due to the much greater number of edges (as a result of a greater number of nodes and the presence of long-range movements in the air traffic network), a lower thinning-threshold was used to ensure the surrogate model can be fitted within a reasonable timeframe at each iteration given the available computational resources. With the considerations as described in Section B.5, the thinning-thresholds of 2.5%, 5% and 7.5% were specified. Again, the same experiments were repeated on each of the resulting graphs to ensure robustness of our results (see Fig. B.6); however, only results from experiments performed on the graph with a thinning-threshold of 5% are presented in Fig. 3.4.

B.7 Degree-assortativity and infection-assortativity

Degree-assortativity of a network, commonly denoted as r_{degree}, is a measure of the tendency for nodes to connect with other nodes with similar degrees. It can take any value between -1 and 1, with a positive value indicating that high-degree nodes are more likely to connect with other high-degree nodes, and similarly for low-degree nodes (assortative mixing by degree). Conversely, a negative value indicates a tendency for high-degree nodes to connect with low-degree nodes, and vice versa (disassortative mixing by degree).

The same idea of assortativity can be extended to other node attributes, including infection status as considered in this study. A positive assortativity by infection status (referred to as infection-assortativity hereafter) indicates a tendency for infected nodes to connect with other infected nodes, and similarly for uninfected nodes (assortative mixing by infection status). We denote the infection-assortativity of a graph with a given underlying disease distribution as $r_{\text{infection}}$.

For a graph with an underlying disease distribution generated by a stochastic SI process (see Section B.1), we generally expect to observe a positive $r_{\text{infection}}$, since a node can only be infected if at least one of its immediate neighbours is also infected. The exact value of $r_{\text{infection}}$ however depends on both the graph structure and the stage of outbreak progression (proportion of nodes infected) (see Figs. 3.3 and 3.4).

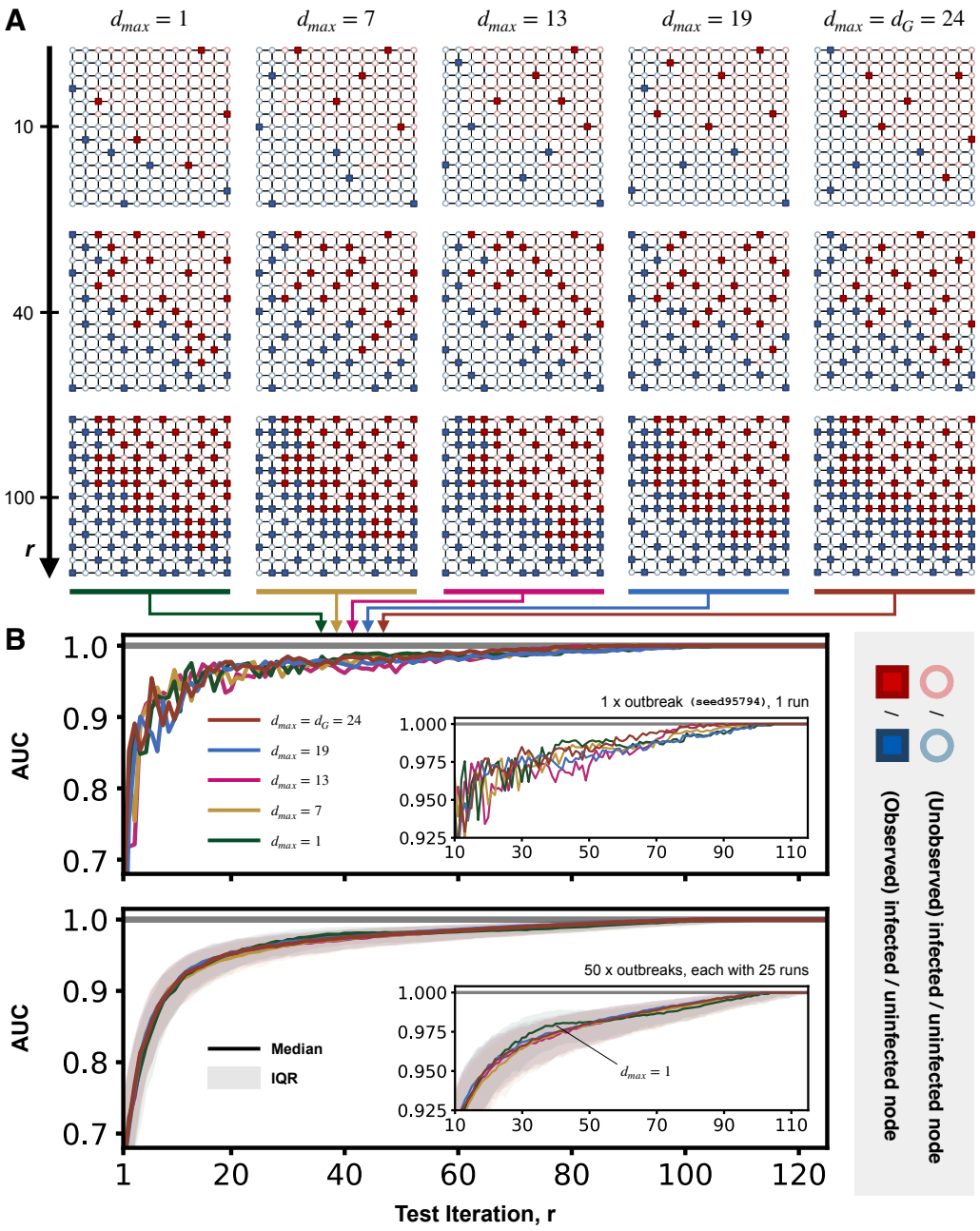


Fig. B.1. Impact of varying d_{max} in Selection by Local-Entropy. (A) Test allocation by five selected agents, each assigned a different policy (LE with $d_{max} = 1, 7, 13, 19, 24$, all with $\lambda = 0$). Each square panel shows the distribution of observed (squares) and unobserved (circles) nodes up to test iterations $r = 10, 40, 100$. Each node is coloured according to its true infection status (red if infected and blue if uninfected, with circles that represent unobserved nodes having a lower opacity). (B) The Top plot shows the performance of the five selected agents for a single outbreak realization, as measured by the AUC. The Bottom plot shows the average performance of the five LE policies with different d_{max} values, each summarised across 1,250 agents (50 outbreak realizations, each with 25 unique initial labelled nodes); the shaded region represents the interquartile range and the solid line represents the median.

Notably, the choice of d_{max} has relatively effect on the performance of LE except for $d_{max} = 1$ (green line), where we observe a slightly higher average AUC at intermediate test iterations ($30 \leq r \leq 50$). This can be explained by the smaller d_{max}-hop neighbourhood of each candidate node (within which the node entropies of surrounding nodes are considered in the LE selection criterion; see Eq. 1 in Chapter 3), which allows the decision-boundaries to be sampled more closely. However, this only persists until there are no remaining candidate nodes along the decision-boundary with unobserved neighbours (therefore with both high Ω_k^{self} and Ω_k^{surr}), after which the policy resumes sampling from relatively unexplored regions, leading to similar performance as other LE policies with $d_{max} > 1$ at larger test iterations. Larger d_{max} values lead to greater sparsity in test allocation, with patterns that resemble parallel diagonal lines that are evenly spaced. These patterns can be explained by the geometry of the d_{max}-hop neighbourhood of each candidate node, which has the shape of a diamond as a result of the geodesic distance between nodes being determined by the Manhattan distance in a lattice graph; however, this effect diminishes as d_{max} increases and approaches the graph diameter $d_G = 24$.

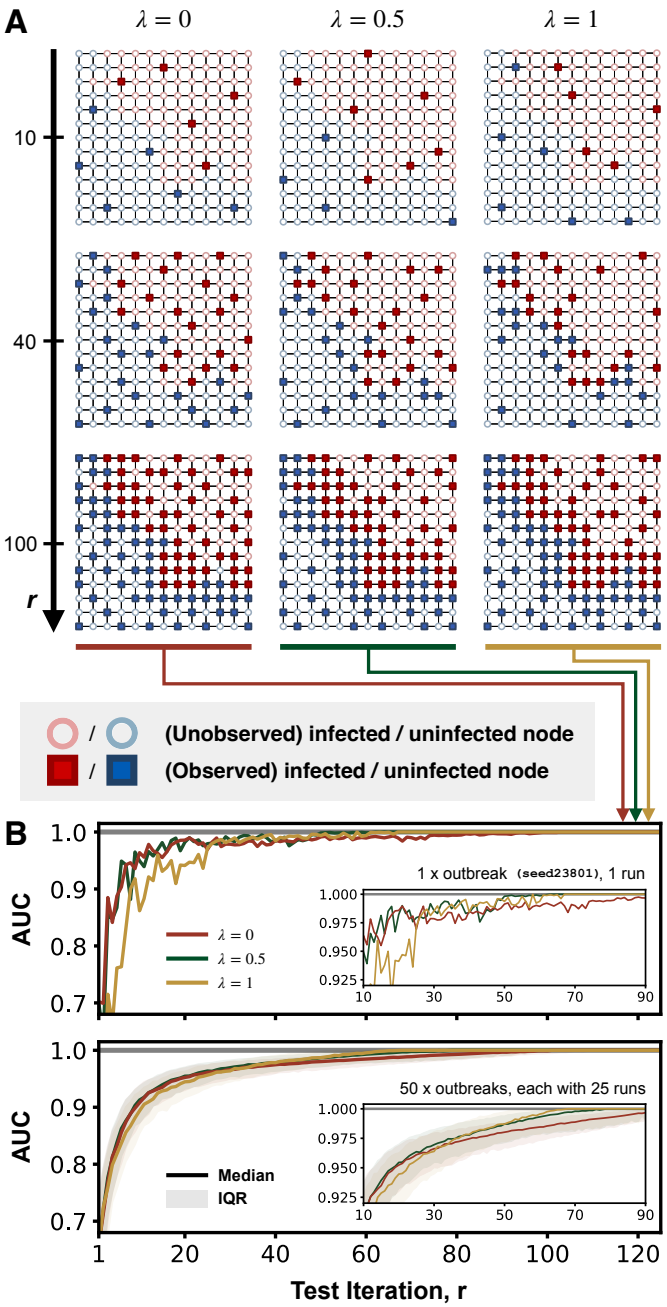


Fig. B.2. Impact of varying λ in Selection by Local-Entropy. (A) Test allocation by three selected agents, each assigned a different policy (LE with $\lambda = 0, 0.5, 1$, all with $d_{max} = d_G = 24$). Each square panel shows the distribution of observed (squares) and unobserved (circles) nodes up to test iterations $r = 10, 40, 100$. Each node is coloured according to its true infection status (red if infected and blue if uninfected, with circles that represent unobserved nodes having a lower opacity). (B) The Top plot shows the performance of the three selected agents for a single outbreak realization, as measured by the AUC. The Bottom plot shows the average performance of the three LE policies with different λ values, each summarised across 1,250 agents (50 outbreak realizations, each with 25 unique initial labelled nodes); the shaded region represents the interquartile range and the solid line represents the median.

As expected, a smaller λ leads to a stronger preference for exploration and a more rapid initial increase in AUC at small test iterations; whereas a larger λ leads to the preferential sampling of nodes close to the decision-boundary and therefore a faster convergence of AUC to 1 at large test iterations, at the cost of slower initial increase in AUC. Interestingly, an increase in λ from 0 to 0.5 results in substantially better performance at large test iteration (faster convergence to a perfect AUC) with comparable initial performance (green line), suggesting that an intermediate λ value could be a suitable default choice in scenarios where the test budget is unknown a priori.

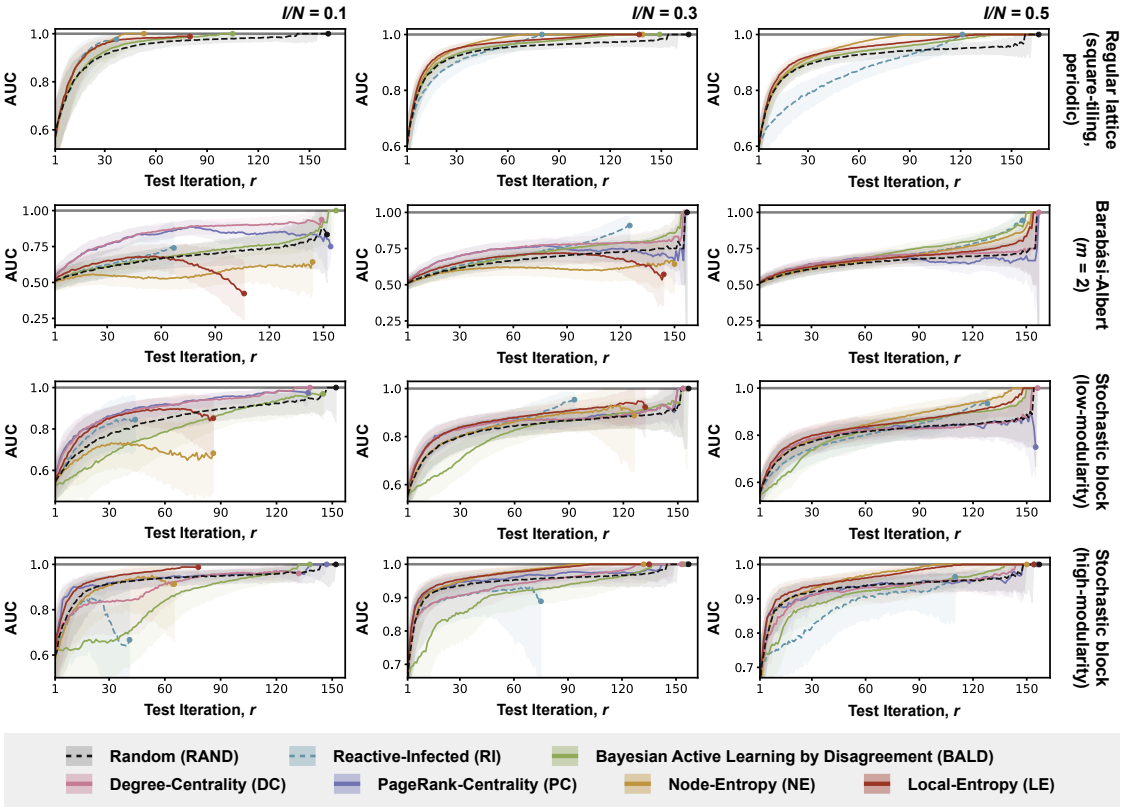


Fig. B.3. Full results from experiments with simulated outbreaks on synthetic graphs. Each row presents results from experiments with simulated outbreaks on a different synthetic graph (as indicated by labels on the right); each column corresponds to simulated outbreaks at a different stage of outbreak progression, as measured by the proportion of nodes infected ($I/N = 0.1, 0.3, 0.5$; as indicated by labels at the top). Each plot shows the performance of policies considered in the corresponding experiment, as measured by the AUC; the shaded region represents the interquartile range and the solid line represents the median. The performance of each policy is shown up to the median number of test iterations required for all infected nodes to be observed among agents assigned to that policy, with the AUC at this cut-off indicated by a colored dot (unlike Fig. 3.3, where the performance of each policy is only shown up to the median number of test iterations required for all infected nodes to be observed among agents assigned to Reactive-Infected (RI)).

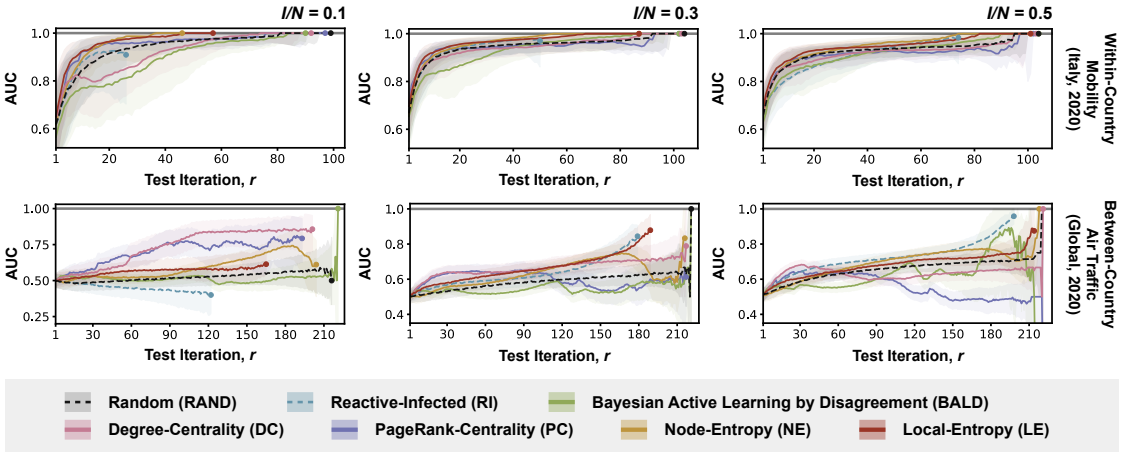


Fig. B.4. Full results from experiments with simulated outbreaks on graphs derived from empirical human mobility data. Each row presents results from experiments with simulated outbreaks on a graph derived from a different empirical human mobility dataset (as indicated by labels on the right); each column corresponds to simulated outbreaks at a different stage of outbreak progression, as measured by the proportion of nodes infected ($I/N = 0.1, 0.3, 0.5$; as indicated by labels at the top). Each plot shows the performance of policies considered in the corresponding experiment, as measured by the AUC; the shaded region represents the interquartile range and the solid line represents the median. The performance of each policy is shown up to the median number of test iterations required for all infected nodes to be observed among agents assigned to that policy, with the AUC at this cut-off indicated by a colored dot (unlike Fig. 3.3, where the performance of each policy is only shown up to the median number of test iterations required for all infected nodes to be observed among agents assigned to Reactive-Infected (RI)).

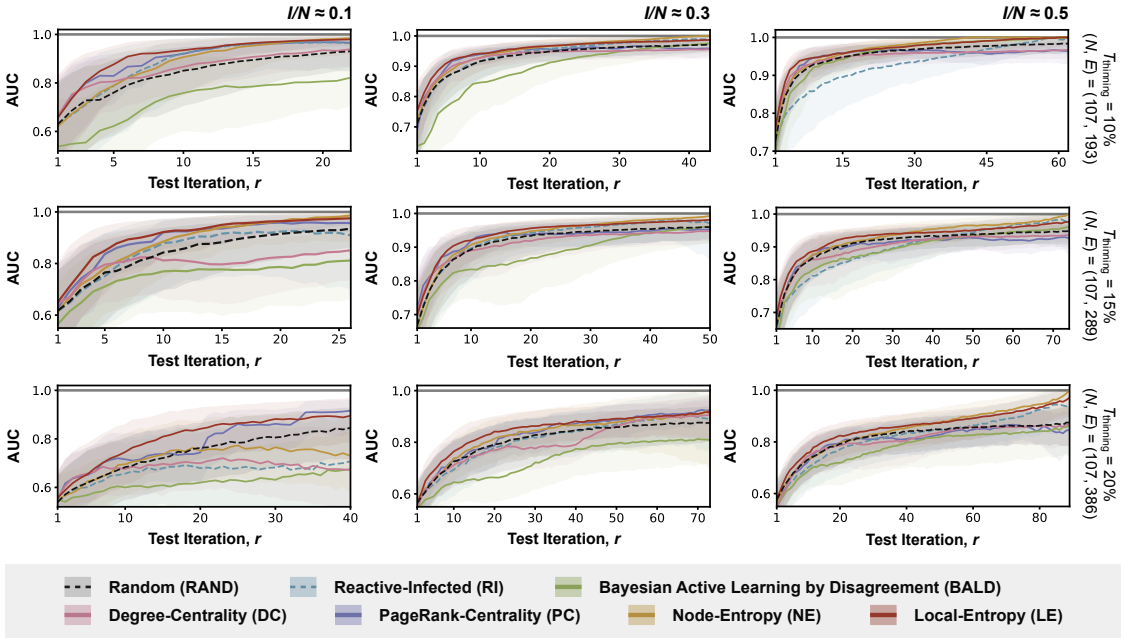


Fig. B.5. Results from sensitivity analyses with simulated outbreaks on graphs derived from aggregated mobility data collected at provincial level in Italy. Each row corresponds to a different thinning-threshold ($T_{\text{thinning}} = 10\%, 15\%, 20\%$; as indicated by labels on the right, with the number of nodes (N) and edges (E) remaining after graph-thinning also shown); each column corresponds to simulated outbreaks at a different stage of outbreak progression ($I/N = 0.1, 0.3, 0.5$; as indicated by labels at the top). Each plot shows the performance of policies considered in the corresponding experiment, as measured by the AUC; shaded regions represent the interquartile range and the solid lines represent the median. Performance of each policy is only shown up to the median number of test iterations required for all infected nodes to be observed under the policy Reactive-Infected (RI).

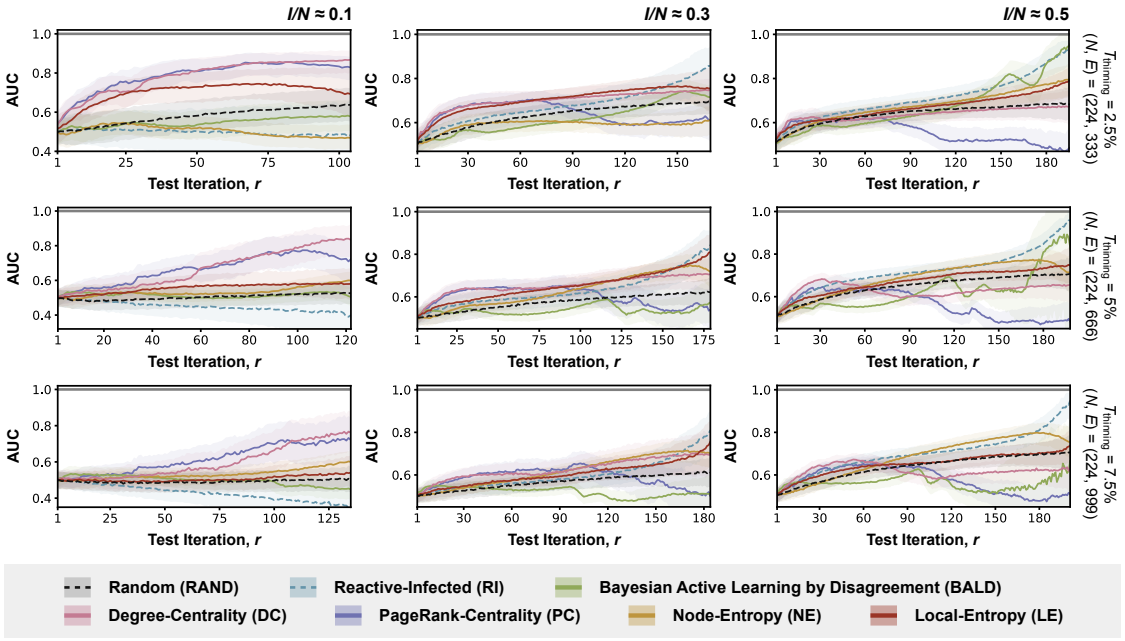


Fig. B.6. Results from sensitivity analyses with simulated outbreaks on graphs derived from air traffic data collected at country level. Each row corresponds to a different thinning-threshold ($T_{\text{thinning}} = 2.5\%, 5\%, 7.5\%$; as indicated by labels on the right, with the number of nodes (N) and edges (E) remaining after graph-thinning also shown); each column corresponds to simulated outbreaks at a different stage of outbreak progression ($I/N = 0.1, 0.3, 0.5$; as indicated by labels at the top). Each plot shows the performance of policies considered in the corresponding experiment, as measured by the AUC; shaded regions represent the interquartile range and the solid lines represent the median. Performance of each policy is only shown up to the median number of test iterations required for all infected nodes to be observed under the policy Reactive-Infected (RI).

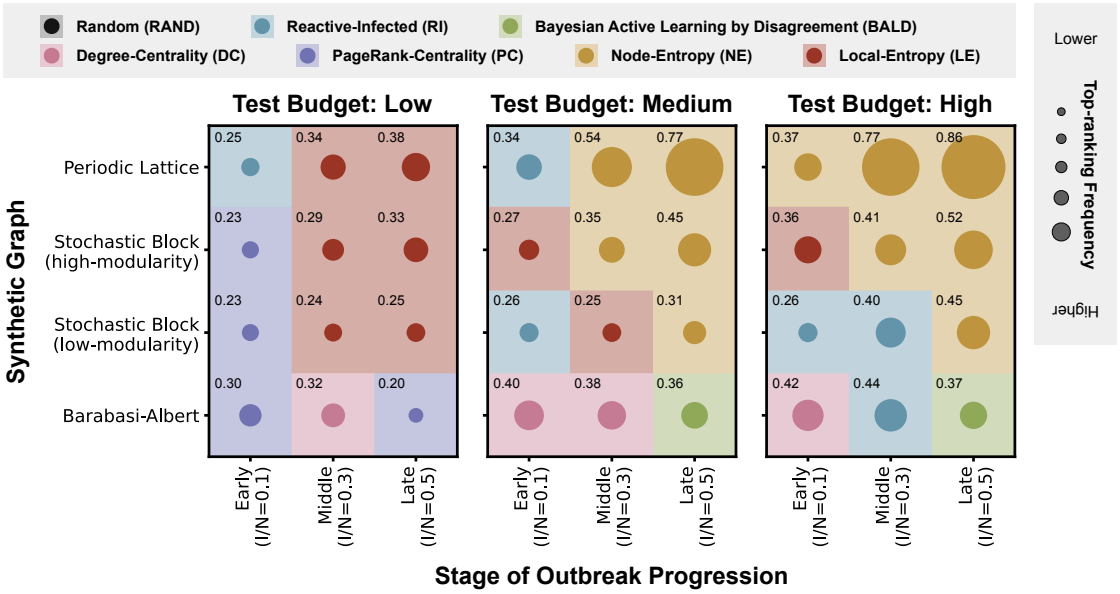


Fig. B.7. Summary of top-ranking policies from experiments with simulated outbreaks on synthetic graphs. Each heat map presents results from experiments considering a different range of test budgets: low (up to one-third of the maximum test budget), medium (between one-third and two-thirds of the maximum), and high (above two-third of the maximum), with the maximum test budget determined by the median number of test iterations required by Reactive-Infected (RI) to identify all infected nodes (see Materials and Methods for more details). Each cell in a heat map corresponds to a specific synthetic graph (as indicated by labels on the left) and stage of outbreak progression (as indicated by labels at the bottom). The colour of each cell indicates the policy that is most frequently ranked top across the test budget range, with the corresponding average top-ranking frequency represented by the size of the enclosed circle; the numerical value (to 2 decimal places) is shown in the top-left corner of each cell.

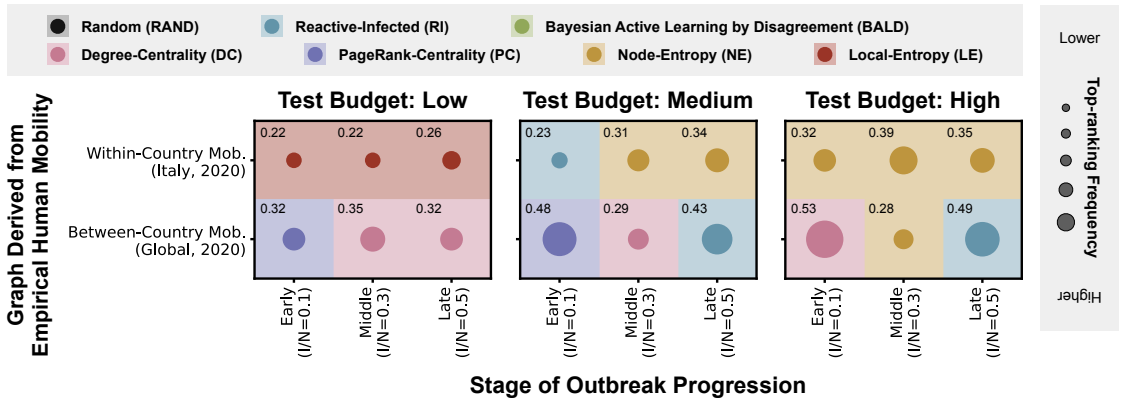


Fig. B.8. Summary of top-ranking policies from experiments with simulated outbreaks on graphs derived from empirical human mobility data. Each heat map presents results from experiments considering a different range of test budgets: low (up to one-third of the maximum test budget), medium (between one-third and two-thirds of the maximum), and high (above two-third of the maximum), with the maximum test budget determined by the median number of test iterations required by Reactive-Infected (RI) to identify all infected nodes (see Materials and Methods for more details). Each cell in a heat map corresponds to a specific synthetic graph (as indicated by labels on the left) and stage of outbreak progression (as indicated by labels at the bottom). The colour of each cell indicates the policy that is most frequently ranked top across the test budget range, with the corresponding average top-ranking frequency represented by the size of the enclosed circle; the numerical value (to 2 decimal places) is shown in the top-left corner of each cell.

Table B.1. Top-ranking frequencies of policies in experiments with simulated outbreaks on a periodic lattice graph (with square-tiling) at different test budget levels.

* highest frequency: **BOLD** ** second highest frequency: <u>underlined</u>		**Test Budget Level**		
Stage of Outbreak Progression	**Policy**	**Low (up to 1/3 of max.)**	**Medium (1/3 to 2/3 of max.)**	**High (above 2/3 of max.)**
$I/N = 0.1$	LE	<u>0.219169</u>	0.2132	0.199100
	NE	0.203815	<u>0.2445</u>	**0.369300**
	BALD	0.164185	0.1051	0.072167
	RI	**0.246215**	**0.3431**	<u>0.302267</u>
	RAND	0.166615	0.0941	0.057167
$I/N = 0.3$	LE	**0.338489**	<u>0.273230</u>	0.063056
	NE	<u>0.248163</u>	**0.538637**	**0.771723**
	BALD	0.169348	0.085289	0.022918
	RI	0.093067	0.051911	<u>0.134338</u>
	RAND	0.150933	0.050933	0.007964
$I/N = 0.5$	LE	**0.380410**	<u>0.15483</u>	<u>0.061580</u>
	NE	<u>0.283737</u>	**0.77407**	**0.858502**
	BALD	0.163385	0.04440	0.017790
	RI	0.027424	0.00158	0.059115
	RAND	0.145044	0.02512	0.003013

Table B.2. Top-ranking frequencies of policies in experiments with simulated outbreaks on a graph generated by the stochastic block model with high-modularity settings at different test budget levels.

* highest frequency: **BOLD** ** second highest frequency: <u>underlined</u>		**Test Budget Level**		
Stage of Outbreak Progression	**Policy**	**Low (up to 1/3 of max.)**	**Medium (1/3 to 2/3 of max.)**	**High (above 2/3 of max.)**
$I/N = 0.1$	LE	<u>0.203914</u>	**0.272657**	**0.363303**
	NE	0.123886	0.172429	<u>0.297600</u>
	BALD	0.040057	0.007029	0.020267
	PC	**0.229943**	0.151914	0.087590
	DC	0.106943	0.061343	0.034769
	RI	0.174371	<u>0.237314</u>	0.118328
	RAND	0.120886	0.097314	0.078144
$I/N = 0.3$	LE	**0.292496**	<u>0.318646</u>	<u>0.242381</u>
	NE	<u>0.188752</u>	**0.346246**	**0.414847**
	BALD	0.023328	0.002985	0.002750
	PC	0.183072	0.093969	0.055608
	DC	0.072512	0.035831	0.033936
	RI	0.103104	0.119338	0.216172
	RAND	0.136736	0.082985	0.034306
$I/N = 0.5$	LE	**0.330959**	<u>0.288663</u>	<u>0.264239</u>
	NE	<u>0.226094</u>	**0.445387**	**0.518526**
	BALD	0.030486	0.032541	0.044090
	PC	0.103805	0.037050	0.006027
	DC	0.105103	0.042423	0.015525
	RI	0.037773	0.079088	0.128737
	RAND	0.165780	0.074847	0.022856

Table B.3. Top-ranking frequencies of policies in experiments with simulated outbreaks on a graph generated by the stochastic block model with low-modularity settings at different test budget levels.

* highest frequency: **BOLD** ** second highest frequency: <u>underlined</u>		**Test Budget Level**		
Stage of Outbreak Progression	**Policy**	**Low (up to 1/3 of max.)**	**Medium (1/3 to 2/3 of max.)**	**High (above 2/3 of max.)**
$I/N = 0.1$	LE	0.169067	0.192613	<u>0.209457</u>
	NE	0.100693	0.075760	0.087429
	BALD	0.040053	0.007093	0.008400
	PC	**0.228880**	<u>0.217120</u>	0.202543
	DC	<u>0.193867</u>	0.200907	0.195543
	RI	0.169707	**0.256640**	**0.257857**
	RAND	0.097733	0.049867	0.038771
$I/N = 0.3$	LE	**0.241110**	**0.252038**	0.162840
	NE	0.107935	0.150200	<u>0.225680</u>
	BALD	0.022271	0.009125	0.032627
	PC	<u>0.208942</u>	0.186113	0.073040
	DC	0.205652	0.144838	0.072720
	RI	0.125871	<u>0.210713</u>	**0.402493**
	RAND	0.088219	0.046975	0.030600
$I/N = 0.5$	LE	**0.251814**	<u>0.171265</u>	0.112317
	NE	0.154642	**0.310391**	**0.447413**
	BALD	0.039507	0.123619	0.137083
	PC	0.161470	0.111767	0.034381
	DC	<u>0.180130</u>	0.104000	0.021162
	RI	0.085674	0.114735	<u>0.222571</u>
	RAND	0.126763	0.064223	0.025073

Table B.4. Top-ranking frequencies of policies in experiments with simulated outbreaks on a graph generated by the Barabási-Albert model at different test budget levels.

* highest frequency: **BOLD** ** second highest frequency: <u>underlined</u>		**Test Budget Level**		
Stage of Outbreak Progression	**Policy**	**Low (up to 1/3 of max.)**	**Medium (1/3 to 2/3 of max.)**	**High (above 2/3 of max.)**
$I/N = 0.1$	LE	0.119113	0.063418	0.041085
	NE	0.050487	0.007673	0.001855
	BALD	0.063043	0.016218	0.007291
	PC	**0.299600**	<u>0.386018</u>	<u>0.354448</u>
	DC	<u>0.296748</u>	**0.398673**	**0.419885**
	RI	0.106557	0.108782	0.165873
	RAND	0.064452	0.019218	0.009564
$I/N = 0.3$	LE	0.114038	0.075800	0.066634
	NE	0.049943	0.010457	0.012517
	BALD	0.045381	0.040610	0.151161
	PC	<u>0.315238</u>	<u>0.304905</u>	0.063405
	DC	**0.319629**	**0.382114**	<u>0.226098</u>
	RI	0.098819	0.159800	**0.435746**
	RAND	0.056952	0.026314	0.044439
$I/N = 0.5$	LE	0.150328	0.073935	0.055966
	NE	0.103792	0.105976	0.177720
	BALD	0.141528	**0.357869**	**0.366313**
	PC	**0.197864**	0.106792	0.031547
	DC	<u>0.190824</u>	0.148465	0.067289
	RI	0.116312	<u>0.152751</u>	<u>0.265550</u>
	RAND	0.099352	0.054212	0.035614

Table B.5. Top-ranking frequencies of policies in experiments with simulated outbreaks on a graph derived from within-country human mobility data ($T_{\text{thinning}} = 15\%$) at different test budget levels.

* highest frequency: **BOLD** ** second highest frequency: <u>underlined</u>		**Test Budget Level**		
Stage of Outbreak Progression	**Policy**	**Low (up to 1/3 of max.)**	**Medium (1/3 to 2/3 of max.)**	**High (above 2/3 of max.)**
$I/N = 0.1$	LE	**0.219754**	0.206427	<u>0.242877</u>
	NE	0.148102	0.219508	**0.318305**
	BALD	0.074824	0.022554	0.033323
	PC	0.158524	<u>0.225227</u>	0.166825
	DC	0.104304	0.034351	0.017657
	RI	<u>0.188401</u>	**0.229539**	0.164540
	RAND	0.106090	0.062394	0.056473
$I/N = 0.3$	LE	**0.219780**	<u>0.239300</u>	0.173316
	NE	0.159239	**0.311060**	**0.394824**
	BALD	0.048376	0.012331	0.045581
	PC	<u>0.162251</u>	0.060553	0.031986
	DC	0.154329	0.108904	0.045688
	RI	0.155200	0.203460	<u>0.270589</u>
	RAND	0.100824	0.064392	0.038016
$I/N = 0.5$	LE	**0.261104**	<u>0.217568</u>	0.170383
	NE	<u>0.196256</u>	**0.337395**	**0.348633**
	BALD	0.090016	0.118931	0.138441
	PC	0.132480	0.032816	0.011147
	DC	0.115040	0.069312	0.041220
	RI	0.066896	0.112088	<u>0.228010</u>
	RAND	0.138208	0.111891	0.062167

Table B.6. Top-ranking frequencies of policies in experiments with simulated outbreaks on a graph derived from between-country air traffic data ($T_{\text{thinning}} = 5\%$) at different test budget levels.

* highest frequency: **BOLD** ** second highest frequency: <u>underlined</u>		**Test Budget Level**		
Stage of Outbreak Progression	**Policy**	**Low (up to 1/3 of max.)**	**Medium (1/3 to 2/3 of max.)**	**High (above 2/3 of max.)**
$I/N = 0.1$	LE	0.131639	0.081161	0.048880
	NE	0.092790	0.026459	0.029920
	BALD	0.116537	0.012156	0.003260
	PC	**0.320390**	**0.478224**	<u>0.362887</u>
	DC	<u>0.220946</u>	<u>0.375717</u>	**0.526677**
	RI	0.061844	0.012937	0.023157
	RAND	0.055854	0.013346	0.005220
$I/N = 0.3$	LE	0.117387	<u>0.229433</u>	0.245539
	NE	0.033120	0.126787	**0.281207**
	BALD	0.033460	0.000773	0.003817
	PC	<u>0.330200</u>	0.194087	0.015254
	DC	**0.352613**	**0.294760**	0.193668
	RI	0.101347	0.132787	<u>0.249539</u>
	RAND	0.031873	0.021373	0.010976
$I/N = 0.5$	LE	0.104297	0.138991	0.072240
	NE	0.066545	<u>0.318454</u>	<u>0.242665</u>
	BALD	0.027958	0.026412	0.133206
	PC	0.135867	0.012967	0.000474
	DC	**0.321073**	0.009600	0.013846
	RI	<u>0.288491</u>	**0.431887**	**0.487163**
	RAND	0.055770	0.061690	0.050406

References

1. Besag, J., York, J. and Mollié, A. (1991) 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the Institute of Statistical Mathematics*, 43(1), pp. 1–20.
2. Bajardi, P., Poletto, C., Ramasco, J.J., Tizzoni, M., Colizza, V. and Vespignani, A. (2011) 'Human Mobility Networks, Travel Restrictions, and the Global Spread of 2009 H1N1 Pandemic', *PLOS ONE*, 6(1), p. e16591.
3. Tegally, H., Wilkinson, E., Tsui, J.L., Moir, M., Martin, D., Brito, A.F., Giovanetti, M., Khan, K., Huber, C., Bogoch, I.I., San, J.E., Poongavanan, J., Xavier, J.S., Candido, D.D.S., Romero, F., Baxter, C., Pybus, O.G., Lessells, R.J., Faria, N.R., Kraemer, M.U.G. and de Oliveira, T. (2023) 'Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern', *Cell*, 186(15), pp. 3277-3290.e16
4. Brockmann, D. and Helbing, D. (2013) 'The Hidden Geometry of Complex, Network-Driven Contagion Phenomena', *Science*, 342(6164), pp. 1337-1342.
5. Zhi, Y.-C., Ng, Y.C. and Dong, X. (2020) 'Gaussian Processes on Graphs via Spectral Kernel Learning', arXiv. Available at https://doi.org/10.48550/arXiv.2006.07361 (Accessed: 13 October 2024).
6. Borovitskiy, V., Azangulov, I., Terenin, A., Mostowsky, P., Deisenroth, M.P. and Durrande, N. (2021) 'Matern gaussian processes on graphs', *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)* (PMLR, 2020), vol. 130, pp. 2593–2601.
7. Zhu, X. and Ghahramani, Z. (2002) 'Learning from labeled and unlabeled data with label propagation', *Technical Report CMU-CALD-02-107,* Carnegie Mellon University.
8. Kipf, T.N. and Welling, M. (2017) 'Semi-supervised classification with graph convolutional networks', *arXiv*. Available at: https://doi.org/10.48550/arXiv.1609.02907 (Accessed: 24 May 2024).
9. Houlsby, N., Huszár, F., Ghahramani, Z. and Lengyel, M. (2011) 'Bayesian Active Learning for Classification and Preference Learning', *arXiv*. Available at: https://arxiv.org/pdf/1112.5745.pdf (Accessed: 24 May 2024).
10. Holland, P.W., Laskey, K.B. and Leinhardt, S. (1983) 'Stochastic blockmodels: First steps', *Social Networks*, 5(2), pp. 109–137.
11. Newman, M.E.J. (2006) 'Modularity and community structure in networks', *Proceedings of the National Academy of Sciences*, 103(23), pp. 8577–8582.
12. Pepe, E., Bajardi, P., Gauvin, L., Privitera, F., Lake, B., Cattuto, C. and Tizzoni, M. (2020) 'COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown', *Scientific Data*, 7, p. 230.
13. Rudolf, S. (2022) *Source for "How to avoid a local epidemic becoming a global pandemic?"*, Zenodo. Available at https://doi.org/10.5281/zenodo.7472836 (Accessed 3 March 2024).

4

Transmission lineage dynamics and the detection of viral importation in emerging epidemics

As my research on improving disease surveillance progressed, my continued contributions to empirical studies of SARS-CoV-2 spread using phylogeography led me to recognise a critical gap in genomic surveillance. Despite our growing capacity for large-scale genomic surveillance, the coordination of sampling efforts remains largely *ad hoc* and is frequently dictated by logistical and infrastructural considerations. Tackling this challenge, however, requires an understanding of how the collected pathogen genomes contribute to downstream analyses, and therefore how different sampling strategies affect the accuracy and robustness of relevant estimates.

In the context of phylogeography specifically, a primary quantity of interest is the frequency of viral importation, which is critical for informing outbreak response and evaluating the effectiveness of containment strategies, as highlighted in Chapter 2. However, it is well recognised that these estimates often substantially underestimate the true number of importation events, given that only a small fraction of infections are typically sampled. Despite such estimates being widely reported in numerous SARS-CoV-2 studies, the extent and underlying mechanisms of their underestimation have remained poorly understood. The work presented in this chapter aims to address this gap.

A manuscript describing this work has been submitted for peer-review and is available as a preprint on medRxiv, under the title "*Transmission lineage dynamics and*

the detection of viral importation in emerging epidemics". It is presented here in full, with minor modifications to ensure consistency of formatting and style within this thesis.

Tsui, J.L.-H., Sambaturu, P., Pena, R.E., Too, L., Gutierrez, B., Inward, R., Kraemer, M.U.G., du Plessis, L. and Pybus, O.G. (2025) 'Transmission lineage dynamics and the detection of viral importation in emerging epidemics', *medRxiv*. Available at: https://doi.org/10.1101/2025.03.05.25323408.

4.1 Abstract

The accurate inference of pathogen movements between locations during an epidemic is crucial for measuring infectious disease spread and for informing effective control strategies. Phylogeographic methods can reconstruct historical patterns of disease dissemination by combining the evolutionary history of sampled pathogen genomes with geographic information. Despite a substantial expansion of pathogen genomics during and since the COVID-19 pandemic, only a small fraction of infections are typically sampled and sequenced, leading to an underestimation of the true intensity of viral importation. Here, we seek to understand the sampling processes underlying this underestimation. We show that the coupling of viral importation and local transmission dynamics can result in local transmission lineages with different size distributions, influencing the probability that individual viral importation events will be detected. Using analytical and simulation approaches, we demonstrate that both the proportion of importation events detected and the temporal patterns of inferred importation are highly sensitive to importation dynamics and local transmission parameters. Our findings highlight the importance of interpreting phylogeographic estimates in the context of outbreak conditions, particularly when comparing viral movements across time and among different epidemic settings. These insights are critical for improving the reliability of genomic epidemiology approaches in the design of public health responses.

4.2 Introduction

Reconstruction of the spatiotemporal spread of an emerging pathogen is needed to inform the design of public health policies that aim to contain and delay further disease spread. Typically, transmission is either measured directly, for example through contact-tracing (1-5) and traveller screening (6-11), or inferred indirectly from epidemiological (e.g., case incidence, hospitalisation rates) and mobility (e.g., mobile devices, flight records) data using spatial transmission models (12-14). For instance, during the 2009 H1N1 influenza pandemic, simulation models that combined intra-country commuting flows, inter-country air traffic, and high-resolution demographic data were used to characterise the dynamics and drivers of global virus spread (15, 16). More recently, an analysis of contact-tracing records for >600,000 COVID-19 cases in New York City during 2020-21 revealed substantial spatial heterogeneities and strong community structures, with frequent non-local transmission events across administrative regions (17). Insights from such studies can identify locations and spatial scales at which targeted interventions will be most effective and assess the potential impacts of interventions that restrict human movement.

However, epidemiological data are often constrained by reporting delays, underreporting, and logistical challenges in data collection and sharing, particularly during large outbreaks. The limited availability of real-time data on human mobility and contact patterns also hinders the accurate inference of pathogen movements, especially at small spatial scales where fine-scale heterogeneity is difficult to capture using standard mobility models (18-20). To address these limitations, pathogen genomic data are increasingly being used to investigate pathogen dissemination, with the potential to uncover cryptic transmission pathways that are not readily observed using traditional epidemiological data alone. Recent advances in high-performance computing, together

with the growing availability of genomic data from public repositories such as GISAID, GenBank and Pathoplexus, have enabled the analysis of large-scale genomic and epidemiological datasets (21). These analyses often reveal complex transmission dynamics across multiple spatial scales, from transmission events among individual households to between-country movements via the global air traffic network. For example, an analysis of 482 SARS-CoV-2 sequences from students and staff at a university in the United Kingdom found limited viral introductions from the wider community, with onward transmission within the university driven primarily by shared student accommodation and in-person course-related interactions (22). At the international-level, a recent study of ~6,000 influenza genomes showed that travel and movement restrictions during the COVID-19 pandemic led to notable shifts in the global dispersal patterns of seasonal influenza lineages, with persistent transmission in South Asia during the pandemic (23).

Viral genomic epidemiology studies often rely on phylogeographic methods, which integrate information about the evolutionary relationships among pathogen genomes and the locations of sampled infections. Historical patterns of pathogen migration are inferred by extrapolating the locations of sampled sequences backward in time, guided by the ancestral relationships among them. Depending on the spatial resolution of the data and model assumptions, locations can be treated as either continuous (24) or discrete (25). Discrete locations are often used to model pathogens spreading among human populations due to heterogeneities in movement resulting from administrative boundaries and long-range transportation systems that span multiple spatial scales (leading to non-linearity between travel time and displacement) (26-28). In discrete phylogeography models, the location of internal nodes in a phylogenetic tree (representing the ancestors of sampled viruses) are commonly inferred by modelling

pathogen movement among locations as a continuous-time Markov process along an estimated phylogeny (25). Once the internal nodes are assigned their inferred locations, local transmission clusters/lineages can be identified, with each representing a series of local transmission events that occurred in a recipient location following a single pathogen importation event. The detection and enumeration of these local transmission lineages (and their associated importation events) provides an opportunity to assess the frequency and dynamics of pathogen movement among locations.

Although analyses of local transmission lineages have been present in the literature for some time (e.g., (29), (30)), their popularity and scale expanded during the COVID-19 pandemic, particularly in settings where traditional travel or contact tracing data were scarce or incomplete. These studies examined how viral importation from different countries contributed to the establishment of new variants and assessed the effectiveness of non-pharmaceutical interventions designed to prevent or delay further dissemination, such as travel restrictions and airport screening (31-34). Estimates of the intensity of viral importation through time at a given location also revealed how local transmission dynamics were influenced by the introduction of new transmission lineages, with implications for the design of local control strategies. However, despite widespread adoption of these approaches, the underlying sampling processes that underpin their inferences remain poorly characterised. For instance, du Plessis et al. (35) investigated the early establishment of SARS-CoV-2 in the UK and found that local transmission lineages varied widely in size and were distributed heterogeneously across space and time – yet, the conditions necessary for the identification and enumeration of these lineages under such heterogeneities have not been explored. Specifically, given that only a small fraction of local infections are typically sampled and sequenced during an outbreak, to what extent is the number of viral importation events underestimated for a genomic

sample of a given size, and how does this discrepancy vary over time and across different outbreak conditions?

In this study, we address this question by considering the mechanisms that underlie the detection of local transmission lineages and their associated viral importation events through discrete phylogeographic reconstruction. We show analytically how variation in viral importation intensity through time can lead to transmission lineages with different size distributions, despite the same local transmission conditions. Using a simple deterministic model, we then verify these analytical results and further show how different lineage size distributions can result in different lineage detection probabilities. Additionally, using simulated data from a stochastic agent-based model, we demonstrate the impact of temporal variation in local transmission intensity on lineage detection, and the resulting bias in the inferred importation intensity over time. We conclude that estimates of viral movements from phylogeographic analyses in the regime of low-intensity sampling ($\lesssim 5\%$) should be interpreted cautiously, and that further work is needed to mitigate such biases with consideration of both the underlying importation dynamics and local transmission conditions.

4.3 Phylogeographic reconstruction and local transmission lineages

Discrete trait analysis (DTA) (25) has emerged as a popular approach for phylogeography due to its computational efficiency, allowing the analysis of thousands of pathogen sequences. In a typical DTA, a time-calibrated phylogeny is first estimated from a set of aligned sequences, each labelled with its location and time of sampling. The most likely location of each internal node in the estimated phylogeny is then inferred using a continuous-time Markov chain (CTMC) model. By tracing a path from the root node of the tree to each leaf node (sampled sequence), a migration or importation event is inferred

to have occurred whenever we observe a change in the label going from one node to another. The inferred age of the ancestral node in a local transmission lineage (often referred to as the Time of Most Recent Common Ancestor, or TMRCA) also provides an estimate of the time of earliest detectable transmission event within that local lineage given the sampled sequences, and therefore an upper bound (most recent estimate) for the timing of viral importation (Fig. 4.1B) (35).

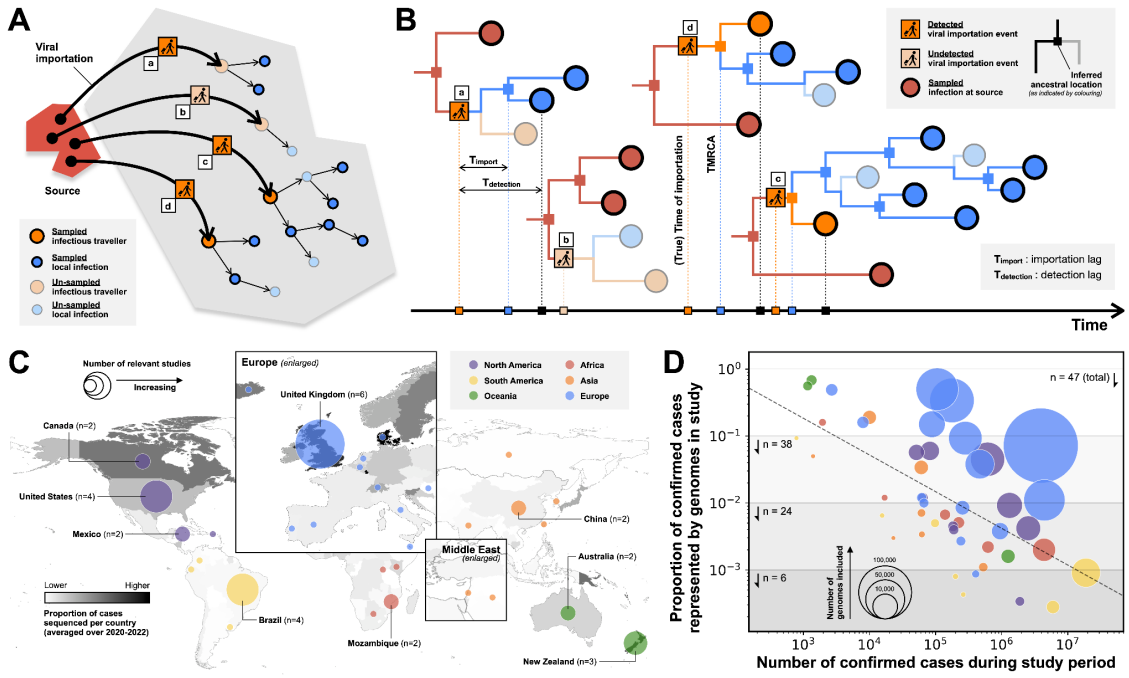


Fig. 4.1. Phylogeographic reconstruction of viral importation and distribution of sampling proportions in COVID-19 studies. (A) Each viral importation event, i.e. the movement of an infectious traveller (represented as a curved arrow) from the source (red polygon) to the recipient (grey polygon) location, can be uniquely mapped to a local transmission lineage. The transmission tree associated with each local cluster is shown, with arrows representing transmission events; orange and blue circles represent arriving infectious travellers and local infections, respectively. (B) Time-scaled phylogeographic reconstruction of the local transmission lineages shown in panel (A). Internal and leaf nodes are represented as squares and circles, respectively, and coloured according to their state (red: infected and detected at source location, orange: imported and detected locally, blue: locally infected and detected). Despite complete sampling of viral genomes at the source location (red circles), only importation events (a), (c), and (d) are detected, as no members of transmission lineage (b) are sampled. The difference between the true time of importation and the TMRCA of a local transmission lineage is known as the importation lag, denoted by T_{import}; the difference between the true time of importation

and the time when a local transmission lineage is first detected is known as the detection lag, denoted by $T_{detection}$. (C) Sampling proportions in previous COVID-19 studies that estimated the number of viral introductions and intensity of viral importation using phylogeographic reconstruction. Countries are coloured according to the average proportion of COVID-19 cases that were sequenced between 2020 and 2022 (darker colours indicate higher sequencing proportions). Countries from which SARS-CoV-2 genomes were collected and analysed are indicated by circles; circle radii are proportional to the number of studies for each country and coloured by continent. (D) Plot of the proportion of confirmed COVID-19 cases that were sequenced versus the total number of confirmed COVID-19 cases for each relevant study. The radius of each circle indicates the number of SARS-CoV-2 genomes included in the phylogeographic analysis of the relevant study, with colouring indicating the continent of the country in question. The black dashed line represents the least squares regression fit (Pearson's r = -0.63) between the log10-transformed number of confirmed cases and the log10-transformed sampling proportion. Note that a study considering four island countries across three continents was omitted from the figure; see Table C.1 in Appendix C for more details.

Given a phylogenetic tree with internal nodes that are annotated with their most likely location, we can then partition the phylogeny into distinct, non-overlapping local transmission lineages. The use of local transmission lineages to represent a chain of transmission within a given location following a single importation event is well established in the literature on phylogenetic epidemiology (29, 30, 36) and was recently popularised by du Plessis et al. (35) and other studies (37-39) which leveraged the increased availability of viral genomic data during the COVID-19 pandemic. This formulation enables a one-to-one mapping between local transmission lineages and viral importation events (Fig. 4.1A, B), provided that (i) virus genetic diversity accumulates at a sufficiently rapid rate, and (ii) viral genomes at the source location (from which the pathogens are imported) are sufficiently densely sampled, such that onward local transmission lineages from distinct imported pathogen carriers can be uniquely resolved phylogenetically. In this study, we assume that both conditions are satisfied and focus our attention instead on the potential biases introduced by the undersampling of local infections.

Here we define a local transmission lineage as a group of individuals who were infected in a given location as a result of onward transmission from a single arriving infectious traveller, with the inclusion of the traveller itself. Additionally, we assume that (i) the infectious traveller remains at the recipient location indefinitely following arrival, and (ii) individuals infected locally do not travel to a new location. These assumptions together imply that:

1. Each local transmission lineage can be uniquely mapped to a single arriving infectious traveller, and therefore a single importation event.
2. Each local transmission lineage has size $l = n + 1$, where n is the total number of secondary infections that occurred locally as a result of onward transmission from the arriving infectious traveller.
3. Each local transmission lineage has a minimum size of 1, i.e. when the arriving infectious traveller fails to establish local onward transmission (i.e. $n = 0$).

Finally, we ignore any uncertainties and biases associated with phylogenetic tree estimation and ancestral state reconstruction (see Discussion). Consequently, we can further assume that (i) there is a non-zero probability that each lineage (and its associated viral importation event) is detected by random sampling of local infections, and (ii) the sampling of at least one member of a given local transmission lineage is a sufficient and necessary condition for the detection of the lineage in a phylogeographic reconstruction.

4.4 Detection of local transmission lineages in the regime of low-intensity local sampling

Despite increases in genomic sequencing worldwide since the beginning of the COVID-19 pandemic, the number of sequenced SARS-CoV-2 genomes remains low compared to

the number of reported cases in most countries. Brito et al. (40) showed that, among 189 countries with active genomic surveillance between 2020 and 2022, the average number of genomic sequences available per confirmed case is just 0.016, with only 13 countries having a sequencing coverage $> 5\%$ and 89 countries having sequencing coverage $< 0.5\%$. We reviewed 48 phylogeographic studies that estimated the number of viral introductions at a given location (see Section C.1 and Table C.1 in Appendix C for review details) and found that in only 9 studies did pathogen genome sequence coverage exceed 10% of the confirmed cases during the corresponding study period; in 24 studies, included sequences represented $< 1\%$ of the confirmed cases (Fig. 4.1B, right). We also observe a negative association between genomic sampling proportion and the number of confirmed cases, likely due to limited sequencing capacity during periods of high case incidence (Fig. 4.1C, D). Importantly, the true proportion of infections that were sequenced is likely lower than these estimates due to the presence of asymptomatic (41-43) and limited testing capacities, as demonstrated by the ratio of seroprevalence to cumulative incidence in many low- and middle-income countries (LMICs) (44, 45).

Given that genome sequencing occurs at such low intensities, the number of transmission lineages detected from a sample of local infections is likely to substantially underestimate the true number of extant transmission lineages in the population. More formally, given S genomic samples collected at random from a local population of infections of size N_I, the probability of detecting a lineage of size l within the sample is given by

$$\Pr(detection) = \Pr\left(n_{sample} \geq 1 \middle| l, S, N_I\right) = 1 - \binom{N_I - l}{S} \Big/ \binom{N_I}{S} \qquad (1)$$

where n_{sample} is the number of lineage members that are sampled. In the regime of low-intensity sampling such that $S \ll N_I$, we obtain the approximation

$$\Pr(detection|l, S, N_I) \approx 1 - (1 - l/N_I)^S \quad (2)$$

Using this approximation, the proportion of lineages in an infected population that we can expect to find in a sample of size S (referred to as lineage detection probability r_d hereafter) is therefore given by

$$r_d \approx \sum_{l=1}^{N_I} [1 - (1 - l/N_I)^S] n(l)/N_{lineages} \quad (3)$$

where $n(l)$, hereafter referred to as lineage density at l, is the number of lineages of size l in the local infected population, such that $\sum_{l=1}^{N_I} n(l) = N_{lineages}$, the total number of lineages in the local infected population. Note that the summation has an upper bound N_I, corresponding to the maximum size of a lineage in the scenario where there is only a single transmission lineage.

From the above result, we can see that the lineage detection probability depends on not only the sample size S (relative to the size of the infected population N_I) (Fig. 4.2B), but also the lineage size distribution, as specified by $n(l)$. We can draw useful insights by considering a hypothetical scenario in which each lineage is of size $l = 1$ (i.e. when none of the arriving infectious travellers are able to establish local onward transmission), for which we obtain $r_d \approx S/N_I$ from Eq. 3, i.e. the expected number of detected lineages is directly proportional to the sampling proportion, $s = S/N_I$. Conversely, in the hypothetical scenario where the entire infected population consists of a single transmission lineage of size N_I, Eq. 3 reduces to $r_d = 1$ for all $S > 0$ and $l \geq 1$, i.e. only a single sampled infection is needed to detect all importation events (with $N_{lineages} = 1$), as expected. In reality, the underlying lineage size distribution, and therefore the expected behaviour of lineage detection, will lie somewhere between these two extremes, as we show below.

Several studies have attempted to measure the size distribution of transmission lineages in an infected population, with most of which reporting observed distributions that were right-skewed (heavy-tailed) (34, 35, 46), i.e. many small lineages and only a few large ones. In practice, however, the true underlying size distribution of local transmission lineages in an infected population cannot be measured directly, especially in the regime of low-intensity sampling, because the observed size of a detected lineage depends on both the sampling proportion and the true lineage size, with the latter being an unknown quantity (see Fig. C.1 in Appendix C). This challenge closely resembles the ecological problem of estimating the number of unseen species, where the observed relative species abundance distribution is skewed by incomplete sampling with rare species being underrepresented (47-51). We leave this non-trivial inference problem for future work. Instead, we use a simple analytical model to investigate how different lineage size distributions may arise as a result of the coupling between viral importation and local transmission.

4.5 Time evolution of transmission lineage size distribution

We consider a hypothetical scenario in which the number of infectious travellers arriving at a location of interest from an arbitrary source location is given by $M(t)$ for $t > 0$. In particular, we focus our attention on the early phase of an epidemic or the emergence of a new immune escape variant of a known pathogen, when it can be reasonably assumed that (i) the population at the location of interest (referred to as the local population hereafter) is completely susceptible, (ii) the rate of local transmission is sufficiently high such that the recovery of infectious individuals has a negligible impact on the overall dynamics of the local outbreak, and (iii) there is no substantial depletion of susceptibles. We will later relax assumptions (ii) and (iii) in a simulation analysis.

Given these assumptions, and following from the abovementioned definition of a local transmission lineage, we propose that the growth of a transmission lineage following the arrival of an infectious traveller can be modelled as the movement of a particle in a one-dimensional continuous lineage size-space (given a sufficiently large infected population) (Fig. 4.2A). The time evolution of the density of these particles (each representing a single local transmission lineage) in this size-space can therefore be considered as solutions to the continuity equation

$$\frac{\partial n}{\partial t} + \frac{\partial}{\partial l}\left(n\frac{dl}{dt}\right) = G(l,t) \tag{4}$$

with the boundary condition $n(l,0) = 0$ for all $l \geq 1$, where $n = n(l)$ is the particle density at l in size-space, i.e. the number of lineages of size between l and $l + \Delta l$, and is analogous to the variable with the same notation in Eq. 3, in the limit of large l. dl/dt represents the instantaneous lineage growth rate, which we assume to be homogeneous across lineages of the same size at a given time. In the fluid dynamics literature (52, 53), the term $G(l,t)$ is commonly referred to as a *source term*, which here describes the rate at which new transmission lineages are being introduced to the location of interest through importation. Assuming further that there is no outward movement of infected individuals, and since each transmission lineage must have a minimum size of $l = 1$ (as described previously), we set

$$G(l,t) = M(t)\delta(l-1) \tag{5}$$

where $\delta(l-1)$ is the Dirac delta function centred at $l = 1$.

Depending on the assumed functional form of the lineage growth rate and importation rate, the above partial differential equation (Eq. 4) can be solved either analytically or numerically (e.g., using finite difference methods). Here we consider a hypothetical scenario in which the local infected population is experiencing exponential

growth such that $dl/dt = rl$, where r is a positive constant, as is commonly assumed during the early stages of an outbreak. It can be shown that Eq. 4 can be solved analytically to give the solution (see Section C.2 in Appendix C)

$$n(l,t) = \frac{M(t - \ln l\,/r)}{l}[\mathrm{H}(l-1) - \mathrm{H}(l - e^{rt})] \tag{6}$$

where $\mathrm{H}(x)$ is the Heaviside step function. The term $\mathrm{H}(l-1) - \mathrm{H}(l - e^{rt})$ is commonly known as a boxcar function, and is zero everywhere except for the interval $l \in [1, e^{rt}]$ in which it takes a value of 1. The right-limit of this interval, e^{rt}, corresponds to the maximum size attainable for a lineage seeded at $t = 0$ under the assumption of local exponential growth and therefore the maximum possible lineage size at a given time t since the first importation event; similarly, the left-limit of this interval represents the minimum size of a lineage (i.e. $l = 1$).

Importantly, we have not made any assumptions in our derivation regarding the functional form of the underlying importation rate, $M(t)$. It is therefore instructive to consider the behaviour of $n(l,t)$ under different assumptions of $M(t)$, specifically (i) a constant rate, (ii) an exponentially decreasing rate, and (iii) an exponentially increasing rate:

1. Assuming a constant rate of viral importation ($M(t) = M_0$), the first term in Eq. 6 reduces to M_0/l, i.e. the lineage density is inversely proportional to lineage size l. This result is not surprising, as larger lineages grow more rapidly compared to smaller lineages under the assumption of exponential growth. In our model of lineage growth, this implies that particles corresponding to larger lineages move at a higher velocity towards the right, resulting in a lower lineage density at large l (Fig. 4.2A).
2. Conversely, the lower velocity of particles corresponding to smaller lineages results in their accumulation at small l; the higher the rate at which new particles (arriving

infectious travellers) are introduced relative to their velocity in size-space (instantaneous lineage growth rate), the more rapidly these particles accumulate. Indeed, if we assume an exponentially decreasing importation rate ($M(t) = M_0 e^{mt}$, with $m < 0$), the first term in Eq. 6 reduces to $M_0 e^{mt} l^{-(1+m/r)}$. This represents a power-law distribution, which increases in density with l only if $m > r$, i.e. only when the rate at which new lineages are introduced is decreasing sufficiently rapidly compared to the local growth rate that we obtain lineage density that increases with lineage size.

3. During the early dissemination of an emerging pathogen, however, the rate of viral importation is likely to be increasing over time as prevalence at the origin location increases, in which case we retrieve a lineage size distribution that decreases in density with increasing lineage size l following a power-law, i.e. $n(l,t) = M_0 e^{mt} l^{-(1+m/r)}$ with $m > 0$.

The derivation of an equivalent expression for $n(l,t)$ under the assumption of local logistic growth (e.g., as a result of the depletion of susceptibles as the outbreak progresses) can be found in Section C.3 in Appendix C.

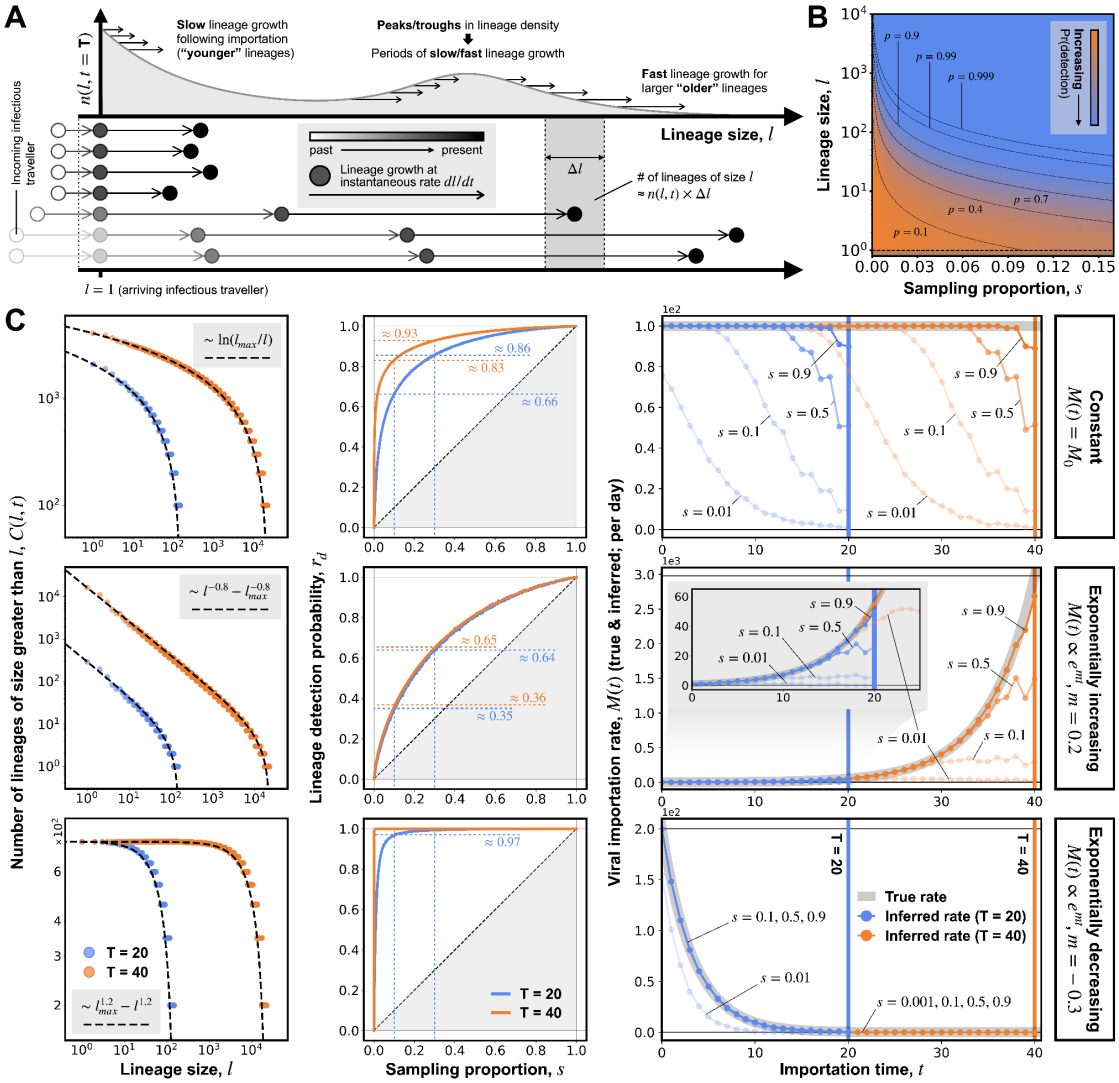


Fig. 4.2. Time evolution of local transmission lineage size distribution and simulated lineage detection under a simple deterministic model with local exponential growth. (A) An illustration of the time evolution of the transmission lineage size distribution in an outbreak modelled as the movement of particles in a 1D lineage size-space. Each particle represents a single local transmission lineage, with instantaneous velocity in size-space given by its growth rate dl/dt, which we assume to be homogeneous across lineages of the same size at a given time. Variation in local growth rate as a function of lineage size leads to different size distributions, e.g., slower growth rate of smaller lineages tends to result in a lineage size distribution that decays with increasing lineage size. (B) Probability of detecting a local transmission lineage from a random sample of local infections (total number of infections $N_I = 10^6$), as a function of lineage size l (y-axis) and sampling proportion s (x-axis). (C) Results from simulated lineage detection under a simple deterministic model assuming no recovery process and no depletion of susceptibles, with local exponential growth at rate $r = 0.25$. Each row presents results from simulations under a different scenario of importation dynamics. The three scenarios (see labelled boxes on the far-right of each row) are: a constant importation rate (top row), an exponentially increasing importation rate (middle row) and an exponentially

decreasing importation rate (bottom row). See figure and text for parameter details. The left hand column compares the cumulative frequency distributions of lineage sizes from simulations (coloured circles) with those obtained from analytical solutions (dashed black lines). Results are shown for two different observation times (blue: T=20; orange: T=40). The middle column shows the median proportion of lineages detected (or lineage detection probability r_d) at different sampling proportions. Results are shown for two different observation times (blue: T=20; orange: T=40). The right hand column compares the true importation rate (solid grey line) with the inferred importation rate (dotted lines) at different sampling proportions. Results are again shown for two different observation times; vertical lines indicate the times when lineage detection is simulated (blue: T=20; orange: T=40). The inset on row 2 enlarges the same data in the interval $0 \leq t \leq 25$.

4.6 Lineage detection under local exponential growth

Having derived an expression (Eq. 6) that describes the time evolution of the underlying lineage size distribution in an infected population experiencing local exponential growth (with no recovery process and no depletion of susceptibles), next we consider the implications of such an evolution on lineage detection. Specifically, we consider the impact of changes in the lineage size distribution on (i) the proportion of lineages and therefore viral importation events that are likely to be detected during a period of observation, and (ii) trends in the inferred importation rate over time. To do so, we construct a simple deterministic model in which new lineages of size $l = 1$ are introduced at rate $M(t)$. Each new lineage is assumed to undergo deterministic exponential growth with no recovery of infected individuals. We run each simulation up to a predefined time T, at which point we perform random sampling of all infected individuals and identify their corresponding local transmission lineages to simulate the inference process of a phylogeographic reconstruction (see Section C.4 in Appendix C for details).

Using this simple model, we first verify the analytical results derived above by examining the lineage size distributions that result from different importation dynamic scenarios. Given the discrete nature of lineage size, it is preferable to compare the

(inverse) cumulative lineage size distribution (denoted by $C(l,t)$, i.e. the number of lineages of size $\geq l$ at a given time t) (35), instead of the actual lineage size distribution $n(l,t)$. The corresponding analytical expression for $C(l,t)$ under the assumption of a constant importation rate can be found by integrating Eq. 6 from l up to $l_{max}(T) = e^{rT}$ (the maximum lineage size attainable at time of observation T), giving

$$C(l,T) = M_0 \ln[l_{max}(T)/l] \tag{7}$$

under the assumption of a constant rate of importation, and

$$C(l,T) = M_0 e^{mT}\left[l_{max}(T)^{-m/r} - l^{-m/r}\right](r/m) \tag{8}$$

under the assumption of an exponentially increasing ($m > 0$) or decreasing ($m < 0$) rate of importation.

From Fig. 4.2C (left column), we see that the simulated lineage size distributions are in good agreement with the analytical predictions, with some deviations at small lineage sizes, likely due to the continuous approximation of discrete lineage size used in our analytical derivation (see Section C.2 in Appendix C). Interestingly, in the case of an exponentially decreasing importation rate, we observe that $C(l,T)$ approaches the same value in the limit $l \to 1$ at both times of observation (Fig. 4.2C). This can be explained by noting that $C(l = 1,T) = N_{lineages}(T)$, i.e. the total number of extant lineages in the local infected population at time T. As a result of the rapidly decreasing rate of importation, the total number of lineages remains fixed between $t = 20$ and $t = 40$ while the lineages that have already been seeded continue to grow to larger sizes (as indicated by the apparent rightward shift of the observed cumulative frequency distribution, $C(l,T)$, between T=20 and T=40). These observations together indicate a lineage size distribution that increases in density with lineage size, as expected given $|m| > r$ ($m = -0.3, r = 0.25$).

Given the different lineage size distributions that result from different (true) importation rates, it is unsurprising that we would observe different lineage detection probabilities r_d and, more specifically, different behaviours in r_d as a function of sampling proportion s (Fig. 4.2C, middle column). We find that when importation rate is exponentially decreasing, r_d rapidly approaches 1 at small s ($\lesssim 0.01$); whereas when importation is exponentially increasing, r_d only exceeds 50% at $s \approx 0.2$. This observed difference can be explained by considering the lineage size distribution: the distribution resulting from an exponentially decreasing importation rate has a greater proportion of larger lineages, whereas a right-skewed distribution with a long tail is predicted for a constant or exponentially increasing importation rate. This difference is accentuated at large T, as lineages that have already been seeded continue to grow, while the total number of extant lineages approaches a finite value if importation rate is exponentially decreasing, but increases over time for a constant or exponentially increasing importation (Fig. 4.2C, left column).

Lineages seeded earlier are more likely to be detected at a given time of observation T given our assumption of local exponential growth, as they have more time to grow to larger sizes than recently seeded lineages. We observe this effect in Fig. 4.2C (right column), where the proportion of detected importation events decreases through time. Consequently the trends in inferred importation are consistently downwardly biased compared to the true importation rate, with greater discrepancy close to the time of observation. This also leads to different inferred importation trends depending on the time of observation; for example, in the case of exponentially increasing importation (Fig. 4.2C, right column, middle row), an early analysis at T=20 at low sampling proportions (e.g. $s \lesssim 0.1$) would conclude erroneously that the intensity of viral importation has started to slow or even decrease, while a later observation and analysis at T=40 would

indicate a continually increasing importation rate at T=20. Note that the extent of this discrepancy also depends on the rate of local lineage growth, with slower growth giving rise to less variation in lineage size and therefore a less pronounced reduction in inferred importation rate close to the time of observation.

4.7 Lineage detection under constant and time-varying local contact rates with recovery

We have assumed so far that the depletion of susceptibles is negligible, which is only likely to be valid during the early stages of an outbreak when the number of infected individuals is small. Further, we have assumed that (i) there is no recovery process, and (ii) each local transmission lineage grows deterministically following its introduction. Together, these two assumptions imply that lineages that are introduced earlier will always grow to larger sizes than those introduced later, regardless of local transmission dynamics. In practice, this is unlikely to be true, especially during periods of low transmission intensity (e.g., due to depletion of susceptibles or epidemic control interventions), when lineages might be subject to stochastic extinction and stop growing soon after introduction (54).

To explore the impact of these assumptions, we construct an agent-based model with a stochastic transmission process in which, at each time step, susceptible individuals become infected with a certain probability upon contact with an infectious individual. Individuals are assumed to mix randomly regardless of their infection status. We also assume a stochastic recovery process, by which infectious individuals recover and become immune to further infection probabilistically at a constant rate. Following the same procedure as in the previous section, we simulate the detection of viral importation events by randomly sampling infected individuals at a given time of observation (T) and

identifying their corresponding local transmission lineages. Here we keep the rate of viral importation constant (100 infectious travellers arriving per day) and instead vary the local transmission dynamics. Specifically, we explore the detection of transmission lineages under two scenarios: (i) fixed local transmission conditions and (ii) time-varying local transmission conditions, in which the contact rate changes through time following a sigmoidal trajectory (see Table C.2 in Appendix C for details).

Results of the first scenario are shown in Fig. 4.3. As in the deterministic model (Fig. 4.2C), the lineage detection probability r_d varies with sampling proportion s (Figs. 4.3A-C). Notably, we find a slower increase of r_d with s when individuals can recover from infection, as compared to without recovery, at both T=10 (Fig. 4.3A) and T=30 (Fig. 4.3B). This is unsurprising, as the recovery of infectious individuals leads to slower lineage growth and therefore smaller lineages at the time of observation. However, this pattern is reversed at T=50 (Fig. 4.3C); this can be explained by the depletion of susceptibles after T=50 under the assumption of no recovery (Fig. 4.3D, blue line), resulting in later-introduced lineages being smaller at observation and therefore less likely to be detected. In some cases these later-introduced lineages immediately become extinct (see Fig. C.2 in Appendix C). In contrast, the rate of depletion of the susceptible population is slower when individuals can recover from infection (Fig. 4.3H, blue line), resulting in larger lineages at later time points.

This effect is also apparent in the inferred importation rates, especially at T=50 under the assumption of no recovery (Fig. 4.3G), for which we observe a sharp drop in the inferred rate at $t \approx 25$ as a result of rapid depletion of susceptibles, resulting in later-introduced lineages being smaller and thus less likely to be detected. Conversely, in the case of a constant recovery rate, we observe a more gradual decline in the inferred importation rate (Fig. 4.3K).

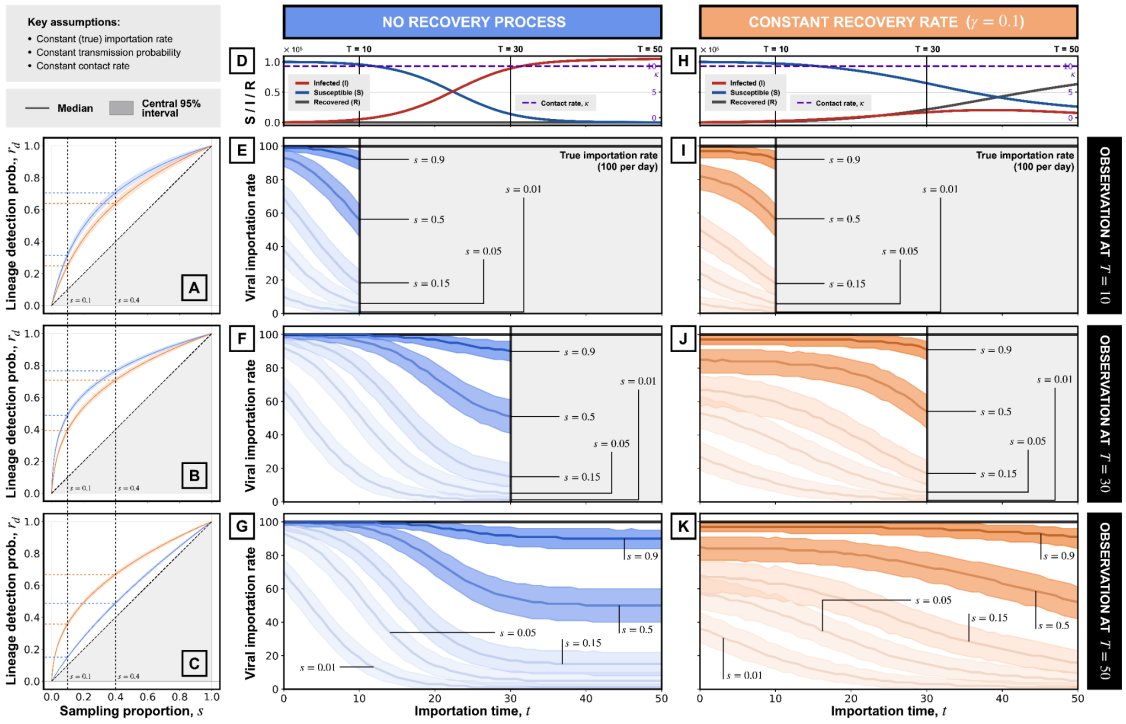


Fig. 4.3. Simulated lineage detection in a stochastic agent-based model assuming a constant local contact rate. (D, H) Simulated epidemic dynamics, showing the number of infected (I; red solid line), susceptible (S; blue solid line) and recovered (R; black solid line) individuals through time, averaged across 200 simulation replicates. Simulation parameters: importation rate $M(t) = M_0 = 100$ per day; transmission probability $\beta = 0.025$ per contact; contact rate $\kappa = 10$ per day. The left column (A-C) shows the proportion of lineages detected for different sampling proportions (across simulation replicates). Solid lines and shaded regions represent the median and the central 95% interval, respectively, of the proportion of detected lineages (with blue and orange indicating results from simulations assuming no recovery process and a constant recovery rate, respectively). The middle (E-G) and right (I-K) columns show the inferred importation rate through time (across simulation replicates), assuming no recovery process (blue) and a constant recovery rate (orange), respectively. Results are shown for different sampling proportions, s (shading transparency varies with s). Solid lines and shaded regions show the median and central 95% interval, respectively, of the inferred importation rate. Solid black lines represent the true importation rate (100 per day). Each row shows results for a different observation time, as indicated by labels on the far-right (T=10, 30, 50).

Fig. 4.4 shows results for the second simulated scenario, in which the contact rate κ varies over time. We observe substantial differences in the lineage detection probability r_d between simulations in which κ increases versus those in which κ decreases. In the case

of increasing κ, there is an almost linear relationship between r_d and s at T=10 (Fig. 4.4A, blue line), i.e. when κ is changing most rapidly (Fig. 4.4D, purple dashed line). This can be explained by a substantially lower contact rate before T=10 which gives rise to slower lineage growth and therefore more frequent lineage extinction and smaller lineages at T=10; consequently the lineage detection probability r_d scales almost linearly with sampling proportion s, as described in Eq. 2. This effect is also reflected in the observation that the inferred importation rate barely changes through time when κ is increasing (Fig. 4.4E), whereas the inferred rate abruptly declines into the recent past when κ is decreasing (Fig. 4.4I). Again, this is because the higher contact rate before $t = 10$ results in earlier-introduced lineages being larger and therefore more likely to be detected compared to later-introduced lineages.

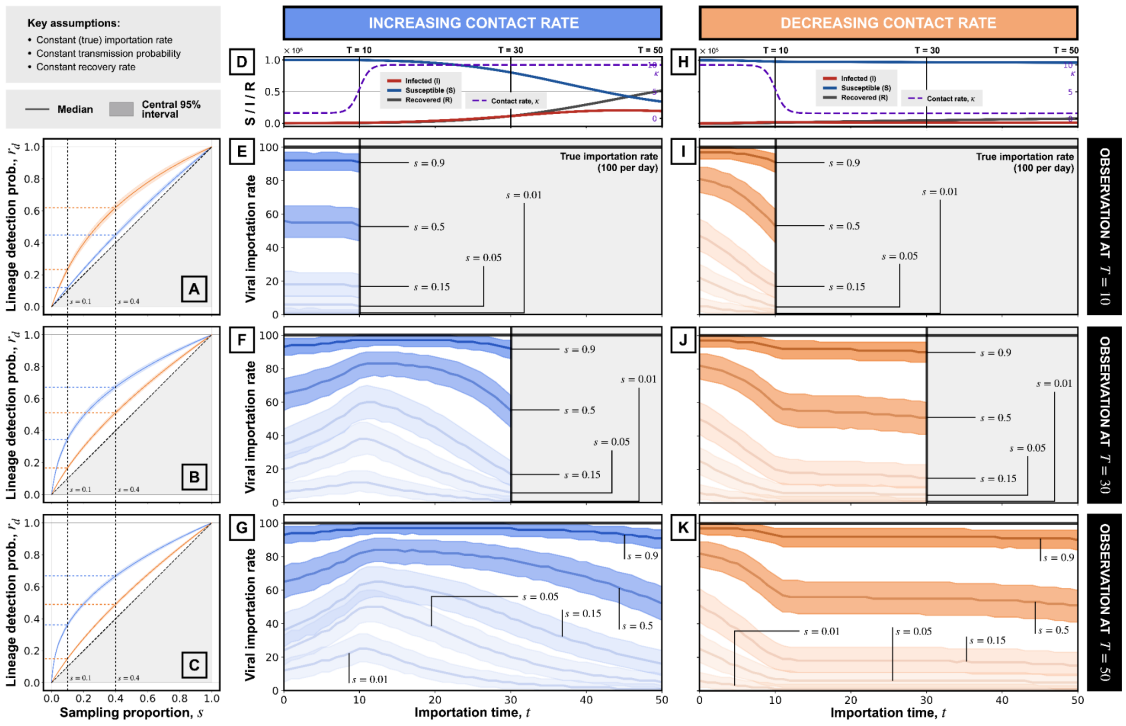


Fig. 4.4. Simulated lineage detection in a stochastic agent-based model assuming time-varying contact rates. (D, H) Simulated epidemic dynamics, showing the number of infected (I; red solid line), susceptible (S; blue solid line) and recovered (R; black solid line) individuals through time, averaged across 200 simulation replicates. Simulation parameters: importation rate $M(t) = M_0 = 100$ per day; transmission probability $\beta =$

0.025 per contact; recovery rate $\gamma = 0.1$ per day. The left column (A-C) shows the proportion of lineages detected for different sampling proportions (across simulation replicates). Solid lines and shaded regions represent the median and the central 95% interval, respectively, of the proportion of detected lineages (with blue and orange indicating results from simulations assuming no recovery process and a constant recovery rate, respectively). The middle (E-G) and right (I-K) columns show the inferred importation rate through time (across simulation replicates), assuming an increasing (blue) and decreasing (orange) local contact rate, respectively. Results are shown for different sampling proportions, s (shading transparency varies with s). Solid lines and shaded regions show the median and central 95% interval, respectively, of the inferred importation rate. Solid black lines represent the true importation rate (100 per day). Each row shows results for a different observation time, as indicated by labels on the far-right (T=10, 30, 50).

At T=30 and T=50, these observed patterns are reversed, as expected given the abrupt changes in κ at T=10. Notably, we see that the lineage detection probability r_d increases more slowly with s in the case of decreasing κ (Fig. 4.4B, C), as a result of new lineages either growing very slowly or becoming extinct soon after introduction. Interestingly, in the case of increasing κ, the higher contact rate following T=10 leads to an apparent increase in the inferred importation rate initially (Fig. 4.4F, G), as later-introduced lineages are more likely to survive past T=10 beyond which they can continue to grow at higher κ. This is then followed by a rapid decline towards the time of observation, again due to variation in lineage size resulting from different importation times; whereas in the case of decreasing κ, we observe a slower decline (Fig. 4.4J, K) due to smaller size variation among lineages introduced at small κ after T=10. Importantly, these abrupt changes in the inferred rate lead to mischaracterisation of the underlying trend of viral importation - despite the true importation rate being constant through time, the degree to which it is underestimated varies substantially over time as a function of both the time-varying local transmission dynamics (the contact rate in the simulations) and sampling proportion s.

4.8 Discussion

Phylogeographic analyses have enabled the detection and enumeration of viral importation events, necessary for understanding the dispersal patterns of an emerging pathogen and evaluating the impact of public health interventions, especially those designed to limit further spatial dissemination. Standard phylogeographic approaches, such as discrete trait analysis, rely on the identification of local transmission lineages consisting of local secondary infections that result from arriving infectious travellers. Given that only a small fraction of local infections are typically sequenced in an outbreak, it is unsurprising that the number of viral importation events inferred from a given genomic dataset underestimates the true number. However, the underlying mechanism that leads to this underestimation and the degree to which this occurs under different outbreak conditions has not been well characterised.

Here, we showed that the proportion of viral importation events detected from a sample of local infections depends on the underlying lineage size distribution, which in turn is determined by the coupled dynamics of viral importation and local transmission. By modelling lineage growth as particle movements in a continuous size-space, we found that the lineage size distribution of an infected population undergoing local exponential growth can be described by a power-law (consistent with empirical observations of SARS-CoV-2 spread made by du Plessis et al. (35) and other similar studies (33, 34, 46)), with the exponent depending on both local lineage growth rate and viral importation intensity over time. More generally, when local lineage growth is slower than the rate at which new lineages are being introduced, there is an accumulation of smaller lineages and, as a result, the proportion of lineages that we expect to detect increases only slowly with sample size. In contrast, when local transmission is intense, lineages grow rapidly

compared to the importation rate, resulting in larger lineages that are more likely to be detected, even at low sampling proportions.

Using an agent-based stochastic model, we found that the inferred importation rate can be substantially downwardly biased, especially at low sampling intensities. This is due to variation in the lineage size distribution at the time of observation, which itself is determined by (i) differences in the time elapsed since introduction, with recently-introduced lineages being smaller and therefore harder to detect, and (ii) stochastic extinction of local lineages during periods of low transmission, potentially resulting in smaller lineages at the time of observation despite early importation.

Our findings have implications for the interpretation of viral movement estimates from phylogeographic analyses. First, while it is known that these estimates represent only approximate lower bounds of the true number of viral importation events, our results indicate that the degree of underestimation can vary substantially between outbreak contexts, especially when the fraction of infections that are genomically sequenced is small. This is important in the context of studies that investigate the source-sink dynamics of virus spread, in which the relative intensities of viral movement between locations are of primary interest. Our findings suggest that such estimates could be substantially biased and their interpretation requires careful consideration of the differences in local transmission dynamics between locations, especially when sampling fractions are low ($\lesssim$ 5%). Second, our observation that inferred viral importation dynamics can depend on local transmission intensity has implications for the interpretation of such estimates in the context of policy evaluation. For instance, a phylogeographic analysis of virus genomes from a country that experienced a recent increase in transmission intensity might conclude erroneously that the resurgence in case numbers was driven by increased viral importation - in reality, the true importation intensity might have remained constant, with the apparent

increase being an artefact resulting from later-introduced lineages being larger and therefore more likely to be detected (see Fig. 4.4F, G). Conversely, an increase in viral importation intensity shortly before the time of genome sampling is unlikely to be detected, due to the lower detection probability of recently-introduced (and therefore smaller) lineages. Studies that evaluate the efficacy of public health policies, especially those intended to prevent or reduce viral importation, should therefore be cautious in interpreting the temporal dynamics of viral movement estimates from phylogeographic analyses when sampling intensities are low.

There are several limitations to our approach. Throughout the study we assumed sufficient sampling of virus genetic diversity at the source location, such that each independent local lineage can be uniquely resolved phylogenetically. This assumption rarely holds in practice due to both limited sequencing capacity and under-reporting, which will apply equally to both the source and recipient location. Violation of this assumption leads to an aggregation of independent local lineages and therefore an underestimation of the true number of viral importation events, even in the limit of complete sampling of infected individuals at the recipient location. Future studies should explore the extent to which the aggregation of lineages occurs given different sampling proportions at the source location and the effect of variation in relevant epidemiological and evolutionary parameters, as well as potential mitigation strategies when presented with such sampling bias (55). Secondly, we also assumed that each infected individual at the recipient location is equally likely to be sampled. This may not hold in practice due to targeted or biased sampling across space (e.g., unequal access to testing or healthcare services, airport screening, intense sampling following contact-tracing) or time (e.g., changes in public awareness, testing fatigue, sequencing capacity being overwhelmed during intense transmission). This will likely lead to further biases in the inferred viral

movements, especially if the sampling probability is correlated with viral importation rate or local transmission intensity; future studies involving empirical data should consider the potential impact of these sampling heterogeneities. Thirdly, by randomly sampling infected individuals only at the end of an outbreak, we have implicitly assumed that local transmission dynamics are unaffected by the sampling process. While this is a reasonable assumption for low-intensity sampling (the focus of this study), future work could consider different types of sampling or sequencing efforts (for example, when infected individuals are required to self-isolate following a positive test result) and their impact on transmission dynamics at high sampling intensities. Finally, we assumed that statistical uncertainties associated with phylogenetic and phylogeographic inference are negligible and can be ignored. In reality, the choice of molecular clock models (56), tree priors (57), and phylogeographic methods (e.g., structured coalescent (58–60), continuous random walk (24)) can all impact estimates of viral lineage movement. Further, we assumed that the timing of importation associated with each local transmission lineage is known - in practice, the exact time when an infectious traveller arrives can be obscured by both importation-lag (i.e. time difference between importation and first inferred local transmission event) and detection-lag (i.e. time difference between importation and first detected local infection; see Fig. 4.1B). These lags remain poorly understood and have received little attention in the literature. A more comprehensive evaluation of our findings in the context of these uncertainties should be explored.

As large-scale pathogen genomic sequencing becomes more common in global public health, phylogeographic analysis is likely to play an increasingly important role in reconstructing and monitoring the spread of emerging pathogens. While this study primarily focuses on characterising the sampling process underlying the detection of viral importation events, our findings lay the groundwork for developing more robust inference

methods to address these biases. Such methods are needed to derive more accurate insights from viral genomic data and to inform the design of effective, timely interventions in response to future epidemics.

4.9 References

1. Adam, D.C., Wu, P., Wong, J.Y., Lau, E.H.Y., Tsang, T.K., Cauchemez, S., Leung, G.M. and Cowling, B.J. (2020) 'Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong', *Nature Medicine*, 26(11), pp. 1714–1719.
2. Diarra, M., Ndiaye, R., Barry, A., Talla, C., Diagne, M.M., Dia, N., Faye, J., Sarr, F.D., Gaye, A., Diallo, A., Cisse, M., Dieng, I., Fall, G., Tall, A., Faye, O., Faye, O., Sall, A.A. and Loucoubar, C. (2023) 'Analysis of contact tracing data showed contribution of asymptomatic and non-severe infections to the maintenance of SARS-CoV-2 transmission in Senegal', *Scientific Reports*, 13(1), pp. 1–9.
3. Faye, O., Boëlle, P.-Y., Heleze, E., Faye, O., Loucoubar, C., Magassouba, N. 'faly, Soropogui, B., Keita, S., Gakou, T., El Hadji Ibrahima, B., Koivogui, L., Sall, A.A. and Cauchemez, S. (2015) 'Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study', *The Lancet Infectious Diseases*, 15(3), pp. 320–326.
4. Sun, K., Wang, W., Gao, L., Wang, Y., Luo, K., Ren, L., Zhan, Z., Chen, X., Zhao, S., Huang, Y., Sun, Q., Liu, Z., Litvinova, M., Vespignani, A., Ajelli, M., Viboud, C. and Yu, H. (2021) 'Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2', *Science*, 371(6526), p. eabe2424.
5. Vazquez-Prokopec, G.M., Montgomery, B.L., Horne, P., Clennon, J.A. and Ritchie, S.A. (2017) 'Combining contact tracing with targeted indoor residual spraying significantly reduces dengue transmission', *Science Advances*, 3(2), p. e1602024.
6. Díaz-Menéndez, M., Angelo, K.M., de Miguel Buckley, R., Bottieau, E., Huits, R., Grobusch, M.P., Gobbi, F.G., Asgeirsson, H., Duvignaud, A., Norman, F.F., Javelle, E., Epelboin, L., Rothe, C., Chappuis, F., Martinez, G.E., Popescu, C., Camprubí-Ferrer, D., Molina, I., Odolini, S., Barkati, S., Kuhn, S., Vaughan, S., McCarthy, A., Lago, M., Libman, M.D. and Hamer, D.H. (2023) 'Dengue outbreak amongst travellers returning from Cuba—GeoSentinel surveillance network, January–September 2022', *Journal of travel medicine*, 30(2), p. taac139.
7. Duvignaud, A., Stoney, R.J., Angelo, K.M., Chen, L.H., Cattaneo, P., Motta, L., Gobbi, F.G., Bottieau, E., Bourque, D.L., Popescu, C.P., Glans, H., Asgeirsson, H., Oliveira-Souto, I., Vaughan, S.D., Amatya, B., Norman, F.F., Waggoner, J., Díaz-Menéndez, M., Beadsworth, M., Odolini, S., Camprubí-Ferrer, D., Epelboin, L., Connor, B.A., Eperon, G., Schwartz, E., Libman, M., Malvy, D., Hamer, D.H., Huits, R. and GeoSentinel Network (2024) 'Epidemiology of travel-associated dengue from 2007 to 2022: A GeoSentinel analysis', *Journal of travel medicine*, 31(7), taae089.
8. Kraemer, M.U.G., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D.M., Open COVID-19 Data Working Group, du Plessis, L., Faria, N.R., Li, R., Hanage, W.P., Brownstein, J.S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O.G. and Scarpino, S.V. (2020) 'The effect of human mobility and control measures on the COVID-19 epidemic in China', *Science*, 368(6490), pp. 493–497.
9. Kucharski, A.J., Chung, K., Aubry, M., Teiti, I., Teissier, A., Richard, V., Russell, T.W., Bos, R., Olivier, S. and Cao-Lormeau, V.-M. (2023) 'Real-time surveillance of international SARS-CoV-2 prevalence using systematic traveller arrival screening: An observational study', *PLOS Medicine*, 20(9), p. e1004283.
10. Mayer, A.B., Consigny, P.H., Grobusch, M.P., Camprubí-Ferrer, D., Huits, R. and Rothe, C. (2023) 'Chikungunya in returning travellers from Bali - A GeoSentinel case series', *Travel medicine and infectious disease*, 52, 102543.
11. Septfons, A., Leparc-Goffart, I., Couturier, E., Franke, F., Deniau, J., Balestier, A., Guinard, A., Heuzé, G., Liebert, A.H., Mailles, A., Ndong, J.R., Poujol, I., Raguet,

S., Rousseau, C., Saidouni-Oulebsir, A., Six, C., Subiros, M., Servas, V., Terrien, E., Tillaut, H., Viriot, D., Watrin, M., Wyndels, K. and the Zika Surveillance Working Group (2016) 'Travel-associated and autochthonous Zika virus infection in mainland France, 1 January to 15 July 2016', *Eurosurveillance*, 21(32), p. 30315.
12. Eggo, R.M., Cauchemez, S. and Ferguson, N.M. (2011) 'Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States', *Journal of The Royal Society Interface*, 8(55), pp. 233-243.
13. Gog, J.R., Ballesteros, S., Viboud, C., Simonsen, L., Bjornstad, O.N., Shaman, J., Chao, D.L., Khan, F. and Grenfell, B.T. (2014) 'Spatial Transmission of 2009 Pandemic Influenza in the US', *PLOS Computational Biology*, 10(6), p. e1003635.
14. Viboud, C., Bjørnstad, O.N., Smith, D.L., Simonsen, L., Miller, M.A. and Grenfell, B.T. (2006) 'Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza', *Science*, 312(5772), pp. 447-451.
15. Bajardi, P., Poletto, C., Ramasco, J.J., Tizzoni, M., Colizza, V. and Vespignani, A. (2011) 'Human Mobility Networks, Travel Restrictions, and the Global Spread of 2009 H1N1 Pandemic', *PLOS ONE*, 6(1), p. e16591.
16. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J.J. and Vespignani, A. (2009) 'Multiscale mobility networks and the spatial spreading of infectious diseases', *Proceedings of the National Academy of Sciences*, 106(51), pp. 21484–21489.
17. Pei, S., Kandula, S., Cascante Vega, J., Yang, W., Foerster, S., Thompson, C., Baumgartner, J., Ahuja, S.D., Blaney, K., Varma, J.K., Long, T. and Shaman, J. (2022) 'Contact tracing reveals community transmission of COVID-19 in New York City', *Nature Communications*, 13(1), pp. 1–8.
18. Camargo, C.Q., Bright, J. and Hale, S.A. (2019) 'Diagnosing the performance of human mobility models at small spatial scales using volunteered geographical information', *Royal Society Open Science*, 6(11), 191034.
19. Masucci, A.P., Serras, J., Johansson, A. and Batty, M. (2013) 'Gravity vs radiation model: on the importance of scale and heterogeneity in commuting flows', *Physical Review E*, 88(2), p. 022812.
20. Yan, X.-Y., Wang, W.-X., Gao, Z.-Y. and Lai, Y.-C. (2017) 'Universal model of individual and population mobility on diverse spatial scales', *Nature Communications*, 8(1), pp. 1–9.
21. Hill, V., Ruis, C., Bajaj, S., Pybus, O.G. and Kraemer, M.U.G. (2021) 'Progress and challenges in virus genomic epidemiology', *Trends in parasitology*, 37(12), pp. 1038–1049.
22. Aggarwal, D., Warne, B., Jahun, A.S., Hamilton, W.L., Fieldman, T., du Plessis, L., Hill, V., Blane, B., Watkins, E., Wright, E., Hall, G., Ludden, C., Myers, R., Hosmillo, M., Chaudhry, Y., Pinckert, M.L., Georgana, I., Izuagbe, R., Leek, D., Nsonwu, O., Hughes, G.J., Packer, S., Page, A.J., Metaxaki, M., Fuller, S., Weale, G., Holgate, J., Brown, C.A., Howes, R., McFarlane, D., Dougan, G., Pybus, O.G., Angelis, D.D., Maxwell, P.H., Peacock, S.J., Weekes, M.P., Illingworth, C., Harrison, E.M., Matheson, N.J. and Goodfellow, I.G. (2022) 'Genomic epidemiology of SARS-CoV-2 in a UK university identifies dynamics of transmission', *Nature communications*, 13(1), pp. 1–16.
23. Chen, Z., Tsui, J.L.-H., Gutierrez, B., Busch Moreno, S., du Plessis, L., Deng, X., Cai, J., Bajaj, S., Suchard, M.A., Pybus, O.G., Lemey, P., Kraemer, M.U.G. and Yu, H. (2024) 'COVID-19 pandemic interventions reshaped the global dispersal of seasonal influenza viruses', *Science*, 386(6722), p. eadq3003.

24. Lemey, P., Rambaut, A., Welch, J.J. and Suchard, M.A. (2010) 'Phylogeography takes a relaxed random walk in continuous space and time', *Molecular biology and evolution*, 27(8), pp. 1877–1885.
25. Lemey, P., Rambaut, A., Drummond, A.J. and Suchard, M.A. (2009) 'Bayesian Phylogeography Finds Its Roots', *PLOS Computational Biology*, 5(9), p. e1000520.
26. Alessandretti, L., Aslak, U. and Lehmann, S. (2020) 'The scales of human mobility', *Nature*, 587(7834), pp. 402–407.
27. González, M.C., Hidalgo, C.A. and Barabási, A.-L. (2008) 'Understanding individual human mobility patterns', *Nature*, 453(7196), pp. 779–782.
28. Kraemer, M.U.G., Sadilek, A., Zhang, Q., Marchal, N.A., Tuli, G., Cohn, E.L., Hswen, Y., Perkins, T.A., Smith, D.L., Reiner, R.C., Jr and Brownstein, J.S. (2020) 'Mapping global variation in human mobility', *Nat Hum Behav*, 4(8), pp. 800–810.
29. Hué, S., Pillay, D., Clewley, J.P. and Pybus, O.G. (2005) 'Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups', *Proceedings of the National Academy of Sciences of the United States of America*, 102(12), pp. 4425–4429.
30. Baillie, G.J., Galiano, M., Agapow, P.-M., Myers, R., Chiam, R., Gall, A., Palser, A.L., Watson, S.J., Hedge, J., Underwood, A., Platt, S., McLean, E., Pebody, R.G., Rambaut, A., Green, J., Daniels, R., Pybus, O.G., Kellam, P. and Zambon, M. (2012) 'Evolutionary Dynamics of Local Pandemic H1N1/2009 Influenza Virus Lineages Revealed by Whole-Genome Analysis', *Journal of virology*, 86(1), pp. 11-18.
31. Gu, H., Xie, R., Adam, D.C., Tsui, J.L.-H., Chu, D.K., Chang, L.D.J., Cheuk, S.S.Y., Gurung, S., Krishnan, P., Ng, D.Y.M., Liu, G.Y.Z., Wan, C.K.C., Cheng, S.S.M., Edwards, K.M., Leung, K.S.M., Wu, J.T., Tsang, D.N.C., Leung, G.M., Cowling, B.J., Peiris, M., Lam, T.T.Y., Dhanasekaran, V. and Poon, L.L.M. (2022) 'Genomic epidemiology of SARS-CoV-2 under an elimination strategy in Hong Kong', *Nature communications*, 13(1), pp. 1–10.
32. Han, A.X., Kozanli, E., Koopsen, J., Vennema, H., RIVM COVID-19 molecular epidemiology group, Aarts, L., Bos, S., van den Brandt, A., van den Brink, S., Cremer, J., Freriks, K., Jaarsma, R., Schmitz, D., Then, E., van der Veer, B., Wijsman, L., Zwagemaker, F., Hajji, K., Kroneman, A., van Walle, I., Klinkenberg, D., Wallinga, J., Russell, C.A., Eggink, D. and Reusken, C. (2022) 'Regional importation and asymmetric within-country spread of SARS-CoV-2 variants of concern in the Netherlands', *eLife*, 11, p. e78770.
33. McCrone, J.T., Hill, V., Bajaj, S., Pena, R.E., Lambert, B.C., Inward, R., Bhatt, S., Volz, E., Ruis, C., Dellicour, S., Baele, G., Zarebski, A.E., Sadilek, A., Wu, N., Schneider, A., Ji, X., Raghwani, J., Jackson, B., Colquhoun, R., O'Toole, Á., Peacock, T.P., Twohig, K., Thelwall, S., Dabrera, G., Myers, R., Faria, N.R., Huber, C., Bogoch, I.I., Khan, K., du Plessis, L., Barrett, J.C., Aanensen, D.M., Barclay, W.S., Chand, M., Connor, T., Loman, N.J., Suchard, M.A., Pybus, O.G., Rambaut, A. and Kraemer, M.U.G. (2022) 'Context-specific emergence and growth of the SARS-CoV-2 Delta variant', *Nature*, 610(7930), pp. 154–160.
34. Tsui, J.L.-H., McCrone, J.T., Lambert, B., Bajaj, S., Inward, R.P.D., Bosetti, P., Pena, R.E., Tegally, H., Hill, V., Zarebski, A.E., Peacock, T.P., Liu, L., Wu, N., Davis, M., Bogoch, I.I., Khan, K., Kall, M., Abdul Aziz, N.I.B., Colquhoun, R., O'Toole, Á., Jackson, B., Dasgupta, A., Wilkinson, E., de Oliveira, T., COVID-19 Genomics UK (COG-UK) consortium¶, Connor, T.R., Loman, N.J., Colizza, V., Fraser, C., Volz, E., Ji, X., Gutierrez, B., Chand, M., Dellicour, S., Cauchemez, S., Raghwani, J., Suchard, M.A., Lemey, P., Rambaut, A., Pybus, O.G. and Kraemer, M.U.G. (2023)

'Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1', *Science*, 381(6655), pp. 336–343.
35. du Plessis, L., McCrone, J.T., Zarebski, A.E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T.R., Faria, N.R., Jackson, B., Loman, N.J., O'Toole, Á., Nicholls, S.M., Parag, K.V., Scher, E., Vasylyeva, T.I., Volz, E.M., Watts, A., Bogoch, I.I., Khan, K., COVID-19 Genomics UK (COG-UK) Consortium, Aanensen, D.M., Kraemer, M.U.G., Rambaut, A. and Pybus, O.G. (2021) 'Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK', *Science*, 371(6530), pp. 708–712.
36. Dudas, G., Carvalho, L.M., Rambaut, A. and Bedford, T. (2018) 'MERS-CoV spillover at the camel-human interface', *eLife*, 7, p. e31257.
37. da Silva Filipe, A., Shepherd, J.G., Williams, T., Hughes, J., Aranday-Cortes, E., Asamaphan, P., Ashraf, S., Balcazar, C., Brunker, K., Campbell, A., Carmichael, S., Davis, C., Dewar, R., Gallagher, M.D., Gunson, R., Hill, V., Ho, A., Jackson, B., James, E., Jesudason, N., Johnson, N., McWilliam Leitch, E.C., Li, K., MacLean, A., Mair, D., McAllister, D.A., McCrone, J.T., McDonald, S.E., McHugh, M.P., Morris, A.K., Nichols, J., Niebel, M., Nomikou, K., Orton, R.J., O'Toole, Á., Palmarini, M., Parcell, B.J., Parr, Y.A., Rambaut, A., Rooke, S., Shaaban, S., Shah, R., Singer, J.B., Smollett, K., Starinskij, I., Tong, L., Sreenu, V.B., Wastnedge, E., COVID-19 Genomics UK (COG-UK) Consortium, Holden, M.T.G., Robertson, D.L., Templeton, K. and Thomson, E.C. (2021) 'Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland', *Nature microbiology*, 6(1), pp. 112–122.
38. Komissarov, A.B., Safina, K.R., Garushyants, S.K., Fadeev, A.V., Sergeeva, M.V., Ivanova, A.A., Danilenko, D.M., Lioznov, D., Shneider, O.V., Shvyrev, N., Spirin, V., Glyzin, D., Shchur, V. and Bazykin, G.A. (2021) 'Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak in Russia', *Nature communications*, 12(1), pp. 1–13.
39. Lemey, P., Ruktanonchai, N., Hong, S.L., Colizza, V., Poletto, C., Van den Broeck, F., Gill, M.S., Ji, X., Levasseur, A., Oude Munnink, B.B., Koopmans, M., Sadilek, A., Lai, S., Tatem, A.J., Baele, G., Suchard, M.A. and Dellicour, S. (2021) 'Untangling introductions and persistence in COVID-19 resurgence in Europe', *Nature*, 595(7869), pp. 713–717.
40. Brito, A.F., Semenova, E., Dudas, G., Hassler, G.W., Kalinich, C.C., Kraemer, M.U.G., Ho, J., Tegally, H., Githinji, G., Agoti, C.N., Matkin, L.E., Whittaker, C., Howden, B.P., Sintchenko, V., Zuckerman, N.S., Mor, O., Blankenship, H.M., de Oliveira, T., Lin, R.T.P., Siqueira, M.M., Resende, P.C., Vasconcelos, A.T.R., Spilki, F.R., Aguiar, R.S., Alexiev, I., Ivanov, I.N., Philipova, I., Carrington, C.V.F., Sahadeo, N.S.D., Branda, B., Gurry, C., Maurer-Stroh, S., Naidoo, D., von Eije, K.J., Perkins, M.D., van Kerkhove, M., Hill, S.C., Sabino, E.C., Pybus, O.G., Dye, C., Bhatt, S., Flaxman, S., Suchard, M.A., Grubaugh, N.D., Baele, G. and Faria, N.R. (2022) 'Global disparities in SARS-CoV-2 genomic surveillance', *Nature Communications*, 13(1), pp. 1–13.
41. Ma, Q., Liu, J., Liu, Q., Kang, L., Liu, R., Jing, W., Wu, Y. and Liu, M. (2021) 'Global Percentage of Asymptomatic SARS-CoV-2 Infections Among the Tested Population and Individuals With Confirmed COVID-19 Diagnosis: A Systematic Review and Meta-analysis', *JAMA network open*, 4(12), p. e2137257.
42. Sah, P., Fitzpatrick, M.C., Zimmer, C.F., Abdollahi, E., Juden-Kelly, L., Moghadas, S.M., Singer, B.H. and Galvani, A.P. (2021) 'Asymptomatic SARS-CoV-2 infection:

A systematic review and meta-analysis', *Proceedings of the National Academy of Sciences*, 118(34), p. e2109229118.
43. Subramanian, R., He, Q. and Pascual, M. (2021) 'Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity', *Proceedings of the National Academy of Sciences*, 118(9), p. e2019716118.
44. Bergeri, I., Whelan, M.G., Ware, H., Subissi, L., Nardone, A., Lewis, H.C., Li, Z., Ma, X., Valenciano, M., Cheng, B., Al Ariqi, L., Rashidian, A., Okeibunor, J., Azim, T., Wijesinghe, P., Le, L.-V., Vaughan, A., Pebody, R., Vicari, A., Yan, T., Yanes-Lane, M., Cao, C., Clifton, D.A., Cheng, M.P., Papenburg, J., Buckeridge, D., Bobrovitz, N., Arora, R.K., Van Kerkhove, M.D. and Unity Studies Collaborator Group (2022) 'Global SARS-CoV-2 seroprevalence from January 2020 to April 2022: A systematic review and meta-analysis of standardized population-based studies', *PLOS Medicine*, 19(11), p. e1004107.
45. Havers, F.P., Reed, C., Lim, T., Montgomery, J.M., Klena, J.D., Hall, A.J., Fry, A.M., Cannon, D.L., Chiang, C.F., Gibbons, A., Krapiunaya, I., Morales-Betoulle, M., Roguski, K., Rasheed, M.A.U., Freeman, B., Lester, S., Mills, L., Carroll, D.S., Owen, S.M., Johnson, J.A., … Thornburg, N.J. (2020) 'Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23–May 12, 2020', *JAMA Internal Medicine* [Online ahead of print]. Available at: https://doi.org/10.1001/jamainternmed.2020.4130.
46. Murall, C.L., Fournier, E., Galvez, J.H., N'Guessan, A., Reiling, S.J., Quirion, P.-O., Naderi, S., Roy, A.-M., Chen, S.-H., Stretenowich, P., Bourgey, M., Bujold, D., Gregoire, R., Lepage, P., St-Cyr, J., Willet, P., Dion, R., Charest, H., Lathrop, M., Roger, M., Bourque, G., Ragoussis, J., Shapiro, B.J. and Moreira, S. (2021) 'A small number of early introductions seeded widespread transmission of SARS-CoV-2 in Québec, Canada', *Genome medicine*, 13(1), pp. 1–17.
47. Chao, A. and Lee, S.-M. (1992) 'Estimating the Number of Classes via Sample Coverage', *Journal of the American Statistical Association*, 87(417), pp. 210-217.
48. Efron, B. and Thisted, R. (1976) 'Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3), pp. 435-447.
49. Fisher, R., Corbet, A. and Williams, C.B. (1943) 'The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population', *Journal of Animal Ecology*, 12(1), pp. 42-58.
50. Hubbell, S.P. (2011) *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press.
51. Volkov, I., Banavar, J.R., Hubbell, S.P. and Maritan, A. (2003) 'Neutral theory and relative species abundance in ecology', *Nature*, 424(6952), pp. 1035–1037.
52. Kundu, P.K., Cohen, I.M. and Dowling, D.R. (2012) *Fluid Mechanics*. Academic Press.
53. White, F.M. (2000) *Fluid Mechanics*. McGraw-Hill Science, Engineering & Mathematics.
54. Curran-Sebastian, J., Andersen, F.M. and Bhatt, S. (2025) 'Modelling the stochastic importation dynamics and establishment of novel pathogenic strains using a general branching processes framework', *Mathematical biosciences*, 380, p. 109352.
55. Lemey, P., Hong, S.L., Hill, V., Baele, G., Poletto, C., Colizza, V., O'Toole, Á., McCrone, J.T., Andersen, K.G., Worobey, M., Nelson, M.I., Rambaut, A. and Suchard, M.A. (2020) 'Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2', *Nature Communications*, 11(1), pp. 1–14.

56. Tay, J.H., Kocher, A. and Duchene, S. (2024) ‘Assessing the effect of model specification and prior sensitivity on Bayesian tests of temporal signal’, *PLOS Computational Biology*, 20(11), p. e1012371.
57. Featherstone, L.A., Zhang, J.M., Vaughan, T.G. and Duchene, S. (2022) ‘Epidemiological inference from pathogen genomes: A review of phylodynamic models and applications’, *Virus Evolution*, 8(1), p. veac045.
58. De Maio, N., Wu, C.-H., O’Reilly, K.M. and Wilson, D. (2015) ‘New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation’, *PLoS genetics*, 11(8), p. e1005421.
59. Müller, N.F., Rasmussen, D.A. and Stadler, T. (2017) ‘The Structured Coalescent and Its Approximations’, *Molecular Biology and Evolution*, 34(11), pp. 2970–2981.
60. Vaughan, T.G., Kühnert, D., Popinga, A., Welch, D. and Drummond, A.J. (2014) ‘Efficient Bayesian inference under the structured coalescent’, *Bioinformatics (Oxford, England)*, 30(16), pp. 2272–2279.

Appendix C:
Supplementary materials for Chapter 4

C.1 Selection criteria for relevant studies published during the COVID-19 pandemic

As described in Section 3 in the main text, we identified and selected peer-reviewed studies estimating the number of SARS-CoV-2 introductions and the intensity of viral importation at specific locations using phylogenetic and/or phylogeographic approaches. Relevant studies satisfying at least one of the following criteria were included:

1. The study provided estimates of the sampling proportion (number of viral genomes included per confirmed case) during the relevant study period.
2. The study provided the number of viral genomes that were included in the analysis and estimates of the total number of confirmed cases during the relevant study period, thereby allowing the sampling proportion to be estimated.
3. The study provided the number of viral genomes that were included in the analysis, and the total number of confirmed cases during the relevant study period could be estimated either from data provided by the authors or from external public repositories, thereby allowing the sampling proportion to be estimated.

For studies investigating the importation of specific SARS-CoV-2 variants, we used relative lineage frequencies (either provided by the study or downloaded from public repositories) to estimate variant-specific reported/confirmed case numbers. For studies considering viral movements among multiple locations, the average sampling proportion weighted by the number of cases from each location is calculated. It is important to note that these sampling proportions are intended as only rough estimates of the proportion of

infections that were sampled, given the substantial variability in reporting practices, as well as changes in case definitions and the proportion of asymptomatic infections over the course of the pandemic.

C.2 Derivation of an analytical expression describing the time evolution of lineage size distribution assuming local exponential growth

As described in Section 4 in the main text, we model the growth of local transmission lineages as particle movement in a one-dimensional continuous lineage size-space. With the definition of a transmission lineage as outlined in Section 4.2 in the Chapter 4, the time evolution of the density of these particles in size-space can be considered as solutions to the continuity equation

$$\frac{\partial n}{\partial t} + \frac{\partial}{\partial l}\left(n\frac{dl}{dt}\right) = M(t)\delta(l-1) \tag{1}$$

with the boundary condition $n(l,0) = 0$ for all $l \geq 1$, where $n = n(l)$ is the particle density at l, $M(t)$ is the true importation rate at time t, and dl/dt is the instantaneous lineage growth rate which is a function of both time and lineage size.

Assuming that the local infected population is undergoing exponential growth (e.g., during the early stages of an outbreak), such that $dl/dt = rl$ where r is a positive constant, Eq. 1 then becomes

$$\frac{\partial n}{\partial t} + \frac{\partial n}{\partial l}rl + nr = M(t)\delta(l-1) \tag{2}$$

which is a first-order PDE with a source term (right hand side).

Applying change of variables with $\epsilon = t$ and $\nu = \ln l - rt$, we obtain

$$\frac{\partial n}{\partial \epsilon} + nr = M(\epsilon)\delta(\nu + r\epsilon) \tag{3}$$

where we have also used the transformation $\delta(g(x)) = \delta(x - x_0)/|g'(x_0)|$, if $g(x)$ has a real root at $x = x_0$.

Eq. 3 can then be solved using an integrating factor to give the general solution

$$n(\epsilon, \nu) = e^{-r\epsilon} \int M(\epsilon')\delta(\nu + r\epsilon')e^{r\epsilon'} d\epsilon' + \Phi(\nu) \tag{4}$$

with $\Phi(\nu)$ being an arbitrary function of ν.

Applying the boundary condition that $n(l, t = 0) = 0$ for all $l \geq 1$, and since $\nu(l, t = 0) = \ln l$, from Eq. 4 we obtain

$$\Phi(\ln l) = -M(-\ln l\,/r)\mathrm{H}(\ln l\,/r)/l \tag{5}$$

where $\mathrm{H}(x)$ is the Heaviside function. If we further let $u = \ln l$, and given that $\mathrm{H}(u/r) = \mathrm{H}(u)$, we get

$$\Phi(u) = -M(-u/r)\mathrm{H}(u)e^{-u} \tag{6}$$

Substituting this back into the general solution (Eq. 4) and expanding the right hand side gives

$$n(\epsilon, \nu) = e^{-(r\epsilon+\nu)}M(-\nu/r)[\mathrm{H}(r\epsilon + \nu) - \mathrm{H}(\nu)] \tag{7}$$

Finally, performing change of variables again from ϵ and ν back to t and l gives the solution

$$n(l, t) = \frac{M(t - \ln l\,/r)}{l}[\mathrm{H}(l - 1) - \mathrm{H}(l - e^{rt})] \tag{8}$$

The second term in final solution is commonly known as a boxcar function with the general form $\mathrm{H}(x-a)-\mathrm{H}(x-b)$, where a and b are constants representing the limits of the interval over which the function gives a value of 1, and 0 otherwise. The corresponding interval of the boxcar function in Eq. 8 is $[1, e^{rt}]$, with 1 being the minimum size that a lineage can have by definition, and e^{rt} being the maximum lineage size attainable up to time t assuming exponential growth at rate r.

C.3 Derivation of an analytical expression describing the time evolution of lineage size distribution assuming local logistic growth

Following from Section C.2, if we now assume that the local infected population is undergoing logistic growth instead such that $dl/dt = rl(1-l/K)$, where r is again a positive constant and represents the initial growth rate when l is small, and K is the maximum lineage size attainable (also known as the carrying capacity), Eq. 1 then becomes

$$\frac{\partial n}{\partial t}+\frac{\partial n}{\partial l}rl\left(1-\frac{l}{K}\right)+nr\left(1-\frac{2l}{K}\right)=M(t)\delta(l-1) \tag{9}$$

Applying change of variables with $\epsilon = t$ and $\nu = \ln[l/(K-l)]$, we obtain

$$\frac{\partial n}{\partial \epsilon}+\frac{\partial n}{\partial \nu}r-nr\tanh\left(\frac{\nu}{2}\right)=M(\epsilon)\left[\frac{K}{K-1}\right]\delta[\nu+\ln(K-1)] \tag{10}$$

where we have again used the transformation $\delta(g(x)) = \delta(x-x_0)/|g'(x_0)|$, giving the factor $K/(K-1)$ in the right hand side of the equation.

It can be shown that any first-order PDE of the form

$$\frac{\partial w}{\partial x}+\frac{\partial w}{\partial y}a=f(x,y)w+g(x,y) \tag{11}$$

has the general solution

$$w(x,y) = F(x,u)\left[\Phi(u) + \int \frac{g(x,u+ax)}{F(x,u)dx}\right] \tag{12}$$

with

$$F(x,u) = \exp\left[\int f(x,u+ax)dx\right] \tag{13}$$

where $u = y - ax$ and $\Phi(u)$ is an arbitrary function of the parameter u.

If we compare Eq. 10 with Eq. 11, it is straightforward to see that the above known result implies the following general solution to our first-order PDE,

$$n(\epsilon,u) = F(\epsilon,u)\left[\Phi(u) + \int M(\epsilon')[K/(K-1)]\frac{\delta[u + r\epsilon' + \ln(K-1)]}{F(\epsilon',u)d\epsilon'}\right] \tag{14}$$

where

$$F(\epsilon,u) = \exp\left[\int r\tanh[(u+r\epsilon')/2]\,d\epsilon'\right] = A\cosh^2[(u+r\epsilon)/2] \tag{15}$$

with A being a constant of integration.

Applying the boundary condition that $n(l, t=0) = 0$ for all $l \geq 1$, and since $u(\nu, \epsilon = 0) = \nu = \ln[l/(K-l)]$, from Eq. 14 we obtain

$$\Phi(q) = -\frac{M\left[-(1/r)[q + \ln(K-1)]\right]\mathrm{H}[q + \ln(K-1)]}{AK/4} \tag{16}$$

where we have let $q = \ln[l/(K-l)]$.

Substituting this back into the general solution (Eq. 12) and expanding the right hand side gives

$$n(\epsilon, u) = \frac{\cosh^2\left[\frac{u + r\epsilon}{2}\right]\left[\mathrm{H}[u + r\epsilon + \ln(K-1)] - \mathrm{H}(u + \ln(K-1)]\right]}{\frac{K}{4}} \cdot M\left[-\frac{[u + \ln(K-1)]}{r}\right] \quad (17)$$

Finally, performing change of variables again from ϵ and ν back to t and l gives the solution

$$n(l,t) = \frac{K}{l(K-l)} M\left[t - \frac{ln\left[\frac{l(K-1)}{K-l}\right]}{r}\right]\left[\mathrm{H}(l-1) - \mathrm{H}\left[l - \frac{Ke^{rt}}{[(K-1)+e^{rt}]}\right]\right] \quad (18)$$

Note that we again have a boxcar function as the second term in our final solution, with the corresponding interval (over which the boxcar function takes a value of 1) being $[1, Ke^{rt}/[(K-1)+e^{rt}]$. By solving $dl/dt = rl(1-l/K)$ with the boundary condition $l(t=0) = 1$, it is easy to show that the maximum lineage size attainable at time t is given by $Ke^{rt}/[(K-1)+e^{rt}]$, i.e. the right-limit of the boxcar function, as expected.

C.4 A simple deterministic model of viral importation and local lineage growth

In this simple model, we assume that the population at the recipient location is completely susceptible initially at $t = 0$. At each subsequent time step t, $M(t)$ infectious travellers arrive per day from the source location, with each traveller introducing a local transmission lineage of size 1 upon arrival. Once introduced, each lineage grows deterministically assuming exponential growth at rate r. We also assume that there is no recovery of infected individuals, such that the size of a local lineage at any given time t is given by $l(t) = e^{r(t-t_0)}$, where t_0 is the time when the lineage was introduced.

Once a predefined amount of time T (referred to as the time of observation) has elapsed since the first viral importation event at $t = 0$, we simulate a sampling process by randomly selecting a proportion s of infectious individuals at the time of observation (with equal probability regardless of their infection time). Assuming that any uncertainties and biases associated with the phylogenetic tree estimation and phylogeographic reconstruction are negligible, a local transmission lineage is considered detected if at least one of its associated members is sampled. The inferred time of importation of each detected lineage corresponds to the time at which the associated infectious traveller entered the local population. To account for the stochasticity in the sampling process, we repeat the random selection of infected individuals 50 times for each sampling proportion s and time of observation T.

C.5 A stochastic agent-based model of viral importation and local lineage growth

Here we construct a stochastic agent-based model with the following key assumptions:

1. The local population consists of N individuals initially.
2. At any given time t, each individual can be in one of three possible states: susceptible (S), infectious (I), or recovered (R).
3. The population is well-mixed, i.e. each individual has an equal probability of coming into contact with any other individual, regardless of their infection status.

At the start (at $t = 0$), we assume that the local population is initially in a fully susceptible state with no infected individuals. At each subsequent time step t, we simulate the following processes:

1. **Viral importation:** $M(t)$ infectious travellers are introduced into the local population.
2. **Contact:** Each infected individual (including infectious travellers) comes into contact with κ individuals randomly selected from the local population, with equal selection probability regardless of their infection status.
3. **Transmission:** Each susceptible individual who comes into contact with an infected individual becomes infected with probability β.
4. **Recovery:** Each infected individual recovers with probability γ.

Importantly, we keep track of who-infected-whom in each transmission event at each time step; this allows us to attribute each local infection to a specific local transmission lineage resulting from a single arriving infectious traveller.

Once a predefined amount of time T (referred to as the time of observation) has elapsed since the first viral importation event at $t = 0$, we again simulate a sampling process by randomly selecting a proportion s of individuals who are either still infectious (I) or have recovered (R) at the time of observation (with equal probability regardless of their infection time). Assuming that any uncertainties and biases associated with the phylogenetic tree estimation and phylogeographic reconstruction are negligible, a local transmission lineage is considered detected if at least one of its members is sampled. The inferred time of importation of each detected lineage corresponds to the time at which the associated infectious traveller entered the local population. To account for the stochasticity in the sampling process, we repeat the random selection of infected individuals 50 times for each sampling proportion s and time of observation T.

The values for parameters β (transmission probability given contact), γ (recovery probability per unit time), and κ (number of contacts per individual per unit time) are specified based on estimates consistent with or similar to those observed during the

COVID-19 pandemic. Details of these parameter values can be found in Table C.2, along with references to studies from which these estimates were extracted.

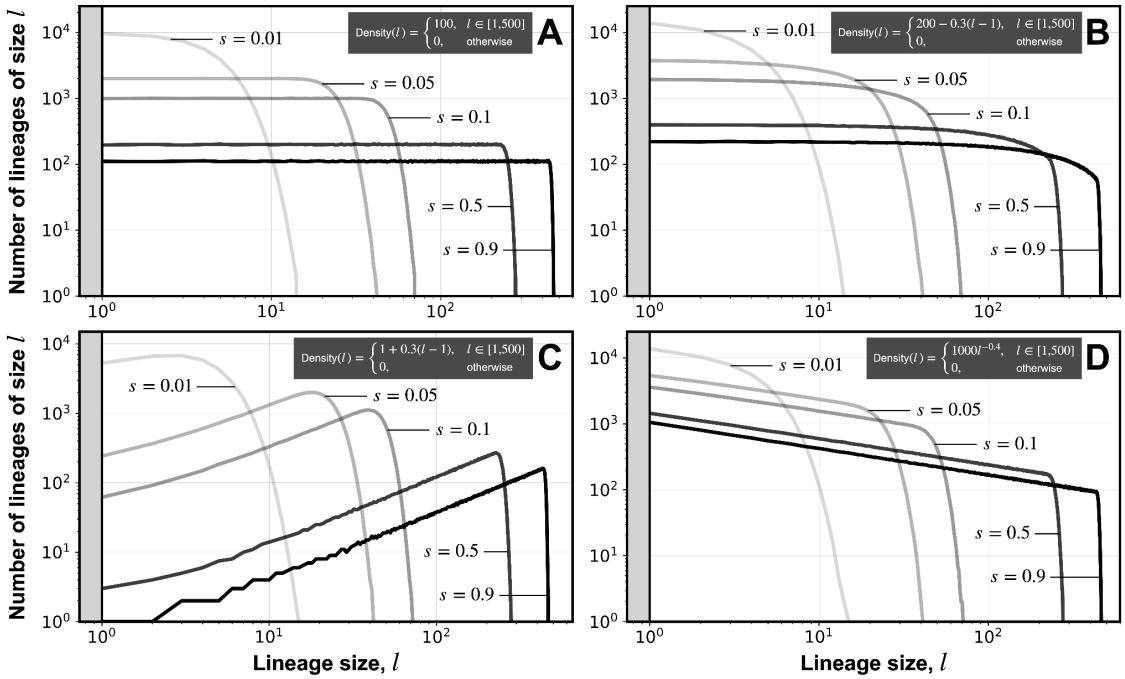


Fig. C.1. Observed lineage size distribution at different sampling proportions on a log-log scale. Each panel shows results from experiments considering a different true lineage size distribution (A: a uniform distribution; B: a linear increase in density with lineage size; C: a linear decrease in density with lineage size; D: a power-law distribution with a negative exponent); the corresponding density function is shown at the top of each panel. All distributions are truncated at maximum lineage size $l = 500$. Solid lines represent the observed lineage size distribution at varying sampling proportions ($s = 0.01, 0.05, 0.1, 0.5, 0.9$), as indicated by their opacity (lower opacity corresponds to a lower sampling proportion).

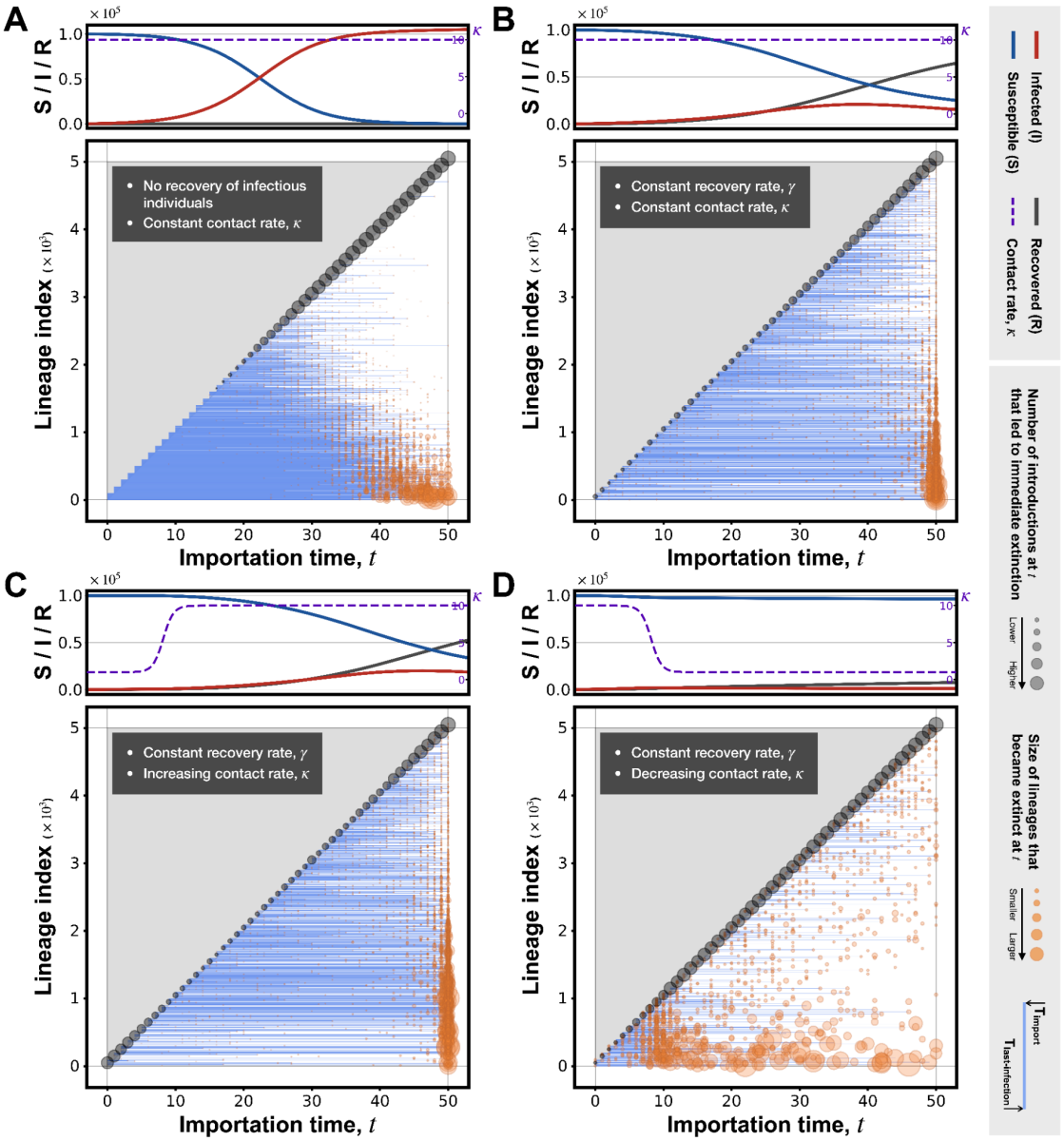


Fig. C.2. Impact of local transmission dynamics on lineage growth and stochastic extinction of lineages. Each panel shows results from a single stochastic agent-based simulation up to time T=50 assuming a constant importation rate ($M(t) = M_0 = 100$ per day) and a constant transmission probability ($\beta = 0.25$ per contact), but either with recovery (B, C, and D; at rate $\gamma = 1$) or without recovery (A) of infectious individuals, and either a constant contact rate (A and B; at rate $\kappa = 10$) or a time-varying contact rate (A: an increasing contact rate, from $\kappa = 1$ to $\kappa = 10$; B: a decreasing contact rate, from $\kappa = 10$ to $\kappa = 1$). The plot at the top of each panel shows the simulated epidemic dynamics, i.e. the number of infected (I; red solid line), susceptible (S; blue solid line) and recovered (R; black solid line) individuals over time. The plot at the bottom shows the time of importation and time of extinction (time of last infection) of each local transmission lineage, as represented by the start- and end-position of a horizontal blue line. Transmission lineages (horizontal blue lines) are positioned along the y-axis according to their order of importation. Extinction events are marked by orange circles

positioned along the x-axis according to the time of last infection, with radius indicating the size of the lineage at extinction (i.e. total number of infected individuals since introduction). Viral introductions resulting in immediate extinction and therefore local transmission lineages of size $l = 1$ are marked by black circles, with radius indicating the number of such events at a given time t.

Table C.1. Key statistics from previous studies estimating number of viral introductions and importation intensity using phylogeography during COVID-19 pandemic. For studies in which phylogeographic analysis was performed for multiple locations, the average sampling proportion (weighted by the estimated number for reported/confirmed cases during study period) is estimated and reported; sampling proportions estimated using numbers of reported/confirmed cases taken from external data from public repositories (as opposed to from the relevant studies themselves) are indicated by an asterisk.

ID	Title (DOI)	No. of genomes included	Estimated sampling proportion	Relevant details and/or external data used
1	A small number of early introductions seeded widespread transmission of SARS-CoV-2 in Québec, Canada (https://doi.org/10.1186/s13073-021-00986-9) (1)	2,921	0.057	N/A
2	Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events (https://doi.org/10.1126/science.abe3261) (2)	772	0.012 *	Case data downloaded from https://github.com/nytimes/covid-19-data (accessed on 23 December 2024) was used to calculate cumulative number of confirmed cases in Massachusetts state between 1 March and 1 May 2020
3	Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland (https://doi.org/10.1038/s41564-020-00838-z) (3)	1,314	0.49	N/A
4	Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK (https://doi.org/10.1126/science.abf2946) (4)	26,181	0.093	N/A
5	Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence (https://doi.org/10.1126/science.abj0113) (5)	17,716	0.038	N/A

6	Context-specific emergence and growth of the SARS-CoV-2 Delta variant (https://doi.org/10.1038/s41586-022-05200-3) (6)	52,992	0.50 *	Total number of Delta cases in England between 12 March and 15 June 2021 was estimated using confirmed case data downloaded from OWID (https://ourworldindata.org/grapher/uk-daily-new-covid-cases?time=earliest..latest&country=~England; accessed on 12 October 2024) and lineage frequency data downloaded from https://covid19.sanger.ac.uk/lineages/raw (accessed on 13 October 2024).
7	Genomic epidemiology of the first two waves of SARS-CoV-2 in Canada (https://doi.org/10.7554/eLife.73896) (7)	27,552	0.045	N/A
8	Genomic epidemiology of SARS-CoV-2 under an elimination strategy in Hong Kong (https://doi.org/10.1038/s41467-022-28420-7) (8)	1,899	0.19	Total number of confirmed cases in Hong Kong during the study period was taken from Supplementary Information provided by the authors.
9	Genomic epidemiology of the SARS-CoV-2 epidemic in Brazil (https://doi.org/10.1038/s41564-022-01191-z) (9)	17,135	0.00091 *	Total number of confirmed cases in Brazil and Paraguay up to 30 June 2021 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 6 October 2024).
10	Phylogenetic estimates of SARS-CoV-2 introductions into Washington State (https://doi.org/10.1016/j.lana.2021.100018) (10)	4,918	0.060	N/A
11	The first wave of the COVID-19 epidemic in Spain was associated with early introductions and fast spread of a dominating genetic variant (https://doi.org/10.1038/s41588-021-00936-6) (11)	2,170	0.0085 *	Total number of confirmed cases in Spain between 25 Feb and 22 June 2020 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 23 October 2024).
12	Genomic epidemiology	649	0.56	N/A

	reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand (https://doi.org/10.1038/s41467-020-20235-8) (2)			
13	Genomic surveillance of SARS-CoV-2 in Puerto Rico enabled early detection and tracking of variants (https://doi.org/10.1038/s43856-022-00168-7) (13)	753	0.0040 *	Total number of confirmed cases in Puerto Rico between 23 March 2020 and 30 September 2021 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 3 December 2024).
14	SARS-CoV-2 introductions and early dynamics of the epidemic in Portugal (https://doi.org/10.1038/s43856-022-00072-0) (14)	1,275	0.16	N/A
15	Genomic epidemiology of SARS-CoV-2 variants during the first two years of the pandemic in Colombia (https://doi.org/10.1038/s43856-023-00328-3) (15)	1,670	0.00028 *	Total number of confirmed cases in Colombia up to February 2022 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed 5 December 2024).
16	Genomic epidemiology of SARS-CoV-2 transmission lineages in Ecuador (https://doi.org/10.1093/ve/veab051) (16)	160	0.0008	Total number of confirmed cases in Ecuador during the study period was taken from Supplementary Information provided by the authors.
17	Genomic epidemiology reveals the reduction of the introduction and spread of SARS-CoV-2 after implementing control strategies in Republic of Korea, 2020 (https://doi.org/10.1093/ve/veab077) (17)	2,065	0.034	N/A
18	Epidemiological dynamics of SARS-CoV-2 VOC Gamma in Rio de Janeiro, Brazil (https://doi.org/10.1093/v	113	0.00043	N/A

	e/veab087) (18)			
19	Sixteen novel lineages of SARS-CoV-2 in South Africa (https://doi.org/10.1038/s41591-021-01255-3) (19)	1,365	0.0022 *	Total number of confirmed cases in South Africa between 6 March and 26 August 2022 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 11 November 2024). Note also that the final number of South African sequences after quality control was used in calculating the sampling proportion.
20	A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa (https://doi.org/10.1126/science.abj4336) (20)	8,746	0.002 *	Total number of confirmed cases in Africa up to 31 March 2021 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 15 November 2024).
21	Evolution and epidemic spread of SARS-CoV-2 in Brazil (https://doi.org/10.1126/science.abd2161) (21)	490	0.005	N/A
22	Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China (https://doi.org/10.1016/j.cell.2020.04.023) (22)	70	0.05	Total number of confirmed cases in Guangdong up to 19 March 2020 (n=1,388; including imported cases), as reported by the authors in the study, was used in calculating the sampling proportion.
23	Tracking the COVID-19 pandemic in Australia using genomics (https://doi.org/10.1038/s41467-020-18314-x) (23)	903	0.68	The number of confirmed cases in Australia between 6 January and 14 April 2020 (n=1,333), as reported by the authors in the study, was used in calculating the sampling proportion.
24	Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak in Russia (https://doi.org/10.1038/s41467-020-20880-z) (24)	211	0.0034	The number of confirmed cases in Russia between 11 March and 23 April 2020 (n=62,745), as reported by the authors in the study, was used in calculating the sampling proportion.
25	Genomic epidemiology of SARS-CoV-2 during	1,142	0.0051	Total number of confirmed cases in Mozambique up to 22 April

	the first four waves in Mozambique (https://doi.org/10.1371/journal.pgph.0001593) (25)			2022 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 20 December 2024).
26	Tracking SARS-CoV-2 introductions in Mozambique using pandemic-scale phylogenies: a retrospective observational study (https://doi.org/10.1016/S2214-109X(23)00169-9) (26)	932	0.0067 *	Although Mozambican sequences were extracted from the GISAID database up to 11 January 2022, the study considered only the Beta and Delta variants which were the dominant lineages up to end of November 2021 (before they were overtaken by the Omicron variant) (Ismael et al., 2023). To calculate the sampling proportion, we have therefore used the total number of confirmed cases in Mozambique between 1 November 2020 and 30 November 2021 downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 5 January 2025), with the assumption that the proportion of confirmed cases that can be attributed to variants other than the Beta and Delta variant is negligible.
27	Tracing the international arrivals of SARS-CoV-2 Omicron variants after Aotearoa New Zealand reopened its border (https://doi.org/10.1038/s41467-022-34186-9) (27)	2,000	0.0016 *	SARS-CoV-2 transmission during the study period was dominated by the Omicron and Delta variants. To estimate the number of confirmed cases that can be attributed to Omicron, we used the total number of confirmed cases in New Zealand between 8 November 2021 and 15 June 2022 downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 15 January 2025). The number of cases attributable to the Delta variant was estimated using lineage frequency data downloaded from https://github.com/ESR-NZ/nz-sars-cov2-variants/tree/main/data (accessed

				on 15 January 2025); this value was then subtracted from the total number of confirmed cases during the study period to obtain the number of relevant Omicron cases.
28	Evolutionary and spatiotemporal analyses reveal multiple introductions and cryptic transmission of SARS-CoV-2 VOC/VOI in Malta (https://doi.org/10.1128/spectrum.01539-23) (28)	666	0.010 *	Total number of confirmed cases in Malta up to 25 January 2022 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 12 January 2025).
29	Genomic epidemiology of early SARS-CoV-2 transmission dynamics, Gujarat, India (https://doi.org/10.3201/eid2804.212053) (29)	434	0.0071 *	Total number of confirmed cases in Gujarat, India between 1 April and 31 July 2020 was calculated using data downloaded from https://github.com/covid19india/api (accessed on 16 January 2025)..
30	Tracking the introduction and spread of SARS-CoV-2 in coastal Kenya (https://doi.org/10.1038/s41467-021-25137-x) (30)	311	0.16	The number of confirmed cases across the coastal counties by 31 July 2020 (n=1,997), as reported by the authors in the study, was used in calculating the sampling proportion.
31	A single early introduction governed viral diversity in the second wave of SARS-CoV-2 epidemic in Hungary (https://doi.org/10.1093/ve/veac069) (31)	352	0.00087	Total number of confirmed cases in Hungary between 29 April 2020 and 23 February 2021 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 20 December 2024).
32	Importation of Alpha and Delta variants during the SARS-CoV-2 epidemic in Switzerland: phylogenetic analysis and intervention scenarios (https://doi.org/10.1371/journal.ppat.1011553) (32)	13,198	0.15 *	Total number of confirmed cases that can be attributed to the Alpha and Delta variants in Switzerland prior to 31 March 2021 and 31 July 2021, respectively, were estimated using surveillance data from Federal Office of Public Health (using an R script provided by the authors in the Supplementary Materials,

				https://github.com/ISPMBern/voc_imports_ch/blob/main/R/imports_100_swissepidemic.R).
33	Introduction and transmission of SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark (https://doi.org/10.1186/s13073-022-01045-7) (33)	60,178	0.34	N/A
34	Genomic epidemiology of the first epidemic wave of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Palestine (https://doi.org/10.1099/mgen.0.000584) (34)	69	0.0030 *	Total number of confirmed cases in Palestine between 4 March and19 August 2020 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 10 October 2024).
35	Genomic reconstruction of the SARS-CoV-2 epidemic in England (https://doi.org/10.1038/s41586-021-04069-y) (35)	281,178	0.072	N/A
36	Genomic epidemiology of the first wave of SARS-CoV-2 in Italy (https://doi.org/10.3390/v12121438) (36)	651	0.0027 *	Total number of confirmed cases in Italy between 29 January and 20 July 2020 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 2 August 2024).
37	Multiple introductions followed by ongoing community spread of SARS-CoV-2 at one of the largest metropolitan areas of Northeast Brazil (https://doi.org/10.3390/v12121414) (37)	101	0.0065 *	Total number of confirmed cases in Pernambuco, Brazil between 12 March and 14 May 2020 (date of latest genome sample) was calculated using data downloaded from https://github.com/henriquemor/covid19-Brazil-timeseries/blob/master/transp-confirmed-new.csv (accessed on 25 October 2024).
38	Phylodynamics reveals the role of human travel and contact tracing in controlling the first wave of COVID-19 in four	793	0.079	Sampling proportion is averaged across the four countries considered in the study, weighted by the number of confirmed cases per location.

	island nations (https://doi.org/10.1093/ve/veab052) (38)			
39	Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1 (https://doi.org/10.1126/science.adg6605) (39)	48,748	0.011 *	Total number of Omicron BA.1 cases in England up to 31 January 2021 was estimated using data on the number of confirmed cases with SGTF, as provided by the authors in the Supplementary Materials (https://github.com/joetsui1994/Omicron-BA.1-invasion-dynamics/blob/main/analyses/epidemiological/GOV.UK_SGTF_BA.1_daily_LTLA_20210901-20220301.csv; accessed on 29 August 2024).
40	Dispersal dynamics of SARS-CoV-2 lineages during the first epidemic wave in New York City (https://doi.org/10.1371/journal.ppat.1009571) (40)	828	0.0045 *	Total number of confirmed cases in New York City (Bronx, Brooklyn, Manhattan, Queens, and Staten Island) up to 10 May 2020 was calculated using data downloaded from https://github.com/sdellicour/sars-cov-2_new_york/blob/master/Scripts_%26_data/NY_epidemiological_data/NY_boroughs_COVID.csv (accessed on 19 December 2024), as provided by the authors in the Supplementary Materials.
41	A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages (https://doi.org/10.1093/molbev/msaa284) (41)	740	0.012 *	Total number of confirmed cases in Belgium up to 10 June 2020 was calculated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 9 December 2024).
42	Regional connectivity drove bidirectional transmission of SARS-CoV-2 in the Middle East during travel restrictions (https://doi.org/10.1038/s41467-022-32536-1) (42)	579	0.0011	N/A
43	Genomic sequencing of SARS-CoV-2 in Rwanda	203	0.012	N/A

	reveals the importance of incoming travelers on lineage diversity (https://doi.org/10.1038/s41467-021-25985-7) (43)			
44	SARS-CoV-2 genomic characterization and clinical manifestation of the COVID-19 outbreak in Uruguay (https://doi.org/10.1080/22221751.2020.1863747) (44)	73	0.093	N/A
45	Comparing the evolutionary dynamics of predominant SARS-CoV-2 virus lineages co-circulating in Mexico (https://doi.org/10.7554/eLife.82069) (45)	10,618	0.0042 *	Total number of confirmed cases that can be attributed to B.1.1.222, B.1.1.519, B.1.1.7, P.1, and B.1.617.2 in Mexico from January 2020 up to 30 November 2021 was estimated using confirmed case data downloaded from https://datos.covid-19.conacyt.mx/ (accessed on 5 January 2025) and lineage frequency data downloaded from https://cov-spectrum.org/explore/Mexico/AllSamples/ (accessed on 5 January 2025).
46	Variant-specific introduction and dispersal dynamics of SARS-CoV-2 in New York City - from Alpha to Omicron (https://doi.org/10.1371/journal.ppat.1011348) (46)	12,093	0.0092 *	Total number of Alpha, Iota, Delta, and Omicron (BA.1) cases in New York City between 2020 and 2022 was estimated using daily confirmed case number and lineage frequency data downloaded from https://github.com/nychealth/coronavirus-data (accessed on 20 January 2024). The final sampling proportion is averaged across the four variants considered in the study, weighted by the number of confirmed cases per variant.
47	Regional importation and asymmetric within-country spread of SARS-CoV-2 variants of concern in the Netherlands	3,596	0.0038	Total number of Alpha, Beta, Gamma, and Delta cases in the Netherlands during the study period was estimated by calculating the total number of non-variant-specific cases

	(https://doi.org/10.7554/eLife.78770) (47)			between December 2020 and August 2021, and subtracting the number of cases that were not attributable to the Alpha variant between December 2020 and April 2021, using data downloaded from https://github.com/AMC-LAEB/nl_sars-cov-2_genomic_epi_2022/tree/main (accessed on 20 January 2025) as provided by the authors in the Supplementary Materials. In the calculation, we have assumed that SARS-CoV-2 transmission during the study period was dominated by the Alpha, Beta, Gamma, and Delta variants, with the proportion of other lineages being negligible.
48	Genomic evolution and early introductions of the SARS-CoV-2 Omicron variant in Mexico (https://doi.org/10.1093/ve/veac109) (48)	641	0.00034 *	Total number of Omicron cases in Mexico between 1 November 2021 and 31 March 2022 was estimated using data downloaded from OWID (https://ourworldindata.org/coronavirus#coronavirus-country-profiles; accessed on 12 January 2025), with the assumption that the number of cases attributable to other lineages was negligible during the study period.

Table C.2. Key parameters used in the stochastic agent-based model. Brief descriptions and references to studies from which values are taken are included where applicable.

Parameter	Description	Value(s)
Population size, N	Size of local population (excluding arriving infectious travellers)	100,000
Transmission probability, β	Probability that a susceptible individual becomes infected upon contact with an infected individual	0.025 per contact (fixed) (49, 50)
Recovery probability, γ	Probability that an infected individual recovers and becomes immune (i.e., I → R) per time step	0.1 per day (fixed; equivalent to an average infectious duration of 10 days) (51, 52)
Contact rate, κ	Average of contacts per individual per time step (regardless of infection status)	Contant κ: 10 per individual per day (53)
		Time-varying κ: increases from 1 to 10 per individual per day, following a sigmoidal trajectory (53, 54)
		Time-varying κ: decreases from 10 to 1 per individual per day, following a sigmoidal trajectory (53, 54)

References

1. Murall, C.L., Fournier, E., Galvez, J.H., N'Guessan, A., Reiling, S.J., Quirion, P.-O., Naderi, S., Roy, A.-M., Chen, S.-H., Stretenowich, P., Bourgey, M., Bujold, D., Gregoire, R., Lepage, P., St-Cyr, J., Willet, P., Dion, R., Charest, H., Lathrop, M., Roger, M., Bourque, G., Ragoussis, J., Shapiro, B.J. and Moreira, S. (2021) 'A small number of early introductions seeded widespread transmission of SARS-CoV-2 in Québec, Canada', *Genome medicine*, 13(1), p. 169.
2. Lemieux, J.E., Siddle, K.J., Shaw, B.M., Loreth, C., Schaffner, S.F., Gladden-Young, A., Adams, G., Fink, T., Tomkins-Tinch, C.H., Krasilnikova, L.A., DeRuff, K.C., Rudy, M., Bauer, M.R., Lagerborg, K.A., Normandin, E., Chapman, S.B., Reilly, S.K., Anahtar, M.N., Lin, A.E., Carter, A., Myhrvold, C., Kemball, M.E., Chaluvadi, S., Cusick, C., Flowers, K., Neumann, A., Cerrato, F., Farhat, M., Slater, D., Harris, J.B., Branda, J.A., Hooper, D., Gaeta, J.M., Baggett, T.P., O'Connell, J., Gnirke, A., Lieberman, T.D., Philippakis, A., Burns, M., Brown, C.M., Luban, J., Ryan, E.T., Turbett, S.E., LaRocque, R.C., Hanage, W.P., Gallagher, G.R., Madoff, L.C., Smole, S., Pierce, V.M., Rosenberg, E., Sabeti, P.C., Park, D.J. and MacInnis, B.L. (2021) 'Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events', *Science*, 371(6529), p. eabe3261.
3. da Silva Filipe, A., Shepherd, J.G., Williams, T., Hughes, J., Aranday-Cortes, E., Asamaphan, P., Ashraf, S., Balcazar, C., Brunker, K., Campbell, A., Carmichael, S., Davis, C., Dewar, R., Gallagher, M.D., Gunson, R., Hill, V., Ho, A., Jackson, B., James, E., Jesudason, N., Johnson, N., McWilliam Leitch, E.C., Li, K., MacLean, A., Mair, D., McAllister, D.A., McCrone, J.T., McDonald, S.E., McHugh, M.P., Morris, A.K., Nichols, J., Niebel, M., Nomikou, K., Orton, R.J., O'Toole, Á., Palmarini, M., Parcell, B.J., Parr, Y.A., Rambaut, A., Rooke, S., Shaaban, S., Shah, R., Singer, J.B., Smollett, K., Starinskij, I., Tong, L., Sreenu, V.B., Wastnedge, E., COVID-19 Genomics UK (COG-UK) Consortium, Holden, M.T.G., Robertson, D.L., Templeton, K. and Thomson, E.C. (2021) 'Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland', *Nature microbiology*, 6(1), pp. 112–122.
4. du Plessis, L., McCrone, J.T., Zarebski, A.E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T.R., Faria, N.R., Jackson, B., Loman, N.J., O'Toole, Á., Nicholls, S.M., Parag, K.V., Scher, E., Vasylyeva, T.I., Volz, E.M., Watts, A., Bogoch, I.I., Khan, K., COVID-19 Genomics UK (COG-UK) Consortium, Aanensen, D.M., Kraemer, M.U.G., Rambaut, A. and Pybus, O.G. (2021) 'Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK', *Science*, 371(6530), pp. 708–712.
5. Kraemer, M.U.G., Hill, V., Ruis, C., Dellicour, S., Bajaj, S., McCrone, J.T., Baele, G., Parag, K.V., Battle, A.L., Gutierrez, B., Jackson, B., Colquhoun, R., O'Toole, Á., Klein, B., Vespignani, A., COVID-19 Genomics UK (COG-UK) Consortium, Volz, E., Faria, N.R., Aanensen, D.M., Loman, N.J., du Plessis, L., Cauchemez, S., Rambaut, A., Scarpino, S.V. and Pybus, O.G. (2021) 'Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence', *Science*, 373(6557), pp. 889–895.
6. McCrone, J.T., Hill, V., Bajaj, S., Pena, R.E., Lambert, B.C., Inward, R., Bhatt, S., Volz, E., Ruis, C., Dellicour, S., Baele, G., Zarebski, A.E., Sadilek, A., Wu, N., Schneider, A., Ji, X., Raghwani, J., Jackson, B., Colquhoun, R., O'Toole, Á., Peacock, T.P., Twohig, K., Thelwall, S., Dabrera, G., Myers, R., Faria, N.R., Huber, C., Bogoch, I.I., Khan, K., du Plessis, L., Barrett, J.C., Aanensen, D.M., Barclay,

W.S., Chand, M., Connor, T., Loman, N.J., Suchard, M.A., Pybus, O.G., Rambaut, A. and Kraemer, M.U.G. (2022) 'Context-specific emergence and growth of the SARS-CoV-2 Delta variant', *Nature*, 610(7930), pp. 154–160.

7. McLaughlin, A., Montoya, V., Miller, R.L., Mordecai, G.J., Canadian COVID-19 Genomics Network (CanCOGen) Consortium, Worobey, M., Poon, A.F.Y. and Joy, J.B. (2022) 'Genomic epidemiology of the first two waves of SARS-CoV-2 in Canada', *eLife*, 11, p. e73896.
8. Gu, H., Xie, R., Adam, D.C., Tsui, J.L.-H., Chu, D.K., Chang, L.D.J., Cheuk, S.S.Y., Gurung, S., Krishnan, P., Ng, D.Y.M., Liu, G.Y.Z., Wan, C.K.C., Cheng, S.S.M., Edwards, K.M., Leung, K.S.M., Wu, J.T., Tsang, D.N.C., Leung, G.M., Cowling, B.J., Peiris, M., Lam, T.T.Y., Dhanasekaran, V. and Poon, L.L.M. (2022) 'Genomic epidemiology of SARS-CoV-2 under an elimination strategy in Hong Kong', *Nature communications*, 13(1), pp. 1–10.
9. Giovanetti, M., Slavov, S.N., Fonseca, V., Wilkinson, E., Tegally, H., Patané, J.S.L., Viala, V.L., San, E.J., Rodrigues, E.S., Santos, E.V., Aburjaile, F., Xavier, J., Fritsch, H., Adelino, T.E.R., Pereira, F., Leal, A., Iani, F.C. de M., de Carvalho Pereira, G., Vazquez, C., Sanabria, G.M.E., Oliveira, E.C. de, Demarchi, L., Croda, J., dos Santos Bezerra, R., Paola Oliveira de Lima, L., Martins, A.J., Renata dos Santos Barros, C., Marqueze, E.C., de Souza Todao Bernardino, J., Moretti, D.B., Brassaloti, R.A., de Lello Rocha Campos Cassano, R., Mariani, P.D.S.C., Kitajima, J.P., Santos, B., Proto-Siqueira, R., Cantarelli, V.V., Tosta, S., Nardy, V.B., Reboredo de Oliveira da Silva, L., Gómez, M.K.A., Lima, J.G., Ribeiro, A.A., Guimarães, N.R., Watanabe, L.T., Barbosa Da Silva, L., da Silva Ferreira, R., da Penha, M.P.F., Ortega, M.J., de la Fuente, A.G., Villalba, S., Torales, J., Gamarra, M.L., Aquino, C., Figueredo, G.P.M., Fava, W.S., Motta-Castro, A.R.C., Venturini, J., do Vale Leone de Oliveira, S.M., Gonçalves, C.C.M., do Carmo Debur Rossa, M., Becker, G.N., Giacomini, M.P., Marques, N.Q., Riediger, I.N., Raboni, S., Mattoso, G., Cataneo, A.D., Zanluca, C., Duarte dos Santos, C.N., Assato, P.A., Allan da Silva da Costa, F., Poleti, M.D., Lesbon, J.C.C., Mattos, E.C., Banho, C.A., Sacchetto, L., Moraes, M.M., Grotto, R.M.T., Souza-Neto, J.A., Nogueira, M.L., Fukumasu, H., Coutinho, L.L., Calado, R.T., Neto, R.M., Bispo de Filippis, A.M., Venancio da Cunha, R., Freitas, C., Peterka, C.R.L., de Fátima Rangel Fernandes, C., Navegantes, W., do Carmo Said, R.F., Campelo de A e Melo, C.F., Almiron, M., Lourenço, J., de Oliveira, T., Holmes, E.C., Haddad, R., Sampaio, S.C., Elias, M.C., Kashima, S., Junior de Alcantara, L.C. and Covas, D.T. (2022) 'Genomic epidemiology of the SARS-CoV-2 epidemic in Brazil', *Nature Microbiology*, 7(9), pp. 1490–1500.
10. Tordoff, D.M., Greninger, A.L., Roychoudhury, P., Shrestha, L., Xie, H., Jerome, K.R., Breit, N., Huang, M.-L., Famulare, M. and Herbeck, J.T. (2021) 'Phylogenetic estimates of SARS-CoV-2 introductions into Washington State', *Lancet regional health. Americas*, 1, p. 100018.
11. López, M.G., Chiner-Oms, Á., García de Viedma, D., Ruiz-Rodriguez, P., Bracho, M.A., Cancino-Muñoz, I., D'Auria, G., de Marco, G., García-González, N., Goig, G.A., Gómez-Navarro, I., Jiménez-Serrano, S., Martinez-Priego, L., Ruiz-Hueso, P., Ruiz-Roldán, L., Torres-Puente, M., Alberola, J., Albert, E., Aranzamendi Zaldumbide, M., Bea-Escudero, M.P., Boga, J.A., Bordoy, A.E., Canut-Blasco, A., Carvajal, A., Cilla Eguiluz, G., Cordón Rodríguez, M.L., Costa-Alcalde, J.J., de Toro, M., de Toro Peinado, I., Del Pozo, J.L., Duchêne, S., Fernández-Pinero, J., Fuster Escrivá, B., Gimeno Cardona, C., González Galán, V., Gonzalo Jiménez, N., Hernáez Crespo, S., Herranz, M., Lepe, J.A., López-Causapé, C., López-Hontangas, J.L., Martín, V., Martró, E., Milagro Beamonte, A., Montes Ros, M., Moreno-Muñoz, R.,

Navarro, D., Navarro-Marí, J.M., Not, A., Oliver, A., Palop-Borrás, B., Parra Grande, M., Pedrosa-Corral, I., Pérez González, M.C., Pérez-Lago, L., Pérez-Ruiz, M., Piñeiro Vázquez, L., Rabella, N., Rezusta, A., Robles Fonseca, L., Rodríguez-Villodres, Á., Sanbonmatsu-Gámez, S., Sicilia, J., Soriano, A., Tirado Balaguer, M.D., Torres, I., Tristancho, A., Marimón, J.M., SeqCOVID-Spain consortium, Coscolla, M., González-Candelas, F. and Comas, I. (2021) 'The first wave of the COVID-19 epidemic in Spain was associated with early introductions and fast spread of a dominating genetic variant', *Nature genetics*, 53(10), pp. 1405–1414.

12. Geoghegan, J.L., Ren, X., Storey, M., Hadfield, J., Jelley, L., Jefferies, S., Sherwood, J., Paine, S., Huang, S., Douglas, J., Mendes, F.K., Sporle, A., Baker, M.G., Murdoch, D.R., French, N., Simpson, C.R., Welch, D., Drummond, A.J., Holmes, E.C., Duchêne, S. and de Ligt, J. (2020) 'Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand', *Nature communications*, 11(1), p. 6351.
13. Santiago, G.A., Flores, B., González, G.L., Charriez, K.N., Huertas, L.C., Volkman, H.R., Van Belleghem, S.M., Rivera-Amill, V., Adams, L.E., Marzán, M., Hernández, L., Cardona, I., O'Neill, E., Paz-Bailey, G., Papa, R. and Muñoz-Jordan, J.L. (2022) 'Genomic surveillance of SARS-CoV-2 in Puerto Rico enabled early detection and tracking of variants', *Communications medicine*, 2, p. 100.
14. Borges, V., Isidro, J., Trovão, N.S., Duarte, S., Cortes-Martins, H., Martiniano, H., Gordo, I., Leite, R., Vieira, L., Portuguese network for SARS-CoV-2 genomics (Consortium), Guiomar, R. and Gomes, J.P. (2022) 'SARS-CoV-2 introductions and early dynamics of the epidemic in Portugal', *Communications medicine*, 2, p. 10.
15. Jimenez-Silva, C., Rivero, R., Douglas, J., Bouckaert, R., Villabona-Arenas, C.J., Atkins, K.E., Gastelbondo, B., Calderon, A., Guzman, C., Echeverri-De la Hoz, D., Muñoz, M., Ballesteros, N., Castañeda, S., Patiño, L.H., Ramirez, A., Luna, N., Paniz-Mondolfi, A., Serrano-Coll, H., Ramirez, J.D., Mattar, S. and Drummond, A.J. (2023) 'Genomic epidemiology of SARS-CoV-2 variants during the first two years of the pandemic in Colombia', *Communications medicine*, 3(1), p. 97.
16. Gutierrez, B., Márquez, S., Prado-Vivar, B., Becerra-Wong, M., Guadalupe, J.J., Candido, D.D.S., Fernandez-Cadena, J.C., Morey-Leon, G., Armas-Gonzalez, R., Andrade-Molina, D.M., Bruno, A., De Mora, D., Olmedo, M., Portugal, D., Gonzalez, M., Orlando, A., Drexler, J.F., Moreira-Soto, A., Sander, A.-L., Brünink, S., Kühne, A., Patiño, L., Carrazco-Montalvo, A., Mestanza, O., Zurita, J., Sevillano, G., Du Plessis, L., McCrone, J.T., Coloma, J., Trueba, G., Barragán, V., Rojas-Silva, P., Grunauer, M., Kraemer, M.U.G., Faria, N.R., Escalera-Zamudio, M., Pybus, O.G. and Cárdenas, P. (2021) 'Genomic epidemiology of SARS-CoV-2 transmission lineages in Ecuador', *Virus evolution*, 7(2), p. veab051.
17. Kwon, J.-H., Kim, J.-M., Lee, D.-H., Park, A.K., Kim, I.-H., Kim, D.-W., Kim, J.-Y., Lim, N., Cho, K.-Y., Kim, H.M., Lee, N.-J., Woo, S., Lee, C.Y., No, J.S., Kim, J., Rhee, J., Han, M.-G., Rhie, G.-E., Yoo, C.K. and Kim, E.-J. (2021) 'Genomic epidemiology reveals the reduction of the introduction and spread of SARS-CoV-2 after implementing control strategies in Republic of Korea, 2020', *Virus evolution*, 7(2), p. veab077.
18. Moreira, F.R.R., D'arc, M., Mariani, D., Herlinger, A.L., Schiffler, F.B., Rossi, Á.D., Leitão, I.C., Miranda, T.D.S., Cosentino, M.A.C., Tôrres, M.C.P., Rmdsc, da C., Gonçalves, C.C.A., Faffe, D.S., Galliez, R.M., Odcf, J., Aguiar, R.S., Dos Santos, A.F.A., Voloch, C.M., Tmpp, C. and Tanuri, A. (2021) 'Epidemiological dynamics of SARS-CoV-2 VOC Gamma in Rio de Janeiro, Brazil', *Virus evolution*, 7(2), veab087.

19. Tegally, H., Wilkinson, E., Lessells, R.J., Giandhari, J., Pillay, S., Msomi, N., Mlisana, K., Bhiman, J.N., von Gottberg, A., Walaza, S., Fonseca, V., Allam, M., Ismail, A., Glass, A.J., Engelbrecht, S., Van Zyl, G., Preiser, W., Williamson, C., Petruccione, F., Sigal, A., Gazy, I., Hardie, D., Hsiao, N.-Y., Martin, D., York, D., Goedhals, D., San, E.J., Giovanetti, M., Lourenço, J., Alcantara, L.C.J. and de Oliveira, T. (2021) 'Sixteen novel lineages of SARS-CoV-2 in South Africa', *Nature medicine*, 27(3), pp. 440–446.
20. Wilkinson, E., Giovanetti, M., Tegally, H., San, J.E., Lessells, R., Cuadros, D., Martin, D.P., Rasmussen, D.A., Zekri, A.-R.N., Sangare, A.K., Ouedraogo, A.-S., Sesay, A.K., Priscilla, A., Kemi, A.-S., Olubusuyi, A.M., Oluwapelumi, A.O.O., Hammami, A., Amuri, A.A., Sayed, A., Ouma, A.E.O., Elargoubi, A., Ajayi, N.A., Victoria, A.F., Kazeem, A., George, A., Trotter, A.J., Yahaya, A.A., Keita, A.K., Diallo, A., Kone, A., Souissi, A., Chtourou, A., Gutierrez, A.V., Page, A.J., Vinze, A., Iranzadeh, A., Lambisia, A., Ismail, A., Rosemary, A., Sylverken, A., Femi, A., Ibrahimi, A., Marycelin, B., Oderinde, B.S., Bolajoko, B., Dhaala, B., Herring, B.L., Njanpop-Lafourcade, B.-M., Kleinhans, B., McInnis, B., Tegomoh, B., Brook, C., Pratt, C.B., Scheepers, C., Akoua-Koffi, C.G., Agoti, C.N., Peyrefitte, C., Daubenberger, C., Morang'a, C.M., Nokes, D.J., Amoako, D.G., Bugembe, D.L., Park, D., Baker, D., Doolabh, D., Ssemwanga, D., Tshiabuila, D., Bassirou, D., Amuzu, D.S.Y., Goedhals, D., Omuoyo, D.O., Maruapula, D., Foster-Nyarko, E., Lusamaki, E.K., Simulundu, E., Ong'era, E.M., Ngabana, E.N., Shumba, E., El Fahime, E., Lokilo, E., Mukantwari, E., Philomena, E., Belarbi, E., Simon-Loriere, E., Anoh, E.A., Leendertz, F., Ajili, F., Enoch, F.O., Wasfi, F., Abdelmoula, F., Mosha, F.S., Takawira, F.T., Derrar, F., Bouzid, F., Onikepe, F., Adeola, F., Muyembe, F.M., Tanser, F., Dratibi, F.A., Mbunsu, G.K., Thilliez, G., Kay, G.L., Githinji, G., van Zyl, G., Awandare, G.A., Schubert, G., Maphalala, G.P., Ranaivoson, H.C., Lemriss, H., Anise, H., Abe, H., Karray, H.H., Nansumba, H., Elgahzaly, H.A., Gumbo, H., Smeti, I., Ayed, I.B., Odia, I., Ben Boubaker, I.B., Gaaloul, I., Gazy, I., Mudau, I., Ssewanyana, I., Konstantinus, I., Lekana-Douk, J.B., Makangara, J.-C.C., Tamfum, J.-J.M., Heraud, J.-M., Shaffer, J.G., Giandhari, J., Li, J., Yasuda, J., Mends, J.Q., Kiconco, J., Morobe, J.M., Gyapong, J.O., Okolie, J.C., Kayiwa, J.T., Edwards, J.A., Gyamfi, J., Farah, J., Nakaseegu, J., Ngoi, J.M., Namulondo, J., Andeko, J.C., Lutwama, J.J., O'Grady, J., Siddle, K., Adeyemi, K.T., Tumedi, K.A., Said, K.M., Hae-Young, K., Duedu, K.O., Belyamani, L., Fki-Berrajah, L., Singh, L., Martins, L. de O., Tyers, L., Ramuth, M., Mastouri, M., Aouni, M., El Hefnawi, M., Matsheka, M.I., Kebabonye, M., Diop, M., Turki, M., Paye, M., Nyaga, M.M., Mareka, M., Damaris, M.-M., Mburu, M.W., Mpina, M., Nwando, M., Owusu, M., Wiley, M.R., Youtchou, M.T., Ayekaba, M.O., Abouelhoda, M., Seadawy, M.G., Khalifa, M.K., Sekhele, M., Ouadghiri, M., Diagne, M.M., Mwenda, M., Allam, M., Phan, M.V.T., Abid, N., Touil, N., Rujeni, N., Kharrat, N., Ismael, N., Dia, N., Mabunda, N., Hsiao, N.-Y., Silochi, N.B., Nsenga, N., Gumede, N., Mulder, N., Ndodo, N., Razanajatovo, N.H., Iguosadolo, N., Judith, O., Kingsley, O.C., Sylvanus, O., Peter, O., Femi, O., Idowu, O., Testimony, O., Chukwuma, O.E., Ogah, O.E., Onwuamah, C.K., Cyril, O., Faye, O., Tomori, O., Ondoa, P., Combe, P., Semanda, P., Oluniyi, P.E., Arnaldo, P., Quashie, P.K., Dussart, P., Bester, P.A., Mbala, P.K., Ayivor-Djanie, R., Njouom, R., Phillips, R.O., Gorman, R., Kingsley, R.A., Carr, R.A.A., El Kabbaj, S., Gargouri, S., Masmoudi, S., Sankhe, S., Lawal, S.B., Kassim, S., Trabelsi, S., Metha, S., Kammoun, S., Lemriss, S., Agwa, S.H.A., Calvignac-Spencer, S., Schaffner, S.F., Doumbia, S., Mandanda, S.M., Aryeetey, S., Ahmed, S.S., Elhamoumi, S., Andriamandimby, S.,

Tope, S., Lekana-Douki, S., Prosolek, S., Ouangraoua, S., Mundeke, S.A., Rudder, S., Panji, S., Pillay, S., Engelbrecht, S., Nabadda, S., Behillil, S., Budiaki, S.L., van der Werf, S., Mashe, T., Aanniz, T., Mohale, T., Le-Viet, T., Schindler, T., Anyaneji, U.J., Chinedu, U., Ramphal, U., Jessica, U., George, U., Fonseca, V., Enouf, V., Gorova, V., Roshdy, W.H., Ampofo, W.K., Preiser, W., Choga, W.T., Bediako, Y., Naidoo, Y., Butera, Y., de Laurent, Z.R., Sall, A.A., Rebai, A., von Gottberg, A., Kouriba, B., Williamson, C., Bridges, D.J., Chikwe, I., Bhiman, J.N., Mine, M., Cotten, M., Moyo, S., Gaseitsiwe, S., Saasa, N., Sabeti, P.C., Kaleebu, P., Tebeje, Y.K., Tessema, S.K., Happi, C., Nkengasong, J. and de Oliveira, T. (2021) 'A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa', *Science*, 374(6566), pp. 423–431.

21. Candido, D.S., Claro, I.M., de Jesus, J.G., Souza, W.M., Moreira, F.R.R., Dellicour, S., Mellan, T.A., du Plessis, L., Pereira, R.H.M., Sales, F.C.S., Manuli, E.R., Thézé, J., Almeida, L., Menezes, M.T., Voloch, C.M., Fumagalli, M.J., Coletti, T.M., da Silva, C.A.M., Ramundo, M.S., Amorim, M.R., Hoeltgebaum, H.H., Mishra, S., Gill, M.S., Carvalho, L.M., Buss, L.F., Prete, C.A., Ashworth, J., Nakaya, H.I., Peixoto, P.S., Brady, O.J., Nicholls, S.M., Tanuri, A., Rossi, Á.D., Braga, C.K.V., Gerber, A.L., de C Guimarães, A.P., Gaburo, N., Alencar, C.S., Ferreira, A.C.S., Lima, C.X., Levi, J.E., Granato, C., Ferreira, G.M., Francisco, R.S., Granja, F., Garcia, M.T., Moretti, M.L., Perroud, M.W., Tmpp, C., Lazari, C.S., Hill, S.C., de Souza Santos, A.A., Simeoni, C.L., Forato, J., Sposito, A.C., Schreiber, A.Z., Santos, M.N.N., de Sá, C.Z., Souza, R.P., Resende-Moreira, L.C., Teixeira, M.M., Hubner, J., Leme, P.A.F., Moreira, R.G., Nogueira, M.L., Ferguson, N.M., Costa, S.F., Proenca-Modena, J.L., Vasconcelos, A.T.R., Bhatt, S., Lemey, P., Wu, C.H., Rambaut, A., Loman, N.J., Aguiar, R.S., Pybus, O.G., Sabino, E.C. and Faria, N.R. (2020) 'Evolution and epidemic spread of SARS-CoV-2 in Brazil', *Science*, 369(6508), pp. 1255-1260.
22. Lu, J., du Plessis, L., Liu, Z., Hill, V., Kang, M., Lin, H., Sun, J., François, S., Kraemer, M.U.G., Faria, N.R., McCrone, J.T., Peng, J., Xiong, Q., Yuan, R., Zeng, L., Zhou, P., Liang, C., Yi, L., Liu, J., Xiao, J., Hu, J., Liu, T., Ma, W., Li, W., Su, J., Zheng, H., Peng, B., Fang, S., Su, W., Li, K., Sun, R., Bai, R., Tang, X., Liang, M., Quick, J., Song, T., Rambaut, A., Loman, N., Raghwani, J., Pybus, O.G. and Ke, C. (2020) 'Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China', *Cell*, 181(5), pp. 997–1003.e9.
23. Seemann, T., Lane, C.R., Sherry, N.L., Duchene, S., Gonçalves da Silva, A., Caly, L., Sait, M., Ballard, S.A., Horan, K., Schultz, M.B., Hoang, T., Easton, M., Dougall, S., Stinear, T.P., Druce, J., Catton, M., Sutton, B., van Diemen, A., Alpren, C., Williamson, D.A. and Howden, B.P. (2020) 'Tracking the COVID-19 pandemic in Australia using genomics', *Nature Communications*, 11(1), pp. 1–9.
24. Komissarov, A.B., Safina, K.R., Garushyants, S.K., Fadeev, A.V., Sergeeva, M.V., Ivanova, A.A., Danilenko, D.M., Lioznov, D., Shneider, O.V., Shvyrev, N., Spirin, V., Glyzin, D., Shchur, V. and Bazykin, G.A. (2021) 'Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak in Russia', *Nature communications*, 12(1), pp. 1–13.
25. Ismael, N., van Wyk, S., Tegally, H., Giandhari, J., San, J.E., Moir, M., Pillay, S., Utpatel, C., Singh, L., Naidoo, Y., Ramphal, U., Mabunda, N., Abílio, N., Arnaldo, P., Xavier, J., Amoako, D.G., Everatt, J., Ramphal, Y., Maharaj, A., de Araujo, L., Anyaneji, U.J., Tshiabuila, D., Viegas, S., Lessells, R., Engelbrecht, S., Gudo, E., Jani, I., Niemann, S., Wilkinson, E. and de Oliveira, T. (2023) 'Genomic

epidemiology of SARS-CoV-2 during the first four waves in Mozambique', *PLOS Global Public Health*, 3(3), p. e0001593.
26. Martínez-Martínez, F.J., Massinga, A.J., De Jesus, Á., Ernesto, R.M., Cano-Jiménez, P., Chiner-Oms, Á., Gómez-Navarro, I., Guillot-Fernández, M., Guinovart, C., Sitoe, A., Vubil, D., Bila, R., Gujamo, R., Enosse, S., Jiménez-Serrano, S., Torres-Puente, M., Comas, I., Mandomando, I., López, M.G. and Mayor, A. (2023) 'Tracking SARS-CoV-2 introductions in Mozambique using pandemic-scale phylogenies: a retrospective observational study', *The Lancet. Global health*, 11(6), pp. e933–e941.
27. Douglas, J., Winter, D., McNeill, A., Carr, S., Bunce, M., French, N., Hadfield, J., de Ligt, J., Welch, D. and Geoghegan, J.L. (2022) 'Tracing the international arrivals of SARS-CoV-2 Omicron variants after Aotearoa New Zealand reopened its border', *Nature communications*, 13(1), p. 6484.
28. Trovao, N.S., Pan, V., Goel, C., Gallego-García, P., Liu, Y., Barbara, C., Borg, R., Briffa, M., Cilia, C., Grech, L., Vassallo, M., Treangen, T.J., Posada, D., Beheshti, A., Borg, J. and Zahra, G. (2023) 'Evolutionary and spatiotemporal analyses reveal multiple introductions and cryptic transmission of SARS-CoV-2 VOC/VOI in Malta', *Microbiology spectrum*, 11(6), p. e0153923.
29. Raghwani, J., du Plessis, L., McCrone, J.T., Hill, S.C., Parag, K.V., Thézé, J., Kumar, D., Puvar, A., Pandit, R., Pybus, O.G., Fournié, G., Joshi, M. and Joshi, C. (2022) 'Genomic Epidemiology of Early SARS-CoV-2 Transmission Dynamics, Gujarat, India', *Emerging infectious diseases*, 28(4), pp. 751–758.
30. Githinji, G., de Laurent, Z.R., Mohammed, K.S., Omuoyo, D.O., Macharia, P.M., Morobe, J.M., Otieno, E., Kinyanjui, S.M., Agweyu, A., Maitha, E., Kitole, B., Suleiman, T., Mwakinangu, M., Nyambu, J., Otieno, J., Salim, B., Kasera, K., Kiiru, J., Aman, R., Barasa, E., Warimwe, G., Bejon, P., Tsofa, B., Ochola-Oyier, L.I., Nokes, D.J. and Agoti, C.N. (2021) 'Tracking the introduction and spread of SARS-CoV-2 in coastal Kenya', *Nature communications*, 12(1), p. 4809.
31. Ari, E., Vásárhelyi, B.M., Kemenesi, G., Tóth, G.E., Zana, B., Somogyi, B., Lanszki, Z., Röst, G., Jakab, F., Papp, B. and Kintses, B. (2022) 'A single early introduction governed viral diversity in the second wave of SARS-CoV-2 epidemic in Hungary', *Virus evolution*, 8(2), p. veac069.
32. Reichmuth, M.L., Hodcroft, E.B. and Althaus, C.L. (2023) 'Importation of Alpha and Delta variants during the SARS-CoV-2 epidemic in Switzerland: Phylogenetic analysis and intervention scenarios', *PLoS pathogens*, 19(8), p. e1011553.
33. Michaelsen, T.Y., Bennedbæk, M., Christiansen, L.E., Jørgensen, M.S.F., Møller, C.H., Sørensen, E.A., Knutsson, S., Brandt, J., Jensen, T.B.N., Chiche-Lapierre, C., Collados, E.F., Sørensen, T., Petersen, C., Le-Quy, V., Sereika, M., Hansen, F.T., Rasmussen, M., Fonager, J., Karst, S.M., Marvig, R.L., Stegger, M., Sieber, R.N., Skov, R., Legarth, R., Krause, T.G., Fomsgaard, A., Danish COVID-19 Genome Consortium (DCGC) and Albertsen, M. (2022) 'Introduction and transmission of SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark', *Genome medicine*, 14(1), p. 47.
34. Qutob, N., Salah, Z., Richard, D., Darwish, H., Sallam, H., Shtayeh, I., Najjar, O., Ruzayqat, M., Najjar, D., Balloux, F. and van Dorp, L. (2021) 'Genomic epidemiology of the first epidemic wave of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Palestine', *Microbial genomics*, 7(6), 000584.
35. Vöhringer, H.S., Sanderson, T., Sinnott, M., De Maio, N., Nguyen, T., Goater, R., Schwach, F., Harrison, I., Hellewell, J., Ariani, C.V., Gonçalves, S., Jackson, D.K., Johnston, I., Jung, A.W., Saint, C., Sillitoe, J., Suciu, M., Goldman, N., Panovska-Griffiths, J., Wellcome Sanger Institute COVID-19 Surveillance Team, COVID-19

Genomics UK (COG-UK) Consortium*, Birney, E., Volz, E., Funk, S., Kwiatkowski, D., Chand, M., Martincorena, I., Barrett, J.C. and Gerstung, M. (2021) 'Genomic reconstruction of the SARS-CoV-2 epidemic in England', *Nature*, 600(7889), pp. 506–511.
36. Di Giallonardo, F., Duchene, S., Puglia, I., Curini, V., Profeta, F., Cammà, C., Marcacci, M., Calistri, P., Holmes, E.C. and Lorusso, A. (2020) 'Genomic Epidemiology of the First Wave of SARS-CoV-2 in Italy', *Viruses*, 12(12), p. 1438.
37. Paiva, M.H.S., Guedes, D.R.D., Docena, C., Bezerra, M.F., Dezordi, F.Z., Machado, L.C., Krokovsky, L., Helvecio, E., da Silva, A.F., Vasconcelos, L.R.S., Rezende, A.M., da Silva, S.J.R., Sales, K.G. da S., de Sá, B.S.L.F., da Cruz, D.L., Cavalcanti, C.E., Neto, A. de M., da Silva, C.T.A., Mendes, R.P.G., da Silva, M.A.L., Gräf, T., Resende, P.C., Bello, G., Barros, M. da S., do Nascimento, W.R.C., Arcoverde, R.M.L., Bezerra, L.C.A., Filho, S.P.B., Ayres, C.F.J. and Wallau, G.L. (2020) 'Multiple Introductions Followed by Ongoing Community Spread of SARS-CoV-2 at One of the Largest Metropolitan Areas of Northeast Brazil', *Viruses*, 12(12), 1414.
38. Douglas, J., Mendes, F.K., Bouckaert, R., Xie, D., Jiménez-Silva, C.L., Swanepoel, C., de Ligt, J., Ren, X., Storey, M., Hadfield, J., Simpson, C.R., Geoghegan, J.L., Drummond, A.J. and Welch, D. (2021) 'Phylodynamics reveals the role of human travel and contact tracing in controlling the first wave of COVID-19 in four island nations', *Virus evolution*, 7(2), p. veab052.
39. Tsui, J.L.-H., McCrone, J.T., Lambert, B., Bajaj, S., Inward, R.P.D., Bosetti, P., Pena, R.E., Tegally, H., Hill, V., Zarebski, A.E., Peacock, T.P., Liu, L., Wu, N., Davis, M., Bogoch, I.I., Khan, K., Kall, M., Abdul Aziz, N.I.B., Colquhoun, R., O'Toole, Á., Jackson, B., Dasgupta, A., Wilkinson, E., de Oliveira, T., COVID-19 Genomics UK (COG-UK) consortium¶, Connor, T.R., Loman, N.J., Colizza, V., Fraser, C., Volz, E., Ji, X., Gutierrez, B., Chand, M., Dellicour, S., Cauchemez, S., Raghwani, J., Suchard, M.A., Lemey, P., Rambaut, A., Pybus, O.G. and Kraemer, M.U.G. (2023) 'Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1', *Science*, 381(6655), pp. 336–343.
40. Dellicour, S., Hong, S.L., Vrancken, B., Chaillon, A., Gill, M.S., Maurano, M.T., Ramaswami, S., Zappile, P., Marier, C., Harkins, G.W., Baele, G., Duerr, R. and Heguy, A. (2021) 'Dispersal dynamics of SARS-CoV-2 lineages during the first epidemic wave in New York City', *PLoS pathogens*, 17(5), p. e1009571.
41. Dellicour, S., Durkin, K., Hong, S.L., Vanmechelen, B., Martí-Carreras, J., Gill, M.S., Meex, C., Bontems, S., André, E., Gilbert, M., Walker, C., Maio, N.D., Faria, N.R., Hadfield, J., Hayette, M.-P., Bours, V., Wawina-Bokalanga, T., Artesi, M., Baele, G. and Maes, P. (2020) 'A Phylodynamic Workflow to Rapidly Gain Insights into the Dispersal History and Dynamics of SARS-CoV-2 Lineages', *Molecular biology and evolution*, 38(4), pp. 1608–1613.
42. Parker, E., Anderson, C., Zeller, M., Tibi, A., Havens, J.L., Laroche, G., Benlarbi, M., Ariana, A., Robles-Sikisaka, R., Latif, A.A., Watts, A., Awidi, A., Jaradat, S.A., Gangavarapu, K., Ramesh, K., Kurzban, E., Matteson, N.L., Han, A.X., Hughes, L.D., McGraw, M., Spencer, E., Nicholson, L., Khan, K., Suchard, M.A., Wertheim, J.O., Wohl, S., Côté, M., Abdelnour, A., Andersen, K.G. and Abu-Dayyeh, I. (2022) 'Regional connectivity drove bidirectional transmission of SARS-CoV-2 in the Middle East during travel restrictions', *Nature communications*, 13(1), p. 4784.
43. Butera, Y., Mukantwari, E., Artesi, M., Umuringa, J.D., O'Toole, Á.N., Hill, V., Rooke, S., Hong, S.L., Dellicour, S., Majyambere, O., Bontems, S., Boujemla, B., Quick, J., Resende, P.C., Loman, N., Umumararungu, E., Kabanda, A., Murindahabi, M.M., Tuyisenge, P., Gashegu, M., Rwabihama, J.P., Sindayiheba, R., Gikic, D.,

Souopgui, J., Ndifon, W., Rutayisire, R., Gatare, S., Mpunga, T., Ngamije, D., Bours, V., Rambaut, A., Nsanzimana, S., Baele, G., Durkin, K., Mutesa, L. and Rujeni, N. (2021) 'Genomic sequencing of SARS-CoV-2 in Rwanda reveals the importance of incoming travelers on lineage diversity', *Nature communications*, 12(1), p. 5705.
44. Elizondo, V., Harkins, G.W., Mabvakure, B., Smidt, S., Zappile, P., Marier, C., Maurano, M.T., Perez, V., Mazza, N., Beloso, C., Ifran, S., Fernandez, M., Santini, A., Perez, V., Estevez, V., Nin, M., Manrique, G., Perez, L., Ross, F., Boschi, S., Zubillaga, M.N., Balleste, R., Dellicour, S., Heguy, A. and Duerr, R. (2021) 'SARS-CoV-2 genomic characterization and clinical manifestation of the COVID-19 outbreak in Uruguay', *Emerging microbes & infections*, 10(1), pp. 51–65.
45. Castelán-Sánchez, H.G., Delaye, L., Inward, R.P.D., Dellicour, S., Gutierrez, B., Martinez de la Vina, N., Boukadida, C., Pybus, O.G., de Anda Jáuregui, G., Guzmán, P., Flores-Garrido, M., Fontanelli, Ó., Hernández Rosales, M., Meneses, A., Olmedo-Alvarez, G., Herrera-Estrella, A.H., Sánchez-Flores, A., Muñoz-Medina, J.E., Comas-García, A., Gómez-Gil, B., Zárate, S., Taboada, B., López, S., Arias, C.F., Kraemer, M.U.G., Lazcano, A. and Escalera Zamudio, M. (2023) 'Comparing the evolutionary dynamics of predominant SARS-CoV-2 virus lineages co-circulating in Mexico', *eLife*, 12, e82069.
46. Dellicour, S., Hong, S.L., Hill, V., Dimartino, D., Marier, C., Zappile, P., Harkins, G.W., Lemey, P., Baele, G., Duerr, R. and Heguy, A. (2023) 'Variant-specific introduction and dispersal dynamics of SARS-CoV-2 in New York City - from Alpha to Omicron', *PLoS pathogens*, 19(4), p. e1011348.
47. Han, A.X., Kozanli, E., Koopsen, J., Vennema, H., RIVM COVID-19 molecular epidemiology group, Aarts, L., Bos, S., van den Brandt, A., van den Brink, S., Cremer, J., Freriks, K., Jaarsma, R., Schmitz, D., Then, E., van der Veer, B., Wijsman, L., Zwagemaker, F., Hajji, K., Kroneman, A., van Walle, I., Klinkenberg, D., Wallinga, J., Russell, C.A., Eggink, D. and Reusken, C. (2022) 'Regional importation and asymmetric within-country spread of SARS-CoV-2 variants of concern in the Netherlands', *eLife*, 11, p. e78770.
48. Castelán-Sánchez, H.G., Martínez-Castilla, L.P., Sganzerla-Martínez, G., Torres-Flores, J. and López-Leal, G. (2022) 'Genome Evolution and Early Introductions of the SARS-CoV-2 Omicron Variant in Mexico', *Virus evolution*, 8(2), p. veac109.
49. Ferretti, L., Wymant, C., Petrie, J., Tsallis, D., Kendall, M., Ledda, A., Di Lauro, F., Fowler, A., Di Francia, A., Panovska-Griffiths, J., Abeler-Dörner, L., Charalambides, M., Briers, M. and Fraser, C. (2023) 'Digital measurement of SARS-CoV-2 transmission risk from 7 million contacts', *Nature*, 626(7997), pp. 145–150.
50. Thompson, H.A., Mousa, A., Dighe, A., Fu, H., Arnedo-Pena, A., Barrett, P., Bellido-Blasco, J., Bi, Q., Caputi, A., Chaw, L., De Maria, L., Hoffmann, M., Mahapure, K., Ng, K., Raghuram, J., Singh, G., Soman, B., Soriano, V., Valent, F., Vimercati, L., Wee, L.E., Wong, J., Ghani, A.C. and Ferguson, N.M. (2021) 'Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Setting-specific Transmission Rates: A Systematic Review and Meta-analysis', *Clinical Infectious Diseases*, 73(3), pp. e754–e764.
51. Byrne, A.W., McEvoy, D., Collins, A.B., Hunt, K., Casey, M., Barber, A., Butler, F., Griffin, J., Lane, E.A., McAloon, C., O'Brien, K., Wall, P., Walsh, K.A. and More, S.J. (2020) 'Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases', *BMJ Open*, 10(8), p. e039856.
52. Cevik, M., Tate, M., Lloyd, O., Maraolo, A.E., Schafers, J. and Ho, A. (2021) 'SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding,

and infectiousness: a systematic review and meta-analysis', *The Lancet. Microbe*, 2(1), pp. e13-e22.
53. Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G.S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M. and John Edmunds, W. (2008) 'Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases', *PLOS Medicine*, 5(3), p. e74.
54. Wong, K.L.M., Gimma, A., Coletti, P., CoMix Europe Working Group, Faes, C., Beutels, P., Hens, N., Jaeger, V.K., Karch, A., Johnson, H., Edmunds, W. and Jarvis, C.I. (2023) 'Social contact patterns during the COVID-19 pandemic in 21 European countries – evidence from a two-year study', *BMC Infectious Diseases*, 23, p. 268.

5

Optimising genomic sampling for phylogeographic inference: a simulation-based evaluation framework

Building on the work presented in Chapter 4, where I examined the effect of undersampling of local infections on the detection of viral importation, this chapter explores the broader impact of heterogeneous sampling on discrete phylogeographic inference by first introducing a new simulation-based evaluation called SOPHI ("*Sandbox for Optimising genomic sampling for PHylogeographic Inference*"). Using this framework, I then investigate: 1) the impact of undersampling at both the source and recipient locations on the detection of viral importation events under realistic outbreak conditions, and 2) the impact of sampling biases on identifying the sources of early viral importation under three commonly used sampling schemes.

The work presented in this chapter is based on a manuscript currently in preparation, under the title "*SOPHI: Sandbox for Optimising genomic sampling for PHylogeographic Inference*".

Tsui, J.L.-H., Inward, R., Kraemer, M.U.G. and Pybus, O.G. (in preparation) '*SOPHI: Sandbox for Optimising genomic sampling for PHylogeographic Inference*'.

5.1 Introduction

Phylogeographic methods have become an essential tool for reconstructing the spatiotemporal dynamics of pathogen spread, especially during large-scale outbreaks

where traditional epidemiological data are scarce or incomplete. By integrating genomic and geographic data, these approaches enable the inference of viral importation events and patterns of pathogen dissemination across multiple spatial scales. This has been showcased in numerous studies examining the local and global transmission dynamics of viruses such as Ebola (1, 2), seasonal influenza (3-6), Zika (7, 8), and most recently, SARS-CoV-2 (9-14), with the work presented in Chapter 2 providing one example (15). In particular, in Chapter 2, I applied a combination of continuous (16) and discrete (17) phylogeographic approaches to characterise both the initial introduction and subsequent local dissemination of SARS-CoV-2 Omicron BA.1 in the UK. While continuous spatial approaches effectively capture patterns of fine-scale diffusive spread within local regions or cities (18, 19), transmission among human populations over larger spatial scales is often better represented as discrete "jumps" between geographically distant locations, reflecting modern human mobility dominated by structured transportation networks such as highways, railways, and international flight routes (3, 20). As a result, discrete phylogeography has generally been the preferred approach for studying outbreaks among populations at larger geographic scales.

Encouraged by the increasing adoption of discrete phylogeography, theoretical advances have led to the development of several distinct approaches. Early phylogeographic studies relied primarily on heuristic methods such as parsimony, which assign geographic states to ancestral nodes by minimising the number of state changes along a fixed phylogeny (21-24). While simple and computationally efficient, these methods do not account for branching times, uncertainty in evolutionary histories, and heterogeneous patterns of pathogen movement (25-28). These limitations led to the development of probabilistic approaches, notably discrete trait analysis (DTA) - which models pathogen movement along phylogeny branches as a continuous-time Markov

process (17) - and structured coalescent (SC) approaches, which explicitly infer how lineages migrate between and coalesce within defined subpopulations (29-31). Both DTA and SC have been implemented in popular phylodynamic software tools (e.g., BEAST1 (32) and BEAST2 (33)); however, DTA is employed more frequently due to its computational efficiency and capacity to handle large datasets, as demonstrated in Chapter 2. Although more efficient approaches that rely on SC approximations such as BASTA (29) and MASCOT (34) have been developed, their applications to real-world outbreak analyses have so far been limited due to the computational demands of analysing increasingly large genomic datasets, especially since the COVID-19 pandemic.

Despite its growing popularity, the use of phylogeography in outbreak response is not without challenges. In particular, it is well recognised that discrete phylogeographic methods, especially DTA, are susceptible to biases introduced by uneven sampling of pathogen genomes across both space and time. For example, analyses of simulated outbreaks have shown that DTA systematically underestimates viral migration from under-sampled locations while overestimating the relative importance of densely-sampled locations (29, 35-37). During the COVID-19 pandemic, intensive genomic surveillance efforts in some European and North American countries led to substantial sampling biases, under-representing viral lineages originating from regions where sequencing was infrequent or absent, and resulting in incomplete or inaccurate reconstruction of global viral dissemination pathways (9, 38-40). These observations motivated the development of a number of mitigation strategies, with some relying on the incorporation of additional data streams, such as reported case incidence and individual travel histories (38, 41). The increased availability of pathogen genomes, driven by the lowering cost of genomic sequencing, has also enabled the use of downsampling approaches to mitigate such biases. For instance, in the continuous phylogeographic

analysis presented in Chapter 2, local BA.1 sequences were randomly subsampled in proportion to weekly case counts at the district level to ensure that the distribution of genome samples reflected the local disease prevalence while maintaining a manageable dataset size; whereas in the importation analysis using discrete phylogeography, an equal number of samples were selected from UK and non-UK countries to balance the need to 1) sample from as many independent local transmission lineages as possible, and 2) ensure that viral genetic diversity in locations that were likely sources of BA.1 importation was sufficiently sampled to prevent the aggregation of independent transmission lineages. Similarly, in a study investigating the transmission dynamics of SARS-CoV-2 lineages in Mexico (42), Castelán-Sánchez et al. employed a mobility-informed strategy where sequences from countries with the highest incoming human mobility flux are prioritised.

Although variations of these sampling (or downsampling) strategies have been applied in numerous SARS-CoV-2 studies, the extent to which they are able to mitigate different sampling biases remains poorly understood and has received limited systematic evaluation. More importantly, the process of deciding which strategy to employ, and how to calibrate relevant parameters for specific outbreak scenarios, has been largely *ad hoc*, with little consensus on best practices. Several factors contribute to this lack of progress. First, the optimal sampling design depends on the specific research question or objective, which may include: a) estimating the true number of migration or importation events; b) inferring temporal trends in migration or importation intensity; c) identifying key migratory pathways or sources of key importation events; or d) characterising source-sink dynamics (i.e. the direction and relative intensity of pathogen movement between locations). The nature and degree of bias introduced by heterogeneous sampling often differ across these objectives, with each requiring a tailored sampling strategy. Second,

even when a research objective is clearly defined, the performance of a given sampling strategy is strongly influenced by context-specific factors - such as underlying migration patterns and local transmission intensities (as demonstrated in Chapter 4), or whether travellers are preferentially sampled and sequenced, as examined by Liu et al. (36). Third, many sampling strategies require the specification of a large number of design parameters (e.g., number of sequences to include, distribution of sampling quotas across space and time), especially when additional data streams such as case incidence data and human mobility patterns are incorporated. Together, these challenges introduce substantial hurdles to efforts to systematically evaluate and optimise the design of sampling strategies at scale. As a result, most simulation-based studies to date have explored only a narrow range of outbreak conditions and design parameters, limiting our ability to draw generalisable insights and robust sampling principles.

To address these challenges, here I develop a new computational framework that simulates each step of a phylogeographic analysis in a virtual environment - from understanding the outbreak context and designing an appropriate sampling strategy, to executing a discrete phylogeographic inference and evaluating its accuracy against a simulated "ground truth". Through an interactive graphical interface with real-time feedback and visualisations of key summary statistics, researchers can rapidly iterate on different sampling strategies and assess their performance with respect to specific research objectives across a wide range of outbreak scenarios. The framework is implemented as an open-source web platform called SOPHI ("*Sandbox for Optimising genomic sampling for PHylogeographic Inference*"), which is freely available at www.sophi-oxf.io. I then apply this framework to address two open questions in spatial genomic epidemiology that arose from the work presented in earlier chapters. Specifically, I investigate 1) how undersampling affects the detection of viral importation

events, building on results from Chapter 4, and 2) how different sampling schemes influence source attribution for early importation events, under varying degrees of sampling bias. In addition to providing insights into how sampling heterogeneities lead to biased phylogeographic inference and how effectively existing mitigation strategies are able to correct for such biases, these applications also demonstrate SOPHI's broader utility as a practical framework for supporting a more systematic and scalable approach to future research inquiries on phylogeographic inference.

5.2 Design of a simulation-based evaluation framework (SOPHI)

In this section, I outline the core methodological components of the SOPHI framework, including outbreak simulation, the design of sampling schemes, discrete phylogeographic reconstruction, and the evaluation of inference accuracy using simulated data. Details related to software implementation are provided in the Appendix D (Section D.1 and Fig. D.1).

5.2.1 Stochastic outbreak and tree simulation using ReMASTER

A core component of SOPHI is the generation of realistic outbreak scenarios, which provide the basis for downstream inferences and evaluation of their outputs. Here I use ReMASTER (43), a software tool implemented within the BEAST2 framework (33) that simulates outbreak dynamics and corresponding phylogenies. Although ReMASTER supports a wide range of compartmental models, I focus here on a relatively simple metapopulation SIR (Susceptible-Infected-Removed) model, in which individuals are distributed among distinct demes (i.e. subpopulation corresponding to geographic locations or other structured groups) and can migrate between them at predefined rates

M_{ij} for each pair of connected demes i and j (Fig. 5.1). Each deme i is associated with a transmission coefficient β_i and recovery rate γ_i, which I assume to be constant for simplicity. Each simulated outbreak produces a set of observables, i.e. data that would typically be available to researchers during an outbreak investigation, such as reported case incidence and sampled pathogen genomes. Depending on predefined assumptions, infectious individuals in each deme i are either: 1) reported at a constant rate ν_i followed by immediate recovery, or 2) reported immediately upon infection (i.e. $\nu_i \to \infty$) without recovery (as shown in Fig. 1). The latter is adopted for the two applications presented later in this chapter, representing an idealised scenario in which true case incidence is known. Similarly, infectious individuals in each deme i are sampled and sequenced at a constant rate δ_i, after which they are immediately removed from the infectious compartment.

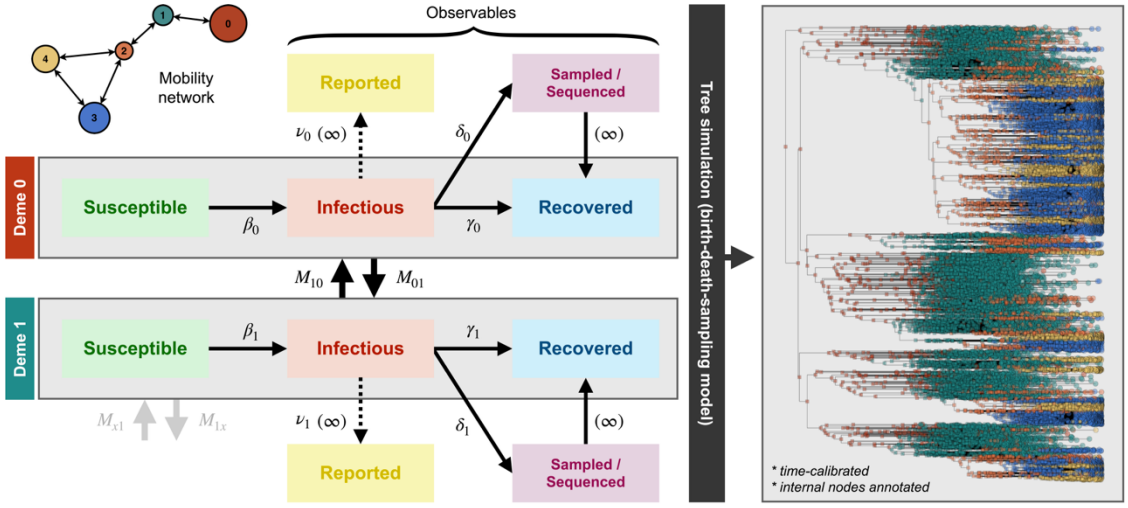


Fig. 5.1. Outbreak simulation using ReMASTER. (Left) Schematic of an outbreak simulation across a mobility network with 5 demes. Local epidemic in each deme i is governed by a standard Susceptible-Infectious-Recovered (SIR) model with deme-specific transmission coefficient β_i and recovery rate γ_i. Infectious individuals in each deme i are reported at a constant rate ν_i and sampled/sequenced at a constant rate δ_i, followed by immediate recovery. In this specific example, it is assumed that infectious individuals are reported immediately upon infection (i.e. $\nu_i \to \infty$) and without recovery, representing an idealised scenario where true case incidence is known. Individuals can migrate freely, regardless of their infection status, between any connected demes i and j

at a predefined constant rate M_{ij}. (Right) In addition to outbreak trajectories, ReMASTER also produces a time-calibrated phylogeny representing the ancestral relationships between sampled infections, with internal nodes annotated by their true corresponding deme (indicated by node colours in this visualisation, generated by SOPHI).

Given these model specifications, ReMASTER produces two primary outputs: 1) population trajectories, recording the size of each SIR compartment after every event (i.e., transmission, recovery, migration, or sampling), and 2) an annotated, time-calibrated phylogeny representing the ancestral relationships between sampled infections, simulated under a birth-death-sampling model (44). These outputs serve both as synthetic observables as described above, and as "ground truth" datasets against which inference accuracy can be evaluated. Importantly, the simulated phylogeny generated by ReMASTER allows the tree estimation process to be emulated by pruning tips corresponding to unsampled infections according to a given sampling strategy. This approach offers substantial gains in computational efficiency by avoiding the need to re-estimate a phylogeny from sequence alignments for each inference - a process that would otherwise take hours or even days for large datasets (>10,000 sequences) using standard software tools. However, this comes at a cost of not accounting for uncertainty in the inferred tree topology and branch lengths. I return to this trade-off between computational tractability and simulation realism in the Discussion.

5.2.2 Design of sampling schemes

Given a simulated outbreak and relevant observables, the next step in a typical phylogeographic analysis is to design a sampling or downsampling strategy to determine which of the available genomes should be included in the inference. To facilitate this process, SOPHI implements a number of sampling schemes commonly employed in

published studies. These sampling schemes are categorised into two domains: spatial and temporal; by combining predefined schemes from each domain, it is possible to construct a wide range of different sampling strategies tailored to specific research questions. For instance, a study investigating how early viral importation contributes to the establishment of local transmission might downsample sequences in proportion to reported case numbers across demes, while prioritising the earliest available samples within each deme (15, 45, 46). In contrast, a study estimating the relative number of importation events from different sources might allocate samples based on importation risk, estimated from reported case numbers and mobility data (42). Table 1 provides an overview of sampling schemes currently implemented in SOPHI and selected studies in which each has been applied.

Once a sampling strategy has been specified for each domain, they can be applied either sequentially or jointly. In a sequential design, sampling is performed in each in a stepwise manner – for instance, a common approach involves allocating sampling quotas evenly across demes to ensure broad geographic coverage, with samples then drawn within each deme in proportion to number of reported cases per day or week (with the importation analysis presented Chapter 2 being one example). Whereas in a joint design, a predefined number or proportion of samples are drawn at random, where each sample is assigned a selection probability determined by some function of its spatial and temporal attributes. Under joint-even-sampling, for example, a sample collected on day t from deme i would be assigned a selection probability inversely proportional to the number of samples available on that day and in that deme - ensuring an even coverage of samples across both space and time.

Table 5.1. Overview of commonly used spatial and temporal sampling schemes that are currently implemented in SOPHI. The shorthand notations (e.g., US, UC, EV) are shown next to the full name of each scheme. The last column shows selected SARS-CoV-2 studies that have applied the corresponding sampling scheme(s).

Domain	**Strategy**	**Brief description**	**Selected relevant studies**
Spatial	Uniform-sample (US)	- Downsample uniformly across all available samples in each deme i - Equivalent to assigning each available sample an equal weight $w_i = 1$ for all deme i	(13, 45, 47-50)
	Uniform-case (UC)	- Downsample in proportion to the number of reported cases in each deme i - Equivalent to assigning each available sample in deme i a weight $w_i = c_i/s_i$, where c_i is the total number of reported cases and s_i is the total number of available samples in deme i	(15, 41, 45, 46)
	Even (EV)	- Downsample in inverse proportion to number of available samples in each deme i - Equivalent to assigning each sample in deme i a weight $w_i = 1/s_i$, where s_i is the total number of available samples in deme i	(10, 15, 42, 49)
Temporal	Uniform-sample (US)	- Downsample uniformly across all available samples in each time-window $[t, t + \Delta t]$ (where Δt could be in units of days or weeks) - Equivalent to assigning each available sample an equal weight $w_t = 1$ for all sampling time t	(13, 45, 47–50)
	Uniform-case (UC)	- Downsample in proportion to the number of reported cases in each time-window $[t, t + \Delta t]$ - Equivalent to assigning each available sample in time-window $[t, t + \Delta t]$ a weight $w_t = c_t/s_t$, where c_t is the total number of reported cases and s_t is the total number of available samples in time-window $[t, t + \Delta t]$	(15, 41, 42, 46)

	Earliest-N (EN)	- Select the earliest N (or a fixed proportion) of available samples by collection date	(15, 45, 46)
	Even (EV)	- Downsample in inverse proportion to number of available samples in each time-window $[t, t + \Delta t]$ - Equivalent to assigning each available sample in time-window $[t, t + \Delta t]$ a weight $w_t = 1/s_t$, where s_t is the total number of available samples in time-window $[t, t + \Delta t]$	(10, 41, 46, 48, 49)

5.2.3 Discrete phylogeographic inference and evaluation metrics

Having defined a sampling strategy, the next step is to perform a discrete phylogeographic inference using the selected samples. In the SOPHI framework, this begins with first pruning tips corresponding to excluded samples from the full simulated tree, emulating the process of tree estimation given a subsampled dataset. The resulting pruned tree, along with relevant metadata for the retained samples (i.e. sampling time and location), is then used as input for discrete trait analysis (DTA).

SOPHI currently supports two widely used DTA methods: 1) parsimony-based inference, implemented via the *phangorn* package in R (51), and 2) maximum-likelihood inference, implemented via TreeTime in Python (52). Both methods produce an annotated pruned tree with internal nodes labelled by their most probably ancestral locations, from which migratory events can be identified (by tracing a path from the root node of the pruned tree to each tip and observing any changes in the inferred or observed state; same procedure as that applied in the importation analysis in Chapter 2). Additionally, SOPHI computes for each inferred migratory event: 1) the inferred origin and destination deme, 2) the size of its associated transmission lineage, defined as the number of sampled infections, 3) Time to the Most Recent Common Ancestor (TMRCA) and Time to the

Parent of the MRCA (TPMRCA) of its associated lineage, and 4) the interval between its estimated time of occurrence (using TPMRCA, TMRCA, or their mid-point as proxy) and the time of its most recent sampled descendant (also commonly known as lineage persistence (11)).

Depending on the research question, the accuracy of an inference can be evaluated either through direct comparison between the inferred value of one of the above estimates and its corresponding true value from the simulated data, or using more tailored metrics. For example, when assessing the degree to which temporal variations in migration rate can be accurately recovered, similarity between the inferred and true importation rates at a given location can be assessed using Pearson correlation or cross-correlation analysis (both automatically computed in SOPHI); whereas distance metrics such as Jaccard similarity or cosine-distance can be used to evaluate whether the relative intensity of viral importation from different source locations can be reliably inferred under varying degrees of sampling bias (as demonstrated in application 2 later in this chapter).

5.3 Applications

In this section, I apply the SOPHI framework described and implemented above to investigate two open questions related to the impact of heterogeneous sampling on discrete phylogeographic inference.

5.3.1 Application 1: impact of undersampling on the detection of viral importation

In Chapter 4, I examined how the coupling between viral importation and local transmission dynamics shapes the size distribution of local transmission lineages, which

in turn influences the number of importation events inferred or detected in a discrete phylogeographic reconstruction. Specifically, assuming complete sampling of all infections at the source location in a 2-deme mobility network, I showed that the lineage detection probability - defined as the proportion of true importation events identified - increases with local sampling proportion at different rates depending on the underlying lineage size distribution at the recipient location. When the importation rate is decreasing over time, the local epidemic at the recipient location is dominated by earlier-introduced lineages which are larger and therefore easier to detect even at low sampling proportions. Conversely, when the importation rate is increasing over time, more recent and thus smaller lineages tend to dominate - resulting in lineage detection probabilities that increase almost linearly with sampling proportion.

However, the analytical and simulation models developed in Chapter 4 made two key assumptions:

1. Equal sampling probability for any previously infected individuals, including individuals who are no longer infectious at the time of observation. In reality, however, sampling occurs continuously through time, with only infectious individuals being eligible for sequencing. This has two important implications: a) infections that occur earlier in the outbreak are more likely to be sampled, simply because they are eligible for sampling for longer, and b) only a subset of infections are typically sampled and sequenced, as infectious individuals who recovered are not eligible. Specifically, given a constant sampling rate δ and recovery rate γ, the expected overall sampling proportion is given by $p_{sample} = \delta/(\delta + \gamma)$, assuming a closed population with no migration.
2. Complete sampling of infections in the source population. This is rarely feasible in practice, due to both competition between sampling and recovery (as discussed in (1))

and logistical constraints. It is generally known that undersampling of genetic diversity in the source population leads to aggregation of independent transmission lineages detectable from the local genome samples, and therefore further underestimation of the number of importation events - even when sampling proportion at the recipient location is high. However, this effect has not been rigorously examined and quantified in existing literature.

In this application, I relax the above assumptions and use SOPHI to explore how sampling proportion at both the source and recipient locations affects the detection of viral importation.

5.3.1.1 Outbreak scenario

Following a setup similar to that in Chapter 4, here I consider two contrasting outbreak scenarios in a 2-deme mobility network. In scenario 1, individuals (regardless of their infection status) are free to travel between the two demes from the beginning of the outbreak (up to day 30, before the local prevalence at the source location reaches a peak), leading to an increasing importation rate; in scenario 2, movement between the two demes is restricted until day 35 (after the peak in local prevalence at the source location), resulting in a decreasing importation rate (with the simulation ending on day 65). In both scenarios, the two demes (each with an initial population size of 100,000) are connected by human movement (500 individuals per day in both directions), with a transmission coefficient $\beta = 0.5$ and recovery rate $\gamma = 0.12$ per day. Sampling of infectious individuals occur at a constant rate $\delta = 0.03$ per day in both demes.

5.3.1.2 Models specifications and experimental setup

Using SOPHI, I perform two sets of experiments under each outbreak scenario. In the first set, I vary the sampling proportion at the recipient location (deme 1) (specifically, $s_1 = 0.01$, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64, and 1, while including all available samples from the source location (deme 0), i.e. $s_0 = 1$. In the second set, I vary the sampling proportion at the source location (deme 0) (specifically, $s_0 = 0.01$, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64, and 1), while including all available samples from the recipient location (deme 1), i.e. $s_1 = 1$. In each experiment, samples are selected uniformly at random (equivalent to the uniform-sample (US) sampling scheme described in Table 1), and 20 replicate inferences are performed using parsimony-based DTA (via the *phangorn* R package, accessed through the SOPHI interface). Inferred importation events with ambiguous origin or destination states are excluded from downstream analysis.

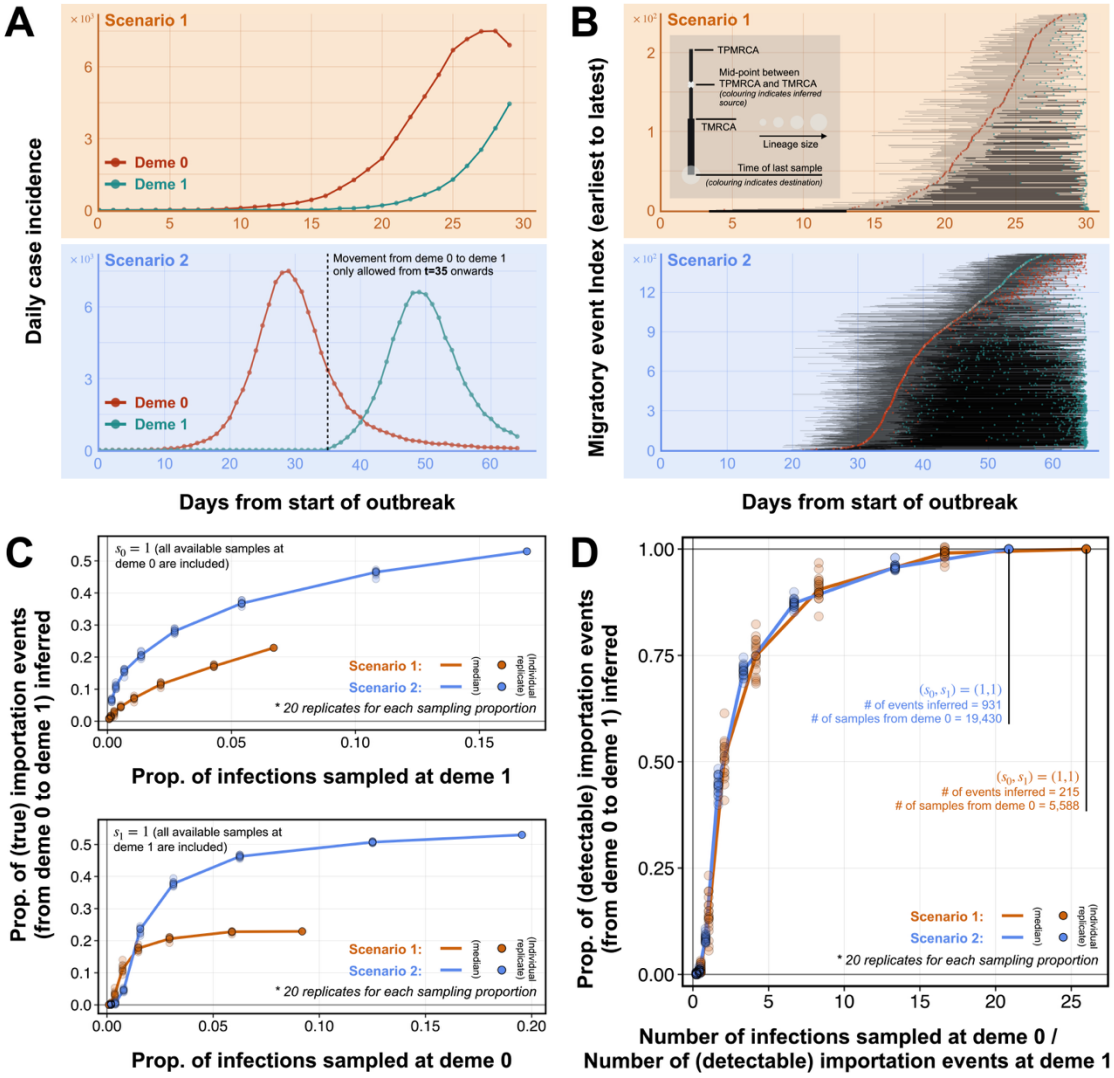


Fig. 5.2. Impact of undersampling on the detection of viral importation (application 1). (A) Daily case incidence in deme 0 (red solid line) and deme 1 (green solid line) under two outbreak scenarios. In scenario 1 (top panel), movement between demes occurs throughout the outbreak; in scenario 2 (bottom panel), movement is restricted until day 35, as indicated by the vertical dashed line. (B) Timeline of migratory events inferred from a phylogeographic inference with $(s_0, s_1) = (1, 1)$, i.e. all available samples at both deme 0 and deme 1 are included. Top and bottom panels correspond to outbreak scenarios 1 and 2, respectively. Each horizontal line represents a single inferred migratory event, with the thinner segment spanning from the Time of the Parent of the Most Recent Common Ancestor (TPMRCA) to the Time of the Most Recent Common Ancestor (TMRCA), and the thicker segment representing the period of sustained local transmission at the destination, from the TMRCA to the time of last sampling. For each event, the cross indicates the mid-point between the TPMRCA and TMRCA, coloured by the inferred origin location; the circle marks the time of last sampling, with radius scaled by lineage size and coloured by the destination location (inferred or observed). (C) (Top) Proportion of true importation events (from deme 0 to deme 1) detected versus proportion of infections sampled at the recipient location (deme 1), with all available samples at the

source location (deme 0) included (i.e. $s_0 = 1$). (Bottom) Proportion of true importation events detected versus proportion of infections sampled at the source location (deme 0), with all available samples at the recipient location (deme 1) included (i.e. $s_1 = 1$). Solid lines indicate the median across 20 replicates for each sampling proportion; individual replicates are plotted as dots. (D) Proportion of detectable importation events (from deme 0 to deme 1) inferred versus the number of infections sampled at the source location (deme 0), normalised by the number of detectable importation events (from deme 0 to deme 1). Solid lines indicate the median across 20 replicates for each sampling proportion; individual replicates are plotted as dots.

5.3.1.3 Results

The top panel in Fig. 5.2C shows that the proportion of true importation events (from deme 0 to deme 1) detected increases with sampling proportion at the recipient location in both outbreak scenarios, although at different rates. Consistent with theoretical predictions from Chapter 4, scenario 2 (where the underlying importation rate is decreasing over time) exhibits a steeper increase in detection probability compared to scenario 1 (where the importation rate is increasing over time). For instance, sampling just 2% of local infections at deme 1 yields a detection probability in scenario 2 that is more than double that in scenario 1. These patterns can be explained by differences in the size distribution of detectable transmission lineages, i.e. lineages inferred when all available samples at both deme 0 and deme 1 are included, or $(s_0, s_1) = (1, 1,)$, as shown in Fig. 5.2B. In scenario 2 (Fig. 5.2B, bottom panel), most transmission lineages introduced into deme 1 (represented by horizontal black lines with green circles at the end) at different times throughout the epidemic persisted and grew to relatively large sizes (as indicated by their longer thick segments and larger green circles). In contrast, in scenario 1 (Fig. 5.2B, top panel), only lineages introduced early on (corresponding to migratory events that occurred before day 25, with indices below 150) during the outbreak persisted. Meanwhile, later-introduced lineages showed relatively little growth or even became extinct (as indicated by the shorter thick segments) soon following their

introduction due to the depletion of susceptibles (Fig. 5.2A, bottom panel). These differences mean that the local epidemic at deme 1 was dominated by mostly small lineages in scenario 1, whereas in scenario 2 larger and more persistent lineages dominated. As a result, a greater proportion of lineages and therefore importation events are detected in scenario 2, even at low sampling proportions.

The bottom panel in Fig. 5.2C shows results from the second set of experiments, where the sampling proportion at the source location (deme 0) is varied. In both outbreak scenarios, increasing the sampling proportion leads to a rapid increase in the proportion of importation events detected, before reaching a plateau where further increases yield diminishing returns. Interestingly, an initial sigmoidal growth (rather than linear growth) is observed under both scenarios, during which doubling the sampling proportion (from $s_0 = 0.01$ to 0.02) results in only marginal increases in the number of importation events detected. This effect arises from the initial substantial overrepresentation of samples from the recipient location ($s_0 = 0.01$ compared to $s_1 = 1$), which causes many importation events to be incorrectly attributed as originating from deme 1 rather than deme 0. Only once the sampling proportion at deme 0 reaches approximately 4% is the correct directionality of viral movement consistently recovered (see Figs. D.2A, D.2B in Appendix D)

Although both outbreak scenarios show a broadly similar pattern in the bottom panel of Fig. 5.2C, the asymptotic values they approach differ substantially. This difference reflects two key factors. First, in scenario 2, the greater proportion of larger and more persistent lineages at the recipient location leads to a higher overall detection probability (53.0% in scenario 2 versus 22.9% in scenario 1 at $(s_0, s_1) = (1, 1)$), as previously discussed. Second, a larger proportion of infections at the source location (deme 0) are sampled in scenario 2 (19.5%) compared to scenario 1 (9.2%), likely

reducing the extent to which independent transmission lineages are aggregated due to undersampling of genetic diversity at the source.

To better illustrate the second factor, Fig. 5.2D presents the same data as in the bottom panel of Fig. 5.2C, but with the x-axis rescaled to show the number of samples included from the source location (deme 0), normalised by the number of importation events detectable from all available samples at the recipient location (deme 1, with $s_1 = 1$). Similarly, the y-axis is rescaled to represent the proportion of these detectable importation events that are successfully recovered in a given inference. Notably, this normalisation causes the two curves to collapse onto a nearly identical trajectory, despite substantial differences in outbreak trajectories and lineage compositions between the two scenarios. This observation suggests that, for a given set of sampled infections at the recipient location - and thus a fixed set of detectable importation events - the proportion of events that can be recovered depends primarily on the number of samples available at the source location per event. For example, it can be inferred from Fig. 5.2D that, if pathogen genomes were sampled from local infections distributed across 100 independent transmission lineages at the recipient location, approximately 1,500 samples from the source location would be needed to recover 90 of the 100 lineages – regardless of their size distribution. However, further investigation is needed to assess whether this relationship generalises across outbreak contexts, including those characterised by different sampling rates and transmission intensities (see Discussion).

5.3.2 Application 2: impact of heterogeneous sampling on source attribution for early viral importation under different sampling schemes

Another common use case of phylogeographic inference is to determine the relative contribution of viral importation from different source locations, particularly during the

early spread of an emerging pathogen or immune-escape variant, as seen in Chapter 2. Such estimates provide critical information for evaluating the effectiveness of control measures aimed at preventing or delaying spatial spread and local establishment. In this application, I assess how different sampling schemes influence the ability to recover the relative proportions of viral importation from multiple sources, under varying degrees of sampling bias.

5.3.2.1 Outbreak scenarios

To explore this, here I consider a network of five demes connected by symmetric human mobility (see Fig. 5.3B). The network structure is designed to resemble a typical real-world transportation system, with a single densely populated location (deme 2, with an initial population of 300,000 individuals) which acts as a travel hub connected to smaller, less populated locations that are only sparsely connected to each other. All demes are initially fully susceptible, except for a single infectious individual in deme 0 (the outbreak origin). I assume the same constant transmission coefficient $\beta = 0.25$ and recovery rate $\gamma = 0.12$ per day in all demes.

As with the first application, I again compare two contrasting outbreak scenarios. In scenario 1, sampling occurs uniformly across space at a constant rate of 0.02 per day in each deme. In scenario 2, sampling is spatially heterogeneous, with deme 1 (outbreak origin) having a higher sampling rate of 0.05 per day (e.g., as a result of heightened genomic surveillance following the detection of a novel pathogen in a neighbouring region (deme 0, the outbreak origin)), while deme 2 (the central travel hub) has a much lower sampling rate of 0.002 per day (e.g., as a result of the local sequencing infrastructure being overwhelmed by high case volumes).

5.3.2.2 Models specifications and experimental setup

For each outbreak scenario, I evaluate the performance of three commonly used sampling schemes: uniform-sample (US), uniform-case (UC), and even (EV) (see Table 1 for more details), across a range of sampling proportions ($s = 0.01$, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64, and 1) which apply equally to each deme; 40 inference replicates are performed under each scheme per sampling proportion to account for sampling stochasticity. For each inference, the distribution of inferred importation events by source location is compared against the corresponding true distribution for each deme, using cosine-distance as a measure of divergence (with lower values indicating more accurate reconstruction). To focus on the early invasion, only importation events inferred to have occurred before the day corresponding to the 10^{th} percentile of true importation times are included in the evaluation. Inferred events with ambiguous origin or destination states are again excluded, as in the first application.

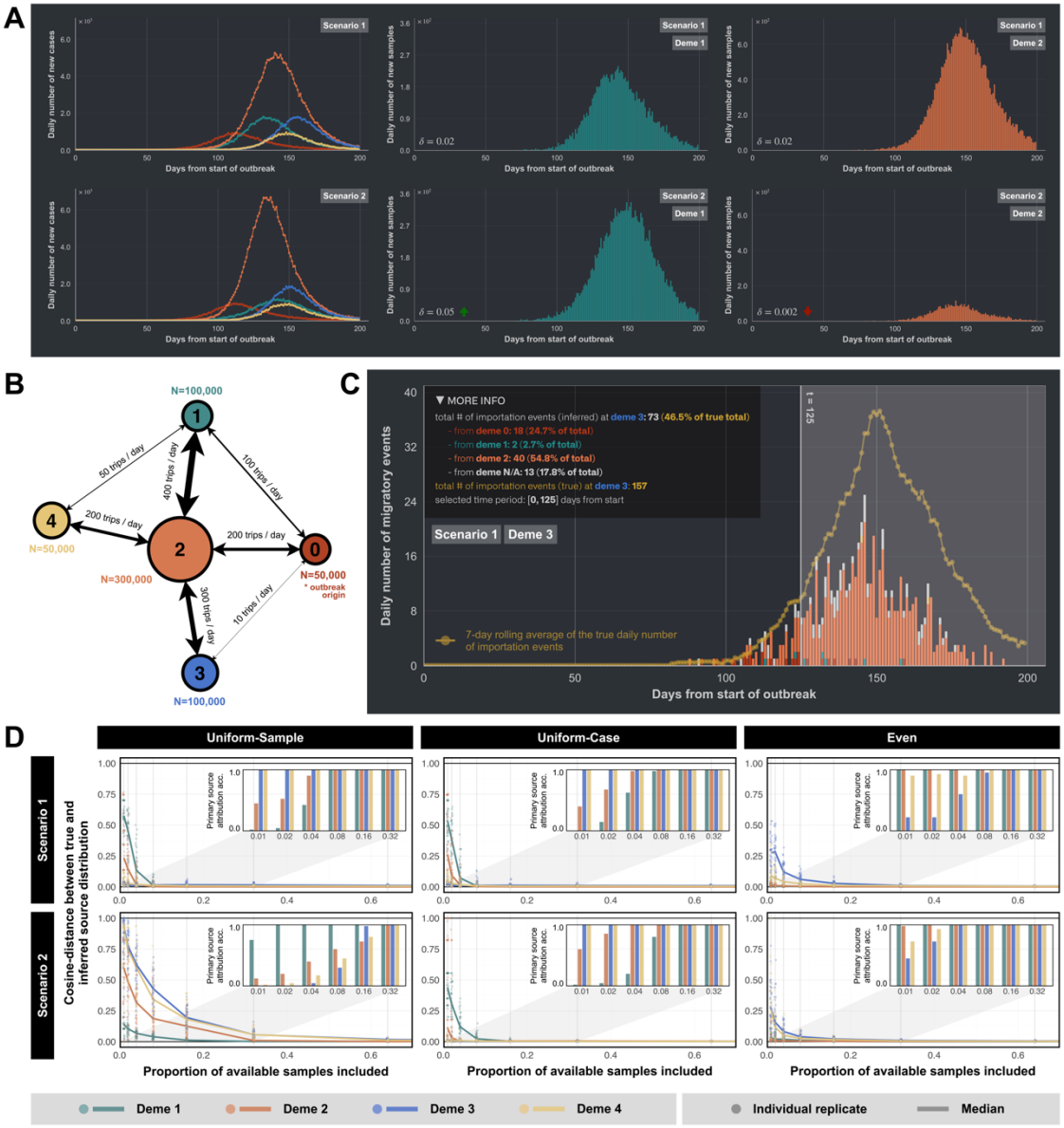


Fig. 5.3. Impact of heterogeneous sampling on source attribution for early viral importation under different sampling schemes (application 2). (A) (Left) Daily case incidence per deme (deme 0: red; deme 1: green; deme 2: orange; deme 3: blue; deme 4: yellow). (Middle and Right) Daily number of samples collected in demes 1 and 2, respectively. Top row corresponds to outbreak scenario 1 which assumes homogeneous sampling at 0.02 per day in each deme; bottom row corresponds to outbreak scenario 2 which assumes heterogeneous sampling, at 0.05 per day for deme 1, 0.002 per day for deme 2, and 0.02 per day in all other demes. All plots are generated by SOPHI, with minor stylistic adjustments for clarity. (B) Mobility network used in Application 2. Node sizes reflect population sizes; edge widths represent the volume of daily movement between connected demes. (C) Data visualisations and summary statistics generated by SOPHI. Solid yellow line shows the 7-day rolling average of true daily number of viral importation events at deme 3. Histogram shows the number of inferred importation events per day, coloured by their inferred source location. Summary statistics are computed

considering only events occurring before day 125 (corresponding to the 10th percentile of true importation times). (D) Cosine-distance between the true and inferred source distributions of viral importation at different sampling proportions (of available samples), under three different sampling schemes (uniform-sample, US; uniform-case, UC; and even, EV) and two outbreak scenarios (scenario 1: top; scenario 2: bottom). Solid lines represent the median across 40 replicates for each sampling proportion, with individual replicates shown as dots (with horizontal jitter added for visual clarity). Insets show the proportion of replicates correctly identifying the primary source of early importation for sampling proportions up to $s = 0.32$. Numerical values for the median cosine-distances are provided in Tables D1-D8 in Appendix D.

5.3.2.3 Results

From Fig. 5.3D, it is observed that cosine-distances generally decrease with increasing sampling proportion, indicating more accurate source attribution as more samples are included regardless of the sampling scheme. There is, however, substantial variability across replicates at low sampling proportion ($s < 0.04$), reflecting the smaller number of importation events that are consistently inferred under sparse sampling.

In outbreak scenario 1 (Fig. 5.3D, top row), uniform-sample (US) and uniform-case (UC) produce nearly identical cosine-distances, as expected given the similar distribution of available samples and reported case incidence under homogeneous sampling (see Fig. 5.3A). Interestingly, the largest cosine-distances under US and UC are observed for demes 1 and 2, especially at low sampling proportions where almost none of the inference replicate correctly identify the primary source of viral importation at $s = 0.01$ and $s = 0.02$. This pattern can be explained by the much higher case numbers and thus larger number of available samples from deme 2, which results in proportionally fewer samples being drawn from smaller demes such as deme 0 (the outbreak origin). With deme 0 being the primary source of viral importation for both demes 1 and 2 (Fig. D.3 in Appendix D), substantial undersampling leads to aggregation of independent transmission lineages introduced from deme 0, causing its contribution to be

underestimated (Figs. D.4, D.6 in Appendix D). This bias diminishes, however, as the overall sampling proportion increases and more samples from deme 0 are included (as previously shown in application 1; also see Figs. D.5, D.7 in Appendix D). While the number of detectable importation events at demes 1 and 2 is expected to also increase with sampling proportion, this likely occurs at a slower rate compared to the rate at which additional samples are included from deme 0, resulting in an overall decrease in the degree of underestimation (equivalent to moving in the positive direction along the x-axis in Fig. 5.2D; see Discussion for further elaboration).

In outbreak scenario 2 (Fig. 5.3D, bottom panel), US now performs notably worse than UC and EV, with consistently larger cosine-distances across all demes except deme 1. This is expected, as the lower sampling rate at deme 2 ($\delta = 0.002$ per day, compared to $\delta = 0.02$ at all other demes and $\delta = 0.05$ at deme 1) leads to underestimation of its contribution, particularly affecting demes 3 and 4, for which deme 2 is the primary source of viral importation (Figs. D.4, D.5 in Appendix D). Interestingly, deme 1 now shows improved performance under scenario 2, as the undersampling at deme 2 results in a more balanced representation of its two primary sources (demes 0 and 2), partially offsetting the underestimation of contribution from deme 0, as observed in scenario 1. These effects are largely mitigated under UC, which adjusts for heterogeneous sampling by downsampling in proportion to reported case numbers. Meanwhile, performance under EV appears mostly unaffected by the sampling bias, with most demes showing cosine-distances that are comparable to those under UC.

5.4 Discussion

It is well known that heterogeneous sampling of pathogen genomes can lead to biased phylogeographic estimates, yet efforts to characterise these effects systematically and to

develop corresponding mitigation strategies have been limited. This is due in part to the challenge of jointly modelling the underlying transmission dynamics, migration patterns, and the sampling process - as well as the complexity of sampling design, which often requires a case-by-case approach with consideration of both the specific outbreak context and research objectives. In this chapter, I explored the impact of heterogeneous sampling on phylogeographic inference by developing a new simulation-based evaluation framework and applying it to two research questions inspired from early work presented in this thesis.

In the first application, I investigated how the undersampling of infections in a 2-deme mobility network leads to different degrees of underestimation of the number of viral importation events. By considering two contrasting outbreak scenarios, I showed that the detection probability increases with sampling proportion at the recipient location (for a given sampling proportion at the outbreak origin), and that the rate of increase depends on whether the underlying migration rate is increasing or decreasing. This finding is consistent with the theoretical predictions derived in Chapter 4, despite two different key assumptions: 1) sampling-through-time, as opposed to the sampling of all infected individuals at the time of inference, and 2) incomplete sampling of infections at the source location. In a second experiment, I also demonstrated that undersampling at the source location results in the aggregation of independent transmission lineages, leading to further underestimation of the number of importation events. This effect diminishes rapidly with increasing sampling proportion, however, with the rate of decay determined primarily by the number of samples available at the source location per detectable importation event, given the sampled infections at the recipient location. This observation reveals an important broader insight into how heterogeneous sampling gives rise to biased estimates of migration rates: for a given outbreak scenario, the number of

migratory events detected between two locations depends on the combined effect of two factors: 1) the number of events that are detectable given the sampled infections at the recipient location, and 2) the number of sampled infections at the source location per detectable event. While increasing the overall sampling proportions generally leads to an increase in the number of migratory events detected, the extent of this increase is not uniform across all deme pairs. Rather, it depends on the relative rates of change of these two factors - i.e. the number of new migratory events that become detectable with additional sampling at the recipient location, and the number of additional samples drawn from the source location.

In the second application, I investigated how different sampling schemes influence the ability to correctly identify the source of early importation events. Specifically, I compared three commonly used sampling strategies (uniform-sample, US; uniform-case, UC; and even, EV) under two outbreak scenarios: one with homogeneous sampling across demes and one with substantial spatial sampling heterogeneities. Under homogeneous sampling, all three schemes performed similarly, with differences largely driven by the number of samples included from key sources of viral importation at low sampling proportions. These differences reflect the degree to which independent transmission lineages at the recipient locations are aggregated, which is in turn determined by the number of sampled infections included at the source location per detectable migratory event (as discussed above). This effect diminishes rapidly, however, resulting in overall more accurate source attribution as sampling proportion increases.

Under heterogeneous sampling, US consistently underperformed, particularly for demes whose primary source of viral importation was undersampled. This bias was largely mitigated under UC, where unequal sampling rates were adjusted for by downsampling in proportion to reported case numbers. Meanwhile, EV showed

comparable performance despite not explicitly adjusting for uneven sampling - providing a practical alternative to UC, especially in outbreak scenarios where reliable case data are unavailable. Beyond sampling design for phylogeographic inference, these results also have broader implications for the coordination of sequencing efforts in genomic surveillance. During the COVID-19 pandemic, for example, only a fraction of samples collected from infected individuals are typically sequenced due to both resource and logistical constraints. Depending on the epidemiological context, these samples were either selected uniformly at random (e.g., a fixed proportion of sample per location and per week for general situational awareness), or dynamically in response to changes in case incidence (e.g., prioritising locations exhibiting rapid local growth to accelerate variant detection) (53-55). Importantly, these practical choices determined the set of genome samples available for downstream analyses, which in turn influenced how samples should be further selected to maximise the accuracy of phylogeographic inference, as explored in this chapter. As such, the design of future genomic surveillance strategies must carefully balance the needs of short-term outbreak analytics and longer-term or retrospective investigations using approaches such as phylogeography, while remaining responsive to shifting priorities over the course of an epidemic.

Several limitations of the above findings should be noted. First, further work is needed to assess the extent to which these results generalise to a broader range of epidemiological contexts and outbreak conditions. For example, although the first application produced results that are consistent with theoretical predictions from Chapter 4, the impact of different volumes of daily movement (beyond just temporal trends, as considered here and in Chapter 4) remains unexplored. Similarly, the finding that the aggregation of independent transmission lineages depends primarily on the number of available samples at the source location per detectable event warrants further

investigation, particularly under outbreak scenarios with different sampling rates or local transmission intensities. Second, both investigations presented in this chapter relied on parsimony-based discrete trait analysis, which was selected for its computational efficiency given the need to perform multiple replicate inferences to account for sampling stochasticity. While this enabled extensive comparative evaluations, the robustness of key findings under alternative inference approaches, such as maximum-likelihood or Bayesian approaches, should be evaluated. Third, in the second application, the timing of inferred importations event was approximated using the mid-point between its corresponding TMRCA and TPMRCA. As a result, whether a given event was classified as "early" (i.e. occurring before the 10^{th} percentile of true importation events) could vary depending on the proxy used (e.g., using TMRCA or TPMRCA directly instead of the mid-point). This potential bias is particularly relevant at low sampling proportions, where the difference between TMRCA and TPMRCA can be large; future work should explore more principled approaches for estimating the timing of importation given a detected transmission lineage.

Beyond these caveats, there are several broader limitations in the current design and implementation of the SOPHI framework. First, SOPHI emulates the process of tree estimation by pruning the full simulated tree generated by ReMASTER (43). While this approach offers greater computational efficiency, it does not account for uncertainties or biases associated with phylogenetic tree inference - such as those resulting from uneven sequencing coverage, limited genetic diversity, or recombination. These factors have been shown to introduce bias in tree topology and evolutionary rates estimates, which are critical for accurate and robust phylogeographic inference. Although Bayesian methods (e.g., BEAST (32, 33)) explicitly incorporate such uncertainties, they are computationally intensive and time-consuming. In light of this trade-off, SOPHI is best viewed as a tool

for rapid prototyping and hypothesis generation, before downstream validation using more rigorous, state-of-the-art inference approaches. Second, SOPHI currently operates on partially observed trees derived only from sampled infections. While ReMASTER can, in principle, simulate fully sampled trees (e.g., by forcing every infection to be sampled and sequenced upon recovery), this quickly becomes computationally intractable for large-scale outbreaks ($\gtrsim 100{,}000$ infections) with multiple demes. As a result, it is not currently possible to evaluate certain key metrics - such as the true size of a detected transmission lineage - which is relevant for understanding the epidemiological impact of viral importation. Future extensions could address this by generating outbreak scenarios using agent-based models or hybrid frameworks such as GLEAM (56). Finally, although SOPHI allows the design of a wide range of sampling strategies, it does not yet support the integration of mobility data and phylogeographic estimates from previous inferences. Future versions will introduce features allowing researchers to design their own sampling schemes programmatically, with full access to underlying datasets including reported case numbers, mobility matrices, and annotated trees (e.g., as Newick strings) from preliminary inferences.

While this chapter has introduced SOPHI primarily as a practical framework for systematically evaluating different sampling strategies under different sampling biases, its potential applications extend beyond manual experimentation. In particular, SOPHI's ability to provide real-time feedback on inference performance in a controlled, virtual environment makes it a natural testbed for machine learning approaches. Optimisation methods such as Bayesian optimisation, active learning, or reinforcement learning could be employed to guide the exploration of different sampling strategies and design choices. By simulating a wide range of outbreaks with diverse transmission conditions, sampling heterogeneities, and mobility network structures, SOPHI can serve as a training

environment to support the development of context-aware, generalised approaches to genomic sampling - ultimately enabling more accurate and robust phylogeographic reconstructions of real-world infectious disease outbreaks.

5.5 References

1. Dudas, G., Carvalho, L.M., Bedford, T., Tatem, A.J., Baele, G., Faria, N.R., Park, D.J., Ladner, J.T., Arias, A., Asogun, D., Bielejec, F., Caddy, S.L., Cotten, M., D'Ambrozio, J., Dellicour, S., Di Caro, A., Diclaro, J.W., Duraffour, S., Elmore, M.J., Fakoli, L.S., Faye, O., Gilbert, M.L., Gevao, S.M., Gire, S., Gladden-Young, A., Gnirke, A., Goba, A., Grant, D.S., Haagmans, B.L., Hiscox, J.A., Jah, U., Kugelman, J.R., Liu, D., Lu, J., Malboeuf, C.M., Mate, S., Matthews, D.A., Matranga, C.B., Meredith, L.W., Qu, J., Quick, J., Pas, S.D., Phan, M.V.T., Pollakis, G., Reusken, C.B., Sanchez-Lockhart, M., Schaffner, S.F., Schieffelin, J.S., Sealfon, R.S., Simon-Loriere, E., Smits, S.L., Stoecker, K., Thorne, L., Tobin, E.A., Vandi, M.A., Watson, S.J., West, K., Whitmer, S., Wiley, M.R., Winnicki, S.M., Wohl, S., Wölfel, R., Yozwiak, N.L., Andersen, K.G., Blyden, S.O., Bolay, F., Carroll, M.W., Dahn, B., Diallo, B., Formenty, P., Fraser, C., Gao, G.F., Garry, R.F., Goodfellow, I., Günther, S., Happi, C.T., Holmes, E.C., Kargbo, B., Keïta, S., Kellam, P., Koopmans, M.P.G., Kuhn, J.H., Loman, N.J., Magassouba, N. 'faly, Naidoo, D., Nichol, S.T., Nyenswah, T., Palacios, G., Pybus, O.G., Sabeti, P.C., Sall, A., Ströher, U., Wurie, I., Suchard, M.A., Lemey, P. and Rambaut, A. (2017) 'Virus genomes reveal factors that spread and sustained the Ebola epidemic', *Nature*, 544(7650), pp. 309–315.
2. Dellicour, S., Baele, G., Dudas, G., Faria, N.R., Pybus, O.G., Suchard, M.A., Rambaut, A. and Lemey, P. (2018) 'Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak', *Nature Communications*, 9(1), pp. 1–9.
3. Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., Russell, C.A., Smith, D.J., Pybus, O.G., Brockmann, D. and Suchard, M.A. (2014) 'Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2', *PLOS Pathogens*, 10(2), p. e1003932.
4. Russell, C.A., Jones, T.C., Barr, I.G., Cox, N.J., Garten, R.J., Gregory, V., Gust, I.D., Hampson, A.W., Hay, A.J., Hurt, A.C., de Jong, J.C., Kelso, A., Klimov, A.I., Kageyama, T., Komadina, N., Lapedes, A.S., Lin, Y.P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A.D., Rimmelzwaan, G.F., Shaw, M.W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R.A. and Smith, D.J. (2008) 'The global circulation of seasonal influenza A (H3N2) viruses', *Science*, 320(5874), pp. 340-346.
5. Rambaut, A., Pybus, O.G., Nelson, M.I., Viboud, C., Taubenberger, J.K. and Holmes, E.C. (2008) 'The genomic and epidemiological dynamics of human influenza A virus', *Nature*, 453(7195), pp. 615–619.
6. Bedford, T., Riley, S., Barr, I.G., Broor, S., Chadha, M., Cox, N.J., Daniels, R.S., Gunasekaran, C.P., Hurt, A.C., Kelso, A., Klimov, A., Lewis, N.S., Li, X., McCauley, J.W., Odagiri, T., Potdar, V., Rambaut, A., Shu, Y., Skepner, E., Smith, D.J., Suchard, M.A., Tashiro, M., Wang, D., Xu, X., Lemey, P. and Russell, C.A. (2015) 'Global circulation patterns of seasonal influenza viruses vary with antigenic drift', *Nature*, 523(7559), pp. 217–220.
7. Faye, O., Freire, C.C.M., Iamarino, A., Faye, O., de Oliveira, J.V.C., Diallo, M., Zanotto, P.M.A. and Sall, A.A. (2014) 'Molecular Evolution of Zika Virus during Its Emergence in the 20th Century', *PLOS Neglected Tropical Diseases*, 8(1), p. e2636.
8. Ebranati, E., Veo, C., Carta, V., Percivalle, E., Rovida, F., Frati, E.R., Amendola, A., Ciccozzi, M., Tanzi, E., Galli, M., Baldanti, F. and Zehender, G. (2019) 'Time-scaled phylogeography of complete Zika virus genomes using discrete and continuous space diffusion models', *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 73, pp. 33-43.

9. Tegally, H., Wilkinson, E., Tsui, J.L., Moir, M., Martin, D., Brito, A.F., Giovanetti, M., Khan, K., Huber, C., Bogoch, I.I., San, J.E., Poongavanan, J., Xavier, J.S., Candido, D.D.S., Romero, F., Baxter, C., Pybus, O.G., Lessells, R.J., Faria, N.R., Kraemer, M.U.G. and de Oliveira, T. (2023) 'Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern', *Cell*, 186(15), pp. 3277-3290.e16
10. du Plessis, L., McCrone, J.T., Zarebski, A.E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T.R., Faria, N.R., Jackson, B., Loman, N.J., O'Toole, Á., Nicholls, S.M., Parag, K.V., Scher, E., Vasylyeva, T.I., Volz, E.M., Watts, A., Bogoch, I.I., Khan, K., COVID-19 Genomics UK (COG-UK) Consortium, Aanensen, D.M., Kraemer, M.U.G., Rambaut, A. and Pybus, O.G. (2021) 'Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK', *Science*, 371(6530), pp. 708–712.
11. Lemey, P., Ruktanonchai, N., Hong, S.L., Colizza, V., Poletto, C., Van den Broeck, F., Gill, M.S., Ji, X., Levasseur, A., Oude Munnink, B.B., Koopmans, M., Sadilek, A., Lai, S., Tatem, A.J., Baele, G., Suchard, M.A. and Dellicour, S. (2021) 'Untangling introductions and persistence in COVID-19 resurgence in Europe', *Nature*, 595(7869), pp. 713–717.
12. Aggarwal, D., Warne, B., Jahun, A.S., Hamilton, W.L., Fieldman, T., du Plessis, L., Hill, V., Blane, B., Watkins, E., Wright, E., Hall, G., Ludden, C., Myers, R., Hosmillo, M., Chaudhry, Y., Pinckert, M.L., Georgana, I., Izuagbe, R., Leek, D., Nsonwu, O., Hughes, G.J., Packer, S., Page, A.J., Metaxaki, M., Fuller, S., Weale, G., Holgate, J., Brown, C.A., Howes, R., McFarlane, D., Dougan, G., Pybus, O.G., Angelis, D.D., Maxwell, P.H., Peacock, S.J., Weekes, M.P., Illingworth, C., Harrison, E.M., Matheson, N.J. and Goodfellow, I.G. (2022) 'Genomic epidemiology of SARS-CoV-2 in a UK university identifies dynamics of transmission', *Nature Communications*, 13(1), pp. 1–16.
13. Wilkinson, E., Giovanetti, M., Tegally, H., San, J.E., Lessells, R., Cuadros, D., Martin, D.P., Rasmussen, D.A., Zekri, A.-R.N., Sangare, A.K., Ouedraogo, A.-S., Sesay, A.K., Priscilla, A., Kemi, A.-S., Olubusuyi, A.M., Oluwapelumi, A.O.O., Hammami, A., Amuri, A.A., Sayed, A., Ouma, A.E.O., Elargoubi, A., Ajayi, N.A., Victoria, A.F., Kazeem, A., George, A., Trotter, A.J., Yahaya, A.A., Keita, A.K., Diallo, A., Kone, A., Souissi, A., Chtourou, A., Gutierrez, A.V., Page, A.J., Vinze, A., Iranzadeh, A., Lambisia, A., Ismail, A., Rosemary, A., Sylverken, A., Femi, A., Ibrahimi, A., Marycelin, B., Oderinde, B.S., Bolajoko, B., Dhaala, B., Herring, B.L., Njanpop-Lafourcade, B.-M., Kleinhans, B., McInnis, B., Tegomoh, B., Brook, C., Pratt, C.B., Scheepers, C., Akoua-Koffi, C.G., Agoti, C.N., Peyrefitte, C., Daubenberger, C., Morang'a, C.M., Nokes, D.J., Amoako, D.G., Bugembe, D.L., Park, D., Baker, D., Doolabh, D., Ssemwanga, D., Tshiabuila, D., Bassirou, D., Amuzu, D.S.Y., Goedhals, D., Omuoyo, D.O., Maruapula, D., Foster-Nyarko, E., Lusamaki, E.K., Simulundu, E., Ong'era, E.M., Ngabana, E.N., Shumba, E., El Fahime, E., Lokilo, E., Mukantwari, E., Philomena, E., Belarbi, E., Simon-Loriere, E., Anoh, E.A., Leendertz, F., Ajili, F., Enoch, F.O., Wasfi, F., Abdelmoula, F., Mosha, F.S., Takawira, F.T., Derrar, F., Bouzid, F., Onikepe, F., Adeola, F., Muyembe, F.M., Tanser, F., Dratibi, F.A., Mbunsu, G.K., Thilliez, G., Kay, G.L., Githinji, G., van Zyl, G., Awandare, G.A., Schubert, G., Maphalala, G.P., Ranaivoson, H.C., Lemriss, H., Anise, H., Abe, H., Karray, H.H., Nansumba, H., Elgahzaly, H.A., Gumbo, H., Smeti, I., Ayed, I.B., Odia, I., Ben Boubaker, I.B., Gaaloul, I., Gazy, I., Mudau, I., Ssewanyana, I., Konstantinus, I., Lekana-Douk, J.B., Makangara, J.-C.C., Tamfum, J.-J.M., Heraud, J.-M., Shaffer, J.G., Giandhari, J., Li,

J., Yasuda, J., Mends, J.Q., Kiconco, J., Morobe, J.M., Gyapong, J.O., Okolie, J.C., Kayiwa, J.T., Edwards, J.A., Gyamfi, J., Farah, J., Nakaseegu, J., Ngoi, J.M., Namulondo, J., Andeko, J.C., Lutwama, J.J., O'Grady, J., Siddle, K., Adeyemi, K.T., Tumedi, K.A., Said, K.M., Hae-Young, K., Duedu, K.O., Belyamani, L., Fki-Berrajah, L., Singh, L., Martins, L. de O., Tyers, L., Ramuth, M., Mastouri, M., Aouni, M., El Hefnawi, M., Matsheka, M.I., Kebabonye, M., Diop, M., Turki, M., Paye, M., Nyaga, M.M., Mareka, M., Damaris, M.-M., Mburu, M.W., Mpina, M., Nwando, M., Owusu, M., Wiley, M.R., Youtchou, M.T., Ayekaba, M.O., Abouelhoda, M., Seadawy, M.G., Khalifa, M.K., Sekhele, M., Ouadghiri, M., Diagne, M.M., Mwenda, M., Allam, M., Phan, M.V.T., Abid, N., Touil, N., Rujeni, N., Kharrat, N., Ismael, N., Dia, N., Mabunda, N., Hsiao, N.-Y., Silochi, N.B., Nsenga, N., Gumede, N., Mulder, N., Ndodo, N., Razanajatovo, N.H., Iguosadolo, N., Judith, O., Kingsley, O.C., Sylvanus, O., Peter, O., Femi, O., Idowu, O., Testimony, O., Chukwuma, O.E., Ogah, O.E., Onwuamah, C.K., Cyril, O., Faye, O., Tomori, O., Ondoa, P., Combe, P., Semanda, P., Oluniyi, P.E., Arnaldo, P., Quashie, P.K., Dussart, P., Bester, P.A., Mbala, P.K., Ayivor-Djanie, R., Njouom, R., Phillips, R.O., Gorman, R., Kingsley, R.A., Carr, R.A.A., El Kabbaj, S., Gargouri, S., Masmoudi, S., Sankhe, S., Lawal, S.B., Kassim, S., Trabelsi, S., Metha, S., Kammoun, S., Lemriss, S., Agwa, S.H.A., Calvignac-Spencer, S., Schaffner, S.F., Doumbia, S., Mandanda, S.M., Aryeetey, S., Ahmed, S.S., Elhamoumi, S., Andriamandimby, S., Tope, S., Lekana-Douki, S., Prosolek, S., Ouangraoua, S., Mundeke, S.A., Rudder, S., Panji, S., Pillay, S., Engelbrecht, S., Nabadda, S., Behillil, S., Budiaki, S.L., van der Werf, S., Mashe, T., Aanniz, T., Mohale, T., Le-Viet, T., Schindler, T., Anyaneji, U.J., Chinedu, U., Ramphal, U., Jessica, U., George, U., Fonseca, V., Enouf, V., Gorova, V., Roshdy, W.H., Ampofo, W.K., Preiser, W., Choga, W.T., Bediako, Y., Naidoo, Y., Butera, Y., de Laurent, Z.R., Sall, A.A., Rebai, A., von Gottberg, A., Kouriba, B., Williamson, C., Bridges, D.J., Chikwe, I., Bhiman, J.N., Mine, M., Cotten, M., Moyo, S., Gaseitsiwe, S., Saasa, N., Sabeti, P.C., Kaleebu, P., Tebeje, Y.K., Tessema, S.K., Happi, C., Nkengasong, J. and de Oliveira, T. (2021) 'A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa', *Science*, 374(6566), pp. 423–431.

14. Paredes, M.I., Perofsky, A.C., Frisbie, L., Moncla, L.H., Roychoudhury, P., Xie, H., Mohamed Bakhash, S.A., Kong, K., Arnould, I., Nguyen, T.V., Wendm, S.T., Hajian, P., Ellis, S., Mathias, P.C., Greninger, A.L., Starita, L.M., Frazar, C.D., Ryke, E., Zhong, W., Gamboa, L., Threlkeld, M., Lee, J., Stone, J., McDermot, E., Truong, M., Shendure, J., Oltean, H.N., Viboud, C., Chu, H., Müller, N.F. and Bedford, T. (2024) 'Local-scale phylodynamics reveal differential community impact of SARS-CoV-2 in a metropolitan US county', *PLOS Pathogens*, 20(3), p. e1012117.
15. Tsui, J.L.-H., McCrone, J.T., Lambert, B., Bajaj, S., Inward, R.P.D., Bosetti, P., Pena, R.E., Tegally, H., Hill, V., Zarebski, A.E., Peacock, T.P., Liu, L., Wu, N., Davis, M., Bogoch, I.I., Khan, K., Kall, M., Abdul Aziz, N.I.B., Colquhoun, R., O'Toole, Á., Jackson, B., Dasgupta, A., Wilkinson, E., de Oliveira, T., COVID-19 Genomics UK (COG-UK) consortium¶, Connor, T.R., Loman, N.J., Colizza, V., Fraser, C., Volz, E., Ji, X., Gutierrez, B., Chand, M., Dellicour, S., Cauchemez, S., Raghwani, J., Suchard, M.A., Lemey, P., Rambaut, A., Pybus, O.G. and Kraemer, M.U.G. (2023) 'Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1', *Science*, 381(6655), pp. 336–343.
16. Lemey, P., Rambaut, A., Welch, J.J. and Suchard, M.A. (2010) 'Phylogeography takes a relaxed random walk in continuous space and time', *Molecular biology and evolution*, 27(8), pp. 1877-1885.

17. Lemey, P., Rambaut, A., Drummond, A.J. and Suchard, M.A. (2009) 'Bayesian Phylogeography Finds Its Roots', *PLOS Computational Biology*, 5(9), p. e1000520.
18. Kraemer, M.U.G., Faria, N.R., Reiner, R.C., Jr, Golding, N., Nikolay, B., Stasse, S., Johansson, M.A., Salje, H., Faye, O., Wint, G.R.W., Niedrig, M., Shearer, F.M., Hill, S.C., Thompson, R.N., Bisanzio, D., Taveira, N., Nax, H.H., Pradelski, B.S.R., Nsoesie, E.O., Murphy, N.R., Bogoch, I.I., Khan, K., Brownstein, J.S., Tatem, A.J., de Oliveira, T., Smith, D.L., Sall, A.A., Pybus, O.G., Hay, S.I. and Cauchemez, S. (2017) 'Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015-16: a modelling study', *The Lancet. Infectious diseases*, 17(3), pp. 330–338.
19. Faria, N.R., Kraemer, M.U.G., Hill, S.C., de Jesus, J.G., de Aguiar, R.S., Iani, F.C.M., Xavier, J., Quick, J., du Plessis, L., Dellicour, S., Thézé, J., Carvalho, R.D.O., Baele, G., Wu, C.-H., Silveira, P.P., Arruda, M.B., Pereira, M.A., Pereira, G.C., Lourenço, J., Obolski, U., Abade, L., Vasylyeva, T.I., Giovanetti, M., Yi, D., Weiss, D.J., Wint, G.R.W., Shearer, F.M., Funk, S., Nikolai, B., Fonseca, V., Adelino, T.E.R., Oliveira, M.A.A., Silva, M.V.F., Sacchetto, L., Figueiredo, P.O., Rezende, I.M., Mello, E.M., Said, R.F.C., Santos, D.A., Ferraz, M.L., Brito, M.G., Santana, L.F., Menezes, M.T., Brindeiro, R.M., Tanuri, A., dos Santos, F.C.P., Cunha, M.S., Nogueira, J.S., I, M Rocco, da Costa, A.C., Komninakis, S.C.V., Azevedo, V., Chieppe, A.O., Araujo, E.S.M., Mendonça, M.C.L., dos Santos, C.C., dos Santos, C.D., Mares-Guia, A.M., Nogueira, R.M.R., Sequeira, P.C., Abreu, R.G., Garcia, M.H.O., Abreu, A.L., Okumoto, O., Kroon, E.G., de Albuquerque, C.F.C., Lewandowski, K., Pullan, S.T., Carroll, M., de Oliveira, T., Sabino, E.C., Souza, R.P., Suchard, M.A., Lemey, P., Trindade, G.S., Drumond, B.P., Filippis, A.M.B., Loman, N.J., Cauchemez, S., Alcantara, L.C.J. and Pybus, O.G. (2018) 'Genomic and epidemiological monitoring of yellow fever virus transmission potential', *Science*, 361(6405), p. 894.
20. Brockmann, D. and Helbing, D. (2013) 'The Hidden Geometry of Complex, Network-Driven Contagion Phenomena', *Science*, 342(6164), pp. 1337-1342.
21. Nakano, T., Lu, L., Liu, P. and Pybus, O.G. (2004) 'Viral gene sequences reveal the variable history of hepatitis C virus infection among countries', *The Journal of infectious diseases*, 190(6), pp. 1098-1108.
22. Slatkin, M. and Maddison, W.P. (1989) 'A cladistic measure of gene flow inferred from the phylogenies of alleles', *Genetics*, 123(3), pp. 603–613.
23. Swofford, D.L. (1999) *PAUP 4.0: Phylogenetic Analysis Using Parsimony (And Other Methods): Software Beta Version and User's Manual for UNIX and VMS*. Sinauer Associates Incorporated.
24. Wallace, R.G., HoDac, H., Lathrop, R.H. and Fitch, W.M. (2007) 'A statistical phylogeography of influenza A H5N1', *Proceedings of the National Academy of Sciences*, 104(11), pp. 4473–4478.
25. Kühnert, D., Wu, C.H. and Drummond, A.J. (2011) 'Phylogenetic and epidemic modeling of rapidly evolving infectious diseases', *Infection, Genetics and Evolution*, 11(8), pp. 1825–1841.
26. Felsenstein, J. (1978) 'Cases in which Parsimony or Compatibility Methods will be Positively Misleading', *Systematic Biology*, 27(4), pp. 401–410.
27. Cunningham, C.W., Omland, K.E. and Oakley, T.H. (1998) 'Reconstructing ancestral character states: a critical reappraisal', *Trends in Ecology & Evolution*, 13(9), pp. 361–366.
28. Ronquist, F. (2004) Bayesian inference of character evolution. *Trends in ecology & evolution*, 19(9), pp. 475-481.

29. De Maio, N., Wu, C.-H., O'Reilly, K.M. and Wilson, D. (2015) 'New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation', *PLOS Genetics*, 11(8), p. e1005421.
30. Müller, N.F., Rasmussen, D.A. and Stadler, T. (2017) 'The Structured Coalescent and Its Approximations', *Molecular biology and evolution*, 34(11), pp. 2970-2981.
31. Vaughan, T.G., Kühnert, D., Popinga, A., Welch, D. and Drummond, A.J. (2014) 'Efficient Bayesian inference under the structured coalescent', *Bioinformatics (Oxford, England)*, 30(16), pp. 2272–2279.
32. Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J. and Rambaut, A. (2018) 'Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10', *Virus evolution*, 4(1), p. vey016.
33. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M.A., Rambaut, A. and Drummond, A.J. (2014) 'BEAST 2: a software platform for Bayesian evolutionary analysis', *PLoS computational biology*, 10(4), p. e1003537.
34. Müller, N.F., Rasmussen, D. and Stadler, T. (2018) 'MASCOT: parameter and state inference under the marginal structured coalescent approximation', *Bioinformatics (Oxford, England)*, 34(22), pp. 3843–3848.
35. Layan, M., Müller, N.F., Dellicour, S., De Maio, N., Bourhy, H., Cauchemez, S. and Baele, G. (2023) 'Impact and mitigation of sampling bias to determine viral spread: Evaluating discrete phylogeography through CTMC modeling and structured coalescent model approximations', *Virus evolution*, 9(1), p. vead010.
36. Liu, P., Song, Y., Colijn, C. and MacPherson, A. (2022) 'The impact of sampling bias on viral phylogeographic reconstruction', *PLOS Global Public Health*, 2(9), p. e0000577.
37. Gámbaro, F., Layan, M., Baele, G., Vrancken, B. and Dellicour, S. (2025) 'Navigating sampling bias in discrete phylogeographic analysis: assessing the performance of an adjusted Bayes factor', *bioRxiv*. Available at: https://doi.org/10.1101/2025.04.23.650183 (Accessed: 1 May 2025).
38. Lemey, P., Hong, S.L., Hill, V., Baele, G., Poletto, C., Colizza, V., O'Toole, Á., McCrone, J.T., Andersen, K.G., Worobey, M., Nelson, M.I., Rambaut, A. and Suchard, M.A. (2020) 'Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2', *Nature Communications*, 11(1), pp. 1–14.
39. Brito, A.F., Semenova, E., Dudas, G., Hassler, G.W., Kalinich, C.C., Kraemer, M.U.G., Ho, J., Tegally, H., Githinji, G., Agoti, C.N., Matkin, L.E., Whittaker, C., Howden, B.P., Sintchenko, V., Zuckerman, N.S., Mor, O., Blankenship, H.M., de Oliveira, T., Lin, R.T.P., Siqueira, M.M., Resende, P.C., Vasconcelos, A.T.R., Spilki, F.R., Aguiar, R.S., Alexiev, I., Ivanov, I.N., Philipova, I., Carrington, C.V.F., Sahadeo, N.S.D., Branda, B., Gurry, C., Maurer-Stroh, S., Naidoo, D., von Eije, K.J., Perkins, M.D., van Kerkhove, M., Hill, S.C., Sabino, E.C., Pybus, O.G., Dye, C., Bhatt, S., Flaxman, S., Suchard, M.A., Grubaugh, N.D., Baele, G. and Faria, N.R. (2022) 'Global disparities in SARS-CoV-2 genomic surveillance', *Nature Communications*, 13(1), pp. 1–13.
40. Worobey, M., Pekar, J., Larsen, B.B., Nelson, M.I., Hill, V., Joy, J.B., Rambaut, A., Suchard, M.A., Wertheim, J.O. and Lemey, P. (2020) 'The emergence of SARS-CoV-2 in Europe and North America', *Science*, 370(6516), pp. 564-570.
41. McCrone, J.T., Hill, V., Bajaj, S., Pena, R.E., Lambert, B.C., Inward, R., Bhatt, S., Volz, E., Ruis, C., Dellicour, S., Baele, G., Zarebski, A.E., Sadilek, A., Wu, N., Schneider, A., Ji, X., Raghwani, J., Jackson, B., Colquhoun, R., O'Toole, Á., Peacock, T.P., Twohig, K., Thelwall, S., Dabrera, G., Myers, R., Faria, N.R., Huber,

C., Bogoch, I.I., Khan, K., du Plessis, L., Barrett, J.C., Aanensen, D.M., Barclay, W.S., Chand, M., Connor, T., Loman, N.J., Suchard, M.A., Pybus, O.G., Rambaut, A. and Kraemer, M.U.G. (2022) 'Context-specific emergence and growth of the SARS-CoV-2 Delta variant', *Nature*, 610(7930), pp. 154–160.

42. Castelán-Sánchez, H.G., Delaye, L., Inward, R.P.D., Dellicour, S., Gutierrez, B., Martinez de la Vina, N., Boukadida, C., Pybus, O.G., de Anda Jáuregui, G., Guzmán, P., Flores-Garrido, M., Fontanelli, Ó., Hernández Rosales, M., Meneses, A., Olmedo-Alvarez, G., Herrera-Estrella, A.H., Sánchez-Flores, A., Muñoz-Medina, J.E., Comas-García, A., Gómez-Gil, B., Zárate, S., Taboada, B., López, S., Arias, C.F., Kraemer, M.U.G., Lazcano, A. and Escalera Zamudio, M. (2023) 'Comparing the evolutionary dynamics of predominant SARS-CoV-2 virus lineages co-circulating in Mexico', *eLife*, 12, e82069.
43. Vaughan, T.G. (2024) 'ReMASTER: improved phylodynamic simulation for BEAST 2.7', *Bioinformatics (Oxford, England)*, 40(1), btae015.
44. Stadler, T. (2010) Sampling-through-time in birth–death trees. *Journal of theoretical biology*, 267(3), pp. 396-404.
45. Giovanetti, M., Slavov, S.N., Fonseca, V., Wilkinson, E., Tegally, H., Patané, J.S.L., Viala, V.L., San, E.J., Rodrigues, E.S., Santos, E.V., Aburjaile, F., Xavier, J., Fritsch, H., Adelino, T.E.R., Pereira, F., Leal, A., Iani, F.C. de M., de Carvalho Pereira, G., Vazquez, C., Sanabria, G.M.E., Oliveira, E.C. de, Demarchi, L., Croda, J., dos Santos Bezerra, R., Paola Oliveira de Lima, L., Martins, A.J., Renata dos Santos Barros, C., Marqueze, E.C., de Souza Todao Bernardino, J., Moretti, D.B., Brassaloti, R.A., de Lello Rocha Campos Cassano, R., Mariani, P.D.S.C., Kitajima, J.P., Santos, B., Proto-Siqueira, R., Cantarelli, V.V., Tosta, S., Nardy, V.B., Reboredo de Oliveira da Silva, L., Gómez, M.K.A., Lima, J.G., Ribeiro, A.A., Guimarães, N.R., Watanabe, L.T., Barbosa Da Silva, L., da Silva Ferreira, R., da Penha, M.P.F., Ortega, M.J., de la Fuente, A.G., Villalba, S., Torales, J., Gamarra, M.L., Aquino, C., Figueredo, G.P.M., Fava, W.S., Motta-Castro, A.R.C., Venturini, J., do Vale Leone de Oliveira, S.M., Gonçalves, C.C.M., do Carmo Debur Rossa, M., Becker, G.N., Giacomini, M.P., Marques, N.Q., Riediger, I.N., Raboni, S., Mattoso, G., Cataneo, A.D., Zanluca, C., Duarte dos Santos, C.N., Assato, P.A., Allan da Silva da Costa, F., Poleti, M.D., Lesbon, J.C.C., Mattos, E.C., Banho, C.A., Sacchetto, L., Moraes, M.M., Grotto, R.M.T., Souza-Neto, J.A., Nogueira, M.L., Fukumasu, H., Coutinho, L.L., Calado, R.T., Neto, R.M., Bispo de Filippis, A.M., Venancio da Cunha, R., Freitas, C., Peterka, C.R.L., de Fátima Rangel Fernandes, C., Navegantes, W., do Carmo Said, R.F., Campelo de A e Melo, C.F., Almiron, M., Lourenço, J., de Oliveira, T., Holmes, E.C., Haddad, R., Sampaio, S.C., Elias, M.C., Kashima, S., Junior de Alcantara, L.C. and Covas, D.T. (2022) 'Genomic epidemiology of the SARS-CoV-2 epidemic in Brazil', *Nature Microbiology*, 7(9), pp. 1490–1500.
46. McLaughlin, A., Montoya, V., Miller, R.L., Mordecai, G.J., Canadian COVID-19 Genomics Network (CanCOGen) Consortium, Worobey, M., Poon, A.F.Y. and Joy, J.B. (2022) 'Genomic epidemiology of the first two waves of SARS-CoV-2 in Canada', *eLife*, 11, p. e73896.
47. Kraemer, M.U.G., Hill, V., Ruis, C., Dellicour, S., Bajaj, S., McCrone, J.T., Baele, G., Parag, K.V., Battle, A.L., Gutierrez, B., Jackson, B., Colquhoun, R., O'Toole, Á., Klein, B., Vespignani, A., COVID-19 Genomics UK (COG-UK) Consortium, Volz, E., Faria, N.R., Aanensen, D.M., Loman, N.J., du Plessis, L., Cauchemez, S., Rambaut, A., Scarpino, S.V. and Pybus, O.G. (2021) 'Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence', *Science*, 373(6557), pp. 889–895.

48. Reichmuth, M.L., Hodcroft, E.B. and Althaus, C.L. (2023) 'Importation of Alpha and Delta variants during the SARS-CoV-2 epidemic in Switzerland: Phylogenetic analysis and intervention scenarios', *PLoS pathogens*, 19(8), p. e1011553.
49. Michaelsen, T.Y., Bennedbæk, M., Christiansen, L.E., Jørgensen, M.S.F., Møller, C.H., Sørensen, E.A., Knutsson, S., Brandt, J., Jensen, T.B.N., Chiche-Lapierre, C., Collados, E.F., Sørensen, T., Petersen, C., Le-Quy, V., Sereika, M., Hansen, F.T., Rasmussen, M., Fonager, J., Karst, S.M., Marvig, R.L., Stegger, M., Sieber, R.N., Skov, R., Legarth, R., Krause, T.G., Fomsgaard, A., Danish COVID-19 Genome Consortium (DCGC) and Albertsen, M. (2022) 'Introduction and transmission of SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark', *Genome medicine*, 14(1), p. 47.
50. Vöhringer, H.S., Sanderson, T., Sinnott, M., De Maio, N., Nguyen, T., Goater, R., Schwach, F., Harrison, I., Hellewell, J., Ariani, C.V., Gonçalves, S., Jackson, D.K., Johnston, I., Jung, A.W., Saint, C., Sillitoe, J., Suciu, M., Goldman, N., Panovska-Griffiths, J., Wellcome Sanger Institute COVID-19 Surveillance Team, COVID-19 Genomics UK (COG-UK) Consortium*, Birney, E., Volz, E., Funk, S., Kwiatkowski, D., Chand, M., Martincorena, I., Barrett, J.C. and Gerstung, M. (2021) 'Genomic reconstruction of the SARS-CoV-2 epidemic in England', *Nature*, 600(7889), pp. 506–511.
51. Schliep, K.P. (2011) 'phangorn: phylogenetic analysis in R', *Bioinformatics (Oxford, England)*, 27(4), pp. 592-593.
52. Sagulenko, P., Puller, V. and Neher, R.A. (2018) 'TreeTime: Maximum-likelihood phylodynamic analysis', *Virus evolution*, 4(1), p. vex042.
53. Viana, R., Moyo, S., Amoako, D.G., Tegally, H., Scheepers, C., Althaus, C.L., Anyaneji, U.J., Bester, P.A., Boni, M.F., Chand, M., Choga, W.T., Colquhoun, R., Davids, M., Deforche, K., Doolabh, D., du Plessis, L., Engelbrecht, S., Everatt, J., Giandhari, J., Giovanetti, M., Hardie, D., Hill, V., Hsiao, N.-Y., Iranzadeh, A., Ismail, A., Joseph, C., Joseph, R., Koopile, L., Kosakovsky Pond, S.L., Kraemer, M.U.G., Kuate-Lere, L., Laguda-Akingba, O., Lesetedi-Mafoko, O., Lessells, R.J., Lockman, S., Lucaci, A.G., Maharaj, A., Mahlangu, B., Maponga, T., Mahlakwane, K., Makatini, Z., Marais, G., Maruapula, D., Masupu, K., Matshaba, M., Mayaphi, S., Mbhele, N., Mbulawa, M.B., Mendes, A., Mlisana, K., Mnguni, A., Mohale, T., Moir, M., Moruisi, K., Mosepele, M., Motsatsi, G., Motswaledi, M.S., Mphoyakgosi, T., Msomi, N., Mwangi, P.N., Naidoo, Y., Ntuli, N., Nyaga, M., Olubayo, L., Pillay, S., Radibe, B., Ramphal, Y., Ramphal, U., San, J.E., Scott, L., Shapiro, R., Singh, L., Smith-Lawrence, P., Stevens, W., Strydom, A., Subramoney, K., Tebeila, N., Tshiabuila, D., Tsui, J., van Wyk, S., Weaver, S., Wibmer, C.K., Wilkinson, E., Wolter, N., Zarebski, A.E., Zuze, B., Goedhals, D., Preiser, W., Treurnicht, F., Venter, M., Williamson, C., Pybus, O.G., Bhiman, J., Glass, A., Martin, D.P., Rambaut, A., Gaseitsiwe, S., von Gottberg, A. and de Oliveira, T. (2022) 'Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa', *Nature*, 603(7902), pp. 679–686.
54. COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk (2020) 'An integrated national scale SARS-CoV-2 genomic surveillance network', *Lancet Microbe*, 1(3), pp. e99–e100.
55. World Health Organization (2021) *Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health*. Available at https://www.who.int/publications/i/item/9789240018440 (Accessed: 28 April 2025).

56. Balcan, D., Gonçalves, B., Hu, H., Ramasco, J.J., Colizza, V. and Vespignani, A. (2010) 'Modeling the spatial spread of infectious diseases: the GLobal Epidemic and Mobility computational model', *Journal of computational science*, 1(3), p. 132.

Appendix D:
Supplementary materials for Chapter 5

D.1 Design and implementation of SOPHI

The SOPHI ("*Sandbox for Optimising genomic sampling for PHylogeographic Inference*") framework consists of two primary components: a backend server implemented in Django as a REST API (with source code available at https://github.com/joetsui1994/sophi-backend), and a frontend web interface built in React (with source code available at https://github.com/joetsui1994/sophi-frontend). At a high level, the backend serves as the core infrastructure responsible for managing data generated by outbreak simulations, performing phylogeographic inferences based on user-defined sampling strategies, and storing data extracted from inference outputs. Meanwhile, the frontend provides an interactive graphic interface organised into four main panels:

1. *Simulation repository*: Displays a list of simulated outbreaks, along with key summary statistics (e.g., number of geographic locations or demes, total population size, and outbreak duration); researchers are prompted to select one of these outbreak scenarios, data from which would then become available for visualisation and further analysis in subsequent panels.
2. *Data panel*: Provides visualisations and summary statistics of various data streams from the selected outbreak scenario - such as daily case incidence, human mobility patterns, and other contextual information (e.g., transmission coefficients, recovery rate, sampling rates).

3. *Inference panel*: Includes a configurable input form where researchers can design their own sampling strategies from existing presets (see later section) and specify key parameters (e.g., number of sequences to include, target demes, minimum number of sequences to sample per deme/day); once finalised, the form is submitted to the backend which executes the requested phylogeographic inference.
4. *Evaluation panel*: Displays the outputs of completed phylogeographic inferences (e.g., transmission lineages that have been identified and their associated importation events, temporal trends in importation intensity at each deme, the annotated phylogeny); researchers can compare these results against the simulated "ground truth" data using a number of evaluation metrics which SOPHI automatically computes and visualises in the graphical interface.

An overview of these different components and how they are integrated within SOPHI is provided in Fig. D.1 in this Appendix.

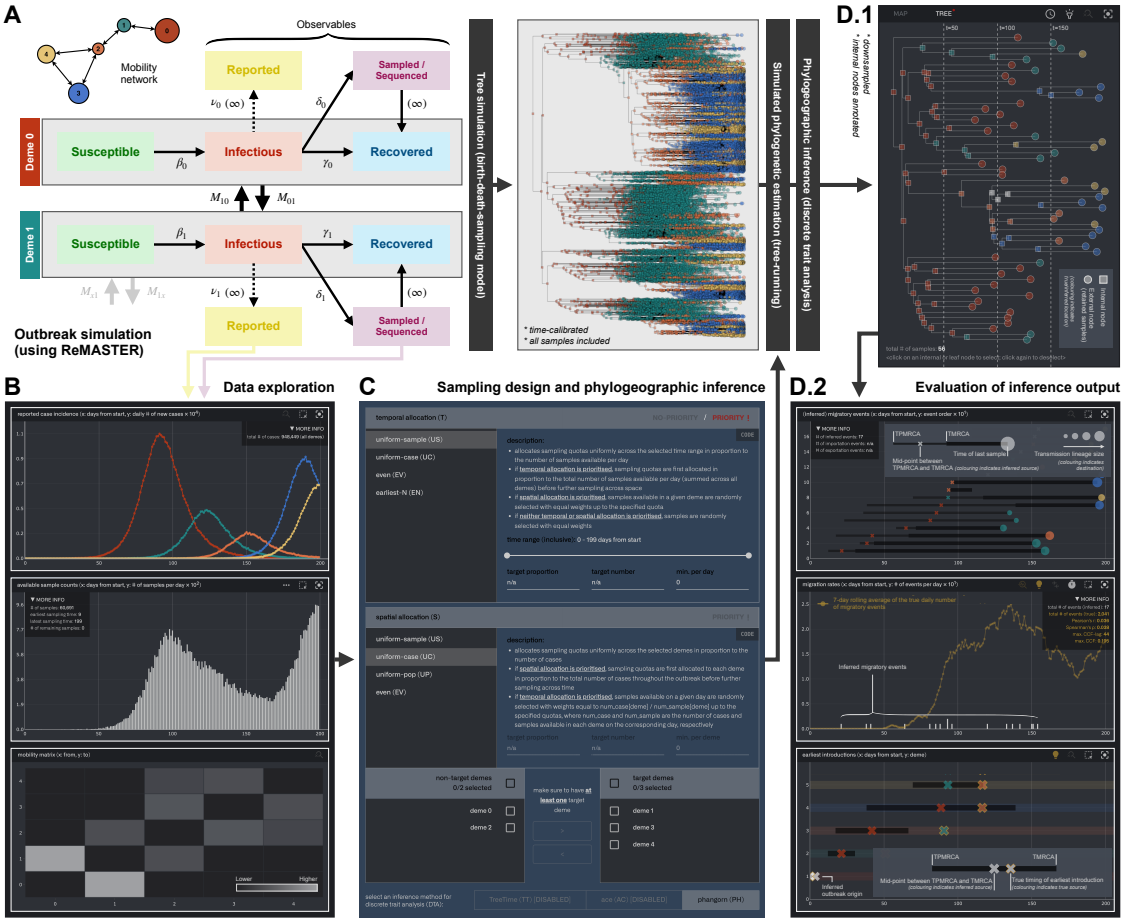


Fig. D.1. Overview of the SOPHI framework. (A) Schematic of an outbreak simulation across a mobility network with 5 demes using ReMASTER. Local epidemic in each deme is governed by the standard Susceptible-Infectious-Recovered (SIR) model with deme-specific transmission coefficient β_i and recovery rate γ_i. Infectious individuals in each deme i are reported at a constant rate ν_i and sampled/sequenced at a constant rate δ_i, followed by immediate recovery. Individuals can migrate freely, regardless of their infection status, between any connected demes i and j at a predefined constant rate M_{ij}. In addition to outbreak trajectories, ReMASTER also produces a time-calibrated phylogeny representing the ancestral relationships between sampled infections, with internal nodes annotated by their true corresponding deme (indicated by node colours in this visualisation, generated by SOPHI). (B) Data panel in the SOPHI frontend interface, showing daily reported case numbers per deme (top), daily available genome samples (middle), and daily movement between demes (bottom; displayed as a heatmap). (C) Inference panel in the SOPHI frontend interface, where users can select from pre-defined spatial and temporal sampling strategies, set relevant design parameters (e.g., sampling proportion, sampling period, target demes), and specify a discrete trait analysis (DTA) method for phylogeographic inference. (D.1) Once a sampling strategy has been specified, tips corresponding to unsampled infections are pruned from the simulated phylogeny to emulate the phylogenetic tree estimation process. The resulting pruned tree is then used as input for DTA, producing a time-calibrated tree with internal nodes labelled by their inferred deme. (D.2) Visualisation of DTA outputs: (top) inferred

migratory events through time; (middle) histogram representing the daily number of inferred migratory events, with solid yellow line representing the 7-day rolling average of the true daily counts; (bottom) earliest inferred importation event for each deme, where crosses with yellow borders indicate the true importation times.

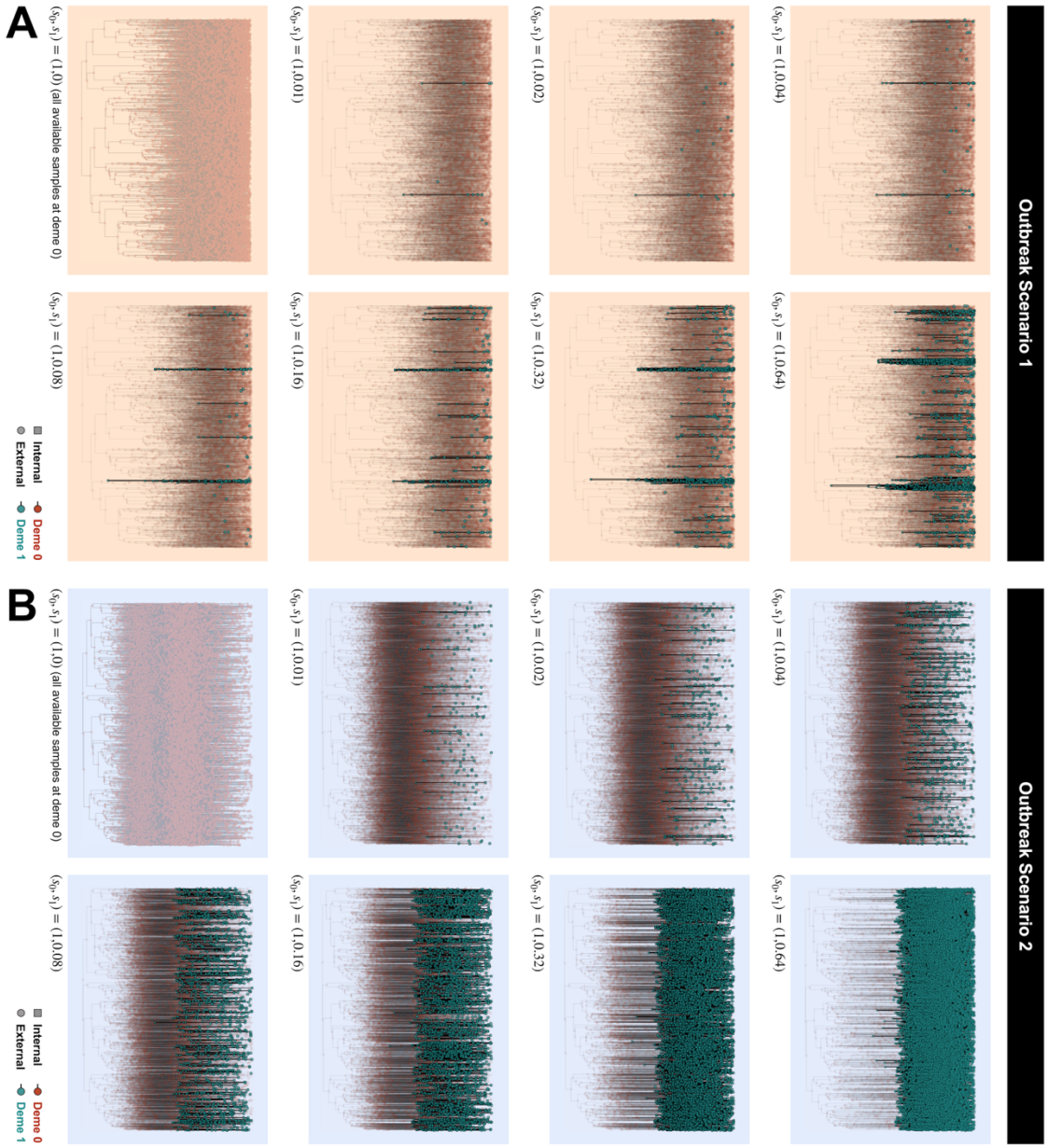


Fig. D.1. Impact of varying sampling proportions at deme 1 on phylogeographic inference under two outbreak scenarios (application 1). (A) Each panel shows the annotated tree from one of 40 inference replicates performed under outbreak scenario 1, with a specific sampling proportion at deme 1 ($s_1 = 0.0, 0.01, 0.02, 0.04, 0.08, 0.16, 0.32,$ or 0.64), while including all available samples at deme 0 (the outbreak origin), i.e. $s_0 = 1$. Squares represent internal nodes; circles represent external nodes. Node colour indicates the inferred or observed location (deme 0: red; deme 1: green). (B) Same as panel (A), but under outbreak scenario 2.

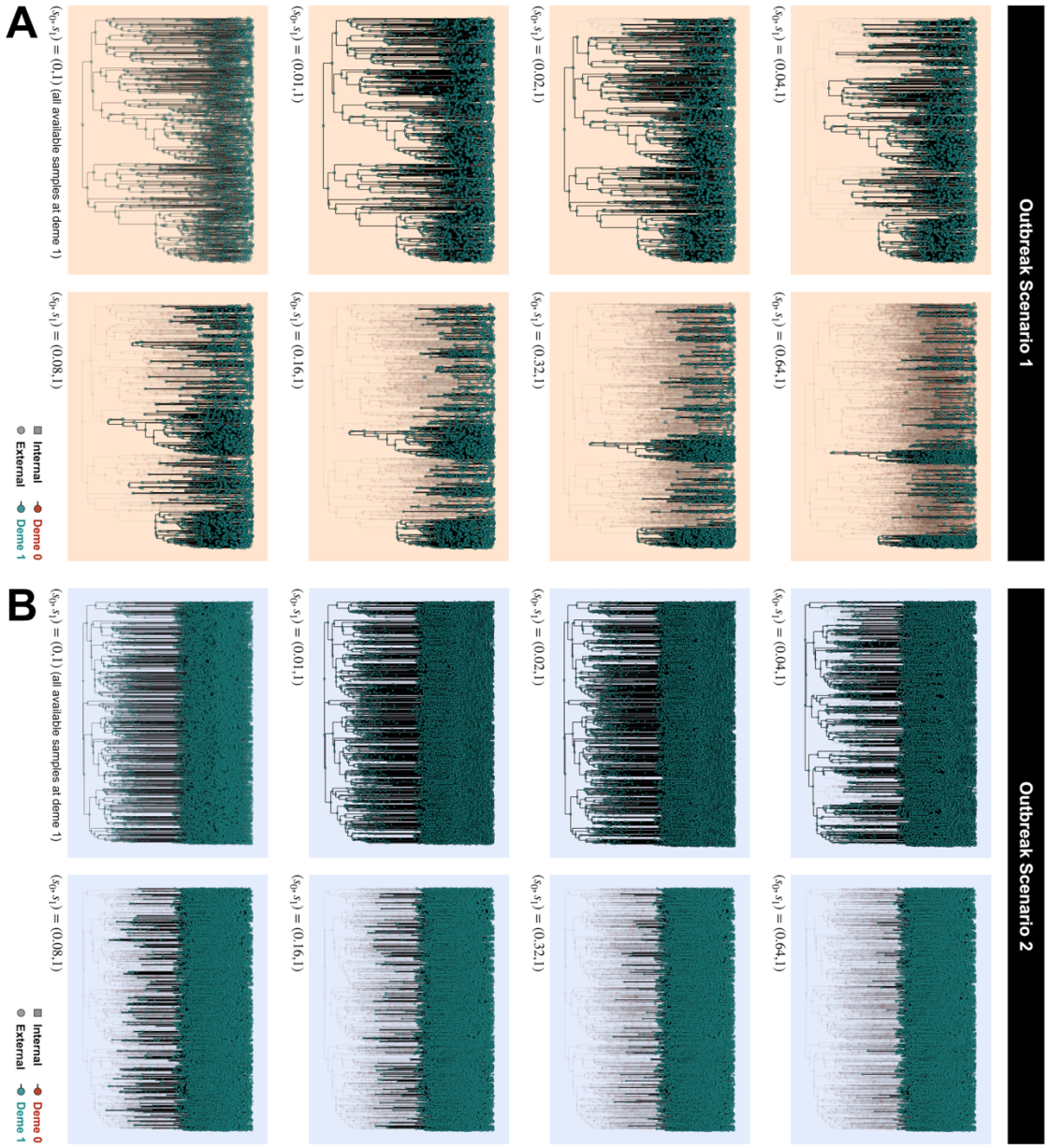


Fig. D.2. Impact of varying sampling proportions at deme 0 on phylogeographic inference under two outbreak scenarios (application 1). (A) Each panel shows the annotated tree from one of 40 inference replicates performed under outbreak scenario 1, with a specific sampling proportion at deme 0 (the outbreak origin; $s_0 = 0.0$, 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, or 0.64), while including all available samples at deme 1, i.e. $s_1 = 1$. Squares represent internal nodes; circles represent external nodes. Node colour indicates the inferred or observed location (deme 0: red; deme 1: green). (B) Same as panel (A), but under outbreak scenario 2.

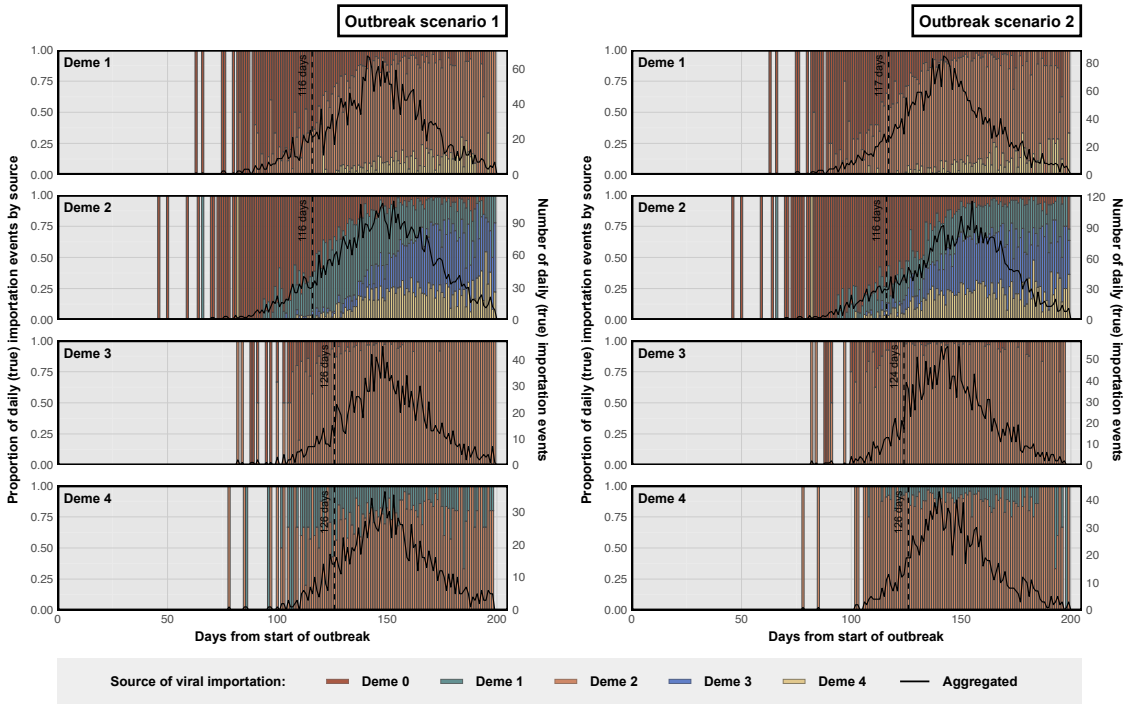


Fig. D.3. Source distribution of true viral importation events (application 2). Each row corresponds to one of four demes (deme 1, 2, 3, and 4, with deme 0 (the outbreak origin) excluded) in the mobility network considered in Section 5.3.2 of Chapter 5, with left and right panels showing results for outbreak scenarios 1 and 2, respectively. Within each panel, stacked bars represent the daily proportion of true viral importation events from each source location (deme 0: red; deme 1: green; deme 2: orange; deme 3: blue; deme 4: yellow), corresponding to the left y-axis. The solid black line shows the total number of daily importation events aggregated over all source locations, corresponding to the right y-axis. The dotted vertical line in each panel marks the time at which the 10th percentile of cumulative importation events occurs in the corresponding deme.

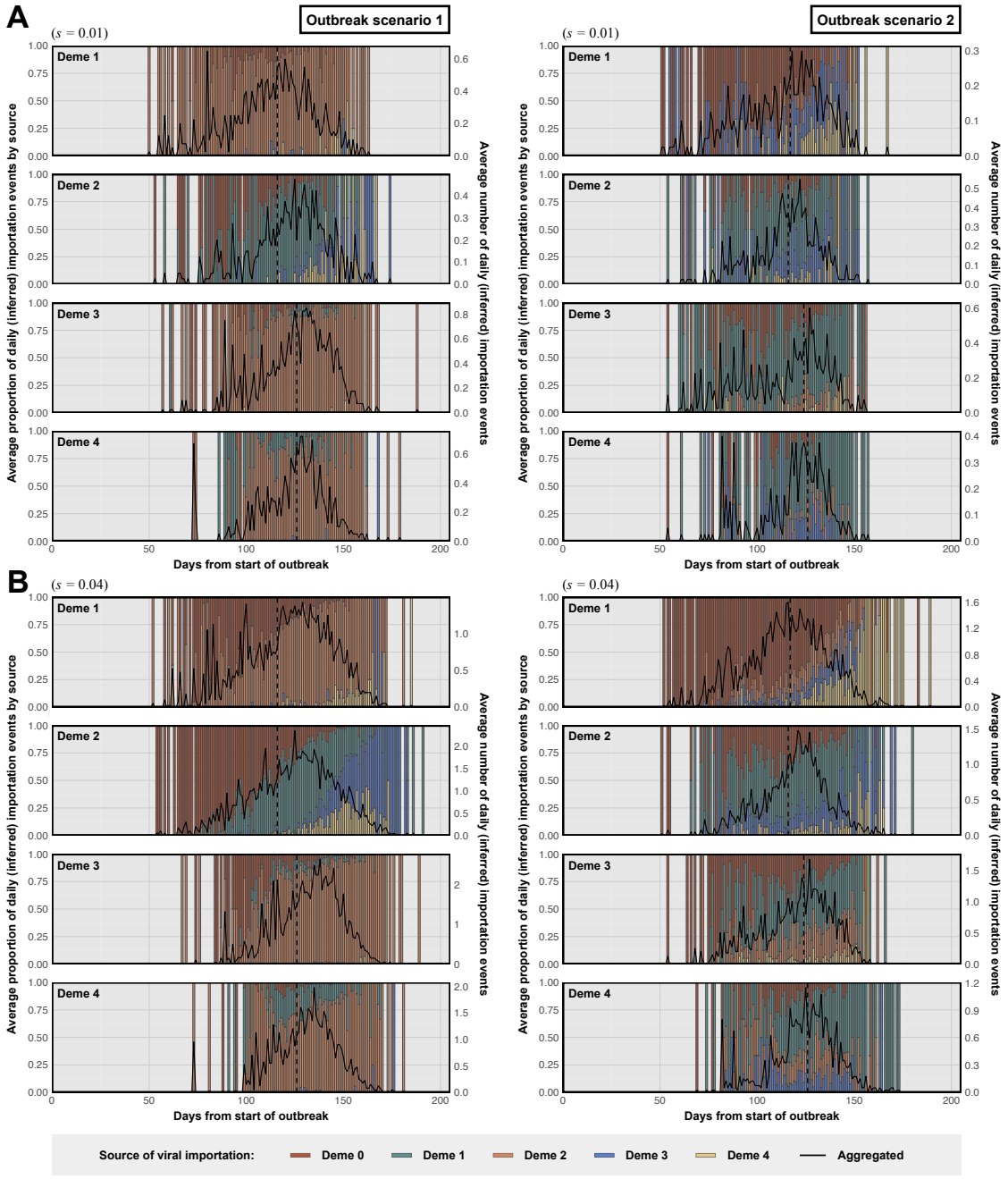


Fig. D.4. Source distribution of inferred importation events under the uniform-sample (US) sampling scheme at selected sampling proportions (application 2). Panels A and B show results from inferences applying the uniform-sample (US) sampling scheme at sampling proportions of 0.01 and 0.04, respectively. (A) Each row corresponds to one of four demes (deme 1, 2, 3, and 4, with deme 0 (the outbreak origin) excluded) in the mobility network considered in Section 5.3.2 of Chapter 5, with left and right panels showing results for outbreak scenarios 1 and 2, respectively. Stacked bars in each panel represent the daily proportion of inferred viral importation events from each source location (deme 0: red; deme 1: green; deme 2: orange; deme 3: blue; deme 4: yellow), averaged across 40 inference replicates (left y-axis). The solid black line shows the average total number of daily inferred importation events aggregated over all source

locations (right y-axis). The dotted vertical line in each panel marks the time at which the 10^{th} percentile of cumulative (true) importation events occurs in the corresponding deme. (B) Same as panel (A), but at a sampling proportion of 0.04.

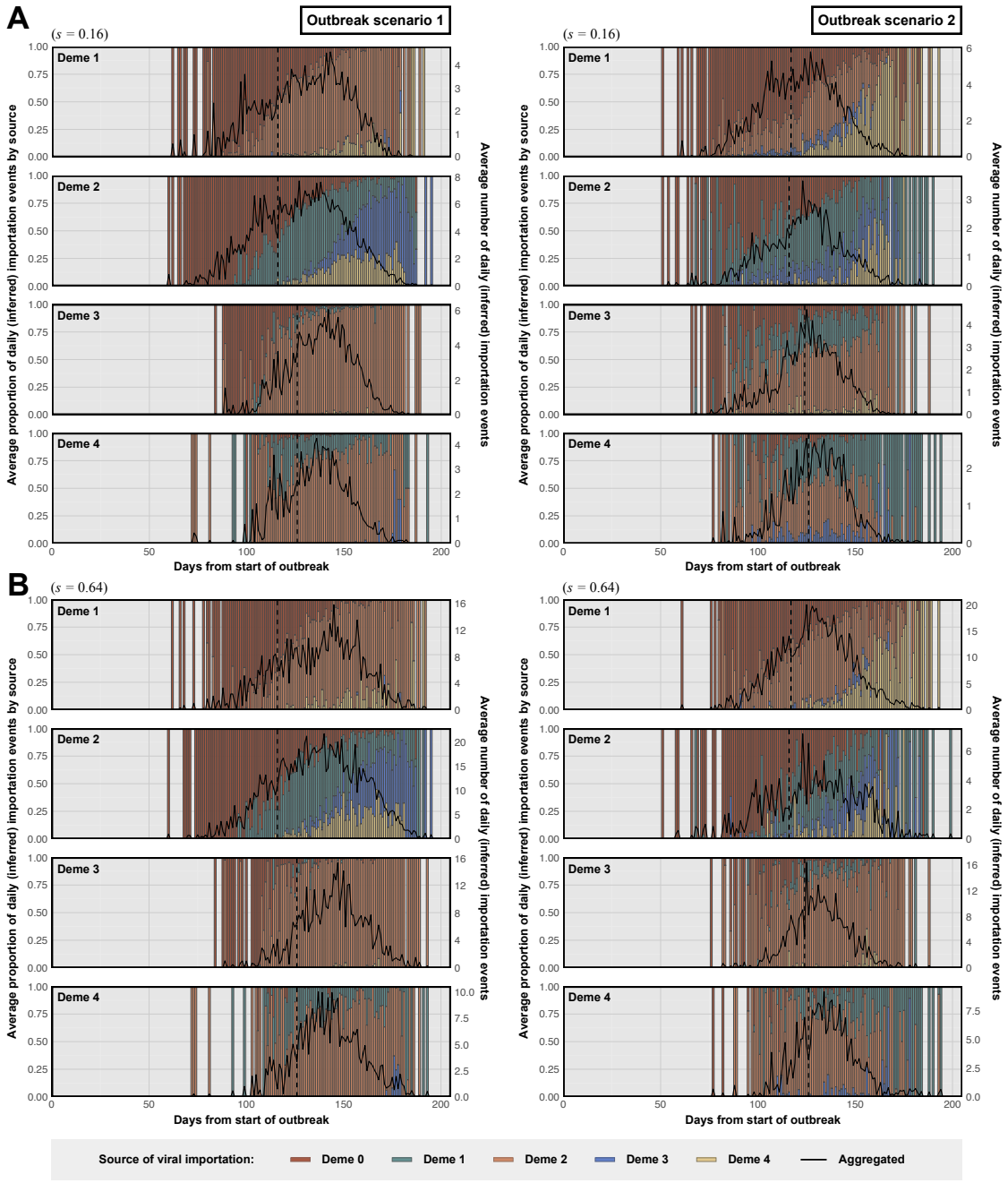


Fig. D.5. Source distribution of inferred importation events under the uniform-sample (US) sampling scheme at selected sampling proportions (application 2). Panels A and B show results from inferences applying the uniform-sample (US) sampling scheme at sampling proportions of 0.16 and 0.64, respectively. (A) Each row corresponds to one of four demes (deme 1, 2, 3, and 4, with deme 0 (the outbreak origin) excluded) in the mobility network considered in Section 5.3.2 of Chapter 5, with left and right panels showing results for outbreak scenarios 1 and 2, respectively. Stacked bars in each panel represent the daily proportion of inferred viral importation events from each source location (deme 0: red; deme 1: green; deme 2: orange; deme 3: blue; deme 4: yellow), averaged across 40 inference replicates (left y-axis). The solid black line shows the average total number of daily inferred importation events aggregated over all source

locations (right y-axis). The dotted vertical line in each panel marks the time at which the 10^{th} percentile of cumulative (true) importation events occurs in the corresponding deme. (B) Same as panel (A), but at a sampling proportion of 0.64.

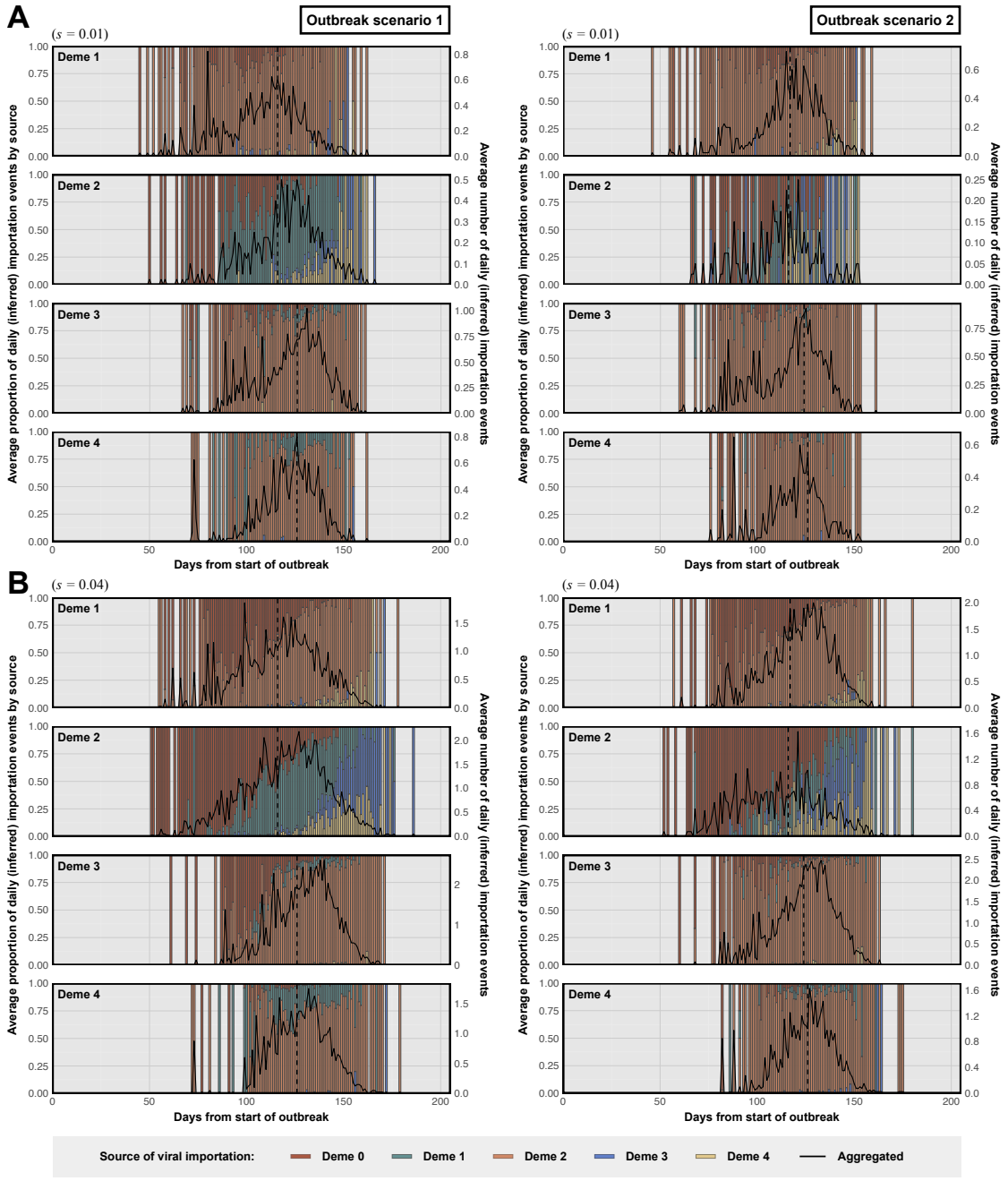


Fig. D.6. Source distribution of inferred importation events under the uniform-case (UC) sampling scheme at selected sampling proportions (application 2). Panels A and B show results from inferences applying the uniform-sample (US) sampling scheme at sampling proportions of 0.01 and 0.04, respectively. (A) Each row corresponds to one of four demes (deme 1, 2, 3, and 4, with deme 0 (the outbreak origin) excluded) in the mobility network considered in Section 5.3.2 of Chapter 5, with left and right panels showing results for outbreak scenarios 1 and 2, respectively. Stacked bars in each panel represent the daily proportion of inferred viral importation events from each source location (deme 0: red; deme 1: green; deme 2: orange; deme 3: blue; deme 4: yellow), averaged across 40 inference replicates (left y-axis). The solid black line shows the average total number of daily inferred importation events aggregated over all source

locations (right y-axis). The dotted vertical line in each panel marks the time at which the 10^{th} percentile of cumulative (true) importation events occurs in the corresponding deme. (B) Same as panel (A), but at a sampling proportion of 0.04.

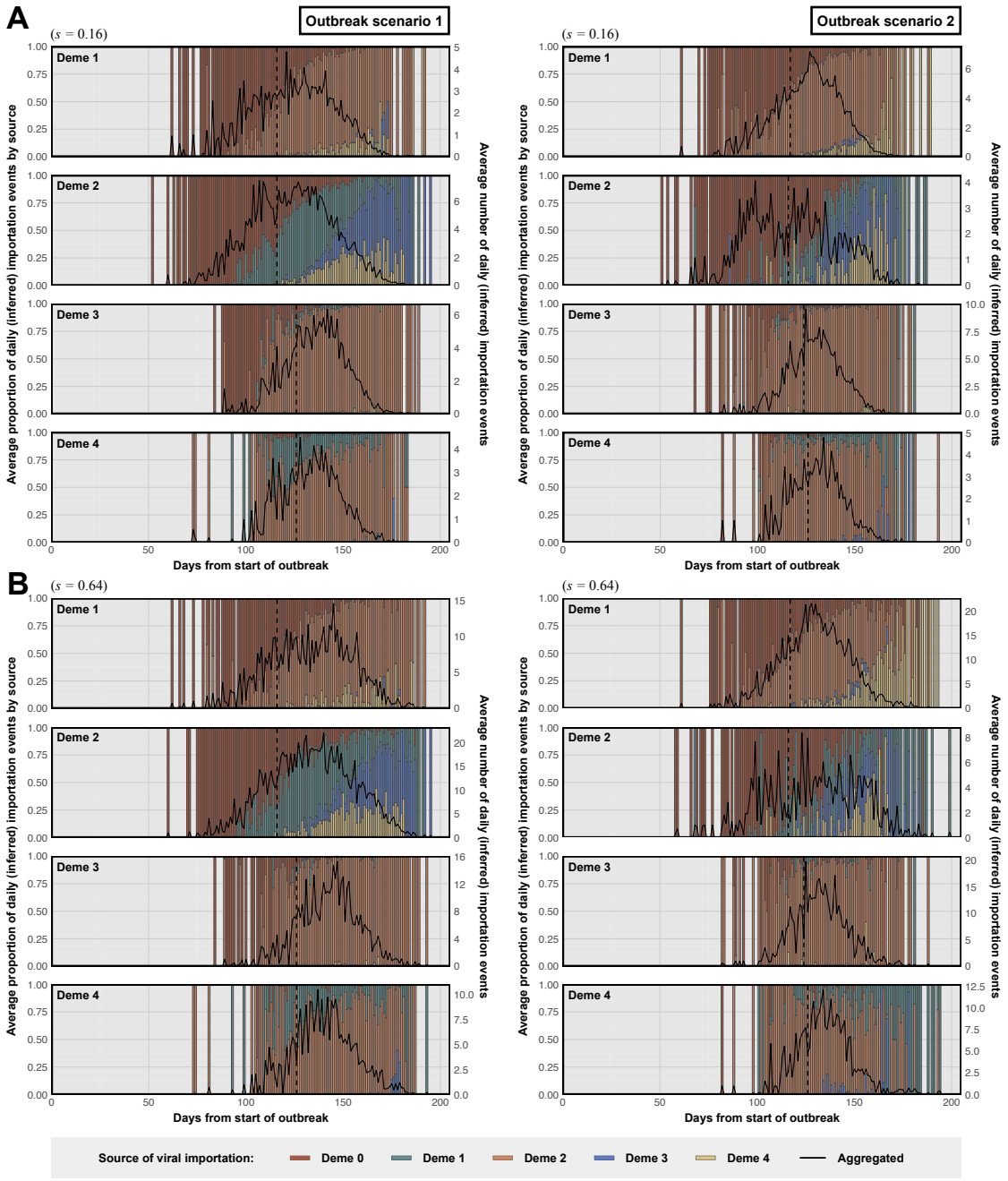


Fig. D.7. Source distribution of inferred importation events under the uniform-case (UC) sampling scheme at selected sampling proportions (application 2). Panels A and B show results from inferences applying the uniform-sample (US) sampling scheme at sampling proportions of 0.16 and 0.64, respectively. (A) Each row corresponds to one of four demes (deme 1, 2, 3, and 4, with deme 0 (the outbreak origin) excluded) in the mobility network considered in Section 5.3.2 of Chapter 5, with left and right panels showing results for outbreak scenarios 1 and 2, respectively. Stacked bars in each panel represent the daily proportion of inferred viral importation events from each source location (deme 0: red; deme 1: green; deme 2: orange; deme 3: blue; deme 4: yellow), averaged across 40 inference replicates (left y-axis). The solid black line shows the average total number of daily inferred importation events aggregated over all source

locations (right y-axis). The dotted vertical line in each panel marks the time at which the 10^{th} percentile of cumulative (true) importation events occurs in the corresponding deme. (B) Same as panel (A), but at a sampling proportion of 0.64.

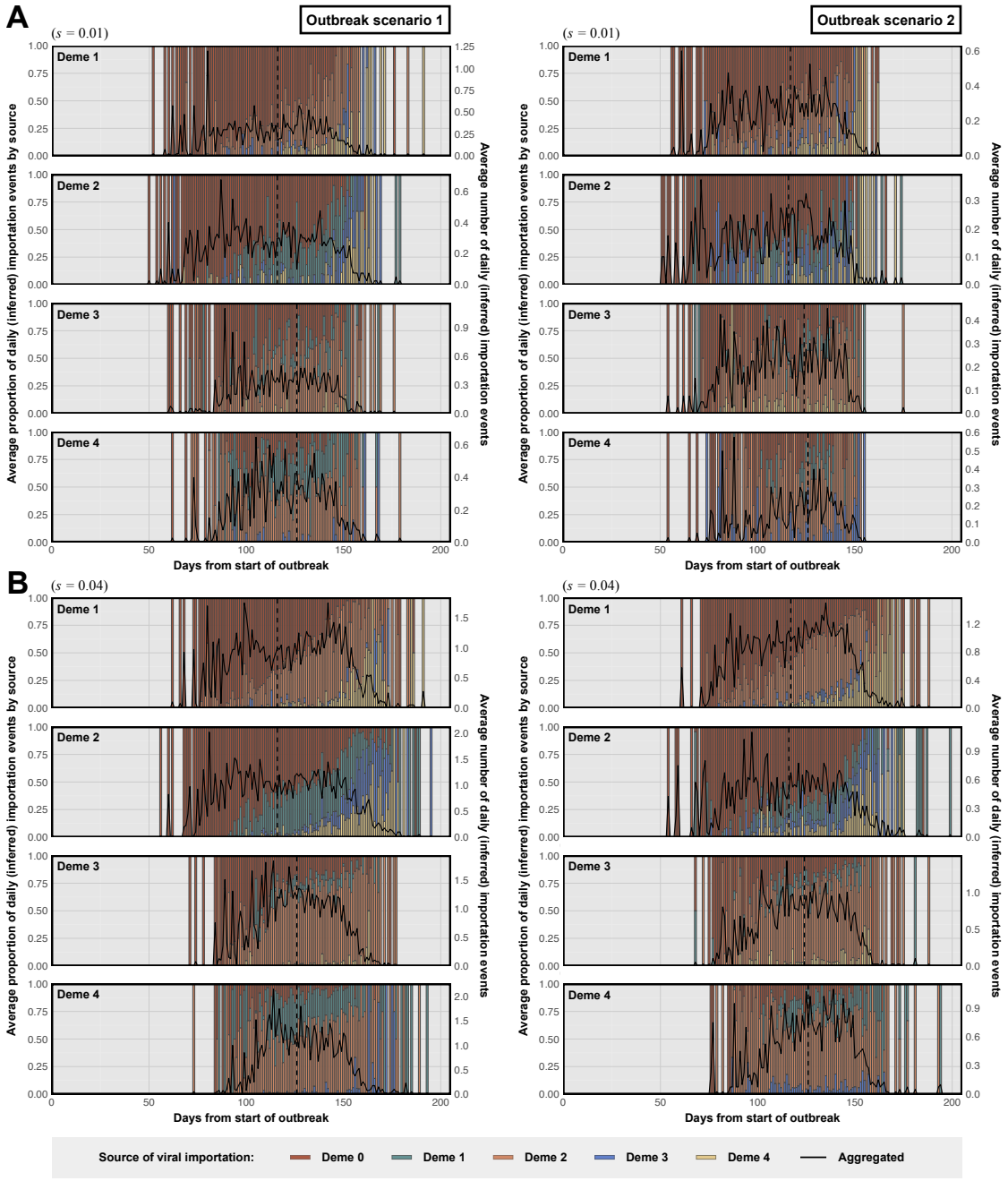


Fig. D.8. Source distribution of inferred importation events under the even (EV) sampling scheme at selected sampling proportions (application 2). Panels A and B show results from inferences applying the uniform-sample (US) sampling scheme at sampling proportions of 0.01 and 0.04, respectively. (A) Each row corresponds to one of four demes (deme 1, 2, 3, and 4, with deme 0 (the outbreak origin) excluded) in the mobility network considered in Section 5.3.2 of Chapter 5, with left and right panels showing results for outbreak scenarios 1 and 2, respectively. Stacked bars in each panel represent the daily proportion of inferred viral importation events from each source location (deme 0: red; deme 1: green; deme 2: orange; deme 3: blue; deme 4: yellow), averaged across 40 inference replicates (left y-axis). The solid black line shows the average total number of daily inferred importation events aggregated over all source

locations (right y-axis). The dotted vertical line in each panel marks the time at which the 10^{th} percentile of cumulative (true) importation events occurs in the corresponding deme. (B) Same as panel (A), but at a sampling proportion of 0.04.

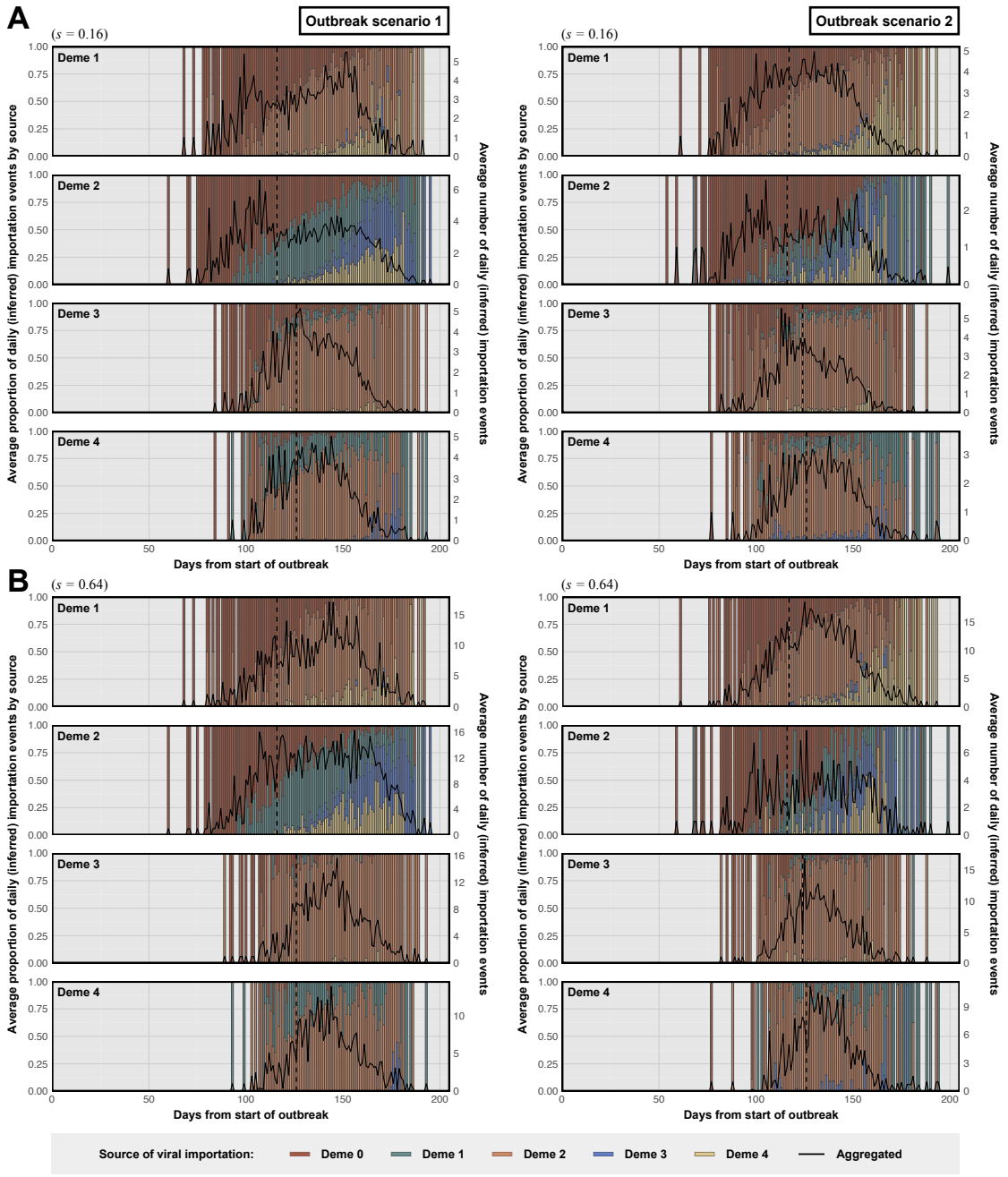


Fig. D.9. Source distribution of inferred importation events under the even (EV) sampling scheme at selected sampling proportions (application 2). Panels A and B show results from inferences applying the uniform-sample (US) sampling scheme at sampling proportions of 0.16 and 0.64, respectively. (A) Each row corresponds to one of four demes (deme 1, 2, 3, and 4, with deme 0 (the outbreak origin) excluded) in the mobility network considered in Section 5.3.2 of Chapter 5, with left and right panels showing results for outbreak scenarios 1 and 2, respectively. Stacked bars in each panel represent the daily proportion of inferred viral importation events from each source location (deme 0: red; deme 1: green; deme 2: orange; deme 3: blue; deme 4: yellow), averaged across 40 inference replicates (left y-axis). The solid black line shows the average total number of daily inferred importation events aggregated over all source

locations (right y-axis). The dotted vertical line in each panel marks the time at which the 10^{th} percentile of cumulative (true) importation events occurs in the corresponding deme. (B) Same as panel (A), but at a sampling proportion of 0.64.

Table D.1. Median cosine-distance between the true and inferred source distributions of early viral importation events at deme 1 under outbreak scenario 1, across different sampling proportions and three different sampling schemes: uniform-sample (US), uniform-case (UC), and even (EV). For each sampling proportion, the smallest median distance is shown in bold, and the second-smallest value is underlined.

Sampling proportion	Uniform-sample (US)	Uniform-case (UC)	Even (EV)
0.01	0.456797	<u>0.455736</u>	**0.148178**
0.02	0.486932	<u>0.439974</u>	**0.098278**
0.04	0.343347	<u>0.274012</u>	**0.097701**
0.08	0.105408	<u>0.066276</u>	**0.105203**
0.16	<u>0.051308</u>	**0.047599**	0.075932
0.32	<u>0.066818</u>	**0.063432**	0.070232
0.64	0.053687	<u>0.041130</u>	**0.040315**

Table D.2. Median cosine-distance between the true and inferred source distributions of early viral importation events at deme 2 under outbreak scenario 1, across different sampling proportions and three different sampling schemes: uniform-sample (US), uniform-case (UC), and even (EV). For each sampling proportion, the smallest median distance is shown in bold, and the second-smallest value is underlined.

Sampling proportion	Uniform-sample (US)	Uniform-case (UC)	Even (EV)
0.01	0.362738	0.362738	**0.289791**
0.02	0.296476	0.287622	**0.183575**
0.04	0.143187	**0.122364**	0.177172
0.08	0.086993	**0.081724**	0.129993
0.16	0.079367	**0.079135**	0.096108
0.32	**0.079113**	0.079130	0.080166
0.64	0.079075	0.079067	**0.058982**

Table D.3. Median cosine-distance between the true and inferred source distributions of early viral importation events at deme 3 under outbreak scenario 1, across different sampling proportions and three different sampling schemes: uniform-sample (US), uniform-case (UC), and even (EV). For each sampling proportion, the smallest median distance is shown in bold, and the second-smallest value is underlined.

Sampling proportion	Uniform-sample (US)	Uniform-case (UC)	Even (EV)
0.01	**0.285735**	**0.285735**	0.442444
0.02	**0.182041**	<u>0.210498</u>	0.429908
0.04	<u>0.198142</u>	**0.163337**	0.423003
0.08	**0.217181**	<u>0.217897</u>	0.371328
0.16	<u>0.190081</u>	**0.172597**	0.282115
0.32	<u>0.176799</u>	**0.165958**	0.190469
0.64	0.161854	**0.156315**	<u>0.158327</u>

Table D.4. Median cosine-distance between the true and inferred source distributions of early viral importation events at deme 4 under outbreak scenario 1, across different sampling proportions and three different sampling schemes: uniform-sample (US), uniform-case (UC), and even (EV). For each sampling proportion, the smallest median distance is shown in bold, and the second-smallest value is underlined.

Sampling proportion	Uniform-sample (US)	Uniform-case (UC)	Even (EV)
0.01	**0.368288**	**0.368288**	0.370421
0.02	<u>0.254595</u>	**0.228404**	0.291353
0.04	<u>0.230048</u>	**0.221780**	0.252752
0.08	0.230732	<u>0.181864</u>	**0.169329**
0.16	<u>0.166667</u>	**0.138885**	0.167746
0.32	0.184998	<u>0.165030</u>	**0.162539**
0.64	0.171805	<u>0.166237</u>	**0.160379**

Table D.5. Median cosine-distance between the true and inferred source distributions of early viral importation events at deme 1 under outbreak scenario 2, across different sampling proportions and three different sampling schemes: uniform-sample (US), uniform-case (UC), and even (EV). For each sampling proportion, the smallest median distance is shown in bold, and the second-smallest value is underlined.

Sampling proportion	Uniform-sample (US)	Uniform-case (UC)	Even (EV)
0.01	0.397721	0.362738	**0.272365**
0.02	0.372048	0.363799	**0.203715**
0.04	0.318495	0.263060	**0.182754**
0.08	0.235550	**0.114811**	0.150199
0.16	0.146771	**0.055821**	0.132892
0.32	**0.082557**	0.084047	0.119922
0.64	**0.045600**	0.056125	0.061157

Table D.6. Median cosine-distance between the true and inferred source distributions of early viral importation events at deme 2 under outbreak scenario 2, across different sampling proportions and three different sampling schemes: uniform-sample (US), uniform-case (UC), and even (EV). For each sampling proportion, the smallest median distance is shown in bold, and the second-smallest value is underlined.

Sampling proportion	Uniform-sample (US)	Uniform-case (UC)	Even (EV)
0.01	0.417021	<u>0.357894</u>	**0.283443**
0.02	0.410439	<u>0.319846</u>	**0.196148**
0.04	0.474342	**0.139563**	<u>0.189232</u>
0.08	0.419118	**0.156086**	<u>0.159320</u>
0.16	0.354825	<u>0.166048</u>	**0.141788**
0.32	<u>0.139492</u>	0.169484	**0.134453**
0.64	**0.094540**	0.153080	<u>0.140768</u>

Table D.7. Median cosine-distance between the true and inferred source distributions of early viral importation events at deme 3 under outbreak scenario 2, across different sampling proportions and three different sampling schemes: uniform-sample (US), uniform-case (UC), and even (EV). For each sampling proportion, the smallest median distance is shown in bold, and the second-smallest value is underlined.

Sampling proportion	Uniform-sample (US)	Uniform-case (UC)	Even (EV)
0.01	0.487440	**0.209936**	<u>0.450504</u>
0.02	0.463771	**0.209936**	<u>0.433071</u>
0.04	0.458832	**0.152414**	<u>0.377291</u>
0.08	0.453511	**0.158575**	<u>0.282724</u>
0.16	0.452123	**0.145898**	<u>0.288982</u>
0.32	0.398494	**0.136006**	<u>0.224980</u>
0.64	0.260218	**0.150890**	<u>0.188718</u>

Table D.8. Median cosine-distance between the true and inferred source distributions of early viral importation events at deme 4 under outbreak scenario 2, across different sampling proportions and three different sampling schemes: uniform-sample (US), uniform-case (UC), and even (EV). For each sampling proportion, the smallest median distance is shown in bold, and the second-smallest value is underlined.

Sampling proportion	Uniform-sample (US)	Uniform-case (UC)	Even (EV)
0.01	0.456199	**0.163283**	<u>0.389872</u>
0.02	0.488375	**0.229762**	<u>0.394936</u>
0.04	0.500000	**0.246330**	<u>0.350340</u>
0.08	0.491505	**0.221468**	<u>0.329914</u>
0.16	0.472981	**0.166667**	<u>0.291855</u>
0.32	0.372692	**0.176697**	<u>0.236695</u>
0.64	0.285501	**0.185776**	<u>0.215427</u>

6
Discussion

This thesis addresses a central question in infectious disease epidemiology and public health: how can sampling design and surveillance strategies be improved to enable more accurate and robust inference of the spatiotemporal dynamics of emerging infectious diseases? The ability to infer where and when a pathogen first emerged, and how it spreads through animal and human populations, is critical to our efforts to anticipate, prepare for, and manage future outbreaks. In recent years, there has been a step-change in our capacity to conduct large-scale disease surveillance, driven by the global response to the COVID-19 pandemic (1-4) and growing concerns over the emergence of other novel pathogens through zoonotic spillover (5-7) linked to climate change, environmental degradation, and rapid urbanisation. As the global community seeks to invest in the development of more robust and adaptive disease surveillance infrastructures (8-10), understanding the limitations and inherent biases in existing data collection protocols and their implications for both scientific research and outbreak response, is vital for ensuring that resources are directed to where they will make the most impact.

Value of data in responding to emerging infectious disease outbreaks

The utility of large-scale, interdisciplinary data collection during an infectious disease outbreak is exemplified in Chapter 2, where I investigated the invasion dynamics of SARS-CoV-2 Omicron BA.1 into the UK. Using pathogen genomic data collected through the national COVID-19 Genomics UK (COG-UK) consortium (1) - one of the

world's most extensive genomic surveillance programmes - together with high-resolution human mobility data from mobile devices, I showed through phylogeography that the invasion process and the subsequent local dissemination can be characterised by distinct stages as a result of both human geography and hierarchical mobility networks. Importantly, I found that a large proportion of local infections can be traced back to early viral introductions, which occurred with increasing frequency despite travel restrictions targeting southern African countries where the variant was first reported. Further analyses integrating air traffic data, individual travel histories, and reported case incidence revealed that these introductions were largely driven by major global transit hubs that were not covered by the travel ban due to undetected local transmission.

These findings contribute to several areas of active research that have received substantial attention since the COVID-19 pandemic, such as the role of human mobility in shaping disease spread at both local (11-13) and global (14-16) scales, the impact of viral importation on local transmission dynamics (17, 18), and the effectiveness of border control as an approach to preventing or delaying the invasion of an emerging pathogen (19-24). In the context of disease surveillance in particular, results from Chapter 2 also raised the question of whether existing systems are designed and deployed in a way that can support effective and timely responses to emerging public health threats. For example, given the disproportionate impact of early viral importation on the subsequent local epidemic, and the substantial contribution from countries prior to their local detection, could our existing surveillance system have provided sufficiently early assessments of importation risks to inform the design of more effective travel restrictions?

Improving surveillance strategies to inform outbreak response

Answering this question is, however, far from straightforward, with key challenges falling into two broad categories. First, there is the need to understand how viral importation influences local transmission dynamics, and the extent to which interventions such as travel restrictions can meaningfully alter the trajectory of an outbreak. This challenge has garnered substantial research attention since the COVID-19 pandemic, with numerous studies assessing the impact of travel restrictions on SARS-CoV-2 spread through either retrospective (19, 20, 22, 23) or simulation (21, 24) analyses. Second, and equally important but less explored, is the challenge of collecting timely and informative data to support the decision-making behind such interventions. Although considerable literature on disease surveillance exists, recent progress has primarily focused on the targeted detection of infected populations (25-27), with little emphasis on accurately inferring the overall spatial distribution of infections - a task that is critical for early risk assessments and outbreak preparedness, particularly in resource-constrained settings. This research gap motivated the work presented in Chapter 3, where I reframed the problem of allocating limited testing resources across a mobility network to maximise the information gained about the underlying disease distribution as an iterative node-classification problem. Importantly, this formulation enables the application of active learning - a subfield of machine learning concerned with the selection of data instances for model training (28). By evaluating the performance of existing active learning algorithms as well as a novel policy developed in this work, I derived a number of key principles that could help guide the design of more cost-effective surveillance policies. These include insights regarding the impact of mobility network structure and outbreak progression on test effectiveness, and the need to consider a trade-off between exploration (testing broadly to identify unobserved infection clusters) and exploitation (targeted

testing to identify the boundaries of known infection clusters) in policy design, especially in low-resource settings where complete detection of all infected populations is impractical.

While the results in Chapter 3 offer promising directions for the development of more effective disease surveillance, they rely on a number of simplifying assumptions. For example, the mobility network was modelled as an undirected and unweighted graph, with each connected location having binary infection status, and the distribution of infections across the network was assumed to be static, rather than changing over time. Further research is therefore needed to assess the robustness of the derived principles under more realistic outbreak scenarios - such as those with evolving disease prevalence, time-varying mobility patterns, and noisy or delayed test feedback. For instance, ongoing work now focuses on developing surrogate models that are better able to account for these dynamics and their complex interactions, while offering greater scalability to larger networks by leveraging recent advances in Graph Neural Networks (GNNs) (29-31) and Transformer-based Neural Processes (TNPs) (32). Building on these extensions, the same adaptive sampling approach could then be applied to a broader range of epidemiological contexts, such as informing the design of sampling strategies for large-scale seroprevalence studies (33, 34), or the implementation of aircraft wastewater surveillance systems for early detection of emerging pathogens (35, 36) – both are areas being actively explored in ongoing work.

Utility of pathogen genomic data under heterogeneous sampling

In parallel to the challenge of collecting timely and informative data for outbreak response, another question that emerged from the work in Chapter 2 concerns the utility of the collected data, specifically pathogen genomes, for informing key estimates in

downstream phylogeographic inference. Indeed, a key result from the analysis of the invasion of SARS-CoV-2 Omicron BA.1 was that a small proportion of early introductions were responsible for the majority of local infections - with the caveat that the number of introductions detected likely represents only a small fraction of the true number due to undersampling of local infections. While this underestimation was a recognised issue reported in numerous studies (17, 37-39), the underlying sampling process that gives rise to this bias remained poorly understood. More broadly, this raises a practical question regarding the utility of genomic samples, and the extent to which increased sampling intensity would lead to improved detection of viral importation through phylogeography. This question motivated the work presented in Chapter 4, where I used a combination of analytical and simulation approaches to develop a mechanistic understanding of how the undersampling of local infections leads to the underestimation of the number of viral importation events. Importantly, I showed that the probability of detecting a viral importation event depends on not only the sampling proportion at the recipient location, but also the underlying migration rate and local transmission conditions. These findings have critical implications for the use of phylogeographic estimates in public health decision-making - particularly when evaluating containment strategies aimed at limiting importation or quantifying the relative contribution of viral importation from different source regions to inform more effective targeted border control.

However, this analysis addresses only one aspect of the problem. A key limitation of Chapter 4 lies in its assumption of complete sampling at the source of viral importation - a simplification that rarely holds in reality. In practice, both the source and recipient locations of viral importation are likely to be subject to varying degrees of undersampling due to asymptomatic infections (40-42), underreporting of cases (43, 44), and limited

sequencing capacity (45, 46). More importantly, this highlights a broader and more complex challenge: how to accurately measure the spatial movement of an emerging pathogen using phylogeography given substantially heterogeneous sampling.

Improving sampling design to inform phylogeographic inference

The impact of heterogeneous genomic sampling is a well-known problem in phylogeography, with implications for the study of many pathogens beyond SARS-CoV-2. Research efforts to address this problem have largely followed three directions: 1) the development of models and inference frameworks that are less sensitive to sampling biases, such as those based on structured coalescent (47-49), 2) the incorporation of external data streams, including reported case counts, mobility data, and travel histories, to inform or augment phylogeographic reconstructions (50-52), and 3) the use of sampling or downsampling strategies to correct for under- or over-representation of infections in available genomic data (53-55). Among these, the third approach has grown in popularity in particular, owing to its practical simplicity and the growing availability of pathogen genomes since the COVID-19 pandemic, as well as recent advances in high-performance computing that allow the analysis of increasingly large datasets. Although a number of sampling strategies have been proposed and applied to empirical studies, systematic evaluations of their effectiveness remain limited, and the development of a generalised, context-aware sampling strategy has so far proved elusive. This is due in part to the complexity of sampling design which often involves many parameters, and the need to tailor each approach to specific outbreak contexts and research objectives.

To advance research in this direction, in Chapter 5 I introduced SOPHI ("*Sandbox for Optimising genomic sampling for PHylogeographic Inference*") - a simulation-based evaluation framework designed to facilitate systematic exploration of different sampling

strategies and their impact on the accuracy of phylogeographic inference. Using SOPHI, I investigated how undersampling of infections at both the source and recipient locations affects the detection of viral importation events, addressing a key limitation of the work in Chapter 4. Additionally, I evaluated and compared the performance of three commonly used sampling schemes in the context of identifying the source of early importation events, under two contrasting outbreak scenarios with different sampling biases. Notably, results from these two case studies provided important insights into the mechanisms by which heterogeneous sampling gives rise to biased phylogeographic estimates of pathogen movement, laying the groundwork for developing more robust and effective mitigation strategies.

SOPHI as an optimisation framework

Beyond its utility as a practical framework for evaluating different sampling designs (as demonstrated in the two case studies presented in Chapter 5), the development of SOPHI was also motivated by its potential as a training environment for data-driven optimisation methods. The large combinatorial space of possible sampling designs and outbreak parameters, coupled with SOPHI's ability to provide real-time feedback on inference performance through a range of evaluation metrics, makes this a natural setting for the application of machine learning approaches. Optimisation methods such as active learning (28), Bayesian optimisation (56), and reinforcement learning (57) can be used to guide the search for sampling strategies that maximise inference accuracy under specific outbreak conditions and research objectives. This approach extends naturally to the work in Chapter 3 as well, where the goal was to optimise the allocation of testing resources to support early risk assessment during an outbreak. As ongoing work introduces greater model complexities and more realistic dynamics and logistic constraints, exhaustive

search for the optimal policies becomes increasingly impractical, while simple heuristics often fail to capture the nuances of key dynamics. In such contexts, optimisation techniques based on simulated data offer a promising and scalable alternative for the systematic exploration of the solution space (58, 59).

At a broader level, this approach of reframing experimental design in epidemiology and decision-making in public health as an optimisation problem points towards a more general framework for directing future research efforts in a way that maximises potential impact. This perspective is exemplified in the problem of optimising test allocation as considered in Chapter 3, where the effectiveness of any given allocation policy depends on a number of interacting processes including the underlying transmission dynamics, human mobility, and logistical constraints such as testing capacity and uptake. A simulation-based optimisation approach requires integrating these processes within a unified environment, leveraging existing domain-specific models that reflect our current best understanding. Within such an environment, it is then possible to not only identify optimal strategies given specific objectives and model assumptions, but also to map how the landscape of optimal solutions changes with different parameterisations of the underlying models. In the context of test allocation, this could reveal insights such as: How does the optimal policy change as a function of transmission characteristics of the pathogen? How sensitive is policy design to seasonal variations in human mobility patterns? Are there regions in transmission parameter space, and therefore for certain pathogens, where even the best-performing policy is unlikely to be effective? Answers to these questions could offer a principled way to prioritise future empirical investigations, by identifying parameters and processes with the most influence on the optimal solutions and therefore warrant more precise measurements and modelling (60, 61).

However, it is important to also acknowledge the risks of applying such an approach to epidemiology and specifically public health, due to its potential to introduce and amplify societal biases. Historically, disease data and empirical studies of past outbreaks systematically underrepresent certain socio-demographic groups (e.g., rural communities, ethnic minorities) as a result of unequal access to healthcare, limited resources for outbreak investigation in remote areas, as well as institutional and systemic prejudice (62-65). Policy models trained on simulations generated by models built from such data and the insights derived from them therefore risk entrenching or even exacerbating pre-existing biases (e.g., neglecting more nuanced heterogeneities associated with marginalised group (66, 67)), leading to policy recommendations that are likely to result in inequitable or even adverse public health outcomes. Such biases can also arise in the absence of biased data, through mis-specified or overly simplistic objective functions. For example, if the overall objective of a vaccination campaign is to maximise the number of individuals vaccinated, the optimal policy might prioritise affluent, well-served populations with high uptake-rate, while neglecting rural communities with lower uptake but potentially greater vulnerability due to limited access to healthcare (68, 69). To mitigate these risks, a number of safeguards are possible: 1) incorporate fairness-aware performance metrics (e.g., worst-group error or demographic-parity gap) (70-72) as explicit constraints or secondary objectives in a multi-objective optimisation framework; 2) co-develop reward functions and operational constraints with public health practitioners and representatives from affected communities; and 3) routinely review and audit simulation models and learned policies for disparate impact during internal validation and post-deployment evaluation. These guardrails can help ensure that the benefits of such optimisation approaches can be leveraged to their full

potential while minimising the risk of reinforcing existing inequities and social disparities.

Advances in AI and implications for public health

Recent advances in artificial intelligence have made the development of such simulation-based optimisation frameworks increasingly feasible (73). Data-driven modelling approaches, such as GNNs (29-31) and time-series foundation models (74, 75), offer the capacity to learn complex, non-linear relationships from large empirical or simulated datasets, while delivering substantially greater computational efficiency and scalability than traditional modelling approaches; equally important, recent advancements in optimisation techniques (28, 56, 57) provide systematic frameworks for effectively navigating high-dimensional decision spaces, guided by reward functions tailored to predefined research or public health objectives (76-78). Meanwhile, the emergence of large language models (79, 80) and multi-modal AI agents (81-83) capable of task planning, tool use, and autonomous execution introduces the possibility of further automating the process of hypothesis generation and validation within such frameworks.

These advances also signal a broader, more fundamental transformation in public health, where policymakers increasingly rely on insights and recommendations from AI-driven systems to navigate difficult policy decisions involving competing priorities and complex trade-offs. During the COVID-19 pandemic, public health decision-making was likened to the infamous trolley problem (84, 85), where the moral dilemma of pulling a lever to divert a runaway trolley mirrors that of choosing one intervention over the other, each with varying effects on different parts of society. Yet this analogy arguably oversimplifies the reality of public health decision-making, where outcomes are often uncertain, non-binary, and lack clear causal pathways. This could change, however, as

advances in AI continue to equip us with increasingly powerful tools capable of modelling, predicting, and designing interventions that can shape outbreak trajectories with greater precision and certainty.

As our technical capabilities expand, it is also vital that we do not overlook the fundamentally human nature of public health decision-making. In attempts to reframe policy design as tractable optimisation problems, there is a risk of technocratic overreach - where nuanced considerations that are inherently political, social, and ethical are oversimplified or ignored entirely in the name of progress. Despite the many advances that AI promises, we must recognise their limitations and contend with the possibility that not all aspects of public health can, or perhaps should, be reduced to quantifiable metrics and optimised. Additionally, as AI assumes an increasingly central role in public health, it is crucial that access to these technologies is democratised, so that their benefits are not limited to only those with the resources to build them - who invariably are least exposed to the threats of emerging infectious diseases. With great power comes greater responsibility – the challenge ahead therefore lies not only in building decision-support systems that leverage these technological advances, but also in establishing effective governance to ensure that these systems are used responsibly, transparently, and in service of equitable public health outcomes.

References

1. COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk (2020) 'An integrated national scale SARS-CoV-2 genomic surveillance network', *Lancet Microbe*, 1(3), pp. e99–e100.
2. Brainard, J., Lake, I.R., Morbey, R.A., Jones, N.R., Elliot, A.J. and Hunter, P.R. (2023) 'Comparison of surveillance systems for monitoring COVID-19 in England: a retrospective observational study', *The Lancet. Public health*, 8(11), pp. e850-e858.
3. Oude Munnink, B.B., Worp, N., Nieuwenhuijse, D.F., Sikkema, R.S., Haagmans, B., Fouchier, R.A.M. and Koopmans, M. (2021) 'The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology', *Nature Medicine*, 27(9), pp. 1518–1524.
4. Tosta, S., Moreno, K., Schuab, G., Fonseca, V., Segovia, F.M.C., Kashima, S., Elias, M.C., Sampaio, S.C., Ciccozzi, M., Alcantara, L.C.J., Slavov, S.N., Lourenço, J., Cella, E. and Giovanetti, M. (2023) 'Global SARS-CoV-2 genomic surveillance: What we have learned (so far)', *Infection, Genetics and Evolution*, 108, 105405.
5. Sharan, M., Vijay, D., Yadav, J.P., Bedi, J.S. and Dhaka, P. (2023) 'Surveillance and response strategies for zoonotic diseases: a comprehensive review', *Science in One Health*, 2, 100050.
6. Bohl, J.A., Lay, S., Chea, S., Ahyong, V., Parker, D.M., Gallagher, S., Fintzi, J., Man, S., Ponce, A., Sreng, S., Kong, D., Oliveira, F., Kalantar, K., Tan, M., Fahsbender, L., Sheu, J., Neff, N., Detweiler, A.M., Yek, C., Ly, S., Sath, R., Huch, C., Kry, H., Leang, R., Huy, R., Lon, C., Tato, C.M., DeRisi, J.L. and Manning, J.E. (2022) 'Discovering disease-causing pathogens in resource-scarce Southeast Asia using a global metagenomic pathogen monitoring system', *Proceedings of the National Academy of Sciences*, 119(11), p. e2115285119.
7. Gardy, J.L. and Loman, N.J. (2017) 'Towards a genomics-informed, real-time, global pathogen surveillance system', *Nature Reviews Genetics*, 19(1), pp. 9–20.
8. UK Health Security Agency (2024) *UKHSA Pathogen Genomics Strategy*. GOV.UK. Available at https://www.gov.uk/government/publications/ukhsa-pathogen-genomics-strategy (Accessed: 12 April 2025).
9. World Health Organization (no date) *International Pathogen Surveillance Network (IPSN)*. Available at https://www.who.int/initiatives/international-pathogen-surveillance-network (Accessed: 13 April 2025).
10. European Commission (2024) *Launching GLOWACON: A global initiative for wastewater surveillance for, Public Health. Available at* https://health.ec.europa.eu/latest-updates/launching-glowacon-global-initiative-wastewater-surveillance-public-health-2024-03-21_en (Accessed: 13 April 2025).
11. Chang, S., Pierson, E., Koh, P.W., Gerardin, J., Redbird, B., Grusky, D. and Leskovec, J. (2020) 'Mobility network models of COVID-19 explain inequities and inform reopening', *Nature*, 589(7840), pp. 82–87.
12. Gutierrez, B., Tsui, J.L.-H., Pullano, G., Mazzoli, M., Gangavarapu, K., Inward, R.P.D., Bajaj, S., Evans Pena, R., Busch-Moreno, S., Suchard, M.A., Pybus, O.G., Dunner, A., Puentes, R., Ayala, S., Fernandez, J., Araos, R., Ferres, L., Colizza, V. and Kraemer, M.U.G. (2024) 'Routes of importation and spatial dynamics of SARS-CoV-2 variants during localized interventions in Chile', *PNAS nexus*, 3(11), p. gae483.
13. Badr, H.S., Du, H., Marshall, M., Dong, E., Squire, M.M. and Gardner, L.M. (2020) 'Association between mobility patterns and COVID-19 transmission in the USA: a

mathematical modelling study', *The Lancet. Infectious diseases*, 20(11), pp. 1247-1254.
14. Kraemer, M.U.G., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D.M., Open COVID-19 Data Working Group, du Plessis, L., Faria, N.R., Li, R., Hanage, W.P., Brownstein, J.S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O.G. and Scarpino, S.V. (2020) 'The effect of human mobility and control measures on the COVID-19 epidemic in China', *Science*, 368(6490), pp. 493–497.
15. Tegally, H., Wilkinson, E., Tsui, J.L., Moir, M., Martin, D., Brito, A.F., Giovanetti, M., Khan, K., Huber, C., Bogoch, I.I., San, J.E., Poongavanan, J., Xavier, J.S., Candido, D.D.S., Romero, F., Baxter, C., Pybus, O.G., Lessells, R.J., Faria, N.R., Kraemer, M.U.G. and de Oliveira, T. (2023) 'Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern', *Cell*, 186(15), pp. 3277-3290.e16
16. Klamser, P.P., Zachariae, A., Maier, B.F., Baranov, O., Jongen, C., Schlosser, F. and Brockmann, D. (2024) 'Inferring country-specific import risk of diseases from the world air transportation network', *PLOS Computational Biology*, 20(1), p. e1011775.
17. du Plessis, L., McCrone, J.T., Zarebski, A.E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T.R., Faria, N.R., Jackson, B., Loman, N.J., O'Toole, Á., Nicholls, S.M., Parag, K.V., Scher, E., Vasylyeva, T.I., Volz, E.M., Watts, A., Bogoch, I.I., Khan, K., COVID-19 Genomics UK (COG-UK) Consortium, Aanensen, D.M., Kraemer, M.U.G., Rambaut, A. and Pybus, O.G. (2021) 'Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK', *Science*, 371(6530), pp. 708–712.
18. Curran-Sebastian, J., Andersen, F.M. and Bhatt, S. (2025) 'Modelling the stochastic importation dynamics and establishment of novel pathogenic strains using a general branching processes framework', *Mathematical biosciences*, 380, p. 109352.
19. Russell, T.W., Wu, J.T., Clifford, S., Edmunds, W.J., Kucharski, A.J. and Jit, M. (2021) 'Effect of internationally imported cases on internal spread of COVID-19: a mathematical modelling study', *The Lancet. Public health*, 6(1), pp. e12-e20.
20. Han, X., Xu, Y., Fan, L., Huang, Y., Xu, M. and Gao, S. (2021) 'Quantifying COVID-19 importation risk in a dynamic network of domestic cities and international countries', *Proceedings of the National Academy of Sciences*, 118(31), p. e2100201118.
21. Stenseth, N.C., Schlatte, R., Liu, X., Pielke, R., Jr, Li, R., Chen, B., Bjørnstad, O.N., Kusnezov, D., Gao, G.F., Fraser, C., Whittington, J.D., Bai, Y., Deng, K., Gong, P., Guan, D., Xiao, Y., Xu, B. and Johnsen, E.B. (2023) 'How to avoid a local epidemic becoming a global pandemic', *Proceedings of the National Academy of Sciences of the United States of America*, 120(10), p. e2220080120.
22. Zhong, L., Diagne, M., Wang, W. and Gao, J. (2021) 'Country distancing increase reveals the effectiveness of travel restrictions in stopping COVID-19 transmission', *Communications Physics*, 4(1), pp. 1–12.
23. Chinazzi, M., Davis, J.T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Pastore y Piontti, A., Mu, K., Rossi, L., Sun, K., Viboud, C., Xiong, X., Yu, H., Elizabeth Halloran, M., LonginiJr, I.M. and Vespignani, A. (2020) 'The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak', *Science*, 368(6489), pp. 395-400.
24. Wells, C.R., Sah, P., Moghadas, S.M., Pandey, A., Shoukat, A., Wang, Y., Wang, Z., Meyers, L.A., Singer, B.H. and Galvani, A.P. (2020) 'Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus

outbreak', *Proceedings of the National Academy of Sciences*, 117(13), pp. 7504–7509.
25. Zhang, D., Ge, Y., Wang, J., Liu, H., Zhang, W.B., Wu, X., B M Heuvelink, G., Wu, C., Yang, J., Ruktanonchai, N.W., Qader, S.H., Ruktanonchai, C.W., Cleary, E., Yao, Y., Liu, J., Nnanatu, C.C., Wesolowski, A., Cummings, D.A.T., Tatem, A. J., and Lai, S. (2024) 'Optimizing the detection of emerging infections using mobility-based spatial sampling', *International Journal of Applied Earth Observation and Geoinformation*, 131, p. 103949.
26. Spott, R., Pletz, M.W., Fleischmann-Struzek, C., Kimmig, A., Hadlich, C., Hauert, M., Lohde, M., Jundzill, M., Marquet, M., Dickmann, P., Schüchner, R., Hölzer, M., Kühnert, D. and Brandt, C. (2024) 'Exploring the Spatial Distribution of Persistent SARS-CoV-2 Mutations - Leveraging mobility data for targeted sampling', *eLife*, 13, RP94045.
27. Oliveira, J.F., Alencar, A.L., Mcls, C., Vasconcelos, A.O., Cunha, G.G., Miranda, R.B., Fmhs, F., Silva, C., Gustani-Buss, E., Khouri, R., Cerqueira-Silva, T., Landau, L., Barral-Netto, M. and Ramos, P.I.P. (2024) 'Human mobility patterns in Brazil to inform sampling sites for early pathogen detection and routes of spread: a network modelling and validation study', *The Lancet. Digital health*, 6(8), pp. e570-e579.
28. Settles, B. (2012) *Active Learning*. Springer Nature Switzerland.
29. Kipf, T.N. and Welling, M. (2017) 'Semi-supervised classification with graph convolutional networks', *arXiv*. Available at: https://doi.org/10.48550/arXiv.1609.02907 (Accessed: 24 May 2024).
30. Liu, Z., Wan, G., Prakash, B.A., Lau, M.S.Y. and Jin, W. (2024) 'A Review of Graph Neural Networks in Epidemic Modeling', *arXiv*. Available at: http://arxiv.org/abs/2403.19852 (Accessed: 24 April 2025).
31. Kapoor, A., Ben, X., Liu, L., Perozzi, B., Barnes, M., Blais, M. and O'Banion, S. (2020) 'Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks', *arXiv*. Available at: http://arxiv.org/abs/2007.03113 (Accessed: 24 April 2025).
32. Jenson, D., Navott, J., Zhang, M., Sharma, M., Semenova, E. and Flaxman, S. (2024) 'Transformer Neural Processes -- Kernel Regression', *arXiv*. Available at: http://arxiv.org/abs/2411.12502 (Accessed: 25 April 2025).
33. Metcalf, C.J., Farrar, J., Cutts, F.T., Basta, N.E., Graham, A.L., Lessler, J., Ferguson, N.M., Burke, D.S. and Grenfell, B.T. (2016) 'Use of serological surveys to generate key insights into the changing global landscape of infectious disease', *Lancet (London, England)*, 388(10045), pp. 728-730.
34. Mina, M.J., Metcalf, C.J.E., McDermott, A.B., Douek, D.C., Farrar, J. and Grenfell, B.T. (2020) 'A Global lmmunological Observatory to meet a time of pandemics', *eLife*, 9, p. e58989.
35. St-Onge, G., Davis, J.T., Hébert-Dufresne, L., Allard, A., Urbinati, A., Scarpino, S.V., Chinazzi, M. and Vespignani, A. (2025) 'Pandemic monitoring with global aircraft-based wastewater surveillance networks', *Nature Medicine*, 31(3), pp. 788–796.
36. Li, J., Hosegood, I., Powell, D., Tscharke, B., Lawler, J., Thomas, K.V. and Mueller, J.F. (2023) 'A global aircraft-based wastewater genomic surveillance network for early warning of future pandemics', *The Lancet. Global health*, 11(5), pp. e791-e795.
37. Nadeau, S.A., Vaughan, T.G., Scire, J., Huisman, J.S. and Stadler, T. (2021) 'The origin and early spread of SARS-CoV-2 in Europe', *Proceedings of the National Academy of Sciences*, 118(9), p. e2012008118.

38. Murall, C.L., Fournier, E., Galvez, J.H., N'Guessan, A., Reiling, S.J., Quirion, P.-O., Naderi, S., Roy, A.-M., Chen, S.-H., Stretenowich, P., Bourgey, M., Bujold, D., Gregoire, R., Lepage, P., St-Cyr, J., Willet, P., Dion, R., Charest, H., Lathrop, M., Roger, M., Bourque, G., Ragoussis, J., Shapiro, B.J. and Moreira, S. (2021) 'A small number of early introductions seeded widespread transmission of SARS-CoV-2 in Québec, Canada', *Genome medicine*, 13(1), p. 169.
39. McLaughlin, A., Montoya, V., Miller, R.L., Mordecai, G.J., Canadian COVID-19 Genomics Network (CanCOGen) Consortium, Worobey, M., Poon, A.F.Y. and Joy, J.B. (2022) 'Genomic epidemiology of the first two waves of SARS-CoV-2 in Canada', *eLife*, 11, p. e73896.
40. Sah, P., Fitzpatrick, M.C., Zimmer, C.F., Abdollahi, E., Juden-Kelly, L., Moghadas, S.M., Singer, B.H. and Galvani, A.P. (2021) 'Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis', *Proceedings of the National Academy of Sciences*, 118(34), p. e2109229118.
41. Ma, Q., Liu, J., Liu, Q., Kang, L., Liu, R., Jing, W., Wu, Y. and Liu, M. (2021) 'Global Percentage of Asymptomatic SARS-CoV-2 Infections Among the Tested Population and Individuals With Confirmed COVID-19 Diagnosis: A Systematic Review and Meta-analysis', *JAMA network open*, 4(12), p. e2137257.
42. Subramanian, R., He, Q. and Pascual, M. (2021) 'Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity', *Proceedings of the National Academy of Sciences*, 118(9), p. e2019716118.
43. Havers, F.P., Reed, C., Lim, T., Montgomery, J.M., Klena, J.D., Hall, A.J., Fry, A.M., Cannon, D.L., Chiang, C.F., Gibbons, A., Krapiunaya, I., Morales-Betoulle, M., Roguski, K., Rasheed, M.A.U., Freeman, B., Lester, S., Mills, L., Carroll, D.S., Owen, S.M., Johnson, J.A., … Thornburg, N.J. (2020) 'Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23–May 12, 2020', *JAMA Internal Medicine* [Online ahead of print]. Available at: https://doi.org/10.1001/jamainternmed.2020.4130.
44. Bergeri, I., Whelan, M.G., Ware, H., Subissi, L., Nardone, A., Lewis, H.C., Li, Z., Ma, X., Valenciano, M., Cheng, B., Al Ariqi, L., Rashidian, A., Okeibunor, J., Azim, T., Wijesinghe, P., Le, L.-V., Vaughan, A., Pebody, R., Vicari, A., Yan, T., Yanes-Lane, M., Cao, C., Clifton, D.A., Cheng, M.P., Papenburg, J., Buckeridge, D., Bobrovitz, N., Arora, R.K., Van Kerkhove, M.D. and Unity Studies Collaborator Group (2022) 'Global SARS-CoV-2 seroprevalence from January 2020 to April 2022: A systematic review and meta-analysis of standardized population-based studies', *PLOS Medicine*, 19(11), p. e1004107.
45. Brito, A.F., Semenova, E., Dudas, G., Hassler, G.W., Kalinich, C.C., Kraemer, M.U.G., Ho, J., Tegally, H., Githinji, G., Agoti, C.N., Matkin, L.E., Whittaker, C., Howden, B.P., Sintchenko, V., Zuckerman, N.S., Mor, O., Blankenship, H.M., de Oliveira, T., Lin, R.T.P., Siqueira, M.M., Resende, P.C., Vasconcelos, A.T.R., Spilki, F.R., Aguiar, R.S., Alexiev, I., Ivanov, I.N., Philipova, I., Carrington, C.V.F., Sahadeo, N.S.D., Branda, B., Gurry, C., Maurer-Stroh, S., Naidoo, D., von Eije, K.J., Perkins, M.D., van Kerkhove, M., Hill, S.C., Sabino, E.C., Pybus, O.G., Dye, C., Bhatt, S., Flaxman, S., Suchard, M.A., Grubaugh, N.D., Baele, G. and Faria, N.R. (2022) 'Global disparities in SARS-CoV-2 genomic surveillance', *Nature Communications*, 13(1), pp. 1–13.
46. Chen, Z., Azman, A.S., Chen, X., Zou, J., Tian, Y., Sun, R., Xu, X., Wu, Y., Lu, W., Ge, S., Zhao, Z., Yang, J., Leung, D.T., Domman, D.B. and Yu, H. (2022) 'Global

landscape of SARS-CoV-2 genomic surveillance and data sharing', *Nature Genetics*, 54(4), pp. 499–507.
47. De Maio, N., Wu, C.-H., O'Reilly, K.M. and Wilson, D. (2015) 'New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation', *PLOS Genetics*, 11(8), p. e1005421.
48. Vaughan, T.G., Kühnert, D., Popinga, A., Welch, D. and Drummond, A.J. (2014) 'Efficient Bayesian inference under the structured coalescent', *Bioinformatics (Oxford, England)*, 30(16), pp. 2272–2279.
49. Müller, N.F., Rasmussen, D.A. and Stadler, T. (2017) 'The Structured Coalescent and Its Approximations', *Molecular biology and evolution*, 34(11), pp. 2970-2981.
50. McCrone, J.T., Hill, V., Bajaj, S., Pena, R.E., Lambert, B.C., Inward, R., Bhatt, S., Volz, E., Ruis, C., Dellicour, S., Baele, G., Zarebski, A.E., Sadilek, A., Wu, N., Schneider, A., Ji, X., Raghwani, J., Jackson, B., Colquhoun, R., O'Toole, Á., Peacock, T.P., Twohig, K., Thelwall, S., Dabrera, G., Myers, R., Faria, N.R., Huber, C., Bogoch, I.I., Khan, K., du Plessis, L., Barrett, J.C., Aanensen, D.M., Barclay, W.S., Chand, M., Connor, T., Loman, N.J., Suchard, M.A., Pybus, O.G., Rambaut, A. and Kraemer, M.U.G. (2022) 'Context-specific emergence and growth of the SARS-CoV-2 Delta variant', *Nature*, 610(7930), pp. 154–160.
51. Lemey, P., Hong, S.L., Hill, V., Baele, G., Poletto, C., Colizza, V., O'Toole, Á., McCrone, J.T., Andersen, K.G., Worobey, M., Nelson, M.I., Rambaut, A. and Suchard, M.A. (2020) 'Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2', *Nature Communications*, 11(1), pp. 1–14.
52. Worobey, M., Pekar, J., Larsen, B.B., Nelson, M.I., Hill, V., Joy, J.B., Rambaut, A., Suchard, M.A., Wertheim, J.O. and Lemey, P. (2020) 'The emergence of SARS-CoV-2 in Europe and North America', *Science*, 370(6516), pp. 564-570.
53. Volz, E., Mishra, S., Chand, M., Barrett, J.C., Johnson, R., Geidelberg, L., Hinsley, W.R., Laydon, D.J., Dabrera, G., O'Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C.V., Boyd, O., Loman, N.J., McCrone, J.T., Gonçalves, S., Jorgensen, D., Myers, R., Hill, V., Jackson, D.K., Gaythorpe, K., Groves, N., Sillitoe, J., Kwiatkowski, D.P., Flaxman, S., Ratmann, O., Bhatt, S., Hopkins, S., Gandy, A., Rambaut, A. and Ferguson, N.M. (2021) 'Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England', *Nature*, 593(7858), pp. 266–269.
54. Alpert, T., Brito, A.F., Lasek-Nesselquist, E., Rothman, J., Valesano, A.L., MacKay, M.J., Petrone, M.E., Breban, M.I., Watkins, A.E., Vogels, C.B.F., Kalinich, C.C., Dellicour, S., Russell, A., Kelly, J.P., Shudt, M., Plitnick, J., Schneider, E., Fitzsimmons, W.J., Khullar, G., Metti, J., Dudley, J.T., Nash, M., Beaubier, N., Wang, J., Liu, C., Hui, P., Muyombwe, A., Downing, R., Razeq, J., Bart, S.M., Grills, A., Morrison, S.M., Murphy, S., Neal, C., Laszlo, E., Rennert, H., Cushing, M., Westblade, L., Velu, P., Craney, A., Cong, L., Peaper, D.R., Landry, M.L., Cook, P.W., Fauver, J.R., Mason, C.E., Lauring, A.S., St. George, K., MacCannell, D.R. and Grubaugh, N.D. (2021) 'Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States', *Cell*, 184(10), pp. 2595–2604.e13.
55. Castelán-Sánchez, H.G., Delaye, L., Inward, R.P.D., Dellicour, S., Gutierrez, B., Martinez de la Vina, N., Boukadida, C., Pybus, O.G., de Anda Jáuregui, G., Guzmán, P., Flores-Garrido, M., Fontanelli, Ó., Hernández Rosales, M., Meneses, A., Olmedo-Alvarez, G., Herrera-Estrella, A.H., Sánchez-Flores, A., Muñoz-Medina, J.E., Comas-García, A., Gómez-Gil, B., Zárate, S., Taboada, B., López, S., Arias, C.F., Kraemer, M.U.G., Lazcano, A. and Escalera Zamudio, M. (2023) 'Comparing the

evolutionary dynamics of predominant SARS-CoV-2 virus lineages co-circulating in Mexico', *eLife*, 12, e82069.
56. Garnett, R. (2023) *Bayesian Optimization*. Cambridge University Press.
57. Sutton, R.S. and Barto, A.G. (2018) *Reinforcement Learning, second edition: An Introduction*. MIT Press.
58. Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D., Donner, C., Fritz, L., Galperti, C., Huber, A., Keeling, J., Tsimpoukelli, M., Kay, J., Merle, A., Moret, J.-M., Noury, S., Pesamosca, F., Pfau, D., Sauter, O., Sommariva, C., Coda, S., Duval, B., Fasoli, A., Kohli, P., Kavukcuoglu, K., Hassabis, D. and Riedmiller, M. (2022) 'Magnetic control of tokamak plasmas through deep reinforcement learning', *Nature*, 602(7897), pp. 414–419.
59. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W. and Abbeel, P. (2017) 'Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World', *arXiv*. Available at: http://arxiv.org/abs/1703.06907 (Accessed: 28 April 2025).
60. Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C.P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., Marks, D., Ramsundar, B., Song, L., Sun, J., Tang, J., Veličković, P., Welling, M., Zhang, L., Coley, C.W., Bengio, Y. and Zitnik, M. (2023) 'Scientific discovery in the age of artificial intelligence', *Nature*, 620(7972), pp. 47–60.
61. Agrawal, A., McHale, J. and Oettl, A. (2024) 'Artificial intelligence and scientific discovery: a model of prioritized search', *Research Policy*, 53(5), 104989.

62. Henly, S., Tuli, G., Kluberg, S.A., Hawkins, J.B., Nguyen, Q.C., Anema, A., Maharana, A., Brownstein, J.S. and Nsoesie, E.O. (2017) 'Disparities in digital reporting of illness: a demographic and socioeconomic assessment', *Preventive Medicine*, 101, pp. 18–22.
63. Burström, B. and Tao, W. (2020) 'Social determinants of health and inequalities in COVID-19', *European journal of public health*, 30(4).
64. Tan, S.B., deSouza, P. and Raifman, M. (2022) 'Structural Racism and COVID-19 in the USA: a County-Level Empirical Analysis', *Journal of racial and ethnic health disparities*, 9(1).
65. Tizzoni, M., Nsoesie, E.O., Gauvin, L., Karsai, M., Perra, N. and Bansal, S. (2022) 'Addressing the socioeconomic divide in computational modeling for infectious diseases', *Nature Communications*, 13(1), pp. 1–7.
66. Jay, J., Bor, J., Nsoesie, E.O., Lipson, S.K., Jones, D.K., Galea, S. and Raifman, J. (2020) 'Neighbourhood income and physical distancing during the COVID-19 pandemic in the United States', *Nature Human Behaviour*, 4(12), pp. 1294–1302.
67. Gauvin, L., Bajardi, P., Pepe, E., Lake, B., Privitera, F. and Tizzoni, M. (2021) 'Socio-economic determinants of mobility responses during the first wave of COVID-19 in Italy: from provinces to neighbourhoods', *Journal of the Royal Society Interface*, 18(181).
68. Viswanath, K., Bekalu, M., Dhawan, D., Pinnamaneni, R., Lang, J. and McLoud, R. (2021) 'Individual and social determinants of COVID-19 vaccine uptake', *BMC public health*, 21(1), 818.
69. Manna, A., Koltai, J. and Karsai, M. (2024) 'Importance of social inequalities to contact patterns, vaccine uptake, and epidemic dynamics', *Nature Communications*, 15(1), pp. 1–11.

70. Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J. and Roth, A. (2017) 'Fairness in reinforcement learning', *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, Sydney, Australia, 6–11 August. Available at: https://dl.acm.org/doi/10.5555/3305381.3305548 (Accessed: 14 July 2025).
71. Ghosh, A., Genuit, L. and Reagan, M. (2021) 'Characterizing intersectional group fairness with worst-case comparisons', *Proceedings of the 2nd Affinity Group Workshop on Diversity in Artificial Intelligence (AIDBEI 2021)*, PMLR. Available at: https://proceedings.mlr.press/v142/ghosh21a/ghosh21a.pdf (Accessed: 14 July 2025).
72. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2012) 'Fairness through awareness', *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012)*, Cambridge, MA, USA, 14–17 January. Available at: https://dl.acm.org/doi/10.1145/2090236.2090255 (Accessed: 14 July 2025).
73. Kraemer, M.U.G., Tsui, J.L.-H., Chang, S.Y., Lytras, S., Khurana, M.P., Vanderslott, S., Bajaj, S., Scheidwasser, N., Curran-Sebastian, J.L., Semenova, E., Zhang, M., Unwin, H.J.T., Watson, O.J., Mills, C., Dasgupta, A., Ferretti, L., Scarpino, S.V., Koua, E., Morgan, O., Tegally, H., Paquet, U., Moutsianas, L., Fraser, C., Ferguson, N.M., Topol, E.J., Duchêne, D.A., Stadler, T., Kingori, P., Parker, M.J., Dominici, F., Shadbolt, N., Suchard, M.A., Ratmann, O., Flaxman, S., Holmes, E.C., Gomez-Rodriguez, M., Schölkopf, B., Donnelly, C.A., Pybus, O.G., Cauchemez, S. and Bhatt, S. (2025) 'Artificial intelligence for modelling infectious disease epidemics', *Nature*, 638(8051), pp. 623–635.
74. Kalahasti, S., Faucher, B., Wang, B., Ascione, C., Carbajal, R., Enault, M., Cassis, C.V., Launay, T., Guerrisi, C., Boëlle, P.-Y., Baldo, F. and Valdano, E. (2025) 'Foundation time series models for forecasting and policy evaluation in infectious disease epidemics', *medRxiv*. Available at: https://doi.org/10.1101/2025.02.24.25322795 (Accessed: 1 May 2025).
75. Das, A., Kong, W., Sen, R. and Zhou, Y. (2023) 'A decoder-only foundation model for time-series forecasting', *arXiv*. Available at: http://arxiv.org/abs/2310.10688 (Accessed: 7 May 2025).
76. Lei, B., Kirk, T.Q., Bhattacharya, A., Pati, D., Qian, X., Arroyave, R. and Mallick, B.K. (2021) 'Bayesian optimization with adaptive surrogate models for automated experimental design', *npj Computational Materials*, 7(1), pp. 1–12.
77. Li, D., Wang, Z., Chen, Y., Jiang, R., Ding, W. and Okumura, M. (2024) 'A Survey on Deep Active Learning: Recent Advances and New Frontiers', *arXiv*. Available at: http://arxiv.org/abs/2405.00334 (Accessed: 10 May 2025).
78. Chen, T., Chen, X., Chen, W., Heaton, H., Liu, J., Wang, Z. and Yin, W. (2021) 'Learning to Optimize: A Primer and A Benchmark', *arXiv*. Available at: http://arxiv.org/abs/2103.12828 (Accessed: 10 May 2025).
79. Lobentanzer, S., Feng, S., Bruderer, N., Maier, A., Wang, C., Baumbach, J., Abreu-Vicente, J., Krehl, N., Ma, Q., Lemberger, T. and Saez-Rodriguez, J. (2025) 'A platform for the biomedical application of large language models', *Nature Biotechnology*, 43(2), pp. 166–169.
80. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) 'Attention Is All You Need', *arXiv*. Available at: http://arxiv.org/abs/1706.03762 (Accessed: 7 May 2025).
81. Wang, H., He, Y., Coelho, P.P., Bucci, M., Nazir, A., Chen, B., Trinh, L., Zhang, S., Huang, K., Chandrasekar, V., Chung, D.C., Hao, M., Leote, A.C., Lee, Y., Li, B., Liu, T., Liu, J., Lopez, R., Lucas, T., Ma, M., Makarov, N., McGinnis, L., Peng, L., Ra, S., Scalia, G., Singh, A., Tao, L., Uehara, M., Wang, C., Wei, R., Copping, R.,

Rozenblatt-Rosen, O., Leskovec, J. and Regev, A. (2025) ‘SpatialAgent: An autonomous AI agent for spatial biology’, *bioRxiv*. Available at: https://doi.org/10.1101/2025.04.03.646459 (Accessed: 2 May 2025).
82. Xie, J., Chen, Z., Zhang, R., Wan, X. and Li, G. (2024) ‘Large Multimodal Agents: A Survey’, *arXiv*. Available at: http://arxiv.org/abs/2402.15116 (Accessed: 7 May 2025).
83. Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., Matias, Y., Carroll, A., Kulkarni, K., Tomasev, N., Guan, Y., Dhillon, V., Vaishnav, E.D., Lee, B., Costa, T.R.D., Penadés, J.R., Peltz, G., Xu, Y., Pawlosky, A., Karthikesalingam, A. and Natarajan, V. (2025) ‘Towards an AI co-scientist’, *arXiv*. Available at: http://arxiv.org/abs/2502.18864 (Accessed: 7 May 2025).
84. Navajas, J., Heduan, F.Á., Garbulsky, G., Tagliazucchi, E., Ariely, D. and Sigman, M. (2021) ‘Moral responses to the COVID-19 crisis’, *Royal Society Open Science*, 8(9), 210096.
85. Di Nucci, E. and Lillehammer, H. (2023) ‘A new trolley problem?’, in *The Trolley Problem*. Cambridge University Press, pp. 231–243.