THE UNIVERSITY
of EDINBURGH

From epidemics to pandemics: Elucidating the dynamics of Ebola Virus and SARS-CoV-2

Verity Hill

Supervisor: Prof. Andrew Rambaut

Institute of Evolutionary Biology

The University of Edinburgh

This dissertation is submitted for the degree of

Doctor of Philosophy

May 2022

Declaration

I declare that this thesis has been composed by myself. I confirm that the work has not been submitted, whole or in part, for any other degree or professional qualification. The work submitted is my own, except where explicitly indicated, and appropriate credit has been given in these cases.

Verity Hill

May 2022

The work presented in chapter 2 was jointly supervised by Prof. Guy Baele.

The work present in chapter 4 has been submitted for publication in *Virus Evolution* and is undergoing review. It can currently be found on BioRxiv as a preprint (Hill et al., 2022). I performed all of the analyses, with the exception of the birth-death skyline analysis, which was designed and run by Dr Louis Du Plessis. Dr Du Plessis also generated figure 4.4C in this chapter (3C in the original paper). I wrote the substantive part of this paper, although minor editing has been undertaken by co-authors.

The work present in chapter 5 is made up of my individual contributions to four separate papers (citations overleaf). The figures and text presented here were made and substantively written by me except where indicated. Dr JT McCrone inferred all of the introductions/transmission lineages for each section which I used as the input into all of my within-country analyses, other than for the Alpha variant analysis as there were no introductions to infer. Supervision was by Prof Andrew Rambaut, except where indicated.

The analysis on the first wave of SARS-CoV-2 in the UK was published as Du Plessis *et al.* (2021), and my input was jointly supervised by Prof. Oliver Pybus and Prof. Moritz Kraemer. Dr Louis Du Plessis also contributed to the inference of the transmission lineages. Dr. Christopher Ruis performed the distance analysis, reported in the paragraph that begins "There is also variation in the spatial range of individual UK transmission lineages". Dr Ruis also made the original version of 5.2B. The sections reporting these results in the original paper were a collaborative effort between Prof Kraemer, Dr Ruis and myself, although the text presented here was substantively written by me.

The D614G analysis was published as Volz et al. (2021b). My contribution to this paper was jointly supervised by Dr JT McCrone who helped me in designing the

"boom-bust" analysis. Prof. Andrew Rambaut contributed to the design of 5.3, and Prof. Oliver Pybus undertook minor text edits for the original paper.

The analysis of the spread of Alpha variant has been published as Kraemer et al. (2021). Dr Simon Dellicour provided an XML for running the continuous phylogeographic analysis and processed the raw output to provide a csv containing the start and end locations of every viral movement. Dr Dellicour also made the original published figure for Figure 5.6A and C although I have remade them here using my own original scripts. This work was jointly supervised by Prof Oliver Pybus and Prof Moritz Kraemer, who undertook minor text edits for the original paper.

The analysis of Delta variant spread has been preprinted on medrxiv as McCrone et al. (2021) and is currently in review at *Nature*. The analyses, text and figures here were all produced by me. Minor text editing was undertaken by Prof. Moritz Kraemer for the original paper, who also co-supervised this project.

The Introduction and Discussion used some concepts and some text (see below) from Hill et al. (2021). I developed and wrote this paper under the supervision of Prof. Moritz Kraemer, and made all of the figures. Ms Sumali Bajaj and Prof. Kraemer wrote the first draft of the section discussing the role of demography of the population and individual-level heterogeneity, but I have since substantively rewritten this during the writing and submission process, as well as editing it significantly for inclusion in this thesis. The section that Dr Chris Ruis wrote is not included in this thesis.

Specific sections of text taken from this paper in the Introduction are: the first paragraph of the introduction, the first two paragraphs of the “individual level heterogeneity” section, and the first two paragraphs of “national and international spread”. In the discussion, the sections “Sampling strategies and downsampling”, “Ethical Data sharing” and the section on “emerging data sources and methods” have been

adapted from that paper. Finally, Fig. 6.1 from the Discussion is originally from this paper.

* indicate co-first authorships

Hill, Verity, Louis Du Plessis, Thomas P. Peacock, Dinesh Aggarwal, Rachel Colquhoun, Alesandro M. Carabelli, Nicholas Ellaby, et al. 2022. “The Origins and Molecular Evolution of SARS-CoV-2 Lineage B.1.1.7 in the UK.” bioRxiv.

Louis du Plessis*, John T. McCrone*, Alexander E. Zarebski*, **Verity Hill***, Christopher Ruis*, Bernardo Gutierrez, Jayna Raghwani, et al. 2021. “Establishment and Lineage Dynamics of the SARS-CoV-2 Epidemic in the UK.” Science 371 (6530): 708–12.

Verity Hill, Christopher Ruis, Sumali Bajaj, Oliver G. Pybus, and Moritz U. G. Kraemer. 2021. “Progress and Challenges in Virus Genomic Epidemiology.” Trends in Parasitology 37 (12): 1038–49.

Erik Volz, **Hill, Verity**, John T. McCrone, Anna Price, David Jorgensen, Áine O’Toole, Joel Southgate, et al. 2021. “Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity.” Cell 184 (1): 64–75.e11.

Moritz U. G. Kraemer*, **Verity Hill***, Christopher Ruis*, Simon Dellicour*, Sumali Bajaj*, John T. McCrone, Guy Baele, et al. 2021. “Spatiotemporal Invasion Dynamics of SARS-CoV-2 Lineage B.1.1.7 Emergence.” Science 373 (6557): 889–95.

John T. McCrone*, **Verity Hill***, Sumali Bajaj*, Rosario Evans Pena*, Ben C. Lambert, Rhys Inward, Samir Bhatt, et al. 2021. “Context-Specific Emergence and Growth of the SARS-CoV-2 Delta Variant.” medRxiv, December, 2021.12.14.21267606.

Verity Hill, Christopher Ruis, Sumali Bajaj, Oliver G. Pybus, and Moritz U. G. Krae-

mer. 2021. “Progress and Challenges in Virus Genomic Epidemiology.” Trends in Parasitology 37 (12): 1038–49.

Acknowledgements

I have to start with Andrew. Thanks for meeting with me in Darwin's (RIP) all those years ago and taking a chance on some undergrad rambling about superspreading for ten minutes. I'm so grateful for the opportunities that being in your group has given me, from the travel to the great people I've met. Thank you for giving me the time, space and support to enable me to explore the field and come out the other side as an independent researcher that, as it turns out, just really loves maps and Skygrids. A really special quality of yours is that you never make people feel like idiots, even when they say really daft things, so thank you for sharing so much of your knowledge with me, even when I probably should have known it to begin with. Also, thanks for letting me know that fuchsia really is the only colour to make figures in.

To the members of Rambaut lab: JT, thanks for being my *mentor* and for letting me know that noses can freeze in the US when it drops below 20C, Emily for consistently rocking up with doughnuts and champagne at 3pm on a Friday, Rachel and Ben for putting up with my geography bugs, and Shawn for being a calming and supportive presence in the office. And generally thanks to you all (and everyone that's come through the lab in the last few years) for being such good fun and always up for a ridiculous chat. Sorry for being so disruptive.

Special thanks must go to Áine, my PhD sister, for putting up with me for all these years. I think you took the best jokes for your acknowledgements, like the *snake áine* that you are, and I hope you're *slightly smiling* now as you *udpate* your CV. Thanks for bullying me into actually enjoying data viz and accumulating far too many references and in-jokes. I'm excited to bemuse people at conferences with a language that sounds like English but turns out to be incomprehensible to those who haven't spent years drinking extensive amounts of red wine together, just to work out what those tannins should sound like.

To the brilliant scientists I've had the opportunity and pleasure to work with over the last four and a half years: Gytis Dudas, for putting up with endless questions on work you did 8 (?!) years ago; Guy Baele for always being on the other side of a slack message when I've broken BEAST for the third time that day and for enabling me to write my first paper; Oli Pybus for inspiring me to go into phylodynamics in the first place, and for helping my papers sound like actual science; Chris Ruis for being a great teammate on all those early COVID papers, I hope that we'll actually get to meet in person someday soon; Christophe Fraser for providing support and guidance on the ABSynthE model design; and Moritz Kraemer for pushing me to achieve my potential and produce some really good (I hope) and definitely really cool science. Thank you to Dhamari for taking me under your wing in Abuja - I really didn't expect to have so much fun, and you taught me so much about what Public Health actually looks like. It's an experience that has really shaped my thinking and will continue to impact my life.

Special mention must go to Ashleigh Griffin, for teaching me how to write (sorry for all those weird essays) and inspiring in me a love of evolutionary biology I didn't know I had. A PhD was never on the cards when I started undergrad, and I have you to thank now that I'm finishing one.

To my original biology crew, Jack and Asher, I'm excited to go for a "quiet drink" soon to chat more ridiculous science, and Megan, Elena and Christine (also my Salty Gals), thank you for keeping me sane with wine, undying support and trashy tv. I can't wait for the six of us to get together in new continents. Thanks to the TYD crew, Amy, Greg, Mary, Eevi, Megan and Lewis for putting up with me spamming weird memes, and generally showing me a ridiculous amount of fun. To Zack (Sack) for cooking me slightly weird pasta. Seriously though, thank you for being endlessly patient, kind and supportive, I couldn't hope for a better cheerleader and I'm so excited to see what comes next. Also thanks for helping me raise our large adult son, Stefan, and to you both for watching some truly awful films with me; and his funny little friend Koo - thanks to you all for helping me not take myself (or anything) too seriously. Finally, to all the other slackers in the DDH for truly spectacular shitchat - who knew any subject could be dissected to the degree that happens in that room.

And to my family! To Auntie Alison, the original Dr Hill, for paving the way for women like me. I'm sorry you didn't get to see me finish, but I'm glad you were around for long enough to know I was staying in science. I hope I can make you proud. To my parents, Jane and Simon, for providing endless delicious food and wine, Alex for providing excellent wine recommendations and loving support (sorry for talking about worms so much) and Oliver for putting up with my basic coding questions, that 1000 line monstrosity that turned out to be a cry for help and being a receiving point for all the coding memes that I understand. Turns out gentle mockery is a good way to support somebody through nearly ten years of higher education, who'd have thought - I wouldn't be the person I am without the four of you (take from that what you will).

Abstract

The advent of large-scale viral genomic sequencing has provided a rich source of data to explore the dynamics of infectious disease epidemics. In combination with the field of phylodynamics, which allows the inference of unobserved patterns from a relatively small sample of the true diversity of a virus, it has been used to great effect in the past decade. The most notable examples were during the West African Ebola Virus Disease (EVD) epidemic in 2013-2016 and the COVID-19 pandemic, still ongoing at the time of writing. The genomic datasets from these epidemics can be used to explore the evolution and transmission of viruses at different scales, from the effect of within-host evolution, to small-scale transmission networks, and national and international epidemic dynamics.

I begin with the national-scale analysis of the dynamics of Ebola virus in Sierra Leone. I developed a phylogeographic analysis in a generalised linear model framework, at two geographical resolutions and in two epochs. I found that the focus of viral movement shifts from the source location in the east, to the capital city in the west. This chapter explores why different locations were important for viral transmission on a national level, and how well the gravity model of infection applies to the spread of Ebola virus in Sierra Leone through time and across different geographical scales.

To address some of the issues in modelling a disease like EVD which has a high degree of superspreading, and to explore the impact of local contact networks, I

created ABSynthE (Agent Based Synthetic Epidemic). ABSynthE is a flexible agent based model, which simulates an EVD epidemic across the population of Sierra Leone. ABSynthE outputs coalescent phylogenies, which are then used to obtain transmission parameters at each contact level by fitting to results from chapter 1. I found that without any intervention, just under half of the population of Sierra Leone may have been infected, regardless of which district the epidemic began in.

There are now well over 8 million genomic sequences of SARS-CoV-2 available for analysis. Within the UK, the sampling is especially dense, allowing detailed epidemiological and phylodynamic analyses to be undertaken. In chapter 3, I explore the origins of the Alpha variant, the first variant of concern, which arose in South East England. I characterise the long ancestral branch, and find that it has a higher evolutionary rate compared to the background and Alpha clades, as well as a single intermediate sequence. I investigate the branches ancestral to the other variants of concern, and explore their mutational profiles, finding that Beta, Gamma and Omicron (but not Delta) have evidence for evolving in a similar manner to Alpha. I explore three different hypotheses for what this manner may be, and conclude that the most likely option is that they evolved within a persistently-infected, but not necessarily immunocompromised, host.

Finally, I use the rich UK genomic SARS-CoV-2 dataset to elucidate the dynamics of the first wave of infection in early 2020, including the emergence of the D614G mutation; the Alpha wave, spreading from Kent and London to the rest of the country in late 2020 and early 2021; and the Delta wave, introduced into multiple regions, but mostly spreading from the North West of England in early 2021. I compare these waves, especially in terms of spreading from multiple introductions versus a single origin; and in the context of tightening or loosening non-pharmaceutical interventions.

Lay Summary

Viruses mutate as they replicate. That is, the letters that make up their genome can change as they make mistakes when they are being copied. When a genome is sequenced, i.e. when we record the letters, or bases, of the virus inside a person, we can compare it to other genomes and make a family tree of these viruses, known as a phylogeny. At the same time, we can use information about where and when the viruses were sampled, combined with mutations, to explore how viruses spread in space and time. This is a part of the field of phylodynamics.

In this thesis, I have explored how two important viruses, Ebola virus and SARS-CoV-2, have spread on small and large scales by applying phylodynamic methods to extensive genome sequence databases. I looked at how Ebola virus spread across Sierra Leone, the worst affected country in the largest Ebola virus epidemic to date. I found that most virus transmission was focused around where the first cases were found in the far east of the country on the border with Guinea, before moving west to the large and densely populated capital city, Freetown, later on in the epidemic. I then used the results from this analysis to provide data for a simulation, called ABSynthE (Agent Based Synthetic Epidemic). This programme works by simulating all of the contacts of every person in Sierra Leone, as well as how their specific infection progresses. This enables me to include differences between individuals in the model, which traditional epidemiological models tend not to do, as they group

individuals together. This simulator is designed to re-run the tape of the epidemic in Sierra Leone so that it is possible to see what would have happened under different conditions. When I re-ran the epidemic without any interventions, I found that Ebola virus could have infected just under half the population before dying out. I also tried simulating the epidemic starting in different areas of the country, and found that it did not make much difference to the overall patterns of where the epidemic started in Sierra Leone.

SARS-CoV-2 has evolved in unexpected ways, especially with the emergence of variants of concern. These are groups of viruses with far more mutations than expected and are more transmissible, virulent or immune-evasive than the virus which originally emerged from China. I explored how the Alpha variant could have evolved by looking at how quickly it evolved before it began spreading between people in the general population of the UK. I also investigated the other variants of concern, for example Omicron, in the same way, and found that Delta variant has different evolutionary signatures to the others. I examined different options for where Alpha variant could have been evolving unobserved, and concluded that the Alpha variant is most likely to have evolved in a person who had a long-term infection. Finally, I used the genomic database to explore and compare the trends in the spread across the UK of each of the major waves of SARS-CoV-2 in 2020 and 2021 and how they interacted with restrictions. I found that major cities and different levels of restrictions were important to the spread of SARS-CoV-2 in the UK.

Using methods and analyses such as these while epidemics are ongoing can help provide another source of information on which to base public health decisions, and examine the effect of different interventions on the spread and evolution of viruses.

Table of contents

List of figures **xix**

List of tables **xxii**

Nomenclature **xxiii**

1 Introduction **1**

1.1 Overview 1

1.2 History of viral genomic epidemiology for public health 3

1.2.1 HIV-1 M 5

1.2.2 SARS-CoV 2002-2003 7

1.2.3 H1N1 pandemic 2009 9

1.2.4 MERS-CoV in Saudi Arabia 2012-present 10

1.2.5 Ebola Virus in West Africa 2013-2016 11

1.2.6 Poliovirus transmission in 2014 14

1.2.7 Zika virus in the Americas 2015-2016 14

1.2.8 Ebola virus in the DRC 2018-2020 15

1.2.9 SARS-CoV-2 pandemic 2019-present 16
1.3 What is phylodynamics? 19
1.4 The scales of viral dynamics and evolution 22
1.4.1 Within-host evolution 23
1.4.2 Transmission between individuals on a local scale 25
1.4.3 National and International transmission 27
1.5 Aims 31

2 The establishment and maintenance of Ebola virus transmission in Sierra Leone 35

2.1 Introduction 37
2.2 Methods 39
2.2.1 Dataset 39
2.2.2 Phylogenetic analysis 40
2.3 Results 42
2.3.1 Exponential growth and establishment 43
2.3.2 Epidemic maintenance 52
2.4 Discussion 55

3 Agent Based Synthetic Epidemic 61

3.1 Introduction 63
3.2 Model design 66
3.2.1 Contact structure and infection parameters 67

3.2.2 Transmission parameters 69
3.2.3 Seeding the epidemic 69
3.2.4 Infection algorithm and coalescent phylogeny generation . . 71
3.2.5 Fitting 74
3.3 Results 79
3.3.1 Base case 79
3.3.2 Application: the importance of starting district 82
3.4 Discussion 85
3.5 Supplementary tables 89

4 The origins and molecular evolution of the SARS-CoV-2 lineage B.1.1.7 in the UK 91

4.1 Introduction 93
4.2 Methods 96
4.2.1 Genomic dataset 96
4.2.2 Evolutionary rate calculation 97
4.2.3 Growth rate calculations 98
4.2.4 Sequencing proportion 100
4.2.5 Rates of evolution in chronically-infected individuals 100
4.2.6 Other variant analyses 101
4.3 Results 103
4.3.1 Characterising the ancestral branch of B.1.1.7 103

4.3.2 Early growth rate of B.1.1.7 in the UK and interaction with November lockdown in England 108
4.3.3 Other variants of concern 111
4.4 Discussion 116
4.5 Supplementary tables 125

5 The spatial dynamics of SARS-CoV-2 in the UK 128

5.1 Introduction 130
5.2 Methods 132
5.2.1 Transmission lineage description and introduction assignment 132
5.2.2 Genomic data 134
5.2.3 Geographical metadata 134
5.2.4 Growth rate of D614G lineages 136
5.2.5 Within-UK spatial dynamics of Alpha and Delta variants . . . 137
5.3 Results 139
5.3.1 First wave and D614G lineages 139
5.3.2 Alpha variant 146
5.3.3 Delta variant 152
5.4 Discussion 158

6 Discussion 163

6.1 Ebola virus versus SARS-CoV-2, a tale of two viruses 166
6.2 The future of genomic epidemiology and phylodynamics in public health 171

6.2.1 Practical issues for integrating non-genomic data into genomic analyses 172

6.2.2 The generation of representative datasets 174

6.2.3 Ethical data sharing 176

6.2.4 Emerging data sources and methods 177

6.3 Concluding remarks 179

References **181**

List of figures

1.1 Timeline showing major public health crises since 2000 5

1.2 Scales of transmission and evolution of viruses. 22

1.3 Geographical spread of sequence data used in this thesis. 32

2.1 Movements of Ebola Virus across Sierra Leone during the emergence of the epidemic 44

2.2 Results of the phylogeographic Generalised Linear Model for Ebola virus transmission on a chiefdom level 46

2.3 Results of the phylogeographic Generalised Linear Model for Ebola virus transmission on a district level 49

2.4 Choropleth map of the values of three asymmetric district-level predictors 51

2.5 Movements of Ebola virus across Sierra Leone in the maintenance and decline of the epidemic 53

3.1 Schematic showing the infection algorithm for a single infected individual in ABSynthE. 71

3.2 Schematic of phylogenetic simulation in ABSynthE 74

3.3 Results of fitting procedure. 77

3.4 Base case of infection based on 100 iterations of ABSynthE using parameters from fitting procedure 81

3.5 Results of simulating the epidemic in Sierra Leone starting from different districts. 84

4.1 Phylogenetic characteristics of B.1.1.7 lineage 104

4.2 Scenarios and timings for the intermediate sequence 107

4.3 Effective population size of B.1.1.7 compared to various other lineages 109

4.4 Coalescent and birth-death growth estimates of B.1.1.7 and background lineages 111

4.5 Phylogenetic characteristics of the Omicron variant 112

4.6 Regressions of root-to-tip genetic distance against time for other variants of concern 113

4.7 Comparing mutation profiles between variants of concern 114

4.8 Results from longitudinally-sampled patients from across seven papers 121

5.1 Timeline of major waves of SARS-CoV-2 infection in the UK 131

5.2 Spatial distribution of the first wave of infections in the UK. 140

5.3 Geographic and temporal distribution of UK phylogenetic clusters in the first wave 143

5.4 Estimated TMRCA and transition times for each 614D or 614G lineage over 40 sequences 145

5.5 Spatial and phylogenetic description of the Alpha variant 147

5.6 Spatial emergence dynamics of the Alpha variant in England 150

5.7 Correlations of imports of the Alpha and SGTF-positive cases per UTLA 151

5.8 Introductions of the Delta variant into each UTLA 153

5.9 Viral lineage movements of the seven largest Delta variant transmission lineages 155

5.10 Effect of NPIs on distance of Delta variant transmission lineage movement 156

5.11 Temporal and spatial growth of seven largest Delta transmission lineages 157

5.12 Distance of viral movements over 50km for each of the largest seven Delta transmission lineages in England. 158

6.1 Schematic showing different types of data and possible actors involved in data production which can be fed into an integrated analysis . . . 172

List of tables

3.1 District characteristics underlying the simulation of the Ebola Virus Disease epidemic in Sierra Leone using ABSynthE. 89

3.2 Infection parameters underlying the simulation of the Ebola Virus Disease epidemic in Sierra Leone using ABSynthE 90

4.1 Non-synonymous mutations and deletions inferred to occur on the branch leading to lineage B.1.1.7. 125

4.2 Marginal Likelihood Estimation of different growth rate models 126

4.3 Information about individuals used in chronic infection analysis . . . 127

Nomenclature

R_0 Basic reproduction number of an infectious disease, the average number of secondary cases from each primary case in a fully susceptible population

R_e Effective reproduction number of an infectious disease, the average number of secondary cases from each primary case in a population with susceptible and non-susceptible individuals

ABM Agent-based model

ACE-2 Angiotensin converting enzyme-2, the human host receptor for the entry of SARS-CoV-2 into the human cell.

Admin2 Administrative level 2, similar to UTLA, but larger counties are split up into smaller parts. Admin2 is a consistent designation across countries, and the sequence data is presented at this geographical level.

COG-UK COVID-19 Genomics UK Consortium

COVID-19 Coronavirus disease 2019

DRC Democratic Republic of the Congo

DTA Discrete Trait Analysis

EVD Ebola Virus Disease

GISAID Global Initiative on Sharing All Influenza Data.

GLM Generalised linear model

HPD Highest Posterior Density

KGH Kenema General Hospital

MCC tree Maximum Clade Credibility tree

ML Maximum likelihood

Molecular clock outliers Sequences which lie far away from the regression line of the of the root to tip genetic distance against sampling time. In other words, sequences which are more or less divergent from the root of the tree than expected given their sampling time

NPI Non-pharmaceutical intervention

NUTS1 Nomenclature of territorial units for statistics, developed by the Office for National Statistics in the UK. Scotland, Northern Ireland and Wales are each one unit, and England is split into 9 units (e.g. North East, North West, Greater London etc

ORF Open reading frame

PHEIC Public Health Emergency of International Concern

RRW Relaxed Random Walk

SARS-CoV-2 Severe Acute Respiratory Syndrome Coronavirus 2

SGTF Spike gene target failure, found when using the Thermofisher TaqPath PCR kit, caused by the 69-70 amino acid deletion in the Spike gene. It can be used

as proxy for Alpha and BA.1 (one of the Omicron variant lineages) infections at the PCR stage

Tier 4 restrictions Strictest tier of restrictions to prevent the spread of SARS-CoV-2 used in England. Restrictions similar to a lockdown: non-essential shops and hospitality are closed, no indoor mixing allowed and only limited outdoor mixing between households permitted.

TMRCA Time of/to to most recent common ancestor

UKHSA UK Health Security Agency, formerly known as Public Health England or PHE

UTLA Upper tier local authority, consists of counties, metropolitan counties, inner/outer London and unitary authorities. Case data is presented at this geographical level

WAR Western Area Rural, the district and chiefdom next to Western Area Urban

WAU Western Area Urban, the district and chiefdom containing Freetown, the capital of Sierra Leone

WHO World Health Organisation

1

Introduction

Viruses are organic species between living and non-living organisms, neither living or non-living.

Anon

2021

1.1 Overview

The decreasing cost of genome sequencing combined with increasing computational power has led to an explosion of interest in the application of whole genome sequencing to public health (Ladner et al., 2019; World Health Organisation, 2021a). This is facilitated by genomic epidemiology, the use of pathogen genomes to study the spread of infectious diseases through populations. Together with the field of phylodynamics - the combination of epidemiology, evolution and immunodynamics (Grenfell et al., 2004) - this rapidly growing field has tackled key questions regarding epidemic preparedness and control, increasingly in real-time.

Over the course of the last two decades, there have been multiple crises in public health: beginning with the first SARS pandemic in 2003, through Ebola virus in multiple countries across Africa, Zika virus in the Americas, and most recently the current SARS-CoV-2 pandemic. Phylodynamics has played a role in each of these, and they have, to some extent, forced the development of new methodologies in bursts of innovation. Further, in order to understand the drivers of spread and persistence of viral epidemics, we can study each of these epidemics on the different levels that transmission and evolution occur on, from the selective and neutral pressures within a host that ultimately drive variation across a population, through individuals and their local contact networks, up to large-scale national and international dynamic studies.

Out of the crises in the last 20 years, two stand out in particular for the use of real-time genomics in public health: the Ebola virus epidemic in West Africa, and the SARS-CoV-2 pandemic. These have led to two of the richest viral genomic datasets to date, allowing us to explore different aspects of how phylodynamics can be applied to help curb the spread of a pathogen. They have also changed much about how genomics is viewed as part of the public health response, and started discussions about how we respond to disease control as a global community.

In 2014, Ebola virus was detected in humans in upper West Africa for the first time. Sustained human-to-human transmission across three countries eventually led to over 28,000 cases and more than 11,000 deaths, causing severe damage to the economies and healthcare systems of three already fragile countries. It was also the first time that genome sequencing was used in a substantial way to supplement traditional epidemiological techniques in close to real-time. 1,610 genomes were collected during the epidemic, making it the largest such dataset at that time. The quality of the metadata, and the representativeness and scale of the sequence dataset provided an opportunity to explore and develop new methodologies for

phylodynamics, as well as investigate the most devastating viral epidemic of the 21st century at that time.

In 2019, a novel coronavirus, subsequently named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), was determined to be causing a cluster of atypical pneumonia cases in Wuhan in China. Despite rapid and decisive actions from Chinese authorities, it spread quickly across East Asia, before seeding Europe, the Americas and Africa. To date, it has caused at least 450 million cases and 6 million deaths globally (Dong, Du, and Gardner, 2020) and is by far the largest viral pandemic since the 1918 H1N1 influenza pandemic. Due to the scale and importance of the pandemic and advancements made in genomic sequencing technology and methodology during the 2010s, the sheer volume of genome sequences and analyses is unprecedented. Further, many countries which did not formerly utilise genomic data have begun to produce and rely on its outputs. With the genomic dataset now numbering more than 10 million sequences from more than 185 countries across the globe (www.gisaid.org), and subsequent pressure from policy makers and funders to see returns on investment in this relatively new field, there has been a rapid development of new methodologies to utilise this dataset to its fullest extent to help control the pandemic.

1.2 History of viral genomic epidemiology for public health

There have been six instances of the World Health Organisation (WHO) announcing a Public Health Emergency of International Concern (PHEIC) since the implementation of the legislation in 2007 (Wilder-Smith and Osman, 2020): H1N1 swine flu in 2009,

Ebola virus disease (EVD) in West Africa in 2013-16, Poliovirus transmission in multiple countries, EVD in the Democratic Republic of the Congo (DRC) in 2018-2020, Zika virus in the Americas in 2016, and COVID-19 (2019-present). The Human Immunodeficiency Virus (HIV) pandemic and the first SARS pandemic began before the classification of PHEICs existed, but are integral to understanding how the field has developed, and so I include them here. I also discuss the MERS-CoV outbreak in Saudi Arabia in 2014, which was not designated a PHEIC, but is considered to be a significant outbreak and is relevant to the background of the current pandemic.

These crises demanded a variety of approaches, including intensive genomic sequencing to understand transmission dynamics during the acute phase of epidemics (Ebola virus in the DRC), and broader genomic surveillance to identify cryptic increases in cases (Polio). They are shown in Fig. 1.1 and provide a useful framework to show the development of the role of genomic analyses in disease control, as well as the changing public health landscape that the EVD and SARS-CoV-2 epidemics have taken place in.

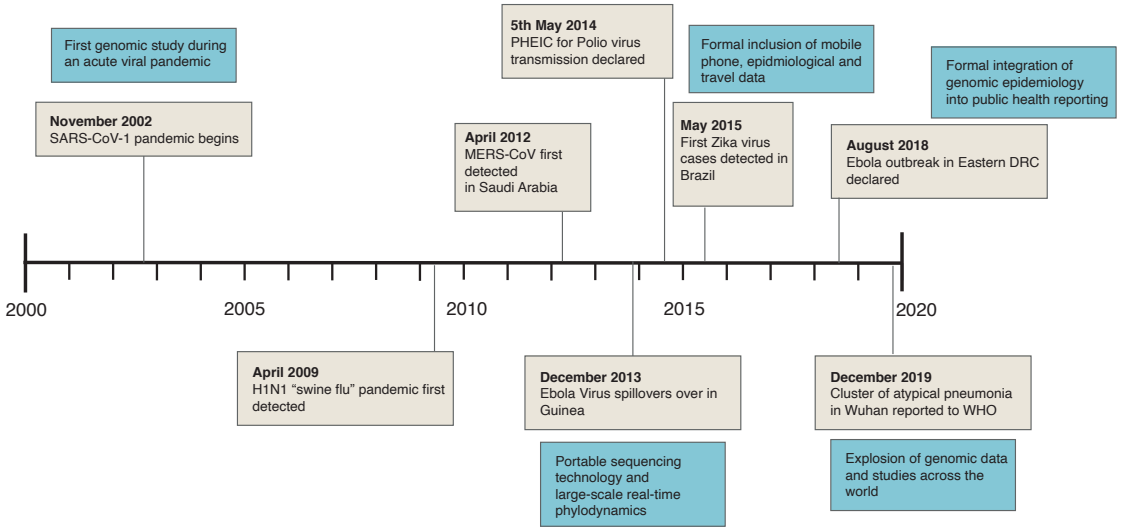

Fig. 1.1 Timeline showing major public health crises since 2000. The beginning of epidemics which would go on to be declared public health emergencies of international concern (PHEICs) are shown, as is the SARS-1 pandemic which took place prior to this legislation. Key advancements in the application of genomic epidemiology and phylodynamics to public health are shown in blue by the relevant epidemic.

1.2.1 HIV-1 M

HIV type 1 group M (HIV-1 M) is the deadliest ongoing viral pandemic, with an estimated 80 million cases and 36 million deaths to date (World Health Organisation, 2021b). The first whole genome sequences were completed in 1985 (Ratner et al., 1985; Wain-Hobson et al., 1985), and since then there have been tens of thousands of genomes sequenced and released. The earliest near-complete genome to have been sequenced is from a lymph node sample, from Kinshasa in the DRC in 1966 (Gryseels et al., 2020), and helped to confirm that dating estimates of the root of the phylogeny, and therefore estimates for the spillover from chimpanzees into humans, were accurate. Genomic analysis for HIV has been thorough, and many of the methodologies and techniques used were developed and honed for HIV. Notably, phylodynamics has been used for the study of HIV origins and early spread, large-

scale analyses of circulating HIV across countries and regions, and high-resolution investigations into person to person transmission.

HIV was established in the human population in the early 20th century, likely from chimpanzees in southern Cameroon (Sharp and Hahn, 2011), which have been found to have genetically very similar Simian immunodeficiency viruses (SIV, Keele et al., 2006). However, there are other HIV viruses of different groups (O, N and P) and HIV-2, resulting from separate cross-species transmission events (Sharp and Hahn, 2011), that have not caused major global pandemics like HIV-1 M and have instead remained for the most part in central Africa. Faria et al. (2014) used gene sequences analysis to explore the initial spread and growth of the HIV-1 M lineage, in order to provide hypotheses for why this subgroup has been substantially more successful than others. By estimating rates of viral migration, they found that early HIV-1 M lineages had spread mainly from Kinshasa, and was present in other key locations in the area in the mid-20th century. Subsequently, they estimated the change in past population dynamics to find that the population entered a phase of rapid growth around 1960, which is when it diverged from the dynamics of the other lineages. In order to explain this difference, they connected this sudden burst of growth to sex workers receiving high numbers of (unsterilised) injections to treat various sexually transmitted infections at the same time (Faria et al., 2014).

On a global scale, HIV-1 M has split into many different subtypes, which themselves recombine into circulating recombinant forms. The most common form globally is subtype C, partially because it is responsible for much of the epidemic in South Africa, which is one of the most affected countries globally. Wilkinson and colleagues (Wilkinson et al., 2019) used a large genomic dataset of sequences from across Southern Africa to elucidate the dynamics of HIV within South Africa, and found that there were many introductions of HIV into South Africa leading to multiple concurrent

epidemics. Further, these separate lineages did not all respond in the same way to large-scale interventions such as the introduction of antiretroviral therapy. This study therefore not only reconstructed the epidemic, but showed direct applicability of genomic epidemiology to public health through the monitoring of the efficacy of public health interventions (Wilkinson et al., 2019).

HIV has also been used to develop methods to examine high-resolution transmission. An early and famous case of this is that of the Florida dentist, an individual with HIV who was thought to have infected a patient. Through a mass testing campaign of the dentist's patients, 6 additional HIV-positive individuals were identified. By sequencing segments of HIV from the dentist, the 7 patients and the local community, the authors found that it was likely that 5 of the patients had likely been infected by the dentist, and the final two had probably received their infection through a different route (Ou et al., 1992). This very early example of genomic epidemiology was foundational to the field, and also highlighted a transmission route which could be severed by providing health care professionals living with HIV with better protective equipment, and the importance of autoclaving instruments. More recently, deep sequencing (i.e. high-depth sequencing of a single individual's infection, useful because within-host HIV is so diverse) paired with novel phylodynamic methods has been shown to be able to reconstruct transmission networks within a population, including direction of transmission, even in the absence of sexual history (Ratmann et al., 2019).

1.2.2 SARS-CoV 2002-2003

The first pandemic of the 21st century began in November 2002 in Guangdong Province in China (Cherry and Krogstad, 2004). It was introduced to Hong Kong in February 2003, and when the ten secondary cases from the Metropole hotel

returned home, SARS-CoV was introduced to multiple countries including Thailand, Canada and Germany (Centers for Disease Control and Prevention, 2003). The pandemic was brought to a halt after 8,096 cases and 774 deaths, and its relatively rapid end was a success of international cooperation and rapid response. However, issues surrounding the reporting of diseases of concern and a reluctance on the part of Chinese authorities to disclose information that could lead to, for example, economic harm through travel bans, highlighted issues in the current international legal framework. Following this, the International Health Regulations (IHR) were updated in 2004 to include, among other things, a demand for base level of pandemic preparedness from every member state; a greatly expanded list of notifiable infectious diseases; and the ability for non-state actors to report cases of diseases of concern.

Next generation sequencing technology was in its infancy in 2003, and rapidly and efficiently generating viral genomes using Sanger sequencing is difficult. However, several hundred sequences were generated reasonably quickly. One study analysed 169 short sequences from the Spike gene to find that the genomic data showed that while there were multiple introductions of SARS-CoV into Hong Kong, most cases were from a single source, specifically the large outbreak at the Amoy Gardens housing complex (Guan et al., 2004). A study addressing the epidemic as a whole identified mutational signatures common to different phases of the epidemic, specifically ORF8 deletions, and also shed light on the origins of the epidemic by including two coronavirus genomes from palm civets (Chinese SARS Molecular Epidemiology Consortium, 2004).

1.2.3 H1N1 pandemic 2009

The next pandemic was the H1N1 influenza A pandemic, known colloquially as "swine flu", was first detected in the US in April 2009. Later research revealed that it likely spilled over into the human population in central Mexico (Mena et al., 2016), and is estimated to have led to over 200,000 deaths worldwide (Dawood et al., 2012).

Swine flu was an early warning sign that global pandemic preparedness was not adequate. While the total number of deaths was lower than normal for seasonal influenza, this is partially because the mortality rate in the elderly was lower than expected, possibly due to their lasting protection from circulating influenza strains pre-1957 (Gostic et al., 2016). The late identification of the pandemic, leading to delayed vaccine production, would have been far more damaging in an influenza pandemic with a higher intrinsic mortality (Collignon, 2011). Finally, in countries such as the UK, that stockpiled vaccines and antiviral medication too late to be of genuine benefit (Boseley, 2010; O'Dowd, 2014), this pandemic became a symbol of government overreaction and waste.

Not many genomic studies were conducted at the time of the pandemic. Arguably the most useful study was the genomic analysis used to identify the origins of the pandemic, with the conclusion that the virus was made of reassorted genome segments from avian, seasonal human and swine influenza across many decades (Smith et al., 2009). This highlighted the role of pigs, specifically their large-scale movement as part of the agricultural industry, in the evolution of pandemic influenza.

1.2.4 MERS-CoV in Saudi Arabia 2012-present

Middle Eastern Respiratory Syndrome Coronavirus (MERS-CoV) was a novel coronavirus discovered in Saudi Arabia in April 2012. There were 190 confirmed cases in Saudi Arabia from 2013 to 2015, mostly in men from the age of 21-60, and 69 deaths (Aleanizy et al., 2017). MERS-CoV is able to transmit between people, and has also caused a large nosocomial outbreak in South Korea resulting in 186 cases across 16 healthcare facilities and 39 deaths (Yang and Jung, 2020). Despite being a novel coronavirus with the ability to spread between humans, it was never declared a PHEIC: the WHO decided over the course of multiple meetings that it was not an "extraordinary event" requiring large-scale international cooperation (Mullen et al., 2020).

During the first outbreak of MERS, there were some genomic studies conducted as the outbreak was ongoing. Cotten and colleagues presented 21 genomes, and combined them with nine other published genomes to describe the outbreak as it progressed (Cotten et al., 2013). This study identified two co-circulating lineages in Riyadh in Saudi Arabia through the reconstruction of the geographical location of the ancestral nodes, and found that the dynamics and genomic diversity observed were due to both human-to-human transmission and continued spillover events. This helped to identify key characteristics about the reservoir population, as it must be both large enough to support substantial viral population, and in regular close contact with humans. Later research identified dromedary camels as a likely reservoir population, due to epidemiological (Reusken et al., 2015) and serological (Müller et al., 2014) evidence. An analysis combining 174 human genomes and 100 camel genomes sought to explore the dynamics of the continuous spillover events observed in Saudi Arabia, and confirmed that hundreds of spillover events from camels into

humans have taken place (Dudas et al., 2018). The authors also found that spillover into humans occurred when the population of susceptible camels had increased sufficiently to cause an outbreak in the camel population, and that this was due to the birth of new calves (Dudas et al., 2018). Therefore, genomic analysis has shown that protecting individuals who interact with camels at specific times of the year and monitoring outbreaks in camel populations may prevent human cases.

1.2.5 Ebola Virus in West Africa 2013-2016

The West African EVD epidemic began in Meliandou, a small village in Guinea near the border with Liberia and Sierra Leone. Its exact origins were never confirmed, but the index case was likely a toddler playing near a hollow tree filled with bats (Marı Saéz et al., 2015) who became ill in December 2013. The child's sister and mother were both infected and died shortly afterwards, and many people were infected caring for his mother during her spontaneous abortion, or during one of their funerals. An individual caring for a relative of the index case then took Ebola virus to Conakry, the capital of Guinea, where he died (Timothy et al., 2019). There was a low index of suspicion for EVD for clinicians in the region as there had never been a case in West Africa, with the exception of a single case in Côte D'Ivoire obtained from the autopsy of a wild chimpanzee (Formenty et al., 1999). It was therefore consistently misdiagnosed, for example as cholera or malaria, until March 2014 when it was reported to the WHO (WHO Ebola Response Team, 2014). A PHEIC wasn't declared until August 2014, by which time there had been almost 1,000 deaths.

From Guinea, Ebola virus spread into Liberia in March 2014. Liberia's first transmission focus was in Lofa county, on the border with Guinea, until Monrovia, the capital, was seeded in mid-June and became the second hub of transmission

(Arwady et al., 2015). There were multiple introductions of EVD into Liberia, but most of the sequences taken fall into a single lineage which was found across the country (Ladner et al., 2015). There were also multiple re-introductions from Liberia back into Guinea from the summer of 2014 onwards (Ladner et al., 2015).

The first case in Sierra Leone was reported on 25th May 2014 in Kenema General Hospital (KGH), following the funeral of a traditional healer in the far east of the country on the border with Guinea (Goba et al., 2016). Similarly to Liberia, the initial focus of the epidemic in Sierra Leone was where the virus had been introduced in the far east of the country, but had a second focus of transmission once the virus reached its capital of Freetown in the west. Across all three countries, the major urban centres of their capital cities were vital for the spread and persistence of the epidemic, and the substantial urban transmission is likely one of the myriad of reasons why this EVD epidemic was larger than every previous EVD epidemic put together (Dudas et al., 2017). There were also sporadic transmission chains in Mali, Nigeria, the USA, and Italy and individual cases in Senegal and Spain.

This epidemic marked a watershed moment for genome sequencing and its applications to public health. Prior to this, genomic datasets collected in the acute phase of epidemics were mostly small and so it was difficult to answer questions about large-scale dynamics of viruses. It was also in this epidemic that the value of genome sequencing beyond purely academic questions and those around origins of viruses began to be clearer to public health actors. The 1,610 whole genome sequences and associated metadata that were collected during the epidemic made it the largest genomic database collected during the acute phase of the pandemic at that time, and it is still the second largest such dataset after SARS-CoV-2.

Part of the reason so many more genomes were generated during this epidemic than before was due to advancements in sequencing technology. The Oxford

Nanopore Technology (ONT) MinION was released in 2014, and has enabled a new era of portable sequencing due to its small size. It can be transported on commercial flights in normal luggage along with all its reagents, is cheaper to run than its more conventional counterparts, and does not require continuous electricity or internet throughout a sequencing run (Quick et al., 2016). Most of the sequences generated in the West African EVD epidemic were not through the ONT MinION, but it helped to kickstart the realisation that real-time sequencing was possible and useful in resource-limited settings, including rural areas without access to extensive laboratory equipment (Ladner et al., 2015; Arias et al., 2016).

Genomics was used extensively to explore this Ebola virus epidemic in real-time. Early papers explored the origins of the epidemic, in particular complementing the shoe-leather epidemiology and reverse contact-tracing effort that found the index case in Meliandou. With whole genome sequencing, they confirmed the timing of the start of the epidemic, and also found that the epidemic arose from a single spillover event followed by sustained human-to-human transmission (Gire et al., 2014; Baize et al., 2014). Later papers explored the dynamics of Ebola virus across Guinea (Carroll et al., 2015), Liberia (Ladner et al., 2015) and Sierra Leone (Arias et al., 2016; Park et al., 2015). Towards the end of the epidemic, genomic data was used to investigate EVD transmission in apparently transmission-free locations and therefore to discover the reactivation of live virus in persistently-infected individuals leading to new transmission chains (Mate et al., 2015).

In this epidemic therefore there was cutting-edge genomic analysis which complemented traditional epidemiological data to elucidate its origins, spread and persistence; and lead to the discovery of novel sequelae relevant to preventing further transmission of the virus.

1.2.6 Poliovirus transmission in 2014

At the same time as MERS-CoV was spreading in Saudi Arabia, and Ebola virus was spreading in West Africa, a PHEIC was declared for Polio transmission in multiple countries (Wilder-Smith and Osman, 2020). While Poliovirus has been a scourge of public health for as long as human civilisation has existed (Galassi, Habicht, and Rühli, 2016), it was pinpointed for global eradication following the development of two highly effective vaccines in the mid-20th century. Surveillance of continuing Poliovirus transmission is therefore vital, as the source of every case must be identified for eradication to be successful.

Genomic surveillance has been used widely to investigate vaccine-derived (Famulare et al., 2016) and wild polio outbreaks (Yakovenko et al., 2014). In Tajikistan, a large outbreak in 2010 led to 463 confirmed cases. Genomic analysis attributed this to wild poliovirus 1 introduced from India, which then spread to several neighbouring countries before being controlled by a rapid vaccination campaign (Yakovenko et al., 2014).

1.2.7 Zika virus in the Americas 2015-2016

The first epidemic that began after the peak of cases of EVD in West Africa was Zika virus in the Americas, another emerging pathogen. The increase of microcephaly cases in Brazil in 2015-2016 led to the discovery of a large Zika epidemic, and the WHO announced that the epidemic was a PHEIC (the first for an arbovirus) in early 2016 (Wilder-Smith and Osman, 2020). The response to the Zika virus epidemic marked the start of a new phase of genomic epidemiology and phylodynamics, as well as public health, and used some of the lessons that had been learnt in EVD.

Substantial genomic investigation has been undertaken to investigate Zika virus epidemics since 2016, including early examples of the principled inclusion of human mobility data. The first major paper used 20 sequences from the Americas and 3 from Pacific islands which had had earlier Zika outbreaks, as well as epidemiological and flight data to identify routes of entry into Brazil (Faria et al., 2016). This study was key to ruling out the introduction of Zika virus into Brazil by mass travel associated with sporting events due to the timing and origins of the introductions into the country, and instead attributed it to normal human movement. This result was confirmed by a later study, which used more genomes collected throughout the epidemic to provide a detailed picture of how Zika spread across the Americas, and highlighted the importance of northeast Brazil in seeding multiple locations in Central and South America and the Caribbean (Faria et al., 2017). It should be noted that this study built on experience of portable sequencing from the West African EVD epidemic, and included sequences from the ZIBRA project, which involved a mobile sequencing unit in a bus travelling around different parts of Brazil. As a final example, a more recent study has shown how Zika may silently circulate by using genomic sequencing from travel-associated cases to identify a hidden epidemic in Cuba in 2017, which was not performing its own genomic surveillance (Grubaugh et al., 2019).

1.2.8 Ebola virus in the DRC 2018-2020

Prior to the SARS-CoV-2 pandemic, the most recent PHEIC declared was the EVD epidemic in Nord Kivu in the north east of the DRC, close to the border with Uganda (Wilder-Smith and Osman, 2020). This became the DRC's largest EVD outbreak to date, with 3,470 cases and 2,280 deaths reported by the time of its official end in July 2020. Usually, EVD outbreaks in the DRC are fairly contained, remaining in rural areas of the country and with strong community engagement instrumental in avoiding

further spread. However, Nord Kivu is an unstable region, with multiple different militia groups rebelling against government control, complicating EVD control and exacerbating the epidemic (Matfess, 2018). Large population movement associated with refugees also increased the risk of international export to neighbouring countries including Uganda and Rwanda. This epidemic was ultimately brought under control with traditional and well-tested EVD control measures, and the introduction of a new and highly effective vaccine (Maxmen, 2020).

In this epidemic, researchers at the Institut National de Recherche Biomedical (INRB) in Kinshasa sought to formally integrate genomic epidemiology with the public health response to make genomic data more directly useful to controlling the disease. They collected 792 whole genomes at regular intervals over the two year period, and released reports in English and French to accompany the data (Kinganda-Lusamaki et al., 2021). These reports led to the inclusion of clergy and motorbike taxi drivers in the pre-emptive vaccination campaign by identifying superspreading events associated with these professions, and identifying a relapse case after 149 days which led to a large transmission chain (Mbala-Kingebeni et al., 2021). The success of this programme with integrating real-time genomic analysis into a more traditional public health response will continue to emphasise the role that genomics can play in response, and help to move it further away from purely academic questions.

1.2.9 SARS-CoV-2 pandemic 2019-present

The COVID-19 pandemic, caused by SARS-CoV-2, was first identified in December 2019 when a cluster of atypical pneumonia cases were connected to a market in Wuhan, China. The virus was rapidly identified as a novel coronavirus, and the first whole genome sequence was shared on the 12th January 2020 (Wu et al., 2020).

By the end of January, there were already 8,000 cases confirmed in 18 countries worldwide, including in North America, Europe and many other countries in Asia (World Health Organisation, 2020b). By the end of March 2020, many countries had instituted stay-at-home orders of varying severity to their citizens to try and slow the spread of disease and avoid overwhelming healthcare systems. The first wave of the pandemic began to decline towards the later half of 2020, however the emergence of the Alpha and Beta variants at the end of 2020 led to new global waves of infection. These were followed by waves caused by the Delta variant in most countries part way through 2021, and most recently there has been a global surge of the Omicron variant at the start of 2022. The mortality of the latter waves was somewhat abated by the roll-out of a series of successful vaccination campaigns, with 63.4% of the world's population having now received at least one dose of a SARS-CoV-2 vaccine (Mathieu et al., 2021). At time of writing, mid-way through 2022, many high-income countries have removed public health measures to prevent spread despite remaining high prevalence. However, with only 13.7% of individuals in low-income countries with a single dose of the vaccine (Mathieu et al., 2021), and a continued lack of genuine political will to remedy this, the pandemic is far from over.

With more than 10 million whole genome sequences for SARS-CoV-2 publicly available, the amount of genomic analysis that has been undertaken on every scale, from small-scale transmission studies to large-scale national and international studies is truly unprecedented. Further, many countries which had little previous sequencing experience have produced detailed genomic analyses of the dynamics of SARS-CoV-2 within their own country (e.g. Butera et al., 2021; Dudas et al., 2021). In the UK, genome sequences were generated by the COVID-19 Genomics UK (COG-UK) consortium. This was a partnership between academic, public health and NHS partners to generate and analyse sequences in real-time. After the set-up of

population-level sequencing, it was able to generate representative samples from across the country, making large-scale analyses possible and reliable. At time of writing, over 2.5 million genomes have been generated by COG-UK.

With all the data that has been produced during the pandemic, it has been impossible for some time to analyse all of the sequences at once or use previously developed methodologies to do so. To make this wealth of data useful to as many people as possible, several tools have been developed. First, a lineage naming system (Rambaut et al., 2020a) and an associated tool, pangolin, to assign sequences to those lineages were developed (O'Toole et al., 2021a). These innovations together enable individuals to track the dynamics at different geographic resolutions without requiring large-scale and complex phylogenetic analysis (as all analysis with this volume of sequence data is complex). By applying pangolin assignments, I have co-developed a global variant tracking system named grinch (O'Toole et al., 2021c), which provides easy-to-understand graphics and descriptions of where new variants are in the world. Finally, to condense the whole genomic dataset into the phylogenetic context around sequences of interest and automate outbreak investigations leading to the generation of interactive reports, I have co-developed the software package civet (O'Toole et al., 2021b). The aim of all of these tools is to remove some of the technical difficulty in analysing SARS-CoV-2 genomes, and ensure that data producers see return on their investment in producing the sequences, and therefore retain buy-in to the whole sequencing programme.

Finally, genomic surveillance has been instrumental in the discovery of variants. Initially, the SARS-CoV-2 viral population exhibited expected evolutionary patterns i.e. it evolved relatively slowly for an RNA virus due to its proof-reading mechanism, and changed incrementally as it spread through the human population. At the end of 2020 however, two distinct viral populations were detected, one in the UK and one

in South Africa, with far more than the expected number of mutations. These were eventually named Alpha and Beta variants (Hill et al., 2022; Tegally et al., 2021). As 2021 began, an additional variant was detected in Brazil (Faria et al., 2021), which became called the Gamma variant. Later in 2021, a concerning growth in cases in India was associated with the Delta variant (Vaidyanathan, 2021; McCrone et al., 2021). Finally, the Omicron variant has been associated with the most recent wave of infection globally (Viana et al., 2022). Without global genomic surveillance, it would be unclear why different countries around the world were suddenly seeing extreme rises in cases, especially those with high population immunity from previous infection or vaccination.

The analyses and innovations able to be undertaken at extreme speed during the SARS-CoV-2 pandemic are built on two decades of theory, technology and methods development. The size and richness of the dataset will help to push the field of phylodynamics to new heights, but it would not be possible without the work done previously to help to demonstrate that genomics is a useful part of the public health response, and how to combine it with non-genomic data.

1.3 What is phylodynamics?

Phylodynamics was defined by Grenfell et al. (2004) as the integration of phylogenetics, immunology and epidemiology, and has been extended over the last ten years to formally include many other types of data into genomic analyses. Viruses, particularly RNA viruses, are well described using phylodynamic methods, as their evolution is on the same timescale as their ecology: they are “measurably evolving” (Drummond et al., 2003). In practice, this means that as they spread through space

and time, we can observe their evolution, and pathogens that are more genetically similar to each other are more likely to be close together spatially and temporally.

I use coalescent methods to generate the phylogenetic trees presented in this thesis. These methods model when sampled lineages join together, or 'coalesce' in a genealogy, going backwards in time (Kingman, 1982). These are an extension of standard population genetic concepts, and have frequently been used to explore population size changes of pathogens (Pybus, Rambaut, and Harvey, 2000). The other commonly used conceptual framework for phylogenetics is birth-death methods, which explicitly include sampling in the phylogenetic inference and can therefore be useful for epidemiological modelling (Stadler et al., 2012).

Coalescent phylogenetics has been greatly expanded over the last few decades. While it was initially developed for constant populations, many parametric and non-parametric growth models have been developed to incorporate changing populations (see Hill and Baele, 2019 for an overview). These all estimate effective population size, a metric which uses genetic diversity under an idealised model of reproduction as a measure of population size. It is not the same as census population size (i.e. the number of individuals), and has a complex relationship with it due to deviations from this idealised model of reproduction. Therefore, changes in effective population size should only be compared to trends in census population size, and not exact numbers.

The most recent of these population growth models is the Skygrid (Gill et al., 2012), a non-parametric smoothed estimate of the change in effective population size over time with externally specified change-points. The Skygrid is useful for tracking viral epidemics, especially when case count data is hard to come by, as it provides an estimate of when the epidemic is growing or shrinking. Further, the externally specified change-points allow the input of data such as the timing of specific policy

changes, and so key timepoints can be highlighted. Finally, this model has been extended to explicitly include external data as covariates in a GLM framework (Gill et al., 2016).

When the sampling time is known, a genetic distance phylogeny can be converted into a time-scaled phylogeny. This uses the concept of a molecular clock, which is that the substitution rate of a given species over time is roughly constant (Zuckerkandl and Pauling, 1965). Using the sampling times of lineages and with sufficient genetic diversity, the substitution rate can be inferred at the same time as the topology of the phylogeny (Drummond and Rambaut, 2007). This has been used to time key events in the history of a pathogen, for example spillover into a human population (Gire et al., 2014), or introduction into a new region (Faria et al., 2017). It can also be used to scale a population growth model in time, and it was this combination of methods that allowed Faria et al. (2014) to detect the sudden population growth in HIV-1 M in the 1960s (see section on HIV-1 M in the "History of viral genomic epidemiology for public health").

When sampling location is known, a phylogeographic analysis can be undertaken (Lemey et al., 2009). This involves modelling the evolution of the spatial location of a lineage across the phylogeny, which should be on a similar timescale to the genetic evolution of the virus. In the initial formulation, locations were viewed as discrete states, and the transitions between these states were reconstructed. This was extended by Lemey et al. (2014) to incorporate data that was not associated with the genomes at all, and instead was connected to the locations. This method includes spatially-heterogeneous data as covariates of a GLM which affects the matrix of the transitions between locations. By examining the coefficient of effect size (which may be negative or positive) and the percentage of iterations of the model that the predictor is included in, it is possible to identify if non-genomic factors affect the rate

of movements between locations. Finally, the change of location along a phylogeny can also be modelled in a continuous manner (Lemey et al., 2010), reconstructing a wave-front of viral spread across a region.

These broad methods have been used historically in the crises discussed, and underpin the work presented in this thesis.

1.4 The scales of viral dynamics and evolution

Analysis of the evolution and spread of viruses can be undertaken at many different scales. Connecting these scales and understanding the through-line in the similarities and differences between them can illuminate patterns to help us understand key principles of viral dynamics and evolution. Different scales also focus on different forms of public health interventions, from ensuring equity of vaccination and antivirals to prevent chronic infection, to infection prevention measures on a single ward of a hospital, to international travel bans and lockdowns (Fig. 1.2).

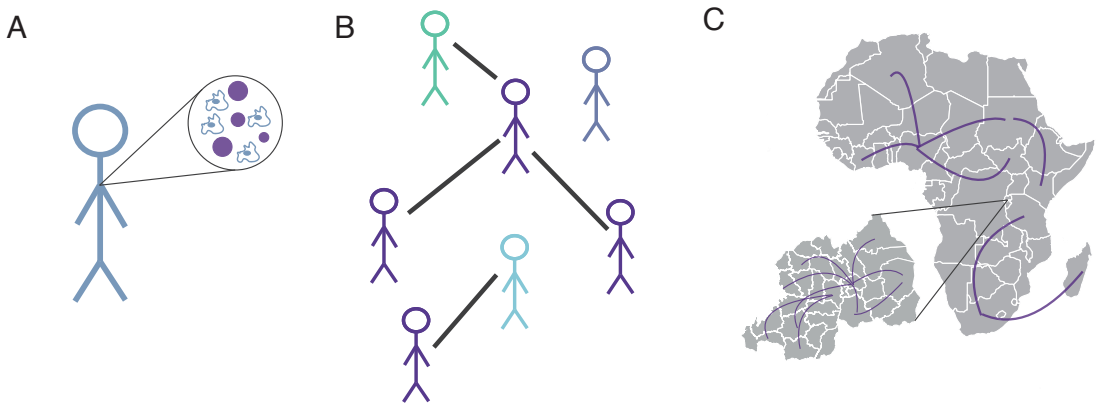

Fig. 1.2 Scales of transmission and evolution of viruses. A) Within-host evolution, especially in response to immune selection. B) Local transmission networks, altered by individual-level heterogeneity in many different factors. C) National and international spread. Lines indicate movement of viruses from one region to another.

This thesis, fundamentally, is an examination of two epidemic viruses at different scales, from evolution within a single host, through spread among individuals and their immediate contacts, and through to national and international dynamics of introduction and maintenance of viral transmission.

1.4.1 Within-host evolution

Evolution at the within-host level is the origin of all viral genetic diversity (Fig. 1.2A). While the importance of this is more obvious for long-term infections such as those caused by HIV, Hepatitis C and cytomegalovirus, it can also be important in acute infections, such as in SARS-CoV-2, influenza and norovirus, especially if these infections become persistent. Persistent infections of normally acute pathogens can lead to the exposure of the virus to a different form of evolutionary pressure, leading to jumps in fitness at the between-host scales (chapter 4). An understanding of within-host pressures is therefore vital in order to understand between-host transmission and adaptation because they underlie the dynamics of every infection.

Selection and drift are both key processes within a host. If the effective population size is large, then selection will be more important, and the evolution of the viral population will be more deterministic, as similar mutations are expected to be adaptive and selected for multiple times (Lauring, 2020). With a smaller effective population size, then genetic drift will take over as a more important factor, and neutral or mildly deleterious mutations that arise may be expected to spread through the population in a stochastic manner. Whether a viral population is large or small varies depending on the nature and strength of the immune response, which is specific to the virus in question, and its methods of evading the immune system, as well as the host itself.

Within-host evolution and between-host evolution have distinct selection pressures. For within-host evolution, the primary selection pressure is the immune system, with antibody and T cell responses driving specific adaptations. For example, escape from CD8+ T cells drives much of the evolution of HIV within a host (Arcia et al., 2017), and selects for mutations which interfere with antigen processing (Draenert et al., 2004), presentation (Leslie et al., 2004) and recognition (Iglesias et al., 2011). In general, immune selection may drive an evolutionary arms race akin to predator-prey interactions. As such, an intermediate immune response is expected to produce the fastest level of viral adaptation within a host, as it is not strong enough to dampen viral replication, but is substantial enough to produce a strong selective pressure (Grenfell et al., 2004). On the other hand, selection for increased between-host fitness may involve a greater amount of viral shedding, or replication in a specific part of the host. As these selection pressures are different, within-host adaptations may come at the cost of between-host fitness. However, an adaptation that increases viral survival or replication within a host may also lead to an increased ability to transmit, and so adaptation to the two scales of transmission are not necessarily mutually exclusive.

Compartmentalisation within a host can lead to increased viral diversity as it leads to selection for different traits to adapt to different environments. Further, different patterns of neutral selection can result in different alleles becoming fixed in compartmentalised viral populations. This compartmentalisation can be anatomical, for example, infection of the soft palate by influenza A has been shown to select for viruses which bind to the sialic acid receptor needed to spread efficiently between human hosts (Lakdawala et al., 2015). It can also be in terms of cells infected within the same anatomical compartment, with HIV infecting different T cell subsets across the course of infection, and this having an impact on the course and severity of the infection (Picchio et al., 1998).

An important stochastic process that can drastically alter the evolutionary trajectory within a host is the genetic bottleneck on transmission (McCrone and Lauring, 2018). Tight bottlenecks lead to a small founder population, emphasising the effect of drift on arrival in the new host. Adaptive mutations may therefore be stochastically lost on entry to the new host, and neutral or deleterious ones may be included in the population. However, it is important to note that the founder population may not be a random sample of the donor. In influenza A, the infection route changes the population composition (Varble et al., 2014); and in HIV, there appears to be some selection for the founder variant (Boeras et al., 2011), perhaps for it to be more similar to the original founder variant and therefore without the within-host specific adaptations it has acquired during the course of a long infection (Lythgoe et al., 2017).

1.4.2 Transmission between individuals on a local scale

Individual-level heterogeneity in transmission can have profound effects on the population dynamics of an infectious disease (Fig. 1.2B). For diseases with a high degree of superspreading, it may be possible to control them more effectively by identifying the locations or individuals connected to such events (Lee et al., 2020; Lloyd-Smith et al., 2005). Revealing the heterogeneity in transmission can be achieved by combining contact tracing data with genomic data, but is difficult to do using genomic data alone. Examples of the pairing of genomic and contact tracing data include two examples of detailed investigations of SARS-CoV-2 transmission lineages that went on to cause the bulk of cases in the mid-2020 wave in New Zealand, from a wedding-related cluster (Geoghegan et al., 2020), and Australia, from a single case of hotel quarantine escape (Lane et al., 2021), both of which showed the importance of early and strict lockdown measures.

Further, understanding the demographic make-up of a population at risk is crucial in determining potential burden on the public health system. Demographic factors like age and socio-economic status are known to be associated with severity of disease outcomes (Ottersen et al., 2014). In the context of COVID-19, economically disadvantaged populations were more severely affected (Mena et al., 2021), as are older individuals (O'Driscoll et al., 2020). At a population level, high quality data can help to explore details of epidemic dynamics, for example the role of individuals between the ages 20-49 in maintaining the SARS-CoV-2 epidemic in the USA (Monod et al., 2021).

If there is sufficient data about the population and the course of the infection, an agent-based model (ABM), can be developed to incorporate some of this individual-level heterogeneity (chapter 3). ABMs monitor the infection status of every individual, as well as every transmission event. By defining parameters on the scale of individuals, local interactions are explicitly captured, and can generate aggregate dynamics as a combination of these local interactions; as well as incorporating rich but qualitative data (Bruch and Atwell, 2015). This metadata can include factors such as travel and how individuals interact with each other, which is useful for pathogens such as influenza that can spread particularly well in schools and workplaces (Rakowski et al., 2010); as well as spatial data and how individuals interact with their environment, which is vital for investigating water-borne diseases such as cholera (Crooks and Hailegiorgis, 2014). This is in comparison to compartmental, or equation-based, models wherein the same equation defining some element of infection applies to the whole compartment; and incorporating complex mixing dynamics in particular becomes computationally and mathematically intractable (Bobashev et al., 2007).

Small-scale transmission is often explored through empirical studies. A key example of where this sort of study is useful is within hospitals, for examining the

infection process between wards and the subsequent implementation of infection control procedures. For example, a whole genome sequencing investigation of a SARS-CoV-2 outbreak in a renal unit found that there were two individuals who had likely infected each other, but had spent no time together on the unit. They had however shared a car to the hospital, and so the renal unit was able to put procedures in place for patient transport to protect against further cases (Francis et al., 2021).

On a slightly larger scale, whole genome sequencing has also been used to examine the spread of SARS-CoV-2 across a town, specifically Cambridge in England (Aggarwal et al., 2022). The authors found that while the cluster in the town had been introduced from elsewhere, two nightclubs were identified as common exposure events for many of the cases. Further, cases in the undergraduate students, for the most part, did not spread to staff at the university or the local community, but there was a cluster involving non-university healthcare workers and medical students (Aggarwal et al., 2022).

1.4.3 National and International transmission

The unprecedented global sequencing effort during the SARS-CoV-2 pandemic has enabled an extremely large number of studies to be undertaken exploring national and international dynamics across the globe. An early example is a study in Guangdong province in China, the secondary focus of transmission at the start of the pandemic, and found that while there were many introductions, effective early interventions made transmission chains short (Lu et al., 2020). A study in Rwanda, the first such Rwandan genomic study with sequencing conducted in-country, highlighted the importance of point-of-entry sequencing, and the impact of cross-border truck drivers from Tanzania, a country with an uncontrolled and unreported SARS-CoV-2 epidemic

(Butera et al., 2021). In New Zealand, the importance of point-of-entry isolation was confirmed, with a cornerstone of their control of the first wave being the prevention of onward transmission after introduction, with only 19% of introductions leading to secondary cases (Geoghegan et al., 2020). There have also been continent-wide studies of SARS-CoV-2 dynamics. Two prominent studies examined the dynamics of SARS-CoV-2 in the summer of 2020 in Europe, and confirmed Spain as the origin of the B.1.177 lineage but with the UK as an important secondary source of infections for the continent (Lemey et al., 2021; Hodcroft et al., 2021). Further, there is an Africa-wide study examining the 8,746 African genomes available at the time to explore lineage dynamics. This was particularly aimed at A.23.1, found mostly in East and Southern Africa, B.1.351 (Beta variant) and C.1, mostly sampled in Southern Africa, and B.1.525, found in West Africa (Wilkinson et al., 2021).

A relatively simple and well-tried way to examine transmission within and between countries (Fig.1.2C) on a larger scale is through a Bayesian phylogeographic analysis. This involves treating a location at any scale (often arbitrarily chosen based on the geographic scale where the samples are collected) as a trait of a sequence and modelling its evolution across the phylogenetic tree in a discrete (chapter 2) or continuous manner (chapter 5).

As discussed previously, discrete phylogeographic approaches are extensions of discrete ancestral trait inferences, and infer the rate of exchange between two or more locations jointly with the topology of the phylogeny (Lemey et al., 2009). Discrete phylogeographic analyses are often used on a national or regional scale to explore drivers behind the spread and persistence of epidemics. A well-known example is that of Dudas et al. (2017), which took the 1,610 Ebola virus genomes and their sampling locations from Guinea, Liberia and Sierra Leone to explore their epidemic dynamics. By reconstructing the locations of the ancestral nodes, they found that

most EVD cases in Sierra Leone resulted from a single introduction and sustained within-country spread; compared to the epidemic in Guinea which persisted due to many introductions into the country from Sierra Leone and Liberia, resulting in different viral population dynamics (Dudas et al., 2017). This information could help with the balance of point-of-entry measures and within-country interventions, and suggests that in some cases a "one-size-fits-all" approach to disease control may not be appropriate between different arms of the same epidemic. They then performed a phylogeographic GLM analysis (see "What is Phylodynamics"), which included 25 environmental, cultural, economic and geographic predictors. Only five were consistently included in the model: distance, population of origin and destination, being in the same country, and sharing an international border. Thus the dynamics of the epidemic roughly fell into a gravity model of transmission and international borders were largely protective. A final key takeaway of this regional-scale analysis for the dynamics of EVD in West Africa was that it mostly followed a metapopulation structure: it was sustained by a series of small and short-lived clusters of infection. A similar phylogeographic analysis conducted on EVD in the most recent DRC epidemic (see "History of viral genomic epidemiology for public health") found that the outbreak was also sustained by many, small clusters (Kinganda-Lusamaki et al., 2021). This implies that this metapopulation structure may be a consistent part of EVD dynamics, and genomic analysis on future outbreaks will help to confirm this.

Discrete phylogeographic approaches have also been adapted to estimate the number of introductions into countries during SARS-CoV-2. While it is possible to use many states in a discrete phylogeography (see chapter 2), the extremely large datasets in SARS-CoV-2 analyses require some simplification in the models used. For the introduction of the Delta variant into the UK, a three state discrete phylogeographic analysis was used to distinguish movement between the UK, India

and all other countries combined (chapter 5, McCrone et al., 2021). Across Europe, a discrete phylogeographic analysis combined with mobile phone and epidemiological data was taken to investigate the spread of the B.1.177 lineage in the summer of 2020. Spain and the UK were found to be important centres of spread for this lineage, and, for most countries, their lineage composition was determined by introductions from other countries. However, if they had a high incidence already, introductions were less successful (Lemey et al., 2021).

Continuous phylogeographic methods model the diffusion of a viral lineage in space (Lemey et al., 2010), rather than as discrete jumps between locations. In comparison to the Brownian diffusion models they are built on, they allow the rate of this diffusion to vary in time and therefore also across the phylogeny. As most locations do not have discrete boundaries (aside from, notably, islands), these methods are more likely to be accurate in most cases. However, high-resolution geographical metadata are required, as these approaches model the change in coordinates, and so if the data is too low-resolution much of the detail will be lost. It is also more computationally intensive than discrete approaches, and so more difficult to use on larger datasets without, at least, fixing the underlying topology of the tree.

Continuous phylogeographic methods have been applied post-hoc to the West African EVD epidemic to build on the analysis of Dudas et al. (2017). Dellicour et al. (2018) aimed to investigate the effect of specific interventions, and to investigate whether key results were different under a continuous diffusion process. They found that stopping longer-range movements had only a small impact on the size of the epidemic, but that preventing movement to all of the capital cities would have reduced the epidemic by two-thirds. Importantly, this analysis agreed with the discrete one that international borders were protective and that border closures slowed the rate of expansion of the epidemic significantly. A more recent paper used these methods to

elucidate the spatial invasion of the USA by West Nile Virus (Dellicour et al., 2020). By examining the lineage diffusion wavefront, the authors were able to estimate the rate of geographical expansion across the tree at 1,200 km/year, although this varied by genotype and this variation was largely down to annual temperature. In the SARS-CoV-2 epidemic in the UK, exceptionally high resolution metadata has been collected for much of the sequence dataset. Further, the UK is a small and well-connected country, and so lends itself much better to a continuous approach as there are not many natural discrete divisions. Using the COG-UK dataset, we conducted continuous phylogeographic analyses to explore the spread of Alpha and Delta variants across England, from the South-East and (predominantly) North-West of England respectively (chapter 5, McCrone et al., 2021; Kraemer et al., 2021).

1.5 Aims

In this thesis, I focus on the application of genomic epidemiology and phylodynamics to two history-making viral epidemics across different scales of evolution and transmission.

The first epidemic I describe in this thesis is the 2013-2016 Ebola virus epidemic in West Africa, with a specific focus on Sierra Leone (Fig. 1.3A). Each of the three main affected countries had different dynamics and trends, and viewing them all at once may obscure details and lead to missing information on a national level. Further, as most of the sequences in Sierra Leone form a single, monophyletic clade, it provides a clean dataset for exploring extensions to phylodynamic methods.

In chapter 2, I perform a detailed phylodynamic analysis of the main viral lineage in Sierra Leone. I explore the factors affecting movement between districts and chiefdoms of the country across different time periods using a time-inhomogeneous

discrete phylogeographic generalised linear model. In particular, it is of interest in how well the gravity model fits the dynamics of EVD when studied on a national, rather than international, level; and what deviations from the gravity model can tell us about how diseases spread across countries from a single introduction.

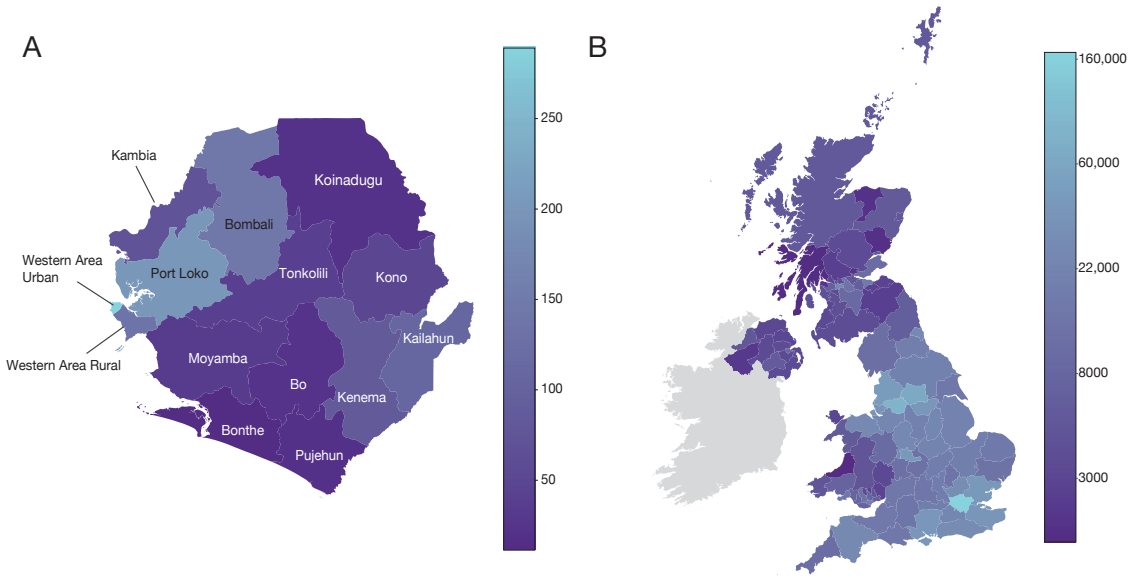

Fig. 1.3 Geographical spread of sequence data used in this thesis. A) Map of Sierra Leone showing the number of Ebola virus sequences per district. District names are also indicated. B) Map of the UK showing the logged number of SARS-CoV-2 sequences collected by the end of 2021 which can be unambiguously assigned to an administrative level 2 region.

In chapter 3, I use this phylodynamic analysis as the real-world data for fitting an ABM to the spread of Ebola virus through Sierra Leone. I have developed a stochastic simulator which has multiple contact levels to mimic differing levels of exposure in order to incorporate some of the individual-level heterogeneity that is so important in the spread of Ebola virus. This simulator also produces phylogenies from the transmission trees, and I use phylogenetic summary statistics to fit the model to the results of chapter 2 in the absence of case or death data. Finally, I use the simulator to explore the role of the seeding location in the EVD epidemic in Sierra Leone as a case study application for the model.

The second epidemic I focus on in this thesis is the SARS-CoV-2 pandemic, particularly on sequences generated in the UK by the COG-UK consortium (Fig. 1.3B). This rich dataset provides us with new opportunities to explore how evolution works on a small- and large-scale; as well as leading to challenges which require solving to work with datasets of this magnitude.

In chapter 4, I explore hypotheses for the evolution of the Alpha variant in the UK. I characterise the extremely long ancestral branch of the clade and examine the patterns in its early growth rates. I formally compare different hypotheses to explain the lack of intermediate transmission chains and elevated rate observed on the ancestral branch, involving the consideration of how within-host selective pressures may result in a variant which has a significant between-host fitness advantage. Finally, I compare all of the variants of concern to date in order to explore how variants evolve, and whether there are any common patterns between them. This highlights the importance of, possibly, a single persistent infection on the trajectory of a pandemic.

Finally, in chapter 5, I explore the dynamics of the major waves of SARS-CoV-2 until the end of 2021 within the UK: the first wave, including D614G viruses, the Alpha variant wave and the Delta variant wave. In comparing them to each other, I identify common themes, such as the effect of introduction into multiple areas at once, and the importance of major urban centres in England. This final chapter therefore connects to the first chapter in the study of introduction and dynamics of a virus at national scale, but uses a much larger dataset and different phylodynamic approaches to do so.

Investigating the evolution and dynamics of viral epidemics across different scales can help us to understand the underlying drivers of transmission and population-level adaptation of viruses. This can help predict how future epidemics and pandemics will

unfold, and therefore which public health interventions may be effective at preventing spread and adaptation. This thesis delves into the evolution and transmission of Ebola virus and SARS-CoV-2 in two specific settings by melding together epidemiological, spatial and genomic data and leveraging two decades of phylodynamic methodological advancement.

2

The establishment and maintenance of Ebola virus transmission in Sierra Leone

Sometimes, life is just too short

Prof. Andrew Rambaut on the deeper meaning of MCC trees

Personal Communication

2021-08-05

2.1 Introduction

The Ebola Virus Disease (EVD) epidemic began in Guinea in late December 2013 (Baize et al., 2014), and spread to Liberia, which had its first recorded cases in March 2014 (Nyenswah et al., 2016). The first case diagnosed in Sierra Leone was on 25th May 2014, in a pregnant woman who had travelled to Kenema General Hospital (KGH) from Kissi Teng Chiefdom in Kailahun district in the far Eastern part of Sierra Leone, on the border with Guinea (Goba et al., 2016). She later became the first survivor of EVD in Sierra Leone. With substantial reverse contact tracing efforts, this case was linked to the funeral earlier that month of a traditional healer who had been treating patients across the border in Guinea (Goba et al., 2016; Wauquier et al., 2015). In the end, 14 cases were found to be connected to this funeral (Goba et al., 2016) and it is thought to be responsible for starting most of the transmission chains in Sierra Leone (World Health Organisation, 2015c).

Cases in Sierra Leone grew rapidly, outpacing both Guinea and Liberia. At the start of the epidemic, transmission was intense in the east of the country, and KGH was rapidly overwhelmed as the only Ebola Treatment Unit (Senga et al., 2016). However, the east of Sierra Leone is relatively sparsely populated, and cases and deaths began to accelerate in earnest once transmission chains had established in Freetown, in Western Area Urban (WAU). Together with Western Area Rural (WAR) on the Freetown Peninsula, and neighbouring Port Loko district, it became the epicentre of the epidemic in Sierra Leone (World Health Organisation, 2015c), along with the capital cities of Monrovia in Liberia and Conakry in Guinea (Dudas et al., 2017; Dellicour et al., 2018). The epidemic was brought to a halt in November 2015, and there was a single flare-up of transmission associated with a persistently-infected

individual in early 2016 which was also controlled. Sierra Leone has therefore been EVD-free since 17th March 2016 (World Health Organisation, 2016).

Two previous studies have explored Ebola virus transmission in Sierra Leone. The first found that the majority of sequences from Sierra Leone fell within a single monophyletic lineage, indicating a single introduction, and only very limited exports to other countries. The authors concluded that transmission was therefore mostly on within-country level for Sierra Leone (Park et al., 2015). The other study investigated two specific widespread outbreaks of EVD: one in Port Loko and Kambia, and the other in Freetown. Of particular importance was the linking of a new transmission chain in Tonkolili district, which had had no cases for approximately four months, to the outbreak in Freetown via phylogenetics (Arias et al., 2016). On a regional level, there are two retrospective studies which explored the general dynamics of EVD spread and persistence across all three countries, one using discrete methods (Dudas et al., 2017), and one using continuous methods (Dellicour et al., 2018). There has, as of yet, not been any detailed studies of the dynamics of Ebola virus in Sierra Leone, or changes in these dynamics over time, across the course of the whole epidemic.

A common way that infectious diseases spread is through the gravity model of infectious disease: that is, large locations and their neighbours have more cases than smaller and farther away ones (Bailey and Gatrell, 1995). Previous work has found that this describes the overall epidemic in the region well (Dudas et al., 2017), as the three capital cities became important foci of transmission. However, in Sierra Leone, there were several months preceding the epidemic in Freetown, and so the dynamics in this period may be different to a gravity model as it becomes established. If the dynamics in this period do not follow expected patterns well, it may indicate that they are more prone to stochasticity, and transmission is more easily disrupted though

small perturbations. By exploring the dynamics of both timeframes, it is possible to explore better timings of interventions.

Thanks to coordinated sequencing efforts, there are over 1,000 genomes from the single introduction into Sierra Leone from a series of different teams. I use these to perform a Bayesian discrete phylogeographic analysis, including predictors in a generalised linear model (GLM) formulation (Lemey et al., 2014), with two epochs in order to tease apart factors which were important in both the initial spread and maintenance of EVD in Sierra Leone.

2.2 Methods

2.2.1 Dataset

The bulk of the genome sequence alignment (n=1,610) was taken from Dudas et al. (2017). These sequences were sampled across multiple countries during the epidemic, and obtained originally from Gire et al. (2014), Park et al. (2015), Bell et al. (2015), Simon-Loriere et al. (2015), Smits et al. (2015), Whitmer et al. (2016), Baize et al. (2014), Hoenen et al. (2016), and Arias et al. (2016). I then supplemented this with genomes sequenced and published after the epidemic had finished, which constituted 60 genomes from Li et al. (2017) and 218 genomes from Jansen van Vuren et al. (2019), resulting in 1,888 genomes. The monophyletic subtree which makes up the majority of sequences from Sierra Leone was then pruned from this wider phylogeny, giving 1,288 sequences. Finally, four sequences were unexpectedly more or less divergent from the root of the tree given their time of sampling, identifying them as molecular clock outliers (Hill and Baele, 2019). These were removed, leaving a final dataset of 1,284 sequences.

Predictor data for the district phylogeographic analyses were a subset of those used in Dudas et al. (2017). For the chiefdom analysis, data was calculated using shapefiles from the Global Administrative Database (gadm.org), and from the 2014 Sierra Leone census. For the distance to cities of more than 50,000 and 100,000 people, these could be in Liberia or Guinea as well as in Sierra Leone, and so to identify these cities, data was taken from the 2008 Liberia census and the 2014 Guinea census. It must also be noted that for the district-level analysis, the predictors involving distance to large cities are travel time, and include resistance measures (e.g. the quality of the roads from destination A to B), whereas for the chiefdom predictors, it is a greater circle distance from the geographic centroid of the chiefdom to the city in question.

2.2.2 Phylogenetic analysis

I first obtained the topology of the tree using a non-phylogeographic inference. The alignment was partitioned into four sections, coding positions 1-3 and intergenic regions, to allow independent estimates of the substitution rate among coding positions. The substitution rate was estimated using the HKY plus Gamma model (Hasegawa, Kishino, and Yano, 1985), to allow further among-site rate heterogeneity. Finally, I used an uncorrelated relaxed clock model and a non-parametric Skygrid model, with 88 gridpoints and a cutoff of 1.7 years prior to the most recent sample date to infer the change in virus effective population size over time (Gill et al., 2012). Two independent chains of 100 million states were run, with 50 million removed from each for burn-in, assessed using Tracer (Rambaut et al., 2018).

The Maximum Clade Credibility (MCC) tree from this initial inference was used as the fixed tree for the discrete phylogeographic GLM analysis. While it is possible

to do a joint phylogenetic and phylogeographic inference (Lemey et al., 2009), the number of discrete states in this analysis, specifically for the chiefdom-level analysis, made the computational burden of the joint inference prohibitive. Inferring a fixed tree topology is a reasonable approximation to this joint inference (e.g. Volz and Frost, 2017; Sagulenko, Puller, and Neher, 2018) and is able to be conducted in a reasonable time-frame.

For the district analysis, 14 districts and 8 predictors were used, with ambiguity codes assigned for sequences with no district information. Two independent chains were run for 40 million states, with 10% removed for burn-in. For the chiefdom analysis, 150 chiefdoms and 9 predictors were used. For sequences with district location but no chiefdom location, chiefdoms from that district were provided as options for the ambiguity codes; and for sequences with no location information, the whole country was provided. 7 chains were run for 500,000 to 5.5 million states, and 10-15% removed for burn-in, giving approximately 10 million post burn-in states for further analysis. For both analyses, there was a 50% probability of no predictors being included.

A two epoch model was used, allowing different predictor matrices to inform the movements between discrete locations. A transition time of the 31st August 2014 (or 1.0314 years prior to the most recent sample date) was used as an estimate of the end of the exponential growth phase of the epidemic in Sierra Leone.

To identify sources and sinks and explore network dynamics, I extracted the jumps between discrete locations from the full posterior set of trees using TreeMarkovJumpHistoryAnalyzer (Lemey et al., 2021), part of the BEAST software package (Suchard et al., 2018). I collated the number of movements between each pair of locations, and used averages from across the posterior set of trees to generate figures and results. For network analyses, I generated two full networks of districts and chief-

doms using the Python package networkx (Hagberg, Swart, and S Chult, 2008), with each location as a node in the network and movements as edges, weighted by the average frequency of the movement across the posterior distribution for that epoch. These average weights were then summed to generate the degree centrality measure used in the Results section. Figures are shown with results from across the whole posterior.

The results of the GLM are given as the effect size. This is the distribution of the coefficient, β, for each parameter, conditioned on that parameter being included in the model.

2.3 Results

In order to reveal the dynamics of Ebola virus in Sierra Leone, especially with respect to different phases of epidemic growth and maintenance, I ran two discrete phylogeographic analyses at the district (administrative level 2, n=14) and chiefdom (administrative level 3, n=150) levels. Districts contain approximately half a million people on average, although this varies significantly, with the smallest district (Bonthe) containing 200,781 people, and the largest (WAU) containing 1.05 million (Fig. 1.3). Chiefdoms are significantly smaller, and contain on average 46,878 people; and vary from 3,548 (Langrama in Kenema district) to 1.05 million in WAU. Note that Western Area Urban and Rural are both the only chiefdom in their district, and so are included at the same resolution for the district and chiefdom analyses.

I ran these analyses with a GLM formulation to include non-genomic data to explore what influences movement from one discrete location to another. Predictors used were those important in the gravity model, i.e. population and distance measures, as well as measures estimating rurality. I included a total of 8 predictors for the

district level analysis and 9 for the chiefdoms. Finally, both of these analyses were split into two epochs: the first epoch lasts until the estimated end of the exponential increase in cases at the end of August 2014, and the second epoch continues until the end of continuous transmission in Sierra Leone (i.e. flare-ups due to persistent infection were excluded). This allows the examination of the dynamics of the spread of Ebola virus across Sierra Leone in different phases of the epidemic.

2.3.1 Exponential growth and establishment

The early phase of the epidemic was characterised by relatively even amounts of movement across all districts; but with disproportionate movements from Kailahun District to Kenema District (Fig. 2.1A and Fig. 2.1C) in the east (17% of between-district movements in the epoch, 95% Highest Posterior Density (HPD): 14% to 20%), and, to a lesser extent, vice versa (4%, 95% HPD: 2% to 7%, Fig. 2.1C). The second highest proportion of movements was from Bombali District to Port Loko District in the north-west (10% of movements, 95% HPD: 6.8% to 13%).

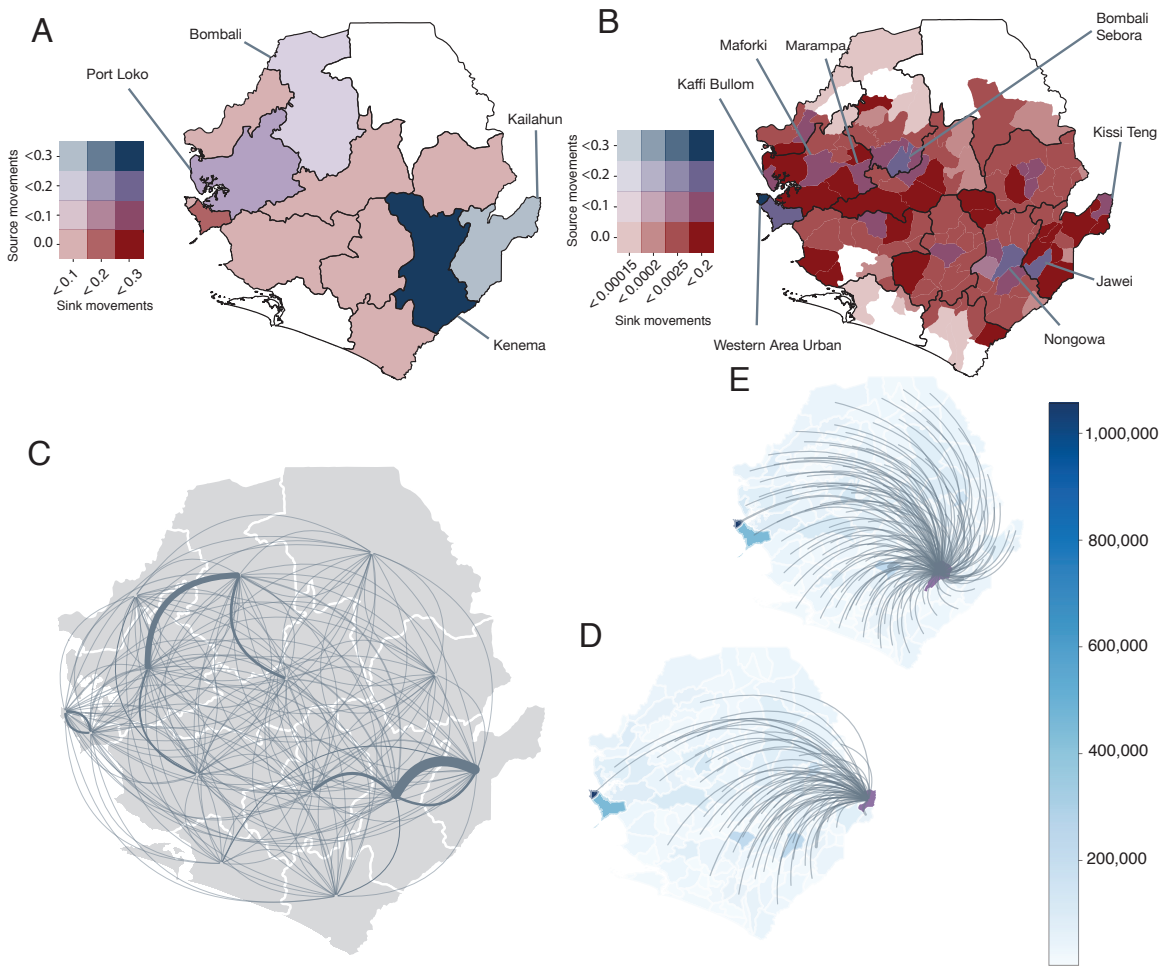

Fig. 2.1 Movements of Ebola virus across Sierra Leone in the establishment phase of epidemic growth across the whole posterior. A) Bivariate choropleth map showing the proportion of movements in the epoch that start (red to blue spectrum) and end (saturation) in specific districts. Important districts are highlighted. B) Bivariate choropleth map showing the proportion of movements in the epoch that start (red to blue spectrum) and end (alpha spectrum) in specific chiefdoms. Chiefdoms discussed in the section are indicated. C) Connections between pairs of districts, the thickness of the line is scaled corresponding to the frequency of movement from one location to the other. Direction is anti-clockwise. D) Movements leaving Kissi Teng chiefdom, choropleth is population of each chiefdom E) Movements leaving Nongowa chiefdom, choropleth is population of each chiefdom.

Out of all of the districts, Kenema and Kailahun were the most source-like: 26.7% of between-district movements in the first epoch started in Kenema (95% HPD: 18.5% to 37.1%), and 23.5% started in Kailahun (95% HPD: 19.6% to 27.4%, Fig. 2.11A).

However, when interrogating this further by examining the movements inferred from the chiefdom-level reconstruction, it becomes clear that there is heterogeneity within each district, and most of the chiefdoms in Kenema and Kailahun acted as sinks relative to other chiefdoms (Fig. 2.1B).

In Kailahun, only two chiefdoms had movements starting in them: one of these was Kissi Teng (3.6% of movements, 95% HPD: 3% to 4.26%), the chiefdom where the epidemic was seeded in Sierra Leone. Movements which began here finished in many other chiefdoms (74/149, distributed across the country), which were not necessarily the largest by population size (Fig. 2.1D), as might be expected under a gravity model. This was confirmed by the GLM analysis, where the population size of the destination had only a weakly positive effect on the frequency of movements when it was included in the model (0.08, 95% HPD: -0.12 to 0.27), which was rare (0.6% of reconstructions, Fig. 2.2D and Fig. 2.2E). The other notable chiefdom in Kailahun was Jawei, which originated 15% of movements (95% HPD: 12.8% to 17.4%, Fig. 2.1B).

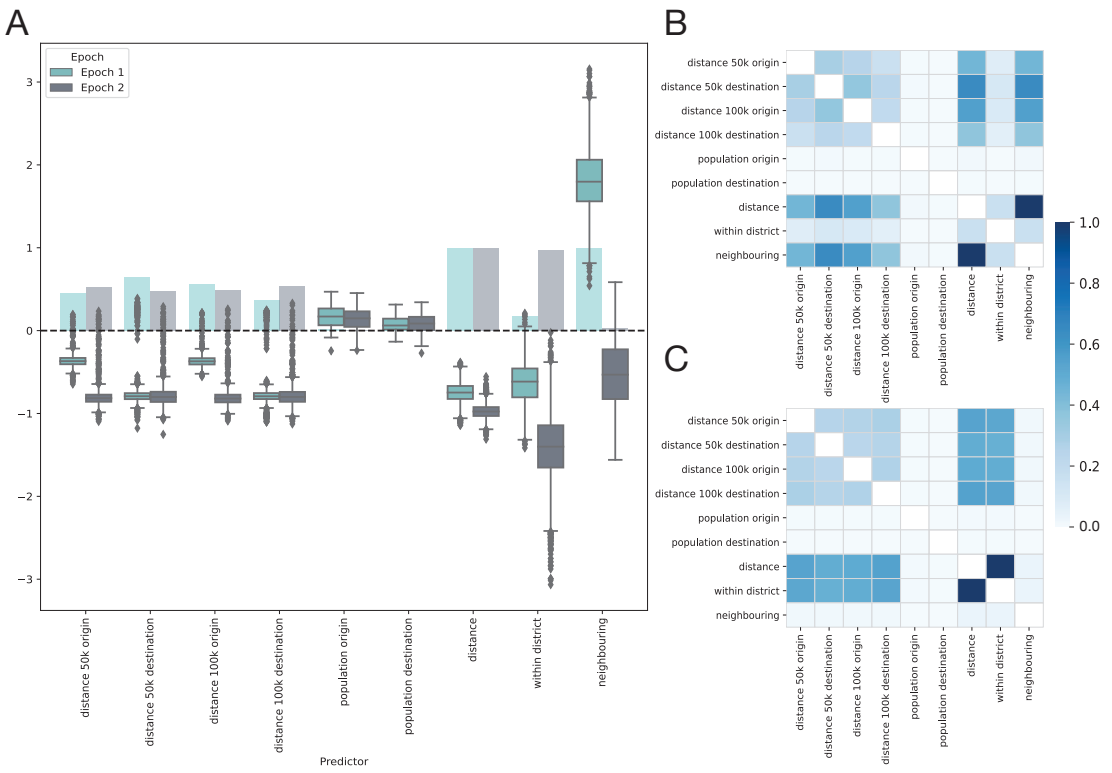

Fig. 2.2 Predictor results for the GLM on a chiefdom level. A) Effect size for the chiefdom level analysis, split by epoch. Bar charts show average inclusion between 0 and 1. B) Co-inclusion matrix for chiefdom level predictors in first epoch, darker colours indicate more co-inclusion. C) Co-inclusion matrix for chiefdom level predictors in second epoch, darker colours indicate more co-inclusion.

In Kenema district, there were several chiefdoms which display source-like behaviour. Of note, 19.5% of between-chiefdom movements in the epoch started in Nongowa (95% HPD: 15.4% to 25.7%), making it the chiefdom with the second highest proportion of originating movements. These movements, like those from Kissi Teng, travelled across much of the country: movements starting in Nongowa ended in 148 out of 149 other chiefdoms (Fig. 2.1E). In comparison, movements starting in a similarly sized chiefdom (Kakua in Bo) only ended in 111 other chiefdoms; and movements beginning in another source-like chiefdom neighbouring Nongowa

(Jawei in Kailahun) ended in 114 other chiefdoms. Nongowa's connectedness to the rest of the country is therefore difficult to attribute solely to size or location.

Further west, movements from Bombali to Port Loko contributed 10% of between-district movements (Fig. 2.1C). On a district level, they both had high proportions of movements ending in them, with Port Loko being slightly more sink-like than Bombali: 14.1% (95% HPD: 10.7% to 17.7%) vs 4.2% (95% HPD: 2.7% to 6.8%) of movements ending in each district respectively. They also both had many movements beginning in them, with Bombali having more movements starting in it (18.7% 95% HPD: 14.1% to 24.2%) than Port Loko (12.5%, 95% HPD: 7.8% to 18.1%, Fig. 2.1A). Similar to Kenema and Kailahun however, a higher resolution analysis revealed differences between chiefdoms within each district. Port Loko mostly had sink-like chiefdoms, which did not have many movements starting in them. There are three chiefdoms however which had a small number of movements originating in them (Fig. 2.1B): Maforki (0.7%, 95% HPD: 0% to 1.9%), Kaffi Bullom (0.8%, 95% HPD: 0% to 3.1%) and Marampa (1.2%, 95% HPD: 0.5% to 3.1%). It is notable that these are all much lower proportions than the equivalent chiefdoms in Kenema and Kailahun, and these locations were still much more sink-like than source-like. In Bombali, again most of the district acted as a sink, with one chiefdom, Bombali Sebora, standing out as a source of viral movements with 17.1% (95% HPD: 11.6% to 23.1%) of movements starting there - the third highest proportion in the country.

Many of the locations with a high proportion of movements starting there also had a higher proportion of movements ending there, therefore it is appropriate to explore reconstructed movements from the perspective of a network rather than dividing them into source or sink. Here, I use degree centrality, which is a simple measure of connectedness: the sum of the average weight of the edges in the network starting or ending with the node in question, where the weight is the frequency of movements

of that pair and direction for each reconstruction. Note that because the weights are averaged across the whole posterior, the average degree may be more than 1, and so should not be interpreted directly as proportions.

As previously discussed, Kenema had the highest proportion of between-district movements starting there, but it also had the highest proportion of movements ending there (22%, 95% HPD: 17.8% to 26.8%). Accordingly, it had the highest degree of centrality in the network of movements across Sierra Leone on a district level (0.31, 95% HPD: 0.25 to 0.45). This is in large part because Nongowa chiefdom in Kenema had the second highest average degree of centrality in the chiefdom network (1.03, 95% HPD: 0.42 - 1.20). Kailahun was the second most connected district, with an average degree centrality of 0.29 (95% HPD: 0.25 to 0.35). While not as significant to the movement network as Nongowa, Kissi Teng was disproportionately connected compared to its population and distance to other chiefdoms in Sierra Leone. It had the 7th highest degree of centrality (0.55, 95% HPD: 0.05 - 0.57) despite it being the 36th largest chiefdom by population. In the north-west group, Bombali Sebora had the third highest network centrality in the country (1.01, 95% HPD: 0.48 - 1.12), and Marampa, in Port Loko, is disproportionately connected compared to its population (0.44, 95% HPD: 0.41 - 0.46) as the 9th most connected but 24th most populous. Kakua, located in Bo district and containing its capital city, was also highly connected (0.78, 95% HPD: 0.32 - 0.85). On the other hand, on a chiefdom level, WAU had the highest degree centrality (1.12, 95% HPD: 0.98 to 1.35), as would be expected under a gravity model of transmission as it is the most populous chiefdom. This is not the case on a district level, where it was the fourth most connected district, after Kenema, Kailahun and WAR.

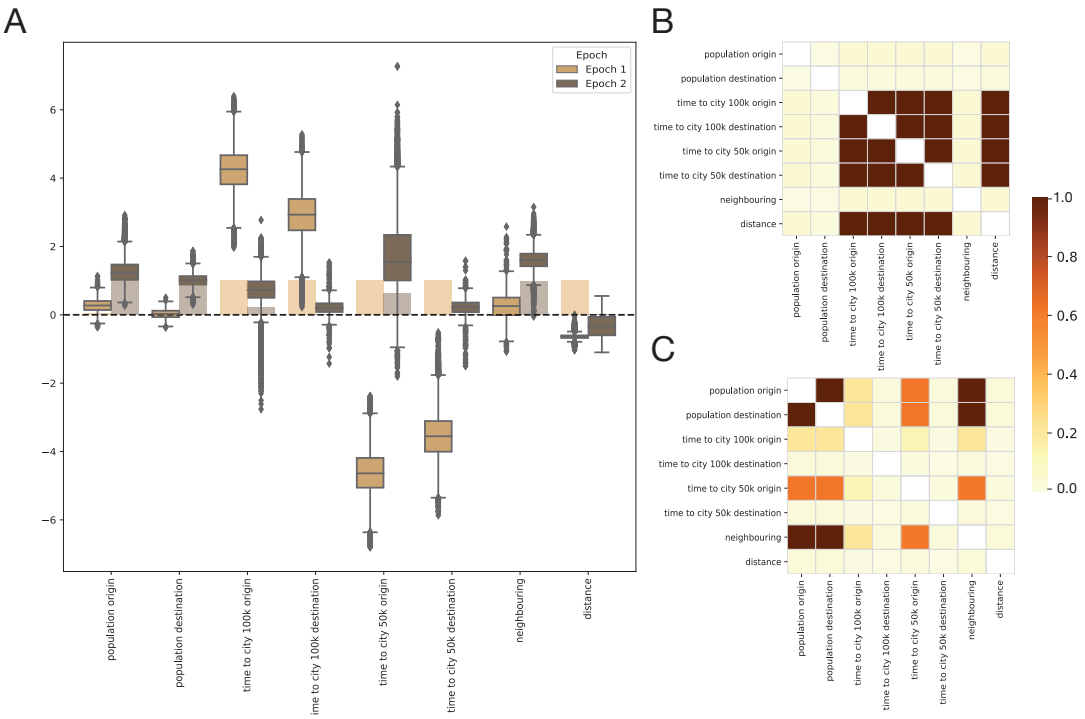

Fig. 2.3 Predictor results for the GLM on a district level. A) Effect size for the district level analysis, split by epoch. Bar charts show average inclusion between 0 and 1. B) Co-inclusion matrix for district level predictors in first epoch, darker colours indicate more co-inclusion. C) Co-inclusion matrix for district level predictors in second epoch, darker colours indicate more co-inclusion.

In general, at the district level, most movements were rural. This is shown by the strong positive effect size (Methods) of the travel time to a city of more than 100,000 people for both origin (4.24, 95% HPD: 3 to 5.4) and destination (2.9, 95% HPD: 1.57 to 4.2, Fig. 2.3A). The former was included in every version of the model, and the latter in 99.9% of them. In other words, if both the origin and destination were more rural, then there were more movements between them. Interestingly, the time to a city of more than 50,000 people had the opposite relationship, with a negative effect size for both the origin (-4.6, 95% HPD: -5.8 to -3.4) and destination (-3.5, 95% HPD: -4.8 to -2.2, Fig. 2.3A). These predictors were both included in every iteration of the model, and so this suggests that they are not competing for the same

signal (Fig. 2.3B). Bombali and Kenema were important districts in the model, and both of these have a higher travel time to cities of 100,000 people than they do to cities of 50,000 people (Fig. 2.4). Therefore, the apparent contradiction here may be explained by the dominance of these two districts in the reconstruction. Distance also had a weakly negative effect on movement (-0.6, 95% HPD: -0.8 to -0.52), and was included 99.9% of the time, indicating a slight preference for closer districts to share movements (Fig. 2.3A).

At the chiefdom level, travel time was not available and so greater-circle distance was used. Geographical distance to cities of 50,000 and 100,000 people for origin and destination all had weakly negative effects (Fig. 2.2A), and were only included in approximately half of the model iterations (distance to 50,000 origin and destination were 0.45 and 0.65 respectively, and distance to 100,000 origin and destination were 0.56 and 0.37 respectively). The more important predictors at the chiefdom level were the distance between locations, which were negatively associated with movements (-0.7, 95 HPD: -0.97 to -0.52, Fig. 2.2A) and supported the district level analysis; and if the chiefdoms were neighbouring then there was a strong correlation with movements (1.8, 95% HPD: 1.14 to 2.5, Fig. 2.2A). Both of these predictors were included in every reconstruction in the posterior. Therefore, even though locations like Nongowa seeded many chiefdoms that are far from it (Fig. 2.1E), if chiefdoms were neighbouring they were more likely to share a movement. It is of note however that this was not dependent on them being in the same district (-0.62, 95% HPD -1.16 to -0.14, included 17% of the time).

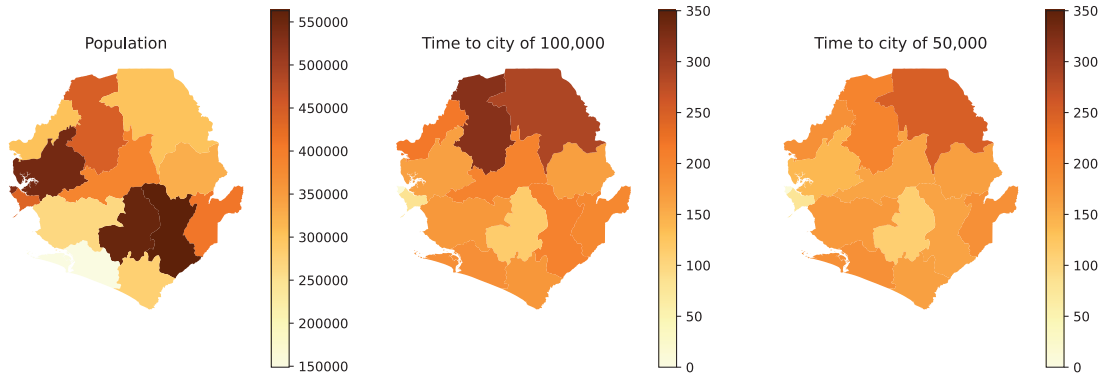

Fig. 2.4 Values of the three asymmetric district-level predictors: population, time to city of 100,000 people and time to city of 50,000 people.

Population size of origin and destination were rarely included at the district and chiefdom level (origin district: 0.04 , destination district: 0.01, origin chiefdom: 0.012, destination chiefdom: 0.006), but it is of note that chiefdoms containing capitals of the district (usually the largest city in the district) were highlighted when examining numbers of movements or network centrality. This applies to Maforki chiefdom in Port Loko, Freetown in WAU, Nongowa in Kenema and Bombali Sebora in Bombali. This implies that population is not necessarily the key to the importance of these locations, implying a role for connectedness or frequent travel to these locations independent of population size.

The first epoch was therefore characterised by movement from the source of the epidemic in Kailahun and from the chiefdom containing a major hospital in Kenema, as well as between major cities in the north of the country.

2.3.2 Epidemic maintenance

After cases stopped growing exponentially in early September 2014, peaked a little over two months later, and then declined, the dynamics of the epidemic changed in several ways. The key difference is that WAU, the district which contains Freetown, the capital of Sierra Leone, became the centre of most movements. It was especially connected with WAR, Port Loko and Bombali (Fig. 2.5). WAU had by far the most movements, with 39.2% (95% HPD: 33% to 48%) of all between-district movements in the epoch originating there; and 37% (95% HPD: 30% to 43%) ending there. This made it the most connected district, with a degree of centrality of 0.36 (95% HPD: 0.24 to 0.50). The next most connected district was WAR, a densely populated district adjacent to WAU on the Freetown Peninsula. This had 25% of movements beginning there (95% HPD: 17% to 31%) and 25% of movements ending there (95% HPD: 20% to 31%); and a degree centrality of 0.26 (95% HPD: 0.17 to 0.5). The whole Western Area therefore dominated the epidemic once established (Fig. 2.5A and Fig. 2.5C).

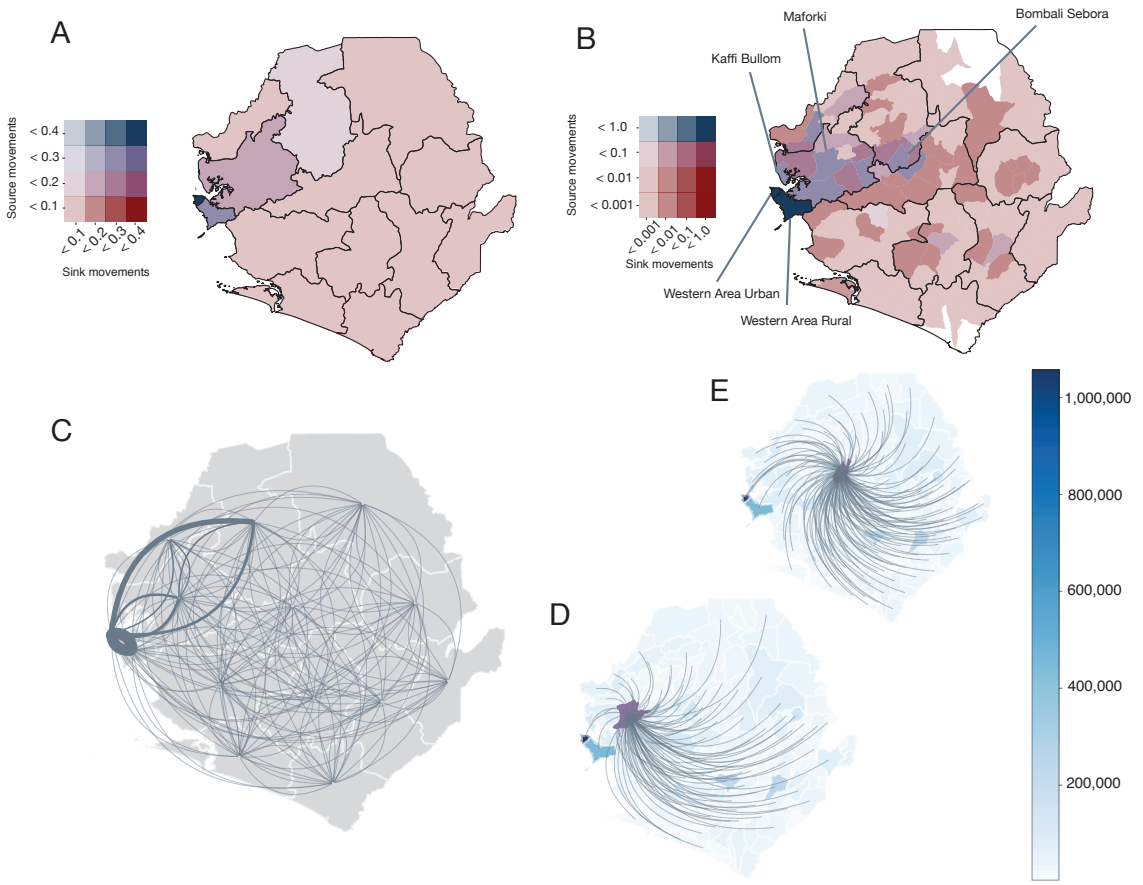

Fig. 2.5 Movements of Ebola virus across Sierra Leone in the maintenance and decline of the epidemic across the whole posterior. A) Bivariate choropleth map showing the proportion of movements in the epoch that start (red to blue spectrum) and end (alpha spectrum) in specific districts. B) Bivariate choropleth map showing the proportion of movements in the epoch that start (red to blue spectrum) and end (alpha spectrum) in specific chiefdoms on a log scale. Chiefdoms discussed in the section are indicated. C) Connections between pairs of districts, the thickness of the line is scaled corresponding to the frequency of movement from one location to the other. Direction is anti-clockwise. D) Movements leaving Maforki chiefdom, choropleth is population of each chiefdom E) Movements leaving Bombali Sebora chiefdoms, choropleth is population of each chiefdom.

There were two other districts which had some significant movement in this phase of the epidemic: Bombali and Port Loko (Fig. 2.5A). Bombali had the third highest percentage of movements starting there in this epoch, at 19.5% (95% HPD: 15% to 24%); and was not very sink-like, with only 6.39% (95% HPD: 4% to 9%) of

movements ending there. Bombali Sebora was an important chiefdom in Bombali (Fig. 2.5B), as it was the third most source-like chiefdom after the two Western Area chiefdoms, with 9.74% of movements beginning there (95% HPD: 7% to 13%). Bombali Sebora also had the third highest degree centrality, with 0.53 (95% HPD: 0.48 to 0.64), and was highly connected to most of the rest of the country, with movements ending in 147 out of 149 other chiefdoms (Fig. 2.5E).

Port Loko district had the third highest percentage of movements ending there in this epoch at 10.9% (95% HPD: 9% to 13%), and was the fourth most source-like district with 10.2% of movements starting there (95% HPD: 8% to 13%). Port Loko was the third most central district in the movement network after the two Western districts, with a centrality measure of 0.12 (95% HPD: 0.08 to 0.17). At a higher resolution, several chiefdoms in Port Loko district had high levels of movement both in and out. However, similar to the first epoch, two chiefdoms which stand out are Kaffi Bullom and Maforki: Kaffu Bullom was the fourth most connected chiefdom (0.46, 95% HPD: 0.44 to 0.51); and Maforki had the third highest percentage of movements ending there (3.9%, 95% HPD: 3% to 5%) and the fourth highest percentage of movements beginning there (6.5%, 95% HPD: 5% to 8%). However, Maforki only had movements ending in 95 out of 149 other chiefdoms (Fig. 2.5E), making it less well connected geographically than Bombali Sebora, the equivalent chiefdom in Bombali.

The second epoch followed the gravity model of transmission much more closely. WAU and WAR both have large populations, Bombali Sebora contains Makeni, the largest city in the Northern Province of Sierra Leone, and Maforki contains Port Loko city, the capital of Port Loko district. This can also be seen in the predictors which had large impacts and high probability of inclusion in the GLM: population of origin (99.9% inclusion) and destination (100% inclusion) were positively associated with between-district movements (Fig. 2.3A, origin: 1.26, 95% HPD: 0.71 to 2.07,

destination: 1.01, 95% HPD: 0.67 to 1.39). However, on a chiefdom level, these predictors are hardly ever included (average indicators 0.009 and 0.007 for origin and destination respectively, Fig. 2.2C). For districts, the time to cities of 100,000 and 50,000 people had a less positive effect on movements compared to the first epoch (Fig. 2.2A), and were less frequently included in the model. Interestingly, neighbouring districts were likely to share many movements (1.6, 95% HPD: 1.06 to 2.16, 99% inclusion), but neighbouring chiefdoms were slightly less likely to share movements with an effect size of -0.5, although it must be noted that the 95% HPD includes 0 (95% HPD: -1.37 to 0.27) and this predictor was only included 3.3% of the time. This is likely because chiefdoms are smaller, so it is easier to travel across multiple chiefdoms to get to e.g. an Ebola Treatment Centre. The two chiefdoms being within the same district also negatively affected movement (-1.4, 95% HPD: -2.2 to -6.89, 97.3% inclusion), likely due to the dominance of movements between WAU and WAR which are not in the same district.

2.4 Discussion

Gravity models can be an extremely useful way of modelling infectious diseases, and have been shown to be a good fit for the Ebola Virus epidemic across the affected region (Dudas et al., 2017). It is clear however that a gravity model does not explain all of the trends observed in the first epoch. Instead, much of this can be explained by the importance of seeding location, and movements between Kenema and Kailahun districts were important for the establishment of the epidemic. The GLM predictor analysis indicated more rural movements, and population indicators were rarely included at district or chiefdom level. This is despite the importance of Nongowa chiefdom, which is the third largest chiefdom in Sierra Leone (after WAU and WAR).

The second epoch fits a gravity model better, with Freetown, a densely populated city of 1.05m people, dominating the movement network. The whole Western Area, consisting of WAU and WAR, contains 21% of the population of Sierra Leone, and is responsible for 64% of the origins and 62% of the destinations of between-district movements. Maforki chiefdom in Bombali, which contains Makeni, the largest city in the Northern province of Sierra Leone, is also an important centre for movements. Population size and distance were also key predictors for movements in the district-level GLM, again expected under a gravity model and supporting results from Dudas et al. (2017) and Dellicour et al. (2018).

The geographical context of the country of interest can help to interpret deviations from the gravity model. In the first epoch chiefdom phylogeography, the distance to a city of 100,000 people (and also to 50,000 people) is negatively correlated with movements, whereas time to the same cities in the district analysis are positively correlated with movements. This apparent contradiction lies in the subtle difference between the two measures: distance is here defined as the greater circle distance between the centroid of the chiefdom to the centre of the city; whereas the travel time includes issues such as road quality. These both support the idea that the signal for these predictors' inclusions come from Kenema and Kailahun movements, as while Kenema city is large (approximately 200,000 people), transport in that part of Sierra Leone is difficult, and travel time in Kailahun in particular is disproportionately long compared to distance (Logistics Cluster, n.d.).

It is of note, that while reduced distance between two chiefdoms or districts increases the number of movements between them, many locations have direct connections to many other parts of the country e.g. Nongowa chiefdom in Kenema had movements ending in most other chiefdoms in the first epoch, despite a non-extensive road network. In Sierra Leone, rural-urban divides are common in families,

and in the words of anthropologist Paul Richards “social intimacy does not always equate with residential proximity” (Richards, 2016). This is made possible by non-traditional modes of transport, such as motorbike taxis, which are able to traverse difficult terrain but have only been present in Sierra Leone since the civil war which finished in 2002. For example, a man infected in Kenema travelled home to see a relative who was a herbalist in the village of Fogbo, near the centre of the country and approximately 130km and two districts away from Kenema, by motorbike taxi. This caused an outbreak in the village, which was then able to spread to Freetown, another approximately 200km, due to attendees at one of the funerals of the cluster (Richards, 2016; Richards et al., 2015). Therefore, even in the absence of public transport or an extensive road system, individuals were able and willing to travel hundreds of kilometres when infected.

Cultural and economic significance can also add additional context. For example, in the west of Sierra Leone, there are two chiefdoms which stand out in both epochs: Maforki and Kaffi Bullom. The former contains Makeni university, the largest private university in Sierra Leone, which would be expected to draw individuals from across the country. Similarly, Kaffi Bullom contains Lungi airport, Sierra Leone’s main international airport. They also both lie along the main highway which connects Freetown to Conakry, and so is an important trading route with regular high volumes of population movement. The impact of the reliance of trade vs subsistence farming for the bulk of the population in a region has been highlighted previously: the self-isolation practise of retreating to “corners”, developed during the civil war, either by returning to the family farm (traditionally) or by physically preventing transport in or out of villages (in one example, by cutting the bridge across a river between two villages), was much more possible in regions less reliant on trade (Richard et al., 2020). Richards *et al* explicitly highlight Port Loko and WAR as districts heavily

involved in trade and transportation and therefore individuals had less recourse to isolate themselves (Richards et al., 2020). In the east of the country, Jawei and Nongowa chiefdoms are important in the movement network, especially in the first epoch. Nongowa chiefdom contains KGH, which was critical in the early spread of the epidemic, with many patients travelling from far away to access healthcare there (the first patient detected in Sierra Leone was a woman treated in KGH after attending the funeral, World Health Organisation, 2015d). It is not immediately obvious as to why Jawei would be the source and sink of many viral movements, but in its capital Daru, there is one of the largest military barracks in Sierra Leone, which provides a possible reason for large scale population movement in and out of the town.

Sampling bias, especially at the chiefdom level, must be taken into account when considering these conclusions. For example, in the first epoch, WAU is not particularly important on a district level with the 4th highest network centrality, but on a chiefdom level has the highest network centrality. This could be because many more sequences have a district level assignment (n=1134) than a chiefdom (n=725), and WAU is both a chiefdom and a district. Therefore, on a chiefdom level, there are more sequences unambiguously assigned to it than other chiefdoms; as much of the chiefdom-level data will be lost and may be assigned to a different chiefdom in the same district in the model. The integration of human mobility data into an analysis like this would be useful to tease apart whether chiefdoms actually have more viral movements, or whether it is biased in the sequence database. Finally, it would of course be vital to include collaborators from Sierra Leone who would have a greater understanding of the importance of different areas in the country.

Previous studies have examined dynamics of the Ebola Virus epidemic on a regional level across all affected countries in West Africa. Here, using chiefdom level data, it is possible to explore dynamics at a deeper level. The higher resolution has

allowed us to see that the role of Kailahun as a source at the start of the epidemic is being driven substantially by Kissi Teng; further emphasising the role of the seeding location. In allowing the pulling apart of the role of different locations and relying on averaging smaller areas, more high-resolution analysis may provide evidence for better informed public health interventions; and shows the importance of connecting high detail epidemiological data to genomic data.

In this analysis, we see more unpredictable transmission while transmission is being established, before it settles into the more stable dynamics predicted by the gravity model. Ebola virus transmission appears therefore in this first epoch to be more impacted by stochastic events, such as seeding larger locations (i.e. Nongowa chiefdom in Kenema) or specific superspreading events. This may be explained by the epidemic being below its outbreak threshold: the number of infected individuals required for deterministic case growth and to ensure that the epidemic is unlikely to go extinct due to stochastic effects (Hartfield and Alizon, 2013). This threshold increases with greater heterogeneity in the onward transmission, which is significant in Ebola virus (Lau et al., 2017), and a lower R_0, which is approximately 1.5 in Sierra Leone (Khan et al., 2015), and so we expect Ebola virus to have a relatively high outbreak threshold. Therefore, before transmission was entrenched in Freetown and displayed the stable dynamics represented by the gravity model, the epidemic in Sierra Leone was more likely to go extinct due to drift. The seeding of Lagos in Nigeria with Ebola virus is an example of these sorts of dynamics: despite being a large and densely populated city, there were only 19 cases (Otu et al., 2017). While this is partially due to successful contact tracing and isolation efforts, it may also be partly attributed to these unstable and stochastic dynamics, and that there were no superspreading events to begin multiple transmission chains.

The interaction of low case counts and stochastic extinction implies that interventions at this more unstable stage of the epidemic will have a greater effect as it is easier to disrupt transmission; and preventing any superspreading events will have a disproportionate impact on the size of the epidemic. This further underscores the vital importance of early, rapid and effective response to epidemic control.

3

Agent Based Synthetic Epidemic

The more I learn about computers, the more it just boggles the mind. Everything is just a directory with stuff in it.

Dr Áine O'Toole

Personal communication

2022-02-02

3.1 Introduction

Understanding individual-level heterogeneity is central to accurately modelling the spread of infectious diseases (Lloyd-Smith et al., 2005; Lee et al., 2020). This is especially important for diseases where there is a high-level of variation in the number of secondary cases caused by each primary case. When an individual infects a much larger number of secondary cases than is expected, this is known as superspreading (Lloyd-Smith et al., 2005). While average factors such as R_0, the basic reproduction number, aid in the design of interventions to reduce the growth of the viral population (e.g. Davies et al., 2020), these measures can obscure vital differences which could result in more efficiently targeted interventions or better predictions when preparing for future waves.

Ebola virus disease (EVD) is one example of a disease with a high amount of transmission heterogeneity between individuals. Lau et al. (2017) found that superspreading was key to the dissemination and maintenance of the EVD epidemic in West Africa, with an age and time component to the level of superspreading. Ebola virus transmits via bodily fluids including blood, faeces, and, in the later stages of infection, RNA has been found in saliva and tears (Judson, Prescott, and Munster, 2015). Thus, it requires specific forms of contact in order to spread, and it is challenging to receive an Ebola virus infection from a casual contact. Common routes of infection include through the caring of an infected person, either at home or in a clinical setting (Wamala et al., 2010; Muyembe-Tamfum et al., 1999), and during funeral preparations and ceremonies (Tiffany et al., 2017), as the body remains infectious for up to seven days after death (Prescott et al., 2015). This means, for example, that individuals who have large funerals have more opportunity to infect large numbers of people. Therefore, due to both the high variation in who is likely to

be infected, and the different opportunities for individuals to infect secondary cases, there is a high level of heterogeneity in the numbers of secondary cases.

Previous EVD outbreaks have demonstrated non-classical dynamics. In the EVD epidemic in the Democratic Republic of the Congo (DRC) in 2018-2020, the case counts did not follow a smooth exponential increase and decrease as would be expected in a homogeneous transmission model. Instead, the case counts increased to a set level of 30-50 cases a day for approximately five months before increasing rapidly to a peak and decreasing to zero a few months later (World Health Organisation, 2020a). A potential mechanism for these dynamics can be found by studying the West African epidemic in 2013-2016, which was shown to have metapopulation dynamics (Dudas et al., 2017). That is, the epidemic consisted of small and short-lived clusters, with the frequent seeding of new susceptible contact networks. This could lead to the oscillation observed in the DRC, as new networks are seeded and depleted either via death or immunity at different time-points throughout.

Traditional epidemiological models may struggle to incorporate the level of heterogeneity and metapopulation dynamics of EVD as they often assume panmictic populations and use fixed transmission terms across large groups of the population. A further complication is that models are usually parameterised by fitting to epidemiological measures such as confirmed case counts, hospitalisation data or death counts (e.g. Kucharski et al., 2015). These all suffer to a greater or lesser extent from under-reporting: in a meta-analysis of 33 studies of notifiable diseases across the US, using all three data sources, Doyle, Glynn, and Groseclose (2002) found that case reporting completeness varied from 9% to 99%, mostly connected to which disease was being reported. Confirmed case counts rely on testing capacity that is often time- and resource-intensive (e.g. PCR), and is regularly not in place at the start of an epidemic; hospitalisation counts can become inaccurate as healthcare

systems become overwhelmed; and death counts often rely on a diagnosis, and so face similar issues to case counts, as well as a lack of all-cause death reporting in many countries (Ghafari, Kadivar, and Katzourakis, 2021). For all three measures, differences in healthcare-seeking behaviour (Oberoi et al., 2016) can lead to drastic underestimates of these measures in a non-homogeneous manner. For example, reduced affordability and accessibility of healthcare and negative beliefs about illness or attending healthcare facilities leading to a reliance on traditional practitioners were important in the undetected spread of EVD in West Africa (Gibbons et al., 2014; World Health Organisation, 2015a). Further, in previous EVD outbreaks, this has been exacerbated by community resistance due to a plethora of factors from misunderstanding of customs and traditions by international teams, to historical exploitation of populations by local and colonial powers (Ntumba et al., 2019; Wilkinson and Fairhead, 2017); and by outbreaks spreading in areas with multiple security issues e.g. Nord Kivu in the DRC, leading to difficulties in setting up testing and contact-tracing systems due to safety concerns (Matfess, 2018).

In order to address both the issues of individual-level heterogeneity and complex dynamics, I have developed ABSynthE (Agent Based Synthetic Epidemic): a stochastic agent based model (ABM). ABMs monitor the infection status of every individual in a population separately, allowing them to have different infection cycles and contacts, and therefore allowing different numbers of onward infections. In theory, this model will work for many different person-to-person viruses and countries, but as a test case I have developed it using the population of Sierra Leone and fitted it using 214 genomes sampled during the 2013-2016 outbreak during the exponential phase of epidemic growth (chapter 2). Previously, a similar approach has been developed to simulate the behaviours and location of the whole population of Poland in order to explore the spread of influenza through the country, with a focus on households

and mixing at school and workplaces (Rakowski et al., 2010). For EVD, Merler and colleagues developed an ABM to model the spread of EVD in Liberia in households, treatment centres and at funerals (Merler et al., 2015). ABSynthE differs from these models by using census data from 2014 to place individuals in three nested contact levels and allowing different contact patterns to exist at each level, and therefore implicitly rather than explicitly including geographical structure.

Instead of using traditional epidemiological data, I have then fitted this model to summary statistics from the phylogeny generated from the viral genomes collected during an epidemic, following Saulnier, Gascuel, and Alizon (2017). Phylogenetics is built around the concept of inferring unobserved events, and so is a useful approach to take when much of the epidemic is unsampled, and provides an additional data source to case, hospitalisation and death data. I use this model and dataset to explore how much larger the epidemic would have grown in the absence of interventions or behaviour change, and how this relates to the herd immunity threshold expected in a homogeneous population. I also build on the network analysis from the last chapter, to examine the importance of the seeding location for the Sierra Leonean epidemic.

3.2 Model design

ABSynthE is a Python-based stochastic ABM. It is pip-installable and called from the command line using, minimally, a directory containing population configuration file which contains the list of all locations and the total population size, and contact structure files, which are JSON files with the location of every individual, and lists of individuals in each contact level. ABSynthE can be found on github at https://github.com/ViralVerity/ABSynthE.

Broadly, the epidemic is run through a fixed contact structure, with a probability (derived from the fitting procedure) of exposing other individuals at each contact level. Each individual has an independent infection trajectory and probability of dying or recovering. Currently, each individual is itself homogeneous, i.e. there is no age or occupation structuring, but the variety in contact numbers based on location and infection trajectory provides individual-level heterogeneity. Each separate epidemic simulation can be run in parallel by specifying the number of subprocesses available, making the running of hundreds of simulations computationally tractable.

3.2.1 Contact structure and infection parameters

Each individual is assigned a household, a chiefdom and a district, which is based on data from the 2014 Sierra Leone census (Statistics Sierra Leone, 2016), fixed between simulation runs and provided as an input to ABSynthE. The names of these levels are specific to Sierra Leone, but is generalisable to administrative level 3 and administrative level 2 for any country for which there was sufficient data. Each district has a different number of individuals, households and average household sizes. Average household size was determined separately for each district by dividing the district population by the number of households. This number was then rounded so that a complete number of people were in each household (Table 3.1). There were slightly incompatible numbers in the census between districts and chiefdoms (i.e. different totals for each). By using chiefdom-level data, there were 0.34% fewer people than using the total census population i.e. 0.34% were unaccounted for in the 2014 census. The total population was taken as the total in chiefdoms (population = 7,068,220, households = 804,536).

When an exposure occurs at a given contact level, the individual selected to be exposed is taken from within the entire level. In other words, the smallest level of the contact structure is the exposure of a secondary household contact, and then, escalating upwards, within the same chiefdom (which may be within the same household), within the same district (may be within the same chiefdom) and finally, anybody in the country.

Each individual has an independent course of infection, the details of which are drawn from a series of probability distributions. The incubation time, which is assumed to be non-infectious (Dowell et al., 1999), and the two distributions for times to death and recovery are all gamma distributions taken from (WHO Ebola Response Team, 2014). Dimensions of all parameters can be found in Table 3.2.

The time to exposure of a secondary case is the time to infection is the mean serial interval (11.6 days) minus the incubation period (8.5 days) to obtain the mean of the gamma distribution, both drawn from WHO Ebola Response Team (2014). The dimensions are related to the infectious period of the individual, which is itself related to whether an individual lives or dies.

The mean recovery time is the time from symptom onset to hospital discharge, the only recovery time measure available. The average of this measure is 17.2 days (WHO Ebola Response Team, 2014), but an individual must receive two negative Ebola tests 24 hours apart in order to be discharged, and so the true mean recovery time is two days earlier, following Kucharski et al. (2015).

The fatality rate is set at 70% (WHO Ebola Response Team, 2014), and individuals remain infectious for seven days after death (Prescott et al., 2015).

3.2.2 Transmission parameters

The probability of infecting someone at each level is a negative binomial process with different infection modifiers for each level.

They are all based on λ, the within household transmission parameter. Therefore we have $\mathcal{A}\lambda$, $\mathcal{B}\lambda$, and $\mathcal{C}\lambda$ for within chiefdom transmission, within district transmission and within country transmission respectively. Within each level, the probability of infecting anyone else is constant i.e. once the secondary case has been assigned a level, there is an equal chance of anyone within that level being infected.

λ is drawn from a negative binomial distribution in order to capture the large variation in the number of secondary cases. As the negative binomial distribution is the combination of a gamma distribution and a Poisson distribution, I began by getting an estimate of the average of a Poisson distribution of cases. In Glynn et al. (2018), there was an average of 0.7 secondary cases in a household per primary case. Dividing this by the estimated over-dispersion parameter of 0.37 from Lau et al. (2017), gives the shape parameter of the gamma distribution of 1.89. The base transmission parameter is therefore a Poisson distribution where the lambda is drawn from a gamma distribution with the scale parameter 0.37, the shape parameter 1.89 and a mean of 0.7.

3.2.3 Seeding the epidemic

While a few imports of EVD into Sierra Leone have been recorded, most of the sequences fall into a single clade and can therefore be attributed to a single introduction. For Sierra Leone, the beginning of the epidemic was traced to the funeral of a traditional healer in Kailahun who had been treating patients across the border in

Guinea (World Health Organisation, 2015d). Due to detailed retrospective contact tracing efforts described in Goba et al. (2016) and Wauquier et al. (2015), many of the individuals infected at the funeral can be identified. For this iteration of ABSynthE, I used the traditional healer as the index case with 2 household contacts, 9 chiefdom contacts and 3 district contacts. No contacts outside of the district were found during contact tracing. All of these secondary cases are infected on day 0, but then are otherwise treated as any other case in the epidemic in terms of their own secondary cases and infection trajectory.

3.2.4 Infection algorithm and coalescent phylogeny generation

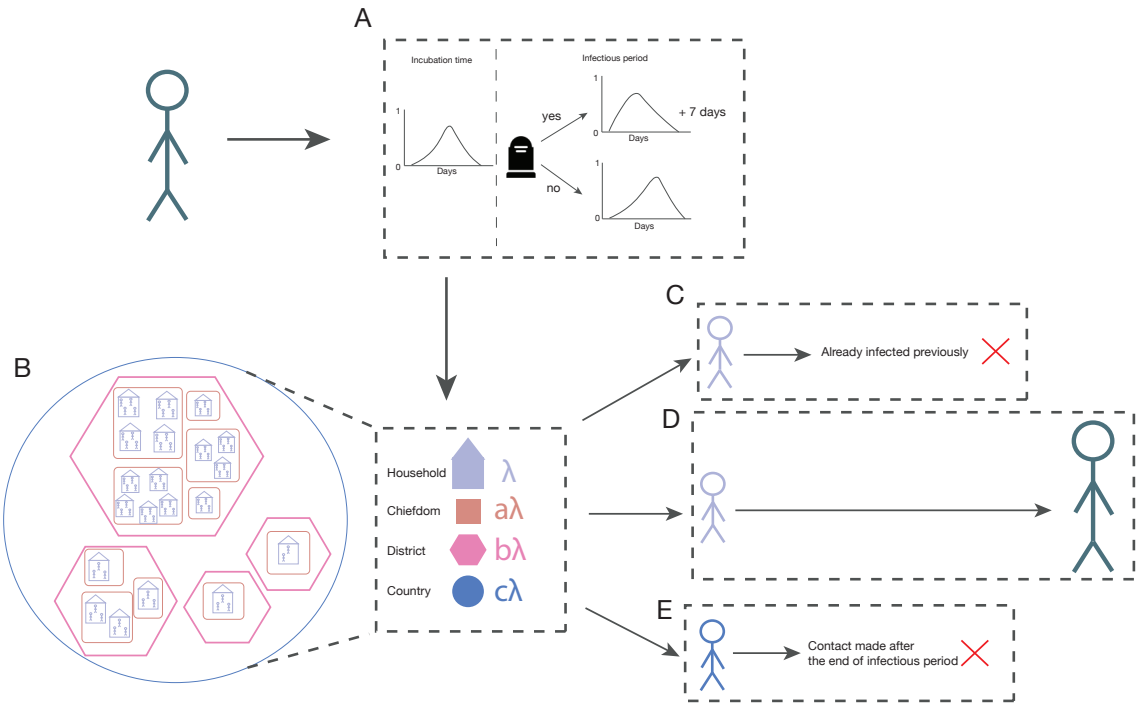

Fig. 3.1 Schematic showing the algorithm for a single infected individual. A) The individual's infection parameters are drawn from a series of distributions. Whether or not an individual dies determines the length of the infectious period, as if they do then the time to death plus 7 days and if not then time to recovery is used. Both time to death and recovery are drawn from gamma distributions. B) Secondary exposures are determined using fitted transmission terms that vary depending on contact level. The underlying contact structure is determined in advance and remains fixed for specific countries and scenarios. In this case, there are two household-level exposures and a country-level exposure. The first household-level exposure (C) has already been a case and so cannot be infected again, the second household-level exposure (D) is successfully converted into a case and begins its own algorithm by determining its infection parameters. The country-level exposure (E) was contacted after the end of the individual's infection period and so does not become a case.

Broadly speaking, there are two core processes which define the epidemic algorithm. The first process is where we define potential cases (i.e. exposed individuals) as case objects, which are then assigned to a specific member of the population if the

exposure is successful. The other process is a temporal one, going through the days of the epidemic and monitoring the progression of individual infections.

The core of the model goes through the days of the epidemic and monitors whether there are any new exposures on that day. If the infector is still infectious, then the case is assigned to a random individual within the appropriate contact level. If this individual has not been successfully infected previously, they now have their infection parameters determined (as above), and any secondary exposures at different contact levels are determined. Those exposures are connected to future days in the epidemic, ready to be evaluated when the algorithm reaches the day they are exposed. If the individual has already been infected previously, or the infector has finished their infection, the case ID is assigned to nobody, and is removed from any relevant data structures at the end of the simulation. This is shown in a schematic in Fig. 3.1.

The end-point of the epidemic can be specified with a case or day limit, or can be left to run until either the whole population is infected or there are no more cases. For monitoring the epidemics, comma-separated value (CSV) files can be produced describing key features of every case in the epidemic, R_0 values can be calculated, and phylogenies can be inferred from the transmission tree. From these phylogenies, lineages through time and skyline data for making plots of these metrics can be generated as well.

The phylogeny is simulated from the full transmission tree using a standard algorithm. Briefly, for each individual infected, a subtree is generated containing: the root, a tip representing the individual (if sampled), any number of tips representing sampled onward transmission, and any coalescent nodes. Each subtree therefore represents the within-host process of infection, with the root representing the point of infection of the focal individual, transmission tips representing the point of infection

of any secondary cases, and the sample tip representing when that individual was sampled (the same day as symptom onset in the base case). Within each subtree, if there are any tips to be joined in a coalescent process, random pairs are selected from whatever lineages are active at a given waiting time (τ) given by the coalescent equation under a constant population model.

Subtrees are then connected root to transmission tip based on parent/child relationships in the transmission tree. Transmission tips and nodes along are removed, leaving only sampled tips and coalscent nodes. If there is no downstream sample, subtrees are removed so that the tips are all samples. This is shown in a schematic in Fig. 3.2.

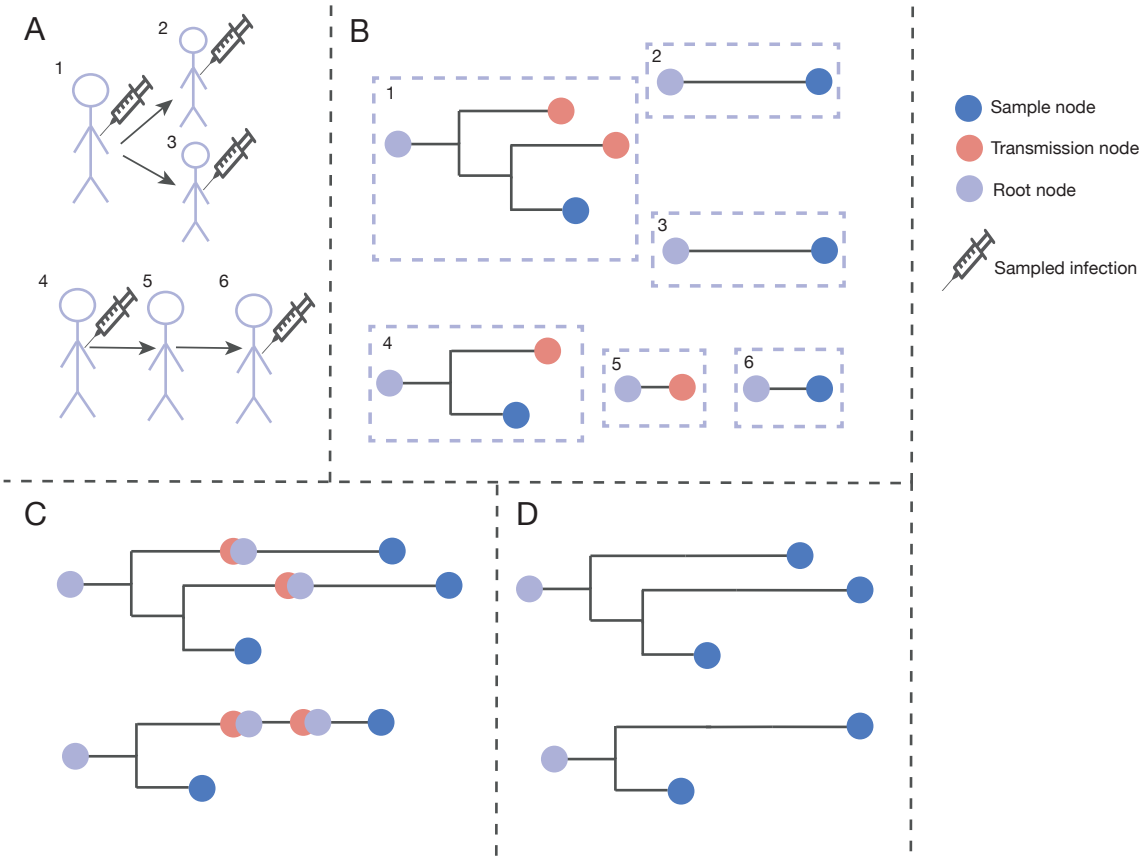

Fig. 3.2 Schematic of phylogenetic simulation. A) Full transmission dynamics from infection simulation, arrows denoting direction of transmission and syringe icons denoting a sampled infection in two different scenarios. B) Subtree is inferred for each individual with the root (purple circle) at the start of infection, if the individual is sampled (e.g. individual 1 but not individual 5) then there's a sample tip at the start of symptoms (in the base case, blue circle). In the case where there are more than two tips, the pairing of lineages is random dependent on what lineages are active at the time. C) Subtrees are connected transmission tips to root nodes, based on the transmission tree. D) Internal nodes (i.e. root and transmission nodes) are removed leaving only a single root and sampled tips of those who are sampled.

3.2.5 Fitting

The observed data is produced in the phylogeographic analysis described in chapter 2. Specifically, for fitting $\mathcal{B}$ and $\mathcal{C}$ (the within-district and within-country transmission modifiers), the observed data is the markov jump counts for the chiefdom and district

phylogeography from the older epoch. For $\mathcal{A}$ (the within-chiefdom transmission modifier), a series of summary statistics from Saulnier, Gascuel, and Alizon (2017) are calculated on the maximum clade credibility (MCC) tree, pruned to the first epoch. Briefly, five sets of statistics were examined: those involving branch lengths (n=12), tree topology (n=8), lineages through time (LTT) statistics (n=7), averaged LTT coordinates (n=40), and all of the above combined (n=67). The number of tips was also added to each set.

For simulated data, ABSynthE was run with a sampling proportion of 16% of cases (calculated as the number of cases in Sierra Leone divided by the number of sequences) for 124 days. This represents an estimate of the number of days from the original funeral that started the main set of transmission lineages in Sierra Leone to the end of the exponential growth phase. I estimated the date of the funeral as the 1st May 2014, as the death was recorded to be on the 30th April 2014, and for the Kissi people (the dominant group in the Kissi Teng chiefdom) funerals happen the day following death at the latest (Fairhead, 2014). I estimate the end of exponential case growth as the 1st September 2014.

This observed data was fitted to simulated data from ABSynthE using approximate Bayesian computation with sequential Monte Carlo sampling (ABC-SMC), using the package pyabc (Klinger, Rickert, and Hasenauer, 2018). Broadly, this works as follows: parameters for $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are drawn from their priors (all uniform distributions between 0.5 and 1 for $\mathcal{A}$, 0 and 0.5 for $\mathcal{B}$ and 0 and 0.5 for $\mathcal{C}$). These are then used to simulate a series of epidemics, and the proposed parameter values are accepted or rejected based on a threshold, ϵ, of the euclidean distance of a normalised vector of each summary statistic between the observed and simulated data. In the SMC formulation of ABC, ϵ is gradually decreased between generations, and is in this case based on the distance of accepted samples from the previous generation. New

parameter values for the current generation are drawn as a weighted sample from the previous generations' posterior distribution.

Each simulation was conditioned on the size of the epidemic. If the number of cases was too low (i.e. if fewer than three individuals were able to be sampled at a 16% sampling rate), then a phylogeny could not be generated and no summary statistics were calculated. Further, epidemics which led to more than 3000 cases in the 124 days allotted were not included and stopped running in the interests of efficiency. Scripts for the fitting procedure can be found in https://github.com/ViralVerity/ABSynthE/tree/master/Fitting/scripts, including scripts for generating summary statistics for observed and simulated data.

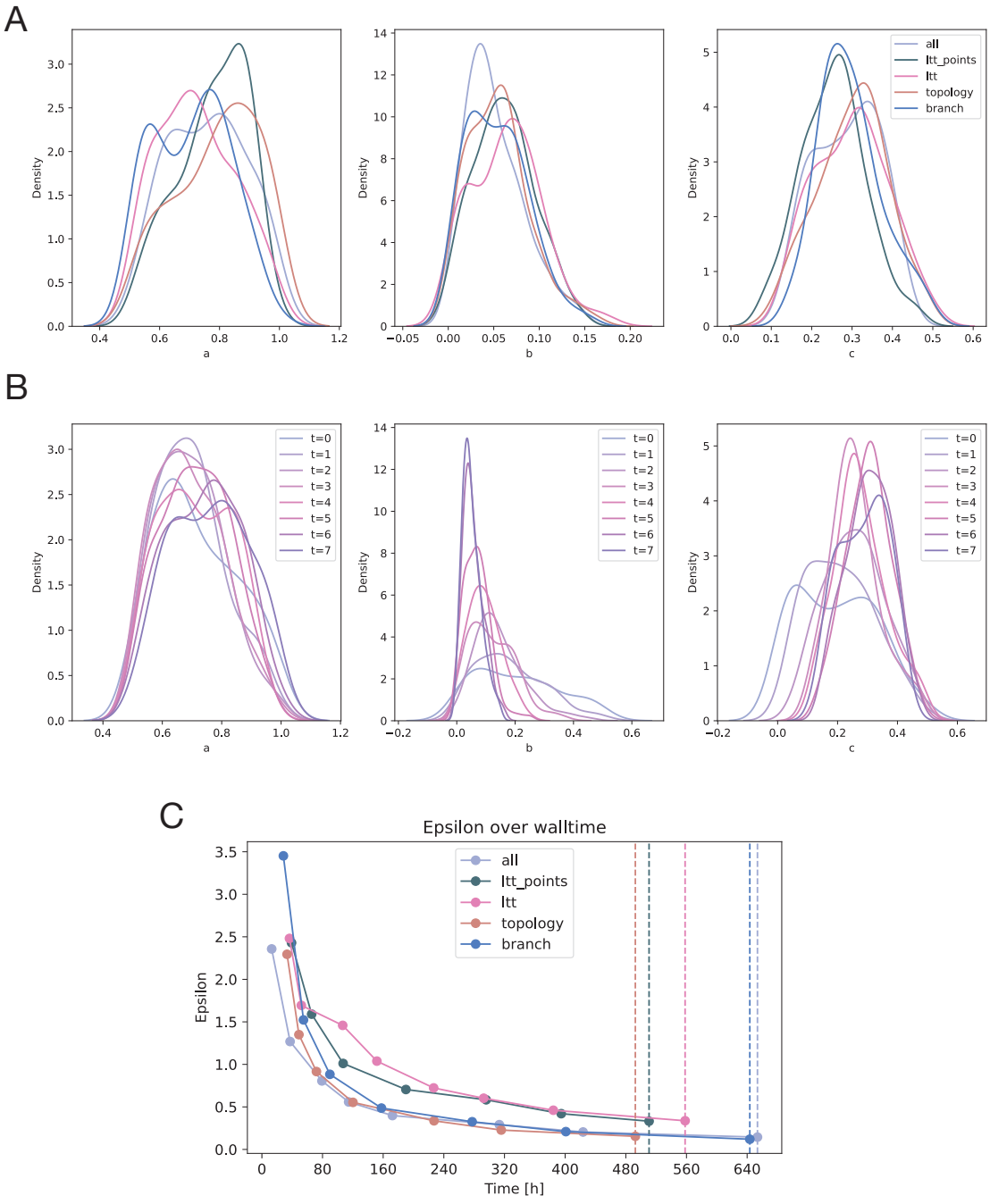

Fig. 3.3 Results of fitting procedure. A) Posterior densities of each parameter for each summary statistic by the final generation of the fitting procedure. Note that all of the final generation posterior densities are similar. B) Posteriors of each generation for all summary statistics combined C) Distance threshold epsilon against wall-time for each set of summary statistics.

Each of the four sets of individual summary statistics was run for seven generations, and the combined set of statistics was run for eight. The final ϵ, i.e. distance threshold between the observed and summary statistics for all the summary statistics combined was 0.14. When examining the summary statistic sets separately, it appears that they all give similar distributions for $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ (Fig. 3.3A), although the final ϵ is different between the different sets of statistics: the topology set was 0.15, the branch length set was 0.12, the LTT metrics set was 0.34 and the averaged LTT points was 0.33 (Fig. 3.3C). Thus it appears that in the combined set of summary statistics, the topology and branch length sets were responsible for much of the signal in the decreased distance between the observed and summary statistics. However, the single statistic provided to fit each of $\mathcal{B}$ and $\mathcal{C}$ (i.e. the jumps between chiefdoms and districts respectively) was more effective than the large sets of statistics provided to fit $\mathcal{A}$. This can be seen in the effective narrowing of the posterior distributions of $\mathcal{B}$ and $\mathcal{C}$ between generations, which is not seen as clearly for $\mathcal{A}$ (Fig. 3.3B). Potentially fitting $\mathcal{A}$ as a separate procedure (and for more generations) and fixing $\mathcal{B}$ and $\mathcal{C}$ would lead to a more accurate estimation of $\mathcal{A}$, as well as excluding the two LTT sets as these do not appear to be as informative for comparing observed and simulated data as the branch length and topology metrics are.

The final values for $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ from combining all of the summary statistics together are $\mathcal{A}$=0.759 (95% HPD: 0.53 to 0.983), $\mathcal{B}$=0.052 (95% HPD: 0.011 to 0.127) and $\mathcal{C}$=0.288 (95% HPD: 0.099 to 0.42). In other words, compared to the within-household rate of infection, the within-chiefdom rate is approximately three quarters, the within district rate is approximately 5%, and the within-country rate is just under a third. The high within-chiefdom rate is expected, as chiefdoms are mostly relatively small geographically, and so many individuals may live nearby within a short distance of each other. More apparently contradictory is the small within-district

value ($\mathcal{B}$), and then larger within-country value ($\mathcal{C}$). However, in chapter 1 when I ran the phylogeographic analysis that provides ABSynthE with observed data to fit to, I found that chiefdoms within the same district were less likely to share movements with each other than ones in different districts, and this was true across the epidemic.

3.3 Results

3.3.1 Base case

Using ABSynthE, I simulated the Sierra Leone arm of the West African EVD epidemic without interventions. That is, I ran 100 iterations of ABSynthE using the parameters $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ from the fitting procedure, which was based on the exponential growth phase of the epidemic and before any substantial interventions were introduced. It must be noted that some spontaneous behaviour change was likely happening in this part of the epidemic, but this is complex to measure. Therefore I cannot be certain that these parameter values include no behaviour change at all, but it is without much of the knowledge and experience that led to the significant community-led action that came later on in the epidemic.

The distribution of size and persistence of these 100 iterations is bimodal (Fig. 3.4A and Fig. 3.4B), as expected under a metapopulation model (Watts et al., 2005), and models taking individual-level heterogeneity into account (Lloyd-Smith et al., 2005). 21% of the iterations were small (fewer than 100 cases), and the smallest epidemic was 18 cases. Noting that all epidemics had to have at least 15 cases as this is fixed as the seeding event (see above), this smallest epidemic therefore only led to three additional cases after the seeding event, and they were each infected by a different person. These small epidemics were not solely restricted to Kailahun as

might be expected, and 8/21 of them reached Western Area Urban. This suggests that while rural epidemics are less likely to become very large, it is not inevitable that once the epidemic reaches a large urban centre it will explode. It is also of note that even when conditioning on an epidemic that did grow quickly, and began with a large super-spreading event, one fifth of the time it did not grow substantially. These small epidemics could sometimes still last longer than anticipated - the longest of these small epidemics was 207 days (approximately 7 months) and was 89 cases, which is slightly longer than historical equivalently-sized epidemics. For example, an EVD outbreak in Uganda in 2007 had 131 cases and lasted approximately 140 days from first case to the resolution of the final case (MacNeil et al., 2011). These discrepancies likely arise from a different context (as there are no previous EVD epidemics in Sierra Leone to compare this to), and because in real epidemics it is more difficult to track the index case, especially in transmission chains which do not have super-spreading events on them, and so the reported epidemic may be artificially truncated.

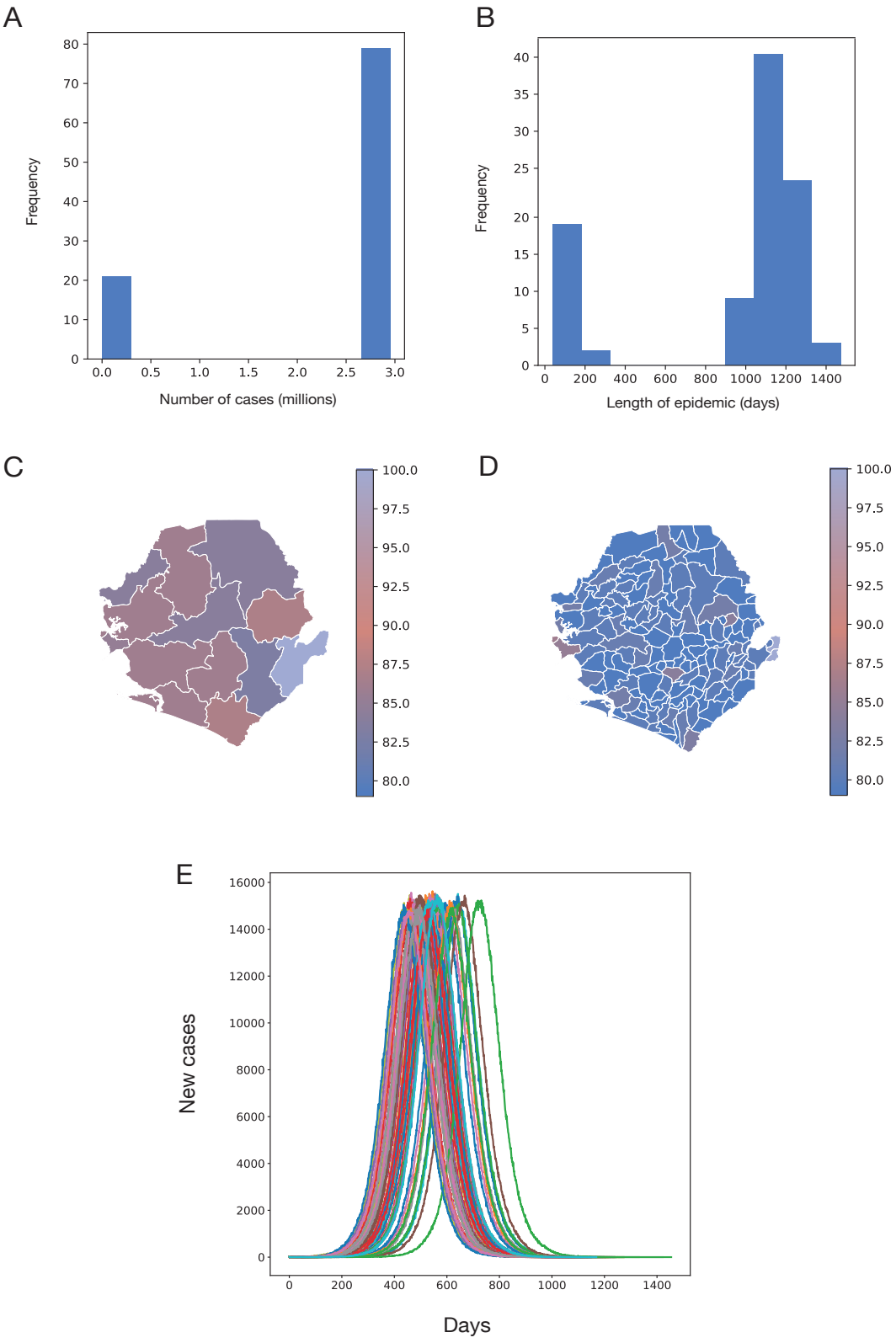

Fig. 3.4 A) Histogram of number of cases B) Histogram of length of epidemic from first case infection to the resolution (i.e. death or recovery) of the last case C) Choropleth of percentage of times each district has at least one case. Note that the legend begins at 80%. D) Choropleth map of percentage of times each chiefdom has at least one case. E) New cases per day for each separate epidemic simulation. Each epidemic that is successful has a smooth exponential increase and decrease.

Of the 79% of epidemics that were above 100 cases, they were all between 2.8 and 3 million cases, with the largest recorded epidemic at 2,950,437 cases. This is approximately 42% of the population of Sierra Leone. The percentage of cases infected at the peak of the epidemic curves for these large outbreaks is 20.7% on average (95% HPD: 18.9% to 22.5%). The herd immunity threshold for this structured population is therefore around a fifth of the population, and the remaining cases (approximately another fifth of the population) represents epidemic overshoot after this threshold has been reached.

The shortest of these large epidemics was 983 days (approximately 2.5 years) from seeding even to the resolution of the last case, and the longest was 1296 days (approximately 3.5 years, Fig. 3.4B). The length and persistence of these simulated epidemics show the vast impact that all the interventions and behaviour change combined had on the epidemic, reducing case and death counts by orders of magnitude, and limiting the main part of the epidemic (not including flare-ups associated with persistent infections) to approximately 18 months. All of the epidemics also show the same pattern of growth and decline (Fig. 3.4E).

On a geographical level, most of the districts are not seeded in the small epidemics and are seeded in the large ones (Fig. 3.4C). This follows the true course of the epidemic in Sierra Leone, even for districts which are poorly connected and very rural. It is similar on a chiefdom level, except that there are many more chiefdoms which are not in the smaller epidemics (Fig. 3.4D).

3.3.2 Application: the importance of starting district

In order to further explore the role of seeding location in the ultimate size and length of the epidemic, and to investigate whether there is anything inherent to

Kailahun that would lead to a large epidemic or constrain the size of the epidemic, I simulated 100 epidemics starting in each district. Each epidemic started with the same superspreading event of 14 secondary cases, and began in a randomly selected chiefdom within the appropriate district.

In general, the patterns for each district are similar to each other. Like the real epidemic starting in Kailahun, epidemics starting in the other districts have bimodal distributions: that is, the epidemics either remain local and small, or they become extremely large. However, where Kailahun either only produced epidemics which were less than 100 cases or more than 2 million, some districts are able to produce epidemics that are slightly larger before they escalate fully. For example, there were eight epidemics in Bonthe between 100 and 200 cases. The largest of these more intermediate outbreaks was one that began in Kenema and infected 585 individuals before ending.

However, most epidemics, no matter which district they begin in, become large (more than 1000 cases), and grow to between 2.9 and 3 million cases (Fig. 3.5A) in 1000 to 1500 days (Fig. 3.5B). There is slight variation in how likely it is that an epidemic will grow large: only 71 of the 100 epidemics that began in Western Area Rural infected more than 1000 people (explaining its much wider interquartile range in Fig. 3.5). On the other hand, 91 out of the 100 epidemics beginning in both Western Area Urban and Pujehun infected more than 1000 people, as did 90 of those starting in Kono, Moyamba and Bombali.

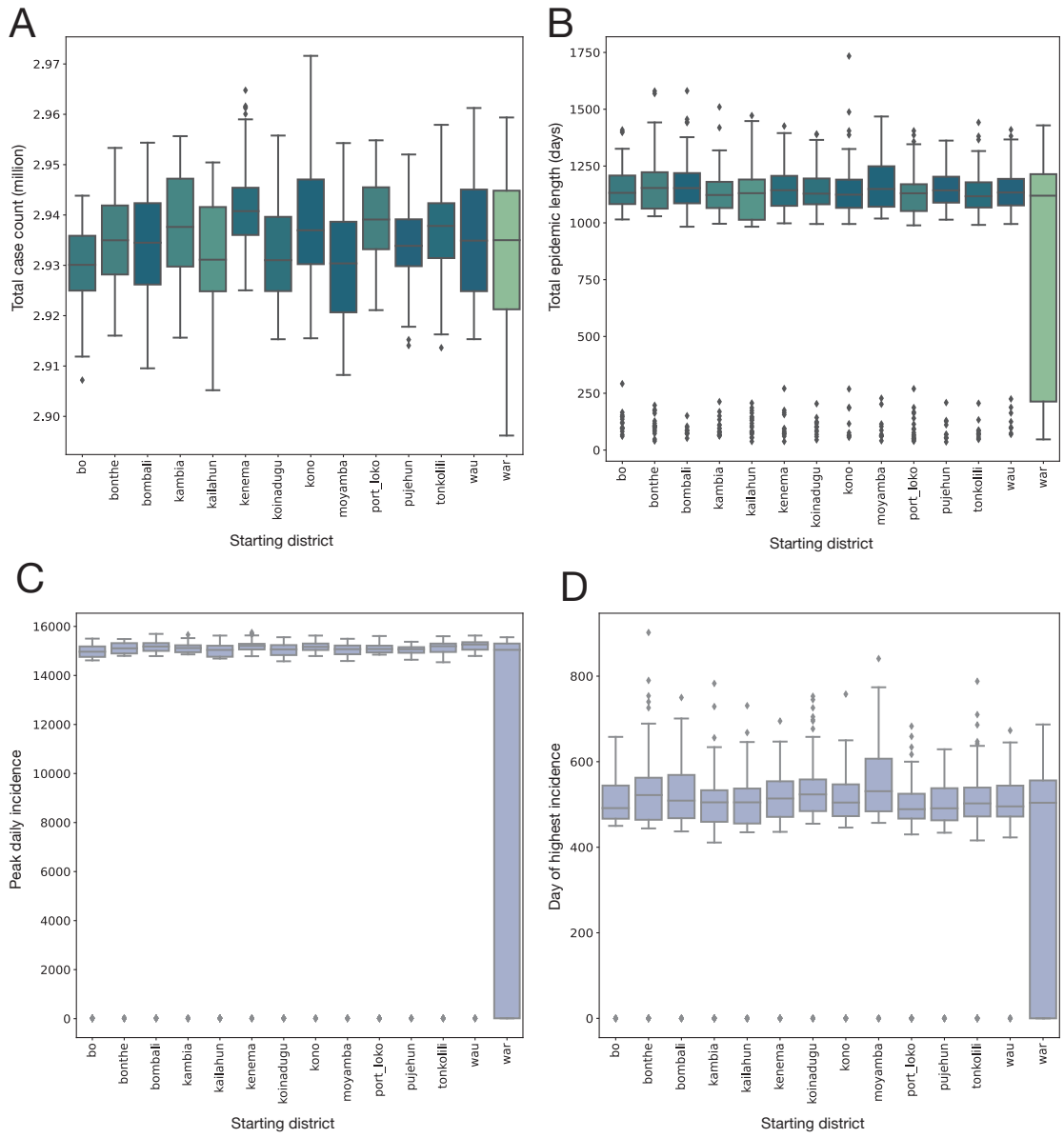

Fig. 3.5 Results of simulating the epidemic in Sierra Leone starting from different districts. A) Distribution of final case count in millions of cases across 100 simulations starting in each different district. Only epidemics with more than 1000 cases are shown here. Boxes are coloured by the percentage of epidemics over 1000 cases, with darker colours indicating more large epidemics. B) Total length in days across 100 simulations starting in each district. Boxes are coloured by the percentage of epidemics over 1000 cases, with darker colours indicating more large epidemics. C) Distribution of highest daily incidence D) Distribution of time in days from the first case of highest daily incidence.

It might be expected that more rural districts, or those with smaller average household sizes, should grow slower, perhaps having the same final epidemic size but taking longer to achieve this. However, in these simulations, this does not appear to be the case, with the average peak of cases at about 15,000 cases per day (Fig. 3.5C) at around 500 days after the first case (Fig. 3.5D) for all starting districts. Therefore, while the starting district does matter in terms of how likely a large epidemic is, the patterns once the epidemics become large are extremely similar regardless of where the epidemic began.

3.4 Discussion

Individual-level heterogeneity in terms of infection progression and contact structures can result in variation in local dynamics leading to surprising aggregate dynamics. This is especially important in EVD epidemics, which have large variations in the number of secondary infections, partially due to the highly specific form of contact required to transmit EVD. ABSynthE simulates every individual in Sierra Leone and places them in a hierarchical contact structure in order to try to capture this heterogeneity and accurately simulate EVD epidemics.

A key result of simulating the epidemic in Sierra Leone, mimicking the starting conditions as closely as possible, is that 79% of the time the epidemics were extremely large. They infected approximately 40% of the population before burning out, lasted for several years and did not reach the peak of infections for 1.5 years. Additionally, approximately half of these cases occurred after the peak of infections, and therefore the herd immunity threshold, had been reached. This highlights the critical significance of the behaviour change and interventions that were put into place, even if some of them were not started until after the optimal time point, and

the importance of not simply relying on herd immunity even with a milder disease. This threshold is lower than the herd immunity threshold expected in a homogeneous population however: this is predicted to be approximately 1/3 of the population of Sierra Leone, as the R_0 is estimated to be 1.5 (Khan et al., 2015). Restricting mixing means that infections are clustered, and so susceptibles are depleted faster and each subsequent infector is less likely to come into contact with a a susceptible individual without seeding a new contact network. Adding population structure has been shown to reduce the herd immunity threshold for SARS-CoV-2 (Britton, Ball, and Trapman, 2020).

ABSynthE does fail to capture dynamics that would be expected from a truly accurate simulation. In particular, every epidemic has a smooth exponential increase and decrease, and so does not show the temporal resurgence that is observed in real epidemics, and in other ABMs. This is possibly in part due to a homogenous district structure, so epidemics cannot get "stuck" in a remote location and then resurge once they reach a densely populated region again; as well as the choice to obtain transmission parameters from fitting to the exponential growth phase of an epidemic. To solve the former, I have designed but not yet implemented an additional transmission parameter which distinguishes districts which are adjacent from those which are not. This would be then fitted using data derived from the phylogeographic analysis in chapter 2, which contains whether districts are adjacent or not as a predictor in the generalised linear model. Therefore the observed data to fit to would be the markov jumps between adjacent and non-adjacent districts. A further extension of ABSynthE for EVD which may make it mimic real epidemics more closely would be the explicit inclusion of age and movement to locations such as treatment centres. The inclusion of age would lead to different interactions within households, for example younger children and parents were highly at risk from

other infections in the household, but older children are less at risk due to different care-giving behaviours (Richards, 2016).

In theory, ABSynthE is currently able to take any four level contact structure, provided the correct input files are supplied. This would enable simulation of an EVD epidemic in a new country, enabling the calculation of level of risk in an epidemic where, like in West Africa, the countries have not dealt with EVD before. However, this would still have the transmission parameters from the Sierra Leonean context, and are conditioned on an epidemic which did grow and spread to every part of the country, and so careful thought would be required to apply the model to the country in question. Finally, ABSynthE could also be used to simulate another virus which transmits person to person, by replacing the parameters which underlie the infectious process.

A useful feature of ABSynthE is that it produces phylogenetic trees. Here, that is used to fit the model to phylogenetic data to see if it is possible to avoid using traditional epidemiological data. For between chiefdom and district movements, the fitting procedure appears to successfully shrink the posterior distribution around a value. For between-household movements, this is less successful, at least partially because the statistics used in this fitting are far more complex. This parameter may simply need more generations than the others to achieve a similarly narrow posterior distribution. This level of contact structure in general is more difficult to estimate than the others because there is currently no direct data to compare it to, as genomic metadata is not commonly collected on this scale. Further, empirical studies do not capture the variation of between-household movement across an entire country,and are far less common than within-household studies.

The true utility of a detailed and flexible simulator that produces phylogenies lies less in the recreation of a specific epidemic, and in the ability to test our assumptions

that underlie phylodynamic methods. Currently, the sampling procedure is random, but this can be changed to explore the percentage of cases required to draw correct conclusions about movement between locations, growth rates, and R_0 under different sampling strategies as both the true and inferred phylogenies are known. This could help to design sampling strategies for future epidemics, as well as provide evidence that the conclusions we draw from coalescent methods are close to reality.

3.5 Supplementary tables

District	Population size	Number of households	Average rounded household size
Bo	575,478	69,009	8
Bombali	606,544	71,056	9
Bonthe	200,781	27,129	7
Kailahun	526,379	53,166	10
Kambia	34,574	37,870	9
Kenema	609,891	64,751	9
Koinadugu	409,372	42,029	10
Kono	506,100	56,770	9
Moyamba	318,588	53,516	6
Port Loko	615,376	69,675	9
Pujehun	531,435	32,421	11
Tonkolili	531,435	54,595	10
Western Area Rural	444,270	63,087	7
Western Area Urban	1,055,964	106,343	10

Table 3.1 District characteristics underlying the simulation of the Ebola Virus Disease epidemic in Sierra Leone, taken from the Sierra Leone census 2014.

Parameter	**Mean (days)**	**Standard deviation (days)**
Time to symptom onset	8.5	7.6
Time to death	8.6	6.9
Time to recovery	15.2	6.2
Time to secondary infection	3.1	2.5

Table 3.2 Infection parameter details, taken from WHO Ebola Response Team (2014). They are then used to obtain the shape and scale parameters of Gamma distributions. NB time to death, recovery and secondary infection are after symptom onset.

4

The origins and molecular evolution of the SARS-CoV-2 lineage B.1.1.7 in the UK

so like what even happened with B.1.1.7

Dr. Emily Scher

Personal communication

2021-05-06

4.1 Introduction

In early December 2020, one of the four UK public health agencies (Public Health England, PHE, now known as the UK Health Security Agency, UKHSA) began tracking and investigating a rapidly-growing cluster of cases in South East England, centred on Kent and East London. Numbers of new cases had grown more rapidly than expected over the previous four weeks, despite an elevated level of non-pharmaceutical interventions (NPIs) in the region, and increased incidence had begun to be observed in other locations in the UK, indicating further spread (Public Health England, 2020c). The cluster was also detected separately within the COG-UK virus genomic surveillance dataset, and the genome sequences carried a substantially larger than usual number of genetic changes (Rambaut et al., 2020b). At a routine PHE meeting on the 8th December 2020, the link between the unusual mutation constellation and the Kent epidemiological situation was made and investigations were rapidly initiated to characterise the mutations and to estimate the growth rate of the cluster. Evidence accumulated that this cluster was growing rapidly and had expanded throughout November, during a national lockdown in England. The cluster was designated B.1.1.7 under the Pango lineage naming system (Rambaut et al., 2020a), and was later labelled as variant of concern (VOC) Alpha under the World Health Organisation (WHO) variant nomenclature (Konings et al., 2021).

Since its discovery, substantial analytical effort has been put into teasing apart the contributions of human behavioural factors and true virological effects on the rapid growth of the lineage (Kraemer et al., 2021). It is now clear that Alpha was associated with a higher transmission rate than the background D614G lineages that dominated in the UK at the time (Volz et al., 2021a; Davies et al., 2021a) as well as a higher case fatality rate (Davies et al., 2021b).

The constellation that defines the Alpha variant contains 14 lineage-specific amino acid replacements and three deletions compared to the then-circulating background lineages (Rambaut et al., 2020b), which was unprecedented in the global virus genomic dataset for the COVID-19 pandemic at the time of its emergence (Table 4.1). This constellation includes several mutations that have arisen independently in other VOCs. For example, N501Y in the Spike protein, also found in Beta (B.1.351), Gamma (P.1) and Omicron (B.1.529 descendants), is a key contact residue in the receptor binding domain (RBD), and experimental data has determined that it increases binding affinity to human and murine ACE2 (Starr et al., 2020; Tian et al., 2021), and has been associated with increased infectivity and virulence in a mouse model (Gu et al., 2020). There are also two deletions of interest in Spike: six base pairs at position 21765 (amino acid positions 69-70) and three base pairs at position 21991 (amino acid position 144). Both have previously arisen in chronically infected individuals (Choi et al., 2020; McCarthy et al., 2021; Avanzato et al., 2020). The former was also associated with the rapid outbreak in mink in Denmark (Oude Munnink et al., 2021), and has been shown *in vitro* to increase infectivity (Meng et al., 2021); and the latter has been shown to prevent monoclonal antibody and, to a lesser extent, convalescent antisera binding (Andreano et al., 2020; Collier et al., 2021), as well as a decrease neutralisation efficiency (Weigang et al., 2021). Further, Alpha contains a 9 base pair deletion in NSP6, also found in the VOCs Beta, Gamma and Omicron, which is on the outside of the autophagy vesicle, theoretically limiting autophagosome expansion (Benvenuto et al., 2020). There is also a mutation in the accessory protein ORF8, which truncates the protein from 121 to 27 amino acids in length, likely resulting in loss of function and allowing further downstream mutations to accrue. Subsequent work has found that the ORF8 deletion has only a modest deleterious effect on virus replication in human primary airway cells compared to viruses without the deletion

(Gamage et al., 2020). However, these mutations and deletions have arisen multiple times during the pandemic, and are not always associated with rapid growth or VOCs. This suggests that there are epistatic effects between many of the mutations present in Alpha that together lead to its increased fitness.

While this constellation of mutations appears to have arisen in one evolutionary leap, two sequences have been identified in the COG-UK genomic surveillance dataset that contain some, but not all, of the Alpha-defining mutations, hence they may represent intermediate steps in the evolution of the Alpha lineage. These sequences could provide clues to the evolutionary processes underlying the evolution and emergence of VOCs and information on the timings of mutational events. The lack of more than two potential intermediate samples must be explained in the evolution of the Alpha variant. Due to the high level of SARS-CoV-2 genomic surveillance in the UK, it is unlikely that the Alpha variant was evolving in a conventional way (i.e. transmitting between individuals in the general UK population) without numerous intermediate genomes being sampled. Instead, it may have been evolving in a cryptic population before being detected in the general population in the UK, with the potential intermediates indicating early introductions from this population into the general population. I propose three possible alternatives for what this population may be: first, the Alpha variant may have evolved conventionally in a location with little or no viral genomic surveillance before being introduced into the UK in Kent; second, it may have evolved in a non-human animal population before a zoonotic event introduced it back into the human population in the UK; or finally, it may have evolved in a single or small number of chronically infected individuals, which were not sampled, before transmitting a single time into the general population.

The sudden appearance in late 2021 of the VOC Omicron has renewed interest in the processes underlying the emergence of variants exhibiting major leaps in

evolution, especially when combined with large fitness increases either through immunological escape or enhanced transmissibility. Whilst the Omicron variant is the most extreme example to have emerged to date, here I consider the origins of B.1.1.7/Alpha and the evidence for this lineage being the result of a similar process. In this chapter, I explore the processes on the branch leading up to the B.1.1.7 lineage, by using a Bayesian phylogenetic analysis to provide temporal and evolutionary rate estimates; as well as examining the two sequences which appear to be evolutionary intermediates. I then conduct coalescent and birth-death analyses to explore any differences in growth rates between Alpha and background lineages. Finally, in order to identify any common patterns among VOCs, I perform similar evolutionary rate analyses on all VOCs and variants of interest (VOIs), and compare their mutation profiles.

4.2 Methods

4.2.1 Genomic dataset

The COG-UK alignment and metadata of all SARS-CoV-2 genomes from 2021-04-21 was restricted to between 2020-08-01 and 2020-12-31 and surveillance (i.e. pillar 2) sequences from England. This dataset was then subsampled in a time-homogeneous way to generate approximately 1000 sequences per sequence set which comprised 50 sequences per week for non-B.1.1.7 sequences and 100 per week for B.1.1.7 sequences. The B.1.1.7 dataset was checked for molecular clock outliers (sequences that have disproportionately too much or too little divergence for its sample time, Hill and Baele, 2019) using TempEST, and one sequence was

identified and removed (England/CAMC-BBDA45/2020). The resulting dataset was 1100 background sequences and 976 B.1.1.7 sequences.

For mutation scanning, the sequences were aligned to the reference sequence Wuhan-Hu-1 using minimap2 (Li, 2018) and gofasta (https://github.com/cov-ert/gofasta). Variants were determined using type_variants.py (https://github.com/cov-ert/type_variants)

4.2.2 Evolutionary rate calculation

First, a maximum likelihood tree was generated using IQTree v2.1.2 (Minh et al., 2020) and an HKY substitution model (Hasegawa, Kishino, and Yano, 1985). This was used to generate the plots showing a linear regression of root-to-tip genetic distance against sampling date and provide estimates of the rate of evolution in background sequences and within the B.1.1.7 clade (slope of the linear regression).

To estimate the rate of evolution in the branch leading up to B.1.1.7, a local clock model in BEAST v1.10.4 (Suchard et al., 2018) was used. Briefly, a strict clock model was applied to each of the three groups so that their clock rates could be estimated independently, each with a Gamma prior. Preliminary analyses suggested that the within-B.1.1.7 evolutionary rate was similar to the background rate, and so the same clock rate model was placed on both the background and the within-B.1.1.7 clade. Nested taxon sets containing the one plausible intermediate sequence (CAMC-946506) were used to estimate the dates of the most recent common ancestor of B.1.1.7 and the intermediate sequence. A nonparametric coalescent Skygrid model (Gill et al., 2012) with 64 change-points spanning 15 months was placed on the background sequences including the intermediate sequence, and an exponential growth coalescent model was placed on B.1.1.7. For both sequence alignments an

HKY substitution model was used. Two chains with 100 million states were run, and following assessment via Tracer (Rambaut et al., 2018), 45 million states from each were removed as burn-in.

4.2.3 Growth rate calculations

To describe general patterns in the growth of B.1.1.7 compared to the background rate, as well as B.1.177 (N=1069) and a Welsh cluster containing N501Y (N=478), I ran a series of non-parametric Skygrid analyses (Gill et al., 2012). These were run independently in BEAST, each for two chains of 100 million states. For B.1.1.7, B.1.177 and the Welsh cluster, 77 change-points were used, spaced equidistantly between the most recent tip and 0.75 of a year before it (approximately 9 months). For the background dataset, 64 change-points were used with 1.25 of a year as the cutoff. All analyses assumed a strict molecular clock model and the HKY substitution model (Hasegawa, Kishino, and Yano, 1985).

To compare parametric growth models, a logistic, exponential and 3-epoch growth model were run only on the B.1.1.7 dataset. The 3-epoch model used fixed transition times and exponential growth rates for within each epoch. Each of them used an HKY substitution model and a strict clock model, and two independent chains of 100 million states were run. A marginal likelihood estimation (MLE) analysis, a commonly used form of Bayesian model selection integrated into the BEAST software package, was used to distinguish between these three models (Suchard, Weiss, and Sinsheimer, 2001).

To infer growth rates using the 3-epoch model, I combined the B.1.1.7 and background lineage datasets to simultaneously estimate growth rates during each of the epochs (N=2076). Transition times were fixed at the start and end of the

lockdown in England (2020-11-05 and 2020-12-02), and an earlier transition time was also placed on 2020-09-01 (the MRCA of B.1.1.7) at the start of the study period to focus on the growth after B.1.1.7 began to diversify. This analysis were run for two chains independently for 100 million states.

To estimate differences in the effective reproduction number (R_e) between B.1.1.7 and background lineages, a Bayesian birth-death skyline (Stadler et al., 2013) model was run independently on the B.1.1.7 and background datasets. An HKY substitution model was used along with a strict clock model, and a Gamma prior with α = 0.001 and β = 1000 was placed on the clock rate. A lognormal prior with mean 0.8 and standard deviation 0.5 was placed on R_e and a Beta prior with α = 2 and β = 1000 on the sampling proportion. R_e was parameterised into 4 epochs with transition times fixed at the start and end of the lockdown in England (2020-11-05 and 2020-12-02), and an earlier transition time placed at 2020-09-04. The sampling proportion was fixed to 0 before the first week containing a sample and then estimated for each week thereafter, resulting in 16 epochs for B.1.1.7 (from 2020-09-19) and 23 for the background dataset (from 2020-08-01). The becoming-uninfectious rate was assumed to be constant and fixed at 36.5, which is equivalent to a mean period of 10 days from infection to loss of infectiousness (through recovery, isolation or death). Analyses were started from initial trees estimated in IQTree v2.1.2 (Minh et al., 2020) and scaled to calendar time using TreeTime (Sagulenko, Puller, and Neher, 2018). For both datasets four chains of 100 million iterations were run independently, sampling states and trees every 10'000 iterations. Chains were combined after removing 10% of states as burn-in and convergence assessed using Tracer (Rambaut et al., 2018). Convergence was assessed using the R-package coda (Plummer et al., 2006) and 10% of states were removed to account for burn-in.

4.2.4 Sequencing proportion

Case data was collated from the UK government dashboard (https://coronavirus.data.gov.uk/). Cases and sequences with sample dates between 2020-04-24 and 2020-09-19 (the first day of the week of the most common recent ancestors of the stem of B.1.1.7 and the clade of B.1.1.7 respectively) were aggregated by week. Cases and sequences with the locations of “Kent” or “Medway”, a county surrounded by Kent, were included.

4.2.5 Rates of evolution in chronically-infected individuals

For each of the eight studies identified containing longitudinal samples of chronically-infected individuals, I counted the number of mutations present per individual in the paper. A mutation to the derived state and back to the reference allele each counted as a separate evolutionary event, relative to the first individual sample available, which was within the first week of infection start for all papers other than Karim et al. (2021) and Avanzato et al. (2020), which were day 12 and 49 respectively. For papers (e.g. Kemp et al., 2021) where proportions of variants were given, a 25% cutoff was used for presence/absence of a mutation. Ambiguous or missing data was treated as the reference allele.

The rate of mutation events was calculated by dividing the number of events by the number of days the individual was followed up divided by 7 to change the denominator to weeks. Note that for Karim et al. (2021), only non-synonymous substitutions were provided in the paper, meaning that the estimate of 2.04 events per week is an underestimate.

Start of infection was taken to be the start of symptoms or the date of the first positive PCR test depending on what was available for each study. End of infection was the date of the final negative PCR test in the study (Karim et al., 2021; Voloch et al., 2020; Williamson et al., 2021; Stanevich et al., 2021; Avanzato et al., 2020; Weigang et al., 2021), death (Choi et al., 2020) or when the individual was lost to follow up (Ramırez et al., 2021).

4.2.6 Other variant analyses

For performing the linear regression of root-to-tip genetic distance against sampling date to provide estimates of the rate of evolution in background sequences and within each VOC lineage, background datasets were obtained from the master COG-UK alignment. For Delta, Lambda and Mu, sequences sampled between 2021-01-01 and 2021-06-01 that were not any of the above variants were downsampled to 50 per week. For Gamma and Beta, the same background dataset as Alpha was used (see above). Then sequences from the correct time period for each variant were taken, and all were downsampled to 50 sequences per week, where possible. Finally, the sequences were run through metadata and sequence quality control. The final dataset sizes were 390 for the background set, 141 for Beta, 31 for Gamma, 373 for Delta, 246 for Lambda, and 208 for Mu.

For Omicron, B.1 sequences sampled between 2021-09-01 and 2022-01-01 and all Omicron sequences until the same cut-off were taken. These were then also downsampled to 50 per week, resulting in 900 background sequences and 390 Omicron sequences. Omicron sequences with any reference calls, any Delta mutations and more than three missing SNPs were excluded, as well as two molecular

clock outliers (see above) in the background dataset. The final dataset comprised 898 background and 328 Omicron sequences.

To undertake the mutation analysis for each variant, the first ten sequences for each variant other than Delta were taken from after the oldest reported sample in the original paper or report describing the variant. These were from 2020-09-20 for Alpha (Rambaut et al., 2020b), 2020-10-15 for Beta (Tegally et al., 2021), 2020-12-17 for Gamma (Faria et al., 2021) and 2021-11-15 for Omicron (Viana et al., 2022). For Delta, due to an unclear starting point and large amounts of missing data in the sequences, all Delta sequences in March 2021 (following the estimate of the start of Delta expansion in March in McCrone et al. (2021)) were extracted from the COG-UK alignment, and run through Scorpio (https://github.com/cov-lineages/scorpio), filtering to allow no reference alleles (from Wuhan-Hu-1) and a maximum of 2 missing alleles in the lineage-definition positions. This resulted in 98 sequences, ten of which were from India, which were taken as the representative group. For all variants, mutations which were common to all representative ten sequences were taken, and supplemented with any lineage-defining mutations which were missing (due to a small amount of missing data), based again on the original defining publication for each variant.

These were then compared to a representative background set. For each variant, this entailed ten sequences from the month of the first reported sample of the parent lineage: B.1.1 for Alpha and Omicron, B.1 for Beta and Omicron, and B.1.1.28 for Gamma.

4.3 Results

4.3.1 Characterising the ancestral branch of B.1.1.7

While the first sequence of B.1.1.7 was sampled on 20th September 2020 (GISAID Accession ID: EPI_ISL_601443), the lineage diverged from other concurrently circulating background lineages in the UK in late April 2020 (time of the most recent common ancestor (TMRCA) 2020-04-24, 95% HPD 2020-03-26 to 2020-05-24). However, it appears that rapid transmission of this lineage within the UK only began later in the year, with the TMRCA of the Alpha clade estimated on the 19th September 2020 (95% HPD 2020-09-18 to 2020-09-20, Fig. 4.1A).

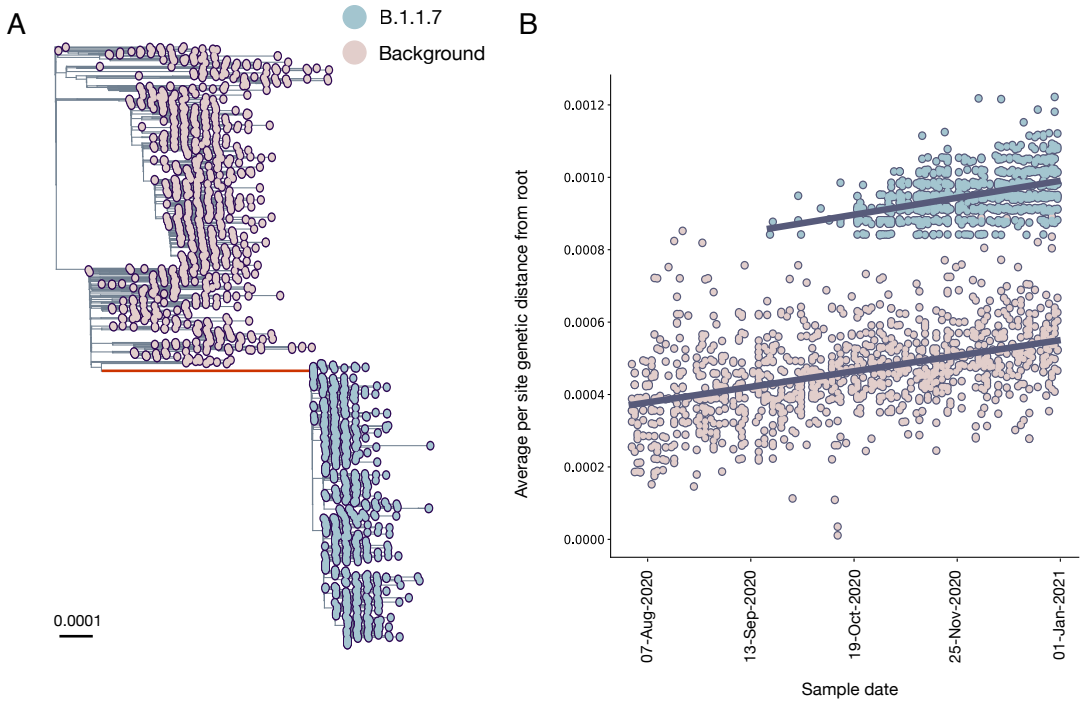

Fig. 4.1 Phylogenetic characteristics of B.1.1.7 lineage A) Maximum likelihood phylogeny showing the well-supported monophyletic clade that constitutes B.1.1.7. The ancestral branch with the higher rate of evolution is highlighted, and branch lengths represent substitutions/site B) Regression of root-to-tip genetic distances against sampling dates, for sequences belonging to lineage B.1.1.7 (blue) and those in its immediate outgroup in the global phylogenetic tree (pink). The regression lines are fitted to the two datasets independently. The regression gradient is an estimate of the rate of sequence evolution. These rates are 4.6x10-4 and 4.3x10-4 nucleotide changes/site/year for the B.1.1.7 and outgroup data sets, respectively

The ancestral branch leading to the B.1.1.7 lineage is exceptionally long for SARS-CoV-2, both in terms of time (mean=147 days, 95% HPD: 112 days to 173 days) and substitutions: there are 23 nucleotide changes with the majority amino acid altering (14 non-synonymous mutations and three deletions). We found that the evolutionary rate of the ancestral branch was an average of 2.3 times higher than the background rate (95% HPD = 1.4 to 3.6). There is however little evidence for an increased rate of evolution within the B.1.1.7 clade: a regression of root-to-tip of genetic distances against genome sampling date (Fig. 4.1B) shows that the rates

within the B.1.1.7 clade and the background sequences are very similar (4.6x10-4 and 4.3x10-4 respectively).

Two sequences lie along the branch leading to the B.1.1.7 clade, and contain some, but not all, of the Alpha-defining mutations. The earlier of the two (COG-UK identifier CAMC-946506, gisaid ID EPI_ISL_556680) was sampled on 15th July 2020, and the more recent genome (MILK-B154B6, gisaid ID EPI_ISL_2735517) was sampled on 23rd October 2020. If these two sequences are truly intermediate - i.e. they represent midpoints in the accrual of the 23 lineage-defining mutations for Alpha - then they may provide insight into the order of mutational accumulation during VOC evolution.

The more recent sequence, MILK-B154B6, is ambiguous at several sites, including at position 501 in Spike, with 80% of reads encoding N (asparagine, found in background lineages) and 20% Y (tyrosine, found in Alpha). These ambiguous sites imply either a coinfection of two different virus populations or laboratory contamination. If a coinfection, the sample could have been an individual who was infected by an early Alpha sequence and a background lineage (possible in late October 2020 in the South East of England, as both were circulating there at the time). MILK-B154B6 contains the synonymous mutation C5986T, which is also found in 971 of the 976 of the early Alpha sequences, but in none of the 1100 background sequences used in this study. It also has two further mutations (C15279T and C913T) that are respectively found in 974 and 970 sequences in the Alpha dataset (out of 976 sequences), but only once in the background dataset. As this sequence contains mutations that are shared by most Alpha sequences and not found frequently in earlier clades, it suggests that its intermediate status is due to a coinfection of an Alpha sequence (which contains these mutations), and a background clade that was co-circulating, and so the consensus sequence contains only some of the Alpha-defining muta-

tions; or cross-contamination in a laboratory handling samples from both Alpha and background lineages. MILK-B154B6 can therefore not be treated as an evolutionary intermediate.

The older sequence, CAMC-946506, contains no ambiguous sites and four of the Alpha-defining mutations: N501Y in Spike, the 9 base pair deletion in NSP6, as well as R52I and Q27 to stop (Q27*) in ORF8. In the UK, prior to 1st September 2020, 37 sequences were sampled with Q27*, 5 with R52I (all in July and August, one of which also had the Q27*), and none with the NSP6 deletion or N501Y (n=34,291). This makes it unlikely that a virus containing all of these mutations either existed in early 2020, prior to the start of the long branch, or that this sequence has convergently acquired these mutations and has been placed erroneously in the tree. Instead, the evidence suggests that CAMC-94506 could be a true intermediate sequence, however it must be noted that it may also simply contain mutations shared by the common ancestor of the hypothesised cryptic population and CAMC-94506 (Fig. 4.2A).

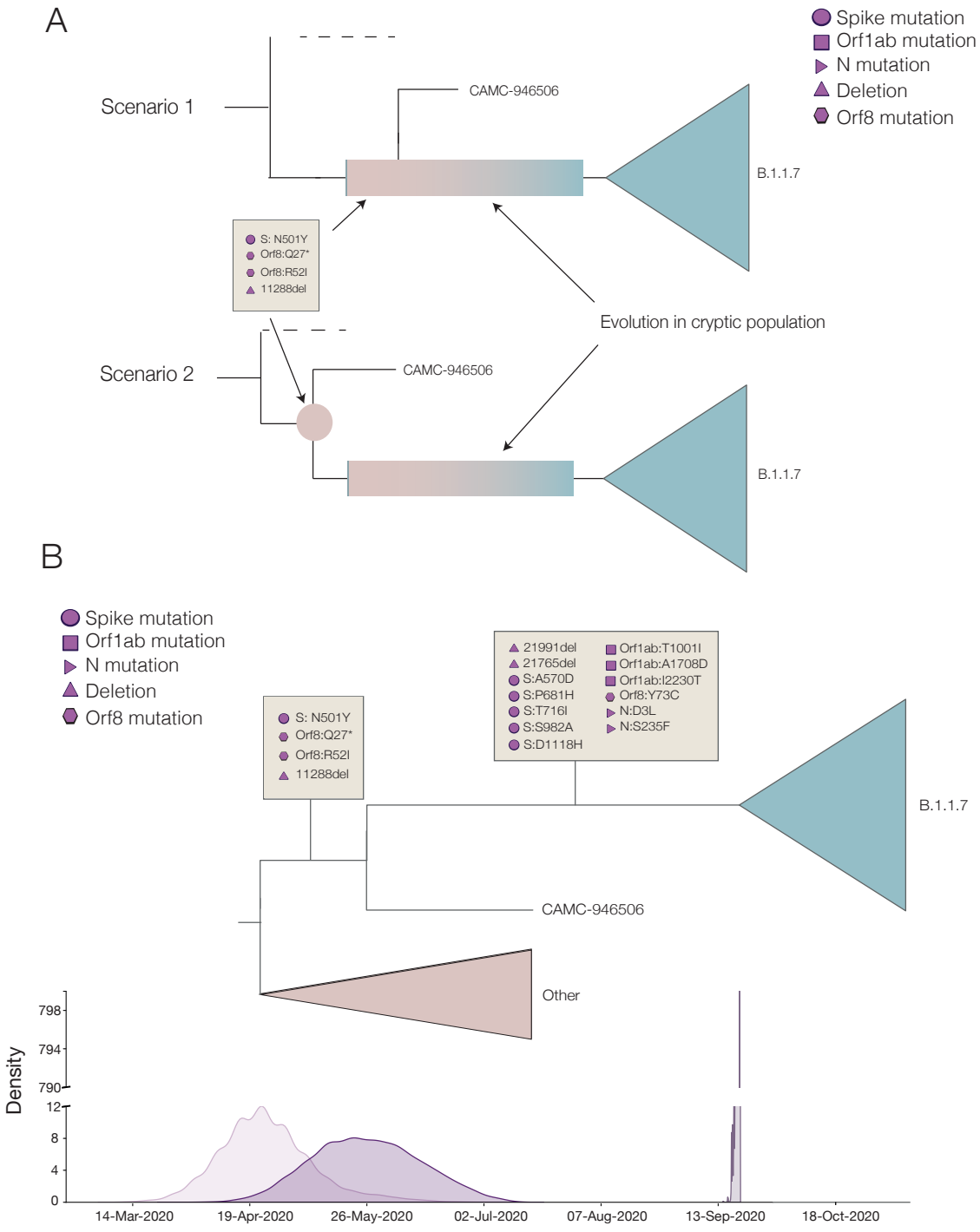

Fig. 4.2 Scenarios and timings for the intermediate sequence A) Two different scenarios of how the shared mutations between CAMC-946506 and the B.1.1.7 clade could have arisen. Scenario 1 shows CAMC-946506 as resulting from a transmission chain spilling over from a chronically infected individual and the mutations arising early in the infection. Scenario 2 shows the mutations as being shared by the common ancestor of CAMC-946506 and a cryptic population. B) Schematic of the time tree showing possible timings for B.1.1.7 lineage-defining mutations. Densities of the most recent common ancestors for (respectively) the background lineages and all B.1.1.7, the intermediate sequence and B.1.1.7, and all B.1.1.7 are shown along the bottom.

This intermediate sequence provides evidence that the spike mutation N501Y, the two ORF8 mutations Q27* and R52I, and the 9 base pair deletion in NSP6 all evolved early in the evolutionary history of Alpha (Fig. 4.2B), between 24th April 2020 (95% HPD 2020-03-26 to 2020-05-24) and 26th May 2020 (95% HPD 2020-04-27 to 2020-06-30), i.e. between the TMRCA of B.1.1.7 and all background sequences, and the TMRCA of CAMC-934506 and B.1.1.7.

4.3.2 Early growth rate of B.1.1.7 in the UK and interaction with November lockdown in England

Using a non-parametric coalescent growth model, I found that the growth of B.1.1.7 in the second half of 2020 in England was extremely rapid compared to the background lineages present at the time (Fig. 4.3A). At the end of the time period, it appears that B.1.1.7 continues to grow while the other lineages are beginning to decrease. These trends are broadly similar when comparing B.1.1.7 to just the B.1.177 lineage (Fig. 4.3C), which spread rapidly across the UK and became the dominant lineage over the summer of 2020 (Hodcroft et al., 2021).

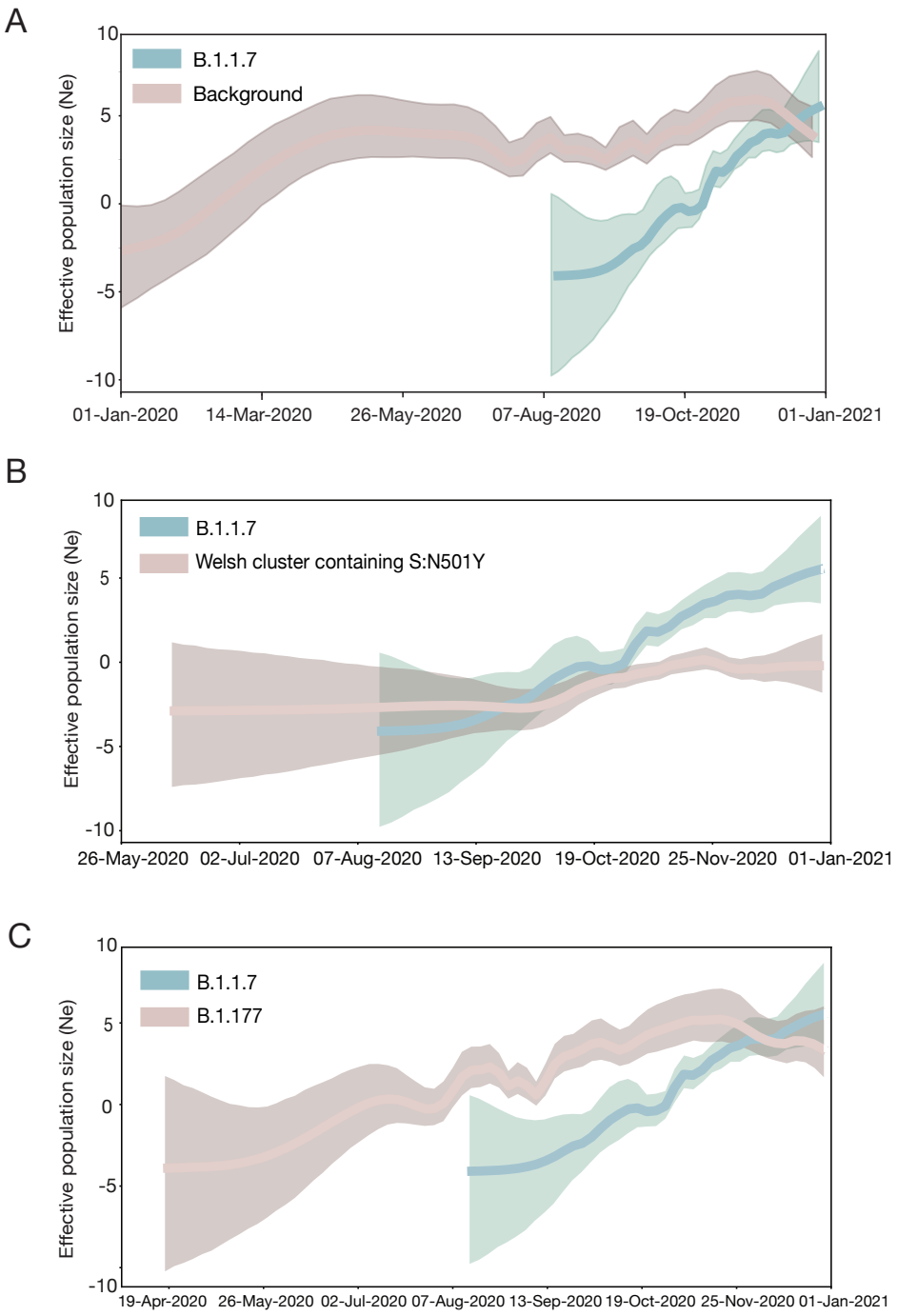

Fig. 4.3 Effective population sizes of B.1.1.7 (blue) against: A) background lineages B) a Welsh cluster defined by N501Y C) B.1.177. All are generated from independent BEAST analyses. 95% HPDs shown as shaded areas.

To further investigate the growth of B.1.1.7, I tested the difference between three standard population growth models: logistic, exponential, and epoch-based. For the last, wherein different epochs are permitted to have different growth rates, I estimated growth rates in the pre-lockdown period (2020-09-01 to 2020-11-04), during lockdown (2020-11-05 to 2020-12-04) and post-lockdown (2020-12-05 to 2020-12-31). Using a marginal likelihood estimation (MLE) approach, I found that the three-epoch model provided the best fit to the genomic data (Table 4.2). For the second time period, B.1.1.7 has a positive growth rate, and the post-lockdown period estimation includes zero; whereas the background lineages have a very strong negative growth rate in the most recent time period (Fig. 4.4A). This suggests that while the national lockdown in England in November was associated with negative growth in background lineages, and caused the B.1.1.7 growth rate to drop significantly, it was not sufficiently strict to push the growth rate of B.1.1.7 below zero. This reduced, but non-negative growth rate for B.1.1.7 during the November lockdown has also been shown on a spatial level (Kraemer et al., 2021).

In order to explore this further, we also estimated R_e, using a birth-death approach, which allowed the sampling proportion to vary to account for changes in genomic surveillance intensity across time (Fig. 4.4B and Fig. 4.4C). This showed that while both the background lineages and B.1.1.7 had an R_e above 1 (i.e. the epidemic was growing) in September and October, the English national lockdown in November was sufficient to push the R_e of the background lineages to around 1 (i.e. the epidemic was stable). However, the R_e of B.1.1.7 remained above 1, matching epidemiological information which showed growth of S-gene target failure positive cases despite the November lockdown (Kraemer et al., 2021).

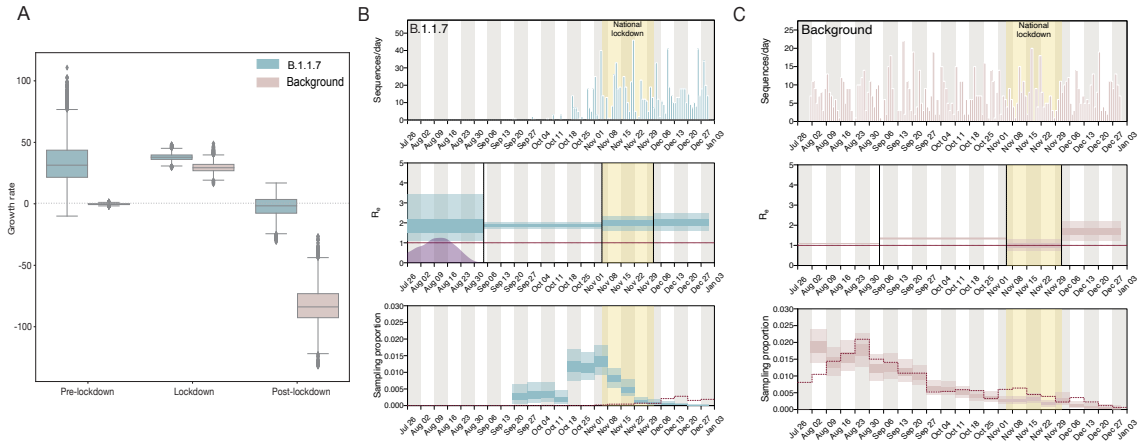

Fig. 4.4 Coalescent and birth-death growth estimates of B.1.1.7 and background lineages. A) Growth rate estimates with fixed transition times at pre-lockdown, lockdown, and post-lockdown, split by background lineage and B.1.1.7. B) Independent birth-death skyline analyses showing the number of sequences per day, the effective reproduction number (R_e) and sampling proportion (which is allowed to vary on a weekly basis), for B.1.1.7 and C) the background. The English national lockdown in November is highlighted in all plots.

4.3.3 Other variants of concern

Under current WHO designations, there are four VOCs other than Alpha: Beta, discovered in South Africa at the end of 2020; Gamma, discovered in Brazil at the start of 2021; Delta, discovered in India at the start of 2021; and Omicron, discovered in South Africa and Botswana at the end of 2021. There are also two VUIs: Lambda, discovered in Peru in mid-2021 and Mu, discovered in Colombia at the start of 2021. Each of these variants has had differing impacts across different regions, but the current Omicron wave has displaced almost all other lineages (outbreak.info).

Similarly to B.1.1.7, Omicron has a long ancestral branch (Fig. 4.5A), and the root-to-tip plot shows that Omicron sequences are distinct from the background diversity (Fig. 4.5B). However, as Omicron did not evolve out of the dominant circulating

variant (i.e. Delta, B.1.617.2 and its descendants), it is more difficult to identify the clear pattern that can be observed in B.1.1.7, which evolved out of the dominant lineage at the time (i.e. B.1.1). Further, Omicron contains two distinct sibling clades, BA.1 and BA.2 which may represent two independent introductions into the general population. There is also a third lineage, which appears to be a recombinant of ancestral BA.1 and BA.2 sequences due to its mixture of some mutations found in both of them (Viana et al., 2022). The circumstances under which Omicron arose are clearly more complex than those that led to the evolution of Alpha. This could indicate a chronically infected individual or individuals with more contact with the general population (Maponga et al., 2022), or perhaps a non-human animal population.

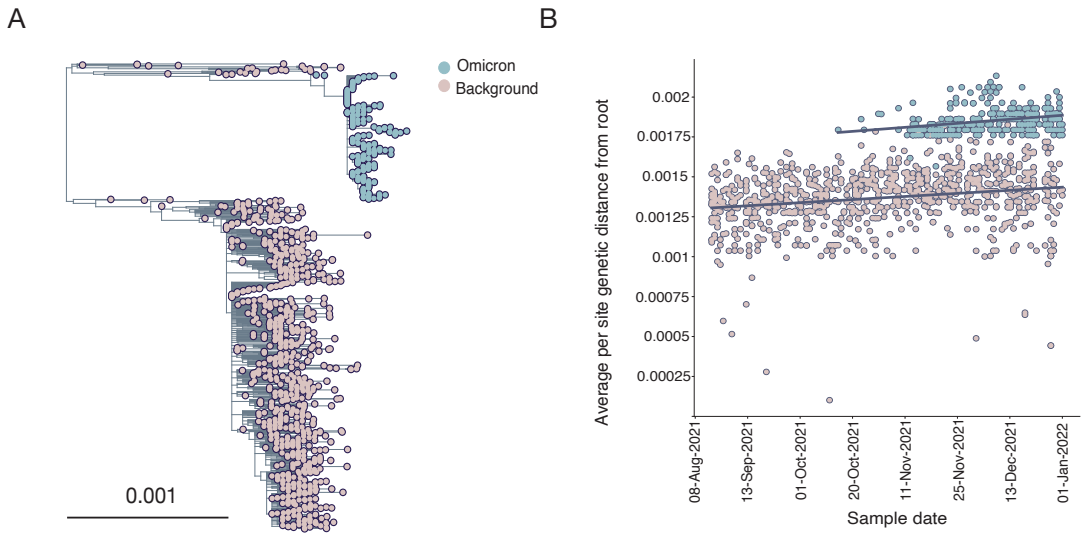

Fig. 4.5 A) Phylogeny showing Omicron in blue and background sequences in pink. The large background group is the Delta variant, the dominant variant globally in the second half of 2021. B) Separate regressions of distance from the root against sample time for background sequences and Omicron sequences. Note that the parallel lines indicate similar rates of evolution within each clade.

There is no clear increase in speed of evolutionary rate on the ancestral branches leading to Delta, Lambda or Mu (Fig. 4.6), suggesting that these may have arisen

under more traditional evolutionary processes involving intense between-host transmission. Beta and Gamma show *some* evidence of an increased evolutionary rate on the ancestral branch (although this stronger in Gamma and Beta remains somewhat inconclusive) and so may have a similar process of emergence involving a potential chronic infection.

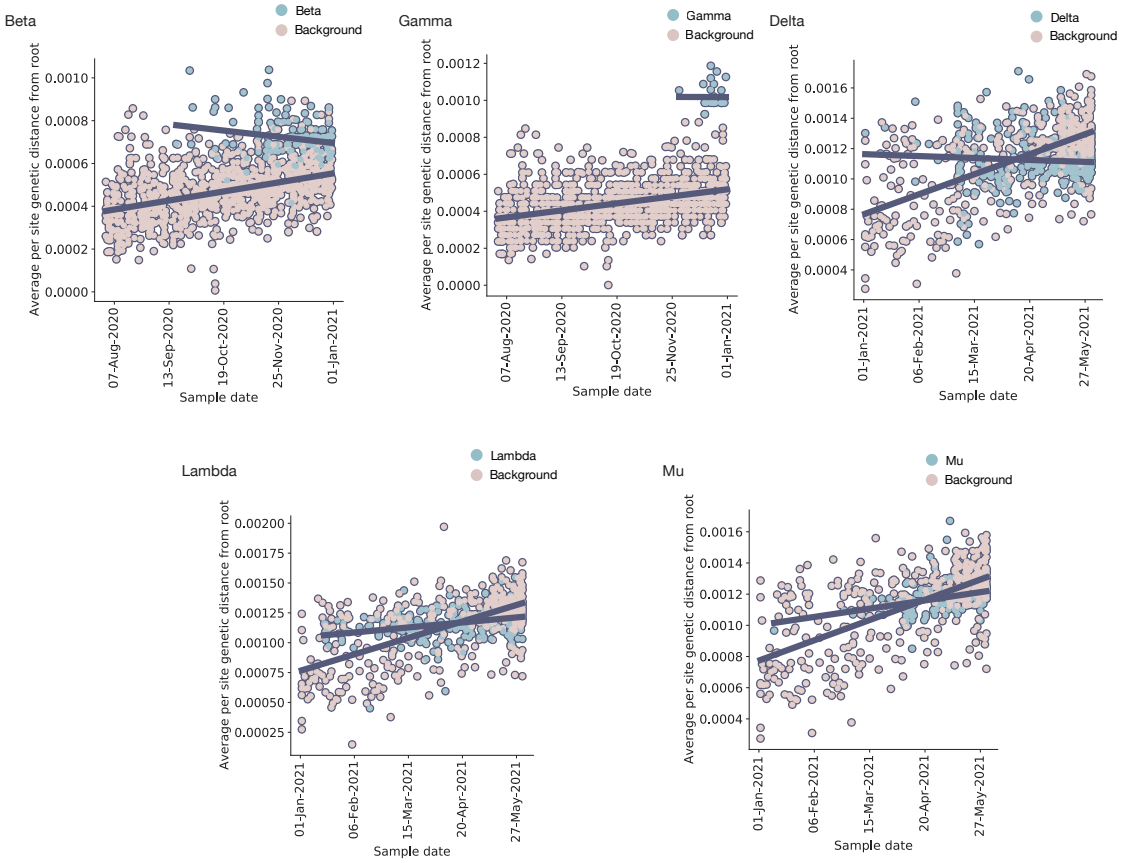

Fig. 4.6 Regressions of root-to-tip genetic distance against time for other variants of concern. A) Root to tip divergence plots for the six other variants of concern (Beta, Gamma and Delta) or variants of interest (Lambda and Mu) as designated by the WHO. None show any noticeable difference in evolutionary rate (shown by the lines) between the background sequences and the relevant variant.

Looking at mutations or patterns shared by variants of concern could provide evidence of common emergence routes or evolutionary pressures. I therefore collated mutations for each VOC (Alpha, Beta, Gamma, Delta and Omicron) compared to the background that they emerged from (see Methods) to identify any similarities. In general, there were very few shared mutations, and in particular, none shared by all variants (Fig. 4.7A). No synonymous mutations were shared between any variants.

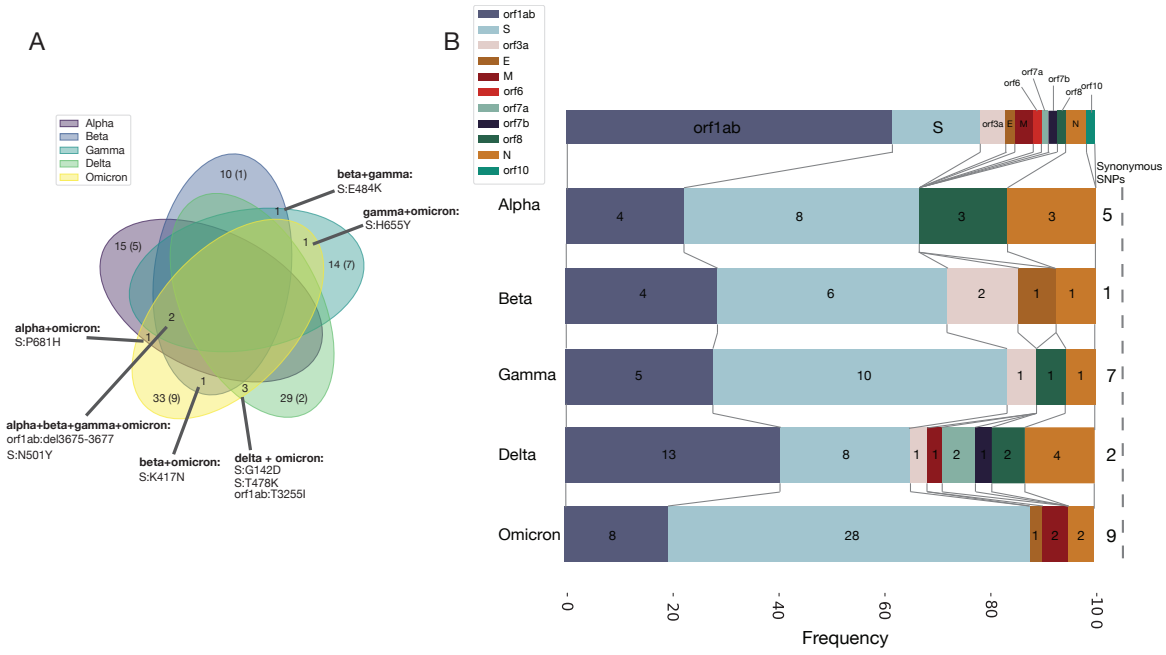

Fig. 4.7 Comparing mutation profiles between variants of concern A)Venn diagram showing numbers of mutations shared between different variants of concern, with synonymous mutations in brackets. Zeroes, denoting no shared mutations acquired on the ancestral branch, are omitted. B) Frequency of non-synonymous mutations acquired on the ancestral branch in different parts of the genome between variants of concern. A schematic of the genome is shown along the top, numbers on each slice represent the absolute numbers of non-synonymous mutations or deletions in that gene, and numbers of synonymous mutations are shown along the right hand side.

The most shared mutations were N501Y in spike, and the nine base pair deletion in NSP6 (Fig. 4.7A), which were both found in Alpha, Beta, Gamma and Omicron: all four of which have evidence of an increased evolutionary rate prior to their emergence.

Notably the intermediate genome for B.1.1.7 also exhibits both N501Y and the NSP6 deletion. N501Y in particular has been monitored throughout the pandemic, due to its ability to increase binding to the ACE2 of human and murine cells. However, it appears that by itself, it is not necessarily enough to create a VOC, as there was a cluster in Wales in late 2020 defined by N501Y but without the NSP6 deletion, which was rapidly outcompeted by Alpha (Fig. 4.3B).

Between Alpha and Omicron, which are the two variants with the most arguments for arising from chronic infections, there is only one unique shared mutation that was acquired during the evolution of the variant: P681H in the furin cleavage site of the Spike protein (Fig. 4.7A), which enhances Spike cleavage (Peacock et al., 2022). It must be noted also that Delta carries P681R, but shares no other mutations compared to the background lineages with Alpha and Omicron. The polybasic furin cleavage site is found in other coronaviruses, although not in any other Sarbecovirus, and it is required for SARS-CoV-2 virus entry into human lung cells (Hoffmann, Kleine-Weber, and Pöhlmann, 2020). Mutations in this area may be an adaptation to the human host, providing evidence for evolution in a human cryptic population. Of note, Peacock et al. (2021) found that mutants with deletions in the furin cleavage site were rare; I speculate that this could be part of the fitness valley that Alpha (and Delta and Omicron) had to cross, as the furin cleavage site appears to be relatively conserved and so possibly many mutations are deleterious. It is also worth noting that while it is not a defining mutation of all sub-lineages of Omicron, the 69-70 deletion in the Spike protein is present in BA.1, which at the time of writing dominates most Omicron epidemics.

Finally, between Beta and Omicron, the variants with the most evidence for immune evasion, the single common mutation is K417N in the Spike protein (with a similar mutation, K417T found in Gamma, Fig. 4.7A). This mutation has been found

to confer reduced susceptibility to neutralisation by specific monoclonal antibody therapies (Starr et al., 2021). This mutation also arose in AY.1, the so-called “Delta plus” variant descended from B.1.617.2 (Kannan et al., 2021), but this variant did not appear to acquire any noticeable advantage compared to the background Delta wave (outbreak.info).

In terms of the frequency of regions of mutations (Fig. 4.7B), all bar Delta have the highest frequency of non-synonymous mutations in the spike protein, and Delta has the highest frequency in orf1ab (approximately 38% of its mutations). Omicron has the highest frequency of spike mutations (56%), Gamma the highest frequency of synonymous mutations (28%) and Alpha and Beta have the highest frequency of deletions (approximately 13%). Overall there doesn’t appear to be a discernable pattern in the types or locations of deletions.

In comparing all of the VOCs, Alpha, Gamma and Omicron share the most evidence for a faster rate of evolution along their ancestral branch and Beta remains somewhat inconclusive, but with some evidence of the lineage having distinctly more root-to-tip divergence than expected. The VOIs and Delta show no evidence of a faster rate of evolution along their ancestral branches. Further, Delta carries mutations spread more evenly across its genome than the other VOCs, and consequently has fewer Spike mutations. This provides evidence for an alternate route of emergence for the Delta variant as compared to the other VOCs.

4.4 Discussion

B.1.1.7/Alpha was first sampled in Kent, in South East England on 20th September 2020, and spread quickly across the rest of the UK (Kraemer et al., 2021; Volz et al., 2021a). It was able to grow rapidly, in part because the English national lockdown in

November was far less strict than previous or later lockdowns (i.e. in March 2020 or January 2021) with fewer restrictions on mixing and schools remaining open; however this weaker lockdown was sufficient to reduce the background lineage growth rate significantly. While these trends have been described previously (Kraemer et al., 2021; Volz et al., 2021a), I have here reproduced them using only phylodynamic techniques and a small but representative genomic dataset. This finding will be useful for investigating VOCs which arise in areas with less genomic sequencing, or tracking those without e.g. SGTF drop-out.

Any proposed origin of B.1.1.7 must explain three observations: first, the long branch leading to the B.1.1.7 clade with at most one intermediate sequence, despite high genomic surveillance; second, an increased evolutionary rate along this branch; and third, a single geographical and evolutionary origin of B.1.1.7 (Kraemer et al., 2021).

In a country with an extensive virus genomic surveillance programme like the UK, which includes random and relatively dense sampling (an average of 7.9% of weekly reported cases in Kent and Medway between 24th April 2020 and 19th September 2020 were sequenced), it is unlikely that a precursor lineage was circulating in Kent over the summer of 2020 and was not detected. It is worth noting also that B.1.1.7 was captured by this surveillance programme within at most two days of its origin - the TMRCA of the clade is the 19th September 2020 (see above), and the first sample was taken on 20th of the same month.

One possible explanation for the lack of detection of the precursor lineage to B.1.1.7 in the UK surveillance data is simply that it was not in the UK prior to its expansion in South East England, but in a region of the world with little or no genomic surveillance. However, this hypothesis requires that the lineage was introduced twice into the UK (firstly for the intermediate sequence and secondly for the B.1.1.7

lineage) without being exported and establishing transmission anywhere else. Genomic surveillance has since been scaled up in many regions, and the fact that no descendants of such a cryptic population have been sampled to date indicates either that this population went extinct (unlikely given the fitness advantages conferred by the lineage-defining mutations) or that no such population existed. Furthermore, transmission between humans, even if rapid, would not explain the higher rate of evolution observed along the branch. These explanations would also apply to a population in the UK which is disproportionately under-sampled, for example vulnerable communities, such as individuals experiencing homelessness or persons who inject drugs, who are unlikely or unable to seek healthcare.

An alternative explanation is a zoonotic event, as SARS-CoV-2 has been shown to spread in non-human animals, for example in mink (Oude Munnink et al., 2021; Oreshkova et al., 2020), white-tailed deer (Chandler et al., 2021), and Syrian hamsters (Yen et al., 2022). In this hypothesis, there would have been a reverse zoonosis from humans, an increased rate of molecular evolution among animals, perhaps due to natural selection for the new host species, followed by at most two zoonotic events in the course of several months (the intermediate and the final clade). For the former, in an animal population that had sufficient contact with humans for a reverse zoonotic event and then two later zoonoses, it is unlikely that there would be only two spillovers in five months: in mink farms in the Netherlands, it was estimated that there were 43 spillovers between April and November 2020 (Lu et al., 2021); and in a pet shop in Hong Kong, there were at least two spillovers in the space of a few weeks (Yen et al., 2022). Further, transmission between animals has not been observed to lead to a higher rate of evolution: even in a large outbreak among densely farmed mink, the rate of evolution was estimated to be similar to what is expected between humans, at approximately 7.9x10-4 nucleotide changes/site/year (Lu et al., 2021).

Further, it is unlikely that a variant so effective at spreading in the human population would have evolved in a non-human population; as the mutations required to be successful in a human population may well be different due to differences in cell receptors, as well as behaviour. Common mutations appear in animal populations, such as N501T and Y453F in mink across different continents (Eckstrand et al., 2021; Lu et al., 2021), but are rarely found in human infections, and not in any of the VOCs. While it would be possible for a two-step evolutionary process wherein there is first some adaptation in an animal population, allowing the crossing of a fitness valley, followed by spillover and human adaptation through conventional transmission, it would once again be difficult to explain the intermediate sample we observe through this transmission process.

I propose that the most likely explanation is that an individual was chronically infected with SARS-CoV-2 over the course of months providing an evolutionary environment conducive to the virus making adaptive jumps. The evolutionary environment within a single host is different to that at the between-host scale, with a large effective population size and the opportunity for recombination (Jackson et al., 2021). This large effective population size can be established and maintained partially due to the different compartments that a respiratory virus can establish infection in, for example, upper and lower respiratory tract, as well as deeper into the lung (e.g. Lakdawala et al. (2015) and Richard et al. (2020). Conversely, the effective population size at the between-host level is small due to tight bottlenecks occurring on transmission (Lythgoe et al., 2021). Furthermore, although a persistent infection will provide the time and selective environment for a period of adaptive evolution, the exact cause of the persistence may affect the traits of the virus that are selected for.

There are a number of studies on chronically or persistently infected individuals which contain longitudinal sampling of the viruses present (Kemp et al., 2021; Karim

et al., 2021; Ramırez et al., 2021; Choi et al., 2020; Williamson et al., 2021; Stanevich et al., 2021; Voloch et al., 2020; Avanzato et al., 2020; Clark et al., 2021; Weigang et al., 2021). Across these studies, there were an average of 4.0 (95% confidence interval: 0.63 to 9.76) evolutionary events (i.e. gaining or losing a mutation compared to the individual's first sample) per week (Table 4.3), compared to the approximately 0.5 mutations expected in between-host transmission. Further, in these studies, deletions (notably, the 69/70 deletion found in Alpha and the BA.1 sub-lineage of Omicron) have been found to both increase and decrease in frequency along the time period of infection (Kemp et al., 2021; Stanevich et al., 2021). As it is unlikely that a deletion could be reverted, this is evidence for multiple coexisting viral populations (Lythgoe et al., 2021).

Much of the discussion of variant emergence from within-host evolution has focused on the idea of individuals with compromised immune systems, either pathologically (e.g., a lymphoma) or medically (e.g., post-transplant suppression or chemotherapy) induced (Karim et al. 2021; Maponga et al. 2022). Persistent infections have also been recorded in apparently immunocompetent individuals (Ramırez et al., 2021; Voloch et al., 2020) and although the average length of these infections is shorter than in immunocompromised cases – from the studies above, an average of 32 days and 174 days respectively (Fig. 4.8B). In one case an immunocompetent individual was infected for approximately 64 days (Voloch et al., 2020). Furthermore, there may be an observation bias towards data from immunocompromised hosts and long term infections in immunocompetent hosts are less likely to be recorded.

The degree of immunocompromisation is variable and will depend on whether antibody or cellular immunity (or both) are affected by the condition of the individual. Grenfell et al. (2004) used a simple population genetic model to posit that the rate of viral adaptation is a non-linear function of immune (selection) pressure, because

of the opposing effects of raised immune pressure on virus population size and average selection coefficients. Hence the highest viral adaptation rate is predicted to arise from intermediate immune pressures (Grenfell et al., 2004). Although there is currently no evidence of a difference in numbers of evolutionary events between immunocompromised and immunocompetent hosts (Fig. 4.8A)), I propose that this may be because these cases lie on either side of the peak of viral adaptation rate in the Grenfell et al. (2004) model. Thus within-host virus evolution in both healthy and immunocompromised hosts could lead to an increased evolutionary rate; and we have not yet observed an individual at the part of the immune spectrum which would lead to the explosive adaptation seen in Alpha and Omicron.

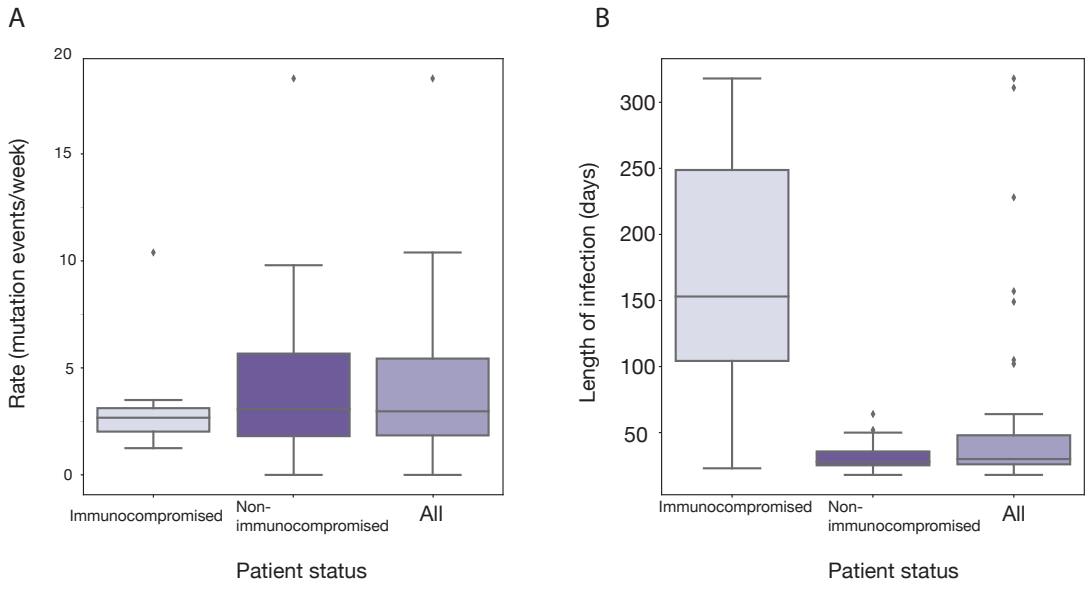

Fig. 4.8 Results from longitudinally-sampled patients from across seven papers. A) Estimated mutation rate in terms of number of mutation events per week. B) Length of infection in days.

The absence of more than one sampled transmission chain arising from intermediate combinations of the Alpha mutations, even on the background of mild NPIs in the summer of 2020 in the UK; as well as this constellation not evolving elsewhere across the phylogeny, suggests that the fitness peak is difficult to reach, despite its large selective advantage compared to background lineages. This implies a complex fitness landscape, and a large fitness valley prior to the peak that Alpha represents for between-host fitness. The different selective pressures inside a host could enable the crossing of such a valley due to a relaxation of selection on transmission-based adaptations in favour of factors such as evading neutralising antibodies, as found in longitudinal samples in Weigang et al. (2021), or focusing on cell entry. The intermediate sequence is likely part of a transmission chain that was ultimately very short, as the virus was still not particularly well adapted to transmitting between hosts at this time point.

However, Alpha and other VOCs clearly became well adapted to spreading between hosts as well as within a host. While a transmission advantage would not be explicitly selected for within hosts, traits which are useful within a host could also lead to better transmission. For example, increased ACE2 binding which increases the efficiency of cell entry (Ozono et al., 2021) would allow a virion to enter and use a host cell faster than its competitor within a host, leading to faster growth; and would also make it easier for a virus population to establish an infection in a new host. The evasion of a host immune response is another clear advantage within a host, allowing fewer virions to be destroyed by the immune system, and on a population level, especially in the immune context of widespread previous infection and vaccination. This is seen most clearly in Omicron, and provides a possible explanation as to why there are three sub-lineages of Omicron which all arose at once: the sub-lineages all acquired immune evasion properties within a host, and then

spread from diverse viral populations within that host. Finally, and non-exclusively, an efficient and host-adapted infection could lead to a large amount of viral shedding, increasing transmissibility.

From the mutational profiles and the evolutionary rates of the VOCs, it appears that Delta is an outlier: it does not contain N501Y or the deletion in NSP6, nor does it appear to have a higher rate of evolution leading up to its emergence, and it has fewer non-synonymous mutations in its Spike protein as well as more mutations in other parts of the genome compared to the other VOCs. I therefore hypothesise that it may have followed an alternative route of emergence, perhaps simply intense between-host transmission in an undersampled location. Given that the estimate of the TMRCA is several months prior to the first sample (McCrone et al., 2021), it is plausible that the lineage-defining mutations could have been acquired sequentially, prior to a larger explosion of cases once the full constellation was present. The common patterns found in the emergence of Alpha, Beta, Gamma and Omicron provide some evidence that Beta, Gamma and Omicron may also have arisen through chronically infected individuals. It is notable that Southern Africa, where Beta and Omicron were first sampled, has a high prevalence of people living with HIV. An individual with a poorly-controlled HIV infection would provide another avenue for a large viral population to be maintained in an individual over a long period of time and therefore new between-host fitness peaks to be explored. Indeed, an individual with an uncontrolled HIV infection accumulated more than 20 mutations in the course of 9 months (Maponga et al., 2022). In this case the HIV virus was controlled and the SARS-CoV-2 cleared through antiretroviral treatment. However, in other circumstances progression of an HIV infection could then allow the partially controlled SARS-CoV-2 infection to proliferate significantly allowing for shedding and therefore transmission. Accessible antiretroviral therapy is therefore a key element

of mitigating the risk of further SARS-CoV-2 variant emergence in countries with significant numbers of people living with HIV, as called for by Maponga et al. (2022). More generally, equitable and universal access to SARS-CoV-2 vaccination and antiviral drugs will be a critical strategy given the apparent diversity of circumstances by which VOCs have emerged thus far.

Chronic COVID-19 cases are relatively rare, but as another wave of transmission sweeps across the world, there will be many more as has been seen recently with the Omicron variant (Viana et al., 2022). If all persistent infections present a risk of a new, highly transmissible or immune evasive variant, then simply shielding the vulnerable and selective vaccination will not be sufficient to prevent the emergence of another wave of morbidity and mortality. Without urgent and truly widespread vaccination efforts, we expect to see the delayed impacts of uncontrolled transmission resulting from vaccine and antiviral inequity into the future.

4.5 Supplementary tables

Gene	Nucleotide	Amino acid	CAMC-946506?	MILK-154B6?	Amplicon
Orf1ab	C3267T	T1001I	No	No	11
	C5388A	A1708D	No	No	18
	T6954C	I2230T	No	Yes	23
	11288-11296 deletion	SGF3675-3677 deletion	Yes	No	37
Spike	21765-21770 deletion	HV 69-70 deletion	No	Yes	72
	21991-21993 deletion	Y144 deletion	No	Yes	72/73
	A23063T	N501Y	Yes	Mixture	76
	C23271A	A570D	No	No	77
	C23604A	P681H	No	Yes	78
	T24506G	S982A	No	Mixture	81
	G24914C	D1118H	No	No	82
Orf8	C27972T	Q27stop	Yes	No	92
	G28048T	R52I	Yes	No	92
	A28111G	Y73C	No	No	92/93
N	28280 GAT -> CAT	D3L	No	No	93
	C28977T	S235F	No	Yes	95

Table 4.1 Non-synonymous mutations and deletions inferred to occur on the branch leading to lineage B.1.1.7.

Model	Path sampling	Stepping stone sampling	Order
Three epoch	-55785.30	-55802.68	Best
Logistic	-55842.26	-55858.27	Second
Exponential	-55844.18	-55861.28	Third

Table 4.2 Marginal Likelihood Estimation of different growth rate models

Paper	Individual number	Length of infection (days)	Days between first and last sequence	Type of immunocompromisation	Number of mutation events
Karim *et al* 2021	1	228	190	HIV positive	57
Weigang *et al* 2021	1	149	140	Organ transplant recipient	59
Ramirez *et al* 2021	1	48	45	-	4
	2	36	30	-	1
Choi *et al* 2020 and Clarke *et al* 2021	1	157	152	B cell depletion	52
Kemp *et al* 2021	1	102	101	Combined immunodeficiency	150
Williamson *et al* 2021	1	311	290	B cell depletion	124
Stanevich *et al* 2021	1	318	308	B cell depletion	79
Avanzato *et al* 2020	1	105	56	B cell depletion	10
Voloch *et al* 2021	5	28	26	-	9
	6	40	28	-	8
	8	42	11	-	14
	9	41	27	-	15
	11	32	23	-	16
	12	45	39	-	14
	13	29	20	-	4
	14	52	10	-	9
	15	35	20	-	15
	19	30	12	-	3
	21	27	21	-	11
	22	24	21	-	7
	23	64	22	-	0
	25	30	21	-	12
	26	23	20	HIV positive	10
	27	23	20	-	13
	29	22	21	-	6
	31	12	17	-	45
	32	21	19	-	3
	33	35	14	-	2
	34	21	19	-	2
	35	31	11	-	9
	36	28	11	-	12
	37	28	11	-	12
	38	27	14	-	11
	39	28	17	-	2
	40	48	20	-	7
	41	20	12	-	3
	43	26	5	-	7
	44	26	6	-	5
	45	25	14	-	6
	46	18	16	-	16

Table 4.3 Information about individuals used in chronic infection analysis

5

The spatial dynamics of SARS-CoV-2 in the UK

Like, mutations and stuff

Dr JT McCrone

Personal communication

2021-01-28

5.1 Introduction

The first case of COVID-19, the disease caused by SARS-CoV-2, in the UK was reported on the 29th January 2020 in York. Further travel-related cases were detected, seeding small clusters in London and Brighton (via France), before spreading into Northern Ireland on the 27th February 2020, Wales on the 28th of February 2020 and finally Scotland on the 1st March 2020. After climbing case and death counts, a UK-wide stay-at-home order was announced on the 23rd March 2020. These restrictions began to be lifted in June 2020, at different speeds between the constituent nations.

The next significant wave of SARS-CoV-2 in the UK began in the Autumn of 2020, and led to the introduction of some localised non-pharmaceutical interventions (NPIs), again different between the four constituent nations, followed by another UK-wide lockdown in January of 2021. This began to be lifted in March 2021, with the removal of all restrictions in England and most restrictions in Scotland, Wales and Northern Ireland in July 2021 despite high case counts. Cases have incrementally increased since then, with an extreme increase towards the end of 2021 due to the latest wave of infection caused by the Omicron variant. This is summarised in Fig. 5.1.

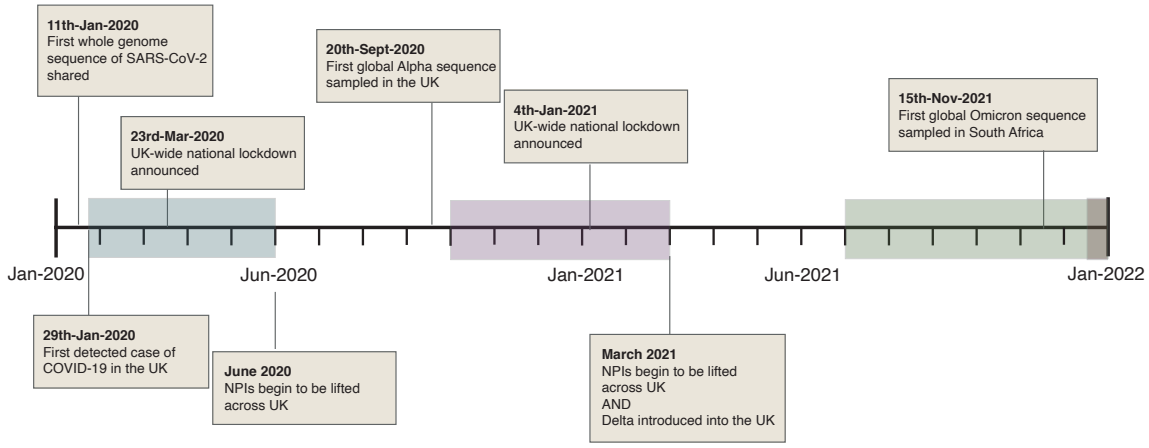

Fig. 5.1 Timeline showing key waves of SARS-CoV-2 infection in the UK (coloured rectangles) and timing of NPIs. Note that NPIs were lifted at different times across the four constituent nations but general trends were similar. The blue rectangle is the first wave, the pink rectangle is the Alpha wave and the green rectangle is the Delta wave. The start of the Omicron wave of infections is indicated at the end of 2021.

Each major wave of COVID-19 cases in the UK has been characterised by different lineages and variants of SARS-CoV-2, has taken place on different backgrounds of interventions, and has different introduction dynamics. The first wave can be characterised by the virus which emerged from Wuhan in China (Andersen et al., 2020) and viruses carrying the D614G substitution, introduced many times into the UK mostly from Europe (du Plessis et al., 2021) on a background of tightening restrictions as the first half of 2020 progressed. The end of 2020 wave can be attributed to the Alpha variant, a variant of concern (VOC) designated B.1.1.7 in the pango nomenclature (Rambaut et al., 2020a), carrying a cluster of 17 non-synonymous amino acid changes. This likely arose in the South East of England (chapter 3), and spread from that single origin point, exacerbated in December 2020 due to a combination of the relaxation of NPIs at the start of December and mass travel associated with the Christmas period (Kraemer et al., 2021). As it did so, it was able to out-compete other lineages due to its transmission advantage, and was

responsible for 95% new infections in England by the end of March 2021 (Davies et al., 2021a). The Alpha variant was displaced in early 2021 by the Delta variant, designated B.1.617.2, introduced multiple times first from India and then from other countries on a background of loosening restrictions and relatively high population immunity from the roll-out of a vaccination campaign as well as high infection rates in previous waves (McCrone et al., 2021).

Due to levels of high sampling and genomic sequencing in the UK, it has been possible to track how these waves have spread, and see the impacts of loosening and tightening restrictions over time. In this chapter, I summarise the results of my contributions to four publications to compare the dynamics of each new wave using sequences collated from the wider COVID-19 Genomics UK (COG-UK) consortium dataset, each of which was the largest group of genomes to be analysed using these techniques when published.

5.2 Methods

5.2.1 Transmission lineage description and introduction assignment

For each of the analyses of variants which did not originate in the UK, which is the first wave, D614G lineages and the Delta variant, the concept of “transmission lineages” was developed. These are similar to introductions, but are likely to underestimate the true number of introductions due to limited global sampling, and are aimed at identifying lineages which represent transmission within the UK.

A few different approaches were used to identify these transmission lineages, and so they have slightly different definitions depending on the analysis in question.

However, the underlying concept is the same. Broadly speaking, maximum likelihood (ML) trees were generated using either IQTree (Minh et al., 2020) or FastTree (Price, Dehal, and Arkin, 2010), rooted on a high-quality sequence from the start of the pandemic, and then transitions between the UK and non-UK states were estimated to generate transmission lineages.

For the first wave analysis, trees were searched depth-first from each UK sequence until a non-UK node is encountered. All tips encountered in the search are classed as part of the same transmission lineage and a single transmission lineage was classed as having at least two sequences in the UK using this method. Singletons were excluded, although it must be noted that with more sampling, the transmission lineages these singletons belong to may have been sampled. Therefore, the estimate of introductions in the UK are likely an underestimate. See du Plessis et al. (2021) for details.

For the D614G analysis, clusters were found using a delayed transition algorithm. This is a form of a maximum parsimony where transitions into the UK farthest from the root are favoured. See Volz et al. (2021b) for details.

For the Delta analysis, the ML tree was split into three subtrees of roughly equal size (n= 28,783, 28,715, and 36,151). These were then scaled into timetrees using TreeTime (Sagulenko, Puller, and Neher, 2018), and used as starting trees to generate an empirical distribution of phylogenies. These empirical distributions were used to estimate lineage transitions into the UK using a three state phylogeographic discrete trait analysis (DTA), assigning sequences to either UK, India or Global (representing every other country in the dataset). Where sequences in the UK had travel history recorded in UK Health Security Agency (UKHSA) datasets, this was included in the inference using a newly developed travel history-aware reconstruction method (Lemey et al., 2020). Introductions were defined as nodes inferred to be in England

with parents in either India or the Global state. The date of importation was assumed to be half-way between a node and its parent. See McCrone et al. (2021) for details.

5.2.2 Genomic data

All genomic data was taken from COG-UK. For the first wave analysis, 26,181 whole genome sequences sampled in the UK before 22nd June 2020 were used. For the D614G analysis, 26,986 whole genome sequences which were unambiguous at 614 locus, and had sample collection dates from the 19th June 2020 data release were included.

For both the Delta and Alpha analyses, whole genome sequences from the "pillar two" (i.e. surveillance) programme with unambiguous postcode districts were used. For both analyses, this only related to sequences from England, as other constituent UK nations did not provide postcodes. For the Alpha variant analysis, this constituted 17,741 whole genome sequences which were assigned B.1.1.7 using pangolin (O'Toole et al. 2021) and sampled between 20th September 2020 and 19th January 2021. For the Delta variant analysis, all sequences assigned Delta using scorpio (https://github.com/cov-lineages/scorpio) on GISAID were downloaded on 15th September 2021 to provide timings of introductions into the UK, and 49,550 sequences from the UK assigned Delta and sampled between 12th March 2021 and 15th June 2021 were used.

5.2.3 Geographical metadata

The first wave and D614G analyses directly used the sampling location of genomes at the administrative level 2 (admin2) level, roughly equivalent to counties in the UK. Alpha and Delta analyses used upper tier local authority (UTLA) data in order

to facilitate comparison to the case database, which is recorded in UTLA. UTLAs are very similar to admin2s and so clean admin2 data can be matched to UTLA data using shapefiles of each and some manual curation; and the same is true with admin3 and local administrative districts (LADs).

To clean the admin2 data in the genomic metadata, i.e. match it to a single standardised set of locations as found in the Global Administrative Database (GADM, https://gadm.org), I developed a set of cleaning scripts, found at https://github.com/COG-UK/geography_cleaning. The cleaning process was run with custom scripts for the first wave and D614G analyses, and had been integrated formally into the COG-UK data processing pipeline (https://github.com/COG-UK/datapipe) in time for the Alpha and Delta variant analyses.

The cleaning process involves several steps. Some sampling locations in the metadata can not be unambiguously mapped to a known location (e.g. “City Centre”), while others were for locations in overseas territories (e.g. Falklands and Gibraltar). Yet other genome sequences had uninformative spatial records (e.g. Yorkshire or Wales), or no admin2 level data at all. I carried out a simple one-to-one mapping where possible, which included correcting spelling mistakes and alternative entries for the same county (e.g. Durham versus County Durham). Locations recorded at a higher spatial resolution were mapped to the corresponding admin2 region (e.g. Solihull was mapped to Birmingham). Where the recorded locations were larger than the admin2 regions (e.g. “West Midlands”), but sequences are commonly from these locations, I combined the smaller regions. I also merge some city authorities with no or very few reported sequences with their surrounding county, on the assumption that the larger county was used to represent the location of city samples (e.g. for Leicester and Leicestershire). The results of these last two processes are recorded in a column

in the metadata called “suggested_adm2_grouping” and the exact groupings used can be found in the readme at https://github.com/COG-UK/geography_cleaning.

For the continuous phylogeographic analyses of Alpha and Delta, postcode districts (i.e. the first half of a UK postcode, e.g. HP22 in HP22 5NS) were used as these are the highest resolution geographical data available in the COG-UK dataset. Some small amount of cleaning can be undertaken (e.g. changing a zero to capital letter O), but usually if a postcode cannot be matched immediately to a real UK postcode, it is discarded from analysis.

5.2.4 Growth rate of D614G lineages

I used a two-epoch coalescent model to estimate a period of exponential growth followed by an independently estimated period of exponential decline on delayed transition clusters of more than 40 sequences from the dataset described above. Note that both of the rates can take either positive or negative values, and the model does not specify growth and decline. The transition time from growth to decline was estimated independently for each cluster using a normal prior with a mean of the 23rd March 2020 (2020.2254), the date of the first lockdown in the UK, and a standard deviation of two weeks.

A normal hyperprior is specified for cluster growth/decline rates for each genotype and the mean and precision of the hyperprior are estimated. The posterior mean growth and decline rates for each genotype are estimated along with the growth/decline rate for each cluster individually. Posterior growth rates within each genotype are therefore correlated. The prior for the mean growth rate is Normal(0,100/year) and the prior of the precision parameter is Gamma(1,0.001).

The model was implemented in BEAST v1.10.5 (Suchard et al., 2018). Four independent chains of 100m states were run for each variant, with 10% removed from each chain to account for burn-in. Convergence was assessed using Tracer (Rambaut et al., 2018) prior to further analysis. The HKY model was used to model nucleotide evolution (Hasegawa, Kishino, and Yano, 1985), and, following Duchene et al. (2020), the evolutionary clock rate was fixed at 0.001 substitutions per site per year.

5.2.5 Within-UK spatial dynamics of Alpha and Delta variants

For both the Alpha and Delta variant analyses, we assigned random coordinates to each sequence within the appropriate postcode district, as the continuous phylogeographic analysis does not permit identical values. This was achieved by choosing random points from shapes of the postcode districts from Pope (2017).

For the Alpha variant analysis, as it arose in the UK (chapter 4), introductions were not inferred as above. Instead, a single time-scaled phylogeny was required to perform the spatial phylogeography. In order to do this, I built an approximately ML tree using FastTree (Price, Dehal, and Arkin, 2010) which I then used as a starting tree for building a more reliable ML tree using the HKY substitution model (Hasegawa, Kishino, and Yano, 1985) in IQTree (Minh et al., 2020). The resulting tree was passed into TreeTime (Sagulenko, Puller, and Neher, 2018) to generate a time-scaled phylogeny, and at this stage molecular clock outliers were removed (Hill and Baele, 2019). The final dataset therefore had 17,716 sequences. The ML and time-scaled tree were used as the inputs for a variation on a common Bayesian phylogenetic analysis using BEAST, using a “thorny tree likelihood” (McCrone, 2021; du Plessis et al., 2021) with a strict clock model and a nonparametric skygrid coalescent prior

(Gill et al., 2012). The tree with the highest likelihood from this analysis was used as the empirical tree for the continuous phylogeographic analysis, shown in Fig. 5.5D.

For the Delta spatial analysis, the seven largest importations (those with >1500 sequences) and all importations with five or more sequences were selected from a representative tree from the posterior set with the same number of total importations as the posterior mean. Only sequences with unambiguous postcode districts were used, and the final dataset was 25,139 sequences for the seven largest transmission lineages and 24,411 across 280 smaller lineages, which were extracted from the master COG-UK alignment, described in the Delta section of "genomic datasets" above.

For both variants, I then reconstructed the geographic movement of nodes on a fixed tree (pruned from the overall Maximum Clade Credibility (MCC) tree for Delta and the time-scaled phylogeny described previously for Alpha) in BEAST v.1.10 (Suchard et al., 2018), using a relaxed random walk (RRW) model (Lemey et al., 2010), and a Cauchy distribution to account for among-branch heterogeneity in dispersal velocity.

For Delta, large lineages were inferred independently, and all small lineages were inferred in a single run, with the shared parameters for likelihood, precision, and covariance of coordinates, but independent estimates of diffusion rate and trait likelihood. Following this run, 22 small introductions were removed due to their chains not converging to the same posterior. An MCC tree was then generated using TreeAnnotator (Suchard et al., 2018) to summarise the posterior tree distribution for all lineages. Visualisations were made using a custom Python script. XML files were generated using beastgen.py (https://github.com/ViralVerity/beastgenpy).

For the export analyses, I compare Greater London to Greater Manchester which consists of the UTLAs Salford, Trafford, Stockport, Oldham, Bolton, Tameside, Bury, Rochdale, Wigan and Manchester.

5.3 Results

5.3.1 First wave and D614G lineages

SARS-CoV-2 was introduced at least 1000 times into the UK before June 2020, mostly from Europe (du Plessis et al., 2021), into multiple different parts of the UK. At this stage of the pandemic, surveillance sampling had not yet been set up, and only hospitalised individuals received a PCR test. Therefore, for the early part of 2020, samples taken for genomic sequencing by the newly formed COG-UK consortium were mostly convenience samples from hospitals. This resulted in good coverage in some areas, but sequencing was not undertaken systematically until community testing was set up in late 2020 (Department of Health and Social Care, 2021). Therefore, for the early part of 2020, it is only possible to perform descriptive analyses due to a lack of nation-wide genomic surveillance.

To illustrate spatial variation in diversity, I characterised the spatial distribution of transmission lineages in the UK at an admin2 level (see Methods). Specifically, I examined the substantial variation among regions in diversity of transmission lineages using Shannon's Index (SI; this value increases as both the number of lineages and the evenness of their frequencies increase; Fig. 5.2A). The highest SIs were in Hertfordshire (4.77), Greater London (4.62) and Essex (4.49); these locations are characterised by frequent commuter travel to and within London, and proximity to major international airports (Greater London Authority Intelligence and Analysis Unit,

2014). Locations with the three lowest non-zero SIs were in Scotland (Stirling=0.96, Aberdeenshire=1.04, Inverclyde=1.32; Fig. 5.2A).

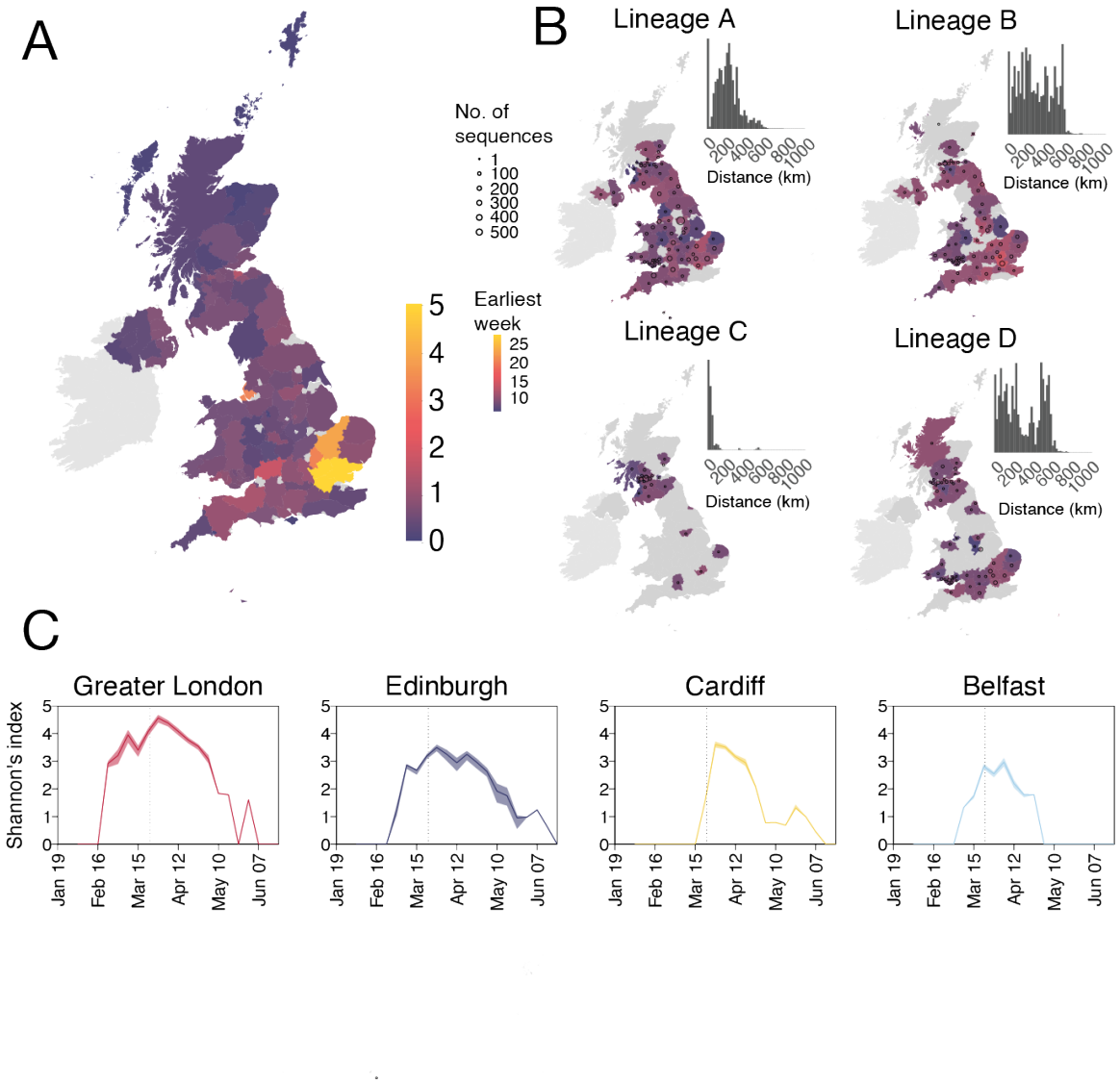

Fig. 5.2 Spatial distribution of the first wave of infections in the UK. A) Map showing Shannon's index (SI) for each region, calculated across the study period (2020-02-02 to 2020-06-26). Yellow colours indicate higher SI values and darker colours lower values. B) Made by Christopher Ruis, an illustration of the diverse spatial range distributions of UK transmission lineages. Colours represent the week of the first detected genome in the transmission lineage in each location. Circles show the number of sampled genomes per location. Insets show the distribution of geographic distances for all sequence pairs within the lineage. C) SI through time for the UK national capital cities.

To examine temporal trends in transmission lineage diversity, I plotted SI through time for each of the UK's national capital cities (Fig. 5.2C). Lineage diversities in each peaked in late March and declined after the UK-wide lockdown. Greater London's epidemic was the most diverse and characterised by an early, rapid rise in SI (Fig. 5.2C), consistent with epidemiological trends there (https://coronavirus.data.gov.uk/, Angus, 2022). Belfast's lineage diversity was notably lower. That Greater London has the highest diversity is expected, given that Greater London has by far the largest population of the capitals (9 million) compared to Cardiff (481,000), Edinburgh (488,000) and Belfast (343,500).

There is also variation in the spatial range of individual UK transmission lineages. Although some lineages are widespread, most are more localised and the range size distribution is right-skewed as most lineages are small (du Plessis et al., 2021). Transmission lineage A (Fig. 5.2B) is spread across the UK, and is only absent from the far north of Scotland and a handful of other dispersed small locations. Transmission lineage B is geographically dispersed (>50% of sequence pairs sampled >234km apart). It is noticeable that transmission lineage B appears to spread up the east coast of England, which has strong road and rail connections from London up to Newcastle, Edinburgh and Aberdeen. In comparison, transmission lineage C is strongly local (95% of sequence pairs sampled <100km apart) and transmission lineage D has multiple foci of sampled genomes (Fig. 5.2B). The pattern of the first wave of infection therefore arose from the combination of heterogeneous lineage-specific dynamics.

During the first wave of infections, systematic genomic variation had begun to appear in the SARS-CoV-2 population. The mutation D614G, the first substitution in the spike protein to sweep through the UK epidemic, was thought to increase transmissibility as it increases infectivity *in vitro* (Zhang et al., 2020; Korber et al.,

2020; Yurkovetskiy et al., 2020); it makes ACE2 binding more likely (Yurkovetskiy et al., 2020); and the observed domination of 614G viruses over 614D viruses globally appeared to suggest a selective advantage (Furuyama et al., 2020; Korber et al., 2020). It arose early on in the pandemic, associated with the Pango lineage B.1 which was first sampled in late January and early February in China, Germany and Italy (Rothe et al., 2020; Lai et al., 2020).

Maximum parsimony methods identified 245 614G and 62 614D clusters containing 10 or more UK virus genomes (Methods). There were more clusters in the UK which carried the 614G variant compared to the wild-type, and on average the 614G clusters had a later first sample date (mean detection date was 16 days later than the 614D lineages, Fig. 5.3B). Variants with 614G also took over spatially over the first half of 2020, with more UTLAs reporting more 614G samples than 614D in April 2020, and very few reporting any 614D samples by June 2020 (Fig. 5.3A).

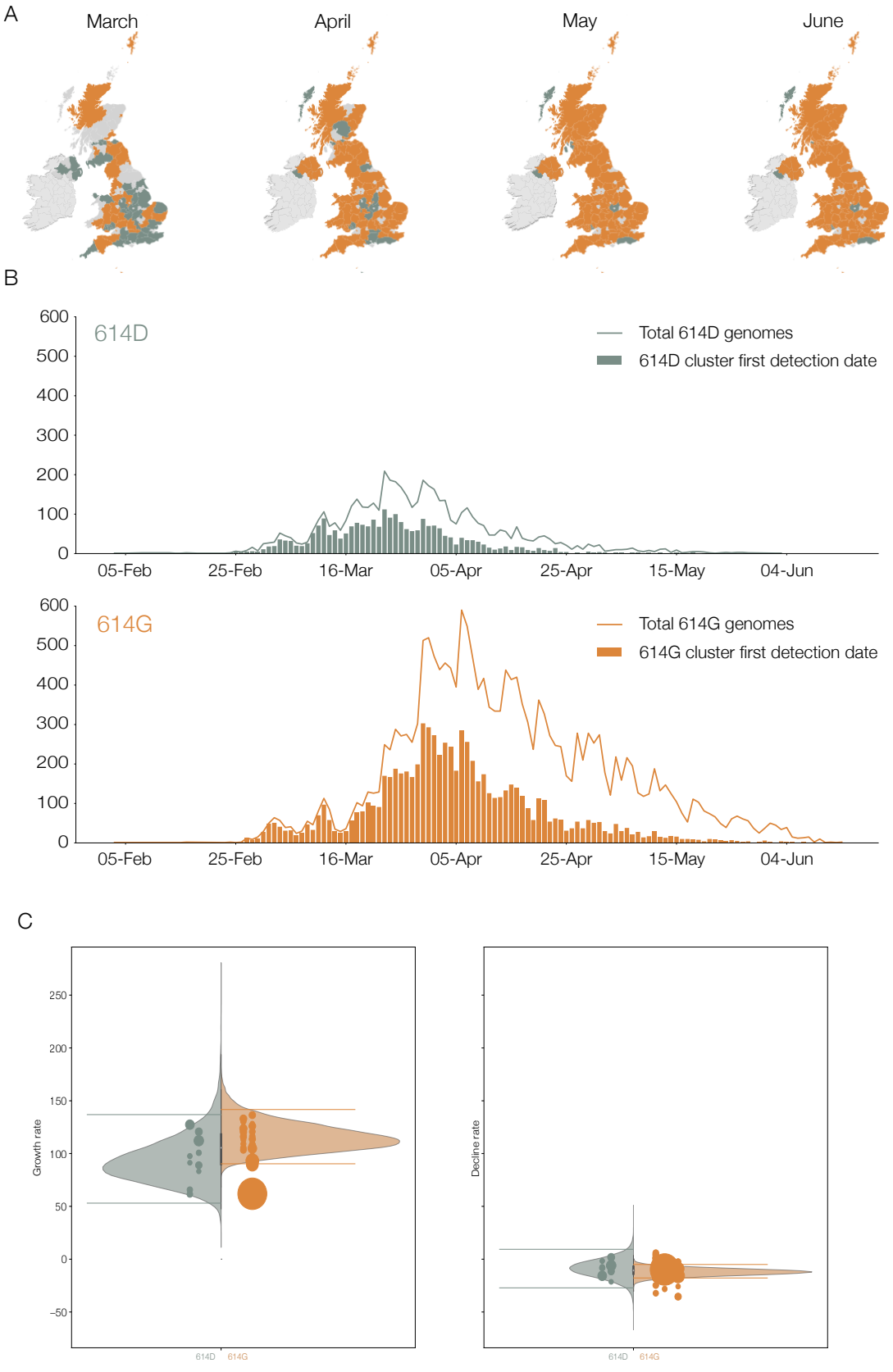

Fig. 5.3 (Caption on next page.)

Fig. 5.3 (Previous page.) Geographic and temporal distribution of UK phylogenetic clusters, classified as 614D or 614G according to the residue they carry at S protein position 614D. A) Shaded regions show the predominant residue in each region on the 15th of each month for March, April, May and June 2020, with orange indicating that 614G was more frequently sampled and green indicating that 614D was more (or equally) frequent. Grey indicates that no sequences had been sampled by that point in time. Light grey indicates the Republic of Ireland. B) The date when each cluster was first detected in the United Kingdom, for variants 614D and 614G. Each cluster contains 2 or more sampled genomes. Solid lines show the total number of sequences collected by day of each 614 variant. C) Distribution of exponential growth and decline rates for Spike 614G (brown) and 614D (grey) in units of 1/year. Solid areas span the 95% credible interval. Points indicate the rates estimated for specific clusters, and are sized by the number of sequences in that cluster.

To investigate the relative growth rates of 614D and 614G lineages, I applied a parametric 'boom-bust' exponential model to all of the lineages containing more than 40 samples (n=50, 614D=11 and 614G=39). This models the virus population as growing exponentially until a transition time, whereupon it shrinks exponentially. These rates vary between each cluster, and a joint estimate for the two variants was obtained using a hierarchical model. Among these larger lineages, those with 614G tended to start later and persist longer than 614D (Fig. 5.4), and 614D lineages tended to have slightly earlier transition times (614D mean = 25th March 2020 versus 614G mean = 1st April 2020). Both growth and decline rates for 614G clusters tended to be larger (mean growth = 114 year-1 versus 93 year-1 and mean decline = -11 year-1 versus -9 year-1), but these differences were non-significant (Fig. 5.3C).

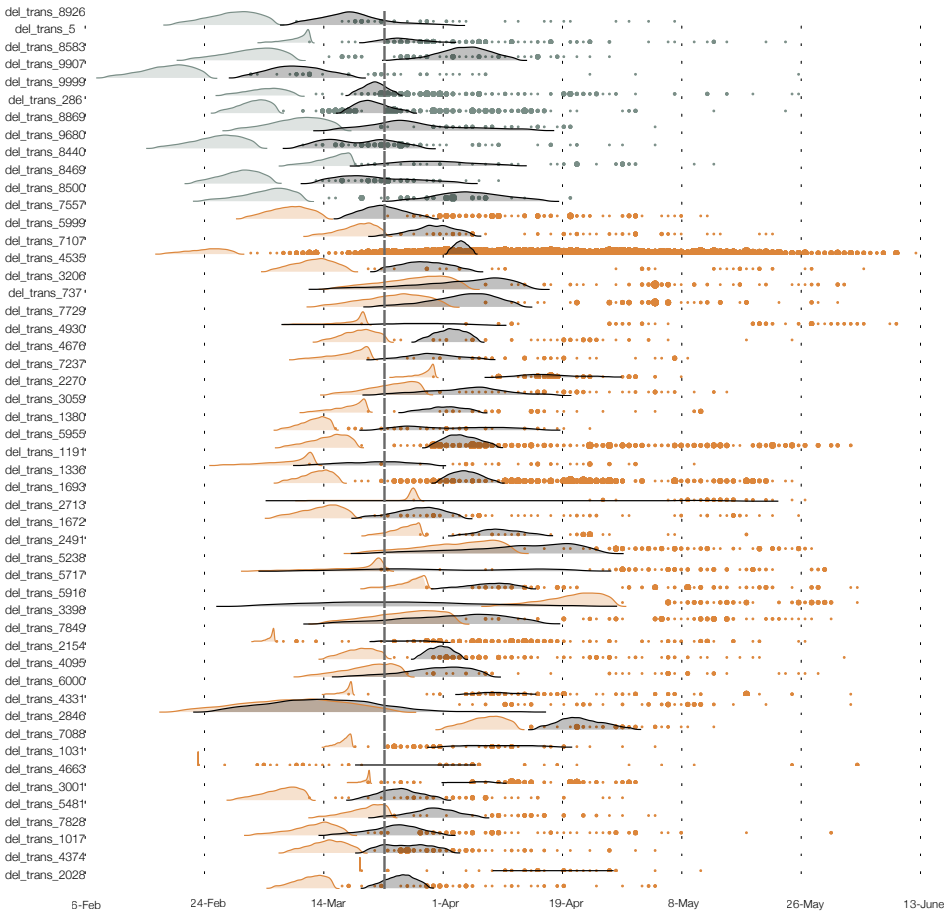

Fig. 5.4 The estimated TMRCA for each of 50 UK clusters (shaded density) and time of each sequence sampled (points). Brown and grey respectively indicate 614G and 614D clusters.The start of UK-wide lockdown on the 23rd March 2020 is indicated by the vertical line. The black density on each line is the distribution of the transition time.

Thus, the differences between lineages with 614G and 614D appears to be mostly in timing, with 614G lineages arriving later, as they reached high enough prevalence in other countries to be able to be exported. It is true that 614G lineages on the whole were larger, although this could be due to factors such as where they were introduced.

As discussed earlier, lineage diversity and case counts in Greater London were very high compared to the rest of the country, and so a lineage introduced here would likely be large as long as it was at least as fit as the wild-type virus. While there is some evidence of a selective advantage for 614G viruses (Volz et al., 2021b), any advantage in transmission was slim and difficult to identify.

5.3.2 Alpha variant

Following the discovery of the B.1.1.7 lineage (later denoted Alpha) in mid-December 2020, retrospective investigation led to the identification of the earliest known case in the COG-UK dataset from 20th September 2020 in Kent in the South East of England. It contains 17 non-synonymous amino acid changes (14 mutations and 3 deletions, chapter 3), and has been shown to be much more transmissible compared to viruses only carrying D614G (Davies et al., 2021a; Volz et al., 2021a). It spread rapidly across the UK, seeding approximately seven new UTLAs per week. Importantly, Alpha had already been detected in several UTLAs before the start of the English national lockdown on the 5th November 2020. When this lockdown ended a month later on the 2nd December 2020, Alpha was already widespread in the UK (Fig. 5.5A).

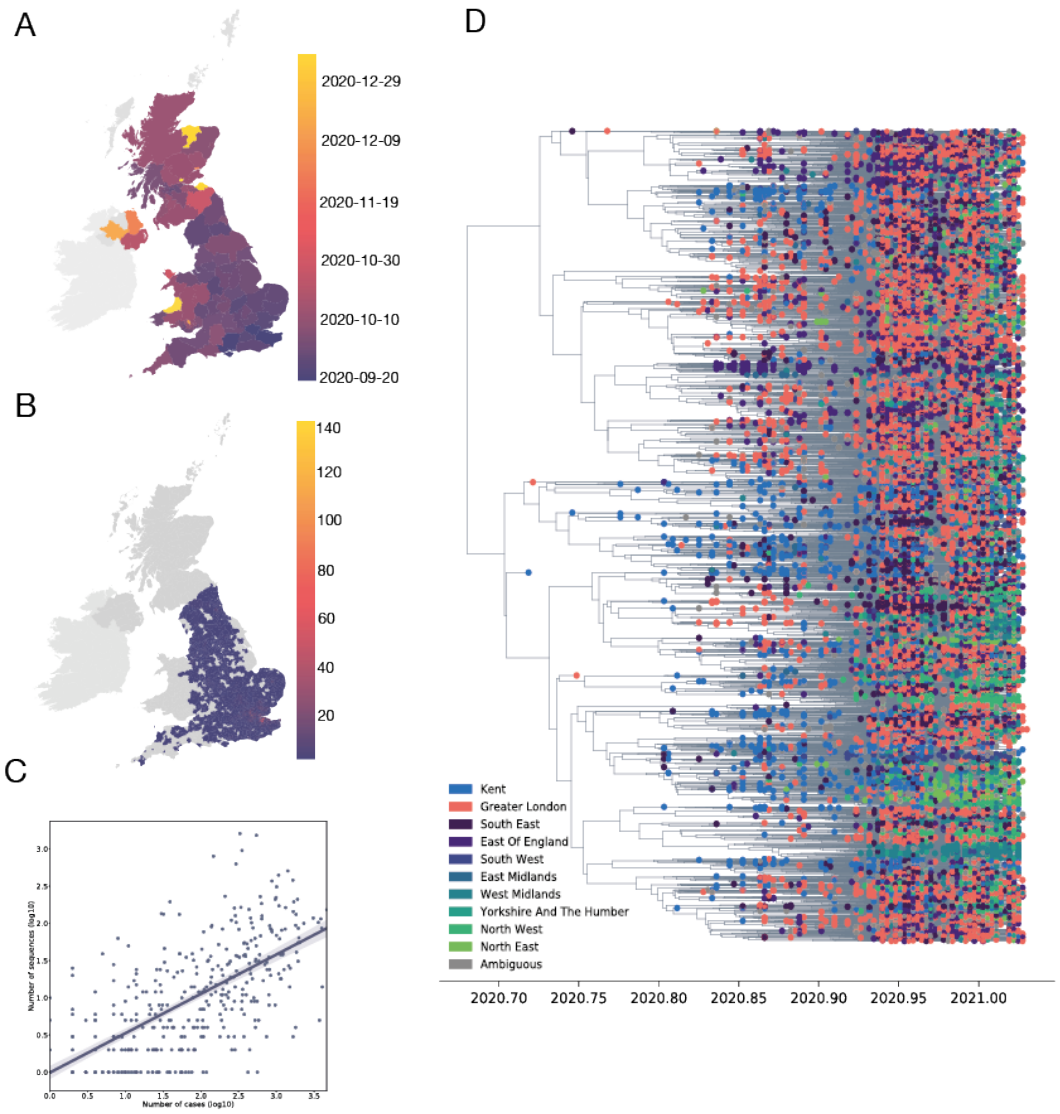

Fig. 5.5 A) Map at the UTLA level (n=115) of arrival dates of the Alpha variant. Darker colours indicate earlier dates and lighter colours later dates. Arrival time is defined as the earliest sampling date of an Alpha genomic sequence in each UTLA. B) Map showing the number of sequences in each postcode region that are Alpha and from pillar two sequencing facilities. Locations in grey have no sequences meeting these criteria, and the Republic of Ireland is shown in a lighter grey. C) Correlation between the logged number of sequences used in the dataset for the continuous phylogeography and the logged number of SGTF positive cases per location and over time (Pearson's r = 0.69, 95% CI: 0.63 - 0.73, p-value < 0.001). D) Time-scaled phylogeny of 17,719 Alpha sequences from England used as input for the continuous phylogeographic analysis. Tips are coloured by NUTS1 administrative region, or highlighted separately if from Kent. Visualised using Baltic (https://github.com/evogytis/baltic).

In comparison to the first wave of SARS-CoV-2 in the UK, the Alpha variant emerged in the South East of England and spread across to the UK from a single point, rather than from multiple introductions. To explore the spread of the Alpha variant as it emerged, I reconstructed viral lineage movements using a continuous phylogeographic approach (Lemey et al., 2010). I analysed 17,719 genomes from surveillance sampling in England collected between 20th September 2020 (the date of the first sequence with the whole Alpha constellation) and 19th January 2021 (Fig. 5.5B and Fig. 5.5D). This represented 4% of all S-gene target failure (SGTF) sequences in the same time period. Samples per location and per week in the SGTF and whole genome datasets are strongly correlated (Pearson's r = 0.69, 95% CI: 0.63 - 0.73, p-value < 0.001, Fig. 5.55C, Volz et al., 2021a), making it feasible to reconstruct Alpha expansion history using phylogeographic approaches (Kalkauskas et al., 2021; Kraemer et al., 2018).

There were distinct phases to the spread of Alpha across England. During the English lockdown in November, almost all (99.2%) of Alpha viral movements started and ended in Greater London or Kent, with infrequent long distance dispersal events (Fig. 5.6A). After the relaxation of NPIs at the start of December 2020, virus movement from the South East of England to other regions increased in frequency, and local transmission between and around larger cities began to be observed. After this, the South East of England was placed into Tier 4 (see Nomenclature) on 20th December 2020, reducing mobility from Greater London. However, the total number of exports of Alpha viruses did not immediately decline, as the large increase in cases in the South East of England compensated for the decline in travel numbers.

In December, the role of Greater London as a source of infection for the rest of England began to decrease: prior to the November lockdown, 49.4% of all between-UTLA movements began in Greater London (Fig. 5.6C). This proportion reduced to

47.2% during the lockdown and then down to 43.6% in December, as other parts of the country began to have large enough local epidemics to maintain regional circulation. The length of movements in kilometres in all three time periods was right skewed, with many more between postcode movements overall in the final time period compared to the other two, as lockdown in England eased and travel associated with Christmas gatherings began. It is also worth noting that the shape of the distance kernels are not smooth declines in frequency as the distance increases: instead, in November and early December, there is a second peak at 250-300km movements (Fig. 5.6D). This is similar to the distance from Greater London to Manchester (approximately 260km) and other major cities in the same area, such as Sheffield.

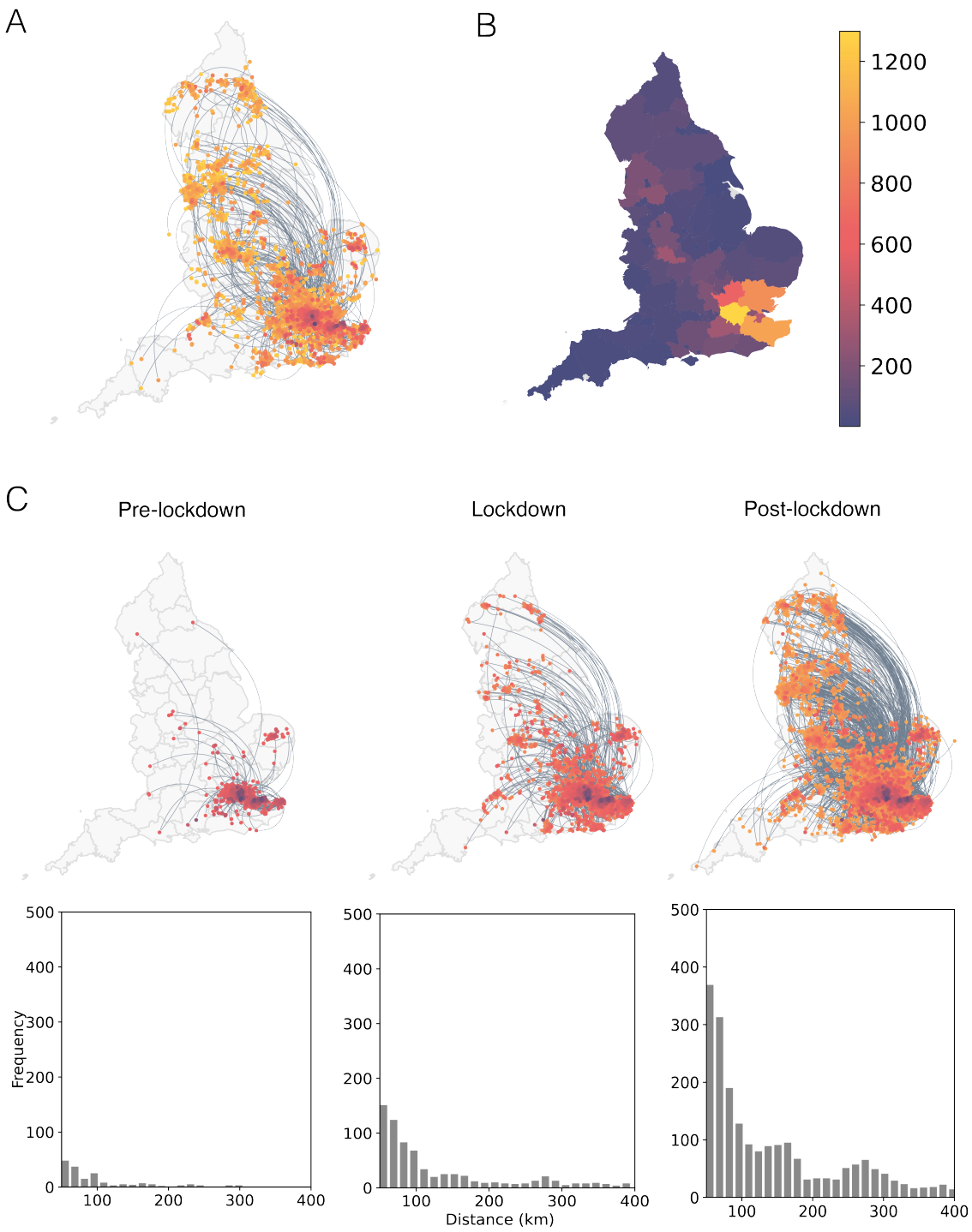

Fig. 5.6 Spatial emergence dynamics of the Alpha variant in England. A) Continuous phylogeographic reconstruction with phylogeny nodes coloured according to their time of occurrence and dispersal direction of phylogeny branches indicated by edge curvature (anti-clockwise) for the whole study period B) Estimated number of cumulative B.1.1.7 introductions inferred from phylogeographic analysis into each administrative area (UTLA). C) Phylogeographic reconstruction limited by time, from left-to-right, data to 2020-09-05, 2020-12-01 and 2020-12-201, respectively. Distance kernels shown underneath the related time period.

Throughout, the weekly number of Alpha cases in a UTLA was positively associated with the number of Alpha introductions into that UTLA during that week (Pearson's r = 0.41, 0.76, 0.91, and 0.73, for October, November, December and January; $p < 0.001$ for all; Fig. 5.67). There was large heterogeneity in the numbers of introductions between UTLAs (Fig. 5.6B), with the largest numbers in Inner London (n=1299), Kent (n=1031) and Essex (n=910); and the lowest in North Lincolnshire (n=1), Hull (n=2) and Torbay (n=3). The next four largest numbers of introductions are in the commuter belt of Greater London, indicating the importance of regular, short journeys for viral movements.

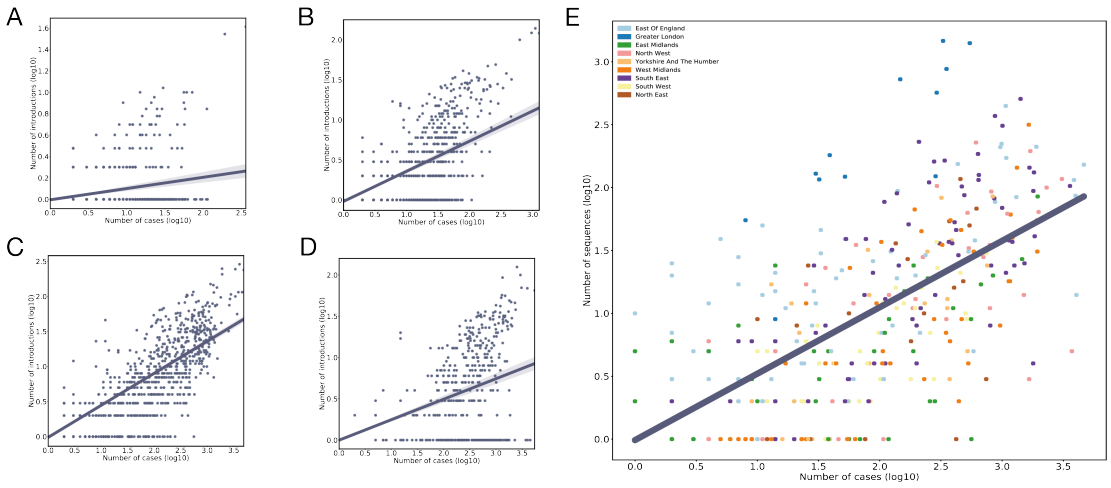

Fig. 5.7 Correlation between the number of imports of Alpha into a UTLA per week and the number of Alpha cases in the same UTLA and same week in October (A, Pearson's r = 0.41, 95% CI: 0.38 - 0.44), November (B, Pearson's r = 0.76, 95% CI: 0.74 - 0.77), December (C, Pearson's r = 0.91, 95% CI: 0.90 - 0.92) and the first two weeks of January (D, Pearson's r = 0.73, 95% CI: 0.71 - 0.92). E) Correlation between numbers of cases (SGTF) and number of sampled Alpha genomes across UTLAs in England, coloured by NUTS1 region. One such bias that was noted when regressing the number of SGTF cases against the number of Alpha sequences in each UTLA was that Greater London had a higher proportion of samples sequenced than other parts of England.

5.3.3 Delta variant

As the spread of Alpha variant was curtailed by strict lockdowns in the first few months of 2021, more individuals in the UK had received vaccines and NPIs were beginning to be relaxed, the Delta variant began to be sampled in the UK. First detected in India in late 2020, Delta carries 30 mutations compared to the wild-type virus, and evidence shows that Delta has increased transmissibility (Public Health England, 2020b; Sonabend et al., 2021; Elliott et al., 2021), rates of hospitalisation (Sheikh et al., 2021; Twohig et al., 2021), and immune evasion (Lopez Bernal et al., 2021; Eyre et al., 2021; Lucas et al., 2021) compared to Alpha. It displaced Alpha variant in India as it became the variant primarily responsible for a wave of transmission and mortality in India in early-mid 2021 (Kupferschmidt and Wadman, 2021; Vaidyanathan, 2021); and was the dominant variant in the UK by mid-May 2021 (Public Health England, 2020b). Subsequently, many countries have had Delta waves which have continued into late 2021.

Similarly to the first wave, Delta was introduced into the UK more than 1,400 times (McCrone et al., 2021), into multiple different locations. Importations of Delta occurred on a background of relaxation of NPIs in England: on 12th April 2021 outdoor dining and non-essential retail reopened, and on 17th May 2021 restrictions on indoor dining and international travel were relaxed (Cabinet Office, 2021). The relative frequency of Delta genomes in England increased rapidly during May and COVID-19 cases subsequently increased (Willis, 2021). Initially, Delta transmission clusters were concentrated in the North West of England and were commonly associated with returning travellers (Challen et al., 2021; SPI-M, 2021; Ferguson, 2021).

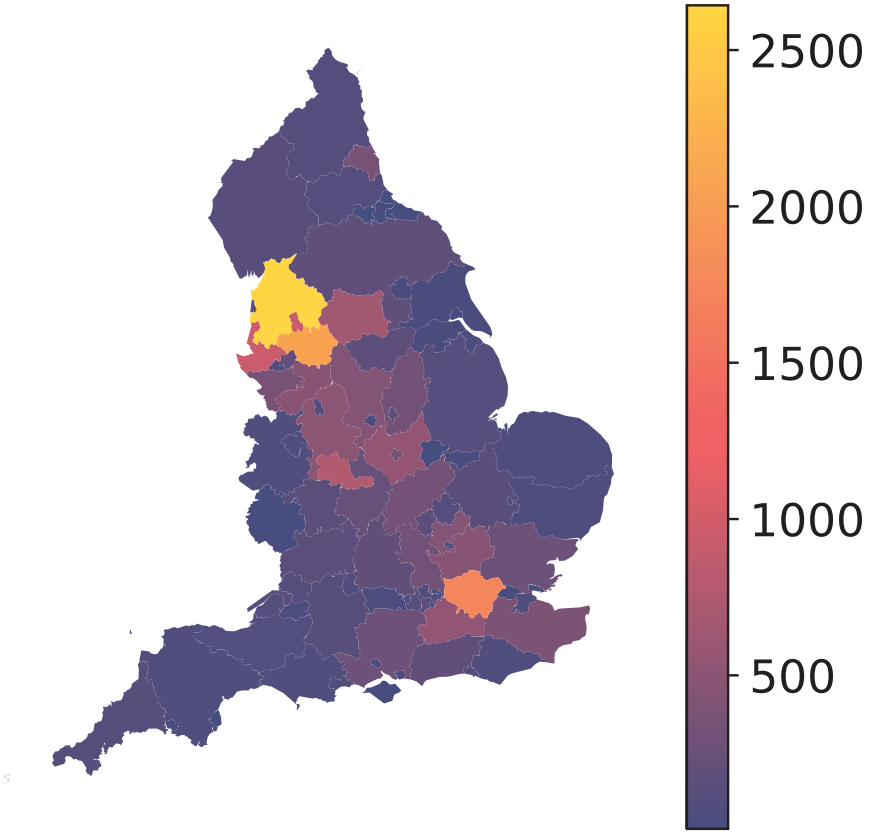

Fig. 5.8 Number of independent introductions per UTLA in England based on continuous phylogeographic analysis of all Delta transmission lineages with >5 sequences.

I analysed all identified Delta transmission lineages in England using continuous phylogeography, thereby reconstructing their dissemination across England. I observed high heterogeneity among UTLAs in the numbers of Delta introductions from other English regions (Fig. 5.8), with Lancashire and Greater Manchester each receiving >2000 estimated independent introductions and Torbay only 9. Greater London also received many Delta cases from elsewhere in England (Fig. 5.8A), as expected, given its population size and connectedness to other metropolitan areas

(Kraemer et al., 2021), similar to the first wave. The majority (n = 11,960) of Delta sequences in England belonged to a single transmission lineage (lineage I, Fig. 5.9), which was sampled mostly in Greater Manchester and Lancashire, and there are many short-range lineage movements among UTLAs in these areas Fig. 5.8). Transmission lineages II and III each comprise 3000-4000 genomes; the former is distributed across multiple urban areas (especially in the North West) whilst the latter is focussed in Greater London and the South East (Fig. 5.9). I also highlight transmission lineage V (Fig. 5.9), originally centred in Bedfordshire, the location of one of the first Delta outbreaks in England and was subjected to surge testing (BBC News 2021). The other three large transmission lineages are also shown in Fig. 5.9, and show similar dynamics to lineage I and II, with foci in the North West, South East and spread to other population centres.

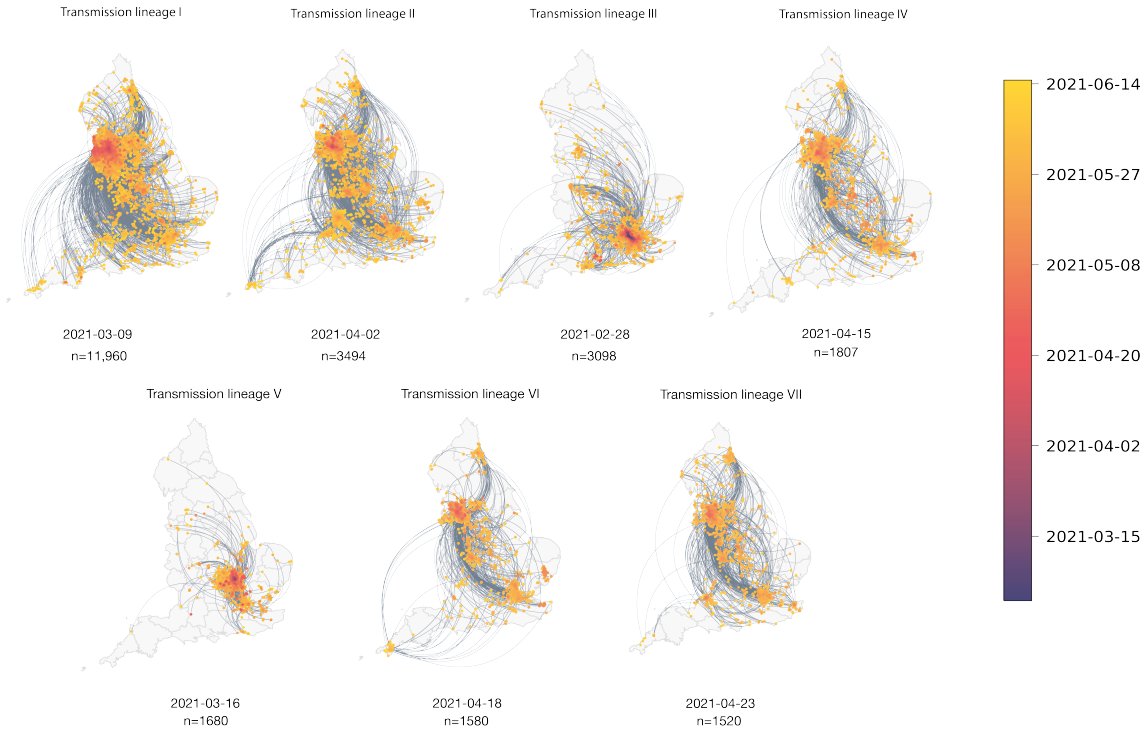

Fig. 5.9 Virus lineage movements inferred by continuous phylogeographic analysis for the seven largest transmission lineages. Direction of lineage movement is anti-clockwise, and dots represent the start and end points of movement, coloured by inferred date. The size and inferred TMRCA of each lineage is shown below each map.

In early May, the number of virus lineage movements among locations accelerated (Fig. 5.10A and Fig. 5.11A and Fig. 5.11B), showing that growth in Delta frequency was associated with regional dissemination. This spread occurred on the background of relaxing NPIs and increased mixing (between mid-January and June 2021, mobility in England increased from 20% to 70% of its pre-pandemic level and estimated mean daily contacts rose from approximately 2 to approximately 5, Jarvis et al., 2020). In general, as NPIs were progressively relaxed through time, long-range viral lineage movements comprised an increasing proportion of all movements (Fig. 5.10B).

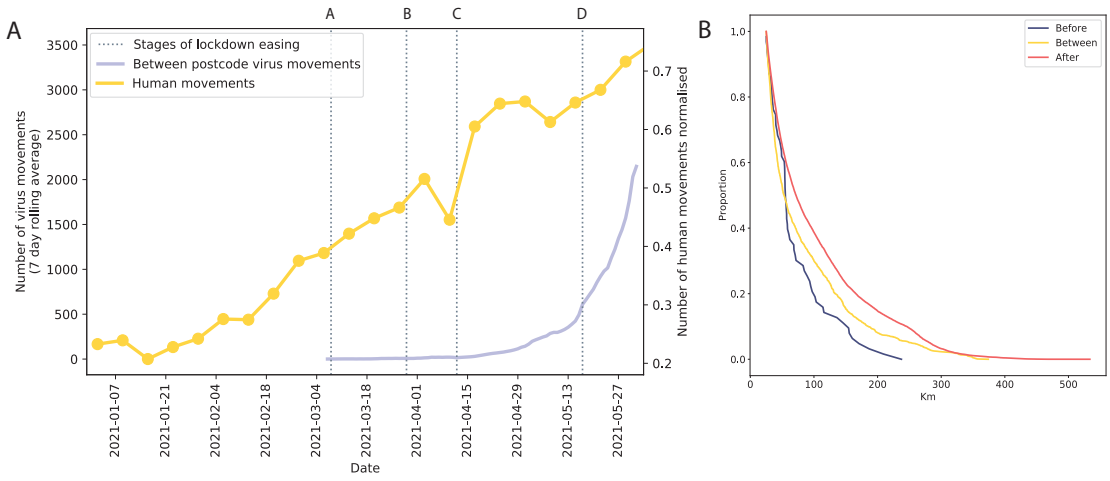

Fig. 5.10 Effect of NPIs on distance of Delta variant transmission lineage movement. A) Trends in aggregate human mobility and the number of virus lineage movements among postcode districts. Letters denote stages of lockdown easing: A (2021-03-08) schools reopen and limited mixing between households outdoors permitted; B (2021-03-29) “Stay at home” directive lifted, more outdoor mixing allowed (up to six people from two households; C (2021-04-12) non-essential retail re-opened, outdoor dining permitted, holiday lets and campsites re-open; D (2021-05-17) indoor hospitality opens, indoor mixing permitted. B) Proportion of virus lineage movements between postcodes >25 km apart: y-axis denotes the proportion of movements that are less than or equal to the value on the x-axis. This is shown for movements before lockdown easing on 2021-04-12 (blue), between 2021-04-12 and 2021-05-17 (yellow) and after 2021-05-17 (red).

For the seven largest Delta transmission lineages in England (I-VII), there were 3 times more exports from Greater Manchester than from Greater London. This difference matches early epidemiological data: the largest and earliest Delta outbreaks were located in the North West (on May 21 Bolton had 452 cases per 100,000 whilst Greater London had 21.6, https://coronavirus.data.gov.uk/). Introductions of Delta into other, smaller urban areas also spread rapidly (e.g. transmission lineage V, Fig. 5.9) and were important for the propagation of the variant across England. Similarly to the Alpha analysis, the frequency of viral movements decrease with distance away from the origin, and there is a second peak at approximately 260km (Fig. 5.12).

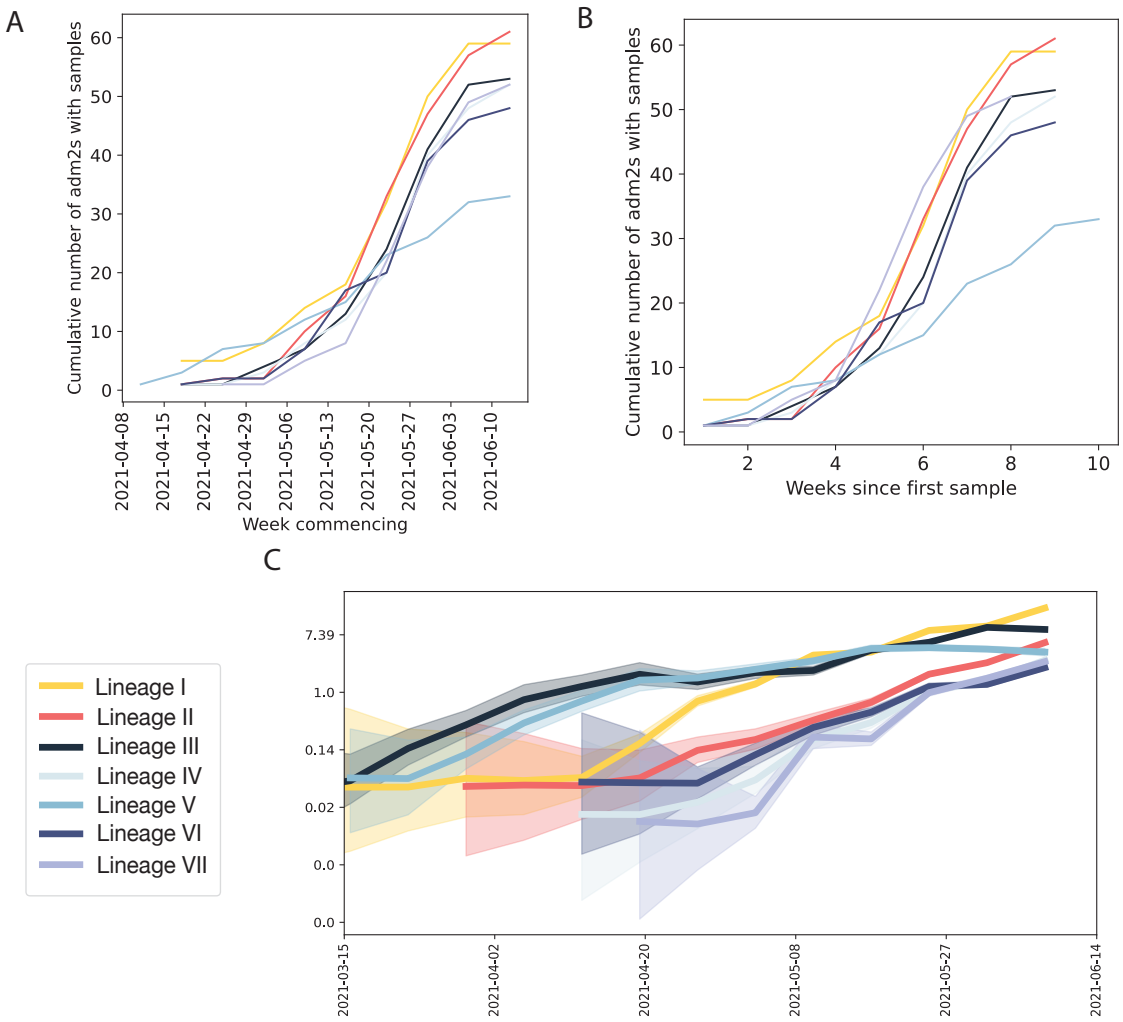

Fig. 5.11 Cumulative number of UTLAs that the seven largest Delta transmission lineages are sampled in A) absolute and B) relative time. C) Viral population growth of the largest seven lineages.

Although North West England was a focus of early Delta transmission, the Delta epidemic in England derived from many successful independent international importations, in contrast to the Alpha variant discussed above. Each of the main Delta transmission lineages in England grew at a similar rate (McCrone et al., 2021, Fig. 5.11C) and time. The spatial expansion of Delta transmission lineages

plateaued after early June, when most UTLAs had established Delta transmission and the relative frequency of Delta genomes in England had exceeded 90% (https://covid19.sanger.ac.uk/lineages/raw).

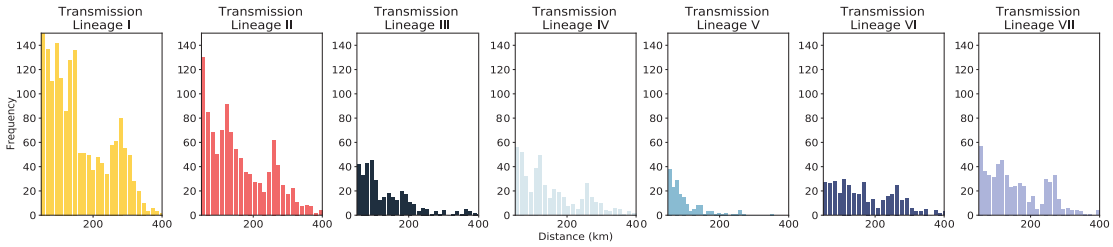

Fig. 5.12 Histograms of the distance of viral movements over 50km for each of the largest seven Delta transmission lineages in England.

Although Scotland, Wales or Northern Ireland are not included here, case count data suggests that cities in England (https://coronavirus.data.gov.uk/details/cases) were the main source of the expanding Delta epidemic in the UK. Further, of the Delta genomes available before 15th June 2021, 57,592 were from England, 9738 from Scotland, 1067 from Wales and 325 from Northern Ireland. I therefore do not anticipate omitting these countries substantially affects the reconstruction of epidemic dynamics in England.

5.4 Discussion

The SARS-CoV-2 epidemic in the UK has consisted of multiple waves with different spatial dynamics. They have all been characterised by different genomic compositions, as well as a mosaic of NPIs and population immunity. However, common themes appear when comparing the three waves together.

When a wave begins elsewhere in the world, but with strong connections to the UK, we see a large number of introductions into multiple different regions at the same time. For both the Delta wave and the first wave, it is clear that the UK was seeded many times. The origin location does make a difference however - with the first wave, the origins of viral lineages were mostly in mainland Europe (du Plessis et al., 2021), whereas for Delta, they were mostly in India (McCrone et al., 2021). While there were many transmission lineages, most of the Delta sequences in England were from a small number of introductions into the North West, which has a large number of people of South Asian ethnicity, and so plausibly spread from individuals returning back to England after visiting family. However, there are many more routes to enter the UK from mainland Europe, including flights to most major cities as well as routes by sea and land. Accordingly, we see more synchronous introduction times, and a slightly less skewed distribution of lineage sizes. In comparison, the Alpha wave began from a single spatial origin in the UK and spread from there, so we observe fewer co-circulating lineages in the early phase of the wave, as the locations were seeded more asynchronously.

The impact of NPIs on each different wave can also be identified by examining the temporal and spatial patterns. The Delta wave and first wave provide a ready comparison, as they both began with many introductions, but the former was on a background of loosening restrictions and the latter caused the first and most strict national lockdown. For the first wave, we see that lineages that were introduced into the UK earlier were able to grow larger (du Plessis et al., 2021) compared to those that were later. The growth rates of both 614D and 614G lineages declined shortly after the stay-at-home order came into effect. This is in comparison to Alpha, where the English national lockdown in November was insufficient to drop the growth rate into negative (chapter 4). For Delta, we see that longer-range movements became more

frequent as it became legal to stay over-night away from home and the in-country holiday sector was reopened. Similarly, with Alpha, we see a large increase in both absolute number and magnitude of movements in the post-lockdown, pre-Christmas period. It is clear that while there are inherent differences in transmissibility between the viral lineages responsible for the three waves, human behaviour and NPIs are still key to their dynamics.

Greater London is central to the dynamics of SARS-CoV-2 in England. This is expected under simple gravity models of disease spread (Bailey and Gatrell, 1995) as it is large, densely populated and well connected to the rest of the UK by air, road and rail. Short-range movements in and around Greater London, likely associated with commuters, were vital for the early maintenance of the Alpha wave, even under relatively strict NPIs. However, while Greater London is also important for the spread of Delta, Greater Manchester and its surroundings are more central in the network, with similar short-range movements. While it is unsurprising that Manchester, another large and well-connected city, can be central to a spatial network, we might expect variants to accelerate spatially once they reach London. For Delta, this was not the case, as while Greater London is the source and destination of many viral movements, it does not take over from Greater Manchester as the centre of the network in this time period. Instead, perhaps while the main source of infection needs to be a well-connected, large and densely populated area, if there is more than one of these locations then the region that the variant is initially introduced into is important. In other words, even though Greater London is larger, the early seeding of Greater Manchester by the Delta variant enables it to remain the major source. Therefore, if a variant originated or was introduced first into another large city, e.g. Birmingham, we would expect that city to become the focus of transmission based on these results, and it would not be reliant on reaching Greater London to accelerate its spread across

the UK. Finally, movements between Greater London and Greater Manchester appear to be important in the spread of both Alpha and Delta variants, with a disproportionate frequency of movements that are the distance between the two cities. These findings may impact the design of more localised interventions, highlighting key routes which may be disrupted to disproportionately impact the spread of the virus.

Comparing these four analyses highlights the importance of a coherent, nationally coordinated testing and genomic surveillance framework. Without population testing and surveillance sequencing, the analyses performed on the first wave and D614G lineages are much less complex and must be limited to descriptive analyses due to sampling bias. In the Alpha and Delta waves, which arose when there was community testing and after pillar two sequencing had been set up, analyses which rely on a more random sample of both the spatial and genomic variation can be performed, and more detailed conclusions drawn. Even with these systems in place, there will still be reduced representation from populations less likely to seek medical care or tests (Public Health England, 2020a) and there is some geographic variation within England in the proportion of cases sequenced. For example, in the Alpha analysis, one such skew that was noted when regressing the number of SGTF cases against the number of Alpha sequences in each UTLA was that Greater London had a higher proportion of samples sequenced than other parts of England (Fig. 5.7E). Sampling biases can drastically affect continuous phylogeographic analyses (Kalkauskas et al., 2021), although the sheer volume of data presented here should help to mitigate underlying skews. As national genomic sequencing efforts are accelerated careful consideration of the genomic sampling framework will be needed to avoid biases (Kraemer et al., 2018).

It is clear that new SARS-CoV-2 variants will continue to evolve in many different countries around the world. The latest, Omicron, has completely dominated most

country's epidemics and led to record numbers of cases. At the same time, political will to enact strong enough NPIs to bring growth rates down has waned. High-detail genomic surveillance programmes can help to identify key spatial routes of transmission, and so provide targets for more localised interventions which may be more palatable to elected officials and the wider population.

6

Discussion

Viruses transmit and evolve across different resolutions and scales, from evolution within a single host, to local transmission and finally to national and international spread. This thesis has been an exploration of how genomic epidemiology and phylodynamics can be applied to explore these different scales and inform public health interventions. Specifically, it has examined the Ebola virus epidemic in West Africa from 2013-2016, and the SARS-CoV-2 pandemic from 2019-present with a focus on Sierra Leone and the UK respectively. These two epidemics are history-making, both in terms of the realisation of how poorly the world is protected from infectious disease, and how genome sequencing can be integrated into public health.

In chapter 2, I expanded on previous work exploring the dynamics of Ebola virus in Sierra Leone (Park et al., 2015; Dudas et al., 2017; Dellicour et al., 2018). Specifically, I used discrete phylogeographic methods to identify drivers of transmission on a national scale, and explored whether the gravity model of infection was a useful way to describe the spread of Ebola virus across Sierra Leone. I found that there were two main foci of infections, first in the far east of the country where the epidemic was seeded, and second in the far west at the capital city of Freetown. The more

stable dynamics after reaching Freetown and settling into the gravity model may be more difficult to disrupt compared to the more stochastic first phase of the epidemic, highlighting the importance of early interventions.

In chapter 3, I investigated the transmission dynamics of Ebola virus by explicitly including individual-level heterogeneity in infection history and contact structure in an agent-based model, ABSynthE. I then fitted this data using phylogenetic summary statistics from the analysis in chapter 1 to obtain transmission parameters at each of the four hierarchical contact levels. Simulating local transmission chains revealed national-scale dynamics: specifically, that while not every epidemic infected many people, they regularly became large enough to infect almost half of the population of Sierra Leone, regardless of which district they began in. The threshold where the case counts began to decline was lower than the expected herd immunity threshold for a homogeneous population, and approximately half of the cases occurred after reaching this threshold.

In chapter 4, I explored the evolution of the Alpha variant of SARS-CoV-2. I characterised the long phylogenetic branch ancestral to the monophyletic clade, and identified a sequence in the COG-UK dataset which I believe is likely to be a true intermediate sequence. Following this, I confirmed using coalescent and birth-death methods that the non-pharmaceutical interventions (NPIs) in England in November 2020 were not sufficient to push the effective reproduction number of the Alpha variant below 1, although they were successful in this respect with the other co-circulating lineages. I compared the evolutionary rates and mutational profiles of the other variants of concern with Alpha, and found that the Delta variant is the outlier both in terms of distribution of mutations across the genome and an apparent lack of signal of increased evolutionary rate on its ancestral branch. This implies that the other variants, Beta, Gamma and Omicron, may have evolved in the same

fashion as Alpha, which I discuss is likely to be during a persistent infection within a single host, who is not necessarily immunocompromised.

Finally, in chapter 5, I explored the dynamics of the major waves of SARS-CoV-2 infection in 2020 and 2021 in the UK. I began by exploring introductions of the first wave of SARS-CoV-2 from Europe, and identify that the main difference in viral lineages carrying the D614G Spike mutation is in the timing of their introductions. Then, I conducted a continuous phylogeographic analysis of the spread of the Alpha variant across England from a single point in the South East; and for the different transmission lineages of the Delta variant. By comparing these waves together, it is possible to identify common themes of SARS-CoV-2 spread across England, specifically that urban centres are important to spread, but that Greater London in particular does not appear to be more important than others. Further, each wave occurred on different levels of tightening and loosening restrictions, and showed that NPIs are key to the dynamics of spread on a national level.

In this chapter, I will compare the West African 2013-2016 Ebola virus epidemic and the SARS-CoV-2 pandemic, exploring wider trends in the epidemics, and whether there are commonalities between the two countries I have focused on. I will then discuss the future of genomic epidemiology for public health, in terms of practical and ethical issues, as well as emerging technologies and data that will help to expand our current knowledge base.

6.1 Ebola virus versus SARS-CoV-2, a tale of two viruses

The diseases caused by Ebola virus and SARS-CoV-2 are extremely different. As discussed in chapter 3, it requires extremely specific contact to contract Ebola virus disease (EVD), as compared to COVID-19 where it is possible, and often likely, that the source of infection is never identified. Pre-symptomatic transmission of EVD has not been documented (Dowell et al., 1999), as compared to COVID-19 where this is a common and important driver of transmission (Casey-Bryars et al., 2021). Finally, while possible in EVD (Timothy et al., 2019), asymptomatic and mild cases are much more common in COVID-19 (Ma et al., 2021). Therefore, Ebola virus is easier to control due to its noticeable and severe symptoms and lack of infectious pre-symptomatic period, but clearly causes much more severe morbidity and mortality.

There are also of course clear differences in the contexts of the two epidemics studied here. Sierra Leone has a mostly agricultural economy, with much of the population reliant on subsistence farming and mining, whereas the UK has a mostly service-based economy (Central Intelligence Agency, 2022). The UK also has an extensive road and rail network, which is important for human and therefore viral movement across the country. Finally, Sierra Leone has one of the worst patient-to-physician ratios in the world, with only two doctors per 100,000 people at the start of the EVD epidemic (Sylvester Squire et al., 2017), compared to 300 doctors per 100,000 people in the UK (World Health Organisation, 2022). However, at each level of transmission and evolution, there are trends that can be compared which may illuminate underlying processes common to different countries and viral epidemics.

At a within-host level, the two viruses provide a good comparison of the different survival tactics against a hostile immune system that pathogens may use, and the impact this has on evolution at a population level. As discussed in chapter 3, during all infections of SARS-CoV-2, there is evidence of a higher evolutionary rate compared to the between-host evolutionary rate, likely due to different selective pressures. It can cause persistent and chronic infections in individuals, who do not have to be immunocompromised, leading to a longer time period at this higher rate of evolution. While the intermediate stages of a persistent infection of SARS-CoV-2 may not be well-adapted to spreading between people (Lythgoe et al., 2017), in some cases the viral population acquires a constellation of mutations that confers a significant fitness advantage at the between-host level over background lineages. I present evidence in chapter 4 that this may be how the Alpha, Beta, Gamma and Omicron variants evolved. In comparison, one of the first indications that Ebola virus can cause persistent infections was due to its lack of evolution compared to the sample time (Diallo et al., 2016). Ebola virus is able to persist by infecting immune-privileged sites such as the testes (Mate et al., 2015), eye (Shantha, Crozier, and Yeh, 2017) or meninges (Jacobs et al., 2016), where its replication slows compared to when it is causing systemic infections. Therefore, the same selection pressure to evade the host immune response can lead to the alternative evolutionary strategies of engaging in a co-evolutionary arms race or hiding in a privileged location; and may result in increased or decreased rates of evolution. The effect on a population scale is important: in the former, it has led to large jumps in fitness and the evolution of variants of concern; whereas in the latter, it has caused small and unexpected transmission chains but no major changes in the evolutionary process.

Both viruses experience a significant amount of individual-level heterogeneity in terms of onward transmission, and superspreading events are key to the development

of outbreaks involving them. As discussed in chapter 3, Ebola virus requires specific forms of contact to transmit, and so some people have more opportunity for onward infection (e.g. if they have a large funeral) and some are at a much higher risk of being infected (e.g. carers) than other members of the population. Superspreading is therefore common, and also important: the epidemic in Sierra Leone started with a superspreading event of 14 secondary cases infected at a funeral (Goba et al., 2016). Though it is easier to spread in general than Ebola virus, there are activities which are particularly effective at spreading SARS-CoV-2 such as singing or speaking loudly (Asadi et al., 2019), and individual choices such as refusing to wear a face-mask also contribute (Howard et al., 2021), and so superspreading is also a key signature of this pandemic (Lewis, 2021). The highest risk events for virus transmission differ between Ebola virus and SARS-CoV-2 because their routes of transmission are very distinct: for Ebola virus, funerals and healthcare settings are the key as these are where infectious bodily fluids are able to come into contact with susceptible individuals. In comparison, SARS-CoV-2 superspreader events have been documented in religious settings, weddings, food-processing factories, hospitals, schools, bars and even at the White House (Majra et al., 2021), as any enclosed, poorly-ventilated, or crowded space is conducive to the inhalation of infectious droplets. The importance of intrinsic host factors in super-spreading is still a matter of debate (Chen et al., 2021), but it is clear that opportunity, the exact type of which depends on transmission route, is vitally important for a superspreading event to occur.

It can be challenging to compare the spread of Ebola virus and SARS-CoV-2 on a local scale due to the radically different contexts of Sierra Leone and the UK, as well as the differences in their infection history. However, both epidemics share an important trait: high levels of infection in healthcare settings. For Ebola virus, the infection route is mostly from patients to healthcare workers (Muyembe-Tamfum

et al., 1999), and by the end of the epidemic in Sierra Leone, 221 health care workers (21% of the workforce) had died (Sylvester Squire et al., 2017). For SARS-CoV-2 in the UK and the US, healthcare workers had approximately three times the risk as the general public of being infected between 24th March and 23rd April 2020, even after adjusting for low population testing (Nguyen et al., 2020). The infection of healthcare workers was exacerbated by a lack of personal protective equipment both for SARS-CoV-2 in the UK (Oliver, 2021) and Ebola virus in Sierra Leone (Raven, Wurie, and Witter, 2018). For SARS-CoV-2, there was also a substantial number of patients infected in healthcare settings (Read et al., 2021), either by other patients, healthcare workers or visitors (Abbas et al., 2021).

Some common trends emerge when considering these epidemics on a national scale. The first is that both spread from major urban population centres: Freetown in Sierra Leone became a transmission hotspot and important source of infections for maintaining the epidemic (chapter 2, Dudas et al., 2017), and Greater London and Greater Manchester were important in the spread and maintenance of diversity of first wave lineages, Alpha variant, and Delta variant (chapter 5). However, in both epidemics, it is also clear that the starting location is important for the initial dynamics, with Kailahun being disproportionately important early on in Sierra Leone (chapter 2), and Greater Manchester being more important than Greater London for the transmission of the Delta variant (chapter 5).

When considering national dynamics, it appears that SARS-CoV-2 in the UK settles into stable dynamics faster than Ebola virus does in Sierra Leone. In other words, the gravity model is more appropriate for broad trends of the spread earlier on. While I did not explicitly examine the SARS-CoV-2 waves in a time-inhomogeneous way, major cities were rapidly colonised by each new wave, and became transmission foci. There are non-biological reasons for this, for example that the UK has better

internal connectivity than Sierra Leone; that lineages from abroad were introduced into major cities; and the one variant that evolved in the UK emerged extremely close to Greater London (as compared to the other side of the country). There are theoretical considerations: the higher R_0 of SARS-CoV-2, which is approximately 2.5 for the original virus from Wuhan (Petersen et al., 2020), in comparison to approximately 1.5 for Ebola virus in Sierra Leone (Khan et al., 2015), lowers the outbreak threshold (Hartfield and Alizon, 2013), meaning that the unstable part of the epidemic where stochastic forces are more impactful, is shorter. Therefore, we expect to observe SARS-CoV-2 in the UK settle into stable gravity model dynamics faster, and we expect the time where fewer interventions are needed to knock the epidemic off-course to be shorter. Importantly, R_0 has increased substantially with each new variant: the Alpha variant is estimated to be 40-90% more transmissible than the wild-type (Davies et al., 2021a), the Delta variant is estimated to be 76% more transmissible as the Alpha variant (Sonabend et al., 2021), and the Omicron variant could be 105% more transmissible than the Delta variant (Sofonea et al., 2022). Therefore, as each new variant emerged, the window of disproportionate impact of interventions was shorter, and it was harder to prevent them from settling into difficult to disrupt dynamics.

The importance based on specific cultural events also led to challenges in disease control in both contexts. Across West Africa, public health teams tried to prevent traditional funeral practices in order to decrease post-mortem transmission of Ebola virus. Measures such as mandatory cremation (Pellecchia et al., 2015) were resisted, and often led to secret burials so that long-standing traditions could be respected (Ryeng, 2015). Individuals in the West have often criticised communities in West Africa for not foregoing ceremony during an epidemic, although some parallels can be drawn from the UK during periods of intense SARS-CoV-2 transmission. The

clearest effect in the genomic data can be seen in the large increase in population movement just before Tier 4 restrictions (see Nomenclature) came into effect in the South East of England in the December of 2020 (chapter 5, Kraemer et al., 2021): this can be attributed to individuals travelling several days early to avoid restrictions and to be able to spend Christmas day with families in other parts of the country. Further, weddings and funerals were exempt from restrictions put into place in late 2021 in the different UK nations, for example the limits on the number of people allowed to mix indoors and physical distancing in Scotland (Sturgeon, 2021), despite being high-risk events. It is important to recognise culturally significant events in disease control, regardless of which pathogen or country is being considered, and find ways to mitigate risk in these settings in a way that is acceptable to the population in question. Further, infectious disease researchers must be mindful to not use culturally - and racially - loaded language when discussing the impact of traditions of other populations, as every population has traditions and ceremonies they view as important enough to risk contracting a disease.

6.2 The future of genomic epidemiology and phylodynamics in public health

The future of genomic epidemiology and phylodynamics as part of public health and its role in understanding viral dynamics is exciting. Many new innovations have occurred out of necessity during the SARS-CoV-2 pandemic, and the size of the datasets as well as the extent of the political buy-in have opened a new door for continued development of the field. There are however several key practical and ethical issues to be overcome when discussing future development, and the challenge

remains to solve several fundamental issues, as well as to explore new avenues of data exploration.

6.2.1 Practical issues for integrating non-genomic data into genomic analyses

In a scenario where genomics is integrated into public health interventions, currently there are many different sources of data from different actors. Fig. 6.1 shows a schematic of this in terms of the progression of an individual's infection.

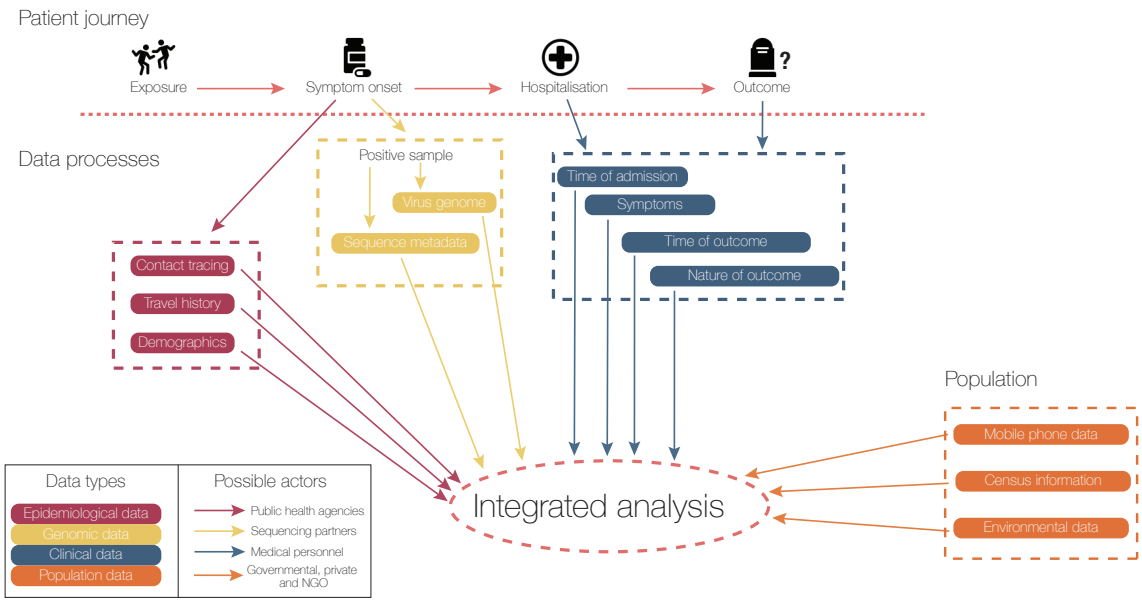

Fig. 6.1 Schematic showing different types of data and possible actors involved in data production which can be fed into an integrated analysis. As a patient moves through an infection cycle, public health agencies may collect data involving their exposure and who else they may have exposed to prevent ongoing transmission. Sequencing partners - who could be academics, within public health agencies, private companies, or a combination of the three - may sequence the positive sample and obtain metadata associated with it. The patient may be hospitalised, at which point clinical metadata can be obtained by clinical partners in the hospital.

In order to extract the most value from a genomic dataset, and to explore factors such as individual-level heterogeneity on a population level, different data types need to be integrated. Formally including human mobility data was one of the innovations of the Zika epidemic (see Introduction). In the COVID-19 pandemic, contact tracing data (Lane et al., 2021), clinical outcomes (Volz et al., 2021b), and mobile phone data (Kraemer et al., 2021 and chapter 5) have been integrated into genomic analyses to explore transmission and mortality patterns. Currently, these data sources are largely independent, and require substantial work to connect individual-level metadata to sequence data; as well as time spent navigating different data protection agreements between the separate actors. This can be remedied by having a single, consistent identifier across all of the databases. This exists to some extent in a country like the UK with a nationalised healthcare system, but is far more difficult in a location with a patchwork of private healthcare providers, like the US. A better solution is to have a single database which stores all individual-level data, and so it could all be downloaded together under a single data agreement. Clearly this would involve a large amount of coordination and cooperation between the different actors, but would drastically speed up the rate, accuracy and depth of analysis in an emergency; and is an obvious and relatively low-cost way to prepare for the next pandemic.

For population-level metrics, an important step would be a standardisation of metadata. For example, in the UK there is no coherent, hierarchical geographical structure that all of the SARS-CoV-2 data uses: the genomic dataset relies on administrative level two (admin2, which itself is somewhat unstandardised) and outer postcodes (which are a different system to the admin2 system and even cross internal borders between devolved nations); and the case data uses upper and lower tier local authorities. In Northern Ireland, there are counties and local government districts, which are on the same geographical scale but not mutually compatible, and data

is collected at both levels. Different population-level metrics would therefore be significantly simpler to combine if they all used the same underlying geographical system.

6.2.2 The generation of representative datasets

In epidemics prior to 2019, much of the sample selection strategy for genomic sequencing involved simply sequencing as much as possible in a fairly opportunistic manner. For Ebola virus in West Africa, this led to a dataset which correlated acceptably well with the case counts (Dudas et al., 2017). This is however far from guaranteed, and opportunistic sampling can mean that analyses such as phylogeography cannot be undertaken with confidence on a dataset. Part of pandemic preparation for governments that wish to integrate genomic analysis into their public health response should therefore be the consideration and development of sampling frameworks.

There are several different options for sampling strategies, and which is most appropriate depends on the sorts of analyses required, and the resources available. For example, even opportunistic sampling can be useful for tracking specific transmission chains of interest or small-scale outbreaks, and can be rapidly set-up without requiring much planning. With sufficient metadata to indicate that cases are connected, larger-scale studies can exclude specific sequences to help mitigate biases. However, random sampling, where a percentage of positive samples are selected for sequencing in a metadata-independent way, is more useful for large-scale studies and surveillance, and remains a relatively simple design. An extension to this is to select samples in a metadata-dependent way in order to take regional and temporal prevalence into account. This is useful as time and location are a proxy for

genetic diversity, and so this method aims to sample as much of the tree as possible. This is difficult to do at a local level, and requires a centralised system which can be time-consuming and challenging to organise. Building relationships between different sectors before an epidemic which can be rapidly leveraged during an emergency helps to mitigate the time-cost for this method. Finally, an alternative approach is to undertake points-of-entry sampling, where incoming travellers at land, air and sea borders are sampled. This is useful for sampling diversity in connected countries to detect future threats, and is relatively easy to set up as significant amounts of data are already collected at borders. However, it cannot be used to study the dynamics within a country. Therefore it works best in combination with some level of within-country sampling, and again with detailed metadata to ensure it is clear that these sequences are actually sampled from the diversity of the originating country of the traveller.

In a dataset that is large enough, such as for SARS-CoV-2 in the UK, sampling biases may be mitigated by downsampling to a representative set of sequences. This therefore reduces the need for extensive pre-pandemic sampling strategy design. Further, it can be hypothesis blind at the data collection point, allowing new methods and questions to be applied to the dataset long after it has been collected. It is however more resource-intensive than a well-designed sampling strategy and can result in wasted effort, as many sequences are likely to remain unused.

Finally, the question of what constitutes a representative dataset for phylodynamics is integral to the question of both sampling and downsampling strategies. Formal testing of the principles underlying this, for example using a phylogenetically-informed epidemic simulator like ABSynthE (chapter 3), would be a useful future direction for the continued production of robust and reliable phylodynamics.

6.2.3 Ethical data sharing

As genome sequencing becomes more accepted by public health actors and more predictive uses are developed, the difference between genomic surveillance, conducted by public health and governmental agencies, and genomic research, conducted by academics, will become more stark. Currently, these two fields overlap significantly, with a large proportion of sequencing conducted by smaller laboratories. The question of who should do sequencing for different purposes requires careful consideration in order to maintain the relationships between hospitals and sequencing partners that have been established during previous epidemics.

Decentralisation of genomic data generation requires a large amount of cooperation between different partners. Each individual sequence represents only a small addition to the dataset, but hundreds of sequences sampled across time-frames and regions provide valuable information for all types of analyses. This requires coordination within and between agencies and countries, and presents enormous challenges of diplomacy and governance. Within countries, setting up consortia and ensuring that data is usable at a local level can help data producers see genuine benefit from their work.

Internationally, the difficulty becomes how to equitably share data - when there is a global crisis, data must be shared globally to produce the most meaningful results, but open data sharing can result in losses for scientists from low and middle income countries (LMICs, Maxmen, 2021). Platforms such as GISAID, originally developed for influenza and now used for SARS-CoV-2 data, which require the signing of a user agreement mandating the crediting of data producers, were set up in an attempt to protect data producers globally. It has helped to encourage equitable data sharing and to mitigate the exploitation of data from LMICs by groups in other countries. The

balance between fast and easy sharing of data and protection of data producers is a difficult issue, and continued discussion is required to provide a solution that maintains data sharing without propagating or widening scientific inequalities.

6.2.4 Emerging data sources and methods

A strength of genomic epidemiology is that it allows the formal integration of different data sources (Fig. 6.1). As discussed and shown previously, this can include location or demographic data as well as cultural and climatic data. As laboratory techniques for sequencing continue to be developed, and the value of genomic epidemiology becomes more apparent to non-scientists, new sources of genomic and associated data continue to emerge.

For viruses that are detectable in faeces, sewage sampling presents a new method to obtain community samples. This is especially useful in scenarios where many infected individuals go unsampled because asymptomatic infections are common. Historically, this has been used to detect poliovirus: in Israel, environmental surveillance through wastewater detected a silent epidemic of wild poliovirus when no cases of Acute Flaccid Paralysis had been detected (Brouwer et al., 2018). More recently, sewage sampling has been used to gather additional information about SARS-CoV-2 strains circulating within the community (to find unsampled groups), and was shown to not only recover the genetic diversity in a community in California, but also indicated a greater level of variation than was present in clinical samples (Crits-Christoph et al., 2021).

In scenarios where more cases are laboratory confirmed than sequenced, additional information can be provided through monitoring the results of PCR assays. For SARS-CoV-2, the Alpha variant contains a deletion (69-70 in the Spike gene)

that results in S-gene target failure (SGTF) in a commonly used diagnostic assay. This allowed the direct analysis of the proportion of cases that tested positive for the Alpha constellation without having to sequence the whole genome, leading to the integration of a much larger and more unbiased source of data to assess whether Alpha variant had a transmission advantage in the UK (Volz et al., 2021a). This is especially important in settings that cannot do the large-scale, relatively representative sequencing that has been possible for SARS-CoV-2 in the UK, and has been applied successfully in Portugal (Borges et al., 2021). Since the SGTF discovery, assays have been developed to distinguish more reliably between the major SARS-CoV-2 variants of concern by using targets that will drop out with specific deletions known to be present in these variants (Vogels et al., 2021). The integration of accurate PCR data into genomic epidemiology increases the utility of case data for untangling genuine growth rate changes. Another additional source of data in these scenarios is occurrence data (i.e. unsequenced case data), which can be included in a birth-death framework to more accurately estimate epidemiological parameters (Featherstone et al., 2021). This is especially useful when the sequencing rate varies over time, as it can help to mitigate sampling bias.

Finally, information on travel is being increasingly included in genomic analyses. du Plessis et al. (2021) used international flight and case count data to calculate a daily "estimated importation intensity". Combining this with a phylogeographic analysis to estimate lineage importation dates revealed that the UK epidemic was driven by >1000 separate (mostly small) introductions that predominantly arrived from other European countries well connected to the UK (du Plessis et al., 2021). Furthermore, thanks to recent advances in the BEAST software package (Suchard et al., 2018), if the start and end points of individual journeys are known, they can be included in the estimation of transition rates among locations and include new

locations in the phylogeographic estimation, indirectly impacting the tree topology inference and providing more reliable estimates of the true location of ancestral nodes (Lemey et al., 2020). This may be useful in resolving internal nodes in phylogenies of viruses such as SARS-CoV-2, for which the signal in the sequence data is less strong. This method can be used even when travel origins have no sequences at all, as in an analysis of SARS-CoV-2 introductions in Rwanda: three connected countries had shared no genome sequences, but introductions from them could be inferred phylogenetically using travel histories (Butera et al., 2021).

6.3 Concluding remarks

The Ebola virus epidemic in West Africa was a warning to the world that, even after all of the papers written on what needed to be learned from the SARS and Swine flu pandemics, it was not prepared to respond quickly and efficiently to a major epidemic. It showed us that long-standing global inequalities and flaws in weak healthcare structures needed to be addressed in order to protect the global population from a fast-spreading virus. Ebola virus did not spread extensively outside of West Africa, but with healthcare workers infected in the USA (Chevalier et al., 2014) and Italy (World Health Organisation, 2015b), and an infectious individual travelling on multiple flights to reach their home in Scotland (Halliday and Carrell, 2014), it is clear that flaws in international response could have led to a larger and more widely-spread epidemic. Disappointingly, some of the same flaws, especially in terms of national and international health inequity, have been evident in the SARS-CoV-2 pandemic.

The trajectory of the use of viral genomic sequencing in public health emergencies has been slightly clearer. The West African Ebola virus epidemic can be viewed as an initial test case for real-time genomic epidemiology, and the beginning of it

being understood by public health actors to be a useful addition to disease control, as opposed to being a purely academic interest. Lessons learned from how to rapidly set up sequencing programmes, as well as simply the expertise acquired globally and the methods generated from sequencing Ebola virus, have been put to use in the sequencing of SARS-CoV-2.

To prepare for the next pandemic, there is significant work that must be done. Above all, health equity must be a global priority. This includes access to affordable healthcare and vaccination, as well as frameworks for data sharing that protect producers at the same time as enabling rapid information dissemination, and fixing inequalities in supply chains which mean that researchers in low income countries cannot access reagents with as much ease as those in high income countries. Further development of efficient sampling strategies and computational models will also help in the democratisation of genome sequencing as they allow these analyses to be less resource intensive. When the next pandemic starts, measures such as these and studies like those presented in this thesis will help us to better predict what the future holds and, hopefully, prepare for it better.

References

Abbas, Mohamed et al. (Jan. 2021). “Nosocomial transmission and outbreaks of coronavirus disease 2019: the need to protect both patients and healthcare workers”. *Antimicrob. Resist. Infect. Control* 10.1, pp. 1–13.

Aggarwal, Dinesh et al. (Feb. 2022). “Genomic epidemiology of SARS-CoV-2 in a UK university identifies dynamics of transmission”. *Nat. Commun.* 13.1, pp. 1–16.

Aleanizy, Fadilah Sfouq et al. (2017). “Outbreak of Middle East respiratory syndrome coronavirus in Saudi Arabia: a retrospective study”. *BMC Infect. Dis.* 17.

Andersen, Kristian G et al. (Mar. 2020). “The proximal origin of SARS-CoV-2”. *Nat. Med.* 26.4, pp. 450–452.

Andreano, Emanuele et al. (Dec. 2020). “SARS-CoV-2 escape in vitro from a highly neutralizing COVID-19 convalescent plasma”.

Angus, Colin (Jan. 2022). *CoVid Plots and Analysis*.

Arcia, David et al. (2017). “Role of CD8+ T Cells in the Selection of HIV-1 Immune Escape Mutations”. *Viral Immunol.* 30.1, pp. 3–12.

Arias, Armando et al. (Jan. 2016). “Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases”. *Virus Evolution* 2.1, vew016.

Arwady, M Allison et al. (Apr. 2015). “Evolution of Ebola Virus Disease from Exotic Infection to Global Health Priority, Liberia, Mid-2014”. *Emerg. Infect. Dis.* 21.4, p. 578.

Asadi, S et al. (Feb. 2019). “Aerosol emission and superemission during human speech increase with voice loudness”. *Sci. Rep.* 9.1.

Avanzato, Victoria A et al. (Dec. 2020). “Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer”. *Cell* 183.7, p. 1901.

Bailey, Trevor C and Gatrell, Anthony C (1995). *Interactive Spatial Data Analysis*. Longman Scientific & Technical.

Baize, Sylvain et al. (Oct. 2014). “Emergence of Zaire Ebola virus disease in Guinea”. *N. Engl. J. Med.* 371.15, pp. 1418–1425.

Bell, A et al. (May 2015). “Genome sequence analysis of Ebola virus in clinical samples from three British healthcare workers, August 2014 to March 2015”. *Euro Surveill.* 20.20.

Benvenuto, Domenico et al. (July 2020). “Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy”. *J. Infect.* 81.1, e24–e27.

Bobashev, Georgiy V et al. (2007). “A Hybrid Epidemic Model: Combining The Advantages Of Agent-Based And Equation-Based Approaches”. *2007 Winter Simulation Conference*, pp. 1532–1537.

Boeras, Debrah I et al. (Nov. 2011). “Role of donor genital tract HIV-1 diversity in the transmission bottleneck”. *Proc. Natl. Acad. Sci. U. S. A.* 108.46, E1156.

Borges, Vıtor et al. (Mar. 2021). “Tracking SARS-CoV-2 lineage B.1.1.7 dissemination: insights from nationwide spike gene target failure (SGTF) and spike gene late detection (SGTL) data, Portugal, week 49 2020 to week 3 2021”. *Eurosurveillance* 26.10.

Boseley, Sarah (Apr. 2010). “Government cancels swine flu vaccine order”. *The Guardian*.

Britton, Tom, Ball, Frank, and Trapman, Pieter (2020). “A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2”. *Science* 369.6505, pp. 846–849.

Brouwer, Andrew F et al. (Nov. 2018). “Epidemiology of the silent polio outbreak in Rahat, Israel, based on modeling of environmental surveillance data”. *Proc. Natl. Acad. Sci. U. S. A.* 115.45, E10625–E10633.

Bruch, Elizabeth and Atwell, Jon (2015). “Agent-Based Models in Empirical Social Research”. *Sociol. Methods Res.* 44.2, pp. 186–221.

Butera, Yvan et al. (Sept. 2021). “Genomic sequencing of SARS-CoV-2 in Rwanda reveals the importance of incoming travelers on lineage diversity”. *Nat. Commun.* 12.1, pp. 1–11.

Cabinet Office (Feb. 2021). *COVID-19 Response - Spring 2021 (Summary)*. https://www.gov.uk/government/publications/covid-19-response-spring-2021/covid-19-response-spring-2021-summary. Accessed: 2021-6-24.

Carroll, Miles W et al. (Aug. 2015). “Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa”. *Nature* 524.7563, pp. 97–101.

Casey-Bryars, Miriam et al. (June 2021). “Presymptomatic transmission of SARS-CoV-2 infection: a secondary analysis using published data”. *BMJ Open* 11.6, e041240.

Centers for Disease Control and Prevention (Mar. 2003). “Update: Outbreak of Severe Acute Respiratory Syndrome — Worldwide, 2003”. *MMWR Morbidity Mortality Weekly Report*.

Central Intelligence Agency (2022). *Sierra Leone*. https://www.cia.gov/the-world-factbook/countries/sierra-leone/#economy. Accessed: 2022-3-18.

Challen, Robert et al. (June 2021). “Early epidemiological signatures of novel SARS-CoV-2 variants: establishment of B.1.617.2 in England”.

Chandler, Jeffrey C et al. (Nov. 2021). “SARS-CoV-2 exposure in wild white-tailed deer (Odocoileus virginianus)”. *Proc. Natl. Acad. Sci. U. S. A.* 118.47.

Chen, Paul Z et al. (Sept. 2021). “Understanding why superspreading drives the COVID-19 pandemic but not the H1N1 pandemic”. *Lancet Infect. Dis.* 21.9, pp. 1203–1204.

Cherry, James D and Krogstad, Paul (July 2004). “SARS: The First Pandemic of the 21st Century”. *Pediatr. Res.* 56.1, pp. 1–5.

Chevalier, Michelle S et al. (Nov. 2014). “Ebola Virus Disease Cluster in the United States — Dallas County, Texas, 2014”. *MMWR Surveill. Summ.* 63.46, p. 1087.

Chinese SARS Molecular Epidemiology Consortium (Mar. 2004). “Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China”. *Science* 303.5664, pp. 1666–1669.

Choi, Bina et al. (Dec. 2020). “Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host”. *N. Engl. J. Med.* 383.23, pp. 2291–2293.

Clark, Sarah A et al. (May 2021). “SARS-CoV-2 evolution in an immunocompromised host reveals shared neutralization escape mechanisms”. *Cell* 184.10, 2605–2617.e18.

Collier, Dami A et al. (May 2021). “Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies”. *Nature* 593.7857, pp. 136–141.

Collignon, Peter (2011). “Swine flu: lessons we need to learn from our global experience”. *Emerg. Health Threats J.* 4.

Cotten, Matthew et al. (2013). “Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study”. *Lancet* 382.9909, p. 1993.

Crits-Christoph, Alexander et al. (Feb. 2021). “Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants”. *MBio* 12.1.

Crooks, Andrew T and Hailegiorgis, Atesmachew B (2014). “An agent-based modeling approach applied to the spread of cholera”. *Environmental Modelling & Software* 62, pp. 164–177.

Davies, Nicholas G et al. (July 2020). “Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study”. *The Lancet Public Health* 5.7, e375–e385.

Davies, Nicholas G et al. (Apr. 2021a). “Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England”. *Science* 372.6538, eabg3055.

Davies, Nicholas G et al. (Mar. 2021b). “Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7”. *Nature* 593.7858, pp. 270–274.

Dawood, Fatimah S et al. (2012). “Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study”. *Lancet Infect. Dis.* 12.9, pp. 687–695.

Dellicour, Simon et al. (June 2018). “Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak”. *Nat. Commun.* 9.1, pp. 1–9.

Dellicour, Simon et al. (Nov. 2020). “Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework”. *Nat. Commun.* 11.1, pp. 1–11.

Department of Health and Social Care (2021). *Community testing: a guide for local delivery*. https://www.gov.uk/government/publications/community-testing-explainer/community-testing-a-guide-for-local-delivery. Accessed: 2022-3-21.

Diallo, Boubacar et al. (Nov. 2016). “Resurgence of Ebola Virus Disease in Guinea Linked to a Survivor With Virus Persistence in Seminal Fluid for More Than 500 Days”. *Clin. Infect. Dis.* 63.10, pp. 1353–1356.

Dong, Ensheng, Du, Hongru, and Gardner, Lauren (May 2020). “An interactive web-based dashboard to track COVID-19 in real time”. *Lancet Infect. Dis.* 20.5, pp. 533–534.

Dowell, S F et al. (Feb. 1999). “Transmission of Ebola hemorrhagic fever: a study of risk factors in family members, Kikwit, Democratic Republic of the Congo, 1995. Commission de Lutte contre les Epidémies à Kikwit”. *J. Infect. Dis.* 179 Suppl 1, S87–91.

Doyle, Timothy J, Glynn, M Kathleen, and Groseclose, Samuel L (May 2002). “Completeness of Notifiable Infectious Disease Reporting in the United States: An Analytical Literature Review”. *Am. J. Epidemiol.* 155.9, pp. 866–874.

Draenert, Rika et al. (2004). *Immune Selection for Altered Antigen Processing Leads to Cytotoxic T Lymphocyte Escape in Chronic HIV-1 Infection*.

Drummond, Alexei et al. (Sept. 2003). “Measurably evolving populations”. *Trends Ecol. Evol.* 18.9, pp. 481–488.

Drummond, Alexei J and Rambaut, Andrew (Nov. 2007). “BEAST: Bayesian evolutionary analysis by sampling trees”. *BMC Evol. Biol.* 7.1, pp. 1–8.

du Plessis, Louis et al. (Feb. 2021). “Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK”. *Science* 371.6530, pp. 708–712.

Duchene, Sebastian et al. (May 2020). “Temporal signal and the phylodynamic threshold of SARS-CoV-2”.

Dudas, Gytis et al. (Apr. 2017). “Virus genomes reveal factors that spread and sustained the Ebola epidemic”. *Nature* 544.7650, pp. 309–315.

Dudas, Gytis et al. (Jan. 2018). “MERS-CoV spillover at the camel-human interface”.

Dudas, Gytis et al. (Oct. 2021). “Emergence and spread of SARS-CoV-2 lineage B.1.620 with variant of concern-like mutations and deletions”. *Nat. Commun.* 12.1, pp. 1–12.

Eckstrand, Chrissy D et al. (Nov. 2021). “An outbreak of SARS-CoV-2 with high mortality in mink (Neovison vison) on multiple Utah farms”. *PLoS Pathog.* 17.11, e1009952.

Elliott, Paul et al. (Nov. 2021). “Exponential growth, high prevalence of SARS-CoV-2, and vaccine effectiveness associated with the Delta variant”. *Science*, eabl9551.

Eyre, David W et al. (Sept. 2021). “The impact of SARS-CoV-2 vaccination on Alpha & Delta variant transmission”.

Fairhead, James (2014). “The significance of death, funerals and the after-life in Ebola-hit Sierra Leone, Guinea and Liberia: Anthropological insights into infection and social resistance”.

Famulare, Michael et al. (Jan. 2016). “Sabin Vaccine Reversion in the Field: a Comprehensive Analysis of Sabin-Like Poliovirus Isolates in Nigeria”. *J. Virol.* 90.1, pp. 317–331.

Faria, N R et al. (May 2017). “Establishment and cryptic transmission of Zika virus in Brazil and the Americas”. *Nature* 546.7658, pp. 406–410.

Faria, N R et al. (May 2021). “Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil”. *Science* 372.6544, pp. 815–821.

Faria, Nuno et al. (Oct. 2014). “The early spread and epidemic ignition of HIV-1 in human populations”. *Science* 346.6205, p. 56.

Faria, Nuno Rodrigues et al. (Apr. 2016). “Zika virus in the Americas: Early epidemiological and genetic findings”. *Science* 352.6283, p. 345.

Featherstone, Leo A et al. (Aug. 2021). “Infectious disease phylodynamics with occurrence data”. *Methods Ecol. Evol.* 12.8, pp. 1498–1507.

Ferguson, Neil (2021). *B.1.617.2 transmission in England: risk factors and transmission advantage*. Tech. rep.

Formenty, Pierre et al. (Feb. 1999). “Human Infection Due to Ebola Virus, Subtype Côte d’Ivoire: Clinical and Biologic Presentation”. *J. Infect. Dis.* 179.Supplement_1, S48–S53.

Francis, Rodric V et al. (Sept. 2021). “The Impact of Real-Time Whole-Genome Sequencing in Controlling Healthcare-Associated SARS-CoV-2 Outbreaks”. *J. Infect. Dis.* 225.1, pp. 10–18.

Furuyama, Taima N et al. (June 2020). “Temporal data series of COVID-19 epidemics in the USA, Asia and Europe suggests a selective sweep of SARS-CoV-2 Spike D614G variant”. eprint: 2006.11609.

Galassi, Francesco M, Habicht, Michael E, and Rühli, Frank J (Sept. 2016). “Poliomyelitis in Ancient Egypt?” *Neurol. Sci.* 38.2, pp. 375–375.

Gamage, Akshamal M et al. (Dec. 2020). “Infection of human Nasal Epithelial Cells with SARS-CoV-2 and a 382-nt deletion isolate lacking ORF8 reveals similar viral kinetics and host transcriptional profiles”. *PLoS Pathog.* 16.12, e1009130.

Geoghegan, Jemma L et al. (Dec. 2020). “Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand”. *Nat. Commun.* 11.1, pp. 1–7.

Ghafari, Mahan, Kadivar, Alireza, and Katzourakis, Aris (June 2021). “Excess deaths associated with the Iranian COVID-19 epidemic: A province-level analysis”. *Int. J. Infect. Dis.* 107, pp. 101–115.

Gibbons, Cheryl L et al. (Feb. 2014). “Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods”. *BMC Public Health* 14.1, pp. 1–17.

Gill, Mandev S et al. (Nov. 2012). “Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci”. *Mol. Biol. Evol.* 30.3, pp. 713–724.

Gill, Mandev S et al. (2016). “Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates”. *Syst. Biol.* 65.6, pp. 1041–1056.

Gire, Stephen K et al. (Sept. 2014). “Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak”. *Science* 345.6202, p. 1369.

Glynn, Judith R et al. (Jan. 2018). “Variability in Intrahousehold Transmission of Ebola Virus, and Estimation of the Household Secondary Attack Rate”. *J. Infect. Dis.* 217.2, pp. 232–237.

Goba, Augustine et al. (Oct. 2016). “An Outbreak of Ebola Virus Disease in the Lassa Fever Zone”. *J. Infect. Dis.* 214.suppl 3, S110–S121.

Gostic, Katelyn M et al. (Nov. 2016). “Potent protection against H5N1 and H7N9 influenza via childhood hemagglutinin imprinting”. *Science* 354.6313, pp. 722–726.

Greater London Authority Intelligence and Analysis Unit (2014). *“Census Information Scheme: Commuting in London”*. Tech. rep.

Grenfell, B T et al. (Jan. 2004). “Unifying the epidemiological and evolutionary dynamics of pathogens”. *Science* 303.5656, pp. 327–332.

Grubaugh, Nathan D et al. (2019). “Travel Surveillance and Genomics Uncover a Hidden Zika Outbreak during the Waning Epidemic”. *Cell* 178.5, 1057–1071.e11.

Gryseels, Sophie et al. (2020). “A near full-length HIV-1 genome from 1966 recovered from formalin-fixed paraffin-embedded tissue”. *Proceedings of the National Academy of Sciences* 117.22, pp. 12222–12229.

Gu, Hongjing et al. (Sept. 2020). “Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy”. *Science* 369.6511, pp. 1603–1607.

Guan, Y et al. (Jan. 2004). “Molecular epidemiology of the novel coronavirus that causes severe acute respiratory syndrome”. *Lancet* 363.9403, p. 99.

Hagberg, Aric, Swart, Pieter, and S Chult, Daniel (Jan. 2008). *Exploring network structure, dynamics, and function using networkx*. Tech. rep. LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Lab. (LANL), Los Alamos, NM (United States).

Halliday, Josh and Carrell, Severin (Dec. 2014). “Ebola screening to be reviewed after doctor attacks ‘inadequate’ measures”. *The Guardian*.

Hartfield, Matthew and Alizon, Samuel (June 2013). “Introducing the Outbreak Threshold in Epidemiology”. *PLoS Pathog.* 9.6, e1003277.

Hasegawa, Masami, Kishino, Hirohisa, and Yano, Taka-Aki (Oct. 1985). “Dating of the human-ape splitting by a molecular clock of mitochondrial DNA”. *J. Mol. Evol.* 22.2, pp. 160–174.

Hill, Verity and Baele, Guy (July 2019). “Bayesian Estimation of Past Population Dynamics in BEAST 1.10 Using the Skygrid Coalescent Model”. *Mol. Biol. Evol.* 36.11, pp. 2620–2628.

Hill, Verity et al. (Dec. 2021). “Progress and challenges in virus genomic epidemiology”. *Trends Parasitol.* 37.12, pp. 1038–1049.

Hill, Verity et al. (Mar. 2022). “The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK”.

Hodcroft, Emma B et al. (July 2021). “Spread of a SARS-CoV-2 variant through Europe in the summer of 2020”. *Nature* 595.7869, pp. 707–712.

Hoenen, Thomas et al. (Feb. 2016). “Nanopore sequencing as a rapidly deployable Ebola outbreak tool”. *Emerg. Infect. Dis.* 22.2, pp. 331–334.

Hoffmann, Markus, Kleine-Weber, Hannah, and Pöhlmann, Stefan (May 2020). “A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells”. *Mol. Cell* 78.4, 779–784.e5.

Howard, Jeremy et al. (2021). “An evidence review of face masks against COVID-19”. *Proceedings of the National Academy of Sciences* 118.4, e2014564118.

Iglesias, Maria Candela et al. (Aug. 2011). “Escape from highly effective public CD8+ T-cell clonotypes by HIV”. *Blood* 118.8, pp. 2138–2149.

Jackson, Ben et al. (Aug. 2021). “Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic”. *Cell*.

Jacobs, Michael et al. (July 2016). “Late Ebola virus relapse causing meningoencephalitis: a case report”. *Lancet* 388.10043, pp. 498–503.

Jansen van Vuren, Petrus et al. (Jan. 2019). “Phylodynamic Analysis of Ebola Virus Disease Transmission in Sierra Leone”. *Viruses* 11.1.

Jarvis, Christopher I et al. (Oct. 2020). *CoMix study - Social contact survey in the UK*. https://cmmid.github.io/topics/covid19/comix-reports.html. Accessed: 2021-11-10.

Judson, Seth, Prescott, Joseph, and Munster, Vincent (Feb. 2015). “Understanding Ebola Virus Transmission”. *Viruses* 7.2, p. 511.

Kalkauskas, Antanas et al. (Jan. 2021). “Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk”. *PLoS Comput. Biol.* 17.1, e1008561.

Kannan, S R et al. (Nov. 2021). “Evolutionary analysis of the Delta and Delta Plus variants of the SARS-CoV-2 viruses”. *J. Autoimmun.* 124.

Karim, F et al. (June 2021). “Persistent SARS-CoV-2 infection and intra-host evolution in association with advanced HIV infection”. *medRxiv*, p. 2021.06.03.21258228.

Keele, Brandon F et al. (July 2006). “Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1”. *Science* 313.5786, p. 523.

Kemp, Steven A et al. (Feb. 2021). “SARS-CoV-2 evolution during treatment of chronic infection”. *Nature* 592.7853, pp. 277–282.

Khan, Adnan et al. (2015). “Estimating the basic reproductive ratio for the Ebola outbreak in Liberia and Sierra Leone”. *Infectious Diseases of Poverty* 4.1.

Kinganda-Lusamaki, Eddy et al. (Apr. 2021). “Integration of genomic sequencing into the response to the Ebola virus outbreak in Nord Kivu, Democratic Republic of the Congo”. *Nat. Med.* 27.4, pp. 710–716.

Kingman, J F C (1982). “The coalescent”. *Stochastic Process. Appl.* 13.3, pp. 235–248.

Klinger, Emmanuel, Rickert, Dennis, and Hasenauer, Jan (Oct. 2018). “pyABC: distributed, likelihood-free inference”. *Bioinformatics* 34.20, pp. 3591–3593.

Konings, Frank et al. (June 2021). “SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse”. *Nature Microbiology* 6.7, pp. 821–823.

Korber, B et al. (July 2020). “Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus”. *Cell*.

Kraemer, M U G et al. (Nov. 2018). “Reconstruction and prediction of viral disease epidemics”. *Epidemiol. Infect.*, pp. 1–7.

Kraemer, Moritz U G et al. (Aug. 2021). “Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence”. *Science* 373.6557, pp. 889–895.

Kucharski, Adam J et al. (Nov. 2015). “Measuring the impact of Ebola control measures in Sierra Leone”. *Proc. Natl. Acad. Sci. U. S. A.* 112.46, pp. 14366–14371.

Kupferschmidt, Kai and Wadman, Meredith (June 2021). *Delta variant triggers dangerous new phase in the pandemic*. https://www.sciencemag.org/news/2021/06/delta-variant-triggers-dangerous-new-phase-pandemic. Accessed: 2021-6-24.

Ladner, Jason T et al. (Dec. 2015). “Evolution and spread of Ebola virus in Liberia, 2014–2015”. *Cell Host Microbe* 18.6, p. 659.

Ladner, Jason T et al. (Feb. 2019). “Precision epidemiology for infectious disease control”. *Nat. Med.* 25.2, pp. 206–211.

Lai, Alessia et al. (July 2020). “Molecular Tracing of SARS-CoV-2 in Italy in the First Three Months of the Epidemic”. *Viruses* 12.8.

Lakdawala, Seema S et al. (Sept. 2015). “The soft palate is an important site of adaptation for transmissible influenza viruses”. *Nature* 526.7571, pp. 122–125.

Lane, Courtney R et al. (Aug. 2021). “Genomics-informed responses in the elimination of COVID-19 in Victoria, Australia: an observational, genomic epidemiological study”. *The Lancet Public Health* 6.8, e547–e556.

Lau, Max S Y et al. (Feb. 2017). “Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic”. *Proceedings of the National Academy of Sciences* 114.9, pp. 2337–2342.

Lauring, Adam S (Sept. 2020). “Within-Host Viral Diversity: A Window into Viral Evolution”.

Lee, Elizabeth C et al. (Oct. 2020). “The engines of SARS-CoV-2 spread”. *Science* 370.6515, pp. 406–407.

Lemey, Philippe et al. (Sept. 2009). “Bayesian Phylogeography Finds Its Roots”. *PLoS Comput. Biol.* 5.9, e1000520.

Lemey, Philippe et al. (Aug. 2010). “Phylogeography Takes a Relaxed Random Walk in Continuous Space and Time”. *Mol. Biol. Evol.* 27.8, p. 1877.

Lemey, Philippe et al. (Feb. 2014). “Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2”. *PLoS Pathog.* 10.2, e1003932.

Lemey, Philippe et al. (Oct. 2020). “Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2”. *Nat. Commun.* 11.1, pp. 1–14.

Lemey, Philippe et al. (June 2021). “Untangling introductions and persistence in COVID-19 resurgence in Europe”. *Nature* 595.7869, pp. 713–717.

Leslie, A J et al. (Mar. 2004). “HIV evolution: CTL escape mutation and reversion after transmission”. *Nat. Med.* 10.3.

Lewis, Dyani (Feb. 2021). “Superspreading drives the COVID pandemic — and could help to tame it”. *Nature* 590.7847, pp. 544–546.

Li, Heng (Sept. 2018). “Minimap2: pairwise alignment for nucleotide sequences”. *Bioinformatics* 34.18, pp. 3094–3100.

Li, Tao et al. (Aug. 2017). “Mapping the clinical outcomes and genetic evolution of Ebola virus in Sierra Leone”. *JCI Insight* 2.15.

Lloyd-Smith, J O et al. (Nov. 2005). “Superspreading and the effect of individual variation on disease emergence”. *Nature* 438.7066.

Logistics Cluster (n.d.). *Sierra Leone Road Network*. https://dlca.logcluster.org/plugins/viewsource/viewpagesrc.action?pageId=5702218. Accessed: 2022-3-5.

Lopez Bernal, Jamie et al. (Aug. 2021). “Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant”. *N. Engl. J. Med.* 385.7, pp. 585–594.

Lu, Jing et al. (May 2020). “Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China”. *Cell* 181.5, 997–1003.e9.

Lu, Lu et al. (Nov. 2021). “Adaptation, spread and transmission of SARS-CoV-2 in farmed minks and associated humans in the Netherlands”. *Nat. Commun.* 12.1, pp. 1–12.

Lucas, Carolina et al. (Oct. 2021). “Impact of circulating SARS-CoV-2 variants on mRNA vaccine-induced immunity”. *Nature*.

Lythgoe, Katrina A et al. (May 2017). “Short-Sighted Virus Evolution and a Germline Hypothesis for Chronic Viral Infections”. *Trends Microbiol.* 25.5, pp. 336–348.

Lythgoe, Katrina A et al. (Apr. 2021). “SARS-CoV-2 within-host diversity and transmission”. *Science* 372.6539.

Ma, Qiuyue et al. (Dec. 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 Netw Open* 4.12, e2137257–e2137257.

MacNeil, A et al. (Nov. 2011). “Filovirus outbreak detection and surveillance: lessons from Bundibugyo”. *J. Infect. Dis.* 204 Suppl 3.

Majra, Dasha et al. (Jan. 2021). “SARS-CoV-2 (COVID-19) superspreader events”. *J. Infect.* 82.1, p. 36.

Maponga, Tongai G et al. (Jan. 2022). “Persistent SARS-CoV-2 Infection with Accumulation of Mutations in a Patient with Poorly Controlled HIV Infection”.

Marı Saéz, Almudena et al. (2015). “Investigating the zoonotic origin of the West African Ebola epidemic”. *EMBO Mol. Med.* 7.1, pp. 17–23.

Mate, Suzanne E et al. (Dec. 2015). “Molecular Evidence of Sexual Transmission of Ebola Virus”. *N. Engl. J. Med.* 373.25, pp. 2448–2454.

Matfess, Hilary (2018). *Layered Insecurity in North Kivu: Violence and the Ebola Response - Democratic Republic of the Congo*. Tech. rep.

Mathieu, Edouard et al. (May 2021). “A global database of COVID-19 vaccinations”. *Nature Human Behaviour* 5.7, pp. 947–953.

Maxmen, Amy (June 2020). “World’s second-deadliest Ebola outbreak ends in Democratic Republic of the Congo”. *Nature*.

— (May 2021). “Why some researchers oppose unrestricted sharing of coronavirus genome data”. *Nature* 593.7858, pp. 176–177.

Mbala-Kingebeni, Placide et al. (Apr. 2021). “Ebola Virus Transmission Initiated by Systemic Ebola Virus Disease Relapse”. *N. Engl. J. Med.* 384.13, p. 1240.

McCarthy, Kevin R et al. (Mar. 2021). “Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape”. *Science* 371.6534, pp. 1139–1142.

McCrone, J T (2021). *Approaches for analyzing large phylogenetic datasets*. https://beast.community/thorney_beast. Accessed: 2022-3-22.

McCrone, J T and Lauring, Adam (Feb. 2018). “Genetic bottlenecks in intraspecies virus transmission”. *Curr. Opin. Virol.* 28, pp. 20–25.

McCrone, John T et al. (Dec. 2021). “Context-specific emergence and growth of the SARS-CoV-2 Delta variant”. *medRxiv*, p. 2021.12.14.21267606.

Mena, Gonzalo E et al. (Apr. 2021). “Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile”. *Science*.

Mena, Ignacio et al. (June 2016). “Origins of the 2009 H1N1 influenza pandemic in swine in Mexico”.

Meng, Bo et al. (June 2021). “Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7”. *Cell Rep.* 35.13, p. 109292.

Merler, Stefano et al. (Feb. 2015). “Spatio-temporal spread of the Ebola 2014 outbreak in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis”. *Lancet Infect. Dis.* 15.2, p. 204.

Minh, Bui Quang et al. (Feb. 2020). “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era”. *Mol. Biol. Evol.* 37.5, pp. 1530–1534.

Monod, Mélodie et al. (2021). “Age groups that sustain resurging COVID-19 epidemics in the United States”. *Science*.

Mullen, Lucia et al. (2020). “An analysis of International Health Regulations Emergency Committees and Public Health Emergency of International Concern Designations”. *BMJ Global Health* 5.6.

Müller, Marcel A et al. (2014). “MERS Coronavirus Neutralizing Antibodies in Camels, Eastern Africa, 1983–1997”. *Emerging Infectious Diseases* 20.12.

Muyembe-Tamfum, J J et al. (Feb. 1999). “Ebola Outbreak in Kikwit, Democratic Republic of the Congo: Discovery and Control Measures”. *J. Infect. Dis.* 179.Supplement_1, S259–S262.

Nguyen, Long H et al. (Sept. 2020). “Risk of COVID-19 among front-line health-care workers and the general community: a prospective cohort study”. *The Lancet Public Health* 5.9, e475–e483.

Ntumba, Harry César Kayembe et al. (Dec. 2019). “Ebola response and community engagement: how to build a bridge?” *Lancet* 394.10216, p. 2242.

Nyenswah, Tolbert G et al. (Feb. 2016). “Ebola and Its Control in Liberia, 2014-2015”. *Emerg. Infect. Dis.* 22.2, pp. 169–177.

O’Dowd, Adrian (Oct. 2014). “Government says it would stockpile Tamiflu again”. *BMJ* 349.

O’Driscoll, Megan et al. (Nov. 2020). “Age-specific mortality and immunity patterns of SARS-CoV-2”. *Nature* 590.7844, pp. 140–145.

O’Toole, Áine et al. (July 2021a). “Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool”. *Virus Evol* 7.2.

O’Toole, Áine et al. (Dec. 2021b). “Genomics-informed outbreak investigations of SARS-CoV-2 using civet”. *medRxiv*, p. 2021.12.13.21267267.

O'Toole, Áine et al. (May 2021c). "Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch". *Wellcome Open Research* 6.121, p. 121.

Oberoi, Simmi et al. (2016). "Understanding health seeking behavior". *Journal of Family Medicine and Primary Care* 5.2, p. 463.

Oliver, David (Feb. 2021). "David Oliver: Lack of PPE betrays NHS clinical staff". *BMJ* 372.

Oreshkova, Nadia et al. (June 2020). "SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020". *Eurosurveillance* 25.23.

Ottersen, Prof Ole Petter et al. (Feb. 2014). "The political origins of health inequity: prospects for change". *Lancet* 383.9917, pp. 630–667.

Otu, Akaninyene et al. (July 2017). "An account of the Ebola virus disease outbreak in Nigeria: implications and lessons learnt". *BMC Public Health* 18.1, pp. 1–8.

Ou, C Y et al. (May 1992). "Molecular epidemiology of HIV transmission in a dental practice". *Science* 256.5060, pp. 1165–1171.

Oude Munnink, Bas B et al. (Jan. 2021). "Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans". *Science* 371.6525, p. 172.

Ozono, Seiya et al. (2021). "SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity". *Nat. Commun.* 12.

Park, Daniel J et al. (June 2015). "Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone". *Cell* 161.7, p. 1516.

Peacock, Thomas P et al. (Apr. 2021). "The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets". *Nature Microbiology* 6.7, pp. 899–909.

Peacock, Thomas P et al. (Jan. 2022). "The SARS-CoV-2 variant, Omicron, shows rapid replication in human primary nasal epithelial cultures and efficiently uses the endosomal route of entry".

Pellecchia, Umberto et al. (Dec. 2015). "Social Consequences of Ebola Containment Measures in Liberia". *PLoS One* 10.12, e0143036.

Petersen, Eskild et al. (Sept. 2020). “Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics”. *Lancet Infect. Dis.* 20.9, e238–e244.

Picchio, Gastón R et al. (Mar. 1998). “The Cell Tropism of Human Immunodeficiency Virus Type 1 Determines the Kinetics of Plasma Viremia in SCID Mice Reconstituted with Human Peripheral Blood Leukocytes”. *J. Virol.* 72.3, p. 2002.

Plummer, Martyn et al. (Mar. 2006). “CODA: convergence diagnosis and output analysis for MCMC”. *R News* 6.1, pp. 7–11.

Pope, Addy (Feb. 2017). *GB Postcode Area, Sector, District*.

Prescott, Joseph et al. (May 2015). “Postmortem Stability of Ebola Virus”. *Emerg. Infect. Dis.* 21.5, p. 856.

Price, Morgan N, Dehal, Paramvir S, and Arkin, Adam P (Mar. 2010). “FastTree 2–approximately maximum-likelihood trees for large alignments”. *PLoS One* 5.3, e9490.

Public Health England (June 2020a). *COVID-19: review of disparities in risks and outcomes*. https://www.gov.uk/government/publications/covid-19-review-of-disparities-in-risks-and-outcomes. Accessed: 2021-3-16.

— (Dec. 2020b). *Investigation of SARS-CoV-2 variants of concern: technical briefings*. https://www.gov.uk/government/publications/investigation-of-novel-sars-cov-2-variant-variant-of-concern-20201201. Accessed: 2021-6-24.

— (2020c). “Public Health England Investigation of novel SARS-COV-2 variant 202012/01: technical briefing 1”.

Pybus, O G, Rambaut, A, and Harvey, P H (July 2000). “An integrated framework for the inference of viral population history from reconstructed genealogies”. *Genetics* 155.3, p. 1429.

Quick, Joshua et al. (Feb. 2016). “Real-time, portable genome sequencing for Ebola surveillance”. *Nature* 530.7589, pp. 228–232.

Rakowski, Franciszek et al. (2010). “Influenza epidemic spread simulation for Poland — a large scale, individual based model study”. *Physica A: Statistical Mechanics and its Applications* 389.16, pp. 3149–3165.

Rambaut, A et al. (Sept. 2018). “Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7”. *Syst. Biol.* 67.5, pp. 901–904.

Rambaut, Andrew et al. (July 2020a). “A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology”. *Nature Microbiology* 5.11, pp. 1403–1407.

Rambaut, Andrew et al. (Dec. 2020b). *Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations*. Tech. rep.

Ramırez, Juan David et al. (Mar. 2021). “Phylogenomic Evidence of Reinfection and Persistence of SARS-CoV-2: First Report from Colombia”. *Vaccines* 9.3, p. 282.

Ratmann, Oliver et al. (Mar. 2019). “Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis”. *Nat. Commun.* 10.1, pp. 1–13.

Ratner, Lee et al. (Jan. 1985). “Complete nucleotide sequence of the AIDS virus, HTLV-III”. *Nature* 313.6000, pp. 277–284.

Raven, Joanna, Wurie, Haja, and Witter, Sophie (2018). “Health workers’ experiences of coping with the Ebola epidemic in Sierra Leone’s health system: a qualitative study”. *BMC Health Serv. Res.* 18.

Read, Jonathan M et al. (Sept. 2021). “Hospital-acquired SARS-CoV-2 infection in the UK’s first COVID-19 pandemic wave”. *Lancet* 398.10305, pp. 1037–1038.

Reusken, Chantal B E et al. (2015). “Occupational Exposure to Dromedaries and Risk for MERS-CoV Infection, Qatar, 2013–2014”. *Emerging Infectious Diseases* 21.8.

Richard, Mathilde et al. (Feb. 2020). “Influenza A viruses are transmitted via the air from the nasal respiratory epithelium of ferrets”. *Nat. Commun.* 11.1, pp. 1–11.

Richards, Paul (Sept. 2016). *Ebola: How a People’s Science Helped End an Epidemic*. Zed Books Ltd.

Richards, Paul et al. (Apr. 2015). “Social Pathways for Ebola Virus Disease in Rural Sierra Leone, and Some Implications for Containment”. *PLoS Negl. Trop. Dis.* 9.4, e0003567.

Richards, Paul et al. (Nov. 2020). “Re-analysing Ebola spread in Sierra Leone: The importance of local social dynamics”. *PLoS One* 15.11, e0234823.

Rothe, Camilla et al. (Mar. 2020). “Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany”. *N. Engl. J. Med.* 382.10, pp. 970–971.

Ryeng, Helene Sandbu (Feb. 2015). *Ebola in Liberia: from secret burials to safe burials*. https://blogs.unicef.org/blog/ebola-in-liberia-from-secret-burials-to-safe-burials/. Accessed: 2022-3-18.

Sagulenko, Pavel, Puller, Vadim, and Neher, Richard A (Jan. 2018). “TreeTime: Maximum-likelihood phylodynamic analysis”. *Virus Evol* 4.1.

Saulnier, Emma, Gascuel, Olivier, and Alizon, Samuel (Mar. 2017). “Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study”. *PLoS Comput. Biol.* 13.3, e1005416.

Senga, Mikiko et al. (Aug. 2016). “Factors Underlying Ebola Virus Infection Among Health Workers, Kenema, Sierra Leone, 2014–2015”. *Clin. Infect. Dis.* 63.4, p. 454.

Shantha, Jessica G, Crozier, Ian, and Yeh, Steven (Nov. 2017). “An update on ocular complications of Ebola virus disease”. *Curr. Opin. Ophthalmol.* 28.6, p. 600.

Sharp, Paul M and Hahn, Beatrice H (Sept. 2011). “Origins of HIV and the AIDS Pandemic”. *Cold Spring Harb. Perspect. Med.* 1.1.

Sheikh, Aziz et al. (June 2021). “SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness”. *Lancet* 397.10293, pp. 2461–2462.

Simon-Loriere, Etienne et al. (Aug. 2015). “Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic”. *Nature* 524.7563, pp. 102–104.

Smith, Gavin J D et al. (June 2009). “Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic”. *Nature* 459.7250, pp. 1122–1125.

Smits, Saskia L et al. (2015). *Genotypic anomaly in Ebola virus strains circulating in Magazine Wharf area, Freetown, Sierra Leone, 2015*.

Sofonea, Mircea T et al. (Jan. 2022). “From Delta to Omicron: analysing the SARS-CoV-2 epidemic in France using variant-specific screening tests (September 1 to December 18, 2021)”. *medRxiv*, p. 2021.12.31.21268583.

Sonabend, Raphael et al. (Nov. 2021). “Non-pharmaceutical interventions, vaccination, and the SARS-CoV-2 delta variant in England: a mathematical modelling study”. *Lancet* 398.10313, pp. 1825–1835.

SPI-M (2021). *SPI-M-O: Consensus Statement on COVID-19*. Tech. rep.

Stadler, T et al. (Jan. 2012). “Estimating the basic reproductive number from viral sequence data”. *Mol. Biol. Evol.* 29.1.

Stadler, Tanja et al. (Jan. 2013). “Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)”. *Proc. Natl. Acad. Sci. U. S. A.* 110.1, pp. 228–233.

Stanevich, Oksana et al. (July 2021). “SARS-CoV-2 escape from cytotoxic T cells during long-term COVID-19”.

Starr, Tyler N et al. (Sept. 2020). “Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding”. *Cell* 182.5, 1295–1310.e20.

Starr, Tyler N et al. (Apr. 2021). “Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016”. *Cell Reports Medicine* 2.4.

Statistics Sierra Leone (2016). *Sierra Leone - Census, Standards & Statistics*. Tech. rep.

Sturgeon, Nicola (2021). *Coronavirus (COVID-19) update: First Minister’s Statement - 21 December 2021*. https://www.gov.scot/publications/coronavirus-covid-19-update-first-ministers-statement-21-december-2021/8. Accessed: 2022-3-18.

Suchard, M A, Weiss, R E, and Sinsheimer, J S (June 2001). “Bayesian selection of continuous-time Markov chain evolutionary models”. *Mol. Biol. Evol.* 18.6, pp. 1001–1013.

Suchard, Marc A et al. (Jan. 2018). “Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10”. *Virus Evol* 4.1, vey016.

Sylvester Squire, J et al. (June 2017). “The Ebola outbreak and staffing in public health facilities in rural Sierra Leone: who is left to do the job?” *Public Health Action* 7.Suppl 1, S47.

Tegally, Houriiyah et al. (Mar. 2021). “Detection of a SARS-CoV-2 variant of concern in South Africa”. *Nature* 592.7854, pp. 438–443.

Tian, Fang et al. (Aug. 2021). “N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2”. *Elife* 10.

Tiffany, Amanda et al. (June 2017). “Estimating the number of secondary Ebola cases resulting from an unsafe burial and risk factors for transmission during the West Africa Ebola epidemic”. *PLoS Negl. Trop. Dis.* 11.6, e0005491.

Timothy, Joseph W S et al. (2019). “Early transmission and case fatality of Ebola virus at the index site of the 2013–16 west African Ebola outbreak: a cross-sectional seroprevalence survey”. *Lancet Infect. Dis.* 19.4, pp. 429–438.

Twohig, Katherine A et al. (2021). “Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: a cohort study”. *Lancet Infect. Dis.* 22.1, pp. 35–42.

Vaidyanathan, G (May 2021). “Coronavirus variants are spreading in India - what scientists know so far”. *Nature* 593.7859, pp. 321–322.

Varble, A et al. (Nov. 2014). “Influenza A virus transmission bottlenecks are defined by infection route and recipient host”. *Cell Host Microbe* 16.5.

Viana, Raquel et al. (Jan. 2022). “Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa”. *Nature*, pp. 1–10.

Vogels, Chantal B F et al. (May 2021). “Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2”. *PLoS Biol.* 19.5, e3001236.

Voloch, Carolina M et al. (Nov. 2020). “Intra-host evolution during SARS-CoV-2 persistent infection”. *medRxiv*, p. 2020.11.13.20231217.

Volz, E M and Frost, S D W (Aug. 2017). “Scalable relaxed clock phylogenetic dating”. *Virus Evol* 3.2.

Volz, Erik et al. (Mar. 2021a). “Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England”. *Nature* 593.7858, pp. 266–269.

Volz, Erik et al. (Jan. 2021b). “Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity”. *Cell* 184.1, 64–75.e11.

Wain-Hobson, Simon et al. (Jan. 1985). “Nucleotide sequence of the AIDS virus, LAV”. *Cell* 40.1, pp. 9–17.

Wamala, Joseph F et al. (July 2010). “Ebola Hemorrhagic Fever Associated with Novel Virus Strain, Uganda, 2007–2008”. *Emerg. Infect. Dis.* 16.7, p. 1087.

Watts, Duncan J et al. (2005). “Multiscale, resurgent epidemics in a hierarchical metapopulation model”. *Proceedings of the National Academy of Sciences* 102.32, pp. 11157–11162.

Wauquier, Nadia et al. (2015). “Understanding the Emergence of Ebola Virus Disease in Sierra Leone: Stalking the Virus in the Threatening Wake of Emergence”. *PLoS Curr.* 7.

Weigang, Sebastian et al. (Nov. 2021). “Within-host evolution of SARS-CoV-2 in an immunosuppressed COVID-19 patient as a source of immune escape variants”. *Nat. Commun.* 12.1, pp. 1–12.

Whitmer, Shannon L M et al. (Oct. 2016). “Preliminary Evaluation of the Effect of Investigational Ebola Virus Disease Treatments on Viral Genome Sequences”. *J. Infect. Dis.* 214.suppl 3, S333–S341.

WHO Ebola Response Team (Oct. 2014). “Ebola Virus Disease in West Africa — The First 9 Months of the Epidemic and Forward Projections”. *N. Engl. J. Med.* 371.16, pp. 1481–1495.

Wilder-Smith, Annelies and Osman, Sarah (Dec. 2020). “Public health emergencies of international concern: a historic overview”. *J. Travel Med.* 27.8.

Wilkinson, Annie and Fairhead, James (Jan. 2017). “Comparison of social resistance to Ebola response in Sierra Leone and Guinea suggests explanations lie in political configurations not culture”. *Crit. Public Health* 27.1, pp. 14–27.

Wilkinson, Eduan et al. (Feb. 2019). “The effect of interventions on the transmission and spread of HIV in South Africa: a phylodynamic analysis”. *Sci. Rep.* 9.1, pp. 1–12.

Wilkinson, Eduan et al. (2021). “A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa”. *Science* 374.6566, pp. 423–431.

Williamson, Maia Kavanagh et al. (June 2021). “Chronic SARS-CoV-2 infection and viral evolution in a hypogammaglobulinaemic individual”. *medRxiv*, p. 2021.05.31.21257591.

Willis, Rhiannon Yapp And (Nov. 2021). *Coronavirus (COVID-19) Infection Survey, characteristics of people testing positive for COVID-19, UK - Office for National Statistics*. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveycharact 3november2021. Accessed: 2021-11-4.

World Health Organisation (2015a). *Factors that contributed to undetected spread*. Tech. rep.

— (May 2015b). “First confirmed Ebola patient in Italy”.

— (2015c). *Sierra Leone: A slow start to an outbreak that eventually outpaced all others*. Tech. rep.

— (2015d). *Sierra Leone: a traditional healer and a funeral*. Tech. rep.

— (2016). *WHO statement on end of Ebola flare-up in Sierra Leone*. Tech. rep.

— (2020a). *Ebola Virus Disease External Situation Report*. Tech. rep.

— (2020b). *Novel Coronavirus situation report 10*. Tech. rep.

— (Jan. 2021a). *Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health*. https://www.who.int/publications/i/item/9789240018440. Accessed: 2021-6-21.

— (2021b). *HIV/AIDS*. https://www.who.int/data/gho/data/themes/hiv-aids. Accessed: 2022-3-8.

World Health Organisation (2022). *World Health Statistics*. http://www.who.int/gho/publications/world_health_statistics/2014/en/. Accessed: 2022-3-18.

Wu, Fan et al. (Feb. 2020). “A new coronavirus associated with human respiratory disease in China”. *Nature* 579.7798, pp. 265–269.

Yakovenko, M L et al. (Feb. 2014). “The 2010 outbreak of poliomyelitis in Tajikistan: epidemiology and lessons learnt”. *Eurosurveillance* 19.7, p. 20706.

Yang, Chang Hoon and Jung, Hyejin (Mar. 2020). “Topological dynamics of the 2015 South Korea MERS-CoV spread-on-contact networks”. *Sci. Rep.* 10.1, pp. 1–11.

Yen, Hui-Ling et al. (Jan. 2022). “Transmission of SARS-CoV-2 (Variant Delta) from Pet Hamsters to Humans and Onward Human Propagation of the Adapted Strain: A Case Study”.

Yurkovetskiy, L et al. (Oct. 2020). “Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant”. *Cell* 183.3.

Zhang, Lizhou et al. (June 2020). “The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity”. *bioRxiv*.

Zuckerkandl, Emile and Pauling, Linus (1965). “Molecules as documents of evolutionary history”. *J. Theor. Biol.* 8.2, pp. 357–366.