Imperial College London

Tools and methods for rapid phylogeographic analysis of pathogen genomic surveillance data with application to poliovirus

David N Jorgensen

Department of Infectious Disease Epidemiology
School of Public Health
Imperial College London

Thesis presented for the degree of Doctor of Philosophy:
Clinical Medicine Research (Public Health)(PhD), April 2023

Abstract

In recent years, the use of phylogenetic analysis to trace infectious diseases has expanded with the increasing accessibility of genetic sequencing globally. Traditional phylogenetic methods tend to consider evolutionary timescales and can be ill-suited to the rapid evolution and movement of human viruses. This thesis presents and tests methods to overcome issues with, and expand upon, existing phylogenetic techniques as well as analysis pipelines and examples of these methods applied to real-world data. The primary focus of the thesis is the application of quick and simple discrete trait Markov models to wild poliovirus transmission and the introduction of a novel method designed to overcome challenges with sampling rate differences by location with this technique. Additionally, extensions to a widely used Bayesian phylodynamic model of SARS-CoV-2 are presented, allowing for more flexible analysis and the inclusion of additional covariate data. These techniques allow us to discern movement patterns between polio-endemic regions over the past decade and changes in SARS-CoV-2 dynamics over the initial outbreak wave with changing non-pharmaceutical interventions in Saudi Arabia. We present these analyses with a mix of novel and existing visualisation techniques designed to aid their interpretation by key stakeholders and policymakers. The new discrete trait technique presented has the potential to become widely adopted due to the simplicity of the proposed correction and ease of implementation in several existing platforms and the extension of a phylodynamic model of SARS-CoV-2 shows promise for better modelling of future pandemic diseases.

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work.

This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Statement of originality

I declare that the work presented in this thesis is my own. Any contributions made by others are acknowledged throughout and in the publications and authorship section below.

Publications and authorship

Parts of this thesis have been previously published. Chapter 1 has been published as an open access review paper and is reproduced here under a Creative Commons Attribution 4 International Licence © Oxford University Press, 2020 [1]. I conducted the included example analysis and drafted all sections of the manuscript with editorial input from the other authors: Margarita Pons-Salort, Alexander Shaw and Nicholas Grassly. An additional section on sequence data availability, added following the PhD viva, is based on work by Darlan da Silva Candido and Isobel Blake with the included figure produced by Darlan da Silva Candido. The work presented in chapter 2 has previously been shared internally with partners in the Global Polio Eradication Initiative and input was sought from Salmaan Sharif (World Health Organization), Abdirahman Mahamud (World Health Organization) and Natalia Molodecky (Formerly of the World Health Organization) on the regional blocks incorporated in the final work presented. Comments on this work from Steve Kroiss and Hil Lyons of the Institute for Disease Modelling and Masroor Alam of the World Health Organization are incorporated in the final work. The Skygrowth model in chapter 5 and figures 5.2 and 5.3 have been previously published. Parts A,B and C of figure 5.2 were developed by Tobias Mourier and part A of figure 5.3 by Raeece Naeem, both of the King Abdullah University of Science and Technology, Saudi Arabia. These figures are reproduced here under a Creative Commons Attribution 4 International Licence © Springer Nature, 2022 [2]. I have also been a co-author on additional relevant publications during my PhD which are referenced where appropriate in the text [3–5].
Comments from my supervisors, Nicholas Grassly and Erik Volz, are incorporated throughout the thesis, as well as comments from Margarita Pons-Salort.

The terms of the licence under which these works are reproduced can be found here:
creativecommons.org/licenses/by/4.0/

Acknowledgements

I would like to thank my supervisors, Nick Grassly and Erik Volz for their help and guidance throughout my PhD and the opportunity to expand my work beyond the original project. It has been a long process but the support throughout has helped to keep me on track. Nick, thank you for all your support over the past 5 years and helping me to set up and complete this PhD, it wouldn't have happened without your input, and I greatly appreciate it.

I would also like to thank the other members of the Vaccine Epidemiology Research Group for their help as well as their friendship. A lot of small comments and contributions from the group are incorporated in this work and have helped to improve and shape my research. I would particularly like to thank Marga Pons-Salort for her invaluable input on this thesis and for taking me under her wing despite her own academic workload and young family, hopefully I can pay it back with some nice research in the future! Thanks also to the members of the Virus Evolution and Genomic Epidemiology group for their input on my SARS-CoV-2 work and chats throughout the pandemic, the collaborative work from the group and the wider department was awe-inspiring and I am grateful to have been a small part of it. I also acknowledge the monumental global sequencing effort for SARS-CoV-2. Without the publicly shared sequence data from labs globally and specifically the Pathogen Genomics Lab at the King Abdullah University of Science and Technology, the SARS-CoV-2 work presented in this thesis would not have been possible.

Discussions with many individuals at the World Health Organization and the Bill and Melinda Gates Foundation and the dedication of these organisations and individuals to polio eradication cannot go unmentioned. I am thoroughly grateful for their input on my work and the privilege to work towards such a unique and inspiring goal.

To my family and Elizabeth, thank you for all the support and putting up with my ramblings about matrices, I promise I won't mention them again. Lots of love to you all.

Selected Abbreviations

AFP Acute Flaccid Paralysis

ARD All Rates Different

BSSVS Bayesian Stochastic Search Variable Selection

CTMC Continuous Time Markov Chain

DTA Discrete Trait Analysis

ES Environmental Surveillance

GPEI Global Polio Eradication Initiative

GPLN Global Polio Laboratory Network

GTR Generalised time Reversible model of DNA evolution

HRMP High Risk Mobile Population

IPV Inactivated Poliovirus Vaccine

LMIC Low or Middle Income Country

LTL Local Transmission Lineage

MBD Mutitype Birth Death model of population dynamics

MJ-DTA Markov Jump Discrete Trait Analysis

Mk K-state Markov transition model

MSLE Mean Square Log Error

Ne Effective population size

NPAFP Non-Polio Acute Flaccid Paralysis

NPI Non-Pharmaceutical Intervention

OPV Oral Poliovirus Vaccine. Can be further classified into t; trivalent (serotype 1,2 and 3) b; bivalent (serotype 1 and 3) and m; monovalent (individual serotypes)

PHEIC Public Health Emergency of International Concern

POLIS Polio Information System

PONS Polio Nucleotide Sequence database

R Reproduction number (R_t: at time t, R_{eff}: effective)

UTR Untranslated Region

VDPV Vaccine-Derived Poliovirus. Can be further classified into c; circulating, a; ambiguous, i; immunodeficiency-related

VP1 Capsid-encoding region of the poliovirus genome

WHO World Health Organization

WPV Wild Poliovirus

Contents

Abstract **3**

Copyright Declaration **3**

Statement of originality **5**

Publications and authorship **5**

Acknowledgements **5**

Abbreviations **7**

Aims **23**

1 The role of genetic sequencing and analysis in the polio eradication programme **25**

1.1 Abstract . . . 25

1.2 Introduction . . . 25

1.3 Poliovirus surveillance and investigation of samples . . . 29

1.3.1 Acute flaccid paralysis surveillance . . . 29

1.3.2 Environmental surveillance . . . 30

1.3.3 Laboratory testing . . . 31

1.4 Investigation of transmission . . . 32

1.4.1 Poliovirus evolution: molecular clock and recombination . . . 32

1.4.2 Using virus sequences to reconstruct virus spread . . . 33

1.5 Vaccine-derived polioviruses . . . 36

1.5.1 Extent of cVDPV transmission . . . 38

1.5.2 New strategies to control cVDPV outbreaks . . . 39

1.6 Availability of poliovirus sequence data . . . 40

1.7 Conclusion . . . 41

1.8 Data availability . . . 42

2 Reconstructing the spatio-temporal spread of serotype 1 wild poliovirus in Pakistan and Afghanistan between 2012 and 2022 43

2.1 Abstract . . . 43

2.2 Introduction . . . 44

2.3 Methods . . . 45

2.3.1 Data . . . 46

2.3.2 Outlying sequences . . . 48

2.3.3 International sequences . . . 48

2.3.4 Discrete locations . . . 48

2.3.5 Bayesian phylogeny generation . . . 50

2.3.6 Markov Jump Discrete Trait Analysis (MJ-DTA) . . . 50

2.3.7 Consensus tree generation . . . 51

2.3.8 Local transmission lineages . . . 51

2.4 Results . . . 53

2.4.1 Local transmission lineages . . . 57

2.5 Discussion . . . 60

2.6 Conclusion . . . 64

3 A simple correction to adjust for sampling biases in phylogeographic discrete trait analysis 65

3.1 Abstract . . . 65

3.2 Introduction . . . 66

3.3 Methods . . . 70

3.3.1 General methodology . . . 70

3.3.2 Proposed corrections . . . 71

3.3.3 Tree simulation . . . 74

3.3.4 Modelled scenarios . . . 76

3.3.5 Fitting the model . . . 77

3.4 Results . . . 79

3.4.1 Correction selection . . . 79

3.4.2 Likelihood surfaces . . . 79
3.4.3 Relative migration rate estimates . . . 80
3.5 Discussion . . . 88
3.5.1 Correction performance . . . 93

4 Application of the proposed DTA correction to real-world polio AFP data 94

4.1 Abstract . . . 94
4.2 Introduction . . . 94
4.3 Methods . . . 95
4.4 Results . . . 98
4.4.1 NPAFP rates by region . . . 98
4.4.2 Sensitivity to gamma rate prior . . . 99
4.4.3 Performance of corrections . . . 100
4.5 Discussion . . . 103

5 Bayesian phylodynamic analysis of the 2020 SARS-CoV-2 outbreak in Saudi Arabia 106

5.1 Background . . . 106
5.2 Introduction . . . 108
5.3 Methods . . . 110
5.3.1 Skygrowth . . . 111
5.3.2 Timing of importations . . . 112
5.3.3 PhyDyn . . . 113
5.4 Results . . . 121
5.4.1 Sampling intensity . . . 121
5.4.2 Skygrowth model . . . 122
5.4.3 Timing of importations . . . 122
5.4.4 PhyDyn model . . . 125
5.5 Discussion . . . 129

6 Discussion **135**

6.1 Polio phylogeography 136

6.2 SARS-CoV-2 phylodynamics 141

6.3 Conclusions 143

Bibliography **144**

Appendices **170**

Appendix A Supplementary information for chapter 2 170

Appendix B Supplementary information for chapter 3 187

Appendix C Supplementary information for chapter 5 190

List of Tables

2.1 Number of sequences by country and sample type (AFP or ES) included in the analysis. 47

3.1 Input parameter values fixed across models. Beta and gamma are set to give an R0 of 6 and infectious period of 10 days and the birth/death rate is set in days to give a life expectancy of 68 years. 76

3.2 Migration and sampling rate ranges for each of the modelled scenarios. 77

5.1 Parameter table for the model shown in figure 5.1 and the equations in 5.3 . . . 116

5.2 Reported seroprevalence estimates for the cities included in this analysis. Lines highlighted in grey are used for fitting. 120

A.1 Supporting information for the inferred transitions with Bayes factor support above 10 with the complete dataset of WPV1 detections by ES and AFP surveillance. 175

List of Figures

1.1 Changes in the spatial and temporal distribution of wild poliovirus endemic countries, 1988-2022. * Cases seen in Mozambique in 2021 and 2022 and Malawi in 2021 were imported from the endemic Pakistan-Afghanistan region and spread locally. This was not determined to be a re-establishment of endemic circulation in southern Africa and the WHO AFRO region remains certified wild polio-free. These maps are adapted and updated from Molodecky 2018 with data from the WHO and GPEI [19] 28

1.2 Simplified structure of the poliovirus genome adapted from Semler et al.[48]. This diagram highlights the position of the VP1 region of the genome, routinely sequenced for typing purposes, and shows the relative positions of the other major regions mentioned in the text. Red bars show major recombinant regions determined by Stern et al. and arrows the positions of the three gatekeeper mutations for reversion to virulence in serotype two vaccine-derived polioviruses [49]. 33

1.3 Movement of poliovirus in the west of Tajikistan during the 2010 WPV1 outbreak inferred from a discrete trait analysis of 113 publicly available sequences [54]. **A)** Inferred phylogenetic tree for the poliovirus sequenced in the three regions in 2010. The x-axis is given in decimal years inferred from the clock rate and sampling date of the sequences at the tips of the phylogeny. The inset map shows the location of the four regions in western Tajikistan. **B)** Inferred movement events and number of sequenced poliovirus stool samples per region. The breadth of the arrows indicates the number of inferred movement events. The end points of the arrows represent the centroid of each region and not the exact location of reported events. The inset map shows the location of the study area in the wider context of Central and South Asia. **C)** Sampling density of sequences over time in the regions shown in A. Sughd has only one sequence collected on 2010-05-10 so does not feature in this density plot. 34

1.4 Countries reporting circulating vaccine-derived poliovirus (cVDPV) between 01/01/2017 and 01/01/2020. The size of the points is proportional to the number of paralysis cases reported over this time period. The red, blue and green represent serotype one, two and three VDPVs respectively. Data from the polio information system of the global polio eradication initiative [90]. * Somalia has reported both serotype 2 cVDPV (eight cases) and serotype three (six cases) in the time period covered by this plot. 39

1.5 Number of publicly available poliovirus sequences per year of collection globally. Bars are coloured by the serotype and intratype of available sequences. This figure was produced by Candido DS et al. (in prep.) and is reproduced here to illustrate sequence availability and recent reduction in available sequences. Data presented contains all VP1 sequences with length > 800 nt which were not reference (vaccine sequences), duplicated, missing country and/or date information, or from animal/drug studies. 41

2.1 Pipeline of data analysis from the initial sequences and metadata to the final visualisation. Much of this analysis is carried out in R (pink) or in software produced and packaged alongside BEAST 1 (blue). The maximum likelihood tree generation (yellow square) is carried out in IQ-TREE, interfaced via command line calls within R, meaning the user does not need to leave the R environment. 46

2.2 Map showing the borders of the discrete locations ('regions') used in this analysis. The borders of the countries of interest are shown in black with the discrete regions coloured and with borders shown in white. The Grey lines within the countries of interest show the admin 2 level subdivisions. The inset map shows the countries included in the dataset in a global context. 49

2.3 Regression of root-to-tip distance on sampling time for 3702 poliovirus VP1 sequences. The linear trend and high R^2 value suggest that these data are consistent with a strict molecular clock at a substitution rate which aligns with previous reporting. 54

2.4 Time-resolved phylogenetic tree for Pakistan and Afghanistan coloured according to the regions shown in figure 2. Internal node locations are inferred using Bayesian Markov jump discrete trait analysis. The x-axis represents time in years. Square tips represent sequences isolated from the environment and round tips from AFP cases. The bar chart below the tree shows monthly counts of reported polio AFP cases, coloured by the collection region with the orange vertical line indicating the date of the first sampled sequence included in the genetic dataset. Two key transmission clusters discussed are labelled on the right-hand side of the tree. 55

2.5 Transitions between regions inferred by Discrete Trait Markov Jump analysis. Both plots represent counts of transition events between a pair of locations with Bayes Factor support above 10. **A** shows the numbers of transitions between regions visualised on a circus plot. The width of the arcs represents the number of inferred movement events, and the arcs are coloured according to the exporting location. Total counts of importations and exportations in each region are given around the circumference of the plot. The map in **B** represents the same movements, again coloured by the exporting region with a width relative to the number of transitions represented. This plot shows the eight regions with supported transitions in the context of Pakistan and Afghanistan. The centroid of each region is used as the connection point and is not influenced by the location of isolations within the region. 56

2.6 Breakdown of the number of local transmission lineages classified into the three categories: dead end, with less than 5 AFP sequences, circulating in a region for less than a year and no onwards transmission to other regions; persistent, with more than 10 AFP sequences and circulating in a region for over a year; and export, including any chains not in the previous two categories with any onwards transmissions to other regions. Chains which did not fall into any of these categories are shown as other. Illustrative examples of local transmission lineages from each category are provided in the supplementary information (supplementary Fig. A.5). 58

2.7 Mean and 95% CI of the time to detection for local transmission lineages extant on each day from 01-01-2014 to the end of the reported sequence data by region. The horizontal black line on each plot represents a 547-day time to detection, the time required to reach the 1.5% threshold for difference to previously circulating virus used by the GPEI to classify orphan sequences assuming a clock rate of 0.01 substitutions site^{-1} year^{-1} 59

3.1 Example of a 100-tip phylogenetic tree with the terminal branches coloured red to show the branches over which the sampling rate correction is applied in the 'tip-only' corrections. Internal branches where the Q matrix is left unaltered are coloured blue. The x-axis shows the number of days from the time of the common ancestor (at zero) of the simulated samples (the root of the phylogeny). 74

3.2 Structure of the simple SIR model implemented in MASTER in one of two populations. Additional populations are added as identical modules with additional migration rates to allow movement between all populations. The compartments represent Susceptible (S), Infected (I) and Recovered (R) individuals with the additional Is compartment representing sampling with removal. 75

3.3 Likelihood surfaces for the base and relative migration rate. These are plotted on a log scale as the maximum likelihood estimation is performed on log-transformed parameters. This plot highlights the ridged likelihood surface and the improvement in maximum likelihood estimation with the proposed correction. This surface is for a single tree with simulation migration rates of 0.01 (pop 1 to 2) and 0.04 (pop 2 to 1) and sampling rates of 0.0039 (pop 1) and 0.000975 (pop 2). The surfaces are split into bins, with bounds shown in the figure legend. The lightest colour represents the maximum likelihood (minimum negative log likelihood) region. The horizontal line on each plot represents the true relative migration rate we are attempting to reconstruct ($0.01/0.04 = 0.25, log(0.25) = -1.39$) and the red dot shows the location of the maximum likelihood estimate from the surface. 80

3.4 Proportion of model runs under the low sampling rate scenarios where the 95% confidence interval of the estimated relative migration rate includes the true relative migration rate from the simulation. We investigate different relative rates of migration and sampling to illustrate the performance of the proposed correction over a range of scenarios. These plots show the prediction under the corrected model (left; plots A and C) and the uncorrected all-rates different Mk model (Right; plots B and D). The top row (plots A and B) shows the performance of the correction and standard model in the simple case, where the transmission rate is equal across the two populations, and the bottom row (plots C and D) the biased transmission case, where transmission is two times higher in population 2 than population 1. Where a confidence interval could not be estimated due to the likelihood surface being flat in at least one direction the tree was dropped from the calculation. The number of trees where confidence intervals could be estimated in each scenario is shown in supplementary figure B.1. 81

3.5 Example phylogeny for the scenario where migration from population 1 to 2 is 4 times lower than the reverse rate and sampling is 2.94 times lower (x= 0.34, y =0.25 on Fig. 3.4C) Branches are coloured according to their simulated location. This phylogeny is presented to highlight the lack of samples from population 2, making migration rate reconstruction difficult. The x-axis shows time in days from the root. 83

3.6 Difference between the true and estimated relative migration rates under the low sampling rate scenarios investigated. These plots show the mean squared logarithmic error (MSLE), a measure of the relative difference between the actual and predicted values. The left-hand panels (A and C) show the ability to reconstruct the true relative migration rate with the proposed correction and the right-hand panels (B and D) without. The top panels (A and B) show the performance where transmission rates are equal across the 2 populations and the bottom panels (C and D) where the transmission rate in population 2 is 2 times higher than in population 1. Values of the MSLE above 2 are shown in black on the above plot to allow better focus on the differences between lower values of the MSLE. 84

3.7 Predictive ability of the DTA model under the "medium sampling rate" scenario. A and B) Proportion of true relative migration rates within the 95% confidence intervals of the inferred relative migration rates under the corrected (A) and uncorrected (B) methods. C and D) Mean squared logarithmic error (MSLE), a measure of the relative difference between the actual and predicted values with the corrected (C) and uncorrected (D) methods. Values of the MSLE above 2 are shown in black on the above plot to allow better focus on the differences between lower values of the MSLE. 86

3.8 Predictive ability of the DTA model under the "high sampling rate" scenario. A and B) Proportion of true relative migration rates within the 95% confidence intervals of the inferred relative migration rates under the corrected (A) and uncorrected (B) methods. C and D) Mean squared logarithmic error (MSLE), a measure of the relative difference between the actual and predicted values with the corrected (C) and uncorrected (D) methods. Values of the MSLE above 2 are shown in black on the above plot to allow better focus on the differences between lower values of the MSLE. 87

4.1 **A)** Annual and **B)** mean non-polio acute flaccid paralysis (NPAFP) rate between 2012 and 2022 across regions of the polio endemic Pakistan-Afghanistan block, shown in situ in the legend. The NPAFP rate is calculated as the number of cases per year per 100,000 inhabitants under the age of 15. 99

4.2 Transition rate matrices for movements between the ten included locations under the uncorrected (top) and tip-only (bottom left) and full-tree (bottom right) corrections. Cells are coloured in a scale to show the difference from zero, with darker colour representing higher rates. Rates with Bayes factor support are highlighted with green cell outlines. The remaining cells represent unsupported transition rates and are not considered in the discussion of the results. 101

4.3 Transition rate matrices for the tip-only (left) and full-tree (right) corrections relative to the inferred transition rates under the uncorrected model (Figure 4.2) Blue cells represent a lower relative rate and red cells a higher relative rate with white cells representing parity with the uncorrected model. Darker colours represent a greater difference from parity. 102

4.4 Bayes factor supported numbers of movements between regions highlighting changes from the baseline uncorrected model with **A)** the proposed correction applied at the tips only and **B)** applied over the full phylogeny. Dark red lines represent a reduction in the number of predicted movements and dark green lines an increase. 103

5.1 Flow diagram representing the modelled populations and the movements between them. The model includes a modified SEIR model in the internal population, accounting for superspreading with an additional I compartment with a higher transmission rate. The model also includes an exponentially growing international reservoir of SARS-CoV-2 infected individuals allowing diversity to be introduced into the local population. 115

5.2 **Sample overview and population genetics. A)** Locations of the sampling cities within Saudi Arabia. **B)** Stacked bars showing the numbers of samples retrieved from the 4 cities and the Eastern region during the first six months of the pandemic. Cities are colored as in panel A. Months are shown at the bottom of the figure, and each month is divided into 5-day intervals. Major restrictions imposed by the Ministry of Health and by Royal decrees are indicated above plots. **C)** Stacked bars showing the average numbers of new daily cases in sampling cities. **D)** Estimate of effective reproduction number (R_{eff}) over time in Saudi Arabia (top) and estimate of effective population size (N_e), the relative population size required to produce the diversity seen in the sample (bottom). Central black lines show median estimates, and gray confidence areas denote the 95% credible intervals. The red horizontal red line represents an R of 1, the level required to sustain epidemic growth. 123

5.3 **A)** Global Time-scaled phylogeny of 952 Saudi Arabia samples colored by Nextstrain clades. Samples are shown as circles and colored according to their genotype at genome positions 28,88128,883. Intensive Care Unit (ICU) status and sampling region are indicated on the right of the tree. **B)** Distributions of importation dates for the five Nextstrain (nextstrain.org) clades found in Saudi Arabia colored by clade. 124

5.4 Effective reproduction number estimates (black line) and 95% credible interval (grey shaded area) over time from the PhyDyn model without seroprevalence for four cities of interest. The red horizontal line represents an R_{eff} of 1, the number required for virus replacement. These estimates have step changes at monthly intervals. 126

5.5 Estimates of the cumulative proportion of the population infected (black line), 95% credible interval (light grey shaded area) and interquartile range (dark grey shaded area) in the model over time for each city. The red error bars and points represent the seroprevalence predicted by the serosurveys and the bounds of the beta binomial distribution applied to these likelihoods. The position of these points on the x-axis represents the end point of the serosurvey from which they were collected. 127

5.6 Estimates of the cumulative proportion of the population infected (black line), 95% credible interval (light grey shaded area) and interquartile range (dark grey shaded area) in the model over time for each city. The red error bars and points represent the seroprevalence predicted by the serosurveys and the bounds of the beta binomial distribution applied to these likelihoods. The position of these points on the x-axis represents the end point of the serosurvey from which they were collected. 128

5.7 Effective reproduction number estimates (black line) and 95% credible interval (grey shaded area) over time from the PhyDyn model with seroprevalence included as an additional covariate for 4 cities of interest. The red horizontal line represents an R_{eff} of 1, the number required for virus replacement. These estimates have step changes at monthly intervals 129

A.1 Neighbour joining tree of the sequence data included in this analysis with tips coloured by region. The outlying sequences of interest are highlighted with a red circle. The x-axis shows genetic distance calculated under a simple JC69 model with the phangorn R package. 170

A.2 Neighbour joining tree of the sequences included in the dataset (as in supplementary figure A.1) with additional background sequences from Genbank highlighted in grey and labelled. These background sequences are KJ460261, a wild poliovirus 1 sequence from the South Corridor region in 2011, a year before the included data, KF992615, a wild poliovirus 1 sequence circulating in Nigeria at the same time, a vaccine virus, Sabin 1 (accession V01159) and KJ170521 a vaccine derived virus circulating in Nigeria in 2011. 171

A.3 Maximum likelihood tree showing the included sequences in the context of historic global transmission of WPV1 which can be reconstructed from publicly available sequence data. 173

A.4 Time-resolved phylogenetic tree of the full dataset split by regions. A new lineage is plotted after each inferred virus movement between regions. The tips represent sequenced viruses, with round tips representing virus isolates from AFP cases and square tips isolations from ES. The root of each lineage is annotated with a diamond coloured according to the inferred ancestral location. As the time of the virus movement from one region to the other cannot be precisely inferred, the ancestral location is plotted at the mid timepoint of the ancestral branch. Where a lineage leads to onwards transmissions in another region, this is shown with a branch leading to an arrow coloured by the region exported to. 177

A.5 Examples of local transmission clusters which would be categorised into each of the classifications based on length of persistence in time, number of AFP tips and onwards transmission to other regions. 183

A.6 Time-resolved phylogenetic tree of sequences from AFP cases for Pakistan and Afghanistan, coloured according to the regions in the map in figure 2. Internal node locations are inferred using Bayesian Markov jump discrete trait analysis and time in years is given on the x-axis. The bar plot below the tree shows monthly counts of reported polio AFP cases, coloured by the collection location and the orange vertical line the date of the first sampled sequence included in the genetic dataset. 185

A.7 **A)** Circus plot of supported migration events between the regions considered in the analysis. The movements are coloured according to the starting location and the width represents the number of transitions inferred, with counts provided around the circumference of the plot. **B)** The same transitions visualised on a map of the discrete regions. 186

B.1 Number of simulations retained after removing those where a maximum likelihood value or standard error for the relative migration rate could not be estimated with both the corrected and uncorrected method. Panel A shows the numbers retained for the low transmission rate scenario and panel B the transmission difference scenario. 188

B.2 Number of simulations retained after removing those where a maximum likelihood value or standard error for the relative migration rate could not be estimated with both the corrected and uncorrected method. Panel A shows the numbers retained for the medium rate scenario and panel B the high transmission rate scenario. 188

B.3 Illustrative example of a phylogeny sampled under the high sampling rate scenario. Here we see the tips are sampled close in time and the root is more recent than in the low sampling rate example shown in figure 3.1. 189

C.1 Effective reproduction number estimates (black line) and 95% credible interval (grey shaded area) over time from the PhyDyn model for 4 cities of interest. The red horizontal line represents an Reff of 1, the number required for virus replacement. These estimates have three step changes set to align with changes in NPIs 191

C.2 Estimates of the cumulative proportion of the population infected (black line), 95% credible interval (light grey shaded area) and interquartile range (dark grey shaded area) in the model over time for each city. 192

Aims

This thesis aims to address some critical challenges and barriers to the implementation of phylogenetic techniques downstream of considerable recent advances in genetic sequencing technology and sequencing capacity. Currently, widely used phylogenetic techniques are often inaccessible to regional sequencing partners due to the computational resources or training required to implement them. This has improved over time with graphic user interface access to tree-building software but advances in simple techniques are required to ensure results returned are representative of true underlying dynamics. Here we attempt to address a key assumption of the simple discrete trait Markov models which are often used as default analyses in these accessible platforms, the assumption that sampling is dependent on prevalence, rather than, as is often the case, occurring under a separate framework influenced by accessibility and capacity. Additionally we present some extensions to the SARS-CoV-2 phylodynamic model of Volz et al., allowing additional covariates to be incorporated in the included outbreak model. Data analysis pipelines are also provided, allowing these analyses to be recreated and updated with ease as new data becomes available or to allow visualisation techniques to be applied to different regions or diseases.

The key aims of this thesis are :

1. Improving the accuracy of rapid phylogenetic inference of viral transmission by incorporating additional surveillance data:
    - Accounting for sampling biases in discrete trait analysis of poliovirus surveillance data.
    - Allowing for change-points and seroprevalence data fitting in phylodynamic models of SARS-CoV-2 transmission.
2. Reproducible R pipelines for Bayesian phylogenetic analysis:
    - A Series of functions, scripts and templates are made available on github to streamline the production of complex BEAST xml files for phylogeographic and pylodynamic analyses.

3. Genetic data visualisations with relevant information for policymakers:

    - Novel visualisations of local transmission chains. With large datasets previous visualisations become too cumbersome. Here we suggest a set of discrete categories and present a timeline of detections.
    - Consistent colouring rather than labelling is a widely used technique which is not often incorporated in internal phylogenetic analyses in the polio programme. Providing consistent and distinct colours for different regions and consistently presenting these can provide an intuitive summary and direct ability to relate different plots.

Chapter 1

The role of genetic sequencing and analysis in the polio eradication programme

1.1 Abstract

Genetic sequencing of polioviruses detected through clinical and environmental surveillance is used to confirm detection, identify their likely origin, track geographic patterns of spread and determine the appropriate vaccination response. The critical importance of genetic sequencing and analysis to the Global Polio Eradication Initiative has grown with the increasing incidence of vaccine-derived poliovirus infections in Africa specifically (470 reported cases in 2019), and globally, alongside persistent transmission of serotype 1 wild-type poliovirus in Pakistan and Afghanistan (197 reported cases in 2019). Adapting what has been learned about the virus genetics and evolution to address these threats has been a major focus of recent work. Here we review how phylogenetic and phylogeographic methods have been used to trace the spread of wild-type polioviruses and identify the likely origins of vaccine-derived polioviruses. We highlight the analysis methods and sequencing technology currently used and the potential for new technologies to speed up poliovirus detection and the interpretation of genetic data. At a pivotal point in the eradication campaign with the threat of anti-vaccine sentiment and donor and public fatigue, innovation is critical to maintain drive and overcome the last remaining circulating virus.

1.2 Introduction

Poliovirus is a species C Enterovirus, with three serotypes (one, two and three). The virus has only one natural host, humans, and is primarily spread between individuals by faecal-oral transmission. In rare cases poliovirus can enter the central nervous system, leading to paralytic poliomyelitis (acute flaccid paralysis). Although genomics was in its infancy when work on

preventing poliomyelitis began, poliovirus was one of the earliest genomes to be sequenced in full, with two groups publishing serotype one sequences in 1981 [6, 7]. Following the certification of smallpox eradication in 1980, and with rapid advancements towards disease control and elimination using efficacious vaccines, poliovirus was selected as a World Health Organization (WHO) eradication target with a global campaign starting in 1988 led by the Global Polio Eradication Initiative (GPEI).

The pre-GPEI history of efforts to understand poliovirus biology have been described in detail previously [8]. Following early failed attempts at vaccine development, successful non-neuronal culture of serotype two poliovirus was achieved in 1949, a major step towards the production of the current vaccines [9]. In the same year, the three serotypes were identified and subsequently codified by the Committee on Typing of the National Foundation of Infantile Paralysis (USA) in 1951 [10, 11]. With in-vitro poliovirus growth a possibility, and knowledge of the three serotypes, successful development of inactivated poliovirus vaccine (IPV) and live-attenuated oral poliovirus vaccine (OPV) followed in 1953 and 1956 respectively [12, 13].

With a cheap, effective and orally administered vaccine available and a virus with no known animal reservoir, eradication seemed like an imminent possibility when the GPEI was launched [14, 15]. At that time, approximately 1,000 people a day were paralysed by wild poliovirus (WPV) [16]. This has since been reduced by over 99.9%, with WPV serotypes two and three eradicated in 2000 and 2012 respectively, and endemic circulation of wild serotype one (WPV1) limited to Pakistan and Afghanistan. Despite this progress, prolonged inaccessibility of children to immunisation programmes, political turmoil and anti-vaccine sentiment over the last few years saw WPV1 cases climb from 28 in 2016 to 163 in 2019, the last major wild poliovirus outbreak. In 2022, 35 cases were recorded, including 8 in Mozambique in Southern Africa which were exported from southern Pakistan [17]. Transmission in southern Africa was short-lived and has not re-established endemic transmission following intense vaccination campaigns. The persistence of WPV1 in politically unstable and inaccessible regions is a WHO Public Health Emergency of International Concern (PHEIC) due to the threat of international spread

[18]. The progress made towards poliovirus eradication since 1988 is summarised in figure 1.1, showing the geographic changes in endemic wild poliovirus transmission since the start of the campaign.

Another major hurdle to eradication is the emergence and spread of vaccine-derived polioviruses (VDPVs), which are OPV strains (often also called Sabin strains, in reference to the researcher who led the development of OPV) that have regained transmissibility and pathogenicity and can cause outbreaks of paralysis similar to their wild counterparts [20]. Although the overall net benefit from the use of attenuated virus vaccines has been huge over the course of the polio eradication campaign, the occurrence of VDPV outbreaks in recent years has highlighted the need to phase out OPV. This transition started with the withdrawal of serotype 2 OPV during a globally coordinated 'switch' from trivalent to bivalent OPV in over 150 countries in April 2016. To mitigate risks associated with this withdrawal of vaccine, WHO recommended that at least one dose of IPV be given to infants during routine immunisation [21]. However, a single dose of IPV results in seroconversion in just 50-80% of children (depending on the age at administration [22, 23]) and its delayed introduction as a result of supply shortages and limited coverage has resulted in a growing cohort of children unimmunised against serotype two poliovirus. These children are exposed to an increasing incidence of serotype two VDPV (VDPV2), which has emerged both from trivalent OPV (tOPV) used before the switch and more recently as a result of monovalent serotype two OPV (mOPV2) used to control VDPV outbreaks [24]. In 2019, there were 270 confirmed VDPV2 paralysis cases across Africa, the Middle East and south Pacific, with 15 countries affected [25].

This review describes how collection and analysis of increasing amounts of genetic data is transforming the long-standing poliovirus eradication programme, highlighting both current activities and future promise. It emphasises how our understanding of poliovirus genetics and evolution will be critical to success in tackling the major challenges facing the GPEI, allowing us to better describe transmission of virus and develop more efficient and stable vaccines.

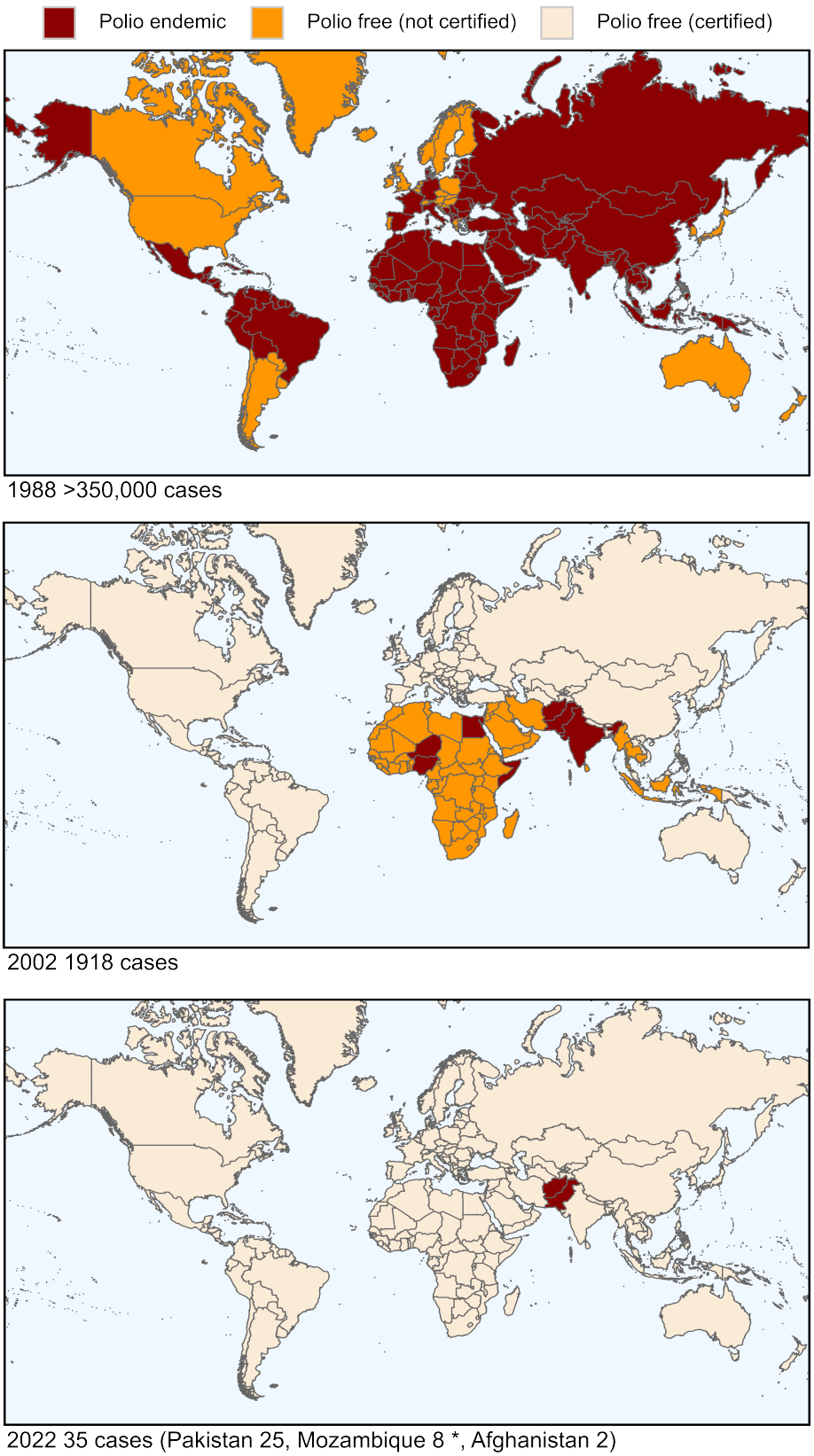


Figure 1.1: Changes in the spatial and temporal distribution of wild poliovirus endemic countries, 1988-2022. * Cases seen in Mozambique in 2021 and 2022 and Malawi in 2021 were imported from the endemic Pakistan-Afghanistan region and spread locally. This was not determined to be a re-establishment of endemic circulation in southern Africa and the WHO AFRO region remains certified wild polio-free. These maps are adapted and updated from Molodecky 2018 with data from the WHO and GPEI [19]

1.3 Poliovirus surveillance and investigation of samples

Identification of poliovirus in the laboratory has had a role in the eradication programme since its inception, with detection of viruses present in stool and environmental samples acting as the major surveillance tool.

1.3.1 Acute flaccid paralysis surveillance

Poliovirus surveillance has traditionally relied on detection and sequencing of poliovirus from stool samples taken from patients with acute flaccid paralysis (AFP). Preceding the global eradication drive, the Pan-American Health Organization recommended reporting all AFP cases in children under 15 years of age and in adults where poliovirus was the suspected cause of paralysis to ensure the detection of paralytic cases [26]. This was incorporated into the global standards of polio reporting adopted by the GPEI. As AFP has several potential viral and non-viral causes, the overall AFP reporting rate is used as an indicator of the sensitivity of the surveillance network to detect poliomyelitis. Gaps in surveillance are a major issue for certification of eradication as poliovirus can circulate undetected for a long time in areas with weak surveillance systems.

As only one in 200 to 1,000 individuals infected with poliovirus is paralysed, the majority of infections are asymptomatic or result in mild illness and are not detected by AFP surveillance. Analysis of the genetic relatedness of polioviruses isolated from AFP cases gives an indication of the sensitivity of AFP surveillance, with viruses found to be distantly genetically related to others referred to as 'orphan' viruses, which are indicative of virus circulation without detection. In Nigeria, a country removed from the endemic list in September 2015, cases of wild poliovirus were detected in 2016 that were directly linked to isolates from virus circulating in the same region in 2011 [27, 28]. This suggests that poliovirus was able to circulate undetected for over 4 years whilst conflict in the north-east of the country prevented adequate AFP surveillance. Highly divergent viruses have also been detected in other settings. For example, in a study

in Pakistan, 7.7% of AFP isolates collected between 2011 and 2013 were orphan viruses with <1.5% divergence from previously collected sequences, suggesting substantial gaps in surveillance [29].

1.3.2 Environmental surveillance

As the incidence of paralytic polio has fallen globally, sampling of sewage networks to look for the virus has expanded; especially in countries with continuing WPV1 transmission or long-term transmission of VDPVs. The majority (>99%) of poliovirus infections are asymptomatic and do not result in paralysis, and so are not detected by acute flaccid paralysis surveillance, but poliovirus shed in stool by infected individuals can still be detected in environmental (sewage) samples. Environmental surveillance (ES) has been shown to be more sensitive than AFP surveillance in areas with convergent sewage networks [30, 31]. For example, inclusion of ES samples in Pakistan reduced the proportion of AFP isolates found to be orphans from 7.7% to 4.9% [29]. Work is ongoing to target these environmental sampling locations to areas where they are likely to have the most impact by investigating catchment areas and sewage conditions.

Sewage sampling is efficient as it allows fewer resources to be used to survey a larger population of possibly affected individuals than direct door to door or local stool surveys. A study in 1991 in Cartagena, Colombia was able to directly sample stool from 196 children in three days in a neighbourhood with approximately 5,300 children. 70% of households in this neighbourhood were served by the sewage network and in the same time period 42 sewage samples were taken. These samples represent repeat samples of the same 70% of the population but much more complete sampling than the 3.6% coverage with stool surveys. When comparing rates of poliovirus detection, 21% of sewage samples were positive compared to 8% of stool samples [32].

Despite the utility of poliovirus ES, it is not without challenges. Poliovirus survival in the environment is highly variable, with a 90% infectivity loss observed in 26 days in sewage at 23

°C [33]. This affects virus detection in cell culture but not necessarily when direct molecular sequencing methods are used. In tropical regions, where poliovirus surveillance is often carried out, the high ambient temperature and bacterial content of the wastewater can further decrease the infective half-life of environmental poliovirus [34].

1.3.3 Laboratory testing

The WHO recommended protocol for detection of poliovirus in stool or ES samples is to isolate the virus using two cell lines, one murine cell line (L20B) that is genetically modified to express the human poliovirus receptor (CD155) and selectively grows poliovirus, and one human rhabdomyosarcoma (RD) cell line which is susceptible to a broad range of enteroviruses [30, 35]. Viruses which grow on the RD cells but not when cross-passaged on L20B cells are classed non-polio enteroviruses and typically not investigated further. Serotype and intratype (vaccine vs other strains) for any polioviruses isolated on L20B cells is determined by quantitative PCR followed by genetic sequencing [36]. Currently, the time from sample arrival at the laboratory and generation of a genetic sequence required to confirm detection is 2-3 weeks, although several activities are aiming to reduce this time by using direct molecular detection methods. For example, the development of sensitive PCR protocols and next generation sequencing (NGS) methods could allow both rapid detection and highly specific typing. Viral RNA can be directly extracted from both stool and ES samples, with PCR amplification and sequencing library preparation performed in as little as two days [37, 38]. PCR primers targeting conserved genomic regions provide detection width, whilst variable regions between the primers allow high resolution typing [39]. Furthermore, multiple polioviruses can be detected simultaneously, with NGS reads corresponding to the mixture of viruses contained in the sample.

1.4 Investigation of transmission

1.4.1 Poliovirus evolution: molecular clock and recombination

The 906-nucleotide VP1 region of the poliovirus genome encoding a portion of the viral capsid is used for typing and is the only section of the genome routinely sequenced (Fig. 1.2). VP1 contributes to the major neutralization sites and amino acid residues on the poliovirus surface, making it highly representative of the interaction of poliovirus with the external environment and host cells [40–44]. Using only this region of the genome it is possible to identify poliovirus amongst other enteroviruses and differentiate between the three serotypes [42]. Poliovirus has a consistently high number of mutations due to its highly error-prone RNA polymerase, which results in a substitution rate on the magnitude of 10^{-2} nucleotide changes per site per year in the VP1 region [45]. Mutations that are fixed in VP1 are usually synonymous as a result of selective constraints on non-synonymous mutations that change the capsid structure. Therefore, they accrue according to a molecular clock, allowing robust phylogenetic trees to be constructed and dates of emergence of new lineages to be estimated [45, 46]. VP1 sequences can therefore be used to investigate poliovirus transmission. Current determination of clusters of transmission follows the US Centers for Disease Control and Prevention guidance with differentiation between circulating viruses based on sequence identity. New clusters are added whenever a detected VP1 sequence is 5% divergent from previously collected samples [47]. Having an indication of the relatedness of viral sequences is a good marker of the progress towards eradication. A reduction in the number of genetic clusters can be indicative of the impact of vaccination campaigns.

In addition to having a high mutation rate, circulating polioviruses frequently recombine with other species C enteroviruses [50]. Much of this recombination occurs in the non-structural regions of the genome during coinfections, meaning recombination does not affect the makeup of the poliovirus surface antigens or the inference of transmission routes from the VP1 sequence. Several projects have been carried out using full genome sequence data to map recombination

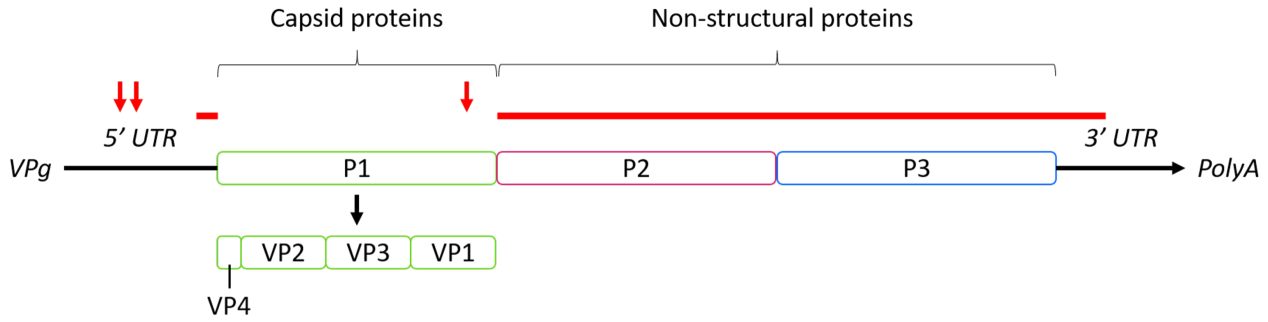


Figure 1.2: Simplified structure of the poliovirus genome adapted from Semler et al.[48]. This diagram highlights the position of the VP1 region of the genome, routinely sequenced for typing purposes, and shows the relative positions of the other major regions mentioned in the text. Red bars show major recombinant regions determined by Stern et al. and arrows the positions of the three gatekeeper mutations for reversion to virulence in serotype two vaccine-derived polioviruses [49].

sites in the genome [49, 51–53]. These primarily occur on the boundary between the 5' UTR and the P1 capsid encoding region and in the P2/P3 non-structural regions (Fig. 1.2). Although this occurs in both wild and vaccine-derived poliovirus in circulation, it is of greater concern in vaccine and vaccine-derived strains due to the loss, through recombination, of attenuating mutations and reversion to a more transmissible pathogenic form.

1.4.2 Using virus sequences to reconstruct virus spread

A number of different phylogeographic methods are available to reconstruct virus spread, incorporating techniques from phylogenetics, coalescent theory and population genomics. Some of the most complex techniques are based on the coalescent theory of Kingman et al. which has been extended to incorporate structured populations and therefore can be used for discrete phylogeographic inference [55–58]. As this technique is computationally costly for large numbers of discrete regions, a number of approximations and extensions have been developed to speed model convergence and include additional parameters [59–64]. Alternate phylogeographic modelling under the birth-death model has also recently become increasingly popular, incorporating parametric rates of infection sampling and recovery in a more familiar framework for infectious disease epidemiologists [65, 66]. Similarly to coalescent methods, structured birth-death methods have been developed, allowing migration between discrete regions [67]. A

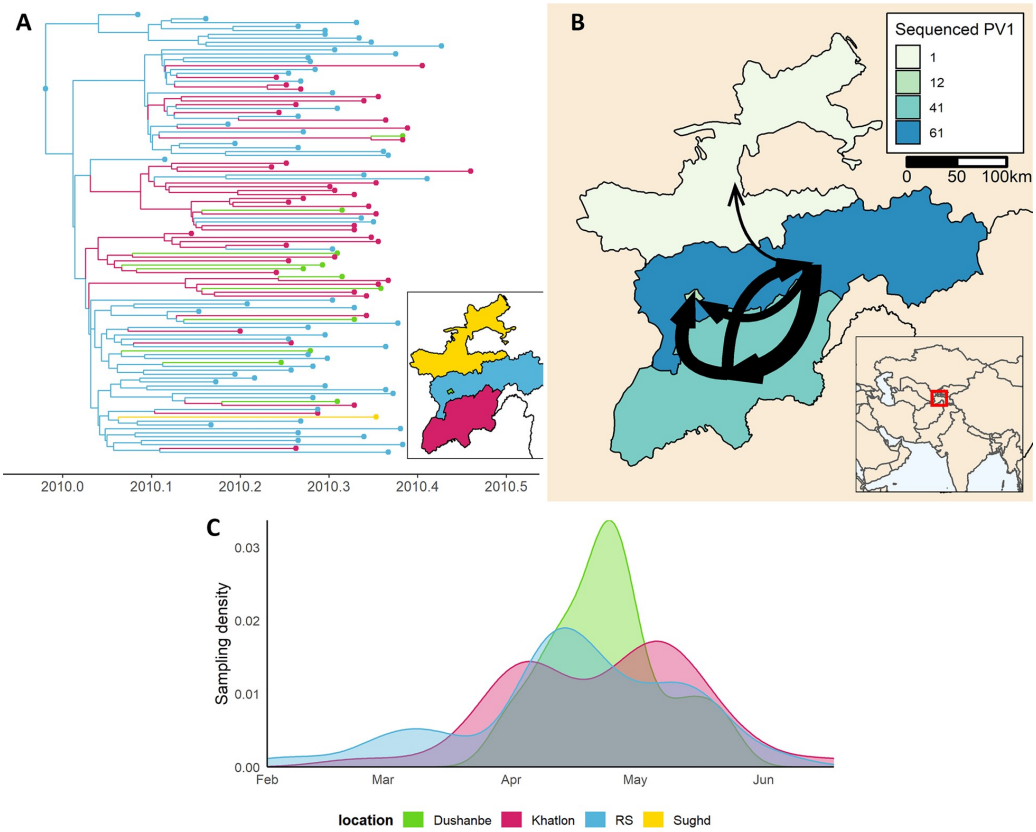


Figure 1.3: Movement of poliovirus in the west of Tajikistan during the 2010 WPV1 outbreak inferred from a discrete trait analysis of 113 publicly available sequences [54]. **A)** Inferred phylogenetic tree for the poliovirus sequenced in the three regions in 2010. The x-axis is given in decimal years inferred from the clock rate and sampling date of the sequences at the tips of the phylogeny. The inset map shows the location of the four regions in western Tajikistan. **B)** Inferred movement events and number of sequenced poliovirus stool samples per region. The breadth of the arrows indicates the number of inferred movement events. The end points of the arrows represent the centroid of each region and not the exact location of reported events. The inset map shows the location of the study area in the wider context of Central and South Asia. **C)** Sampling density of sequences over time in the regions shown in A. Sughd has only one sequence collected on 2010-05-10 so does not feature in this density plot.

discussion of phylodynamic methods based on these two models is presented at the beginning of chapter 5. Migration between discrete regions is most commonly considered with the discrete trait analysis technique, a simple model which treats migration akin to mutation, and can efficiently incorporate large numbers of discrete regions and sequences, this technique and the limitations of treating migration in this way are introduced in more detail in chapter 3. Beyond discrete regions, where detailed location information is available, migration can also be modelled with continuous phylogeographic approaches, avoiding the need to spatially aggregate available sequences into discrete geographic regions and the inability to infer ancestors outside

of the included regions, key issues with the simple discrete trait approach. The continuous phylogeographic approach is outlined in Dellicour et al. 2021 and can also be extended to use a generalised coalescent allowing more than two lineages to merge at the same coalescent event, a limitation of the Kingman coalescent [68–70].

Investigation of the genetic relatedness of isolates through VP1 sequencing has been critical in distinguishing areas with persistent transmission from those suffering repeated introductions of poliovirus. Application of methods from phylogeography has allowed complex virus movement dynamics to be unravelled, and sources and sinks of infection to be identified. To date Discrete Trait Analysis (DTA), a discrete phylogeographic approach based on models of DNA evolution and Felsenstein's pruning algorithm, has been applied to polio; this provides information on the geographic history of viral movement and associated rates of movement between discrete regions [71]. Starting from geolocated virus sequences, this approach allows inference of the location of internal nodes in phylogenetic trees. DTA has previously been applied to inform viral diffusion for several infectious diseases including rabies, influenza and Ebola [72–74]. An illustrative example with a small dataset is presented here using data from the 2010 WPV1 Tajikistan outbreak, available on GenBank (KC880365–KC880521)(Fig. 1.3)[54]. Case reporting from this outbreak shows a large number of cases in both the region of Republican Subordination (RS) and Khatlon, as well as a number of temporally clustered cases in the capital city Dushanbe (separated from RS for this analysis). When a phylogeographic model that incorporates sampling location and time is implemented, it becomes clear that transmission within Tajikistan likely originates in RS, which also supports the majority of virus transmission. All sequenced cases in Dushanbe are inferred to be imported with no onwards transmission within the city or to other regions. Earlier phylogenetic analysis also points to a single importation of WPV1 into Tajikistan from India and onwards transmission, which cannot be inferred from case data alone [54].

The DTA approach has recently been applied to WPV1 sequence data from Pakistan and Afghanistan, where WPV1 persists with complex circulation dynamics and geographical spread

due to mobile and hard to reach populations. These preliminary analyses have identified cycles of transmission between the south-west of each country (Kandahar/Quetta and Karachi). They also highlight repeated introductions of virus from the tribal regions (between Islamabad and Kabul) into the Punjab province of Pakistan and show uninterrupted virus circulation within these tribal regions. Analysis of the distribution of clusters of transmission without formal DTA analysis also highlights these three, key, high-burden regions but does not construct probable ancestral locations or virus movements [75].

Phylogenetic visualisations that split trees into clusters of transmission based on those presented by Dudas et al. for the 2014 Ebola outbreak and interactive phylogenetic trees using the Nextstrain platform have also been developed, and are provided in internal WHO reports, but are not yet publicly available [73, 76]. Regular updates of the polio surveillance sequence data using these new visualisation tools may provide insights into changes in the polio epidemiology, which can in turn guide the implementation of more strategic supplementary immunisation activities. These analyses may be particularly relevant to halt the remaining circulation of wild poliovirus.

Given the declaration of polio as a PHEIC and the value of poliovirus genetic sequence data to interpreting its epidemiology, there is an urgency to share poliovirus sequences and phylogenetic trees more rapidly and more widely to allow broad scrutiny by scientists and policy makers. This is being pursued with the creation of a restricted access global polio nucleotide sequence database (PONS) by the US CDC.

1.5 Vaccine-derived polioviruses

Live-attenuated OPV strains replicate in the gut to induce both local mucosal and systemic immunity, the former being important to prevent transmission of wild viruses. This is a major benefit over injected IPV, which provides systemic immunity against paralysis but limited

intestinal mucosal immunity. Replication of OPV strains is, however, associated with virus shedding and transmission to contacts. Selection for loss of the attenuating mutations which reduce OPV virulence occurs during virus replication through mutation and recombination. In rare cases this can lead the vaccine virus to regain pathogenicity and transmissibility similar to that seen in wild-type polioviruses. Vaccine strains are classified according to their source and their relatedness to the Sabin strain and other polioviruses. Current reporting standards define serotype two VDPVs as those viruses with at least six nucleotides difference from the Sabin strain in the VP1 region (at least 10 nucleotides for serotypes one and three). Vaccine-related viruses not reaching this threshold are classified as 'Sabin-like'. Although it is possible for Sabin-like viruses to cause poliomyelitis, this occurs much less frequently than with viruses which have crossed the threshold to be classified as VDPVs [77, 78].

Replication and shedding of the vaccine virus in the intestine and throat can occur long term in individuals with a primary B cell immunodeficiency who have been vaccinated with OPV as they are not able to clear vaccine virus infection. These immunodeficient VDPVs (iVDPV) may pose a risk to the wider population and have required the introduction of a primary immunodeficiency surveillance programme by the GPEI [79]. Long-term shedding of iVDPV can result in highly divergent viruses. For example, one individual being monitored in the UK was shedding a virus with 17.7% VP1 sequence divergence from the Sabin strain in 2015, approximately 28 years after vaccination [80].

VDPVs are classed as ambiguous (aVDPVs) when they are not linked to ongoing chains of transmission or to any known immunodeficient individuals. When linked cases are subsequently identified, many aVDPVs are re-classed as circulating VDPVs (cVDPVs). cVDPVs may cause large outbreaks of poliomyelitis and are a major threat to the eradication of poliovirus. They are likely to emerge where population immunity is low, and sanitation is poor [81].

A number of studies have highlighted key 'gatekeeper' mutations required for Sabin poliovirus

to regain wild-type fitness in the human intestine. Recent studies of serotype two poliovirus agree on three specific codon positions for these reversions, two in the 5' UTR and the third in the protein coding VP1 region (Fig. 1.2) [49, 82], two of which were first published in 1991 [83, 84]. Alongside specific point mutations, studies have found that the majority of cVDPVs are recombinant with other species C enteroviruses [85, 86]. This often provides revertant viruses with more efficient viral protease and RNA-dependent RNA polymerase. Burns et al. found that 95% of cVDPV2 isolates from northern Nigeria between 2005 and 2011 had lost the major attenuating mutations in the 5' UTR by recombination rather than by point mutation and all isolates had lost the major attenuating mutation in VP1 by recombination [86].

1.5.1 Extent of cVDPV transmission

Since the 2016 switch from trivalent to bivalent OPV, the incidence of VDPV2 outbreaks has increased with 29 VDPV2 outbreaks in 15 countries reported between January 2018 and June 2019 compared to nine in six countries in the same period in 2017-18 [87]. This increase in incidence is largely the result of the replacement of trivalent with bivalent OPV in routine and supplementary vaccination activities and the risk of cVDPV2 transmission associated with falling population immunity to serotype two poliovirus [88, 89]. Analysis of recent cVDPV2 outbreaks has shown that they mostly emerged from mOPV2 given in response to previous outbreaks [24]. This raises the serious concern that mOPV2 is not a suitable tool for cVDPV2 outbreak response in the context of a growing global cohort of children who have not been immunised with OPV2.

cVDPV cases have been reported in 15 countries between January 2017 and January 2020 (Fig. 1.4) [90–92]. Using genetic data, it is possible to infer which of these cases are due to local transmission and which are more likely to be linked to chains of cVDPV transmission in other countries. The polio eradication programme has been able to link cVDPV2 cases in Benin, Cameroon, Ghana and Niger to ongoing circulation in Northern Nigeria, whereas cases in Angola and Democratic Republic of the Congo were found to be the result of unlinked local

emergences [87]. Understanding which cases are due to under-vaccinated local populations and which are due to international spread is important to determine the correct vaccine response and to investigate the cause of the outbreak. In addition, genetic divergence from the originating Sabin strain is used to determine how long a virus has been circulating at the time of detection which influences the scale of the response mounted. Some studies have highlighted a faster mutation rate of sites under strong positive selection pressure early in VDPV emergence. This leads to a slight acceleration in the molecular clock corresponding to a shortening of inferred dates of divergence by about 30 days [49, 82].

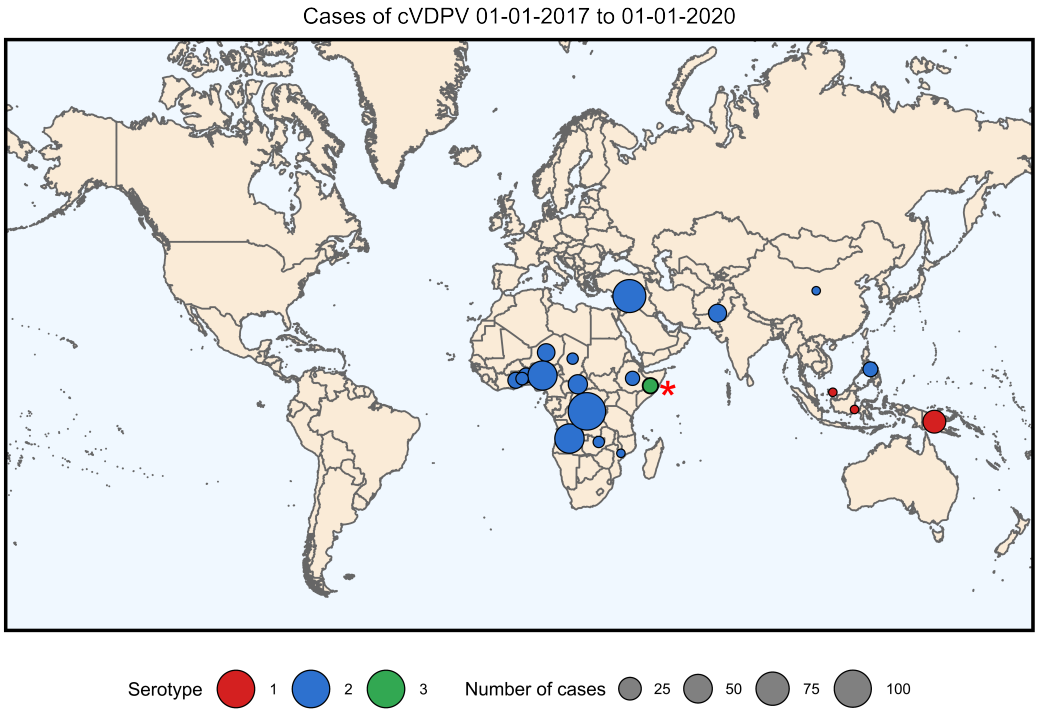


Figure 1.4: Countries reporting circulating vaccine-derived poliovirus (cVDPV) between 01/01/2017 and 01/01/2020. The size of the points is proportional to the number of paralysis cases reported over this time period. The red, blue and green represent serotype one, two and three VDPVs respectively. Data from the polio information system of the global polio eradication initiative [90]. * Somalia has reported both serotype 2 cVDPV (eight cases) and serotype three (six cases) in the time period covered by this plot.

1.5.2 New strategies to control cVDPV outbreaks

When the live-attenuated Sabin oral poliovirus vaccine was in development almost 70 years ago, potential vaccine strains were selected empirically based on their immunogenicity and

neuropathogenicity in animals prior to testing in clinical trials. Current projects to update vaccines can now be informed by our knowledge of the structure and function of the genome. More genetically stable oral poliovirus vaccines are in development with genomes edited to reduce the risk of reversion to neurovirulence [93]. Human phase one trials of novel OPV2 candidate vaccines were carried out in 2018 with improvements to the stability of the major attenuating mutation in the 5' UTR [94]. Each candidate also had further mutations to inhibit recombination and reduce replicative fitness respectively. Small quantities of novel OPV2 vaccines are expected to enter the vaccination programme for emergency use in mid-2020 [87, 94]. These may become a key tool for the programme to stop the current VDPV2 outbreaks while avoiding seeding new ones. At this critical point post-switch, the reintroduction of routine tOPV vaccination is also being debated as a possible solution to control the increasing number of serotype 2 vaccine-derived cases reported. Re-introduction of tOPV would represent a major step back in the polio eradication campaign, away from the phase out of current live vaccines, but would also reintroduce type two immunity in children, helping to control future VDPV2 outbreaks.

1.6 Availability of poliovirus sequence data

Despite ongoing work within the Global Polio Eradication Programme to centrally store sequence data and link this data to associated AFP cases, access to this sequence data is limited. Recent work by Candido DS et al. (in prep.) extracted 7227 publicly accessible poliovirus sequences from Genbank, the majority of which were missing serotype (61.9%), subnational location (92.4%) and sample type (68.2%) information from the attached metadata. In some instances this information could be partially reconstructed from the sequence name, source publication or phylogenetic reconstruction but the publicly available sequences still represent a small subset of the sequence data produced routinely by the eradication programme.

Of the available sequences, 5057 (70%) covered the whole VP1 region routinely used in phylogenetic reconstruction for enteroviruses, with over 50% of these sequences collected from just three countries, Nigeria, Egypt and the Democratic Republic of the Congo. The year of col-

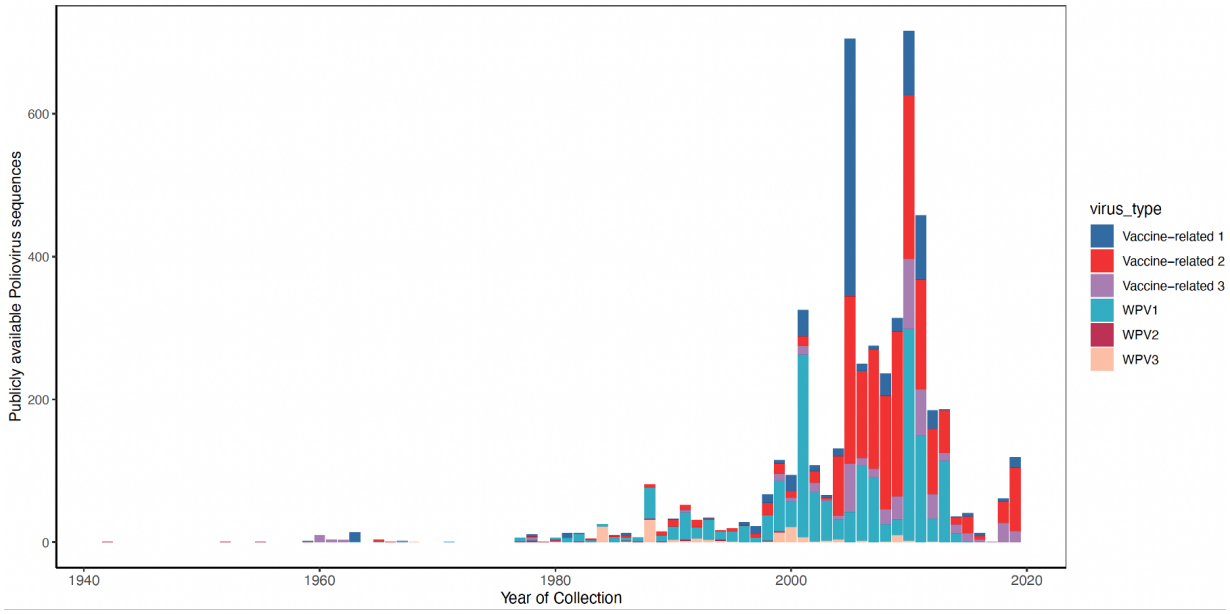


Figure 1.5: Number of publicly available poliovirus sequences per year of collection globally. Bars are coloured by the serotype and intratype of available sequences. This figure was produced by Candido DS et al. (in prep.) and is reproduced here to illustrate sequence availability and recent reduction in available sequences. Data presented contains all VP1 sequences with length $>$ 800 nt which were not reference (vaccine sequences), duplicated, missing country and/or date information, or from animal/drug studies.

lection and sample type of publicly available sequences is shown in figure 1.5, highlighting the reduction in available sequences following the switch from trivalent to bivalent OPV in 2016, and the lack of any recent publicly available WPV1 sequences. The restricted public availability of sequence data limits the contribution of the wider phylogenetic community to poliovirus eradication and differences in the public availability of sequence data by country can influence the inferences made from poliovirus sequence data.

1.7 Conclusion

With an upsurge in the number of wild poliovirus cases in 2019, and nine countries affected by ongoing VDPV outbreaks at the beginning of 2020, the GPEI is at a critical stage. Recent improvements in genetic sequencing and analysis techniques are helping support efforts to achieve a polio-free world. The main current lines of research include the development of direct detection and sequencing protocols that allow rapid detection and typing of poliovirus from stool and environmental samples (thus reducing the time from sample collection to reporting);

the implementation of intuitive pipelines for bioinformatics and genetic analysis that speed and standardise the production of genetic data; the implementation of phylogeographic analysis and development of visualisations tools that allow analysis and interpretation of the genetic data by planners and policymakers; and the production of new oral vaccines based on newly engineered strains that have a lower risk to revert and regain virulence.

1.8 Data availability

No new data is presented in this chapter. An illustrative example is presented using publicly accessible data from GenBank (KC880365–KC880521)

Chapter 2

Reconstructing the spatio-temporal spread of serotype 1 wild poliovirus in Pakistan and Afghanistan between 2012 and 2022

Phylogeographic analysis points to a shift from persistent poliovirus transmission within local regions to frequent elimination and re-introduction driven by population movement between regions. This shift was associated with the extinction and replacement of a major circulating viral cluster.

2.1 Abstract

Despite concerted vaccination efforts, wild poliovirus remains endemic in Pakistan and Afghanistan (22 reported poliomyelitis cases and 59 environmental detections in 2022). Mass population movements and inaccessible reservoirs pose a challenge to polio eradication, with understanding the geographic distribution and movement of poliovirus key to interrupting transmission long-term. Here we investigate poliovirus genetic sequence data (n=3702 sequences), from the past decade (January 2012 to July 2022) which are routinely collected for virus typing and genetic clustering, to reveal long term transmission patterns and recent changes in dynamics using Bayesian discrete trait phylogeographic methods. This analysis highlights the importance of the city of Karachi in the south of Pakistan and the border regions between Pakistan and Afghanistan to wild poliovirus persistence over the past decade, with these regions historically supporting large and persistent transmission clusters and regularly seeding transmission in other regions. We also find a change in dynamics following a 2019 peak in transmission, from long-term, geographically distinct, persistent transmission in four key transmission regions to shorter, often interrupted, local transmission lineages with frequent movements between regions. This process coincides with the die out of one of two major genetic clusters of transmission in the endemic region. More recently, in 2022, very few movements between regions have been

observed with all reported cases limited to the south of Khyber Pakhtunkhwa province. Our analysis provides additional granularity to viral movement and detection data, useful epidemiological information to guide immunisation activities and surveillance in the drive towards polio eradication.

2.2 Introduction

Since the introduction of effective vaccines in the 1950s and 60s and the launch of the Global Polio Eradication Initiative (GPEI) in 1988, all three serotypes of wild poliovirus have been in decline globally, with serotypes 2 and 3 declared eradicated in 2015 and 2019 respectively [12–14, 95]. The remaining serotype 1 wild poliovirus only circulates endemically in Pakistan and Afghanistan with this circulation considered a Public Health Emergency of International Concern by the World Health Organization due to the risk of international spread [18, 96].

The most recent major outbreak of wild poliovirus in these endemic countries occurred between 2019 and 2021, with current transmission limited in scope. In 2022, WPV1 primarily persisted in the border region between the two endemic countries, with large mobile and hard-to-reach populations, political instability and economic migration between and within the two countries contributing to persistence [97]. In 2022, paralytic poliomyelitis cases were primarily limited to the North Waziristan district of Pakistan, with an additional case in North East Afghanistan. WPV1 was also detected in wastewater collected across parts of Northern and Eastern Pakistan and North-Eastern Afghanistan as well as from a sample collected in Karachi in Southern Pakistan [96]. The reduced geographical range of detections since the beginning of 2021 compared to the outbreak period suggests that interventions have been effective in most of the endemic region and there may currently be a good opportunity to push towards eradication by interrupting transmissions.

Recently we have seen a return of wild poliovirus to Africa, with an outbreak in Malawi and

Mozambique in 2022. This has been linked back to transmission in the south of Pakistan in 2019, underscoring the importance of achieving wild type poliovirus eradication in endemic regions to protect children globally from poliomyelitis [17].

In this study, we use Bayesian discrete trait phylogeographic methods to reconstruct the circulation of wild serotype 1 poliovirus over the last 10 years (2012-2022) in the remaining endemic countries. We use genetic sequences of the VP1 region, encoding a portion of the viral capsid, which is routinely sequenced for poliovirus typing by labs in the Global Polio Laboratory Network [42]. This region mutates according to a consistent molecular clock on the order of 10^{-2} nucleotide changes per site per year due to the error prone RNA polymerase encoded in the poliovirus genome, allowing robust phylogenetic inferences to be made with simple methods [45]. On top of reconstructing long term patterns of movement, inferring the locations of unobserved ancestral virus from the known locations of sampled tips allows us to link poliovirus sequences to each-other via their inferred ancestors and determine individual lineages of related sequences at a regional level [98]. As a result, we can distinguish between different transmission dynamics, including long term persistent virus circulation, dead ends for viral transmission (where imported virus is detected but with few associated cases and without onwards exportation) and regions that regularly seed virus transmission elsewhere. In this analysis, acute flaccid paralysis (AFP) data are supplemented with additional WPV1 isolates from environmental surveillance (ES) of wastewater systems. ES is an important supplemental surveillance system for detecting poliovirus transmission and provides additional detail on transmission routes and geographic spread of the virus [99].

2.3 Methods

In this analysis we infer a phylogeny of wild serotype 1 poliovirus with genetic sequence data and use discrete trait analysis (DTA) to estimate the likely locations of internal nodes in the phylogeny [98]. With this information, we then estimate movement events between regions in Pakistan and Afghanistan and the duration of transmission within those regions. To this aim,

we first reconstruct a set of empirical trees with BEAST 1.10.5 before implementing the Markov Jump method of Minin and Suchard to estimate ancestral locations and transitions between regions [98, 100]. The R pipeline created for this analysis recreates aspects of the baltic python pipeline (`https://github.com/evogytis/baltic`) with R code to generate animated and fixed migration maps, phylogenetic trees with overlaid case data, and local transmission lineages from a nexus consensus phylogeny produced in BEAST. The code is built on a number of R packages, primarily the phylogenetic package ape and the ggplot2 and ggtree plotting packages and is available on github (`https://github.com/JorgensenD/PhD_thesis`). A visualisation of the complete pipeline used from initial metadata matching to final analysis and visualisation is shown in figure 2.1.

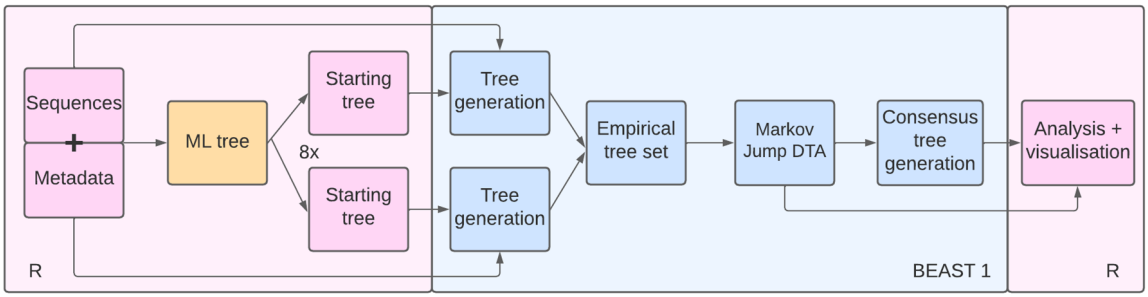


Figure 2.1: Pipeline of data analysis from the initial sequences and metadata to the final visualisation. Much of this analysis is carried out in R (pink) or in software produced and packaged alongside BEAST 1 (blue). The maximum likelihood tree generation (yellow square) is carried out in IQ-TREE, interfaced via command line calls within R, meaning the user does not need to leave the R environment.

2.3.1 Data

We analysed VP1 gene sequences of serotype 1 wild poliovirus detected across Pakistan and Afghanistan through AFP and ES surveillance. These sequences were generated by the Regional Reference Laboratory for Poliomyelitis at the National Institutes of Health, Islamabad, Pakistan as part of routine poliovirus surveillance. This laboratory performs virus culture and sequencing for both Pakistan and Afghanistan. The culture and sequencing of these viruses followed the WHO gold-standard protocol outlined in chapter 1, section 1.3.3, with isolation on both RD cells sensitive to enteroviruses (the genus to which polioviruses belong), and adapted

L20B cells which express the human poliovirus receptor to identify polioviruses. Viruses which have a cytopathic effect on L20B cells are amplified with quantitative PCR and poliovirus detections are referred for VP1 sequencing [36].

The sequences are 906 bases long with mostly synonymous amino acid changes which occur according to a molecular clock at a high rate compared with most viral pathogens [8, 40, 101]. The dataset used in this analysis includes 3702 sequences which could be matched to location and date metadata which cover just over 10 years, from 1st January 2012 to 20th July 2022 (Table 2.1).

Table 2.1: Number of sequences by country and sample type (AFP or ES) included in the analysis.

Country	ES	AFP	Total
Pakistan	2272	836	3108
Afghanistan	331	263	594
Total	2603	1099	3702

The two data sources, AFP and ES, have different collection methodologies, with AFP sampling generally distributed according to transmission and ES sampling restricted to specific sampling sites [31, 102]. ES samples can also often return a large number of closely related or identical sequences due to the extraction and sequencing protocols used, meaning there are a large number of duplicate sequences in the numbers included in Table 2.1, here we retain duplicate sequences as they do not affect inferred transitions although they could be removed to reduce computational cost. The number of ES sampling sites has expanded over the course of the data from 23 in 2012 (All in Pakistan) to 233 in 2022 (201 in Pakistan and 32 in Afghanistan). These sites are targeted to known reservoirs of WPV1 transmission, transportation hubs and areas with large mobile populations [31]. To assess the potential impact on the results of using sequences from ES, we performed phylogeographic analysis on two datasets, a full dataset including all sequences (AFP and ES) presented in the main text and an AFP only dataset, a priori considered as less biased, included in the supplementary material.

2.3.2 Outlying sequences

Prior to attempting to reconstruct ancestral locations with discrete trait analysis, the sequence data were aligned with MUSCLE in AliView and checked for outliers [103, 104]. A maximum likelihood phylogeny was generated with IQ-Tree (version 1.6.12) and re-rooted with the R package ape (version 5.3) to give a phylogeny most compatible with a strict molecular clock [105, 106]. Root-to-tip divergence was calculated in R with EpiGenR (version 1.1.0) [107].

2.3.3 International sequences

As the last remaining endemic countries for poliovirus transmission (Figure 1.1), no reservoirs of wild poliovirus exist outside of the Pakistan-Afghanistan region. In this analysis we do not include historic sequence data from other regions, making the explicit assumption that the common ancestor of all wild poliovirus detections in the endemic region over the past decade was also in the endemic region, with no importations from other regions. A phylogeny is presented in supplementary appendix A with the included sequences in the context of historic publicly available sequence data, showing that the included sequences cluster together with older sequences from the same region, distinct from the majority of WPV1 seen circulating elsewhere. A cluster of detections from Israel and Egypt in 2013 published by Shulman et al. and a single detection from China in 2011 (GenBank accession: MH750912) cluster together with the Pakistan-Afghanistan sequences, although these have previously been inferred to have been seeded by transmission in the Pakistan-Afghnaistan region, as were WPV1 detections in Mozambique and Malawi more recently which are not publicly available [17, 108].

2.3.4 Discrete locations

For the analysis in this paper, we divided Pakistan and Afghanistan into ten regions based on the identification of discrete areas of linked poliovirus circulation ('epidemiological blocks') by the GPEI. These included three corridor regions along the border between the two countries:

the South Corridor, centred around the major border crossing between Quetta and Kandahar; the North Corridor, centred around a major border crossing between Peshawar and Jalalabad linking the capitals of the two countries; and the Central Corridor, incorporating a more porous border with various smaller crossing points. In addition to these corridors, the city of Karachi on the coast of Pakistan in the South, was included as a discrete region alone due to its importance as a hub for poliovirus transmission. Beyond these previously identified regions, the two countries were divided into a further three discrete locations each based on their population density and incorporating major transportation routes between the key regions outlined (Figure 2.2).

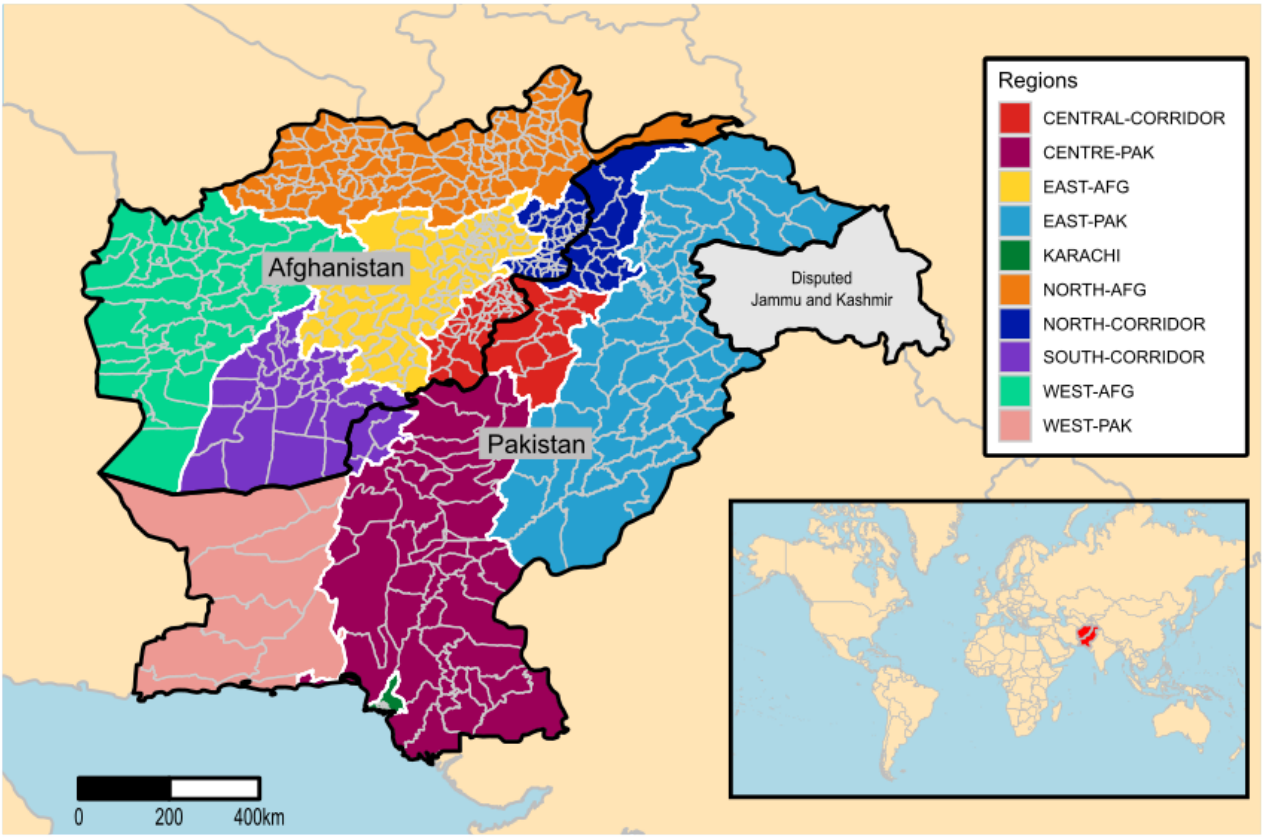


Figure 2.2: Map showing the borders of the discrete locations ('regions') used in this analysis. The borders of the countries of interest are shown in black with the discrete regions coloured and with borders shown in white. The Grey lines within the countries of interest show the admin 2 level subdivisions. The inset map shows the countries included in the dataset in a global context.

2.3.5 Bayesian phylogeny generation

Prior to Bayesian tree generation, a single maximum likelihood tree was inferred from the available sequence data with IQ-TREE [105]. From this initial tree, branches were collapsed and polytomies resolved randomly to give eight unique starting trees for Bayesian phylogeny generation. These starting trees were dated using the R package treedater with the molecular clock rate estimated using initial bounds of 0.005 and 0.015 substitutions per site per year and a starting value of 0.01, informed by previous analyses of the VP1 region [45, 109].

Eight Markov chains, initiated with each of the unique starting trees, were run in parallel for 15 million iterations with a phylogeny sampled every 1000 iterations. A strict clock with a strongly informative normal prior centred on a clock rate of 0.01 with a standard deviation of 0.001 was used, based on the calculated root-to-tip divergence (see results section). Multiple chains were used to speed up the analysis and also to allow convergence to be assessed by ensuring effective sample sizes for each parameter were over 200 and 95% credible intervals were not significantly different across chains. After the iteration limit was reached, the initial 30% of each chain was removed as burn-in and the tree files combined with logcombiner [110]. For the complete dataset of ES and AFP data, two of the eight chains failed to converge and were dropped from further analysis. For the AFP dataset alone, all chains were retained. The trees produced were down-sampled to give a set of 1000 trees to be used as an empirical tree set for DTA.

2.3.6 Markov Jump Discrete Trait Analysis (MJ-DTA)

Discrete trait transitions over the phylogeny were estimated using the Markov jumps and rewards method of Minin and Suchard, included in BEAST 1.10.5 [98, 100]. This method was chosen over the traditional DTA as it can be used to estimate numbers of transitions between locations directly in the BEAST model. Despite the large number of parameters to estimate, the method can be sped up significantly by supplying a pre-estimated empirical tree set to

estimate over as we did in this analysis. As the numbers of transitions are sampled directly as parameters in BEAST, statistical support and bounds for numbers of transitions can be trivially calculated in post-analysis of the Markov Chain Monte Carlo (MCMC) trace files. The code used to produce the BEAST XML files used in this analysis is available on github. The Markov jump DTA was run for 10 million iterations with one of the 1000 empirical trees randomly sampled at each iteration. A 10% burn-in was removed after the iteration limit was reached. Spread3 v0.9.7 was used to assign Bayes Factor support to the inferred number of transitions between pairs of discrete locations [111]. Transitions with Bayes Factor support above 10 are presented in the results section.

2.3.7 Consensus tree generation

To allow for the visualisation of geographic clusters of transmission, the set of phylogenetic trees output from BEAST were combined to produce a representative consensus tree. A maximum clade credibility tree with common ancestor heights was generated with TreeAnnotator, including location assignments at each internal node from the Markov jump DTA [110]. Common ancestor heights were used to ensure only positive branch lengths were reconstructed in the final maximum clade credibility phylogeny. This consensus tree was used to produce the phylogenies and transmission chain information presented in the results section.

2.3.8 Local transmission lineages

When inferring local transmission lineages (LTLs) we use a phylogeny illustrated with the most probable internal node locations for each internal node and the known locations of the sequenced tips. We divided the consensus phylogeny into LTLs by splitting the phylogeny wherever a movement between regions was inferred by discrete trait analysis (two connected nodes of the phylogeny are inferred to be in different regions). Here we use the midpoint of the branch between connected nodes with different inferred locations as the estimated time of this movement. This technique was based on visualisations presented by Dudas et al. for the

2013-2016 West African Ebola outbreak [73]. LTLs are summarised in this analysis by classifying them into three distinct types. The first, 'dead-end' lineages, are defined as circulating for less than one year with fewer than 5 associated AFP sequences and without onwards transmission to other regions. The second, 'persistent' lineages are detected for greater than one year with greater than 10 associated AFP cases. The third are 'export' lineages which include any remaining lineages not included in the previous two groups which seed onwards transmission in another region. Classification of LTLs into these summary groups uses AFP case data as a measure of lineage size as this is less influenced by sampling intensity by region than ES data. All sequences are considered when assessing movement between regions and time to detection or die-out.

We calculate a mean time to detection and 95% CI for each region using the inferred LTLs. Time to detection is calculated as the time from the midpoint of the branch connecting an LTL to a node in a different region to the first detection in the LTL by either ES or AFP surveillance. This is calculated as a daily rolling mean where any LTL which is extant at time t is included in the mean and 95% CI presented for that time. This includes any LTLs which were detected before time t which are still extant or any LTLs which were yet to be detected at time t but were later detected. When assessing time to detection we start from the beginning of 2014 due to the inflated time to detection for earlier sequences which sit close to the root. This inflation is due to the lack of samples in the dataset prior to these early detections, meaning that this part of the tree has fewer internal nodes and less information available to reconstruct the importation date of a lineage. There will also be edge effects due to censoring of the data at the end of the period covered, with any extant lineages which are yet to be detected at the endpoint not included. Code to extract transmission lineages and recreate the plots presented from a BEAST consensus tree is provided on github (`github.com/JorgensenD/PhD_thesis/`).

2.4 Results

A linear regression of divergence on time (Figure 2.3) showed that the data was consistent with a clock rate of 0.011 substitutions per site per year, with no major outliers from the linear relationship (R^2 =0.96) [107]. This is consistent with previously published data on WPV1 and allowed a fixed molecular clock with narrow bounds to be used for the remainder of the analysis [45].

Two large outbreaks of WPV1 occurred in Pakistan and Afghanistan over the period covered by the sequence data, with the first peaking in the fall of 2014 and the second peaking in the fall of 2019 (Figure 4). In both cases, the Central Corridor contributed significantly to early circulation and recorded a high proportion of AFP cases. Both outbreaks then spread primarily to the South corridor, with the 2014 outbreak also leading to a large number of cases in the North Corridor. The North Corridor supported the long-term persistence of a WPV1 lineage between the outbreaks which went on to seed the majority of circulation in the later outbreak. This lineage is the direct ancestor of all currently circulating WPV1. A large outbreak dominated by cases in the South corridor preceded the sequences included in this analysis (2010-2011) (Figure 2.4).

We infer the root of all included sequences to be in early 2006 (2006-04-17; 95% HPD: 2005-05-08 to 2007-06-04), although this includes a group of outlying sequences collected in Afghanistan in 2012 with no recent ancestors. Investigation of these sequences shows some shared non-synonymous mutations which are not present in the rest of the dataset (Investigated further in the supplementary information). When this group is excluded, the root of all other sequences is found to be in late 2010 (2010-10-30; 95% HPD: 2010-06-17 to 2011-03-25). We estimate that there were 640 (95% CrI 619 to 659) movement events between regions over the 10 years covered.

We see strong positive correlation between the number of sequences included in the dataset

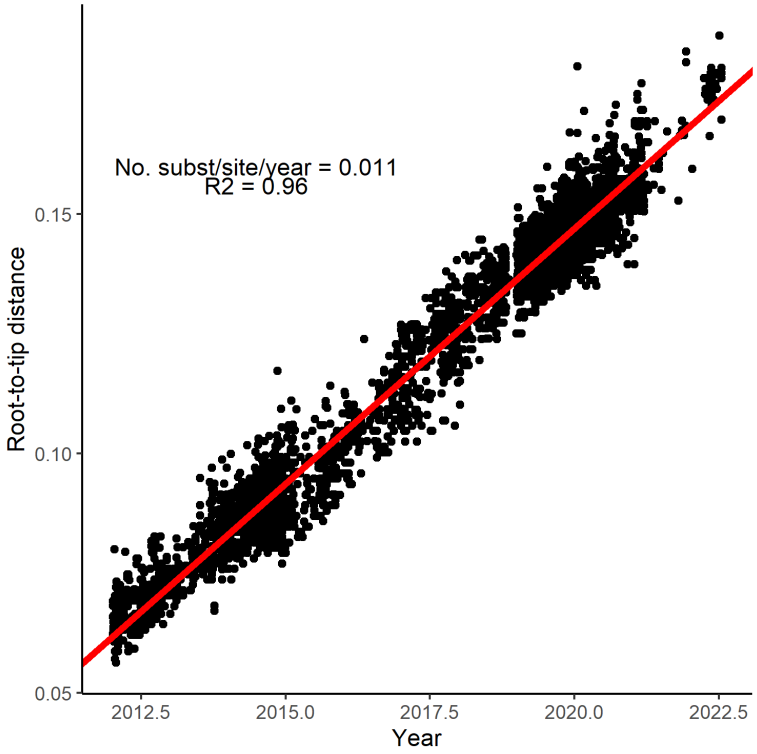


Figure 2.3: Regression of root-to-tip distance on sampling time for 3702 poliovirus VP1 sequences. The linear trend and high R^2 value suggest that these data are consistent with a strict molecular clock at a substitution rate which aligns with previous reporting.

and the number of cases of acute flaccid paralysis due to WPV1 across regions (R = 0.75, p = 0.007) and across years (R = 0.82, p = 0.002) with the Pearson test. This suggests we have a representative sample of cases in the study area across the time period.

The phylogeny with inferred internal node locations presented in figure 2.4 reveals repeated movements between the city of Karachi and the South Corridor region (Karachi to South Corridor n = 53, 95% CrI 42 to 61; South Corridor to Karachi n = 20, 95% CrI 14 to 26), as well as from these regions to Centre Pakistan (Karachi to Centre Pakistan n = 106, 95% CrI 91 to 119; South Corridor to Centre Pakistan n = 35, 95% CrI 27 to 42). Outside of this southern transmission loop, there are also repeated movements from the North Corridor to the Central Corridor (n = 19, 95% CrI 14 to 23), and to a lesser extent the reverse (n = 11, 95% CrI 8 to 14). Exports from the Central Corridor region are low considering the large burden of AFP cases in this region shown in figure 2.4, reporting primarily long term persistent transmission lineages and dead end groups of linked isolates without onwards transmission (Figure 2.6). Light blue branches, representing transmission in East Pakistan (primarily covering the highly populous Punjab province) are interspersed in both transmission loops, suggesting consistent spill-over from the primary reservoirs to this region throughout the decade. The analysis of

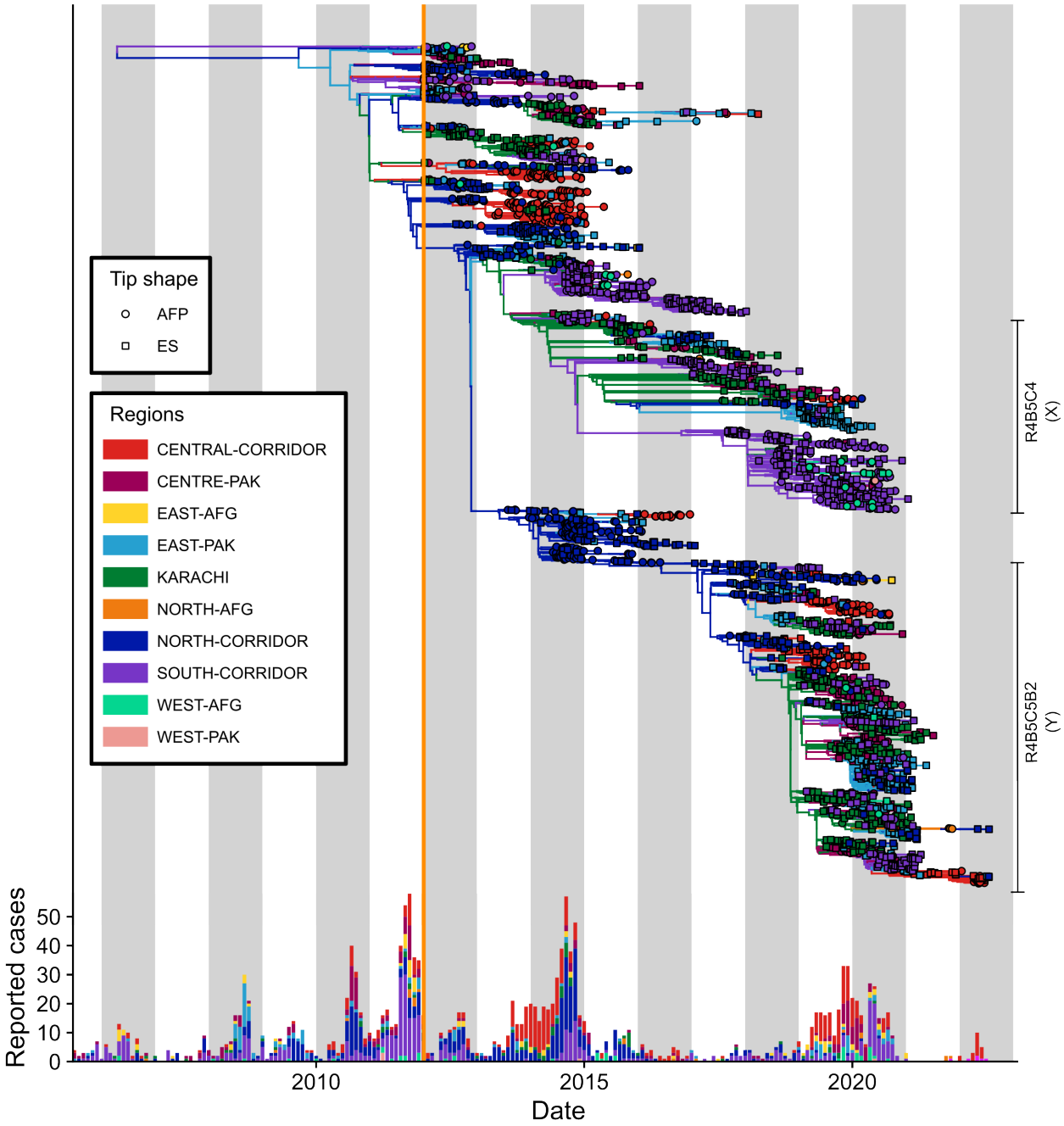


Figure 2.4: Time-resolved phylogenetic tree for Pakistan and Afghanistan coloured according to the regions shown in figure 2. Internal node locations are inferred using Bayesian Markov jump discrete trait analysis. The x-axis represents time in years. Square tips represent sequences isolated from the environment and round tips from AFP cases. The bar chart below the tree shows monthly counts of reported polio AFP cases, coloured by the collection region with the orange vertical line indicating the date of the first sampled sequence included in the genetic dataset. Two key transmission clusters discussed are labelled on the right-hand side of the tree.

AFP data alone presented in the supplementary shows very little exportation linked to East Pakistan compared to what we see here with both ES and AFP data. East Pakistan has a large number of ES sampling sites with large catchments which influences the inferred contribution of the region to virus transmission elsewhere. The exportations from each region, shown in figure 2.5, highlight Karachi as a major exporter of virus to other regions which is supported when ES data are not included in supplementary figure A.7. A full table of movement events with Bayes Factor support above 10 between pairs of regions is provided in supplementary table A.1

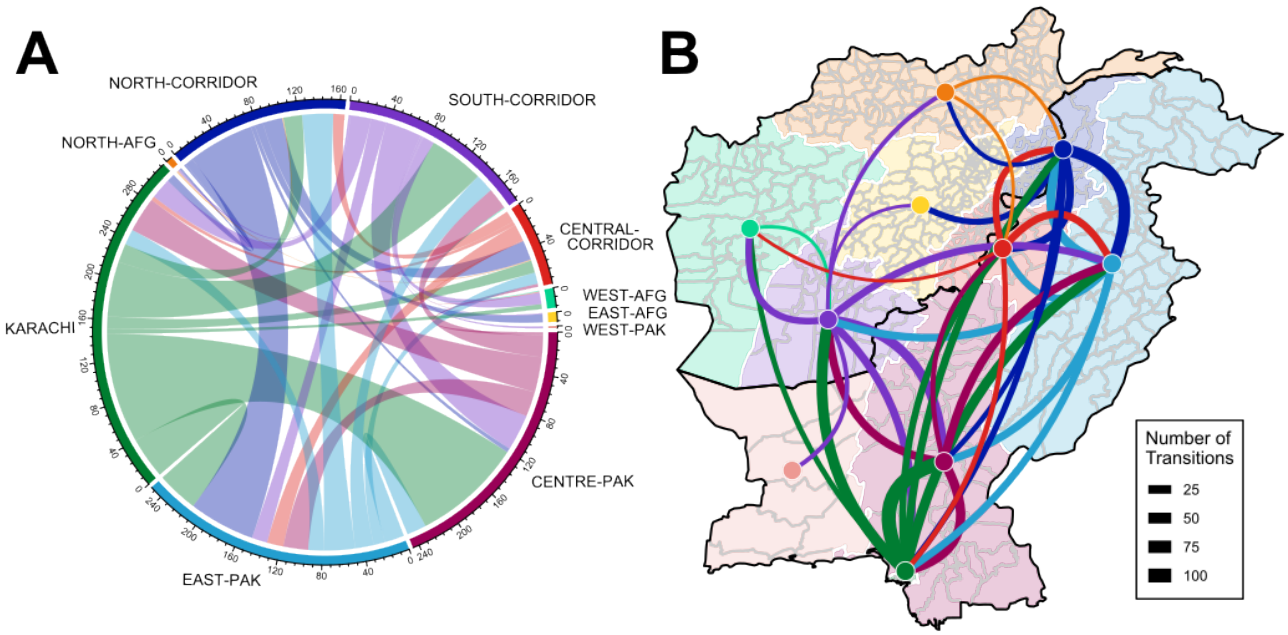


Figure 2.5: Transitions between regions inferred by Discrete Trait Markov Jump analysis. Both plots represent counts of transition events between a pair of locations with Bayes Factor support above 10. **A** shows the numbers of transitions between regions visualised on a circus plot. The width of the arcs represents the number of inferred movement events, and the arcs are coloured according to the exporting location. Total counts of importations and exportations in each region are given around the circumference of the plot. The map in **B** represents the same movements, again coloured by the exporting region with a width relative to the number of transitions represented. This plot shows the eight regions with supported transitions in the context of Pakistan and Afghanistan. The centroid of each region is used as the connection point and is not influenced by the location of isolations within the region.

Since the 2014 outbreak, all reported sequences have belonged to two main WPV1 lineages that coexisted for over eight years, with a common ancestor in September 2012 (Figure 2.4). However, all sequences collected in 2022 have been related to one of these lineages, known under the US Centers for Disease Control and Prevention naming scheme as R4B5C5B2 or 'Y', suggesting that the other lineage (R4B5C4 or 'X') died out in early 2021 (lineages labelled

on Figure 2.4). The extinct X lineage circulated mostly in the South corridor, Karachi and East Pakistan. The Y lineage mostly circulated in the North Corridor until the 2019 outbreak, where it spread to the southern regions (Karachi, Centre Pakistan and the South Corridor), co-circulating with the X lineage throughout 2019 and 2020. The Y lineage was seen to replace the X lineage towards the end of 2020. When investigating numbers of transmissions between locations we see a large amount of Y cluster circulation supported by chains of transmission between the southern regions, despite it originating in the North Corridor and currently only circulating in the Central and North corridors. This suggests that the widespread expansion of this cluster to the southern regions in 2020 has recently reversed.

Analysis including only sequences from AFP cases shows similar results in terms of the general observation of a clear transmission loop in the south and lesser, but still important, movement between the North and Central Corridors (Supplementary Information). However, the contribution of East Pakistan and Karachi in particular in the full analysis is boosted by the inclusion of ES data due to the large number of ES sites in these regions.

2.4.1 Local transmission lineages

Local transmission lineages provide useful information on time before detection and type of clusters circulating in each region, full plots showing all the LTLs by region are presented in supplementary figure A.4. These show the location a cluster was imported from, the sampling method for each sequenced tip and the location of any onwards transmission.

Figure 2.6 highlights changes over time in the circulation patterns, with transmission shifting from persistent transmission lineages within regions before 2019 to smaller dead-end transmission lineages in most locations and repeated reintroduction. In the Central Corridor, the current focus for WPV1 transmission, there are continued persistent lineages with long term uninterrupted transmission. This figure also highlights the increase in the number of extant

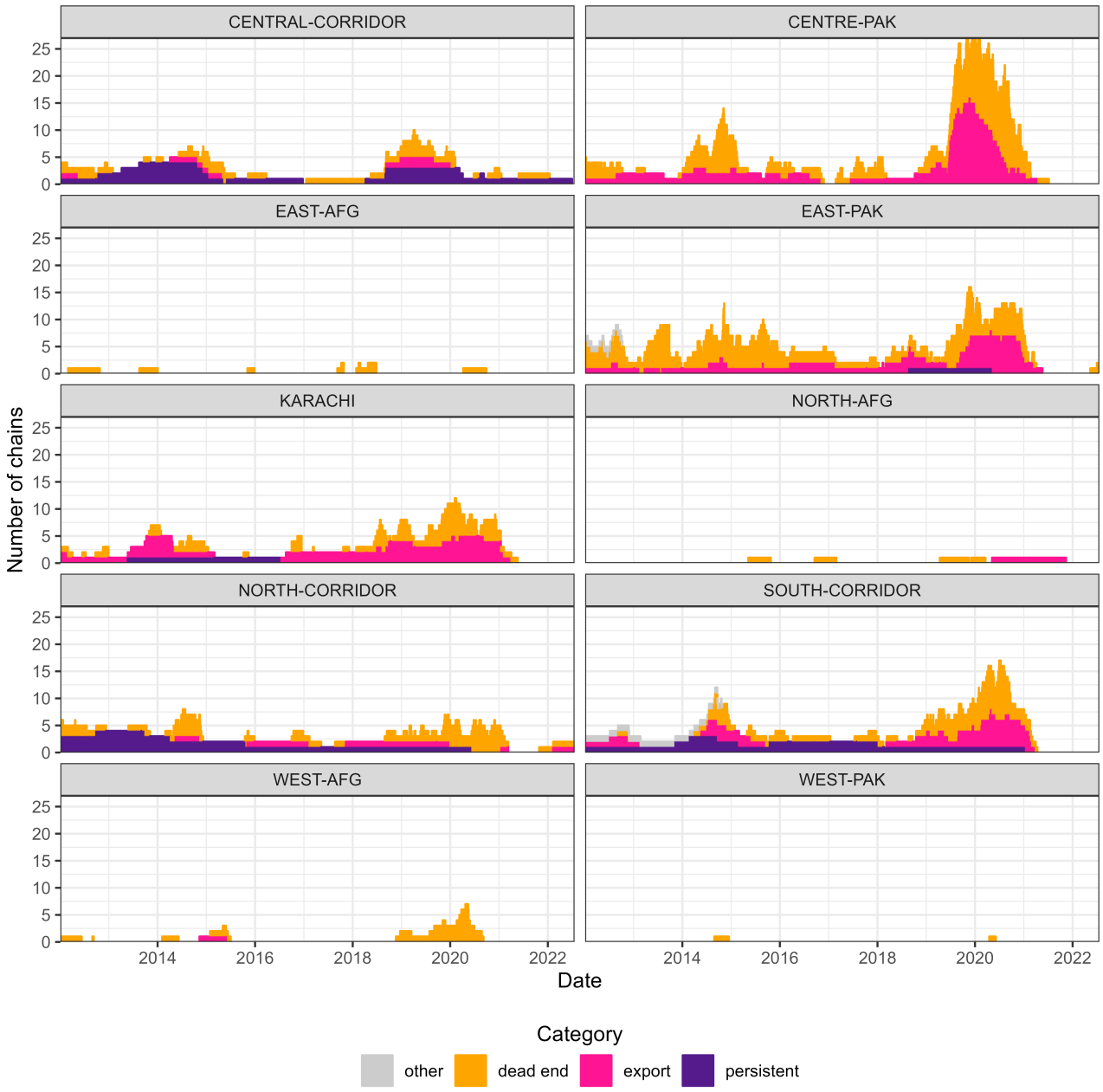


Figure 2.6: Breakdown of the number of local transmission lineages classified into the three categories: dead end, with less than 5 AFP sequences, circulating in a region for less than a year and no onwards transmission to other regions; persistent, with more than 10 AFP sequences and circulating in a region for over a year; and export, including any chains not in the previous two categories with any onwards transmissions to other regions. Chains which did not fall into any of these categories are shown as other. Illustrative examples of local transmission lineages from each category are provided in the supplementary information (supplementary Fig. A.5).

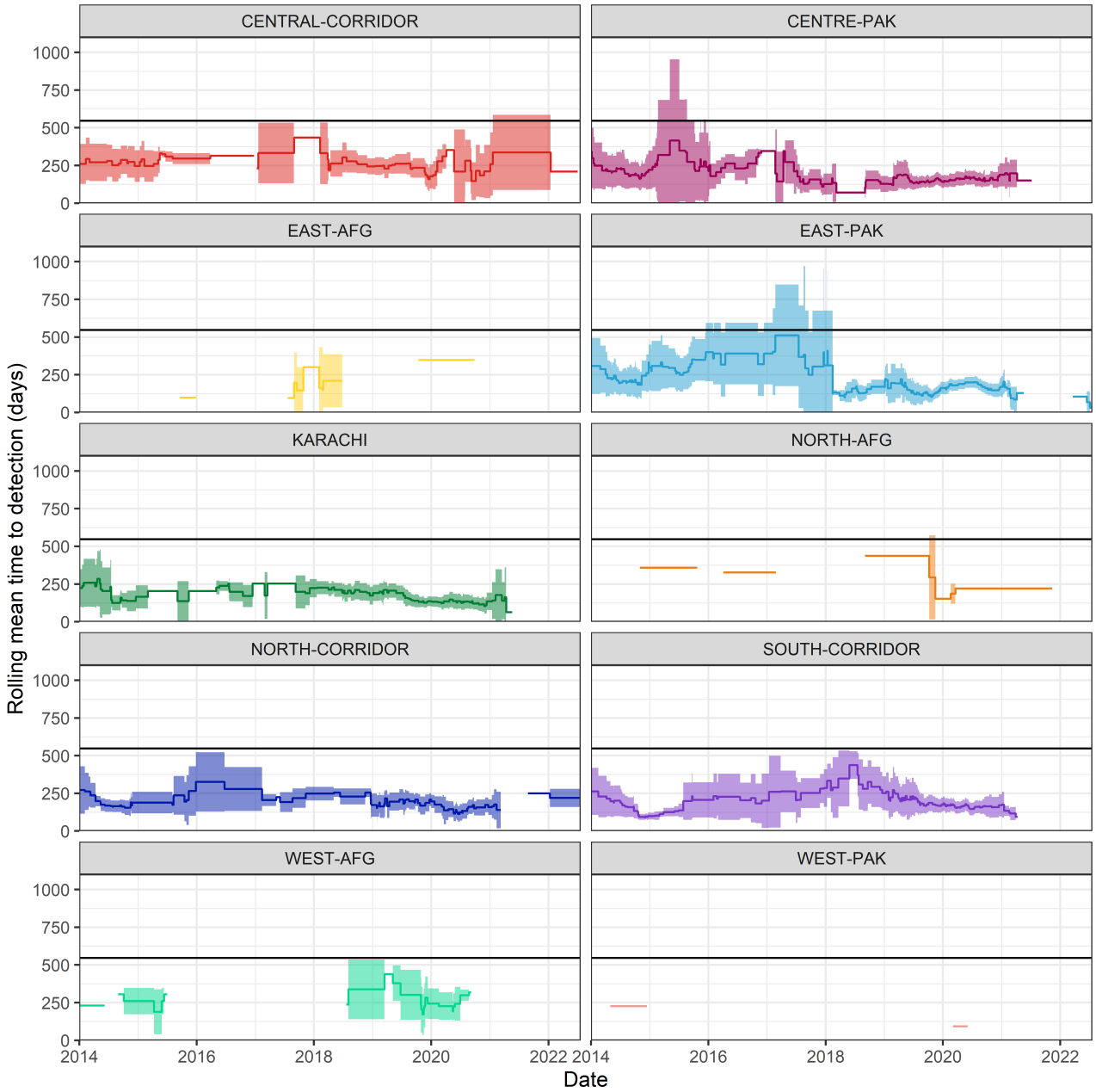


Figure 2.7: Mean and 95% CI of the time to detection for local transmission lineages extant on each day from 01-01-2014 to the end of the reported sequence data by region. The horizontal black line on each plot represents a 547-day time to detection, the time required to reach the 1.5% threshold for difference to previously circulating virus used by the GPEI to classify orphan sequences assuming a clock rate of 0.01 substitutions site^{-1} year^{-1}

lineages associated with the 2019 outbreak. More inter-regional transmission during the outbreak period, with a greater number of importations and exportations, led to more inferred LTLs. The high number of dead-end lineages during this period indicates that these lineages are generally shorter with fewer associated cases (Figure 2.6). All of the persistent LTLs shown in figure 2.6 have resulted in onward transmission to other regions.

Time to detection of LTLs over time in each region is presented in figure 2.7, with the expected time to reach the threshold of 1.5% sequence difference from previous isolates (used by policymakers to determine orphan detections) overlaid. We see general improvement in time to sampling in all regions with large samples of LTLs over the course of the included data and all are, on average, below the orphan detection threshold over the whole period from 2014 to 2022. Time to sampling in 2022 varies across regions; In particular, we estimate a time to sampling of a transmission lineage over the year to July 2022 of 193 days (95% CI 163-222) in the Central and North Corridor regions compared to 89 days (95% CI 37-142) in the remainder of the endemic region.

2.5 Discussion

By reconstructing the spread of wild serotype 1 poliovirus across Afghanistan and Pakistan using sequence data we are able to identify key regions which support the persistence of endemic WPV1 transmission. We identified the corridor regions between the two countries to be key to continued persistence of transmission, all supporting recent persistent lineages and with a long term persistent lineage still circulating in the central corridor at the end of the included data; This supports the findings of previous research [112, 113]. Other key exporting regions, Karachi and Centre Pakistan, have also been linked to transmission of wild poliovirus in two further outbreak countries in Southern Africa, highlighting the importance of interrupting endemic transmission to transmission globally [17]. Despite the expansion of poliovirus surveillance in the endemic region, continued circulation means the risk of spread to regions with less intense surveillance persists. Weaker surveillance in non-endemic regions means that any imported wild virus could circulate for longer without detection than in the endemic regions, requiring larger

vaccination responses which bring an increased risk of seeding vaccine derived transmissions. Surveillance gaps also still exist in the endemic region, highlighted by the time to sampling presented in figure 2.7 and are likely amplified by political instability in Afghanistan and the border regions of Pakistan [97, 112].

We see continued importance of the Karachi, Central Pakistan and South Corridor regions to spread of the virus into other regions in recent years. Despite only a single ES detection in these three regions since July 2021, it is important to continue targeting vaccination to these south-western regions to prevent reintroduction and long-term circulation between and from these regions, a key driver of poliovirus transmission in the past decade. The Central Corridor region supports the only remaining persistent transmission lineage as defined in this study, with changes in transmission dynamics in other regions towards more rapid detection and smaller chains of transmission. Continued circulation is often linked to long-term under vaccinated and hard to reach sub-populations, which can be masked when coverage data are aggregated. This is highlighted by recent inference of 98% population immunity from vaccination campaign coverage data despite the persistent transmission seen in this study [114]. The North Corridor, particularly on the Afghanistan side, has also acted as a key poliovirus reservoir over the time period covered here and continued to support two LTLs in June 2022, one of which was exported and detected in the adjacent East Pakistan region. The long time to detection in the Central and North Corridor and links to other regions suggest they remain key target regions for increased surveillance and vaccination.

The highlighted corridor regions and Karachi have also been previously linked to high-risk mobile populations (HRMPs), presenting additional challenges for vaccination targeting [97]. Together with interruptions to vaccination campaigns associated with the COVID-19 response and violence against vaccination staff in recent years, a number of barriers to maintaining population immunity and eventual poliovirus eradication remain [115, 116]. Recrudescence of polio in hard to reach communities was a major factor in the delay of certification of WPV1 eradication in Nigeria and considering these communities is key when pushing towards eradication

[27, 28]. The ancestral linkage information and break-out of the phylogenetic tree into chains of transmission in this study can help to identify otherwise invisible movements and shed light on lineages which circulated for a long time prior to detection, such as those currently circulating in the Central and North Corridors. Here we calculate a simple time to sampling metric with the inferred LTLs, the time to detection and response will be longer than this due to the time taken for transportation, culture and sequencing of samples as well as reporting and response delays [117].

Our analysis also highlights the additional information provided by performing routine surveillance of wastewater systems over relying on symptom-based AFP surveillance alone. As AFP is a rare outcome, some regions can support transmission of poliovirus whilst not reporting symptomatic cases [30, 31]. The analysis including ES data presented here highlights the importance of regions reporting few AFP cases, such as East Pakistan and Karachi, to supporting chains of transmission and exporting virus to other regions. Despite the utility of ES where few symptomatic cases are seen, the fixed nature of the surveillance points and the lack of standardisation of the site sensitivity in terms of population density, catchment area, water flow rates, and other factors, mean that care needs to be taken when interpreting ES data. Comparison of the analysis of the complete data and AFP only data suggests that the ES data is not introducing a large amount of additional bias in the model, a risk with the DTA model which assumes the number of samples in a region is directly proportional to the number of cases. It would still be an interesting and important area of future work to incorporate ES and AFP data as separate data sources to account for sampling differences in both, whilst retaining the quick and simple DTA framework used here. ES can give an early warning of expanded transmission where circulation is not occurring at a high enough rate to be detected via AFP surveillance and it will become an increasingly key resource as we reach the end point of the poliovirus eradication campaign in the coming years.

A number of simplifying steps were made in this analysis to speed up the pipeline. The speed of analysis is important to allow real-time feedback to regional stakeholders in the eradication

programme and maintaining speed with increasing numbers of sequences can be challenging. The consistent molecular clock of poliovirus lends itself well to quick and simple phylogenetic analysis [45]. The Markov Jump DTA method used here, when compared to the standard DTA, allows us to reconstruct numbers of transitions between locations and the support for these transitions in the model. Producing an empirical tree distribution over which to apply the Markov Jump method provides a considerable speed increase over a standard Bayesian DTA model where the phylogeny is inferred and adjusted at each step in the Markov chain, bringing the computational time down by around 45% to 2 hours per million steps. The chains are also quicker to converge, with little burn-in required, meaning a smaller number of steps can be used. This empirical tree distribution is an appropriate simplifying assumption with the DTA as the movement events inferred have no effect on the phylogenetic tree structure. This framework can also be used to estimate a complete history of jumps between regions, although coming with considerable additional computational cost, allowing time and likelihood distributions to be estimated for each individual transition between regions. This may be of interest in the future when investigating important transitions such as the movement from Southern Pakistan to the non-endemic Malawi-Mozambique border region. Using discrete locations is an additional simplifying assumption, with rate reconstruction directly dependent on the borders of the regions included. Here we attempt to describe epidemiologically distinct regions of interest but arbitrary borders between regions based on administrative boundaries will always introduce additional sources of possible bias depending on the segregation or combination of linked populations [118, 119]. The treatment of these discrete populations in this analysis is also at odds with widely used population movement models such as the gravity and radiation models which consider distance between locations and population density when reconstructing movement probabilities. It would be useful to quantify these biases with continuous phylogeography or by incorporating a more complex transition model between the discrete blocks.

Through investigation of the poliovirus sequence data with simple and fast to implement phylogeographic methods, and by producing effective visualisations and summaries of these data, we can provide additional epidemiological information over what can be gained from current

routinely implemented methods, linking clusters of viruses on genetic similarity alone. The speed of analysis and interpretability of the information provided was a particular focus in this analysis as we aim to produce a routinely implemented analysis pipeline which can provide succinct information of use to stakeholders and decision makers planning future vaccination and surveillance responses.

2.6 Conclusion

This analysis corroborates previous analysis of WPV1 transmission in Pakistan and Afghanistan with both epidemiological and sequence data, supporting the importance of corridor regions between the countries as key WPV1 reservoirs. It also highlights the die out of a major cluster of poliovirus transmission in the region, which was last seen in early 2021, improving time to detection and smaller geographic lineages of related virus, all indicators of progress towards eradication of wild poliovirus.

Chapter 3

A simple correction to adjust for sampling biases in phylogeographic discrete trait analysis

3.1 Abstract

Discrete trait analysis (DTA) methods are widely used to infer patterns of movement between discrete geographic locations from genetic sequence data as they are simple and fast to implement. With the increasing interest in phylogeography for pathogen genomic epidemiology, accounting for and understanding the limitations of the methods used is becoming increasingly important. Applying DTA to location makes several strong simplifying assumptions, including the assumption of equal sampling probability across locations. This may lead to biased estimates of migration rates and ancestral locations where sampling rates vary. Here we investigate these biases using simulated migration and transmission data. We propose a simple correction to DTA for unequal sampling probability across locations when sampling rates are known or can be estimated. This correction to the migration rate matrix can be applied only to phylogenetic branches leading to sampled tips or to the whole phylogeny. We assess the performance of this correction using simulated data for a range of endemic transmission scenarios with varied sampling and migration rates. In a simple, two-location, case the correction performs significantly better than a standard DTA method in three of the four sampling rate distributions investigated. Applying the correction across only the tips of the phylogeny and over the full depth of the tree gives similar results, with the correction only at the tips performing marginally better in this simple two-location simulation. As migration is a key driver of infectious disease spread, improving estimates of migration with simple phylogenetic methods is a valuable development in infectious disease phylogeography.

3.2 Introduction

With the increasing availability of high quality and geographically diverse genetic sequence data from emerging and endemic diseases, phylogeographic analysis is becoming increasingly popular to estimate likely sources and routes of viral transmission. Recent studies have applied phylogeographic methods to study Influenza, Rabies, HIV, Ebola, Zika and the SARS-CoV-2 pandemic, amongst other viral infections [73, 74, 76, 120–125]. Observed viral sequences are often annotated with location information and phylogeographic research aims to combine this information with the inheritance information provided by the genetic sequences themselves to infer the likely location of unobserved ancestral virus and the rate of movement of virus between locations. A number of methods have been developed to reconstruct ancestral states on a fixed phylogeny [71, 126]. Traditionally, this mapping was carried out using parsimony methods, minimising the number of state transitions required to explain the observed traits at the tips of the phylogeny [126]. This can be misleading where rates of migration are high or state exchange probabilities are unequal, both important considerations in infectious disease phylogeography [127, 128].

A simple and widely used alternative is discrete trait analysis (DTA) (also often referred to as 'mugration' as the migration process is modelled using similar methods to those used to model mutation, with each location modelled as a state of the discrete trait) [128]. Using a probabilistic method allows a model of location or 'state' change to be explicitly specified and transition rate parameters to be estimated. DTA does not directly use the location data during inference of the phylogeny, but infers migration along the branch lengths and estimates the geographic location of internal nodes. This allows the evolution of the discrete trait to be informed by both a model of dispersion patterns over time and the structure of the phylogeny previously inferred from the genetic sequence data [128]. DTA methods typically employ continuous-time Markov chain models from the 'Mk' family of Markov models of discrete character evolution [129]. Pagel (1994) was the first to propose the use of an Mk model for discrete trait data with a probability distribution of character states inferred at each internal node on the phylogeny,

the basic structure used for contemporary maximum likelihood discrete trait analysis [130].

Most discrete trait models consider 'reversible' transition rates, where the forward and backward rate of movement between a pair of discrete states are equal, as these are more commonplace in the substitution models already widely implemented in phylogenetic software. The TreeTime mugration method of Sagulenko et al., for example, implements the generalised time reversible model of Tavaré to estimate migration and forms the basis of the internal node location estimates in the popular, open-source, Nextstrain pipeline [76, 131, 132]. In this chapter we consider a non-reversible or 'all-rates-different' (ARD) Mk model applied to a single discrete trait, location. The aim of employing a non-reversible model is to better capture asymmetric migration patterns. This has an associated cost of doubling the number of transition rate parameters which need to be inferred over a reversible model.

When modelling human movement in epidemiology it is common to consider either the radiation or gravity models, where movement of individuals is incorporated as a directional process [133]. These models were developed to account for short-term movement of individuals such as commuting and retail, key drivers of the spread of infectious diseases between populations. This directionality is also an intuitive property of medium-term human migration with certain regions gaining and losing population over time within a country [134]. This differs from traditional molecular ecology where we are considering the development of traits over 'ecological time', measured in generations, and drivers of individual movement are of less importance to trait reconstruction [135]. As the spread of human viruses between populations is driven primarily by the movement of individuals between locations, it follows that movement of virus between populations will be similarly directional. Directional discrete trait models are implemented in the ace function from the R phylogenetics package ape and are supported for discrete trait reconstruction in the Bayesian phylogenetic software BEAST [106, 110, 136].

Although DTA is computationally efficient and easy to implement with a simple maximum like-

lihood implementation of Felsenstein's pruning algorithm, the method was designed to model the substitution of alleles at a genetic locus. Here we instead aim to discern the movement of individuals between discrete locations [71]. The primary difference between the two is that phenotypes used in traditional discrete trait models are directly linked to the underlying genetic data informing the phylogeny whereas genetic data has no direct causative effect on the location of a sample.

Assuming that migration is analogous to mutation inherits assumptions from mutation models at odds with traditional models of migration, outlined succinctly by De Maio et al. [61]. Briefly, DTA assumes i) the shape of the phylogeny is independent of migration, ii) sampling intensity in each state is proportional to the infected population in that state at the time of sampling, iii) prevalence of states can drift with states able to become fixed or extinct over time, iv) any states not sampled at the tips are not able to be inferred as ancestral locations. These factors can lead, amongst other issues, to biased estimates of migration rates and the location inferred at internal nodes where there is unequal sampling between states [137].

Alternative phylogeographic methods which are better suited to structured populations (such as the spatial and temporal subdivisions present in epidemiological models) are also becoming more widespread. One such model is the structured coalescent, introduced by Hudson and Notahara in 1990 and based on the earlier coalescent method of Kingman, 1982 which is implemented in a range of phylogenetic software [55, 57, 138, 139]. The structured coalescent is more appropriate than discrete trait analysis for structured populations as it incorporates fixed effective population sizes, allowing sampling rates to differ across populations, and allows the population structure to affect the shape of the inferred phylogeny. The structured coalescent can still result in improper posterior densities for migration rate and population size parameters where sensible bounds are not given or uninformative priors are used for demographic parameters [58, 140]. Full structured coalescent models can also be slow to, or fail to, converge where more than three or four discrete states are used, which can also be an issue when incorporating complex population structure [58, 63]. For this reason, and to incorporate predictor

variables, a number of approximations and extensions to the structured coalescent have been proposed in recent years. These include the approximations proposed by Volz, De Maio et al. and Muller et al. which avoid the sampling of migration histories in the structured coalescent by integrating over possible migration histories [59, 61, 63]. Differences between these approximations are described in Muller et al. 2017 [60]. Despite these simplifying assumptions, there are still limitations to the numbers of sequences and locations which can be considered with these methods, in the hundreds of sequences and fewer than 10 discrete location states. This precludes their application to large datasets and complex populations. Both the DTA and the marginal approximation of the structured coalescent proposed by Muller et al. have also been extended to allow additional migration predictor information to be provided with a generalised linear model [63, 72]. This can improve inference estimates where demographic or distance parameters, for example, affect the movement probability between locations.

Other methods such as the birth-death process and the generalised coalescent can be used to combine epidemiological models of infection and recovery with genetic data and analyse phylogenies constructed with data from structured populations [59, 141]. These both allow more complex models of birth, death, migration and sampling to be specified, although these models may also be biased and can be slow and difficult to implement. Alternatively, location data can be treated as continuous rather than discrete information, which has the benefit of not requiring homogeneous discrete regions to be defined [68, 142, 143]. Defining these regions is a key step in any discrete phylogenetic approach, with results strongly dependent on the regional boundaries and extent of the study area considered [118, 119]. As DTA continues to be used widely due to the speed and scalability of the method to large numbers of states, context-specific amendments are important to produce reliable results. Here we propose a simple change to the transition matrices central to the likelihood calculations used in discrete trait analysis to correct for unequal sampling rates between locations.

3.3 Methods

Simple maximum likelihood discrete trait analysis uses k-state Markov models (Mk models) of migration between locations [129]. Using a Markov transition matrix and Felsenstein's tree pruning algorithm it is possible to estimate the likely locations of internal nodes in a phylogeny. This process treats migration akin to mutation of a discrete trait, hence the name 'mugration' is often used to describe discrete trait analysis applied to migration [61, 71, 144]. These Mk models are traditionally used in molecular ecology as genetic substitution models, with the simplest equal rates and equal frequency model introduced by Jukes and Cantor in 1969 [145]. Here we consider the ARD Mk model where the migration rate from location i to j can differ from the reverse rate from j to i. Models of this type are not typically used as substitution models, although they were introduced for this use by Barry and Hartigan in 1987 [146, 147]. Statistical identifiability issues for tree generation with the non-reversible general Markov model were described by Zou et al. in 2011 [148]. In the case of DTA, as we are considering changes in a discrete trait independently of the process of tree generation, the non-reversible model can be used and is offered as a method in popular phylogenetic software [106].

3.3.1 General methodology

In an Mk model, the instantaneous rate of change between state i and j is denoted by q_{ij}. General Markov models of discrete character evolution with k states can be represented with a transition rate matrix, $\mathbf{Q}$ [129, 149] (Eq. 3.1):

$$\mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1k} \\ q_{21} & q_{22} & \cdots & q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k1} & q_{k2} & \cdots & q_{kk} \end{bmatrix} \tag{3.1}$$

Where each element on the main diagonal is equal to the negative of the sum of all other elements on the same matrix row ($q_{kk} = -\sum_{i \neq k} q_{ki}$). The transition rate matrix can be used to compute the transition probability matrix $\mathbf{P}(t)$ for a time interval t by taking the

matrix exponential [150](Eq. 3.2). The resulting matrix ($\mathbf{P}(t)$) is a right stochastic Markovian transition matrix.

$$\mathbf{P}(t) = e^{\mathbf{Q}t} \tag{3.2}$$

The likelihood for an ancestral node is given by the product of this transition matrix ($\mathbf{P}(t)$) and a vector (v_{desc}) corresponding to the location probability of each descendent node. In the case of a tip, where the sampling location is known, this probability is 1 for the sampled location and 0 for all other locations. As each ancestor has two descendants, we need to calculate $\mathbf{P}(t) \cdot v_{desc}$ for each descendent branch. The product across these two branches gives the overall likelihood at the ancestral node. Once the ancestral node likelihood is assigned the descendent branches from that node can then be 'pruned' and the likelihood calculated recursively up the tree using Felsenstein's pruning algorithm [71]. The overall tree likelihood is calculated by taking the product of these values across all internal nodes.

3.3.2 Proposed corrections

We propose and test the performance of four corrections to attempt to account for sampling rate differences between locations in DTA. When testing the corrections, we consider sampling rate as the proportion of all currently infected individuals sampled per unit time at endemic equilibrium. The four corrections use a similar method to the equilibrium base frequency vector supplied in the generalised time reversible DNA evolution model of Tavaré [131]. These four proposed corrections are made up of two related corrections applied only at branches leading to the tips of the phylogeny (Fig. 3.1) or over the full depth of the phylogeny, as in the GTR model of Tavaré.

The first correction applies the sampling rate directly, with a diagonal matrix of sampling rates multiplied to the Q matrix via matrix multiplication (Eq. 3.3).

$$\begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1k} \\ q_{21} & q_{22} & \cdots & q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k1} & q_{k2} & \cdots & q_{kk} \end{bmatrix} \cdot \begin{bmatrix} \delta_1 & 0 & 0 & 0 \\ 0 & \delta_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \delta_k \end{bmatrix} \tag{3.3}$$

where each of the k rates of moving into a location i (q_{ki}) are multiplied by the corresponding sampling rate in that location (δ_i). We then set the diagonal elements to the negative of the sum of all other elements on the same matrix row (Eq. 3.4).

$$\mathbf{Q} = \begin{bmatrix} q_{11} & \delta_2 q_{12} & \cdots & \delta_k q_{1k} \\ \delta_1 q_{21} & q_{22} & \cdots & \delta_k q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_1 q_{k1} & \delta_2 q_{k2} & \cdots & q_{kk} \end{bmatrix} \tag{3.4}$$

$$where : q_{kk} = -\sum_{i \neq k} \delta_i q_{ki}$$

The second correction takes the additional step of normalising the sampling rates. We normalise the sampling rates to sum to one, akin to the treatment of equilibrium base frequencies in the GTR model. As the magnitude of the sampling rates is already reflected in the tree structure; with branch lengths and internal node positions in time influenced by the number of samples taken per day, this normalisation is designed to reduce the order of magnitude differences in sampling rates between scenarios to a set of relative rates on the same scale across scenarios. As all sampling rates are positive, the normalisation is carried out by dividing the rate in each location by the sum of rates across the k locations (Eq. 3.5) to give normalised sampling rates $\rho_1 \cdots \rho_k$.

$$\begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1k} \\ q_{21} & q_{22} & \cdots & q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k1} & q_{k2} & \cdots & q_{kk} \end{bmatrix} \cdot \begin{bmatrix} \rho_1 & 0 & 0 & 0 \\ 0 & \rho_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \rho_k \end{bmatrix} \tag{3.5}$$

$$where : \rho_i = \frac{\delta_i}{\sum_k \delta}$$

The diagonal elements of the **Q** matrix are given by the negative of the sum of all other elements on the same matrix row (Eq. 3.6), as in the previous proposed correction.

$$\mathbf{Q} = \begin{bmatrix} q_{11} & \rho_2 q_{12} & \cdots & \rho_k q_{1k} \\ \rho_1 q_{21} & q_{22} & \cdots & \rho_k q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_1 q_{k1} & \rho_2 q_{k2} & \cdots & q_{kk} \end{bmatrix} \tag{3.6}$$

$$where : q_{kk} = -\sum_{i \neq k} \rho_i q_{ki}$$

These two correction methods are applied either at branches leading to tips, as illustrated in Fig. 3.1, or over the full depth of the tree, giving the four proposed corrections. We propose these two alternative implementations as, where migration and mutation occur at a higher rate than sampling, it holds that the phylogeny produced will have a more diverse sample at the tips than where sampling rates are higher than migration and mutation rates. We posit that where a more diverse sample is collected, and thus the root of the tree is deeper and internal branches are longer, sampling rate differences at the tips will have a greater effect on the inferred location of the nodes directly ancestral to the tips than the deeper tree. Vice-versa where the sample is less diverse, and the root of the tree is less deep in time we suggest that sampling may have a greater effect on the structure of the full phylogeny. Where the correction is applied only at branches leading to tips the general methodology is followed for internal branches, with the **Q** matrix applied as given in equatxion 3.1. Work by Faria et al. applied a similar branch partitioning approach to account for differences between successful host-shifts and dead-end cross-species transmissions in a model of zoonotic infection[151].

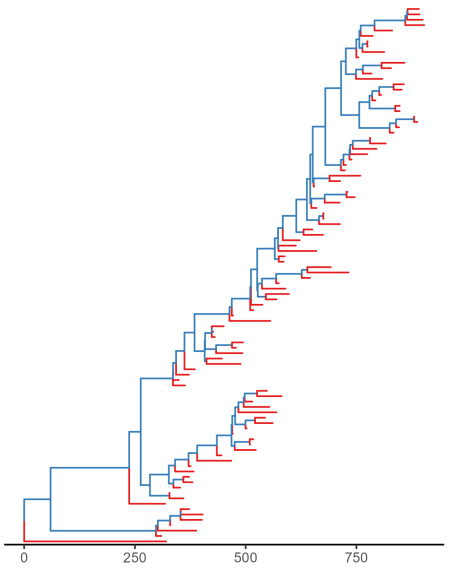


Figure 3.1: Example of a 100-tip phylogenetic tree with the terminal branches coloured red to show the branches over which the sampling rate correction is applied in the 'tip-only' corrections. Internal branches where the Q matrix is left unaltered are coloured blue. The x-axis shows the number of days from the time of the common ancestor (at zero) of the simulated samples (the root of the phylogeny).

Applying the adjustments prior to calculating the diagonal elements of the **Q** matrix preserves the properties of the transition rate matrix as outlined in Norris 1997 [152] (Eq. 3.7), namely:

$$\begin{aligned} 0 \leq -q_{ii} < \infty \\ 0 \leq q_{ik} \text{ for } i \neq k \\ \sum_k q_{ik} = 0 \text{ for all } i \end{aligned} \tag{3.7}$$

Giving rows which sum to zero and no negative off-diagonal or positive diagonal elements.

The proposed corrections are not perfect as sampling events also influence the tree structure and the timing of internal nodes. In the DTA, phylogenies are estimated via a separate process and the topology of the phylogeny is not updated to account for the inferred migration model. This means that the influence of migration on tree topology cannot be accounted for, and this would require a more complex phylogeographic model.

3.3.3 Tree simulation

To test the proposed model corrections, we consider a range of simulated phylogenies with different migration and sampling rates. The simulated phylogenies are produced under a multitype

birth-death (MBD) model using the BEAST2 package MASTER [153, 154].

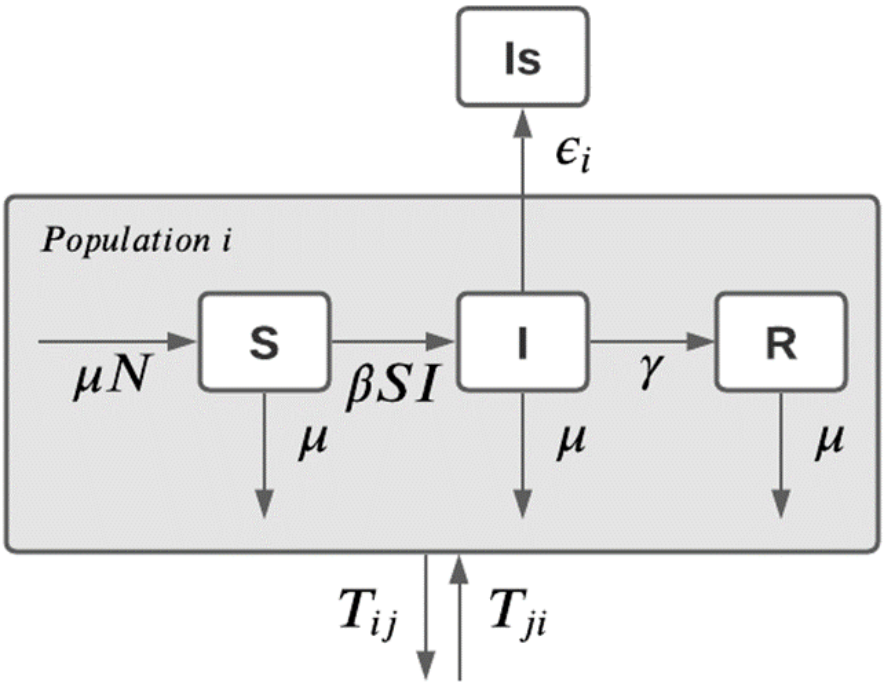


Figure 3.2: Structure of the simple SIR model implemented in MASTER in one of two populations. Additional populations are added as identical modules with additional migration rates to allow movement between all populations. The compartments represent Susceptible (S), Infected (I) and Recovered (R) individuals with the additional Is compartment representing sampling with removal.

A simple epidemic model is simulated with a single infection seeded in one of the locations at time zero. For simplicity we assume that migration occurs at a constant rate over time and that there is no substructure within the states, with all individuals within each state having the same fitness and same probability of being sampled. The model is allowed to progress for 10000 time-steps (days) without sampling to ensure an endemic equilibrium is reached prior to sampling. A sampling compartment is used with no outflows to ensure sampled individuals are an endpoint for transmission. This is a simplification of real sampling dynamics for mathematical convenience, preventing a sampled lineage from being directly ancestral to another sample which would result in a zero length branch. When reconstructing the tree, the time of movement into the sampling compartment is taken as the time of the sampled tip. Recovered individuals are considered immune and no longer contribute to the transmission as infectors or

susceptible individuals. Each sampled phylogeny is taken from an independent simulation and contains 100 sampled tips across the two sampled populations, regardless of sampling rates. We do not simulate genetic sequences for these phylogenies, instead we use only the simulated labelled trees and branch lengths. In other words, we do not account for phylogenetic uncertainty relating to the quality, length, and characteristics of any sequence.

Figure 3.2 shows a flowchart representation of one of the populations included in the MBD model. Rates shown in Figure 3.2 but not in table 3.1 are varied to test the effectiveness of the model corrections under different scenarios outlined below.

Table 3.1: Input parameter values fixed across models. Beta and gamma are set to give an R0 of 6 and infectious period of 10 days and the birth/death rate is set in days to give a life expectancy of 68 years.

Parameter	Meaning	Value (per day)
μ	birth/death rate	0.00004
β	transmission rate (per susceptible individual per infectious individual)	0.6
γ	recovery rate	0.1

3.3.4 Modelled scenarios

For ease of computation and understanding, we test two-state implementations of the correction methods. Further tests with more states will be required to further validate the correction methods for use with greater numbers of states, the primary use case for DTA. For each of the two-state MBD models we consider a range of relative migration and sampling rates between the two modelled locations. We test relative rates up to 4 times higher in each location over the other by randomly selecting migration rates between 0.005 and 0.05 per person per day to satisfy the relative rates we are aiming to reconstruct.

To test the methods under different sampling scenarios we vary the sampling and transmission rates. These scenarios are the "high sampling rate" scenario, where the sampling rates are an order of magnitude higher than migration rates; the "medium sampling rate", where the

migration and sampling rates are on the same order of magnitude; the "low sampling rate", where the sampling rates are an order of magnitude lower than the migration rates; and the "transmission difference" scenario where sampling rates are the same as in the low sampling rate scenario but transmission in population 2 is twice as fast as in population 1, leading to a larger number of infected individuals at equilibrium in this population (Table 3.2). This covers a realistic range of relative migration and sampling rates between populations.

Table 3.2: Migration and sampling rate ranges for each of the modelled scenarios.

Model scenario	Migration rate range	Sampling rate range
High sampling rate	0.005 to 0.05	0.05 to 0.5
Medium sampling rate	0.005 to 0.05	0.005 to 0.05
Low sampling rate	0.005 to 0.05	0.0005 to 0.005
Transmission difference *	0.005 to 0.05	0.0005 to 0.005

* The transmission difference scenario has identical migration and sampling rates to the low sampling rate scenario but different transmission rates between locations (Transmission difference: $\beta_1 = 0.6$ and $\beta_2 = 1.2$; Low sampling rate: $\beta_1 = \beta_2 = 0.6$)

3.3.5 Fitting the model

We test the adjusted model by adapting the 'ace' function from the R phylogenetics package 'ape' [106]. This method uses maximum likelihood optimisation which, although simple and fast to implement, can often have issues with the Mk model. Previous work has reported a flat or ridged likelihood surface for a two-state model, making it impossible to reconstruct exact values for the q parameters [61, 149]. As we find similar issues with the reconstruction of directional migration rates, we reparameterise the migration model to estimate a relative migration rate between a pair of locations rather than attempting to estimate each migration rate. In order to estimate a relative rate directly from the likelihood maximisation, we amend the Q matrix (Eq. 3.8):

$$\mathbf{Q} = \begin{bmatrix} 0 & q_{21} \cdot x \\ q_{21} & 0 \end{bmatrix} \tag{3.8}$$

Giving a base rate of migration from location 2 to location 1 (q_{21} in eq. 3.8) and a rate relative to this base rate for the reverse migration (x in eq. 3.8 = $\frac{q_{12}}{q21}$). This minimises convergence issues resulting from collinearity between forwards and backwards rates as the relative rate remains consistent for a range of base rates along a ridge, meaning a constant relative rate should be returned as long as likelihood maximisation results in a point on the ridge. Although this change does not lead to a convex likelihood surface, the ideal scenario, it does allow for comparison between model fits (Fig. 3.3). This change is carried out prior to applying corrections and setting the diagonal elements, as outlined in equations 3.3 to 3.6.

As relative rates have a lower bound of zero and, where both rates are equal, have a value of 1, the natural logarithm of the relative rate will approximately follow a normal distribution. We perform likelihood maximisation on the log-relative-rate, allowing standard errors to be estimated on the normal distribution, resulting in correctly distributed approximate confidence intervals when transformed back to a linear scale. This transformation has the added benefit of a $\pm\infty$ scale, meaning a wider range of likelihood maximisation techniques can be used. We can reconstruct the approximate standard error for this log-relative-rate from the hessian of the maximum likelihood region of the likelihood surface. Here we use the Hessian method implemented in the 'optim' function from the 'stats' R package. Unfortunately, the eigenvalue-based solution used assumes a convex likelihood surface around the maximum likelihood point estimate and can fail to reconstruct a standard error where the likelihood surface is flat in one or more directions, as is the case with the models tested. Where confidence intervals could not be reconstructed with either the corrected or uncorrected methods, or both, the phylogeny was dropped for mean squared log error (MSLE) and confidence interval comparisons.

We assess the ability of the corrections to reconstruct relative migration rates by comparing the MSLE between the true and estimated relative migration rates across all simulations. Significance is assessed with a p-value threshold of 0.001 due to the very large sample size of 10,000 trees used.

3.4 Results

3.4.1 Correction selection

For the corrections using the exact rather than the normalised sampling rate, the overall MSLE was 0.26 applied across all nodes and 0.19 applied only at the branches leading to tips. Where the sampling rates were normalised to sum to one, the overall MSLE was 0.18 applied across all nodes and 0.14 applied only at branches leading to tips. As the correction with normalised sampling rates applied to branches leading to tips performed best across all sampling scenarios, the results from this method are presented in the main text below. The normalised correction applied to the full depth of the tree performed better than either of the non-normalised corrections and is applied to example polio data alongside the tip-only correction in chapter 4.

3.4.2 Likelihood surfaces

The likelihood surface for the migration rate from 2 to 1 and the relative rate of return (q_{21} and x in eq. 3.8 respectively) showed a repeated trend of heavy ridges, where the maximum likelihood relative migration rate is constant across a large range of base migration rates. This ridged maximum likelihood region is brought more into line with the true relative migration rate with the proposed normalised tip-only correction. An example of the performance of the correction for a single phylogeny presented in figure 3.3. In both the corrected and uncorrected cases shown in figure 3.3 the maximum likelihood ridge is consistent for high values of the log-transformed base migration rate, becoming distorted at lower values. This trend is seen over the range of sampling and migration rates investigated, although the effect is more pronounced with more heavily biased rates.

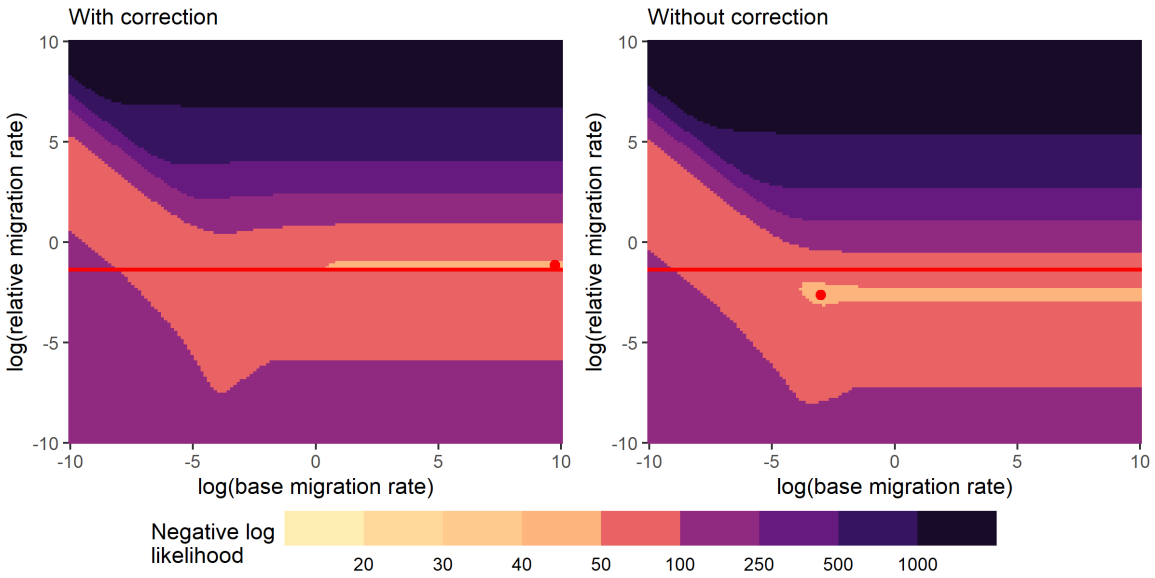


Figure 3.3: Likelihood surfaces for the base and relative migration rate. These are plotted on a log scale as the maximum likelihood estimation is performed on log-transformed parameters. This plot highlights the ridged likelihood surface and the improvement in maximum likelihood estimation with the proposed correction. This surface is for a single tree with simulation migration rates of 0.01 (pop 1 to 2) and 0.04 (pop 2 to 1) and sampling rates of 0.0039 (pop 1) and 0.000975 (pop 2). The surfaces are split into bins, with bounds shown in the figure legend. The lightest colour represents the maximum likelihood (minimum negative log likelihood) region. The horizontal line on each plot represents the true relative migration rate we are attempting to reconstruct ($0.01/0.04 = 0.25, log(0.25) = -1.39$) and the red dot shows the location of the maximum likelihood estimate from the surface.

3.4.3 Relative migration rate estimates

Low sampling rate and transmission difference scenarios

As the low sampling rate and transmission difference scenarios both use sampling rates an order of magnitude lower than the migration rates, we present the results together. The scenarios differ as the transmission difference scenario incorporates a 2-fold higher transmission rate in one population than the other ($\beta_1 = 0.6, \beta_2 = 1.2$) to assess the ability of the method to correct for sampling rate differences where transmission is unequal across regions (Table 3.2). Fig. 3.4 shows the proportion of model runs where the true relative migration rate was within the 95% confidence interval of the estimated relative migration rate under the corrected and uncorrected models. The number of trees from the 100 simulations retained for each parameter set is shown in supplementary figure B.1.

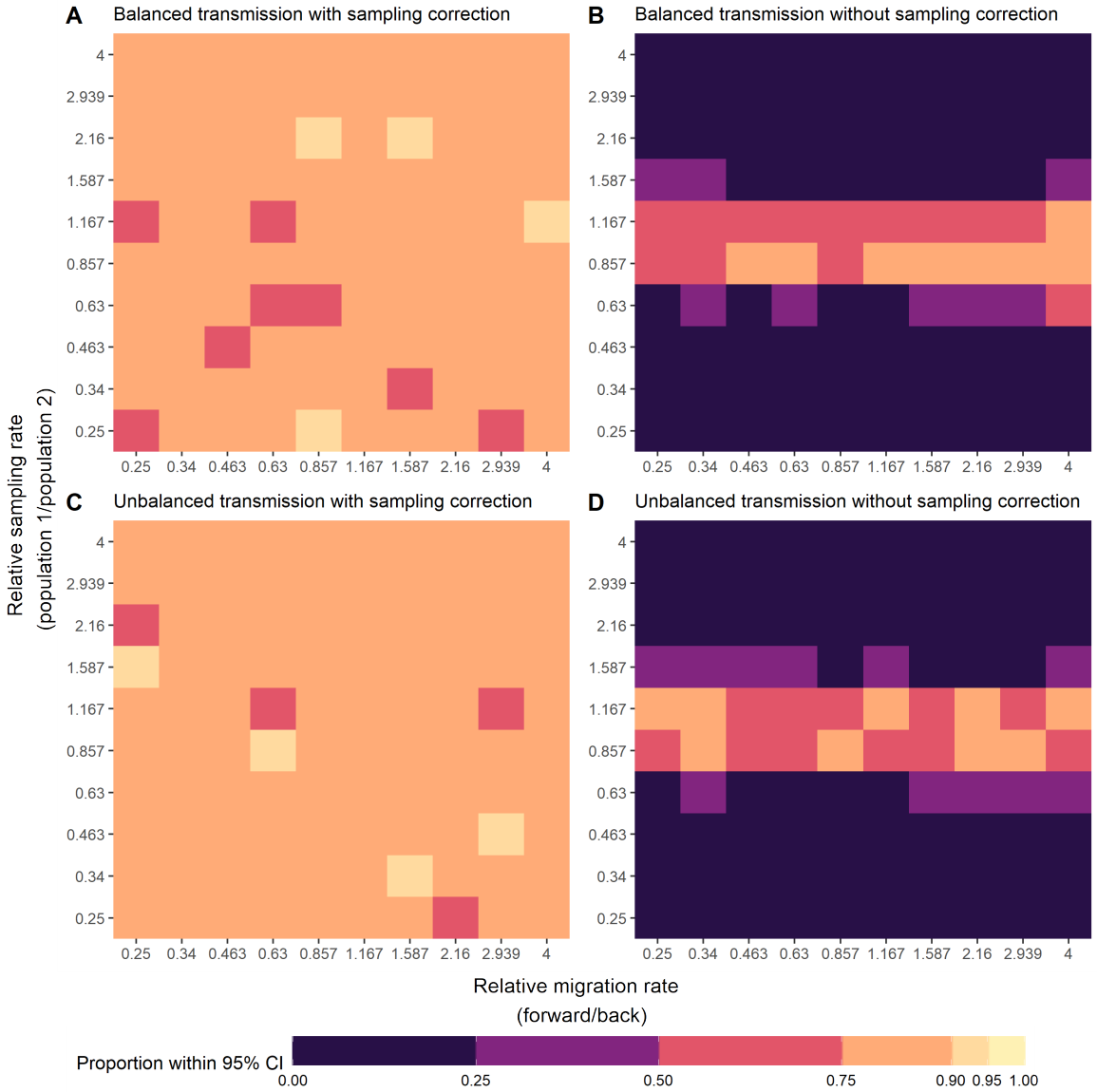


Figure 3.4: Proportion of model runs under the low sampling rate scenarios where the 95% confidence interval of the estimated relative migration rate includes the true relative migration rate from the simulation. We investigate different relative rates of migration and sampling to illustrate the performance of the proposed correction over a range of scenarios. These plots show the prediction under the corrected model (left; plots A and C) and the uncorrected all-rates different Mk model (Right; plots B and D). The top row (plots A and B) shows the performance of the correction and standard model in the simple case, where the transmission rate is equal across the two populations, and the bottom row (plots C and D) the biased transmission case, where transmission is two times higher in population 2 than population 1. Where a confidence interval could not be estimated due to the likelihood surface being flat in at least one direction the tree was dropped from the calculation. The number of trees where confidence intervals could be estimated in each scenario is shown in supplementary figure B.1.

Figure 3.4 shows that there is significant improvement in the confidence interval prediction with the correction. Where transmission rates are equal in the 2 populations, the true relative migration rate is included in the 95% CI of the predicted relative migration rate in the majority of simulations for every scenario with the correction. The minimum proportion of 95% CI predictions including the true value for any migration and sampling scenario is 0.68 compared to 0.00 with the uncorrected model. In the unequal transmission scenario, where transmission is 2 times higher in population 2 than in population 1, the correction continues to perform well with a minimum proportion of 0.73 of the reconstructed 95% confidence intervals including the true relative migration rate. Without the correction, this minimum value is again 0.00.

Under the transmission difference scenario, we see examples where very few tips are sampled from one of the two locations, as illustrated in Fig. 3.5. As migration rate reconstruction with DTA relies on the known location of sampled tips, it is much more difficult to infer a migration rate where one location is heavily under-represented in the phylogeny, but the correction appears to perform well at reconstructing the relative migration rate even in this scenario. In this scenario, however, a greater number of phylogenies are dropped due to inability to reconstruct 95% confidence intervals with both the corrected and uncorrected methods, outlined in supplementary B.

Figure 3.6 shows the MSLE for each modelled sampling and migration rate scenario. As with the confidence intervals, we see an overall improvement in prediction with the corrected method when sampling rates differ between locations. Under the equal transmission rate scenario, we see a highest MSLE of 0.33 with the correction and 2.56 without the correction (Fig. 3.6 A and B). In the transmission difference scenario, the correction performs similarly well, with a maximum MSLE of 0.32 with the correction and 2.43 without. We can also calculate the MSLE across the full range of true and predicted values in order to compare the corrected and uncorrected models. The MSLEs are the same for the transmission difference and low sampling rate scenarios with an overall MSLE of 0.07 with the correction and 0.86 without.

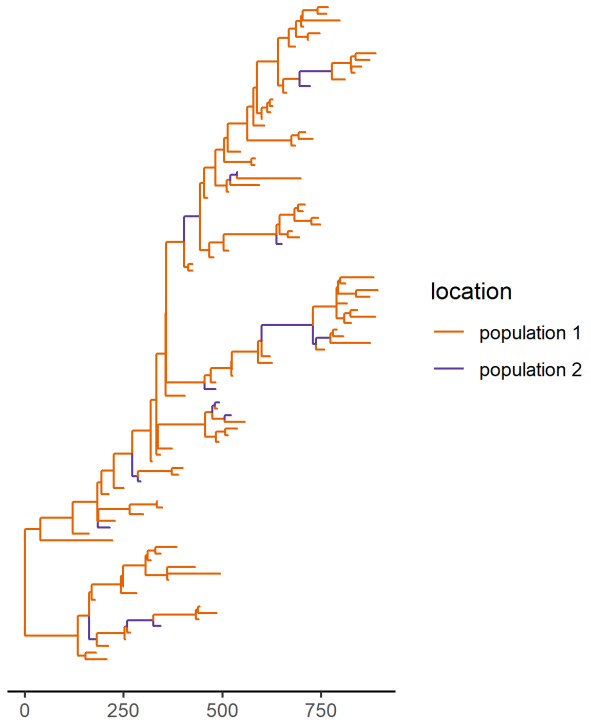


Figure 3.5: Example phylogeny for the scenario where migration from population 1 to 2 is 4 times lower than the reverse rate and sampling is 2.94 times lower (x= 0.34, y =0.25 on Fig. 3.4C) Branches are coloured according to their simulated location. This phylogeny is presented to highlight the lack of samples from population 2, making migration rate reconstruction difficult. The x-axis shows time in days from the root.

With the uncorrected method we see a clear horizontal band on the plots of both the proportion of 95% confidence intervals including the true value and MSLE for the predicted relative migration rates (Fig. 3.4 and 3.6). This suggests that the standard uncorrected method performs well at predicting the relative migration rate and the bounds on relative migration rate where the relative sampling rates are close to one. Further from the equal sampling case, where the differences in sampling rate are influencing the tip distribution of the phylogeny, the uncorrected method has much worse performance whereas the correction continues to perform well.

It is also evident that we see the worst performance with both the corrected and uncorrected methods where the phylogenies have very few tips from one location. This tends to occur most often where the rate of leaving a location is high and the rate of sampling in the same location is low or in the opposite case where the rate of leaving a location is low and the rate of sampling in that location is high (bottom right and top left of the plots in fig. 3.4 and 3.6).

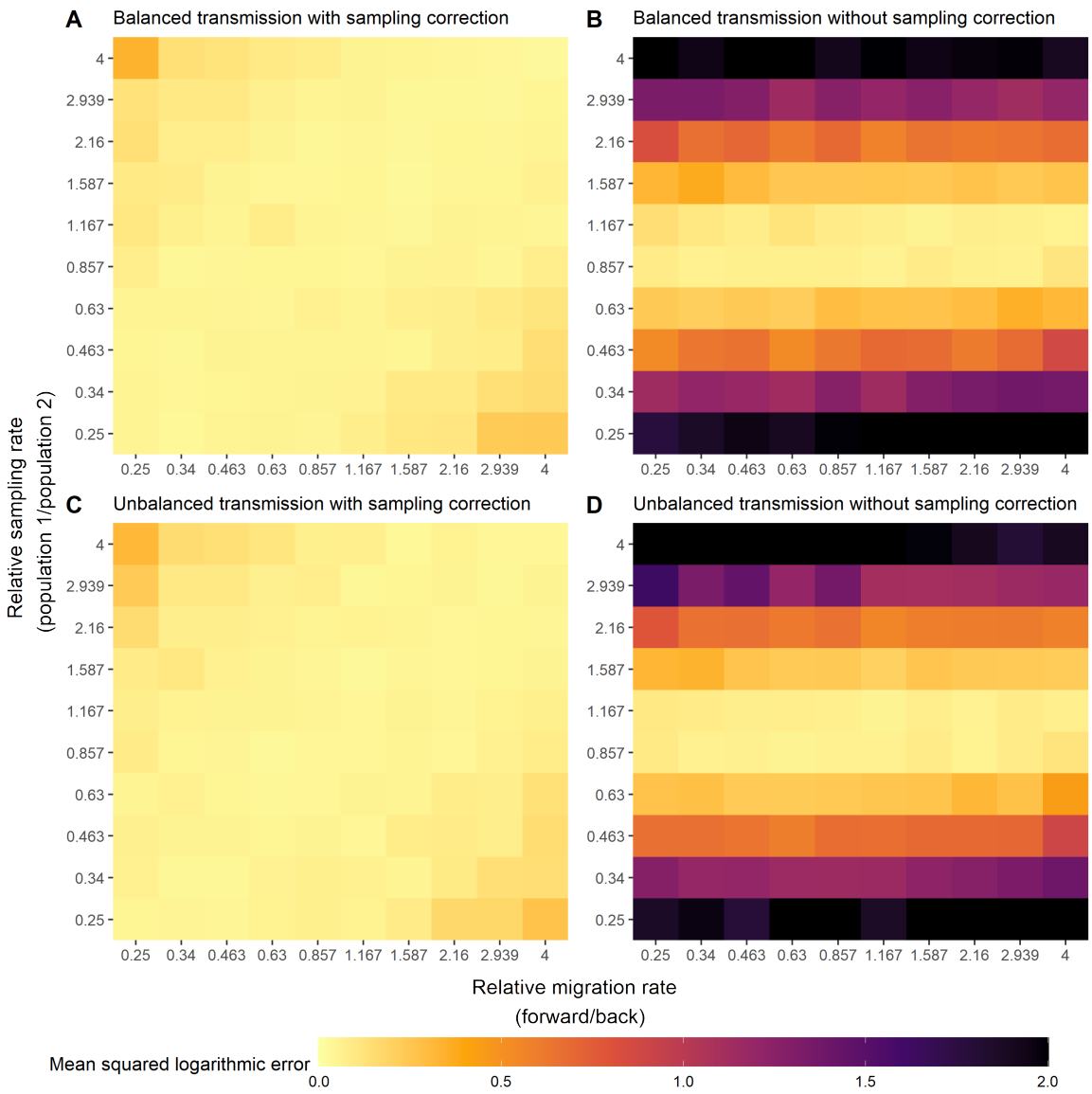


Figure 3.6: Difference between the true and estimated relative migration rates under the low sampling rate scenarios investigated. These plots show the mean squared logarithmic error (MSLE), a measure of the relative difference between the actual and predicted values. The left-hand panels (A and C) show the ability to reconstruct the true relative migration rate with the proposed correction and the right-hand panels (B and D) without. The top panels (A and B) show the performance where transmission rates are equal across the 2 populations and the bottom panels (C and D) where the transmission rate in population 2 is 2 times higher than in population 1. Values of the MSLE above 2 are shown in black on the above plot to allow better focus on the differences between lower values of the MSLE.

Medium sampling rate scenario

Under the medium sampling rate scenario, where sampling rates are on the same order of magnitude as migration rates, the correction performs similarly well (Fig 3.7) to in the low sampling scenarios above (Fig 3.6) with an overall MSLE of 0.08. We see an improvement with the uncorrected DTA in this scenario with an overall MSLE of 0.72. The corrected model performs significantly better than the uncorrected model when comparing overall MSLE ($p < 0.001$). 95% confidence intervals including the true relative migration rate and MSLEs under this scenario are shown in figure 3.7.

High sampling rate scenario

With a high sampling rate, the sampled tips in the phylogeny are very close together in time and phylogenies tend to be rooted more recently than with lower sampling rates. An example phylogeny is presented in supplementary figure B.3. In this case, the uncorrected DTA method is better able to reconstruct the relative migration rate than in the other sampling scenarios presented previously. The corrected (overall MSLE = 0.31) and uncorrected (overall MSLE = 0.29) methods perform equivalently ($p=0.047$). We posit that this is due to the much shallower phylogenies including fewer transmission events and less chance for alternative inference of migration events. In figure 3.8, with both the corrected and uncorrected methods, inference appears better close to the centre of the y-axis, where sampling rates are close to equal and sampling differences have less influence on the migration events. This suggests that sampling rate is still influencing the phylogenies produced. As both methods infer discrete location independently of tree topology, they are not able to correct for any topology differences introduced by biased sampling with high sampling rates.

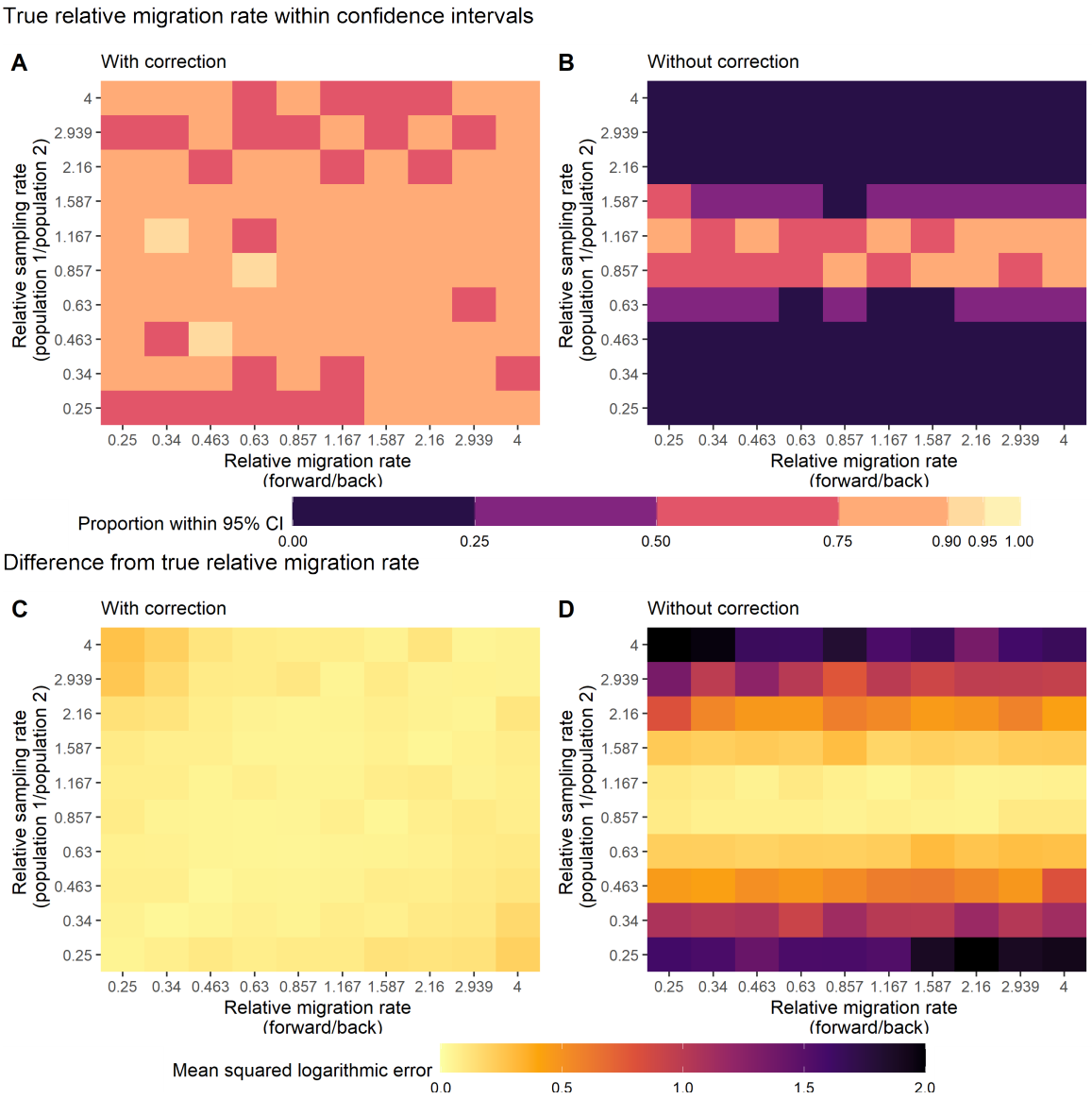


Figure 3.7: Predictive ability of the DTA model under the "medium sampling rate" scenario. A and B) Proportion of true relative migration rates within the 95% confidence intervals of the inferred relative migration rates under the corrected (A) and uncorrected (B) methods. C and D) Mean squared logarithmic error (MSLE), a measure of the relative difference between the actual and predicted values with the corrected (C) and uncorrected (D) methods. Values of the MSLE above 2 are shown in black on the above plot to allow better focus on the differences between lower values of the MSLE.

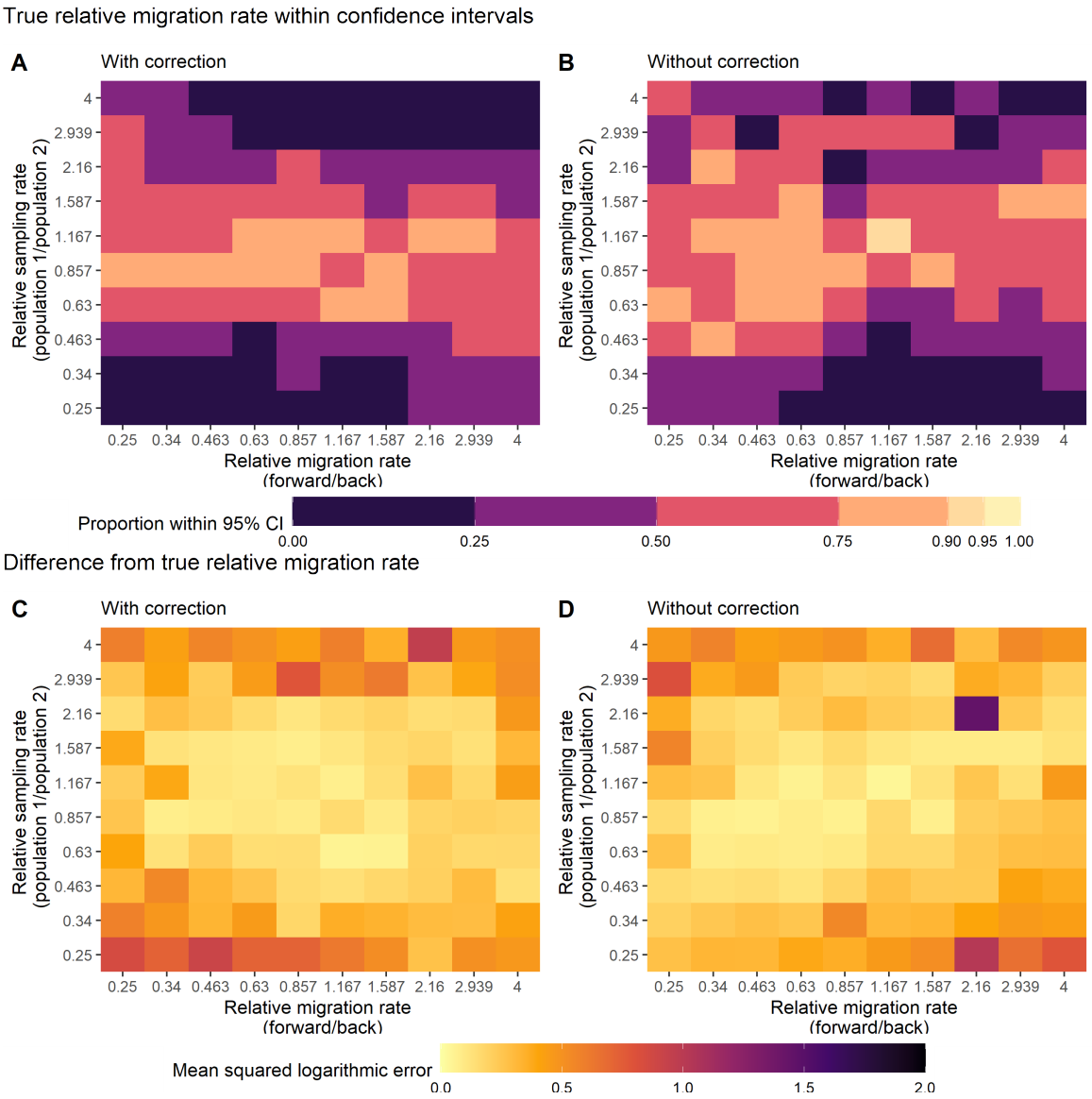


Figure 3.8: Predictive ability of the DTA model under the "high sampling rate" scenario. A and B) Proportion of true relative migration rates within the 95% confidence intervals of the inferred relative migration rates under the corrected (A) and uncorrected (B) methods. C and D) Mean squared logarithmic error (MSLE), a measure of the relative difference between the actual and predicted values with the corrected (C) and uncorrected (D) methods. Values of the MSLE above 2 are shown in black on the above plot to allow better focus on the differences between lower values of the MSLE.

3.5 Discussion

We have developed a correction for discrete trait phylogeographic analysis that improves estimates of migration rates between locations when sampling differs between them, thereby allowing more accurate inference of disease transmission across geographies. This correction can be simply implemented in standard phylogenetic software with minimal increase in computational time (mean of 0.19 seconds to reconstruct maximum likelihood migration rates and ancestral locations on a 100-tip, two-location phylogeny in the ace function from the R package ape compared to 0.12 seconds without the correction). The results presented in this chapter show that the proposed correction can improve the inference of relative migration rates on phylogenies drawn from a simulation of an endemic disease where sampling rates are known, and vary by location. This is an important step as a major issue with the use of discrete trait analysis or mugration models is the assumption that sampling in a state is directly proportional to the underlying number of individuals in that state. In a two state scenario we show that the uncorrected Mk model, the base for standard discrete trait analysis, is only capable of reconstructing relative migration rates where sampling is approximately equal in each state as others have inferred previously [61, 149]. In the most biased low sampling rate scenario, the non-reversible DTA implemented fails to reconstruct a relative migration rate with a 95% confidence interval overlapping the true value in 100% of the 100 simulated phylogenies, we can reduce this to 42% with the proposed correction. This improves when larger sample sizes or larger numbers of phylogenies are considered.

As DTA continues to be widely used in phylogeography, it is important to consider and account for the limitations of the method, in the same way as approximations to the structured coalescent have been provided previously to account for the issues with speed of reconstruction and limitations on the number of states which can be feasibly considered [60]. The proposed correction is a first step, focusing on decoupling the sampling rate from the underlying population sizes. Several additional assumptions still exist with DTA which are at odds with traditional models of migration and can affect the tree structure and inferred parameters [61].

The proposed correction builds upon existing substitution models which account for the equilibrium frequency of bases with a similar method. The most widely used of these is the GTR model and using the weights in the GTR model to account for sampling rate differences when reconstructing geographic discrete traits has previously been suggested as a possibility [155]. The proposed amendment differs from the use of the GTR to model migration and sampling in two key ways: 1) The proposed amendment is based on a non-reversible transition rate matrix and 2) The amendment can be applied only at branches leading to sampled tips of the phylogeny. These differences are designed to better capture migration between discrete states on a short time scale, as is required when modelling geographic spread of viruses.

The correction can be easily implemented with code based on the 'ace' function from the 'ape' R package, including an eigen decomposition method which provides a significant fitting speed increase over performing matrix exponentiation on each fitting run. A Bayesian implementation of the proposed correction is presented in chapter 4 where both the tip-only and whole tree implementations are incorporated in a Markov Jump discrete trait analysis method. In future maximum likelihood use cases, selection of a 95% confidence interval estimation technique which is more robust to ridged likelihood surfaces or a transformation to produce a non-ridged surface would be preferable.

It is important to consider the limitations of the correction and of DTA and phylogeography more widely when selecting a method for different disease transmission scenarios. The method shown here is applied in an endemic scenario, where migration rate parameters are more simple to estimate as other parameters are steady over time. In an outbreak scenario, sampling and transmission rates can change dramatically over a short time period and previous studies have shown it can be hard to reconstruct migration history under these conditions [59]. This has been shown to also be the case with more complex structured coalescent models which have assumptions that can be inappropriate with complex population structures and emerging viruses

and have been shown to perform more poorly than uncorrected DTA even without sampling rate differences [59, 156, 157]. The Mk model implemented here is applied on an empirical fixed set of simulated phylogenies. This highlights another issue with DTA, namely that the location data does not directly influence tree reconstruction and is, instead, inferred over a separately estimated phylogeny. As we consider sampling to have a significant effect on the migration rates estimated between locations, this bias is likely to also influence the depth and number of internal nodes estimated in the phylogeny itself. As a phylogeny only represents the relationships between the samples presented at the tips, any internal branching leading to unsampled lineages, for example, will not be included and unsampled locations will not be inferred at any ancestral nodes with DTA. Likewise, internal nodes along a branch are not included if both descendent lineages are not sampled. This can only be accounted for with a more complex, integrated method such as the structured coalescent or birth-death process. Sampling can also be geographically or phenotypically limited within each discrete region and not representative of the wider population, for example where environmental samples are only collected in a central city or from the outflows of a single hospital. This information can be lost when aggregating up to larger discrete regions and requires careful assessment of the discrete regions used or the treatment of location as a continuous rather than a discrete trait.

Continuous phylogeography offers the ability to infer an ancestral node outside of the sampled regions, an integrated inference of the tree and location, and produces dispersal velocities rather than migration rates between discrete regions which can be of interest when investigating the differential movement of virus clusters or the emergence of new viruses or serotypes [69]. Recent studies have indicated that continuous phylogeography can also be affected by differential sampling in space, although amendments to incorporate less certain spatial location data have been recently introduced [143, 158]. Current widely used continuous phylogeographic methods may be inappropriate for the use case of polio transmission, where distinct seasonal human migration patterns between rural and urban regions, distinct travel routes, religious movement events and nomadic populations within certain regions in the endemic Pakistan-Afghanistan region are at odds with the random Brownian motion viral movement assumption made. Different

phylogeographic diffusion models developed by Barton, Etheridge and colleagues and outlined by Guindon et al. previously, have been implemented in recent continuous phylogeographic approaches [68, 158, 159]. These models are based on a generalised Wright-Fisher diffusion process known as the lambda-Fleming-Viot process, a model of extinction and recolonisation on a spatial continuum. The reverse of this model is the segregated lambda coalescent, allowing multiple lineages to coalesce at one time, rather than only two under the Kingman coalescent.[70, 160, 161].

Sub-sampling is a widely used alternative method to account for sampling biases in phylogenetic analysis. Here only a portion of the available data is included in analysis to produce a more representative dataset. Sub-sampling can be informed by a variety of demographic and epidemiological parameters such as the population size or number of reported cases in a region and recent work testing sampling schemes has been carried out on a set of Bayesian skyline methods [162]. Testing the same sub-sampling schemes on simulated structured population data and comparing the inferences under these sub-sampling schemes to the inferences with the proposed corrections would be of interest in the future, with improvement over sub-sampling vital to widespread adoption of the proposed correction.

Estimating the rate of genetic sampling for human viruses can often be challenging, especially where symptoms are not readily detected or are hard to differentiate from other infections. In reality, it is often not possible to know exact sampling rates and inferring them from collected data can be complex and introduce biases of its own. Where incidence and prevalence estimates can be made, via seroprevalence surveys or epidemiological modelling for example, the proportion of cases sequenced can be estimated and adjusted for by region. These processes introduce additional uncertainty not present under the modelled scenarios presented in this chapter, further limiting the application of the proposed correction to real-world surveillance data.

The proposed correction provides a significant boost to the reliability of mugration model results where the sampling rate is known, particularly on a large number of phylogenies or where a large number of samples have been collected. As the sampling rates used are explicitly stated in the model, and there is little computational cost over standard maximum likelihood DTA, it would make sense to use the correction alongside the standard method to show the influence of potential sampling biases of different magnitudes on inferred migration rates. This comparison across a range of magnitudes of sampling bias between regions has been carried out in work to map the dispersal of HIV using a similar correction previously [163].

An example of a possible use case for the proposed correction is presented in chapter 4, with the phylogeographic analysis of polio sequences from acute flaccid paralysis data. As non-polio acute flaccid paralysis is monitored and reported globally, we expect the global distribution of cases to be approximately even according to population [164]. Differences from the expected non-polio paralysis rate by population size can be indicative of sampling differences and can be incorporated in the proposed correction when considering data sampled from polio paralysis cases.

As the results of DTA applied to migration are widely reported in literature, improving these methods without introducing a large amount of additional inference time or computational cost is important for phylogeography more widely and further research to address other flaws of the DTA method for migration is needed. Recent events also show a strong interest globally in phylogeography for outbreaks and epidemic disease and testing and modifying the proposed correction on simulated epidemic data would be a useful next step in development of more robust DTA models. Going forward it would also be useful to compare the inference under more complex structured coalescent and multitype birth-death methods to the proposed DTA correction, in addition to the comparison to standard DTA model presented here. A recent analysis by Seidel et al., and other previous studies, conclude that the birth-death and structured coalescent models perform similarly well in an endemic scenario with balanced sampling [165]. Comparison of these methods and the proposed correction under an endemic biased sam-

pling scenario would be a useful comparison to make, as well as future extensions to incorporate more complex epidemic dynamics.

The proposed correction, although not able to address all of the issues with the use of DTA for migration rate inference, provides a significant boost to the precision of inferred migration rates in our simulated scenarios. As infectious disease genetic data collection expands, it is increasingly important to outline the limitations of specific methods and to provide updates and corrections to improve their inference ability.

3.5.1 Correction performance

As stated in the results section, the corrections with normalised sampling rates both performed better than the non-normalised sampling rate corrections when tested on simulated endemic data. This is likely due to the gradient descent (negative log likelihood minimisation) method used in this example. As normalising the sampling rates brings the base migration rates estimated (q_{21} in equation 3.8) onto a more similar scale to the relative migration rates for the reverse movements (x in equation 3.8), gradient descent updates are likely less dominated by the base migration rate. It would be interesting in the future to also rigorously test a true relative sampling rate with a reference population although this would require adaptations to existing model implementations rather than borrowing from the widely implemented equilibrium distribution of a DNA evolution model as we do here. The tip-only corrections perform better in these examples although further simulations with more clustered samples would be of interest to test the corrections on more heavily structured populations. We apply both the tip-only and full tree normalised corrections in the next chapter to assess changes in the rates and numbers of movements between locations on real-world endemic transmission data.

Chapter 4

Application of the proposed DTA correction to real-world polio AFP data

4.1 Abstract

The discrete trait analysis (DTA) method applied in chapter 2 makes several strong simplifying assumptions about migration between discrete locations. In this chapter we apply the DTA correction outlined in chapter 3 to real-world polio acute flaccid paralysis (AFP) data, using non-polio AFP rates as a measure of the sampling rate in each discrete location. The maximum likelihood correction presented previously is translated to a Bayesian Markov jump implementation in BEAST, allowing the correction to be applied over the same set of phylogenies used in the AFP analysis in supplementary appendix A. The model is also extended to allow for variable selection, meaning a zero rate of transition can be inferred between an unlinked pair of locations. We find that both the tip-only and full-tree implementations of the correction change the number of movements inferred between locations, with a much greater effect when applied over the whole tree. In this example, both corrections provide estimates of the number of transitions between locations within the credible intervals of an uncorrected DTA model.

4.2 Introduction

Assessing and accounting for the impact of sampling biases on the simple DTA phylogeographic method is a key step for its continued use in fast and simple analysis of genetic sequence data. This is particularly important for use in early epidemic response planning or in the late stage of eradication, such as in the endemic polio example presented previously. In these scenarios speed of response is of utmost importance to contain transmission and determining sources and sinks of transmission quickly can help to target this response effectively. To investigate the potential application of the movement rate correction presented in chapter 2 to real-world

data, we test it on a subset of the serotype 1 poliovirus VP1 data analysed in chapter 3. Our correction method uses data on relative sampling intensity to correct the inference of movement rates. As estimating sampling sensitivity for poliovirus environmental surveillance would require detailed information on catchment populations, sewage flow dynamics and water quality, here we consider only AFP data [31, 166]. AFP affects the spinal cord, indicated by weakness and flaccidity in one or more limbs and data is considered a priori less biased than ES due to this noticeable associated symptomatology [167].

Despite nationwide AFP surveillance being considered the gold standard for poliomyelitis detection, detection rates can differ between regions due to a number of factors including accessibility and resource requirements [102, 168, 169]. As AFP can have a number of causes other than poliovirus infection, rates of paralysis associated with non-poliovirus causes are used in the Global Polio Eradication Initiative (GPEI) as a measure of the ability of the surveillance system to detect polio paralysis were it present [170]. Here we use this non-polio AFP (NPAFP) rate as a proxy for the sampling rate of polio AFP for each of the ten regions of Pakistan and Afghanistan presented in chapter 2. This sampling rate is applied with a Bayesian Markov Jump implementation of the proposed corrections from chapter 3.

4.3 Methods

The DTA correction proposed in chapter 2 is here applied to AFP data from the polio-endemic Pakistan-Afghanistan block from 2012 to 2022. As the sequence data from polio AFP cases used in this chapter have already been described in detail in chapter 2, we refer the reader to that chapter for more details. Reporting of AFP cases is mandatory, and samples are tested for poliovirus at accredited laboratories that are part of the global polio laboratory network. Data from these labs, including the reporting date and location of NPAFP cases is available via the global database of polio data (polio information system; POLIS) [90]. Here we aggregate NPAFP case data from Pakistan and Afghanistan up to the 10 regions used for the discrete trait analysis of polio data in chapter 3 and produce an annual count.

Estimates of population by age group and gender from worldpop.org for 2020 are aggregated to give a denominator of all children under the age of 15 in each region [171, 172]. Annual NPAFP cases in each region were divided through by the number of children under the age of 15 measured in 100,000s to calculate the NPAFP rate. As a simple proof of concept, NPAFP rates are aggregated across the whole time-period for each region and normalised for use as a proxy for the sampling rate of polio AFP cases in the proposed model corrections.

To test the corrections in the most likely use case, they have been translated from the maximum likelihood model in R, presented previously, to a Bayesian model in BEAST. To allow for a non-reversible model with a diagonal multiplier at branches leading to tips, a pair of generalised substitution models were coded for the "internal" (non-tip) and "external" (tip) branches. The non-reversible transition model of Edwards et al. has previously been implemented in BEAST 1 and forms the basis of the implementation used here [136]. We make use of the "frequencies" parameter of the "generalSubstitutionModel" to incorporate sampling rates as a diagonal multiplier on the transition matrix in the "external" model as follows in box 4.1.

Box 4.1: BEAST xml code for a corrected all-rates-different substitution model using the generalSubstitutionModel format. k is the number of discrete regions included and rho1...rhok the normalised sampling rates (rho) from equation 2.5 in chapter 2.

```
<generalSubstitutionModel id="state.external.model" randomizeIndicator="false">
    <generalDataType idref="state.dataType"/>
    <frequencies>
        <frequencyModel id="state.freqs"  normalize="false" >
            <generalDataType idref="state.dataType"/>
            <frequencies>
                <parameter id="state.frequencies" dimension="k" value="rho1,rho2,...,rhok"/>
            </frequencies>
        </frequencyModel>
    </frequencies>
    <rates>
        <parameter id="state.rates" dimension="k(k-1)" value="1.0" lower="0.0"/>
    </rates>
    <rateIndicator>
        <parameter id="state.indicators" dimension="k(k-1)" value="1.0"/>
    </rateIndicator>
</generalSubstitutionModel>
```

A similar Bayesian discrete trait phylogeographic correction in BEAST 1 was proposed in Faria et al. 2012, where the same non-reversible approach from Edwards et al. 2011 was implemented with prior gamma distributions on the transition rates out of under-sampled locations set to give movements 10, 100 and 1000 times less likely than the reverse movements into under-sampled locations [136, 163]. Numbers of transitions inferred under these three scenarios were incorporated into reported estimates of the number of transitions between locations in the study [163].

To apply the uncorrected model at internal nodes an additional substitution model is implemented with the "normalize" parameter of the frequency model (highlighted in Box 4.1) set to true, giving k evenly distributed state frequencies as in the uncorrected DTA method. The internal model uses the same "rates" and "rateIndicators" as the external model in box 4.1. The external and internal models are combined using the "externalInternalBranchModel" evolution model in Beast 1.10.5pre (Box 4.2) [110, 151].

Box 4.2: BEAST xml "externalInternalBranchModel" chunk which allows a different substitution model to be applied to external (leading to tips) and internal (leading to internal nodes) branches.

```
<externalInternalBranchModel id="branchSiteModel">
    <treeModel idref="treeModel"/>
    <generalSubstitutionModel idref="state.internal.model"/>
    <externalBranches>
        <generalSubstitutionModel idref="state.external.model"/>
    </externalBranches>
</externalInternalBranchModel>
```

Where the correction is applied over the whole tree, only the "external" substitution model needs to be specified, without the additional "externalInternalBranchModel" code. XML files for the analysis are available on github (`github.com/JorgensenD/PhD_thesis/`).

The corrections are applied to the same empirical tree distribution used in the MJ-DTA analysis of AFP data (Supplementary appendix A with forward and backward rates of movement between each pair of locations estimated (non-reversible rates)). As we use ten discrete lo-

cations, 90 movement rates were included in the model. Bayesian stochastic search variable selection (BSSVS) is used to limit the number of rates inferred to only those where evidence of movement between the locations was present in the data [110].

To allow the model to converge efficiently, a gamma prior was placed on each of the transition rates with post-analysis of BSSVS indicator variables used to determine which rates and movements have significant support from the data [111]. To test the sensitivity of the results to the gamma rate prior, the analysis was performed three times, with a gamma(1,1), gamma(1,5) and gamma(2,1) prior distribution on the transition rates.

BEAST models were run for ten million iterations on the set of 1000 phylogenies inferred under the model described in chapter 2. Following the estimation of transition rates and numbers of transitions in BEAST, a ten percent burn-in was discarded for each model run. Median transition rates and the median and 95% credible interval for the number of movements between locations were determined using tracer v1.7.1 and Bayes factor supports for these values were extracted with SpreaD3 v0.9.7 [111]. Transition rates and numbers of transitions with Bayes factor support above ten are presented in the results section.

4.4 Results

4.4.1 NPAFP rates by region

NPAFP rates per 100,000 inhabitants under 15 years of age were calculated for each year from 2012 to 2022 (Figure 4.1A) and aggregated to give an overall mean rate across the period (Figure 4.1B). Of the 157,944 NPAFP cases recorded, 787 were dropped from the dataset due to ambiguous location (missing admin 2 information for Pakistan or admin 1 for Afghanistan), as they could not be assigned to one of the 10 regions. We can see from the rates presented in figure 4.1 that there is variation between the regions, highlighting the importance of cor-

recting for sampling even in the case of AFP rates, which are generally considered to be the least biased polio surveillance method. The lowest rates were in the sparsely populated West Pakistan region and the major port city of Karachi in South-West Pakistan. These mean rates were normalised to sum to 1 when applied in the corrected DTA model.

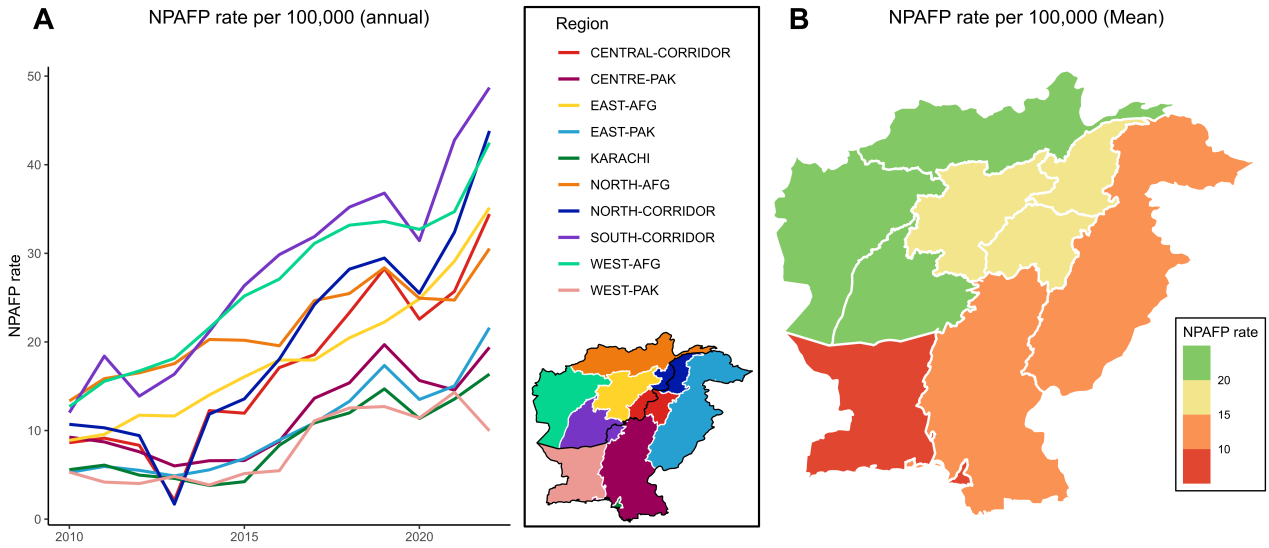


Figure 4.1: **A)** Annual and **B)** mean non-polio acute flaccid paralysis (NPAFP) rate between 2012 and 2022 across regions of the polio endemic Pakistan-Afghanistan block, shown in situ in the legend. The NPAFP rate is calculated as the number of cases per year per 100,000 inhabitants under the age of 15.

4.4.2 Sensitivity to gamma rate prior

Although the inferred transition rates differ under different prior distributions, we found a strong linear relationship between the estimated rates from the full tree model under each of the three transition rate prior distributions tested (A gamma(1,1), B gamma(1,5), C gamma(2,1)) (R2 [A,B] = 0.995, [B,C] = 0.979, [A,C] = 0.976). The median number of inferred transitions between states (A: 189 (95% HPD: 177-202), B: 189 (177-202), C: 187 (176-201)) and median number of non-zero transition rates (A: 35 (95% HPD: 30-41),B: 35 (30-41),C: 33 (27-38)) were not significantly different between the three analyses.

The tip-only correction also returned a strong linear relationship between rates estimated un-

der each of the prior distributions (R^2 [A,B] = 0.996, [B,C] = 0.980, [A,C] = 0.979). Applying the correction only at the tips resulted in a lower number of transitions A: 181 (95% HPD: 169-192), B: 180 (169-192), C: 179 (169-190) and greater number of non-zero transition rates A: 38 (95% HPD: 32-44), B: 38 (32-44), C: 36 (30-41) than applying over the full tree but no significant difference between the three gamma rate priors.

When compared to uncorrected models run under the same prior conditions, the relative movement rates, indicating the change attributable to the correction, were similar under all prior conditions tested. For the remainder of this analysis, rate estimates under the widely used gamma(1,1) rate prior are presented.

4.4.3 Performance of corrections

The uncorrected, tip-only correction, and whole-tree correction were all applied to the same distribution of empirical trees to assess the effect of the corrections on the inferred numbers and rates of transition between locations.

The median inferred rates across the nine million post burn-in iterations are shown in figure 4.2 for each model. 29 transition rates had Bayes factor support above ten without the correction, reducing to 28 with both the correction applied only to branches leading to tips and over the whole tree. The supported rates differed between models and are highlighted in green on figure 4.2. The unsupported rates and their associated movements are considered to be zero and are not included in the remainder of the analysis.

Inferred movement rates under the two proposed corrections relative to the uncorrected model are presented in figure 4.3. This highlights the small rate changes with the correction applied only at branches leading to sampled tips, and larger changes where the correction is applied to the full depth of the phylogenies. All median rates fall within the credible interval of the

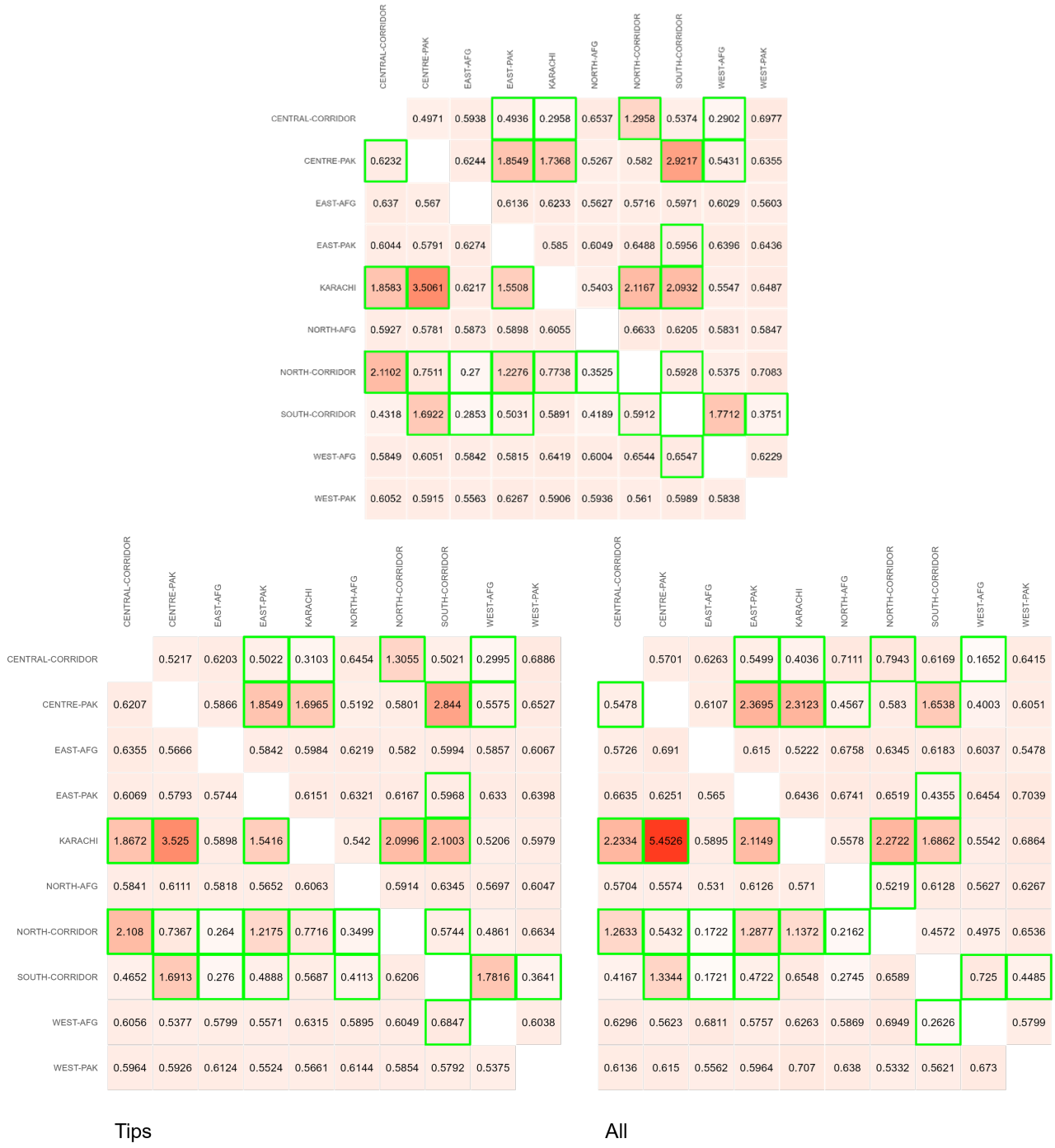


Figure 4.2: Transition rate matrices for movements between the ten included locations under the uncorrected (top) and tip-only (bottom left) and full-tree (bottom right) corrections. Cells are coloured in a scale to show the difference from zero, with darker colour representing higher rates. Rates with Bayes factor support are highlighted with green cell outlines. The remaining cells represent unsupported transition rates and are not considered in the discussion of the results.

uncorrected rate estimates.

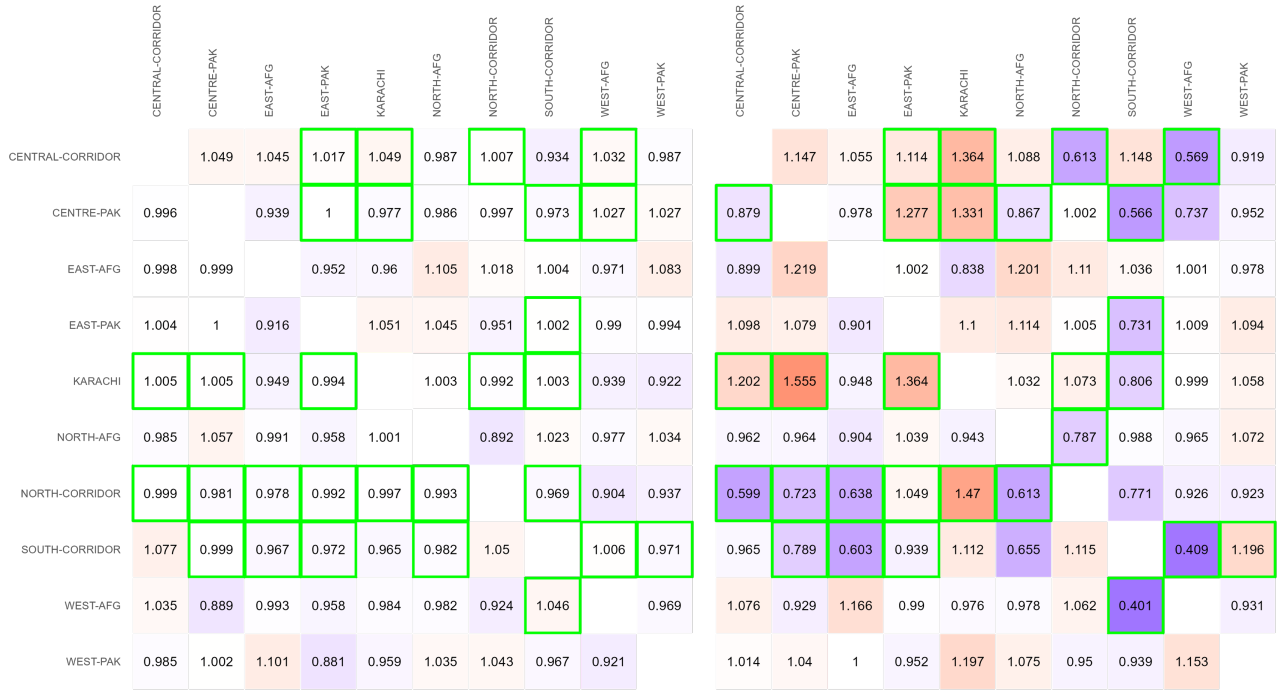


Figure 4.3: Transition rate matrices for the tip-only (left) and full-tree (right) corrections relative to the inferred transition rates under the uncorrected model (Figure 4.2) Blue cells represent a lower relative rate and red cells a higher relative rate with white cells representing parity with the uncorrected model. Darker colours represent a greater difference from parity.

Changes in the number of estimated transitions between locations from the baseline uncorrected model to the corrected models are shown in figure 4.4. There are changes in the number of transitions under both proposed corrections, with one fewer movement from Centre Pakistan to the South Corridor inferred with the correction applied to branches leading to tips of the phylogeny. As expected, a larger number of differences are inferred with the correction applied over the full depth of the tree, with the most noticeable of these being an increase in all of the predicted transitions out of the under-sampled Karachi region (Figure 4.4)(highlighted in dark green). Although the median values for these transitions have changed, all transition estimates fall within the 95% credible intervals of the uncorrected model estimates.

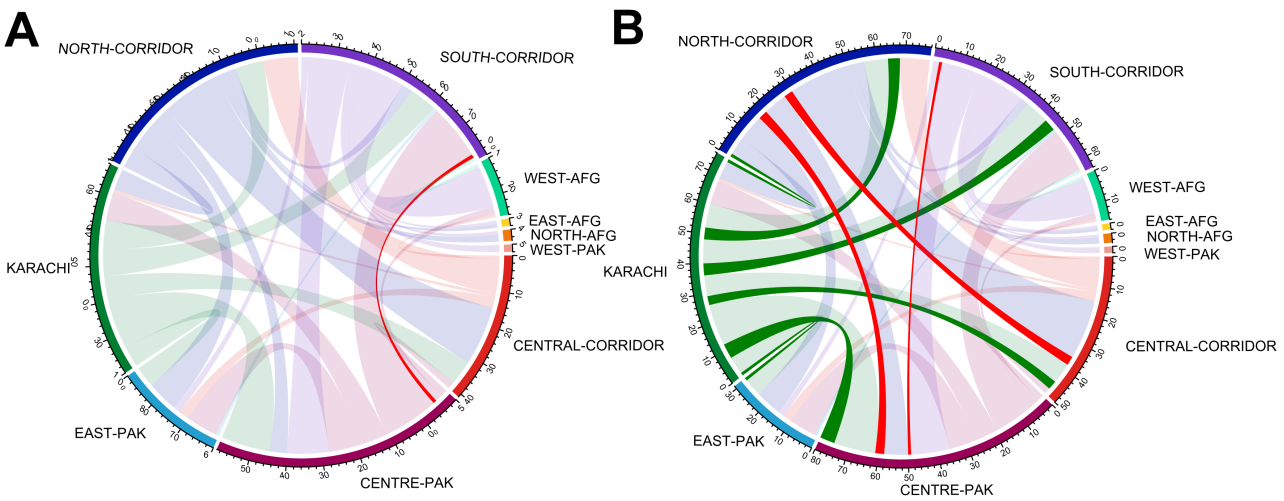


Figure 4.4: Bayes factor supported numbers of movements between regions highlighting changes from the baseline uncorrected model with **A)** the proposed correction applied at the tips only and **B)** applied over the full phylogeny. Dark red lines represent a reduction in the number of predicted movements and dark green lines an increase.

4.5 Discussion

This work shows that sampling rate differences between locations likely do have a noticeable, albeit small, effect on the inference of movements between locations under discrete trait analysis. In this example, where the reported sampling rates are small, on the order of 1 in 10,000, and similar (8.1-28.1 per 100,000, a 3.5-fold difference from highest to lowest) the DTA corrections that we propose do not significantly alter the outcome and interpretation of the analysis. The median numbers of movements between locations, even under the correction applied to the full depth of the phylogeny, fall within the 95% credible intervals of the movement estimates from the uncorrected model. This supports the use of the currently widely used classical DTA approach unless a large difference in sampling rates is suspected, although assessing sampling differences between locations is an important step when applying the method and interpreting the results.

The small difference inferred between the tip-only correction and the uncorrected model is likely partly driven by the modelling process used. Phylogenies in discrete trait analysis are

inferred solely from the genetic relatedness of the samples and do not incorporate additional information provided by the location data. Location inference is an entirely separate procedure, as highlighted by the ability to infer phylogenies and locations in two separate steps in this work. This means that changes to sampling rates in the proposed corrections do not have any influence on the underlying structure of the phylogenies, including the position of internal nodes in time and in relation to pairs of sampled tips, changes which are possible with more complex phylodynamic structured coalescent and birth death models. The real-world data also tended to be more closely clustered by location than the simulations presented in Chapter 2. This means that sampled tips tend to be directly related to other samples from the same location, negating any effect of applying a sampling rate correction only at branches leading to sampled tips. This is another limiting factor of the DTA model used, as only locations sampled at the tips can be inferred at the direct ancestor node. This clustering has less of an effect where the sampling rates-based corrections are applied over the full depth of the phylogeny in the second proposed correction as deeper nodes are less likely to be localised to descendants from a specific region.

Although here we treat the difference seen in NPAFP rates as a proxy for the sampling rate in each location, the difference may not be entirely explained by sampling rate differences. Non-polio causes of AFP include other transmissible infectious diseases which can be unevenly distributed in space, contributing to differences in detection by region. The annual NPAFP data also shows a time trend, with detection rates broadly improving over time in all locations, as well as increasing disparity between high and low detection rates over time. In order to capture these temporal dynamics, it would be interesting to extend the proposed corrections to a more complex epoch model in the future, where sampling rates are able to differ over discrete blocks of time [173].

These proposed corrections show promise as simple quick estimates of underlying dynamics, taking around five hours to run 10 million iterations of the discrete trait estimation process on the 1099 included 906-base sequences, with pre-inferred empirical phylogenies. Further

investigation of tip clustering and epoch models on more complex simulated data would be of interest prior to future application of the methods to real-world data.

Chapter 5

Bayesian phylodynamic analysis of the 2020 SARS-CoV-2 outbreak in Saudi Arabia

5.1 Background

The term "phylodynamics" was coined in 2004 to describe the blend of immunology, epidemiology and evolutionary biology required to integrate genetic data with epidemiological modelling [174]. Genetic data complements routinely collected and analysed epidemiological data and is able to provide information on pathogen diversity and linkage between sampled pathogens not available from other sources. Genetic diversity of collected sequences can also be used as a proxy for prevalence of infections, as we expect more mutation in a larger infected population.

In the previous chapters we have focused on incorporating time and discrete location data into a simple phylogeographic framework. Adaptations can improve the utility of these methods to study short-term viral transmission but, as we and others have highlighted previously, simple phylogenetic methods make a range of assumptions which can limit the utility of the results returned [61]. Phylodynamic methods generally incorporate the viral diversity over time provided by Bayesian phylogenetic reconstructions as an indicator of the size of the infected population at the same time point with various link functions used depending on the method. As with the application of DTA to discrete geographic information, several challenges remain with complex phylodynamic methods and a range of developments have been made to overcome these. Some of the main challenges were outlined by Frost et al. in 2015 [175].

One key challenge, to which a large amount of development has been devoted since the 2015 paper referenced, is accounting for sampling rate differences over time when attempting to reconstruct the effective population size (Ne). The majority of skyline methods are designed to

deal with a single population, meaning they do not need to account for sampling rates which differ between populations, but extensions are required to account for changes in sampling rate over time where population size is not constant. The classic skyline of Pybus et al. (2000), and later revisions which smooth Ne estimates over longer periods, infer Ne over time using the coalescent event times of phylogenies reconstructed from sampled sequences [176, 177]. The Kingman coalescent process underlying the coalescent skyline is asymptotically unbiased for a Wright-Fisher process (with constant population size N) irrespective of sampling rate but, if the effective population size varies over time, sampling intensity can bias the estimates [138, 178].

A variety of parametric and non-parametric models have been developed to extend the skyline to account for sampling over time, outlined by Volz and Frost in 2014 [178]. More recently, the 2016 method of Karcher et al. addresses linkage between Ne and sampling (e.g. when sampling is increased in response to circulation), by modelling sampling times as an inhomogeneous Poisson process with a rate dependent on the effective population size. Gill et al. address the same issue with the incorporation of independent epidemic size covariates in an extension of the Skygrid method [62, 64, 179]. Skyline methods based around a birth-death model rather than a coalescent model have also been implemented, incorporating sampling over time with an explicit parametric sampling rate [65]. Birth-death and coalescent skyline models incorporating a parametric sampling rate (Including Karcher et al. and related methods and birth-death models) can be biased if this sampling rate is misspecified but can improve performance where a correct sampling rate model is specified.

In this chapter we use the Skygrowth method, which allows additional sampling covariates to be incorporated in the same way as the Skygrid method but differs from other smoothed skyline methods due to the first order stochastic process used to smooth the estimates. In Skygrowth this stochastic process is defined in terms of the growth rate of the effective population size Ne rather than Ne itself [180]. The growth rate of Ne is here assumed to have a simple relationship with the growth rate of the epidemic from which the samples are collected (the reproduction

number; R(t))(eq. 5.2) [180]. Since the publication of Skygrowth in 2018, further methods have also been developed relaxing the assumptions made in Karcher et al. to allow for sampling rates to vary across epochs (Parag et al. 2020) and a smoothed adaptive sampling method built on this (Cappello and Palacios 2022)[181, 182].

More complex phylodynamic models have also been developed (using birth-death and coalescent phylogenetic reconstruction methods) incorporating mechanistic infectious disease models. Information on the effective population size can be used to calibrate traditional compartmental epidemiological models, and vice-versa, allowing the model parameters to influence the phylogenetic reconstruction. Integrated phylodynamic models incorporating compartmental models allow information on interventions, timing of events and birth, death, infection and recovery dynamics to be explicitly included and estimated, difficult parameters to implement in GLM-based phylodynamic analyses. A number of these methods implemented in BEAST2 are outlined and referenced in Volz and Siveroni (2018) where the PhyDyn framework, used in this chapter alongside skygrowth, is introduced [183]. PhyDyn extends other implementations, allowing for a structured coalescent model approximation with complex population structure [183].

In this chapter we apply the PhyDyn and Skygrowth methods to the early outbreak of SARS-CoV-2 in Saudi Arabia, reconstructing the effective population size and reproduction number over time with both methods.

5.2 Introduction

Saudi Arabia has a first-wave SARS-CoV-2 transmission profile similar to that seen in much of the world in early 2020, with importations inferred to have begun in late February, leading into a peak in cases in June following national and regional non-pharmaceutical interventions [2]. The first case of coronavirus in Saudi Arabia (KSA) was confirmed on 2nd March 2020,

in a traveller returning from Iran, with the first confirmed transmission within the country reported on March 10th [184]. Following the initial reporting of cases, interventions were put in place quickly, with the closure of Mosques and public places occurring on March 5th. Further movement and business restrictions were introduced throughout March. A number of early virus introductions were centred around the religious cities of Makkah (Mecca) and Madinah (Medina) which attract global visitors year-round. These were the first cities to enter partial lockdowns on April 2nd due to the seeding of onwards transmission locally [185]. Curfews were imposed in all cities by April 15th, tightening around the Eid holiday in late May, and were lifted in all cities by June 21st.

Saudi Arabia has an approximately 84% urban population, making effective restrictions in cities particularly important to reduce the effective reproduction number and slow viral spread [186]. Here we attempt to reconstruct the effective reproduction number in Saudi Arabia using genetic sequence data from viral samples collected over time, to assess how transmission changed with changing non-pharmaceutical interventions. Viral genetic data can give an indication of diversity of circulating virus and, with a random sample, this diversity data gives an indication of the size of the viral population at a given time. This data can be used to reconstruct virus transmission in the same way as incidence data is used in traditional epidemiological modelling.

As case data is often under-reported and can be biased by local testing strategies, estimates of epidemic trajectories from alternate data sources are of interest for outbreak analysis. Here we use viral sequence data but the number of reported deaths was another widely used alternate data source for case comparison between locations as death reporting is more complete than direct case counts in most healthcare settings [187, 188]. Genetic data, although less widely available than death or case counts, was pivotal in the reconstruction of the Australian SARS-CoV-2 outbreak, inferring transmission clusters and effective reproduction numbers before and after the introduction of quarantine measures, and has been a key tool in tracing new coronavirus variants globally [189].

Genetic data on circulating viruses from asymptomatic and hospitalised individuals were collected throughout the first wave of SARS-CoV-2 in Saudi Arabia and deposited on GISAID (gisaid.org) by the national CDC (cdc.gov.sa) and the pathogen genomics laboratory at King Abdullah University (pgl.kaust.edu.sa). This data was primarily collected from four cities in Saudi Arabia (Madinah, Makkah, Jeddah and Riyadh), the focus of this analysis. These data are used here to estimate effective population size and effective reproduction number using Bayesian coalescent approximations.

Rapid analysis with the Skygrowth model implemented in R shows utility in reconstructing general trends throughout the outbreak [180]. A more complex model implemented in PhyDyn incorporating a traditional ordinary differential equation based epidemiological model together with coalescent phylogenetic reconstruction is fitted on smaller scale data from Major cities in Saudi Arabia [183]. The addition of the extra complexity in the form of an epidemiological model allows the introduction of time-bounded transmission rate parameters to investigate changes over time in viral transmission rates. The method presented here is an extension of the model presented previously in a report by Volz et al. from 2020, adapted here to allow R_{eff} estimates to be split at discrete time points as well as fit to seroprevalence data [3, 4, 190, 191]. This model includes a set of international background sequences which seed new transmission chains in the country to give a better representation of the real-world dynamics of SARS-CoV-2 in an internationally connected centre such as Saudi Arabia.

5.3 Methods

A set of full genome genetic sequences of SARS-CoV-2 deposited on GISAID prior to December 31$^{\text{st}}$ 2020 are used throughout this analysis. A total of 892 sequences were collected by the Pathogen Genomics Laboratory at the King Abdullah University of Science and Technology (KAUST) from 5 different areas of Saudi Arabia, the cities of Riyadh (n = 69), Jeddah (n = 260), Madinah (n = 307) and Makkah (n = 227) and the eastern coastal region (n = 29) [192]. All SARS-CoV-2 sequences used in this analysis are publicly available on GISAID

(10.5587/gis8.230619.tc)[192]. A supplemental table describing the data used is provided in supplementary appendix C.

5.3.1 Skygrowth

The model

To give an indication of the progression of the first wave of the SARS-CoV-2 epidemic in Saudi Arabia as a whole, we employ the Skygrowth non-parametric method of Volz 2018 [183]. As mentioned in the background section, this method calculates effective viral population through time based on the relatedness of sampled viruses, from which the growth rate of this effective population size can be estimated. The method assumes a direct link between the growth rate of the effective population size *Ne* and the growth rate of the epidemic, often referred to as r. In a simple SIR model we can calculate the effective reproduction number with the epidemic growth rate (r) and the generation time (ψ) as follows in equation 5.1.

$$R(t) = 1 + \psi r(t) \tag{5.1}$$

The skygrowth model uses this same calculation, replacing the epidemic growth rate with the growth rate of *Ne* (equation 5.2).

$$R(t) \approx 1 + \psi \frac{\dot{Ne}_{(t)}}{Ne_{(t)}} \tag{5.2}$$

Where $\dot{Ne}_{(t)}$ is the derivative of $Ne_{(t)}$ with respect to time and $\frac{\dot{Ne}_{(t)}}{Ne_{(t)}}$ is the growth rate of *Ne*. The model used in this analysis included 35 timesteps and an exponential prior on the smoothing parameter (precision) corresponding to a 1% change in growth per week. The growth rate output was converted to an estimate of R over time using an assumed generation time (ψ) of 9.5 days [193]. The model is fitted from March 5$^{\text{th}}$ to ensure exponential growth is established in the population prior to model fitting as we cannot estimate the growth rate or effective reproduction number when virus circulation is driven purely by importation events.

Phylogenies used

The skygrowth model relies on pre-inferred phylogenies to estimate $Ne_{(t)}$. Here we generated a set of contemporary phylogenies based on the same data to allow smoothed average estimates to be made in the skygrowth model. As the skygrowth model does not allow for population structure, the model could not include additional background sequence data or importations from outside of the region of interest, a strong assumption for SARS-CoV-2 where much of the early transmission was driven by repeat introductions.

Phylogenies for the Saudi Arabian SARS-CoV-2 sequences were generated with IQTree, using an HKY substitution model and the ultrafast bootstrap method to produce 1000 pseudo-replicate trees [105, 194, 195]. For each of these 1000 trees an additional tree was produced by collapsing short branches (less than 1×10^{-5} changes per site) and randomly resolving the resulting polytomies with the 'ape' R package [106]. The resulting 2000 phylogenies were dated with the R package 'treedater' with a strict molecular clock constrained between 0.0009 and 0.0015 substitutions per site per year [109, 196]. The reproduction number estimates and bounds presented in the results section are averaged across this set of 2000 inferred trees.

5.3.2 Timing of importations

To assess the suitability of the skygrowth model for SARS-CoV-2 in Saudi Arabia, which doesnt allow for multiple importation events to be inferred, we estimated the importation dates of the different SARS-CoV-2 clades found in Saudi Arabia in the included data.

Subsets of the Saudi Arabia sequence data were generated based on their nextstrain clades which are assigned using the nextstrain pipeline and provided in the GISAID metadata when sequences are downloaded [76, 192]. These sequences were combined with a global background set of 500 sequences per clade, extracted from GISAID. The background sequence set used to

select these exogenous sequences consists of over 19,000 sequences collected between 24th December 2019 and 14th September 2020 in 117 countries, available at `10.5587/gis8.230619.tc`. The Saudi Araia sequences used in this analysis can also be found in the same GISAID dataset. To select a subset of 500 sequences for each clade, pairwise genetic distances were calculated for all sequences using the TN93 software; implementing the distance method of Tamura and Nei (`github.com/veg/tn93`)[197]. The 2000 most closely related sequences were selected by genetic distance and stratified by time, with the 500 sequences used in the analysis evenly sampled in time across the 2000 selected sequences.

For each clade, IQTree was used to generate maximum likelihood phylogenies with short branches collapsed and polytomies resolved randomly in the same way as was used in the skygrowth analysis. In this analysis, due to the greater number of sequences used, IQtree was run only once with polytomies resolved 10 times. Treedater was then run 10 times on each of the 10 randomly resolved phylogenies to produce 100 phylogenies for each clade. The state of each internal node in the phylogeny was reconstructed by maximum parsimony with the R package 'phangorn' [198]. As this method cannot directly infer the timing of an importation event, importations were estimated to have occurred at the midpoint of a branch along which a state change was inferred to have happened between the background population and the internal samples. These timings were used to produce the smoothed probability density of importation events over time into Saudi Arabia presented in figure 5.3B.

5.3.3 PhyDyn

As we inferred that multiple importation events were likely to have occurred for each clade and at different times between clades, we implemented a more complex phylodynamic model which allows for a structured population with importations from a global reservoir. Here we consider each of the four cities included in the dataset individually to assess differences in the epidemic trajectory between them. PhyDyn was not run on the eastern region due to the small number of sequences available and the much more limited time over which samples were collected than

the city regions.

The model

This analysis uses a compartmental structured coalescent model in the BEAST2 package PhyDyn to estimate epidemiological parameters from sequence data [153, 183]. Here we are primarily interested in estimates of the effective reproduction number (R_{eff}), complete model parameters are outlined in table 5.1. The model used is an SEIR model with two I compartments to account for overdispersion of transmission rates (superspreading individuals) described in early studies of SARS-CoV-2 transmission and previous description of the SARS epidemic [107, 199]. Exponential growth is assumed in the external population with the internal susceptible population set to the population of the city of interest. This basic model structure has previously been applied during the SARS-CoV-2 epidemic on a number of locations worldwide by the Imperial College COVID-19 response team but is adapted here to include step changes in R_{eff} [3, 4, 190, 191]. Step changes in R_{eff} were implemented by estimating an independent β variable for each month modelled, with a simple time-based if-else statement used to switch between β parameters in the model.

$$
\begin{aligned}
\frac{dI_{EX}}{dt} &= \beta_{EX} I_{EX} - \gamma_{EX} I_{EX} \\
\frac{dS}{dt} &= -\beta S \frac{I_h \tau + I_l}{N} \\
\frac{dE}{dt} &= \beta S \frac{I_h \tau + I_l}{N} - \gamma_0 E \\
\frac{dI_h}{dt} &= \gamma_0 p_h E - \gamma_1 I_h \\
\frac{dI_h}{dt} &= \gamma_0 (1 - p_h) E - \gamma_1 I_l \\
\frac{dR}{dt} &= \gamma_1 (I_h + I_l)
\end{aligned}
\tag{5.3}
$$

$$
N = S + E + I_h + I_l + R
$$

$$
\beta_{EX} = \gamma_{EX} + \omega_{EX}
$$

The equations in 5.3 represent the temporal dynamics for each of the boxes in figure 5.1. Flow to

and from the external population with rate ϵ is dependent on the size of the internal populations only and not the external population, allowing for importation events without affecting the underlying number of infections in the internal population. As these movements do not change the numbers of infections they are not included in equations 5.3. The birth rate in the external population (β_{EX}) is set to allow replacement of those leaving the I compartment (γ_{EX}) and additional growth (ω_{EX}) to give an exponential increase in the external I_{EX} compartment over time.

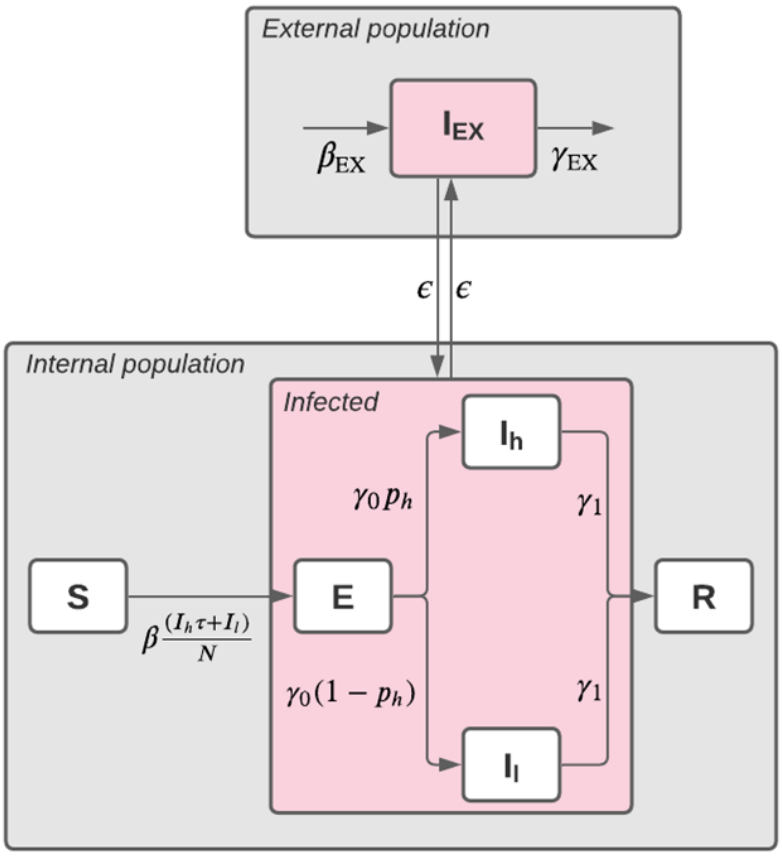


Figure 5.1: Flow diagram representing the modelled populations and the movements between them. The model includes a modified SEIR model in the internal population, accounting for superspreading with an additional I compartment with a higher transmission rate. The model also includes an exponentially growing international reservoir of SARS-CoV-2 infected individuals allowing diversity to be introduced into the local population.

Table 5.1: Parameter table for the model shown in figure 5.1 and the equations in 5.3

Parameter	Meaning	Starting value	Bounds	Prior
ϵ	Import / export rate (movement from and to the external population)	7	0-10	$\sim \exp(10)$
γ_{EX}	Recovery rate for I_{EX}	44	-	*fixed*
ω_{EX}	Growth rate for I_{EX}	36.5	0 to ∞	$\sim$ lognormal$(3.6, 0.25)$
β_0	Initial transmission rate (In the first month of the outbreak)	22	20 to ∞	$\sim$ lognormal$(2.7, 0.5)$
$\beta_{i\neq 0}$	Transmission rates for the other one month blocks	15	5 to ∞	$\sim$ lognormal$(2.7, 0.5)$
τ	Infectiousness of those in the high transmission (I_h) compartment relative to those in the low transmission (I_l) compartment	74	0 to ∞	$\sim$ lognormal$(4.3, 0.1)$
γ_0	Progression rate from exposed to infectious	73	-	*fixed*
γ_1	Recovery rate of infectious individuals (internal)	121.667	-	*fixed*
p_h	proportion of progressions to the high transmission compartment (I_h)	0.2	0 to 1	$\sim$ beta$(50, 200)$

Note: rates are expressed per year

Model parameterisation

The start time for growth in the external population was set to 2019.7 to precede the most recent common ancestor of SARS-CoV-2 based on phylogenies fit to data available on GISAID. Progression to the infectious compartments in the internal population (I_h and I_l) was set to zero before 30 days prior to the 10^{th} reported case in the region of interest, a rule of thumb introduced early in the outbreak [200]. A lower bound of 5 on the transmission rate parameters ($\beta_{i\neq 0}$) was used to prevent stochastic die-out of the virus and to aid convergence, with a lower bound of 20 on β_0 to speed convergence. This value was influenced by initial runs of the model where β was varied at fewer time points, presented in supplementary figure C.1, and by previous modelling work from Alshammari et al. estimating an R_0 of 2.7 for SARS-CoV-2 transmission in Saudi Arabia [201].

Fixed progression and recovery/death rates were used in the model to give an average duration of infection of 8 days in the external population and an incubation period of 5 days and infectious period of 3 days in the internal population [202, 203]. These estimates are based on early reports of SARS-CoV-2 disease progression [204, 205]. The starting value of the external growth rate was set to give a doubling time of 10 days. The starting value of τ (the increased transmissibility of the I_h compartment compared to I_l) was set to give a negative binomial offspring distribution with $k = 0.22$ at $R_0 = 2$. The starting value of ph corresponds to 20% of infected individuals in the internal population progressing to the superspreading I_h compartment.

The tree fitting process uses a strict clock with a uniform prior bounded by 5×10^{-4} and 0.005, based on maximum likelihood estimates of the clock rate and an HKY substitution model. A 'free tree' model is implemented in PhyDyn with tree topology and node heights estimated.

Starting trees and international reservoir sequences

Starting trees for PhyDyn were produced with IQTree, as in the previous sections, with the addition of 80-100 reservoir sequences per region selected by TN93 distance as above. In this analysis, only sequences with 2 or fewer substitutions difference from the regional sequences were included. For Riyadh only 80 matched sequences were available, but for the other regions a time stratified set of 100 sequences was selected from the closely related sequences to act as the international reservoir. As with the work on the timing of importations, only a single maximum likelihood tree was inferred with IQtree; with the set of 20 starting trees used in PhyDyn generated by collapsing short branches and randomly resolving polytomies 10 times and dating each resolved phylogeny twice with treedater.

Running the analysis

The Beast2 package CoupledMCMC version 1.0.2 was used to run MCMC chains in parallel, with one hot and 3 cold chains on each of the 20 starting phylogenies for each location. Sub-chain swapping was used to minimise chain sticking at local optima [206]. These chains were run to a length of 4 million iterations with 50% of each removed as burn in. Where the individual runs had not converged sufficiently with 4 million iterations, they were extended for a further 4 million iterations, repeating until convergence. Chains were logged every 10,000 steps, with sub-chain swaps allowed every 1000 steps.

Reconstructing R from the epidemic trajectory

A reproduction number for each time bounded beta (β_i) was calculated for each location using equation 5.4 at each time step.

$$R_i \;=\; (1-p_h)\frac{\beta_i}{\gamma_1} \;+\; \tau p_h \frac{\beta_i}{\gamma_1} \tag{5.4}$$

Where p_h is the proportion of infectious individuals in the superspreading I compartment (I_h), τ is the relative infectiousness of these individuals compared to those in the lower I compartment (I_l) and γ_1 is the rate of leaving these compartments. This is used to calculate an effective R value (R_{eff}) over time by incorporating the susceptible population size at each time point (Eq. 5.5).

$$R_{\text{eff}(it)} \;=\; R_{it} p S_t \tag{5.5}$$

Where pS_t is the proportion of the total population in compartment S at time t and R_{it} is the R_i value at time t.

Serosurvey data

To test the performance of the PhyDyn models for each region, we compare the estimated cumulative proportion of the population infected from the model to serosurvey data from the

same 4 cities. Serosurvey data from Saudi Arabia are presented in table 5.2. As we are investigating four locations, local level data is limited and the data which is available is sampled from multiple risk groups. When comparing the modelled proportion infected to serosurveys, survey data from healthcare workers were dropped and only data from blood donors were used, highlighted in grey in table 5.2. Although blood donors are likely to be more representative than healthcare workers of general population-level risk, they are skewed heavily to the young male population in Saudi Arabia and migrant workers were also slightly under-represented in the major survey published by Banjar et al [207]. Migrant workers often reside in dormitories or larger households where contacts are likely to have been less limited by the restrictions implemented than in other populations. Banjar et al. reported seroprevalence for migrant workers and native Saudis separately at a national level showing a seroprevalence of 4.3% in migrants and 0.3% in citizens respectively in May 2020. The bounds for each serosurvey were adjusted to account for sampling biases by using a beta binomial rather than binomial distribution.

To estimate bounds for each serosurvey, we assume a binomial distribution for the number of positive samples from a serosurvey with variance $\sigma^2 bin = N\pi(1-\pi)$ where N is the number of individuals tested and π the probability of success. We can approximate the variance of a beta binomial as $\sigma^2 bin \cdot (1 + \rho(N-1))$ where we draw the probability of success π from a beta distribution with overdispersion parameter ρ equal to $\frac{1}{a+b+1}$ where a and b are the parameters of the beta distribution. As sample sizes in serosurveys were small, and repeat samples were not carried out, it was not possible to estimate overdispersion from the data. The overdispersion parameter applied in the beta binomial was fixed at 0.005 to allow for variance in the probability of finding a positive serosample in each survey. The bounds estimated from this assumption are shown in table 5.2.

Table 5.2: Reported seroprevalence estimates for the cities included in this analysis. Lines highlighted in grey are used for fitting.

Seroprevalence (%)	Estimated SD	Location	Population	Date	Reference
24.5	3.56	Makkah city	Blood donors	11/2020	Alharbi et al. [208]
12.9	2.80	Jeddah city	Blood donors	11/2020	Alharbi et al. [208]
2.4	1.15	Riyadh region	Blood donors	06/2020	Alharbi et al. [208]
6.3	-	Riyadh region	Non-COVID patients	06/2020	Alharbi et al. [208]
19.3	2.15	Madinah region	Blood donors (male)	15/04/2020 - 15/07/2020	Mahallawi & Al-Zalabani [209]
0	0.38	Jeddah city	Blood donors	01/01/2020 - 30/04/2020	Alandijany et al. [210]
4.1	2.70	Madinah region	Blood donors	20/05/2020 - 25/05/2020	Banjar et al. [207]
8.1	3.50	Makkah city	Blood donors	20/05/2020 - 25/05/2020	Banjar et al. [207]
2.3	0.79	Jeddah city	Blood donors	20/05/2020 - 25/05/2020	Banjar et al. [207]
0	0.54	Riyadh region	Blood donors	20/05/2020 - 25/05/2020	Banjar et al. [207]
6.4	-	Makkah city	Healthcare workers	06/2020 - 07/2020	Ahmed et al. [211]
1.1	-	Riyadh region	Healthcare workers	20/05/2020 - 30/05/2020	Alserehi et al. [212]
4.6	-	Madinah region	Healthcare workers	20/05/2020 - 30/05/2020	Alserehi et al. [212]
6.3	-	Makkah region	Healthcare workers	20/05/2020 - 30/05/2020	Alserehi et al. [212]

Fitting to serosurvey data

The PhyDyn model was also updated to accept an additional gaussian likelihood to integrate serosurvey data as a fitting parameter. The ability to estimate an epidemic size directly in the PhyDyn model with population size variables from the compartmental model allows a proportion infected to be estimated and fit to serosurvey data which would not be possible with many other population genetic phylodynamic models. Due to limitations with distributions implemented in the PhyDyn package, the mean and variance of the beta-binomial model were approximated with a gaussian distribution in the model. These were assumed to be equivalent for model fitting. The Gaussian likelihood implementation used is included in PhyDyn version 1.7.3 patch 3, fitting the estimated number of infections from the model to a surveyed proportion infected by dividing by a user-supplied fixed population estimate.

As serosurveys are carried out over a period of time and not at a single time point an additional assumption needed to be made to fit the model at a particular time. For this investigation the representative time point selected was the end of the survey period as seroprevalence was likely to have increased over the course of early surveys, meaning a standard assumption that reported seroprevalence was representative of the midpoint of the survey would likely be an overestimation.

5.4 Results

5.4.1 Sampling intensity

Figure 5.2 shows sampling over time by location in comparison to the cases recorded in the same one week periods. Although early May is underrepresented in all datasets, Madinah shows the most consistent viral sampling over time, with sequenced genomes produced during all months of the first wave of SARS-CoV-2. Jeddah has consistent sampling up to the lifting of restrictions at the end of June but data does not extend beyond this point and Makkah has strong early

sampling with few additional sequences later in the outbreak. For Riyadh, the largest city in the analysis, few sequenced genomes are available on GISAID and all are sampled in June, July, and August despite the first reported cases occurring in March, in line with the other cities. Nucleotide diversity was not found to be significantly correlated with the number of included sequences each week using a simple Pearson correlation estimate. Madinah: r(12)=-.11, p=.71, Makkah: r(6)=-.45, p=.27, Riyadh: r(4)=-.13, p=.80, Jeddah: r(10)=-.06, p=.84 where the number in brackets is the degrees of freedom on the Pearson correlation coefficient (r).

5.4.2 Skygrowth model

The Skygrowth model shown in figure 5.2 shows a downward trend in the effective reproduction number over time with the introduction and maintenance of non-pharmaceutical interventions. Following the lifting of restrictions towards the end of June the model estimates that R remained below or at 1 to the end of the period covered by the genetic data presented in this paper. The effective population size represents the relative diversity of the sequences collected in Saudi Arabia over the course of the outbreak, suggesting that diversity peaked at the start of June, ahead of the peak in cases in the country. This is likely driven by the earlier peak in the cities contributing the most genetic sequences to the sample used (Madinah, Makkah, Riyadh and Jeddah).

5.4.3 Timing of importations

The estimates of importation date with phangorn suggest the global connectivity of Saudi Arabia led to early importation of all of the clades detected in the dataset. Clade 19 A and B are estimated to have been imported with highest probability in late February 2020 and the other reported clades were most likely imported in March (Fig. 5.3B). The tree in Figure 5.3A shows that clade 20A was imported on multiple separate occasions in February, with importation date estimates (Fig. 5.3B) suggesting the importations which resulted in onwards transmission in the country occurred in late March.

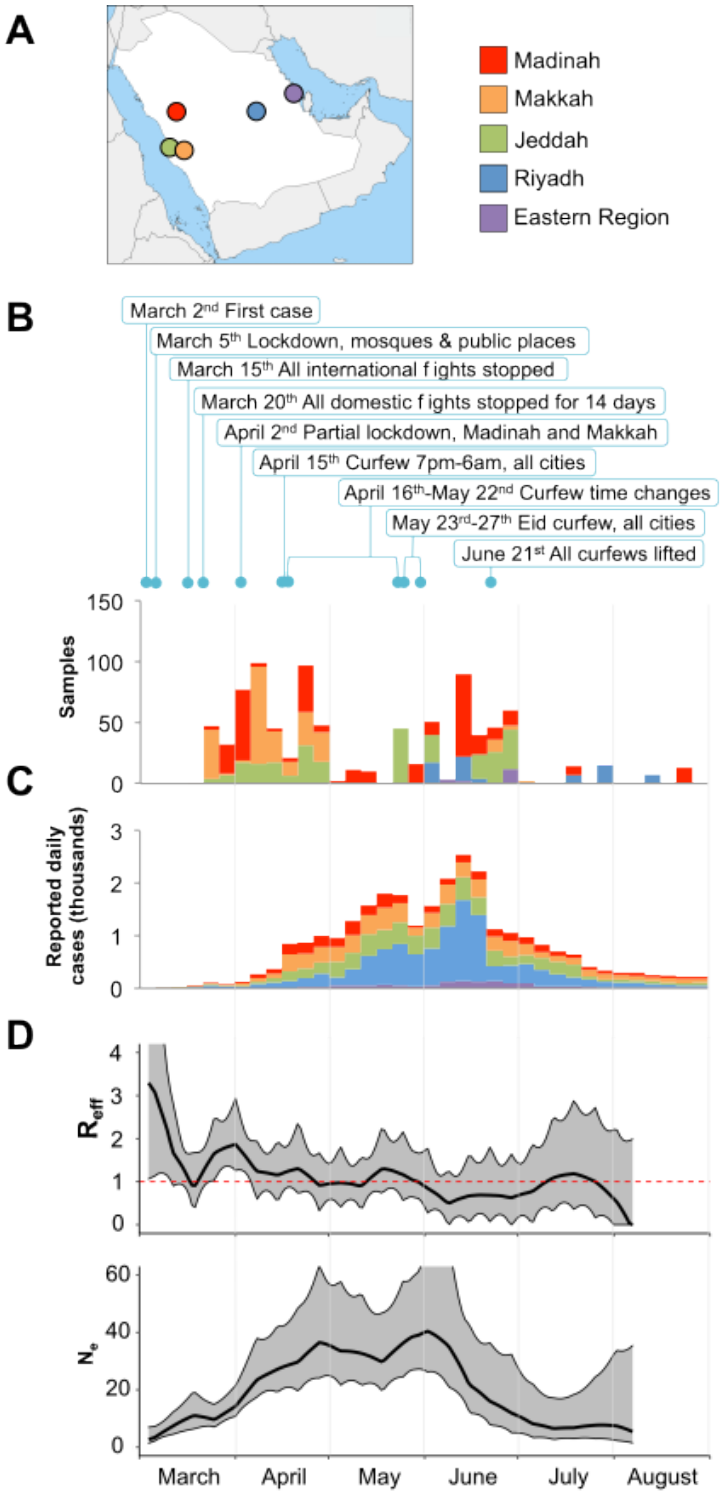


Figure 5.2: **Sample overview and population genetics. A)** Locations of the sampling cities within Saudi Arabia. **B)** Stacked bars showing the numbers of samples retrieved from the 4 cities and the Eastern region during the first six months of the pandemic. Cities are colored as in panel A. Months are shown at the bottom of the figure, and each month is divided into 5-day intervals. Major restrictions imposed by the Ministry of Health and by Royal decrees are indicated above plots. **C)** Stacked bars showing the average numbers of new daily cases in sampling cities. **D)** Estimate of effective reproduction number (R_{eff}) over time in Saudi Arabia (top) and estimate of effective population size (N_e), the relative population size required to produce the diversity seen in the sample (bottom). Central black lines show median estimates, and gray confidence areas denote the 95% credible intervals. The red horizontal red line represents an R of 1, the level required to sustain epidemic growth.

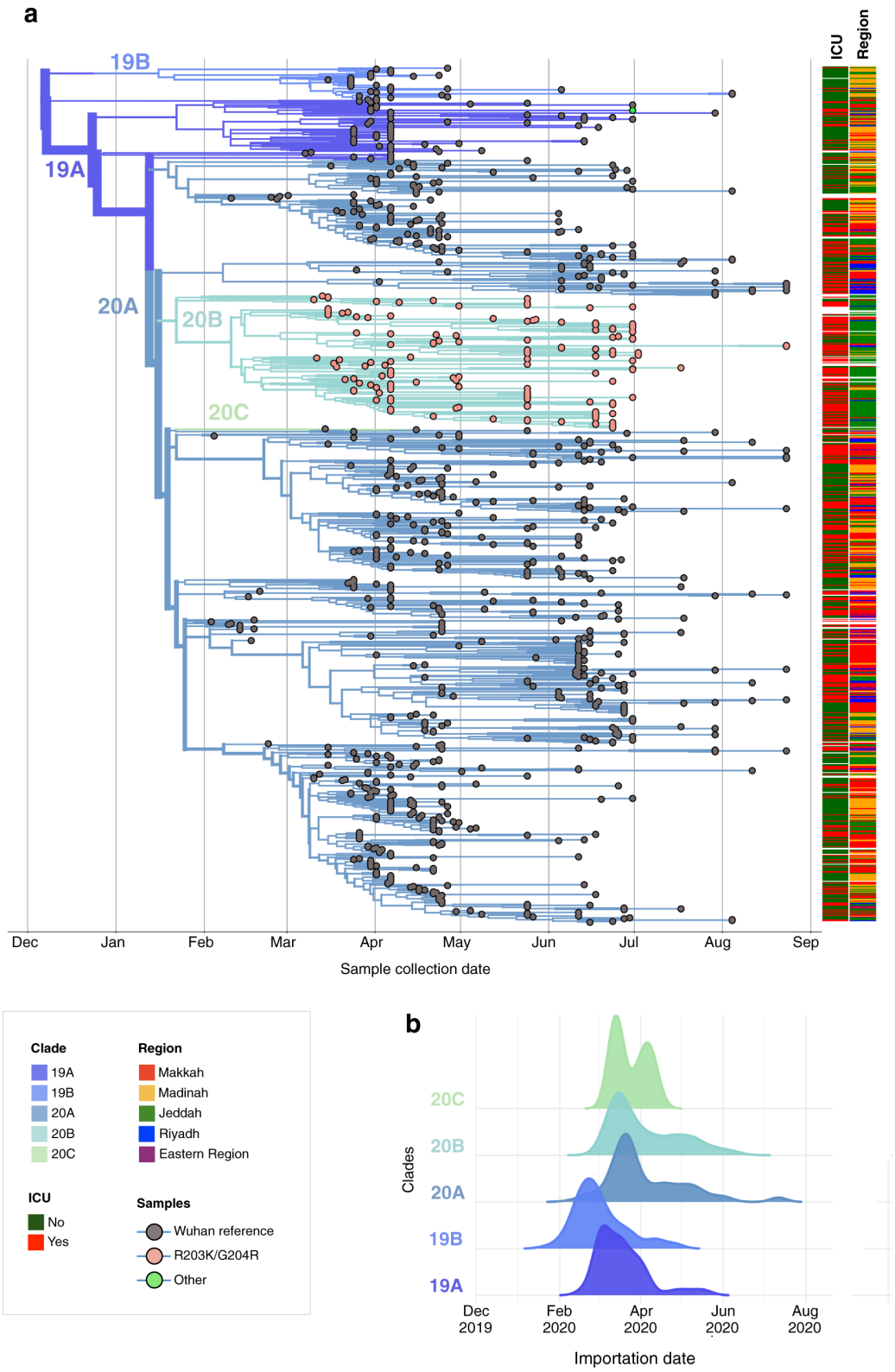


Figure 5.3: **A)** Global Time-scaled phylogeny of 952 Saudi Arabia samples colored by Nextstrain clades. Samples are shown as circles and colored according to their genotype at genome positions 28,88128,883. Intensive Care Unit (ICU) status and sampling region are indicated on the right of the tree. **B)** Distributions of importation dates for the five Nextstrain (nextstrain.org) clades found in Saudi Arabia colored by clade.

5.4.4 PhyDyn model

Assuming that the underlying model structure is well specified, the more complex PhyDyn model should provide better precision on the smaller regional datasets than a more general Skygrowth model. Here we investigate effective reproduction number and proportion infected over the course of the epidemic in four cities individually.

Genetic data fitting

Fitting the model to genetic data alone provided estimates of effective reproduction number for each of the one month periods covered by the data. For all locations estimates decrease following the introduction of restrictions in early March (fig 5.4). In the β interval including the introduction of these initial restrictions the 95% credible interval of the R_{eff} for Makkah (R1 = 1 (CrI: 0.7-1.4)) and Jeddah (R1 = 1.2 (0.9-2.0)) includes one, suggesting no significant prediction of expanding transmission. In Riyadh and Madinah this drop in R_{eff} occurs more gradually, with the credible interval not including one until the succeeding interval representing April 2020 (Fig 5.4)(March: Madinah R1 = 1.5 (1.0-2.0) and Riyadh R1 = 1.9 (1.3-2.8), April: Madinah R2 = 1.1 (0.8-1.4) and Riyadh R2 = 1.5 (1.0-2.0)). Following the initial decline, only Makkah has a one-month period where the credible interval of the R_{eff} estimate does not include one for the remainder of the fitted period (R3 = 1.9 (1.2-2.7)). This occurs in May (Fig 5.4) and coincides with the peak in reported cases in Makkah but also during a period where no genetic sequences were collected in the city (Fig 5.2). The median R estimate is higher in July, directly following the lifting of restrictions at the end of June, than in the previous month in all locations; although, with 95% credible intervals including 1 in all locations at this time, this should not be overinterpreted.

The mean rate of movement to and from the external reservoir estimated for each of the four cities was variable, with high rates of 7 per year for Riyadh and 6.27 for Jeddah and low rates of 1.69 for Madinah and 1.60 for Makkah. This suggests that the larger, more internationally

connected cities of Riyadh and Jeddah had more interaction with the outside world whereas local transmission was more central to persistence in the holy cities of Madinah and Makkah which were without visitors and under stricter lockdowns for much of the first wave of the pandemic.

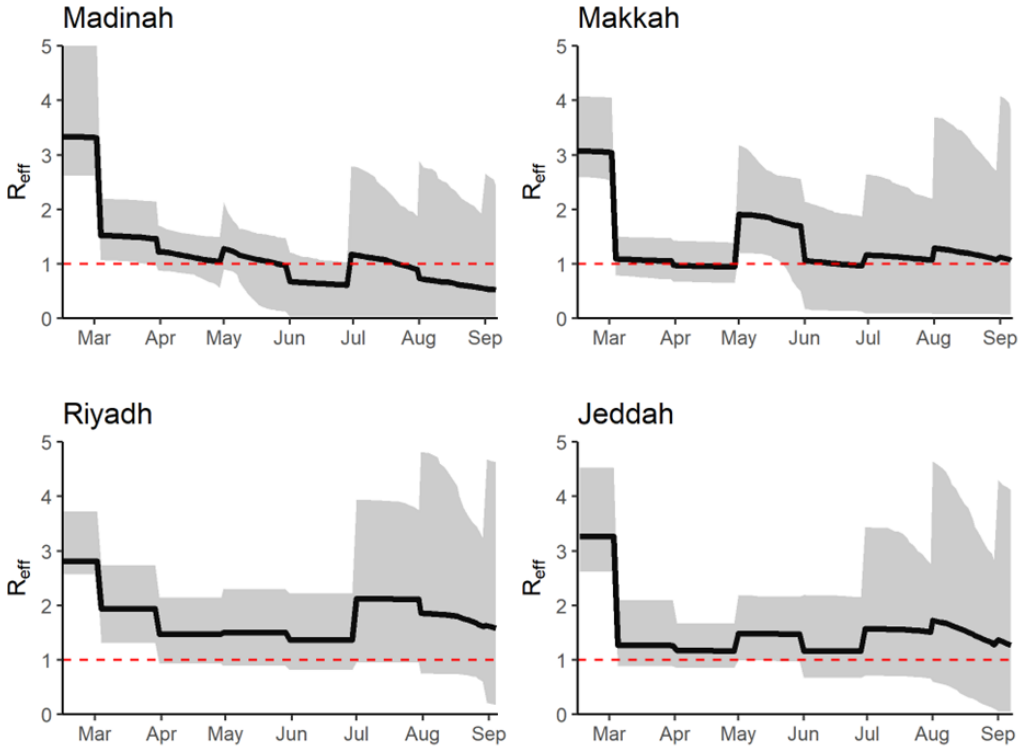


Figure 5.4: Effective reproduction number estimates (black line) and 95% credible interval (grey shaded area) over time from the PhyDyn model without seroprevalence for four cities of interest. The red horizontal line represents an R_{eff} of 1, the number required for virus replacement. These estimates have step changes at monthly intervals.

Comparison to serosurvey data

Estimates of the cumulative proportion of the population infected were compared to serosurvey estimates of the proportion of the population found to be seropositive at the same point in time (Fig 5.5). Comparison of fitted estimates to serosurvey data shows that serosurveys estimate a higher infection rate than models based on available genetic data in mid-July for Madinah. In the other locations the median modelled estimate falls within the bounds of the serosurveys. This suggests that the model is performing well at predicting viral transmission in Riyadh, Makkah and Jeddah using genetic data alone, although the serosurvey data used for

comparison are limited and the estimates from the model have wide credible intervals towards the end of the modelled period.

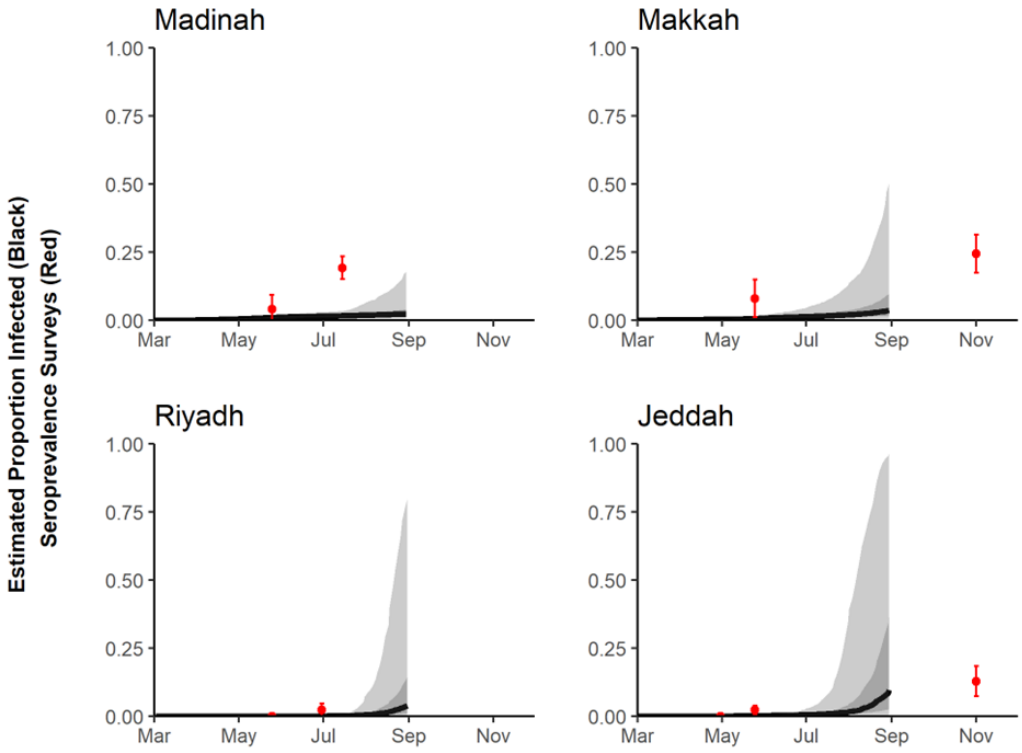


Figure 5.5: Estimates of the cumulative proportion of the population infected (black line), 95% credible interval (light grey shaded area) and interquartile range (dark grey shaded area) in the model over time for each city. The red error bars and points represent the seroprevalence predicted by the serosurveys and the bounds of the beta binomial distribution applied to these likelihoods. The position of these points on the x-axis represents the end point of the serosurvey from which they were collected.

Incorporating serosurvey data

As some discrepancy is seen between the serosurvey data and the model estimated cumulative proportion of the population infected, serosurvey data were considered as an additional covariate for model fitting. Using the serosurvey data as additional parameters when fitting the phylodynamic model significantly increases the estimated proportion infected in Makkah and Madinah at the end of the first wave of coronavirus transmission (Fig. 5.5 & 5.6). For Makkah the new estimate falls within the 95% credible interval estimated without the additional serosurvey data but for Madinah the second serosurvey estimated a significantly higher proportion

of the population had been infected by mid-July than was estimated by the model fitted solely to genetic data. Supplementary figure C.2 shows that a model fit with fewer beta changes leads to an estimate which better fits this serosurvey datapoint.

The model fitted to serosurvey data estimates a spike in effective reproduction number in July in Madinah, although the 95% credible interval of this estimate still includes 1 (Fig. 5.7). This estimate is bimodal with beta estimates of 9.8 and 43.5 being the most common. The 20 chains run were not able to converge to one of these values after 24 million iterations each. The estimated importation and exportation rates are consistent with the rates reported for the model fitted to genetic data alone presented above. Overall estimates of R_{eff} with serosurvey

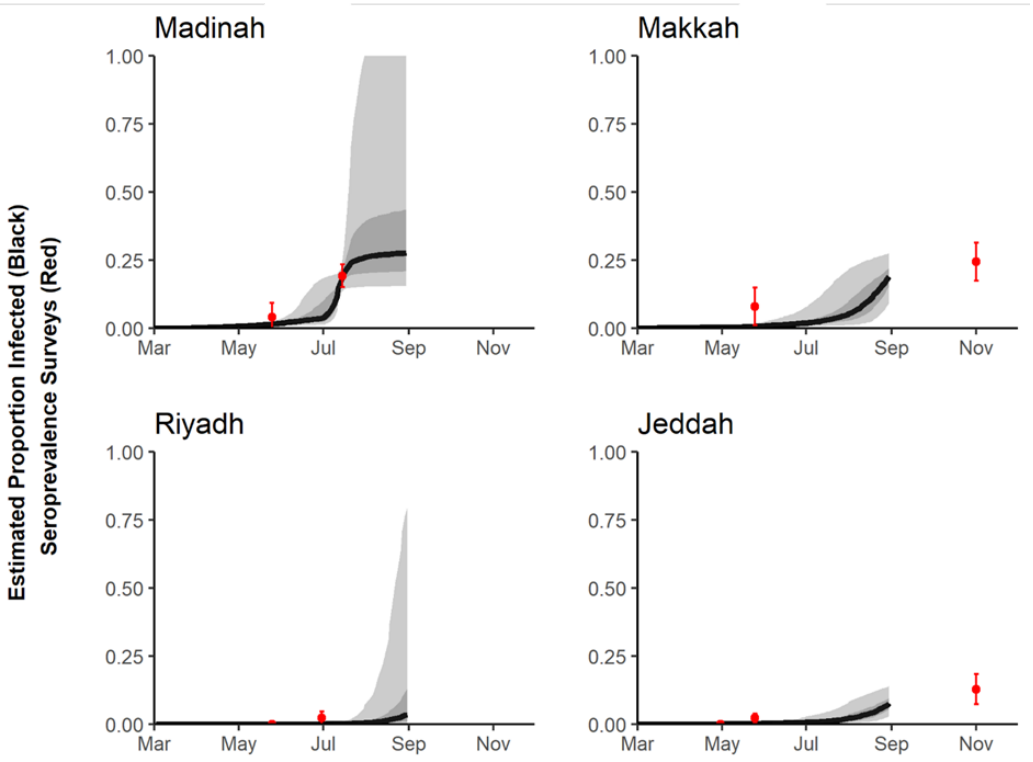


Figure 5.6: Estimates of the cumulative proportion of the population infected (black line), 95% credible interval (light grey shaded area) and interquartile range (dark grey shaded area) in the model over time for each city. The red error bars and points represent the seroprevalence predicted by the serosurveys and the bounds of the beta binomial distribution applied to these likelihoods. The position of these points on the x-axis represents the end point of the serosurvey from which they were collected.

data included were in line with those without, with slight increases in the lower bound of the 95% credible interval for some time periods. There were no time periods where 95% intervals

including one when fitted to genetic data alone, did not include one with the addition of serosurvey data.

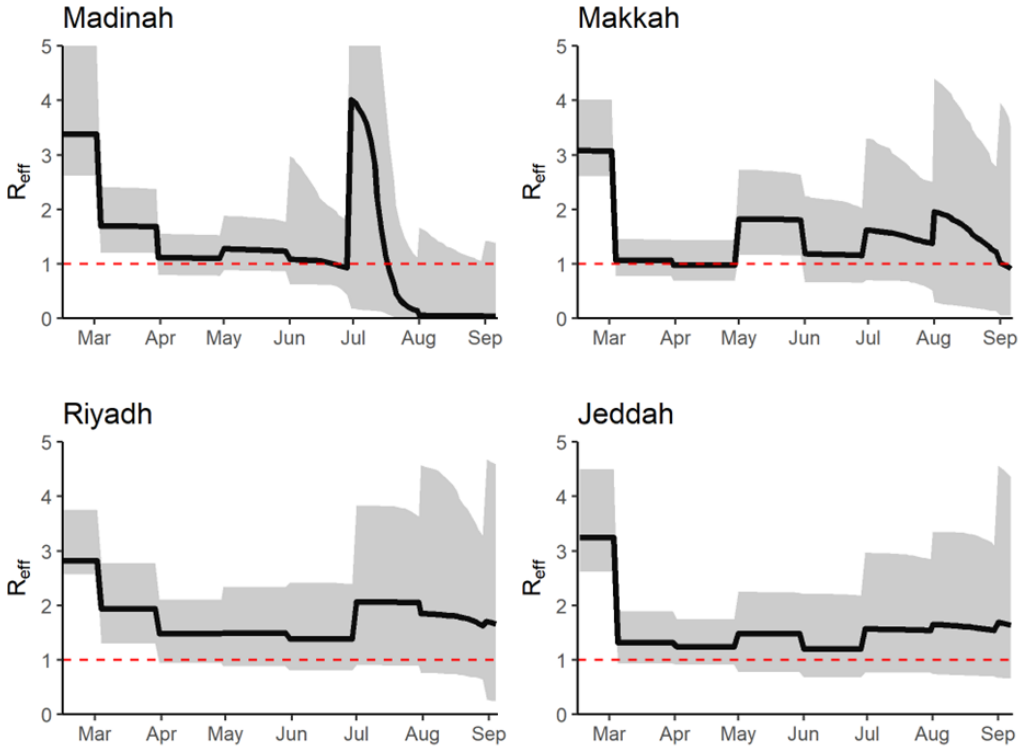


Figure 5.7: Effective reproduction number estimates (black line) and 95% credible interval (grey shaded area) over time from the PhyDyn model with seroprevalence included as an additional covariate for 4 cities of interest. The red horizontal line represents an R_{eff} of 1, the number required for virus replacement. These estimates have step changes at monthly intervals

5.5 Discussion

This study highlights the utility of genetic data collected during the SARS-CoV-2 pandemic beyond the common use case of tracking and investigating viral variants and individual level transmission events. By leveraging the flexibility of the coalescent we are able to use genetic data providing much more information on the interconnectedness of the viral population as a proxy for case data in a traditional epidemiological model. This is a useful avenue of investigation and becomes increasingly topical with the move towards ubiquity of rapid, full genome sequencing techniques in viral genomics.

The Skygrowth model presented gives a good overall picture of the progress of the early SARSCoV2 epidemic in Saudi Arabia with changes in R_{eff} appearing to follow the implementation of non-pharmaceutical interventions. Although the coalescent and the assumptions of skygrowth are relatively robust to heterogenous sampling through time, the method can still be biased by significant changes in effective population size or unequal sampling in space or across risk groups. Computation of reproduction numbers is premised on equivalence of the growth rate of the epidemic and the growth rate of the effective population size which does not hold when the transmission rate is highly variable. Work by Inward et al. suggests that, as with the birth-death skyline, the Skygrowth model performs best when sampling is proportional to the incidence in each time period [162]. For this sampling scheme we would expect to see positive association between nucleotide diversity and sample size; The lack of correlation between nucleotide diversity and number of samples in this analysis and the lack of alignment between the samples in figure 5.2B and cases in figure 5.2C suggest that sampling was more opportunistic in this dataset, the worst performing strategy for reconstructing R_{eff} with the skygrowth method in Inward et al. 2022. [162]. Additionally, Skygrowth makes an implicit assumption that the included data is representative of the full picture of transmission, which is not the case for SARS-CoV-2 where we know that virus circulation was global with multiple repeat introductions of different clades into various regions. The repeat introductions and different introduction timings across clades shown in the timing of importation analysis presented in this chapter highlight that this was also the case for Saudi Arabia and using the Skygrowth model for this scenario is likely to overestimate the transmission rate in the region required to generate the diversity seen.

PhyDyn R_{eff} estimates for all locations show a general decline from the initial outbreak to after the introduction of interventions. In Makkah and Jeddah R_{eff} was estimated to fall lower in the first phase of epidemic response than in the later curfew phases although the 95% credible interval includes 1 for all of the time following the first introduction of restrictions in both cities. As these cities are close geographically, and share a common regional government, it is expected that they follow similar patterns. Alrasheed et al. applied a network-based epidemic

simulation model to early SARS-CoV-2 case data from Saudi Arabia, structured based on age, gender, nationality and location [213]. In this analysis they assess concordance between modelled and observed data to assume much higher compliance with early blanket measures such as school closures and domestic flight shutdowns, which occurred for 14 days during this early period, compared to later individual-level restrictions such as curfews [213]. This aligns with the observed strong decline in March and April in all locations. The changes in R_{eff} during the curfew period (April to mid-June) may represent changes in curfew over time but are also influenced by sampling differences in each time period, hence the wide credible intervals. Curfew measures varied by time and location from 11 hour overnight curfews or neighbourhood lockdowns preventing movement within cities to full 24 hour national lockdowns [214].

In Madinah, where the initial R_{eff} estimate prior to the introduction of restrictions was higher than in the other regions, the PhyDyn model does not predict a fall below one prior to the curfew but the more strict curfew measures compared to the other cities appear to push R_{eff} lower throughout the curfew period (Fig. 5.7). Four districts in Madinah had 24 hour curfews imposed on April 10^{th} which ended slightly before the nationwide Eid curfew from the 23rd to 27th May meaning the majority of the curfew period from early April to June 21^{st} was spent at the highest restriction level [215]. The large peak in R_{eff} shown in figure 5.7 following the lifting of restrictions is not reflected with a peak in reported cases (Fig. 5.2). Figure C.2 shows that with fewer beta periods we are able to reconstruct a case trajectory which fits these seroprevalence estimates without this transmission spike in July. Uncertainty in this large peak in transmission rate is supported by the bimodal nature of the beta prediction for this time period in Madinah. Limited genetic data from July in Madinah may also play a role in this prediction.

Most published modelling studies from Saudi Arabia did not attempt to estimate starting R_0 or effective R over time estimates [213, 216]. Alshammari estimates an R_0 of 2.7 using a deterministic compartmental model fitted on early case data [201]. The starting values and bounds used in this study were based on these findings together with the median R_0 estimates of 2.86 in Jed-

dah, 2.92 in Madinah and 2.81 in Makkah with fewer beta periods presented in supplementary figure C.1. To improve convergence in this model with a large number of β compartments, the lower bound for β_0 was set to 20 provide a lower bound for R_0 of 2.3. Ideally these additional constraints would not be required but here we attempt to balance computational efficiency and number of parameters predicted.

Although the current PhyDyn model assumes continued transmission from the global reservoir to the internal population to account for migration, it is currently at a fixed rate throughout the epidemic. Introducing a step change in this value in a similar way to that introduced for beta would give a better representation of the dynamics in a country such as Saudi Arabia where strong controls on international travel were implemented. This could result in fewer estimated importation events in some locations, requiring higher internal transmission rates to reconstruct the observed internal viral diversity.

There is also varied evidence from viral culture experiments to suggest that infectiousness may precede the detection of symptoms and that mean duration of infectiousness may be longer than the estimates used in this analysis [217]. As the movements between compartments are defined as rates, there is flexibility in the time for progression through the compartments between individuals. Additionally, the flexible PhyDyn modelling language would allow a more complex compartmental model incorporating symptomatic and asymptomatic infectiousness to be developed. The model also depends strongly on the fixed progression and recovery rates (γ_0 and γ_1) which determine the generation time, and changing these rates can dramatically alter the final size of the epidemic returned.

The addition of multiple β values over the course of the epidemic provides an advantage over the initial SARS-CoV-2 PhyDyn model introduced in Volz et al. and applied more widely in Geidelberg et al., Ragonnet-Cronin et al. and Miller et al. [3, 4, 190, 191]. This initial model was used to predict changes in viral dynamics in the initial exponential phase of an outbreak,

only allowing R_{eff} to decrease over the fitted period as susceptible individuals are depleted. By maximising a set of individual β parameters in this work, the effective reproduction number estimated is able to increase or decrease at change points throughout the modelled period. This change introduces additional parameters which need to be estimated but provides a model which is more adaptable to changeable transmission dynamics over the course of an epidemic.

Fitting the PhyDyn model additionally to seroprevalence results in little difference in R_{eff} prediction in this example as there are few serosurveys over the course of the outbreak. The ability to include additional likelihood data in this way could provide additional clarity in locations with more complete serosurvey data or more complex outbreak dynamics over time. As the example of Madinah highlights, fitting to additional data in this way can dramatically alter the estimated proportion infected, suggesting transmission dynamic estimates from incomplete genetic data should be interpreted with caution.

Serosurvey data also have their own limitations. In Saudi Arabia, where these data are primarily sampled from young male blood donors (age range 24-55, over 90% male) and healthcare workers, the sampled individuals are unlikely to be representative of the infection risk profile of the population as a whole [218, 219]. A random sample of the population would provide a better subset for comparison although this would require a targeted study rather than analysis of coincidentally collected blood samples.

For many SARS-CoV-2 datasets collected globally, genetic sampling was also targeted at certain time periods or demographics, or only small numbers of samples were sequenced as sequencing is not required for diagnosis. Saudi Arabia implemented a strong quarantine and isolation procedure for positive tests and hospitalised cases with a large number of genetic samples collected from these populations. This provides a more robust sample to fit this model on than would be seen in other regions where sampling was more heavily targeted to travellers or contact tracing, both of which could affect the diversity found in the sample, making it less representative of

the population. The sampling here was still not ideal due to the opportunistic nature of the samples collected as outlined previously.

We see concordance between the two coalescent based estimates of R_{eff} presented in this work from Skygrowth and PhyDyn, showing a general decrease over time and a non-significant increase following the lifting of restrictions at the end of June. As genetic data without the introduction of non-pharmaceutical interventions is not available for comparison, the degree to which the observed declines are due to interventions alone cannot be ascertained but the fitting provides evidence that the stronger curfew procedures in Madinah are associated with lower effective reproduction number estimates over the course of the curfew period and R_{eff} estimates during the curfew period were significantly lower than before the introduction of interventions in all locations. A published modelling study predicts 104% higher case numbers without school closures, 113% higher without mosque closures, 83% higher without flight shutdowns and 23% higher without curfews in the country [213].

With the increasing global capacity for sequencing and the wealth of sequence data generated throughout the SARS-CoV-2 pandemic it is clear that phylodynamic methods will play an increasingly important role in future epidemics. Estimating epidemiological parameters with a combined model incorporating sequence data alongside traditional methods can provide more robust insights into transmission dynamics and well-sampled genetic datasets should be comparable across countries with less dependence on the attribution of cases and deaths by country which can affect traditional estimates of R_{eff}. The incorporation of a compartmental transmission model in the PhyDyn framework makes this method a good candidate for further integration with existing epidemiological techniques and the updates outlined in this work expand the utility of the method in this role.

Chapter 6

Discussion

With the advent of fast, accurate and inexpensive molecular techniques, phylogenetics is becoming an increasingly important discipline for the tracing and control of infectious diseases [220]. Genetic data provide additional information on the evolution and spread of viruses which can be hard to ascertain from traditional epidemiological data [221]. This data can be linked with other covariates, especially the location and date of sampling, providing unique and valuable inferences. The wealth of variant and phylogeographic inference provided by collaborative analysis of the high volume and publicly accessible SARS-CoV-2 sequence data is evidence of this utility [222].

Rapid analyses of emerging viruses, and new detections of eradication targets such as polio, are crucial to implement timely intervention response [223]. These two scenarios, elimination of a previously widely endemic disease and response to a newly emerging pathogen have several parallels, especially the importance of geographic spread to response, and the influence of sampling rates on conclusions drawn from analyses. This thesis aims to retain rapid phylogenetic analysis techniques, often implemented in early outbreak response or the analysis of new detections of eradication targets, improving the reliability of conclusions drawn by addressing sources of error in estimates made. A technique for incorporating sampling data in discrete trait analysis is proposed and tested, based around existing k-state Markov transition models, and providing a significant reliability boost to transition rates reconstructed on simulated data. The thesis also introduces visualisation techniques and pipelines which will allow for rapid and interpretable analysis of new poliovirus detections in the future, key goals of the initial proposal. Additionally, novel phylodynamic analysis of early-outbreak SARS-CoV-2 in Saudi Arabia is presented, incorporating additional data and parameters not included in published work.

6.1 Polio phylogeography

Following a long-term global eradication effort, wild-type poliovirus is now endemic to only Pakistan and Afghanistan, with only one of the three serotypes (type 1) remaining in circulation [18]. The current WHO eradication target is 2026, following several delays to the initial target of the turn of the millennium, and many difficulties; especially with providing vaccinations in hard-to-reach populations and falling public and government interest as the eradication campaign has led to dwindling paralysis cases [224]. Reintroduction of WPV1 to Africa, with eight cases reported in Mozambique and Malawi in 2022, has dampened expectations of eradication further, alongside declining vaccination rates (80% globally in 2021), and vaccine-derived polio detections in 2022 in a number of western countries [17, 224, 225]. Recent analysis has also shown that the current speed and scope of response to vaccine-derived outbreaks in Africa is not sufficient to interrupt transmission without seeding further outbreaks [226]. In this context, new techniques and better integration of existing data sources become increasingly important to help drive the eradication campaign forwards.

Surveillance for poliovirus is difficult as the majority of infections (around 70%) are asymptomatic, with paralysis only occurring in 1 in 200 to 1 in 1000 infections. Combined with the complex movement dynamics in Pakistan and Afghanistan driven by nomadic populations, seasonal migrant workers, cross-border movement and political instability, this low case-to-infection ratio means pockets of transmission can be missed entirely with AFP surveillance [75, 97]. Over the course of the last decade, environmental surveillance of wastewater systems has expanded in the block, particularly in Pakistan, which provides invaluable additional information on infections [30]. Wastewater surveillance allows a large population to be covered by sampling a small number of confluence points, providing detail on otherwise undetected asymptomatic infections, as outlined in the introduction.

As both ES and AFP detections require sequencing for virus typing, a large number of virus

sequences are produced annually in the endemic countries. Currently neighbour-joining phylogenies are routinely produced from these sequences to assign 'transmission clusters' based on genetic similarity. Here we aim to extend on this regularly produced analysis with phylogeography, enabling us to better understand the spatial spread of the virus with routinely collected sequence data.

Discrete trait analysis was selected for this work following its application to several high-profile epidemic and endemic diseases in recent years, including flu, Zika and Ebola [66, 73, 123]. As with previous analyses, the primary motivation for selecting DTA was the speed and ease of implementing the method on large numbers of sequences and discrete regions. With an eradication campaign timing is everything, new data are sparse due to the small number of detections but quick response to new detections is of utmost importance to prevent further spread [226]. DTA can provide information on geographic spread in a matter of hours to days where robust pipelines for data analysis, subsetting and cleaning are implemented and, due to the simple technique, can incorporate greater numbers of discrete locations than other, more complex, phylogeographic methods such as structured coalescent or birth-death models.

Despite the migration assumptions made with DTA, treating migration akin to evolution of a phenotypic discrete trait, recent work has shown that the method performs equivalently to more complex methods, under both endemic and epidemic conditions [61, 157, 165]. A major assumption made in DTA is that sampling intensity in a region is proportional to the infected population, one which is often violated in real-world surveillance [61]. Here we propose a simple correction to account for sampling rate by location, based on a non-reversible Mk model and incorporating sampling rate as a multiplier of the transition rate matrix in the same way as equilibrium base frequencies are incorporated in the GTR model of Tavaré [129, 131]. This means that the correction can be easily implemented in a range of existing phylogenetic software, with implementations in the popular R package 'ape' and the BEAST1 Bayesian phylogenetics software presented here. The treetime method implemented in Nextstrain, which has recently been applied to wild poliovirus data, could also be easily adapted to allow migration to

be modelled with the proposed correction with the addition of a non-reversible transition model.

A large amount of recent effort in phylogeography and phylodynamics has been aimed at accounting for changes in sampling rates over time, an issue which affects birth-death and mugration models if not considered in the model. Approximate coalescent models based on deterministic demographic processes can also be biased where this is a bad approximation of the true stochastic process in the population [178]. The method suggested in this thesis does not account for this, although future work to generate an epoch version of the model, incorporating changing sampling rates over time, could allow for better integration of changing sampling rates. This would also need to be tested and compared with other proposed corrections, specifically approximations of the birth-death and coalescent processes. Recent work by Layan et al. has suggested that a Bayesian implementation of DTA, even without correction for sampling bias, performed equivalently to a pair of structured coalescent approximations in an epidemic scenario [157]. Further testing with time-structured implementations would be required as an epoch DTA model is developed, with a birth-death simulator for time-varied sampling introduced by Volz and Frost in 2014 available to allow these models to be tested against a known sampling process [178]. With the benefits of DTA in terms of computational power required and number of discrete locations considered, future work to assess the performance of my proposed correction to the DTA over changes in dynamics during an epidemic would be of particular interest to test on simulated data in the future; An extention to the endemic disease simulations included in this thesis.

Further testing of the correction method proposed is also required on a larger number of locations to ensure that the correction continues to perform well in more complex structured populations. The BEAST implementation presented in chapter 4 with the addition of BSSVS rate selection should make this process simpler, allowing for zero rates of transition between some pairs of locations which is not possible with the simple relative rate maximum likelihood method as implemented in chapter 3. There is also a case for continued development of the maximum likelihood implementation due to the strong influence of the priors used on the move-

ment rates reconstructed under the Bayesian model, requiring the use of Bayes Factor support to select significant rates. A penalised maximum likelihood rate selection procedure penalising non-zero rates could be implemented without specifying priors, although convergence would likely be an issue without prior rate information. This is worth investigating further as the speed and lower computational cost of simple maximum likelihood optimisation implemented in R over a Bayesian MCMC implementation may mean it is more widely used in computational-resource-limited or time-limited scenarios such as outbreak response and eradication campaigns.

Applying the proposed correction to polio data in this case resulted in corrected rates and numbers of transitions within the 95% credible intervals of the estimates made with the simple DTA method. This shows that DTA is likely appropriate for the analysis of polio data going forward. More work is needed to ensure that ES data is incorporated correctly into future models due to differences in sampling intensity over time and location. For example, Punjab province in the East Pakistan region is overrepresented in the ES data, inflating the importance of this region to transmission in the analysis presented in chapter 3. This importance is strongly diminished when only considering AFP data in Appendix B and chapter 4, suggesting that treating both data sources as equivalent leads to overemphasis of certain well-sampled regions. This also re-emphasises the difficulty of surveillance for polio where, were environmental surveillance complete, we would see a much greater extent of movement and transmission of the virus which cannot be detected by AFP surveillance alone.

Here we introduce and investigate a discrete method, where sequences collected are grouped into subpopulations based on geographic location. This assumption helps to produce interpretable results such as rates and numbers of movements between regions, but efforts need to be made to produce representative subpopulations. In this analysis a large amount of input was sought from collaborators familiar with the demographic and epidemiological situation in the region, requiring regions to be constructed incorporating parts of different administrative provinces. Increasing the number of regions is always of interest, especially to policymakers who need to work at smaller scales than those represented in this thesis, but this disaggregation introduces

additional variability in sampling between regions, leading to a smaller amount of information available for each region and inferences which are more likely to be biased by sampling rate differences. Introducing artificial boundaries between discrete regions, or boundaries to other linked regions not included in the analysis, can additionally lead to missing data problems or the spatial amalgamation or segregation of linked populations depending on where boundaries are drawn (the "boundary" and "modifiable unit" problems of spatial analysis) [118, 119]. In the future it may make sense to integrate continuous space in the analysis of polio sequence data, as is often used with other epidemiological data. Including additional covariate data on population dynamics, such as population movement models, travel, and mobile phone data can also provide additional information on movements, increasing statistical power over a basic genetic-based DTA model as implemented here.

The visualisations presented in this work provide an update to existing routine poliovirus visualisations. Here consistent regional colours are retained throughout the analyses, providing uniformity across a range of visualisations. This is an improvement over current visualisations which often report the complete name of a sequence, including basic metadata, which can more elegantly be presented with shapes and colours, improving interpretability. These visualisations are based on previous work (primarily the visualisations presented by Dudas et al. for the 2013-2016 Ebola outbreak and available in the baltic python library [73]) with adaptions made to better suit the larger poliovirus database included and the two different sampling sources represented in the data. I coded the included plotting techniques in R, around the ggplot and ggtree packages, rather than building on the python framework used in the baltic pipeline to provide a visualisation framework which can be easily integrated with existing R-based phylogenetic techniques [227, 228]. R is the statistical tool of choice for many epidemiologists and provides implementations of many phylogenetic techniques as well as integrations with the BEAST software suite.

6.2 SARS-CoV-2 phylodynamics

Phylodynamics has played a large role in the epidemiological analysis of SARS-CoV-2, not seen before in outbreak response. Phylodynamic analyses were produced early in the epidemic, primarily used to assess transmissibility and estimate reproduction numbers [190, 229]. This role expanded throughout the pandemic with phylogenetic and phylodynamic analyses applied to discovery and characterisation of new variants, variant tracking, and assessing the impact of interventions [4, 230–232]. The SARS-CoV-2 pandemic was the first major outbreak in the new era of pathogen genomics, with faster and cheaper sequencing methods proliferating around the world, including the expansion of sequencing capacity in low and middle income countries (LMICs) (despite this expansion, huge disparities in the proportion of SARS-CoV-2 cases sequenced between high income countries and LMICs were still seen [233]). The amount of open-access sequence data produced in the first year of the pandemic, reaching one million sequences on the GISAID platform in March 2021, emphasised the need for fast, accurate and simple to implement phylodynamic analyses [234]. By January 2022 the capability for sequencing had expanded further, with 68% of WHO member states having the capability to sequence SARS-CoV-2 in country, an increase of 14% on a year earlier and a good basis for enduring genomic surveillance for other pathogens [235].

The basis for the model presented in this thesis was developed early in the pandemic by the Imperial College COVID-19 response team and released in a report in late February 2020 [190]. This original technique, although widely used throughout the pandemic, was designed to model the initial outbreak phase of an epidemic in a location, allowing for the estimation of an initial reproduction number followed by a decrease as susceptible individuals were depleted. In order to apply the model across a longer time range, here I split the model into discrete 1-month time periods, allowing the reproduction number to change more flexibly over the course of the epidemic and allowing the value to increase and decrease over time. This allows us to, for example, detect an increase in the reproduction number following the lifting of non-pharmaceutical interventions in Saudi Arabia in late June of 2020, matching more closely what we see with the

Bayesian skyline model [2].

Fitting the model to serosurvey data, a gold standard measure of acquired immunity, gave a measure of the proportion of the population that had been infected (prior to the introduction of COVID-19 vaccines) with which to calibrate our models. Although serosurvey data were generally published later in the outbreak, meaning they were not available for fitting during the epidemic phase itself, this could be an area where greater collaboration between immunological investigators and genomic epidemiologists should be sought in future outbreaks, allowing this fitting to be done closer to real time.

Here the model is applied to data from Saudi Arabia due to the availability of routinely sampled sequence data for the country. Many countries globally relied on sequencing primarily as a detection tool for new introductions, targeting travellers specifically, and thus their sequence data did not provide a complete picture of the virus variants circulating in the population. The Saudi sequences used in this study were instead collected from quarantine hotels which were used to house positive detections in the community (a resource intensive and not widely used quarantine procedure) and hospitalised patients. This makes the data collection and fitting presented a best-case scenario of detections, especially early in the pandemic before the proliferation of community sequencing in places such as the UK. The lack of changes in diversity in line with the distribution of cases in the dataset suggests sampling was incomplete even in this scenario and may introduce additional bias in the R_{eff} estimates presented [162]. More complete analysis of a range of different intervention techniques with the original method from Volz et al. is presented in Ragonnet-Cronin et al. [4, 190].

The simple improvements to a widely used SARS-CoV-2 model presented here may be of interest in the future if similar phylodynamic models are applied to different outbreak situations. The PhyDyn platform used provides a simple and flexible framework to incorporate traditional epidemiological and phylogenetic analysis, a tool which becomes increasingly important with

the proliferation of genetic sequencing data, and these improvements present a further boost in its utility.

6.3 Conclusions

Pathogen genomics is an ever-expanding field with faster, cheaper and simpler sequencing methods introduced in recent years [236]. Of particular note is the availability and affordability of these techniques in LMICs and even the ability to sequence in field conditions rather than purpose built laboratories [237, 238]. The GPEI and funders are increasingly interested in rolling out next generation sequencing techniques for polio, including direct detection of polioviruses from stool and sewage samples without the need to first culture the virus [239]. This technique greatly improves the delay from collection of a sample to identification of a poliovirus serotype and production of a VP1 sequence and trial use is currently underway [5]. We have also seen next generation sequencing techniques used widely in outbreak responses since the SARS-CoV-2 pandemic, with MinIon sequencing data published for the recent Ebola outbreaks in Guinea and Uganda as well as MinIon and Illumina sequencing being widely used in response to the ongoing mpox outbreak [240–242]. Additionally, the WHO has published new guidance on the use of genomics for outbreaks and potential pandemic pathogens, pushing for further training and use in LMICs as well as increased sharing of sequence data produced [235].

Improving the simplicity and speed of phylogenetic and phylodynamic methods downstream of these huge sequencing advancements, as well as producing interpretable and intuitive results and visualisations, are key steps for the future of phylogenetic analysis in epidemiology. It is important that limitations and downfalls of methods are investigated, accounted for, and spelled out plainly in phylogenetic literature, particularly where methods are introduced, making the methods accessible to genetic researchers carrying out sequencing and data collection.

Phylodynamics and phylogeography as fields have two conflicting aims, the implementation

of increasingly complex dynamic models and the ability to analyse increasingly large and detailed datasets, including much of the new data now available as a result of more widespread pathogen sequencing. These aims are hard to unify as both increase computation times and introduce barriers to their use in real-time epidemiology. There are valid methods to attempt to overcome these issues, through sample selection and downsampling to improve the representativeness of the analysis dataset and reduce its size, and using increasingly large amounts of computing resources to run more complex models. These of course have their costs, in terms of the use of resource to produce unused sequence data and the energy cost of computational effort. These costs, and the appetite for simple visualisation pipelines such as that provided by the Nextstrain platform, suggest that simplifying assumptions are increasingly important to consider, and work to validate and improve simple phylogeographic methods will have a widespread influence on the analysis of pathogen genomic data going forward.

Bibliography

1. Jorgensen, D., Pons-Salort, M., Shaw, A. & Grassly, N. The role of genetic sequencing and analysis in the polio eradication programme. *Virus Evolution* **6.** doi:10.1093/VE/VEAA040 (July 2020).

2. Mourier, T. *et al.* SARS-CoV-2 genomes from Saudi Arabia implicate nucleocapsid mutations in host response and increased viral load. *Nature Communications* **13.** ISSN: 20411723. doi:10.1038/s41467-022-28287-8 (2022).

3. Geidelberg, L. *et al. Genomic epidemiology of a densely sampled COVID-19 outbreak in China* Sept. 2020. doi:10.1101/2020.03.09.20033365.

4. Ragonnet-Cronin, M. *et al.* Genetic evidence for the association between COVID-19 epidemic severity and timing of non-pharmaceutical interventions. *Nature Communications* **12,** 1–7. ISSN: 20411723 (Apr. 2021).

5. Shaw, A. *et al. Early detection of vaccine-derived poliovirus outbreaks using nested PCR and nanopore sequencing in the Democratic Republic of Congo, 2021-2022* 2023.

6. Kitamura, N. *et al.* Primary structure, gene organization and polypeptide expression of poliovirus RNA. *Nature* **291,** 547–553. ISSN: 0028-0836 (June 1981).

7. Racaniello, V. R. & Baltimore, D. Molecular cloning of poliovirus cDNA and determination of the complete nucleotide sequence of the viral genome. *Proceedings of the National Academy of Sciences of the United States of America* **78,** 4887–91. ISSN: 0027-8424 (Aug. 1981).

8. Koch, F. & Koch, G. in *The Molecular Biology of Poliovirus* 3–14 (Springer-Verlag, Vienna, 1985). doi:10.1007/978-3-7091-7000-7{_}1.

9. Enders, J. F., Weller, T. H. & Robbins, F. C. Cultivation of the Lansing Strain of Poliomyelitis Virus in Cultures of Various Human Embryonic Tissues. *Science (New York, N.Y.)* **109,** 85–7. ISSN: 0036-8075 (Jan. 1949).

10. Bodian, D., Morgan, I. M. & Howe, H. A. Differentiation of types of poliomyelitis viruses: III. The grouping of fourteen strains into three basic immunological types. *American Journal of Epidemiology* **49,** 234–247. ISSN: 00029262 (1949).

11. Committee on Typing of National Foundation of Infantile Paralysis. Immunologic classification of poliomyelitis viruses: I. A cooperative program for the typing of one hundred strains. *American Journal of Hygiene* **54,** 191–204. ISSN: 0002-9262 (1951).

12. Salk, J. E. Studies in human subjects on active immunization against poliomyelitis. I. A preliminary report of experiments in progress. *Journal of the American Medical Association* **151,** 1081–98. ISSN: 0002-9955 (Mar. 1953).

13. Sabin, A. B. Present status of attenuated live-virus poliomyelitis vaccine. *Journal of the American Medical Association* **162,** 1589–1596. ISSN: 23768118 (Dec. 1956).

14. World Health Organization. *Forty-first World Health Assembly, Geneva, 2-13 May 1988: verbatim records of plenary meetings, reports of committees.* tech. rep. (Geneva, 1988). <https://apps.who.int/iris/handle/10665/164198>.

15. John, T. J. & Dharmapalan, D. The moral dilemma of the polio eradication programme. *Indian Journal of Medical Ethics,* 01–04. ISSN: 09748466 (Oct. 2019).

16. Bandyopadhyay, A. S., Garon, J., Seib, K. & Orenstein, W. A. *Polio vaccination: Past, present and future* 2015. doi:10.2217/fmb.15.19.

17. Rachlin, A. *et al.* Progress Toward Polio Eradication Worldwide, January 2020April 2022. *MMWR. Morbidity and Mortality Weekly Report* **71,** 650–655. ISSN: 0149-21951545-861X (May 2022).

18. World Health Organization. *WHO statement on the second meeting of the International Health Regulations Emergency Committee concerning the international spread of wild poliovirus* tech. rep. (2014). <https://www.who.int/mediacentre/news/statements/2014/polio-20140505/en/>.

19. Molodecky, N. A. *Modelling of poliovirus transmission in Pakistan and Afghanistan to identify optimal vaccination strategies for eradication* PhD thesis (2018). doi:10.25560/81502. <http://spiral.imperial.ac.uk/handle/10044/1/81502>.

20. Jenkins, H. E. *et al.* Implications of a circulating vaccine-derived poliovirus in Nigeria. *New England Journal of Medicine* **362,** 2360–2369. ISSN: 15334406 (June 2010).

21. World Health Organization. *Summary of the April 2017 meeting of the Strategic Advisory Group of Experts on Immunization* Geneva, 2017. <http://www.who.int/imunization/sage/en/index.html>.

22. Grassly, N. C. *Immunogenicity and effectiveness of routine immunization with 1 or 2 doses of inactivated poliovirus vaccine: Systematic review and meta-analysis* Nov. 2014. doi:10.1093/infdis/jit601.

23. Bandyopadhyay, A. S., Modlin, J. F., Wenger, J. & Gast, C. Immunogenicity of New Primary Immunization Schedules with Inactivated Poliovirus Vaccine and Bivalent Oral Polio Vaccine for the Polio Endgame: A Review. *Clinical Infectious Diseases* **67,** S35–S41. ISSN: 15376591 (2018).

24. Macklin, G. R. *et al.* Evolving epidemiology of poliovirus serotype 2 following withdrawal of the type 2 oral poliovirus vaccine. *Science,* eaba1238. ISSN: 0036-8075 (Mar. 2020).

25. World Health Organization. *Circulating vaccine-derived poliovirus GPEI* 2020. <http://polioeradication.org/polio-today/polio-now/this-week/circulating-vaccine-derived-poliovirus/>.

26. Pan American Health Organization. *Fourth meeting of the EPI Technical Advisory Group (TAG) on eradiaction of poliomyelitis in the Americas* tech. rep. (Pan American Health Organization, Antigua, Guatemala, 1987). <https://www.paho.org/hq/dmdocuments/2016/tag04-1987-FinalReport-Eng.pdf>.

27. World Health Organization. *WHO Removes Nigeria from Polio-Endemic List* 2015. <http://www.who.int/mediacentre/news/releases/2015/nigeria-polio/en/>.

28. Nnadi, C. *et al.* Continued endemic wild poliovirus transmission in security-compromised areas Nigeria, 2016. *Morbidity and Mortality Weekly Report* **66,** 190–193. ISSN: 1545861X (Feb. 2017).

29. Cowger, T. L. *et al.* The role of supplementary environmental surveillance to complement acute flaccid paralysis surveillance for wild poliovirus in Pakistan 20112013. *PLoS ONE* **12.** ISSN: 19326203. doi:10.1371/journal.pone.0180608 (2017).

30. Kroiss, S. J. *et al.* Assessing the sensitivity of the polio environmental surveillance system. *PLoS ONE* **13.** ISSN: 19326203. doi:10.1371/journal.pone.0208336. <https://doi.org/10.1371/journal.pone.0208336> (2018).

31. OReilly, K. M. *et al.* Field sensitivity of acute flaccid paralysis and environmental surveillance for poliovirus in Pakistan: an observational study. *Lancet Infectious Disease* **18,** 176. ISSN: 14712334 (Dec. 2018).

32. Tambini, G. *et al.* Direct Detection of Wild Poliovirus Circulation by Stool Surveys of Healthy Children and Analysis of Community Wastewater. *Journal of Infectious Diseases* **168,** 1510–1514. ISSN: 0022-1899 (Dec. 1993).

33. Dowdle, W. R. & Birmingham, M. E. The biologic principles of poliovirus eradication. *The Journal of infectious diseases* **175 Suppl,** 286–92. ISSN: 0022-1899 (Feb. 1997).

34. Pan American Health Organization. *Poliomyelitis eradication field guide, Third Edition* ISBN: 9275316074. <www.paho.org/en/field-guides> (Washington, D.C., 2006).

35. Vaccine Assessment and Monitoring of the Department of Vaccines and Biologicals. *Guidelines for environmental surveillance of poliovirus circulation* tech. rep. 03 (World Health Organization, Geneva, 2003). <www.who.int/vaccines-documents/>.

36. World Health Organization. *Polio laboratory manual* tech. rep. August (World Health Organization, Geneva, 2004), 1–157. <https://apps.who.int/iris/bitstream/handle/10665/68762/WHO_IVB_04.10.pdf?sequence=1&isAllowed=y>.

37. Victoria, J. G. *et al.* Metagenomic analyses of viruses in the stool of children with acute flaccid paralysis. *J Virol* (2009).

38. Majumdar, M. *et al.* Isolation of Vaccine-Like Poliovirus Strains in Sewage Samples from the United Kingdom. *Journal of Infectious Diseases* **217,** 1222–1230. ISSN: 15376613 (Mar. 2018).

39. Kilpatrick, D. R. *et al.* Poliovirus serotype-specific VP1 sequencing primers. *Journal of Virological Methods* **174,** 128–130. ISSN: 01660934 (June 2011).

40. Minor, P. D. *et al.* Location and primary structure of a major antigenic site for poliovirus neutralization. *Nature* **301,** 674–679. ISSN: 0028-0836 (Feb. 1983).

41. De, L. *et al.* Identification of vaccine-related polioviruses by hybridization with specific RNA probes. *Journal of Clinical Microbiology* **33,** 562–571. ISSN: 00951137 (1995).

42. Oberste, M. S., Maher, K., Kilpatrick, D. R. & Pallansch, M. A. Molecular evolution of the human enteroviruses: correlation of serotype with VP1 sequence and application to picornavirus classification. *Journal of virology* **73,** 1941–8. ISSN: 0022-538X (1999).

43. Chen, Z. *et al.* Cross-neutralizing human anti-poliovirus antibodies bind the recognition site for cellular receptor. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 20242 (2013).

44. Shaw, J. *et al.* Dynamics of Evolution of Poliovirus Neutralizing Antigenic Sites and Other Capsid Functional Domains during a Large and Prolonged Outbreak. *Journal of Virology* **92,** 01949–17. ISSN: 0022-538X (May 2018).

45. Jorba, J., Campagnoli, R., De, L. & Kew, O. Calibration of Multiple Poliovirus Molecular Clocks Covering an Extended Evolutionary Range. *Journal of Virology* **82,** 4429–4440. ISSN: 0022-538X (May 2008).

46. Gavrilin, G. V., Cherkasova, E. A., Lipskaya, G. Y., Kew, O. M. & Agol, V. I. Evolution of Circulating Wild Poliovirus and of Vaccine-Derived Poliovirus in an Immunodeficient Patient: a Unifying Model. *Journal of Virology* **74,** 7381–7390. ISSN: 0022-538X (Aug. 2000).

47. Jorba, J. in, 227–237 (Humana Press, New York, NY, 2016). doi:10.1007/978-1-4939-3292-4. <http://link.springer.com/10.1007/978-1-4939-3292-4>.

48. Semler, B. L. Poliovirus proves IRES-istible in vivo. *Journal of Clinical Investigation* **113,** 1678–1681. ISSN: 0021-9738 (June 2004).

49. Stern, A. *et al.* The Evolutionary Pathway to Virulence of an RNA Virus. *Cell* **169,** 35–46. ISSN: 10974172 (Mar. 2017).

50. Bessaud, M., Joffret, M. L., Blondel, B. & Delpeyroux, F. Exchanges of genomic domains between poliovirus and other cocirculating species C enteroviruses reveal a high degree of plasticity. *Scientific Reports* **6.** ISSN: 20452322. doi:10.1038/srep38831 (2016).

51. Duggal, R., Cuconati, A., Gromeier, M. & Wimmer, E. Genetic recombination of poliovirus in a cell-free system. *Proceedings of the National Academy of Sciences of the United States of America* **94,** 13786–13791. ISSN: 00278424 (Dec. 1997).

52. Runckel, C., Westesson, O., Andino, R. & DeRisi, J. L. Identification and Manipulation of the Molecular Determinants Influencing Poliovirus Recombination. *PLoS Pathogens* **9** (ed Diamond, M. S.) e1003164. ISSN: 15537366 (Feb. 2013).

53. Holmblat, B. *et al.* Nonhomologous recombination between defective poliovirus and coxsackievirus genomes suggests a new model of genetic plasticity for picornaviruses. *mBio* **5,** 1–12. ISSN: 21507511 (2014).

54. Yakovenko, M. L. *et al.* The 2010 outbreak of poliomyelitis in Tajikistan: Epidemiology and lessons learnt. *Eurosurveillance* **19.** ISSN: 15607917. doi:10.2807/1560-7917.ES2014.19.7.20706 (2014).

55. Kingman, J. F. C. On the genealogy of large populations. *Journal of Applied Probability* **19,** 27–43. ISSN: 0021-9002 (July 1982).

56. Takahata, N. The coalescent in two partially isolated diffusion populations. *Genetical Research* **52,** 213–222. ISSN: 14695073 (1988).

57. Notohara, M. The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology* **29,** 59–75. ISSN: 14321416 (Oct. 1990).

58. Vaughan, T. G., Kühnert, D., Popinga, A., Welch, D. & Drummond, A. J. Efficient Bayesian inference under the structured coalescent. *Bioinformatics* **30,** 2272–2279. ISSN: 1367-4803 (Aug. 2014).

59. Volz, E. M. Complex population dynamics and the coalescent under neutrality. *Genetics* **190,** 187–201. ISSN: 00166731 (Jan. 2012).

60. Müller, N. F., Rasmussen, D. A. & Stadler, T. The Structured Coalescent and Its Approximations. *Molecular biology and evolution* **34,** 2970–2981. ISSN: 15371719 (2017).

61. De Maio, N., Wu, C. H., OReilly, K. M. & Wilson, D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genetics* **11,** e1005421. ISSN: 15537404 (Aug. 2015).

62. Gill, M. S., Lemey, P., Bennett, S. N., Biek, R. & Suchard, M. A. Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates. *Systematic Biology* **65,** 1041. ISSN: 1076836X (Nov. 2016).

63. Müller, N. F., Rasmussen, D. & Stadler, T. MASCOT: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics* **34,** 3843–3848. ISSN: 1367-4803 (Nov. 2018).

64. Gill, M. S. *et al.* Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular biology and evolution* **30,** 713–724. ISSN: 1537-1719 (Mar. 2013).

65. Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences of the United States of America* **110.** ISSN: 00278424. doi:10.1073/pnas.1207965110 (2013).

66. Boskova, V., Bonhoeffer, S. & Stadler, T. Inference of Epidemiological Dynamics Based on Simulated Phylogenies Using Birth-Death and Coalescent Models. *PLOS Computational Biology* **10,** e1003913. ISSN: 1553-7358 (Nov. 2014).

67. Kühnert, D., Stadler, T., Vaughan, T. G. & Drummond, A. J. Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data. *Molecular Biology and Evolution* **33,** 2102–2116. ISSN: 0737-4038 (Aug. 2016).

68. Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography takes a relaxed random walk in continuous space and time. *Molecular biology and evolution* **27,** 1877–1885. ISSN: 1537-1719 (2010).

69. Dellicour, S. *et al.* Relax, Keep Walking A Practical Guide to Continuous Phylogeographic Inference with BEAST. *Molecular Biology and Evolution* **38,** 3486–3493. ISSN: 15371719 (July 2021).

70. Pitman, J. Coalescents with multiple collisions. *Annals of Probability* **27.** ISSN: 00911798. doi:10.1214/aop/1022874819 (1999).

71. Felsenstein, J. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Biology* **22,** 240–249. ISSN: 1063-5157 (Sept. 1973).

72. Lemey, P. *et al.* Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS Pathogens* **10.** doi:10.1371/JOURNAL.PPAT.1003932 (2014).

73. Dudas, G. *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544,** 309–315. ISSN: 14764687 (2017).

74. Yang, J., Xie, D., Nie, Z., Xu, B. & Drummond, A. J. Inferring host roles in bayesian phylodynamics of global avian influenza A virus H9N2. *Virology* **538,** 86–96. ISSN: 0042-6822 (Dec. 2019).

75. Akhtar, R. *et al.* Genetic Epidemiology reveals three chronic reservoir areas with recurrent population mobility challenging poliovirus eradication in Pakistan. *Clinical Infectious Diseases*. ISSN: 1058-4838. doi:10.1093/cid/ciz1037 (Oct. 2019).

76. Hadfield, J. *et al.* NextStrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34** (ed Kelso, J.) 4121–4123. ISSN: 14602059 (Dec. 2018).

77. Zurbriggen, S. *et al.* Isolation of Sabin-like polioviruses from wastewater in a country using inactivated polio vaccine. *Applied and Environmental Microbiology* **74,** 5608–5614. ISSN: 00992240 (Sept. 2008).

78. Korotkova, E. A. *et al.* A Cluster of Paralytic Poliomyelitis Cases Due to Transmission of Slightly Diverged Sabin 2 Vaccine Poliovirus. *Journal of Virology* **90,** 5978–5988. ISSN: 0022-538X (July 2016).

79. World Health Organization. *Global Polio Surveillance Action Plan, 2018-2020* tech. rep. (World Health Organization, Geneva, 2019).

80. Dunn, G. *et al.* Twenty-Eight Years of Poliovirus Replication in an Immunodeficient Individual: Impact on the Global Polio Eradication Initiative. *PLoS Pathogens* **11.** ISSN: 15537374. doi:10.1371/journal.ppat.1005114 (2015).

81. Pons-Salort, M. *et al.* Preventing Vaccine-Derived Poliovirus Emergence during the Polio Endgame. *PLoS Pathogens* **12.** ISSN: 15537374. doi:10.1371/journal.ppat.1005728 (July 2016).

82. Famulare, M. *et al.* Sabin Vaccine Reversion in the Field: a Comprehensive Analysis of Sabin-Like Poliovirus Isolates in Nigeria. doi:10.1128/JVI.01532-15 (2015).

83. Macadam, A. I. *et al.* The 5 noncoding region of the type 2 poliovirus vaccine strain contains determinants of attenuation and temperature sensitivity. *Virology* **181,** 451–458. ISSN: 10960341 (1991).

84. Ren, R., Moss, E. G. & Racaniello, V. R. Identification of two determinants that attenuate vaccine-related type 2 poliovirus. *Journal of Virology* **65,** 1377–1382. ISSN: 0022-538X (Mar. 1991).

85. Jegouic, S. *et al.* Recombination between polioviruses and co-circulating coxsackie A viruses: Role in the emergence of pathogenic vaccine-derived polioviruses. *PLoS Pathogens* **5.** ISSN: 15537366. doi:10.1371/journal.ppat.1000412 (May 2009).

86. Burns, C. C. *et al.* Multiple Independent Emergences of Type 2 Vaccine-Derived Polioviruses during a Large Outbreak in Northern Nigeria. *Journal of Virology* **87,** 4907–4922. ISSN: 0022-538X (May 2013).

87. Jorba, J. *et al.* Update on Vaccine-Derived Poliovirus Outbreaks - Worldwide, January 2018-June 2019. *MMWR. Morbidity and mortality weekly report* **68,** 1024–1028. ISSN: 1545861X (Nov. 2019).

88. Pons-Salort, M. *et al.* Population Immunity against Serotype-2 Poliomyelitis Leading up to the Global Withdrawal of the Oral Poliovirus Vaccine: Spatio-temporal Modelling of Surveillance Data. *PLoS Medicine* **13.** ISSN: 15491676. doi:10.1371/journal.pmed.1002140 (Oct. 2016).

89. Blake, I. M. *et al.* Type 2 Poliovirus Detection after Global Withdrawal of Trivalent Oral Vaccine. *New England Journal of Medicine* **379,** 834–845. ISSN: 0028-4793 (Aug. 2018).

90. World Health Organization. *Polio Information System (POLIS)* 2014. <extranet.who.int/polis/public/CaseCount.aspx>.

91. Greene, S. A. *et al.* Progress toward polio eradication-worldwide, january 2017-march 2019. *Morbidity and Mortality Weekly Report* **68,** 458–462. ISSN: 1545861X (May 2019).

92. Thornton, J. Polio returns to the Philippines. *The Lancet* **394,** 1217. ISSN: 01406736 (Oct. 2019).

93. Knowlson, S. *et al.* New Strains Intended for the Production of Inactivated Polio Vaccine at Low-Containment After Eradication. *PLoS Pathogens* **11** (ed Racaniello, V.) e1005316. ISSN: 15537374 (Dec. 2015).

94. Van Damme, P. *et al.* The safety and immunogenicity of two novel live attenuated monovalent (serotype 2) oral poliovirus vaccines in healthy adults: a double-blind, single-centre phase 1 study. *The Lancet* **394,** 148–158. ISSN: 1474547X (July 2019).

95. World Health Organization. *Statement of the Twenty-first IHR Emergency Committee Regarding the International Spread of Poliovirus* 2019. <https://www.who.int/news-room/detail/29-05-2019-statement-of-the-twenty-first-ihr-emergency-committee>.

96. World Health Organization. *Certification of Poliomyelitis Eradication. Fifteenth meeting of the European Regional Commission for the Certification of Poliomyelitis Eradication* tech. rep. (World Health Organization, Geneva, Sept. 2005), 136. <http://www.euro.who.int/en/health-topics/communicable-diseases/poliomyelitis/activities/polio-certification-activities/certification-of-the-regions-polio-free-status-in-2002>.

97. Molodecky, N. A. *et al.* Quantifying movement patterns and vaccination status of high risk mobile populations in Pakistan and Afghanistan to inform poliovirus risk and vaccination strategy. *Vaccine* **39,** 2124–2132. ISSN: 1873-2518 (Apr. 2021).

98. Minin, V. N. & Suchard, M. A. Fast, accurate and simulation-free stochastic mapping. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363,** 3985. ISSN: 09628436 (Dec. 2008).

99. Nathanson, N. & Martin, J. R. The epidemiology of poliomyelitis: Enigmas surrounding its appearance, epidemicity, and disappearance. *American Journal of Epidemiology* **110,** 672–692. ISSN: 00029262 (Dec. 1979).

100. Minin, V. N. & Suchard, M. A. Counting labeled transitions in continuous-time Markov models of evolution. *Journal of mathematical biology* **56,** 391–412. ISSN: 1432-1416 (Mar. 2008).

101. Ward, C. D., Stokes, M. A. & Flanegan, J. B. Direct measurement of the poliovirus RNA polymerase error frequency in vitro. *Journal of virology* **62,** 558–62. ISSN: 0022-538X (Feb. 1988).

102. Asghar, H. *et al.* Environmental surveillance for polioviruses in the Global Polio Eradication Initiative. *The Journal of infectious diseases* **210 Suppl,** S294–S303. ISSN: 1537-6613 (Nov. 2014).

103. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32,** 1792–7. ISSN: 1362-4962 (2004).

104. Larsson, A. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30,** 3276–3278. ISSN: 14602059 (Nov. 2014).

105. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32,** 268–274. ISSN: 1537-1719 (Jan. 2015).

106. Paradis, E. & Schliep, K. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35** (ed Schwartz, R.) 526–528. ISSN: 1367-4803 (Feb. 2019).

107. Li, L. M., Grassly, N. C. & Fraser, C. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. *Molecular biology and evolution* **34,** 2982–2995. ISSN: 15371719 (Nov. 2017).

108. Shulman, L. M. *et al.* Molecular epidemiology of silent introduction and sustained transmission of wild poliovirus type 1, Israel, 2013. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* **19.** ISSN: 1560-7917. doi:10.2807/1560-7917.ES2014.19.7.20709. <https://pubmed.ncbi.nlm.nih.gov/24576471/> (2014).

109. Volz, E. M. & Frost, S. D. W. Scalable relaxed clock phylogenetic dating. *Virus Evolution* **3.** ISSN: 2057-1577. doi:10.1093/ve/vex025 (July 2017).

110. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* **4.** ISSN: 2057-1577. doi:10.1093/ve/vey016 (Jan. 2018).

111. Bielejec, F. *et al.* SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Molecular biology and evolution* **33,** 2167–2169. ISSN: 1537-1719 (Aug. 2016).

112. Mbaeyi, C. *et al.* Progress Toward Poliomyelitis Eradication Pakistan, January 2021July 2022. *MMWR. Morbidity and Mortality Weekly Report* **71,** 1313–1318. ISSN: 0149-21951545-861X (Oct. 2022).

113. Mendes, A. *et al.* Spatial analysis of genetic clusters and epidemiologic factors related to wild poliovirus type 1 persistence in Afghanistan and Pakistan. *PLOS Global Public Health* **2,** e0000251. ISSN: 2767-3375 (June 2022).

114. Institute for disease Modeling. *Polio Immunity Mapping* <https://pim.bmgf.io/>.

115. Al-Mandhari, A. *WHO EMRO — Statement condemning killing of 8 polio health workers in Afghanistan — News — Media centre* 2022. <https://www.emro.who.int/media/news/statement-condemning-killing-of-8-polio-health-workers-in-afghanistan.html>.

116. Din, M., Asghar, M. & Ali, M. Delays in polio vaccination programs due to COVID-19 in Pakistan: a major threat to Pakistan's long war against polio virus. *Public Health* **189,** 1. ISSN: 14765616 (Dec. 2020).

117. Blake, I. M. *et al.* Faster Detection of Poliomyelitis Outbreaks to Support Polio Eradication. *Emerging Infectious Diseases* **22,** 449. ISSN: 10806059 (Mar. 2016).

118. Openshaw, S. *The modifiable areal unit problem* ISBN: 0860941345 (Concepts and Techniques in Modern Geography, 1983).

119. Griffith, D. A. The boundary value problem in spatial statistical analysis. *Journal of Regional Science* **23,** 377–387. ISSN: 14679787 (Aug. 1983).

120. Bedford, T. *et al.* Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* **523,** 217. ISSN: 14764687 (July 2015).

121. Bourhy, H. *et al.* The origin and phylogeography of dog rabies virus. *Journal of General Virology.* ISSN: 00221317. doi:10.1099/vir.0.2008/003913-0 (2008).

122. Faria, N. R. *et al.* The early spread and epidemic ignition of HIV-1 in human populations. *Science (New York, N.Y.)* **346,** 56. ISSN: 10959203 (Oct. 2014).

123. Grubaugh, N. D. *et al.* Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature 2017 546:7658* **546,** 401–405. ISSN: 1476-4687 (May 2017).

124. Motayo, B. O., Oluwasemowo, O. O., Olusola, A., Opayele, A. V. & Faneye, O. Phylogeography and evolutionary analysis of African Rotavirus a genotype G12 reveals district genetic diversification within lineage III. doi:10.1016/j.heliyon.2019.e02680 (2017).

125. Ehichioya, D. U. *et al.* Phylogeography of Lassa Virus in Nigeria. *Journal of Virology* **93,** 929–948. ISSN: 0022-538X (Nov. 2019).

126. Wiens, J. J. Character analysis in morphological phylogenetics: Problems and solutions. *Systematic Biology* **50,** 689–699. ISSN: 10635157 (2001).

127. Cunningham, C. W., Omland, K. E. & Oakley, T. H. Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology & Evolution* **13,** 361–366. ISSN: 0169-5347 (Sept. 1998).

128. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS Computational Biology.* ISSN: 1553734X. doi:10.1371/journal.pcbi.1000520 (2009).

129. Lewis, P. O. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Systematic Biology* **50,** 913–925. ISSN: 1076-836X (Nov. 2001).

130. Pagel, M. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B: Biological Sciences* **255.** ISSN: 14712970. doi:10.1098/rspb.1994.0006 (1994).

131. Tavaré, S. *Some probabilistic and statistical problems in the analysis of DNA sequences* 1986. doi:10.3/JQUERY-UI.JS.

132. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution* **4.** ISSN: 20571577. doi:10.1093/VE/VEX042 (Jan. 2018).

133. Simini, F., González, M. C., Maritan, A. & Barabási, A. L. A universal model for mobility and migration patterns. *Nature* **484,** 96–100. ISSN: 00280836 (Feb. 2012).

134. Poot, J., Alimi, O., Cameron, M. P. & Maré, D. C. The Gravity Model of Migration: The Successful Comeback of an Ageing Superstar in Regional Science (2016).

135. Slobodkin, L. *Growth and Regulation of Animal Populations* (eds Holt, R. & Holt, W.) ISBN: 0030104254 (New York, NY, 1961).

136. Edwards, C. J. *et al.* Ancient hybridization and an irish origin for the modern polar bear matriline. *Current Biology* **21.** ISSN: 09609822. doi:10.1016/j.cub.2011.05.058 (2011).

137. Bloomfield, S. *et al.* Investigation of the validity of two Bayesian ancestral state reconstruction models for estimating Salmonella transmission during outbreaks. *PLOS ONE* **14,** e0214169. ISSN: 1932-6203 (July 2019).

138. Kingman, J. F. C. The coalescent. *Stochastic Processes and their Applications* **13,** 235–248. ISSN: 03044149 (Sept. 1982).

139. Beerli, P. & Felsenstein, J. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152,** 763–773. ISSN: 0016-6731 (1999).

140. Ewing, G., Nicholls, G. & Rodrigo, A. Using Temporally Spaced Sequences to Simultaneously Estimate Migration Rates, Mutation Rate and Population Sizes in Measurably Evolving Populations. *Genetics* **168,** 2407. ISSN: 00166731 (Dec. 2004).

141. Stadler, T. Lineages-through-time plots of neutral models for speciation. *Mathematical Biosciences* **216,** 163–171. ISSN: 0025-5564 (Dec. 2008).

142. Guindon, S., Guo, H. & Welch, D. Demographic inference under the coalescent in a spatial continuum. *Theoretical Population Biology* **111.** ISSN: 10960325. doi:10.1016/j.tpb.2016.05.002 (2016).

143. Dellicour, S., Lemey, P., Suchard, M. A., Gilbert, M. & Baele, G. *Accommodating sampling location uncertainty in continuous phylogeography* 2022. doi:10.1093/ve/veac041.

144. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17,** 368–376. ISSN: 00222844 (1981).

145. Jukes, T. H. & Cantor, C. R. in *Mammalian protein metabolism* 21–132 (1969).

146. Barry, D. & Hartigan, J. A. Asynchronous Distance between Homologous DNA Sequences. *Biometrics* **43,** 261. ISSN: 0006341X (June 1987).

147. Barry, D. & Hartigan, J. A. Statistical Analysis of Hominoid Molecular Evolution. *https://doi.org/10.1214/ss/1177013353* **2,** 191–207. ISSN: 0883-4237 (May 1987).

148. Zou, L., Susko, E., Field, C. & Roger, A. J. The Parameters of the Barry and Hartigan General Markov Model Are Statistically NonIdentifiable. *Systematic Biology* **60,** 872–875. ISSN: 1063-5157 (Dec. 2011).

149. Harmon, L. in *Phylogenetic Comparative Methods* chap. 7 (2019). ISBN: 978-1719584463. <https://lukejharmon.github.io/pcm/chapter7_introdiscrete/>.

150. Weyenberg, G. & Yoshida, R. in *Algebraic and Discrete Mathematical Methods for Modern Biology* (ed Robeva, R. S.) 1–368 (Elsevier, Mar. 2015). ISBN: 9780128012710. doi:10.1016/C2013-0-18496-6.

151. Faria, N. R., Suchard, M. A., Rambaut, A., Streicker, D. G. & Lemey, P. Simultaneously reconstructing viral crossspecies transmission history and identifying the underlying constraints. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368.** ISSN: 14712970. doi:10.1098/rstb.2012.0196 (2013).

152. Norris, J. R. Markov Chains. doi:10.1017/CBO9780511810633 (Feb. 1997).

153. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology* **15** (ed Pertea, M.) e1006650. ISSN: 15537358 (Apr. 2019).

154. Vaughan, T. G. & Drummond, A. J. A stochastic simulator of birth-death master equations with application to phylodynamics. *Molecular Biology and Evolution* **30,** 1480–1493. ISSN: 07374038 (June 2013).

155. Neher, R. *Inference of transition between discrete characters and mugration models Tree-Time 0.9.0-b.1 documentation* 2019. <https://treetime.readthedocs.io/en/latest/tutorials/mugration.html>.

156. MacPherson, A., Louca, S., McLaughlin, A., Joy, J. B. & Pennell, M. W. Unifying Phylogenetic BirthDeath Models in Epidemiology and Macroevolution. *Systematic Biology* **71,** 172–189. ISSN: 1063-5157 (Dec. 2021).

157. Layan, M. *et al.* Impact and mitigation of sampling bias to determine viral spread: Evaluating discrete phylogeography through CTMC modeling and structured coalescent model approximations. *Virus Evolution* **9.** ISSN: 2057-1577. doi:10.1093/VE/VEAD010 (Mar. 2023).

158. Kalkauskas, A. *et al.* Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLoS Computational Biology* **17.** ISSN: 15537358. doi:10.1371/JOURNAL.PCBI.1008561 (2021).

159. Barton, N. H., Etheridge, A. M. & Véber, A. A new model for evolution in a spatial continuum. *Electronic Journal of Probability* **15.** ISSN: 10836489. doi:10.1214/EJP.v15-741 (2010).

160. Freeman, N. The segregated Λ-coalescent. *Annals of Probability* **43.** ISSN: 00911798. doi:10.1214/13-AOP857 (2015).

161. Montano, V. Coalescent inferences in conservation genetics: Should the exception become the rule? *Biology Letters* **12.** ISSN: 1744957X. doi:10.1098/rsbl.2016.0211 (2016).

162. Inward, R. P., Parag, K. V. & Faria, N. R. Using multiple sampling strategies to estimate SARS-CoV-2 epidemiological parameters from genomic sequencing data. *Nature Communications* **13.** ISSN: 20411723. doi:10.1038/s41467-022-32812-0 (2022).

163. Faria, N. R. *et al.* Phylogeographical footprint of colonial history in the global dispersal of human immunodeficiency virus type 2 group A. *Journal of General Virology* **93.** ISSN: 00221317. doi:10.1099/vir.0.038638-0 (2012).

164. Global Polio Eradication Initiative. *Polio endgame Strategy 2019-2023* tech. rep. (Geneva, 2019).

165. Seidel, S., Stadler, T. & Vaughan, T. G. Estimating disease spread using structured coalescent and birth-death models: A quantitative comparison. doi:10.1101/2020.11.30.403741 (2020).

166. Hamisu, A. W. *et al. Characterizing Environmental Surveillance Sites in Nigeria and Their Sensitivity to Detect Poliovirus and Other Enteroviruses* in *Journal of Infectious Diseases* **225** (2022), 1377–1386. doi:10.1093/infdis/jiaa175. <https://academic.oup.com/jid/article-abstract/doi/10.1093/infdis/jiaa175/5818305>.

167. UK Health Security Agency. *Acute flaccid paralysis (AFP) Information for patients* tech. rep. (2021). <https://www.gov.uk/government/publications/acute-flaccid-paralysis-information-for-patients/acute-flaccid-paralysis-afp-information-for-patients>.

168. World Health Organization. Acute flaccid paralysis (AFP) surveillance : the surveillance strategy for poliomyelitis eradication = Surveillance de la paralysie flasque aiguë (PFA) : stratégie de surveillance pour léradication de la poliomyélite. *Weekly Epidemiological Record = Relevé épidémiologique hebdomadaire* **73,** 113–114 (1998).

169. World Health Organization. *Global Polio Surveillance Action Plan 2022-2024* tech. rep. (World Health Organization, Geneva, 2022). <https://apps.who.int/iris/bitstream/handle/10665/354479/9789240047310-eng.pdf?sequence=1&isAllowed=y>.

170. Tangermann, R. H., Lamoureux, C., Tallis, G. & Goel, A. The critical role of acute flaccid paralysis surveillance in the Global Polio Eradication Initiative. *International Health* **9,** 156–163. ISSN: 1876-3413 (May 2017).

171. Bondarenko, M., Kerr, D., Sorichetta, A. & Tatem, A. Census/projection-disaggregated gridded population datasets for 189 countries in 2020 using Built-Settlement Growth Model (BSGM) outputs. doi:10.5258/SOTON/WP00684 (Sept. 2020).

172. Global Polio Eradication Initiative. *Best practice for monitoring the quality of polio eradication campaign performance* tech. rep. (World Health Organization, Geneva, 2018). <http://polioeradication.org/wp-content/uploads/2018/02/best-practices-20180207-02.pdf>.

173. Bielejec, F., Lemey, P., Baele, G., Rambaut, A. & Suchard, M. A. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. *Systematic biology* **63,** 493–504. ISSN: 1076-836X (2014).

174. Grenfell, B. T. *et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens* 2004. doi:10.1126/science.1090727.

175. Frost, S. D. *et al.* Eight challenges in phylodynamic inference. *Epidemics* **10,** 88–92. ISSN: 18780067 (2015).

176. Pybus, O. G., Rambaut, A. & Harvey, P. H. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155,** 1429. ISSN: 00166731 (July 2000).

177. Strimmer, K. & Pybus, O. G. Exploring the Demographic History of DNA Sequences Using the Generalized Skyline Plot. *Molecular Biology and Evolution* **18,** 2298–2305. ISSN: 0737-4038 (Dec. 2001).

178. Volz, E. M. & Frost, S. D. Sampling through time and phylodynamic inference with coalescent and birth-death models. *Journal of the Royal Society, Interface* **11.** ISSN: 1742-5662. doi:10.1098/RSIF.2014.0945 (Dec. 2014).

179. Karcher, M. D., Carvalho, L. M., Suchard, M. A., Dudas, G. & Minin, V. N. Estimating effective population size changes from preferentially sampled genetic sequences. *PLoS Computational Biology* **16.** ISSN: 15537358. doi:10.1371/journal.pcbi.1007774 (2020).

180. Volz, E. M. & Didelot, X. Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. *Systematic Biology* **67,** 719–728. ISSN: 1076836X (July 2018).

181. Cappello, L. & Palacios, J. A. Adaptive Preferential Sampling in Phylodynamics With an Application to SARS-CoV-2. *Journal of Computational and Graphical Statistics* **31.** ISSN: 15372715. doi:10.1080/10618600.2021.1987256 (2022).

182. Parag, K. V., Du Plessis, L. & Pybus, O. G. Jointly Inferring the Dynamics of Population Size and Sampling Intensity from Molecular Sequences. *Molecular Biology and Evolution* **37,** 2414. ISSN: 15371719 (2020).

183. Volz, E. M. & Siveroni, I. Bayesian phylodynamic inference with complex models. *PLoS Computational Biology* **14.** ISSN: 15537358. doi:10.1371/journal.pcbi.1006546 (2018).

184. Ministry of Health - Kingdom of Saudi Arabia. *MOH Reports First Case of Coronavirus Infection* 2020. <https://www.moh.gov.sa/en/Ministry/MediaCenter/News/Pages/News-2020-03-02-002.aspx>.

185. Memish, Z. A., Aljerian, N. & Ebrahim, S. H. *Tale of three seeding patterns of SARS-CoV-2 in Saudi Arabia* Jan. 2021. doi:10.1016/S1473-3099(20)30425-4.

186. U.N. Population Division. *World urbanization prospects : the 2018 revision* ISBN: 9789211483192 (New York, 2019).

187. Jombart, T. *et al. Inferring the number of COVID-19 cases from recently reported deaths* Mar. 2020. doi:10.1101/2020.03.10.20033761.

188. Russell, T. W. *et al.* Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. *BMC Medicine* **18,** 332. ISSN: 17417015 (Dec. 2020).

189. Seemann, T. *et al.* Tracking the COVID-19 pandemic in Australia using genomics. *Nature Communications* **11.** ISSN: 20411723. doi:10.1038/s41467-020-18314-x (2020).

190. Volz, E. M. *et al. Report 5: Phylogenetic analysis of SARS-CoV-2, 15 February 2020* tech. rep. February (Imperial College London, 2020), 2–9. doi:10.25561/77169. <https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-5-phylogenetics-of-sars-cov-2/>.

191. Miller, D. *et al.* Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nature Communications 2020 11:1* **11,** 1–10. ISSN: 2041-1723 (Nov. 2020).

192. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data from vision to reality. *Eurosurveillance* **22,** 30494. ISSN: 1560-7917 (Mar. 2017).

193. Hu, Z. *et al.* Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in Nanjing, China. *Science China Life Sciences* **63,** 706–711. ISSN: 18691889 (May 2020).

194. Hasegawa, M., Kishino, H. & Yano, T. a. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22,** 160–174. ISSN: 00222844 (Oct. 1985).

195. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35,** 518–522. ISSN: 0737-4038 (Feb. 2018).

196. Duchene, S. *et al.* Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evolution* **6.** doi:10.1093/VE/VEAA061 (July 2020).

197. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10,** 512–526. ISSN: 0737-4038 (May 1993).

198. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27,** 592–593. ISSN: 1460-2059 (Feb. 2011).

199. Endo, A., Abbott, S., Kucharski, A. J. & Funk, S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Research* **5,** 67. ISSN: 2398-502X (July 2020).

200. Flaxman, S. *et al.* Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. doi:10.25561/77731. <https://doi.org/10.25561/77731.> (2020).

201. Alshammari, F. S. A Mathematical Model to Investigate the Transmission of COVID-19 in the Kingdom of Saudi Arabia. *Computational and Mathematical Methods in Medicine* **2020.** ISSN: 17486718. doi:10.1155/2020/9136157 (2020).

202. Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. Superspreading and the effect of individual variation on disease emergence. *Nature* **438,** 355–359. ISSN: 0028-0836 (Nov. 2005).

203. McAloon, C. *et al.* Incubation period of COVID-19: A rapid systematic review and meta-analysis of observational research. *BMJ Open* **10,** e039652. ISSN: 20446055 (Aug. 2020).

204. Nie, X. *et al.* Epidemiological characteristics and incubation period of 7015 confirmed cases with coronavirus disease 2019 outside Hubei Province in China. *Journal of Infectious Diseases* **222,** 26–35 (June 2020).

205. Lauer, S. A. *et al.* The incubation period of coronavirus disease 2019 (CoVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine* **172.** ISSN: 15393704. doi:10.7326/M20-0504 (2020).

206. Müller, N. F. & Bouckaert, R. R. Adaptive metropolis-coupled MCMC for BEAST 2. *PeerJ* **8.** ISSN: 21678359. doi:10.7717/PEERJ.9473/SUPP-6 (2020).

207. Banjar, A. *et al.* Seroprevalence of antibodies to SARS-CoV-2 among blood donors in the early months of the pandemic in Saudi Arabia. *International Journal of Infectious Diseases* **104,** 452–457 (2021).

208. Alharbi, N. K. *et al.* Nationwide Seroprevalence of SARS-CoV-2 in Saudi Arabia. *Journal of Infection and Public Health* **14.** ISSN: 1876035X. doi:10.1016/j.jiph.2021.04.006 (2021).

209. Mahallawi, W. H. & Al-Zalabani, A. H. The seroprevalence of SARS-CoV-2 IgG antibodies among asymptomatic blood donors in Saudi Arabia. *Saudi Journal of Biological Sciences* **28.** ISSN: 1319562X. doi:10.1016/j.sjbs.2020.12.009 (2020).

210. Alandijany, T. A. *et al.* Lack of Antibodies to SARS-CoV-2 among Blood Donors during COVID-19 Lockdown: A Study from Saudi Arabia. *Healthcare* **9,** 51. ISSN: 2227-9032 (2021).

211. Ahmed, W. A. *et al.* Seroprevalence of Neutralizing Antibodies to Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) among Healthcare Workers in Makkah, Saudi Arabia. *Journal of King Saud University - Science* **33,** 101366. ISSN: 10183647 (Feb. 2021).

212. Alserehi, H. A. *et al.* Seroprevalence of SARS-CoV-2 (COVID-19) among healthcare workers in Saudi Arabia: comparing case and control hospitals. *Diagnostic Microbiology and Infectious Disease* **99,** 115273. ISSN: 18790070 (Mar. 2021).

213. Alrasheed, H., Althnian, A., Kurdi, H., Al-Mgren, H. & Alharbi, S. COVID-19 spread in Saudi Arabia: Modeling, simulation and analysis. *International Journal of Environmental Research and Public Health* **17,** 1–24. ISSN: 16604601 (2020).

214. Sayed, A. A. The progressive public measures of Saudi Arabia to tackle COVID-19 and limit its spread. *International Journal of Environmental Research and Public Health* **18,** 1–9. ISSN: 16604601 (2021).

215. Godinho, V. *Saudi declares complete curfew across the kingdom during Eid Al Fitr holidays* May 2020.

216. Al-Khani, A. M., Khalifa, M. A., Almazrou, A. & Saquib, N. The SARS-CoV-2 pandemic course in Saudi Arabia: A dynamic epidemiological model. *Infectious Disease Modelling* **5,** 766–771. ISSN: 24680427 (Jan. 2020).

217. Walsh, K. A. *et al. The duration of infectiousness of individuals infected with SARS-CoV-2* Dec. 2020. doi:10.1016/j.jinf.2020.10.009.

218. Golding, J., Northstone, K., Miller, L. L., Smith, G. D. & Pembrey, M. Differences between blood donors and a population sample: Implications for case-control studies. *International Journal of Epidemiology* **42,** 1145–1156. ISSN: 03005771 (Aug. 2013).

219. Guarnaccia, C., Giannone, F., Falgares, G., Caligaris, A. O. & Sales-Wuillemin, E. Differences in social representation of blood donation between donors and non-donors: An empirical study. *Blood Transfusion* **14,** 490–499. ISSN: 17232007 (Nov. 2016).

220. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology 2021 39:11* **39,** 1348–1365. ISSN: 1546-1696 (Nov. 2021).

221. Grad, Y. H. & Lipsitch, M. *Epidemiologic data and pathogen genome sequences: A powerful synergy for public health* 2014. doi:10.1186/s13059-014-0538-4.

222. Attwood, S. W., Hill, S. C., Aanensen, D. M., Connor, T. R. & Pybus, O. G. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nature Reviews Genetics 2022 23:9* **23,** 547–562. ISSN: 1471-0064 (Apr. 2022).

223. Rivas, A. L. *et al.* Critical response time (time available to implement effective measures for epidemic control): Model building and evaluation. *Canadian Journal of Veterinary Research* **67,** 307. ISSN: 08309000 (Oct. 2003).

224. The Lancet. *Polio eradication: falling at the final hurdle?* Oct. 2022. doi:10.1016/S0140-6736(22)01875-X.

225. Klapsa, D. *et al.* Sustained detection of type 2 poliovirus in London sewage between February and July, 2022, by enhanced environmental surveillance. *The Lancet* **400,** 1531–1538. ISSN: 1474547X (Oct. 2022).

226. Cooper, L. V. *et al.* Risk factors for the spread of vaccine-derived type 2 polioviruses after global withdrawal of trivalent oral poliovirus vaccine and the effects of outbreak responses with monovalent vaccine: a retrospective analysis of surveillance data for 51 countries i. *The Lancet Infectious Diseases* **22,** 284–294. ISSN: 14744457 (Feb. 2022).

227. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* 221. ISBN: 0387981411 (Springer Science & Business Media, 2009).

228. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8** (ed McInerny, G.) 28–36. ISSN: 2041210X (Jan. 2017).

229. Rockett, R. J. *et al.* Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. *Nature Medicine 2020 26:9* **26,** 1398–1404. ISSN: 1546-170X (July 2020).

230. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372.** ISSN: 10959203. doi:10.1126/science.abh2644 (2021).

231. Tegally, H. *et al.* Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. *Nature Medicine* **28,** 1785–1790. ISSN: 1546170X (June 2022).

232. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature 2021 593:7858* **593,** 266–269. ISSN: 1476-4687 (Mar. 2021).

233. Brito, A. F. *et al.* Global disparities in SARS-CoV-2 genomic surveillance. *Nature Communications 2022 13:1* **13,** 1–13. ISSN: 2041-1723 (Nov. 2022).

234. Maxmen, A. One million coronavirus sequences: popular genome site hits mega milestone. *Nature* **593,** 21. ISSN: 14764687 (May 2021).

235. World Health Organization. *Global genomic surveillance strategy for pathogens with pandemic and epidemic potential 20222032* tech. rep. (Geneva, 2022). doi:10.2471/BLT.22.288220. <https://www.who.int/publications/i/item/9789240046979>.

236. Lefterova, M. I., Suarez, C. J., Banaei, N. & Pinsky, B. A. Next-Generation Sequencing for Infectious Disease Diagnosis and Management: A Report of the Association for Molecular Pathology. *Journal of Molecular Diagnostics* **17.** ISSN: 19437811. doi:10.1016/j.jmoldx.2015.07.004 (2015).

237. Belman, S., Saha, S. & Beale, M. A. SARS-CoV-2 genomics as a springboard for future disease mitigation in LMICs. *Nature Reviews Microbiology 2021 20:1* **20,** 3–3. ISSN: 1740-1534 (Nov. 2021).

238. Gowers, G. O. F. *et al.* Entirely Off-Grid and Solar-Powered DNA Sequencing of Microbial Communities during an Ice Cap Traverse Expedition. *Genes* **10.** ISSN: 2073-4425. doi:10.3390/GENES10110902 (Nov. 2019).

239. Shaw, A. *et al.* Time Taken to Detect and Respond to Polio Outbreaks in Africa and the Potential Impact of Direct Molecular Detection and Nanopore Sequencing. *The Journal of infectious diseases* **226.** ISSN: 15376613. doi:10.1093/infdis/jiab518 (2022).

240. Balinandi, S. *et al. September 2022 Sudan Ebola virus disease outbreak in Uganda - Ebolavirus - Virological* Oct. 2022. <https://virological.org/t/september-2022-sudan-ebola-virus-disease-outbreak-in-uganda/902/3>.

241. Isidro, J. *et al.* Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus. *Nature medicine* **28,** 1569–1572. ISSN: 1546-170X (Aug. 2022).

242. Keita, A. *et al.* Resurgence of Ebola virus in 2021 in Guinea suggests a new paradigm for outbreaks. *Nature* **597.** ISSN: 14764687. doi:10.1038/s41586-021-03901-9 (2021).

243. Furlone, M. *et al.* Polioviruses with natural recombinant genomes isolated from vaccine-associated paralytic poliomyelitis. *Virology* **196,** 199–208. ISSN: 10960341 (1993).

244. Emini, E. A., Elzinga, M. & Wimmer, E. Carboxy-terminal analysis of poliovirus proteins: termination of poliovirus RNA translation and location of unique poliovirus polyprotein cleavage sites. *Journal of virology* **42,** 194–199. ISSN: 0022-538X (Apr. 1982).

245. Gumede, N. *et al. Phylogeny of imported and reestablished wild polioviruses in the democratic republic of the Congo from 2006 to 2011* Nov. 2014. doi:10.1093/infdis/jiu375.

246. Shaukat, S. *et al.* Molecular characterization and phylogenetic relationship of wild type 1 poliovirus strains circulating across Pakistan and Afghanistan bordering areas during 2010-2012. *PLoS ONE* **9,** e107697. ISSN: 19326203 (2014).

Appendix A

Supplementary information for chapter 2

A.1 Investigation of outlying sequences

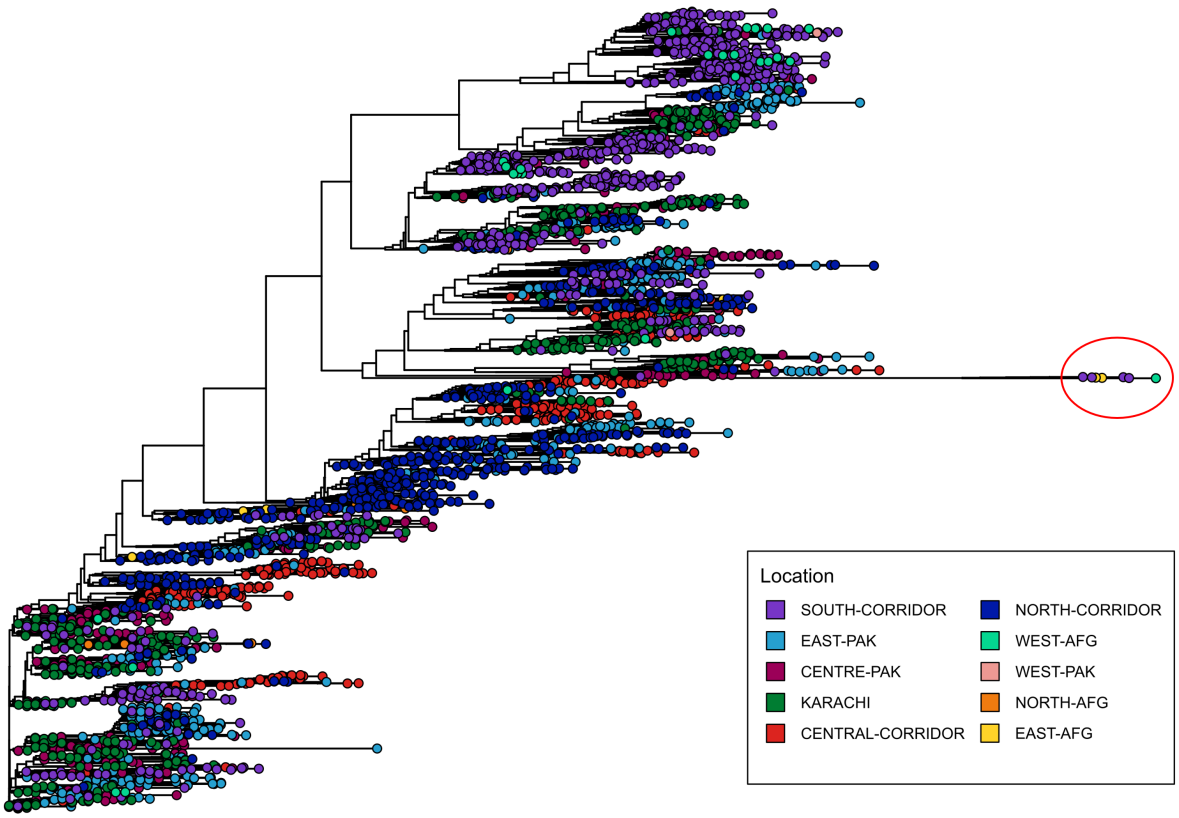


Figure A.1: Neighbour joining tree of the sequence data included in this analysis with tips coloured by region. The outlying sequences of interest are highlighted with a red circle. The x-axis shows genetic distance calculated under a simple JC69 model with the phangorn R package.

Creating a simple neighbour joining tree of the sequence data available results in a long branch to a small cluster of related sequences (supplementary figure A.1). This may suggest misclassification or sequencing error. The selected sequences share non-synonymous differences to the remaining sequences; an ACA to GCC at position 58-60, a GCC to AGC at position 64-66 a TCC to GCC at position 94-96 and a GTC to ATC at position 412-414. These sequences were all collected in Afghanistan in 2012 and are associated with paralysis cases with the seven

sequences linked to six individuals.

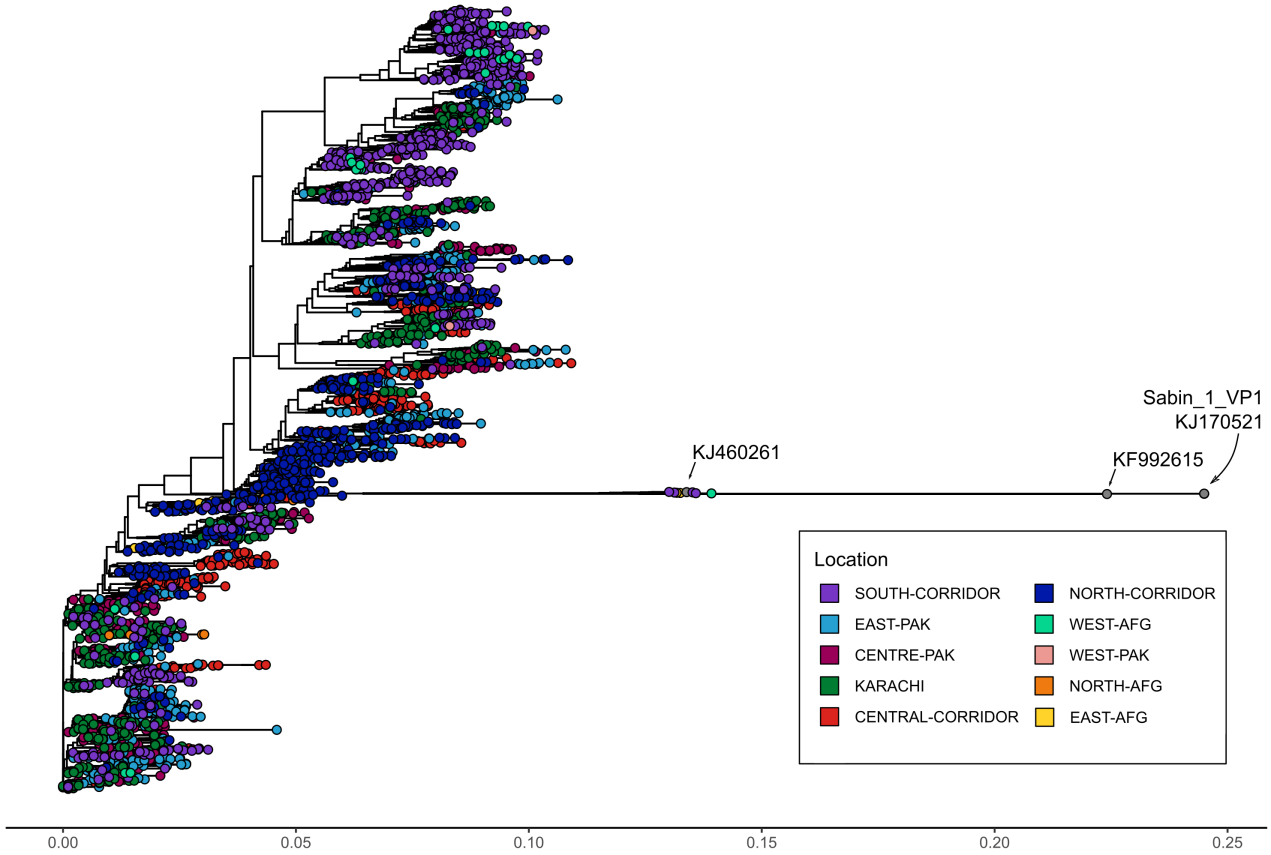


Figure A.2: Neighbour joining tree of the sequences included in the dataset (as in supplementary figure A.1) with additional background sequences from Genbank highlighted in grey and labelled. These background sequences are KJ460261, a wild poliovirus 1 sequence from the South Corridor region in 2011, a year before the included data, KF992615, a wild poliovirus 1 sequence circulating in Nigeria at the same time, a vaccine virus, Sabin 1 (accession V01159) and KJ170521 a vaccine derived virus circulating in Nigeria in 2011.

A number of possible reasons for these outlying sequences were investigated. The first was the possibility that these could be circulating vaccine derived poliovirus misclassified as wild type. As the Sabin oral polio vaccine is inactivated it can regain virulence and begin to spread in the population after administration [243]. Adding the VP1 region from the Sabin vaccine strain for type 1 poliovirus (Sabin 1) (aligned and trimmed from a full genome sequence, Genbank accession no. V01150) results in the common ancestor of Sabin 1 and the wild type sequences included here being within this cluster but with a long branch to Sabin, suggesting that these are unlikely to be vaccine derived unless they had been circulating for a long period prior to

the data included here [244]. Including wild and circulating vaccine derived poliovirus VP1 sequences from Nigeria, collected in 2011 (Genbank accession nos. KF992615 and KJ170521 respectively) gives a similar result, suggesting these are not likely to be imported [82, 245]. We also investigated the possibility that this cluster could have been previously circulating more widely in Afghanistan and was replaced by the more closely related sequences seen in the remainder of the phylogeny. An additional 2011 WPV1 VP1 sequence from Kandahar, Afghanistan (Genbank accession no. KJ460261) clusters with these sequences, suggesting these represent the end of a previously circulating cluster of transmission in the region [246]. These additional background sequences are shown in grey and labelled in supplementary figure A.3. Ultimately, in a maximum likelihood tree with a deep root to the remaining sequences, in 2006, these sequences are consistent with a clock rate on the order of 10^{-2} substitutions per site per year, as expected for poliovirus, and as such were retained for the full analysis [45].

A.2 International sequences

As the Pakistan-Afghanistan region is currently the only endemic region, importation from other reservoir regions, as we see for other infectious diseases, is not possible. Towards the beginning of the included data, transmission of WPV1 was seen in other countries. Here we align the included sequences with publicly available international sequences from GenBank extracted by Candido DS et al. (in prep) and produce a maximum likelihood phylogeny with IQTree which is dated with the R package treedater [105, 109]. This phylogeny shows that the majority of international transmission clusters distinctly from the included sequences, with a small number of international sequences from Shulman et al. 2014, collected in Israel and Egypt, which cluster together with the Pakistan-Afghanistan sequences [108]. The virus transmission in this region has previously been inferred to most likely have originated in Pakistan and to have spread via Egypt to Israel, distinct from other ongoing WPV1 outbreaks in Africa at the time [108].

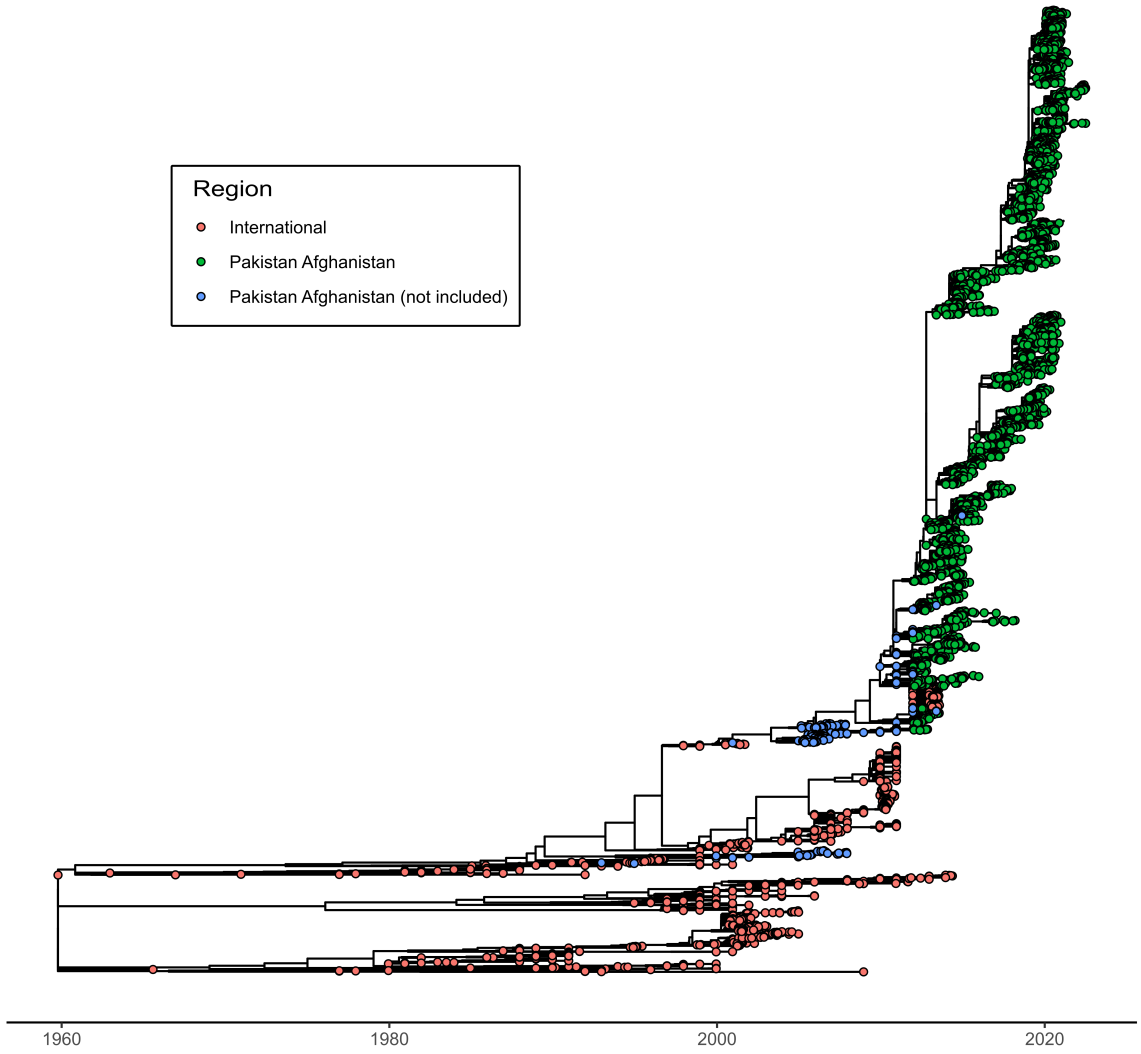


Figure A.3: Maximum likelihood tree showing the included sequences in the context of historic global transmission of WPV1 which can be reconstructed from publicly available sequence data.

This work uses a background set of sequences available on GenBank (`ncbi.nlm.nih.gov/genbank/`). The accession numbers are reproduced on github (`github.com/JorgensenD/PhD_thesis/blob/main/background_accession_WPV1`)

A.3 Supporting information for inferred transitions

In the main text, transitions are included on figure 2.5 where there is Bayes factor support of 10 or more for a transition rate greater than zero. The Bayes factors, probabilities of transition rates greater than zero and median inferred number of transitions are included in table A.1. The highest posterior density interval is calculated from the posterior distribution of inferred transitions between locations across the set of 100 pre-estimated phylogenies.

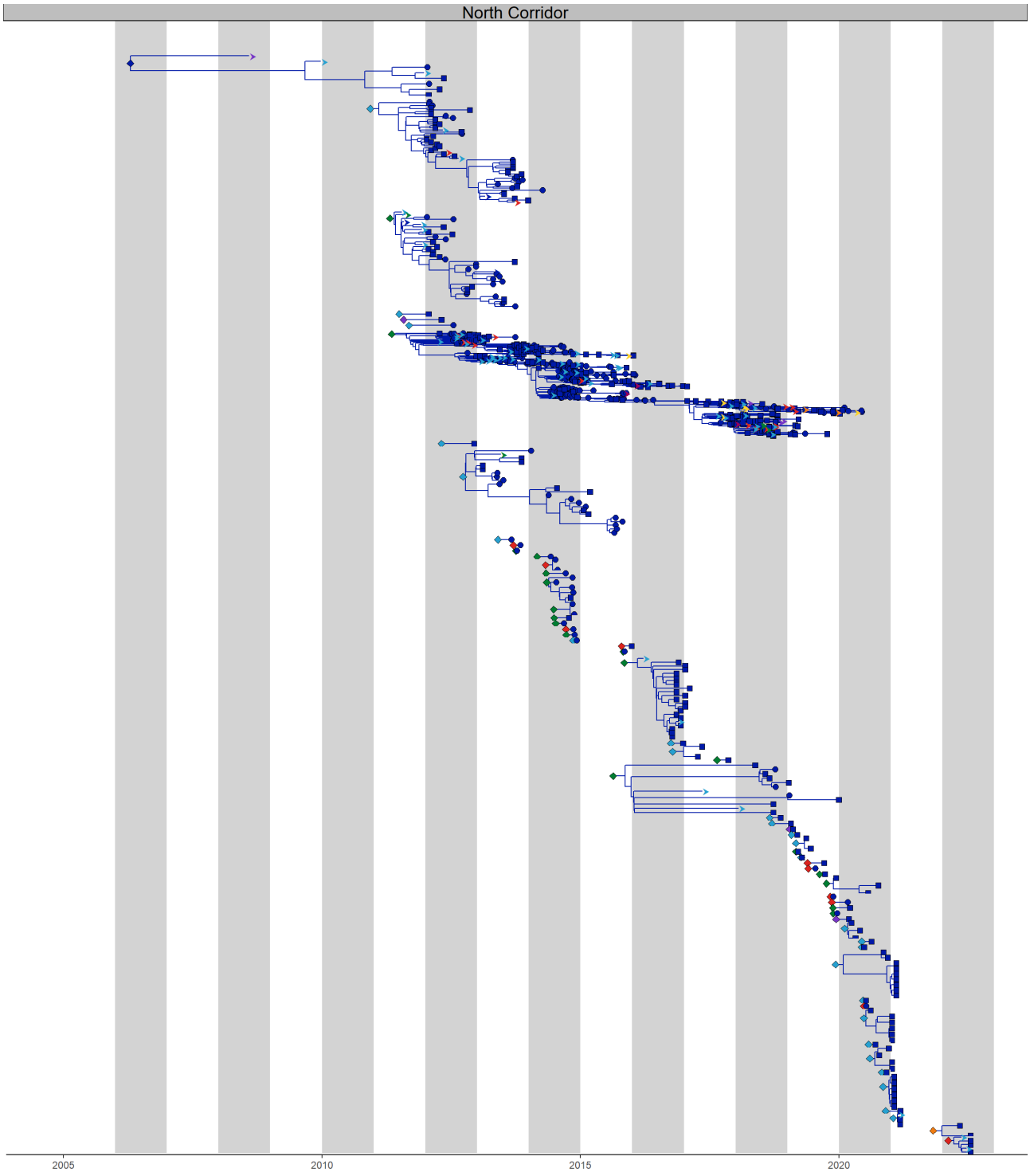
Table A.1: Supporting information for the inferred transitions with Bayes factor support above 10 with the complete dataset of WPV1 detections by ES and AFP surveillance.

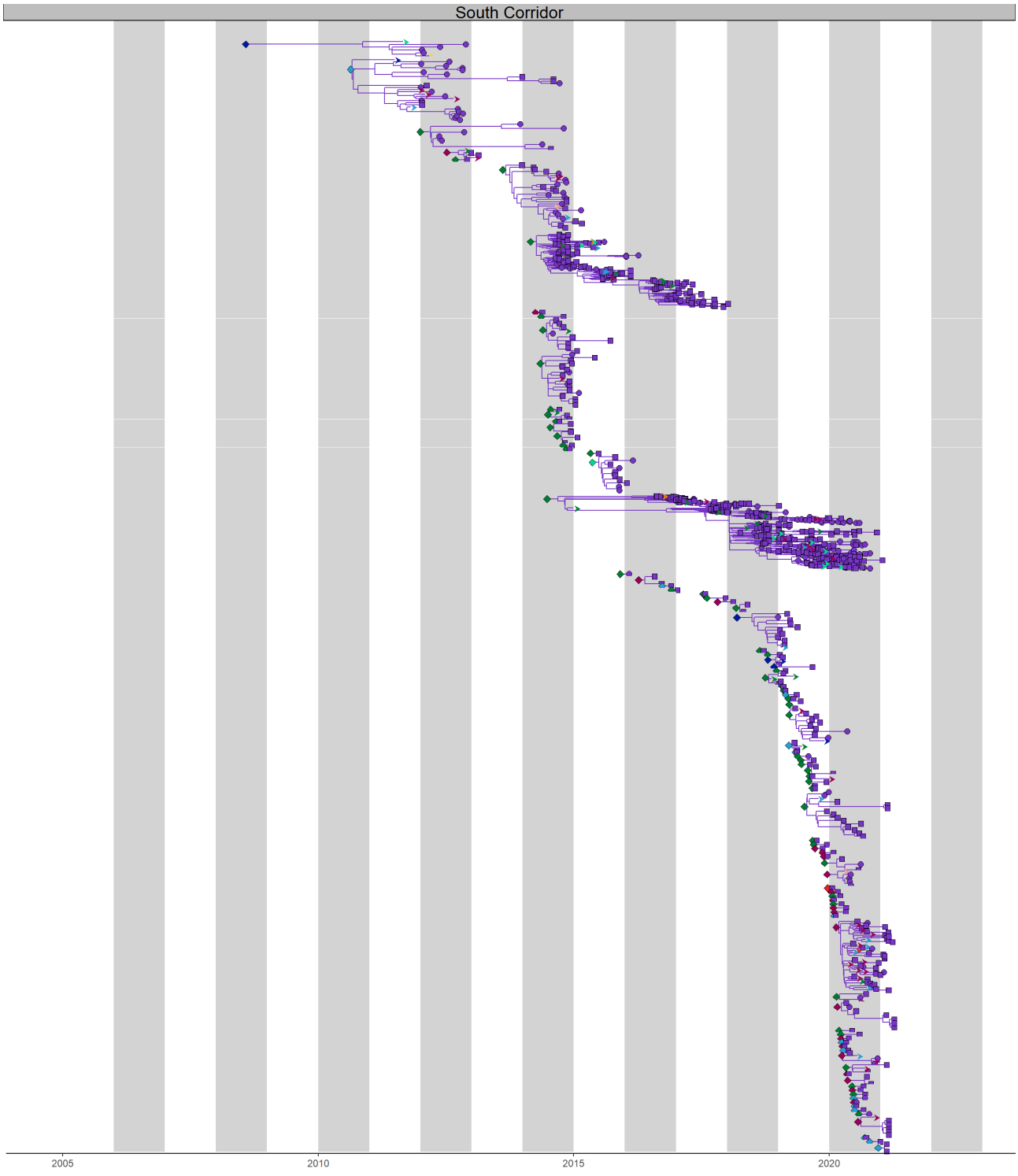
From	To	Bayes Factor Support	Posterior Probability	Median Transitions	(95% HPD)
CENTRAL-CORRIDOR	EAST-PAK	57165.88	1.00	16	(12-19)
CENTRAL-CORRIDOR	KARACHI	1579.89	0.99	4	(1-6)
CENTRAL-CORRIDOR	NORTH-CORRIDOR	57165.88	1.00	11	(8-14)
CENTRAL-CORRIDOR	WEST-AFG	23.25	0.74	1	(0-2)
CENTRE-PAK	CENTRAL-CORRIDOR	53.73	0.87	3	(0-6)
CENTRE-PAK	EAST-PAK	57165.88	1.00	24	(14-31)
CENTRE-PAK	KARACHI	57165.88	1.00	34	(23-44)
CENTRE-PAK	SOUTH-CORRIDOR	57165.88	1.00	21	(12-30)
EAST-PAK	CENTRAL-CORRIDOR	57165.88	1.00	11	(5-17)
EAST-PAK	CENTRE-PAK	57165.88	1.00	19	(13-25)
EAST-PAK	KARACHI	57165.88	1.00	15	(9-21)
EAST-PAK	NORTH-CORRIDOR	57165.88	1.00	30	(21-39)
EAST-PAK	SOUTH-CORRIDOR	57165.88	1.00	18	(12-24)
KARACHI	CENTRAL-CORRIDOR	57165.88	1.00	12	(6-17)
KARACHI	CENTRE-PAK	57165.88	1.00	106	(91-119)
KARACHI	EAST-PAK	57165.88	1.00	40	(31-48)
KARACHI	NORTH-CORRIDOR	57165.88	1.00	18	(13-26)
KARACHI	SOUTH-CORRIDOR	57165.88	1.00	53	(42-61)
KARACHI	WEST-AFG	320.30	0.97	5	(2-7)
NORTH-AFG	CENTRAL-CORRIDOR	12.88	0.61	1	(0-2)
NORTH-AFG	NORTH-CORRIDOR	18.70	0.69	1	(0-2)
NORTH-CORRIDOR	CENTRAL-CORRIDOR	57165.88	1.00	19	(14-23)
NORTH-CORRIDOR	CENTRE-PAK	283.42	0.97	4	(0-6)
NORTH-CORRIDOR	EAST-AFG	57165.88	1.00	8	(8-9)
NORTH-CORRIDOR	EAST-PAK	57165.88	1.00	64	(56-72)
NORTH-CORRIDOR	KARACHI	744.01	0.99	8	(3-13)
NORTH-CORRIDOR	NORTH-AFG	871.32	0.99	2	(2-5)
SOUTH-CORRIDOR	CENTRE-PAK	57165.88	1.00	35	(27-42)
SOUTH-CORRIDOR	EAST-AFG	12.00	0.59	1	(0-1)
SOUTH-CORRIDOR	EAST-PAK	57165.88	1.00	15	(10-19)
SOUTH-CORRIDOR	KARACHI	57165.88	1.00	20	(14-26)
SOUTH-CORRIDOR	NORTH-AFG	86.37	0.91	2	(0-3)
SOUTH-CORRIDOR	WEST-AFG	57165.88	1.00	11	(10-14)
SOUTH-CORRIDOR	WEST-PAK	774.92	0.99	2	(1-2)
WEST-AFG	SOUTH-CORRIDOR	110.09	0.93	1	(0-3)

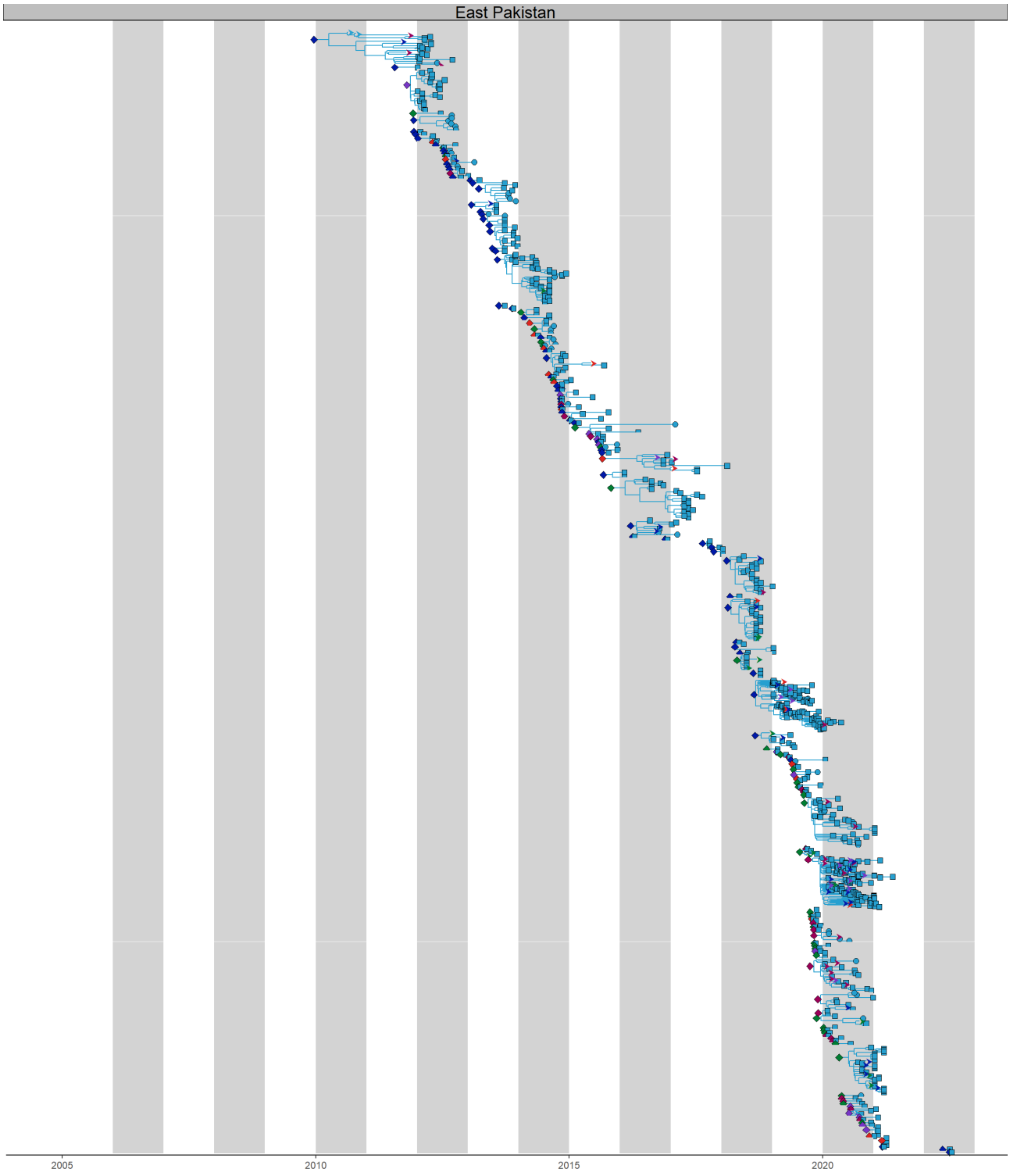
A.4 Local transmission lineages

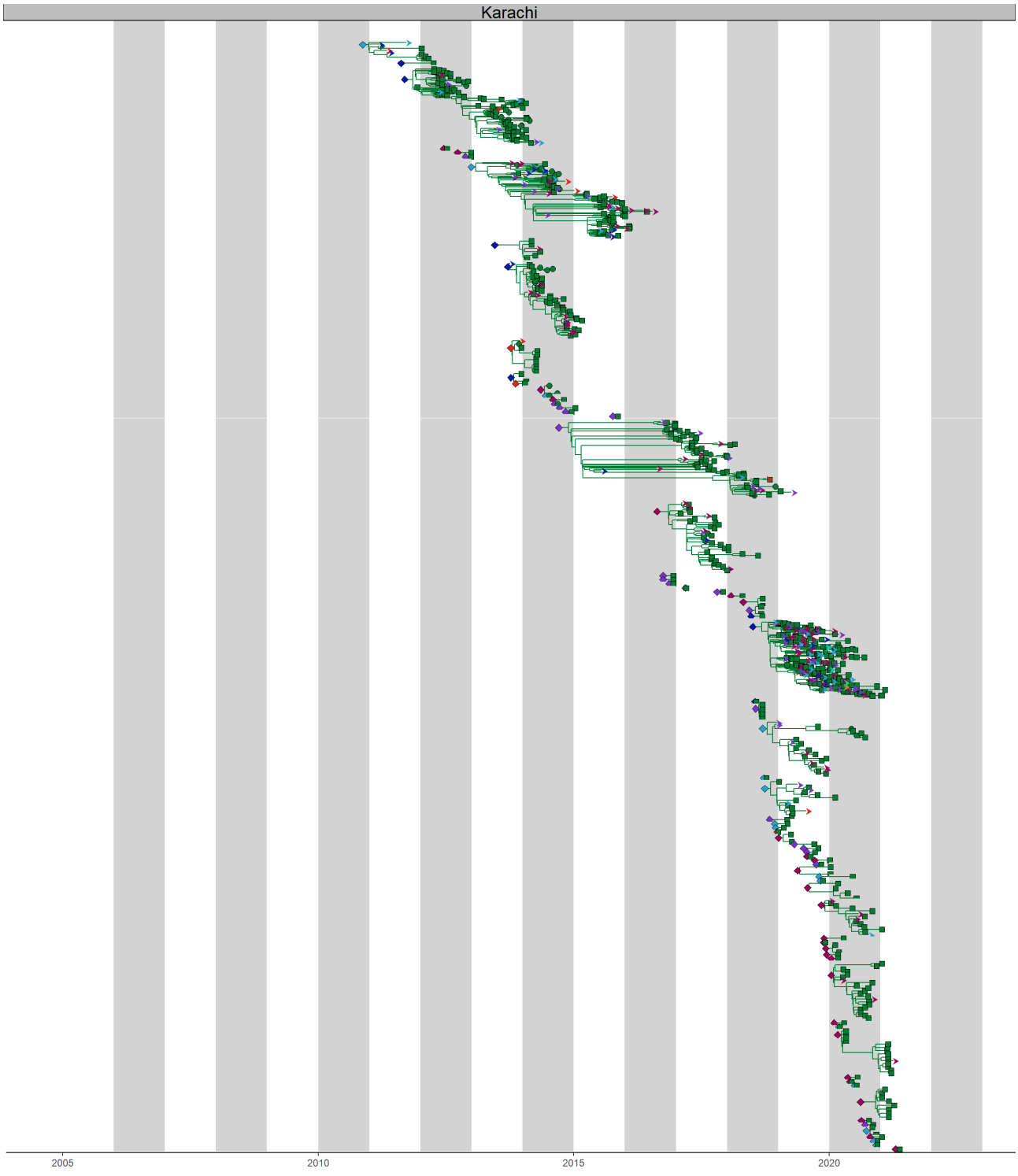
Complete local transmission lineages across the period of the data are presented for each of the 10 regions defined in the methods section (Supplementary Fig. A.4). These plots highlight the individual local transmission lineages observed in each region, including he location they were imported from and any exportations from the cluster to other regions. Here we clearly see the changes in transmission dynamics from longer term circulation to shorter lineages of fewer isolates in the majority of regions, as outlined in figure 2.6 in the main text. We also see, in the Central Corridor, continued long term persistence and large lineages of linked isolates through to 2022.

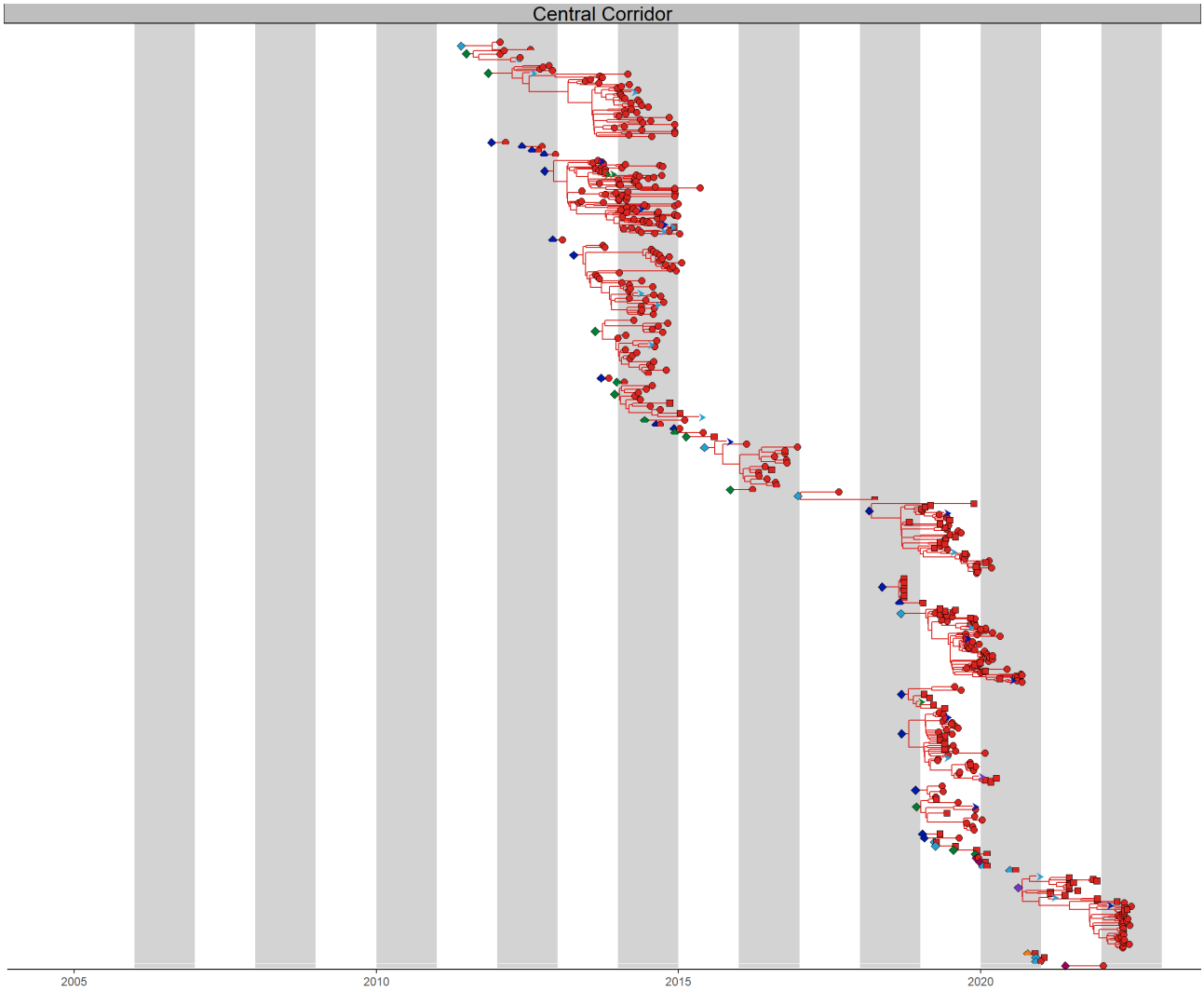
Figure A.4: Time-resolved phylogenetic tree of the full dataset split by regions. A new lineage is plotted after each inferred virus movement between regions. The tips represent sequenced viruses, with round tips representing virus isolates from AFP cases and square tips isolations from ES. The root of each lineage is annotated with a diamond coloured according to the inferred ancestral location. As the time of the virus movement from one region to the other cannot be precisely inferred, the ancestral location is plotted at the mid timepoint of the ancestral branch. Where a lineage leads to onwards transmissions in another region, this is shown with a branch leading to an arrow coloured by the region exported to.

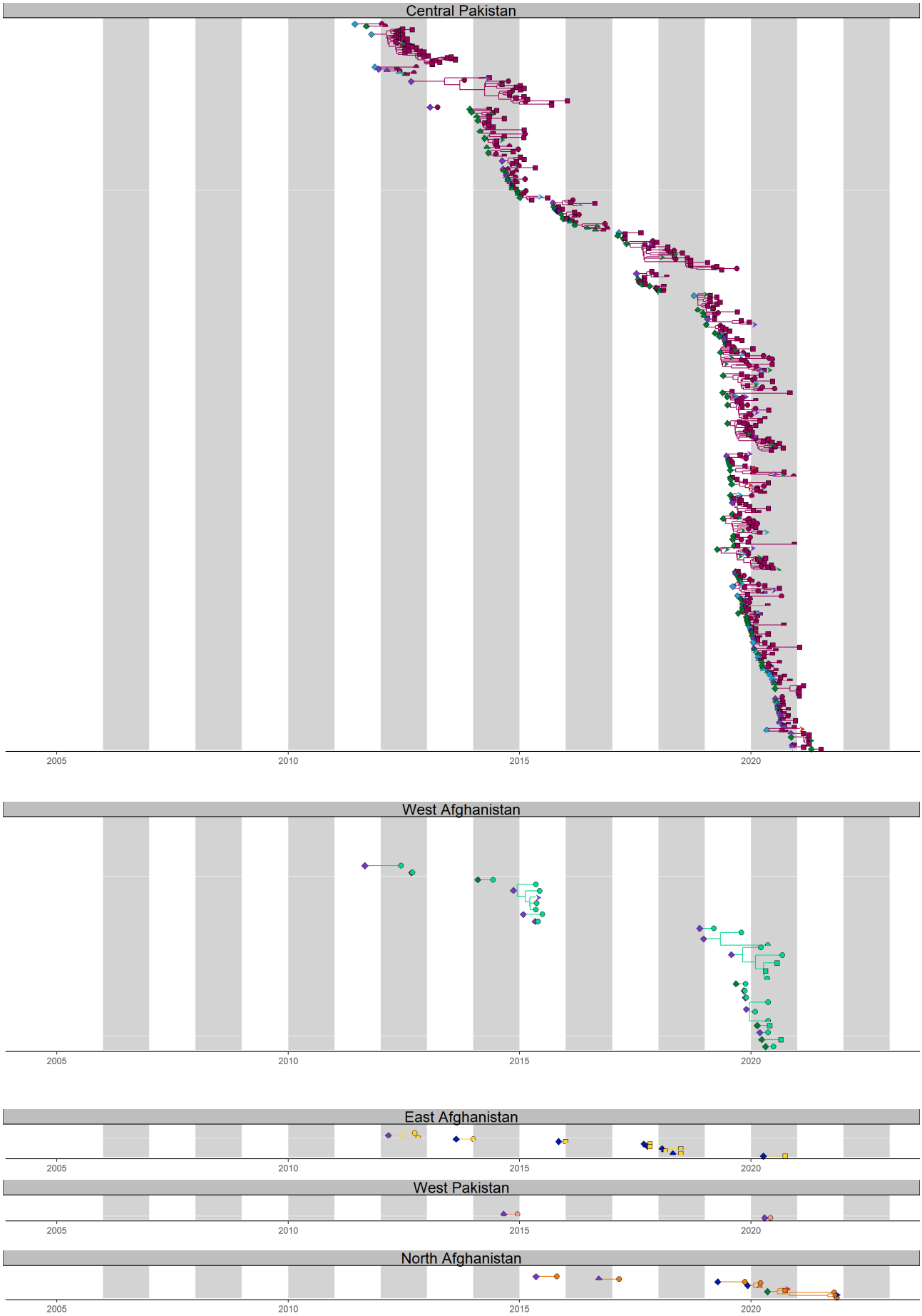



South Corridor
2005
2010
2015
2020


East Pakistan
2005
2010
2015
2020


Karachi
2005
2010
2015
2020


Central Corridor
2005
2010
2015
2020


Central Pakistan
2005
2010
2015
2020
West Afghanistan
2005
2010
2015
2020
East Afghanistan
2005
2010
2015
2020
West Pakistan
2005
2010
2015
2020
North Afghanistan
2005
2010
2015
2020

A.5 Classifying local transmission lineages

To produce a summary figure of local transmission lineages in the main text they were classified into three groups. These are determined by length of circulation, onwards transmission to other regions and number of sequenced tips as outlined in chapter 2 An example set of classifications for local transmission lineages as presented in supplementary figure A.4 is presented in supplementary figure A.5.

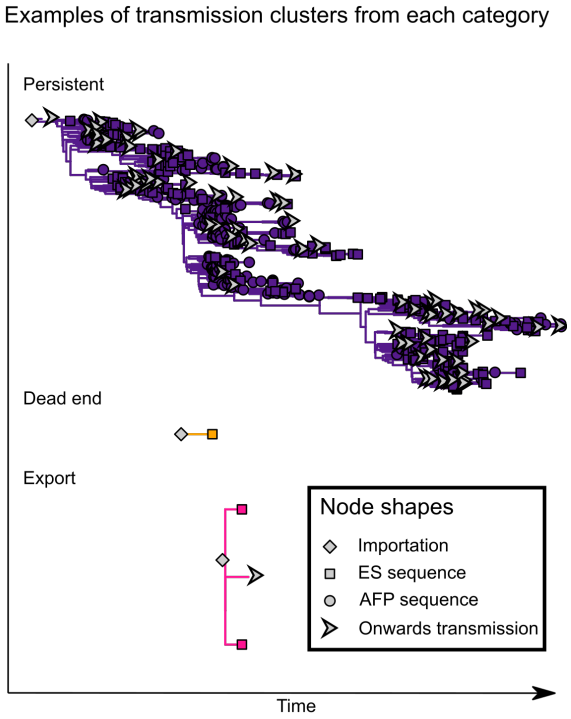


Figure A.5: Examples of local transmission clusters which would be categorised into each of the classifications based on length of persistence in time, number of AFP tips and onwards transmission to other regions.

A.6 Investigation of AFP sequences alone

A.6.1 Sequencing intensity

As highlighted in the full analysis with the epidemic curve presented under the phylogeny in figure 2.4, the epidemiology of poliovirus has changed over the course of data collection repre-

sented in this analysis with high and low transmission periods. AFP case data are extracted from the Polio Information System (POLIS), a global database of polio data maintained by the World Health Organization [90]. Preliminary analysis of cases and sequences over time showed that the sequence dataset is representative of reported AFP cases. 53 cases were dropped from the comparison as incomplete metadata prevented a region from being assigned. If we assume that AFP case reporting is approximately complete in all regions due to the severity of the outcome and that AFP occurs as a symptom of polio at the same rate regardless of region, this sample should represent sampling proportional to the number of infected cases, as required for DTA analysis.

A.6.2 Location estimates

A phylogeny with inferred ancestral locations estimated using sequences from AFP cases alone is shown in Figure A.6. Without the large number of ES sequences the patterns reconstructed differ, but overall we see similar trends. The first of these is the three southern regions of Karachi, Centre Pakistan and the Southern Corridor supporting cyclical chains of transmission. Without ES, the role of Eastern Pakistan in sustaining transmission is diminished as very few paralysis cases are reported in this region. Looking at AFP data alone, this region acts as a sink of virus transmission, primarily from the Central Pakistan region. This is supported by the short transmission chains with few isolates seen in the analysis of the full dataset.

The AFP data alone supports the observation of the replacement of distinct lineages in the north and south with a single lineage rooted in the Northern Corridor. This lineage was exported to the South during the expansion of poliovirus transmission during the 2019 outbreak. All recent detections of AFP are linked to this cluster of transmission. Without ES data we miss recent spill over events from the current reservoir population in the Central Corridor which have resulted in ES detections but not paralysis cases in surrounding regions. This again highlights the importance of ES as an early detection system, particularly in regions with close geographic or cultural ties to regions with higher levels of virus circulation.

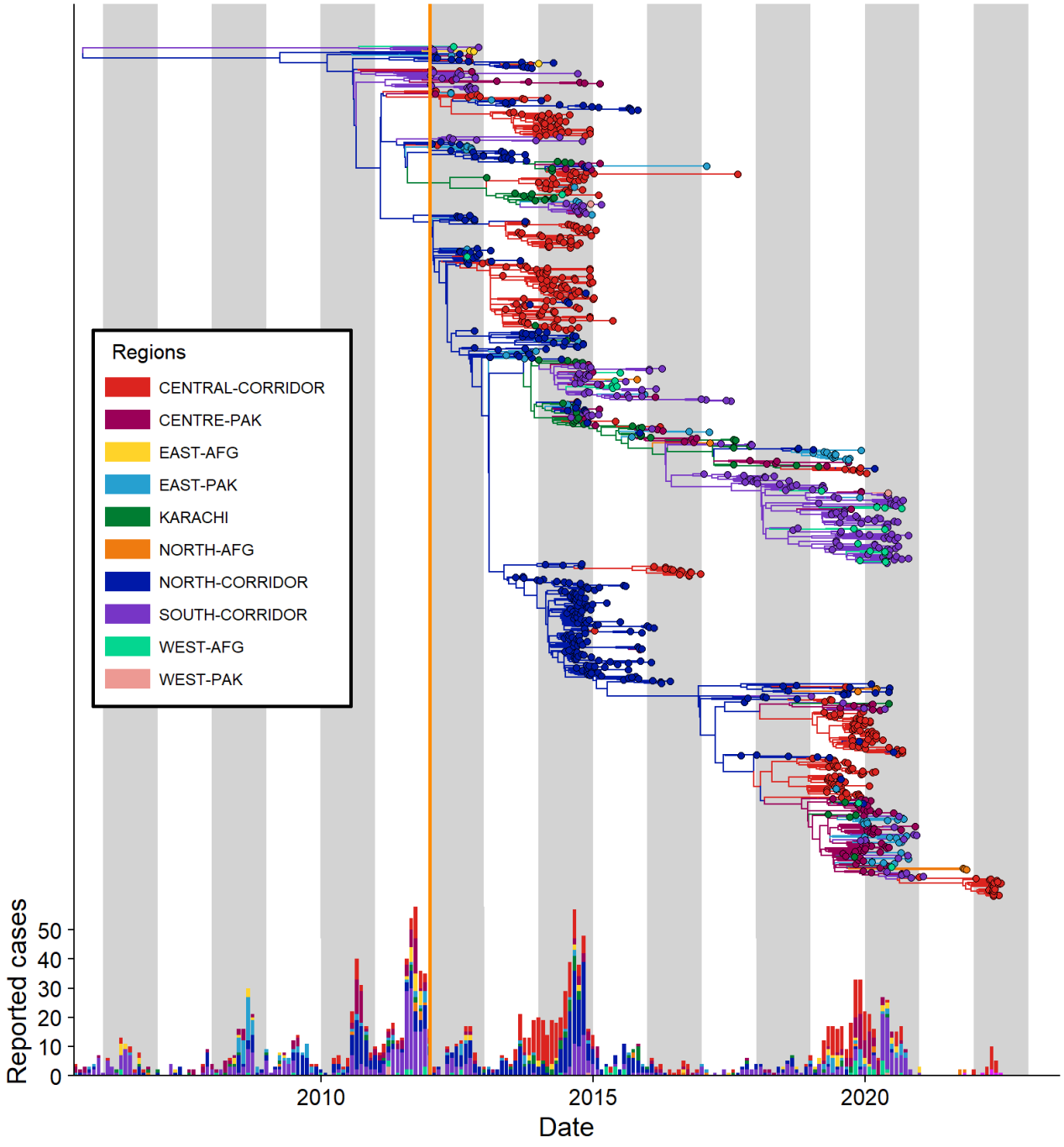


Figure A.6: Time-resolved phylogenetic tree of sequences from AFP cases for Pakistan and Afghanistan, coloured according to the regions in the map in figure 2. Internal node locations are inferred using Bayesian Markov jump discrete trait analysis and time in years is given on the x-axis. The bar plot below the tree shows monthly counts of reported polio AFP cases, coloured by the collection location and the orange vertical line the date of the first sampled sequence included in the genetic dataset.

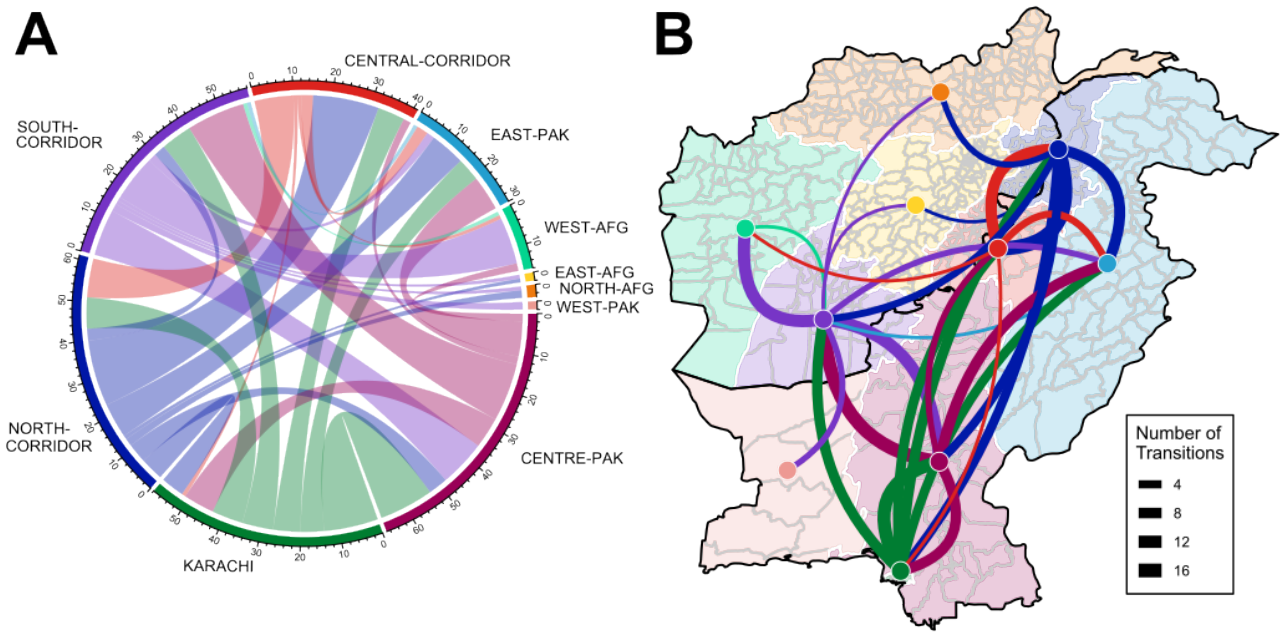


Figure A.7: **A)** Circus plot of supported migration events between the regions considered in the analysis. The movements are coloured according to the starting location and the width represents the number of transitions inferred, with counts provided around the circumference of the plot. **B)** The same transitions visualised on a map of the discrete regions.

A.6.3 Movement estimates

Using AFP data alone we estimate numbers of movements between locations with the same method outlined for the complete dataset. Here we again present movements with Bayes factor support above 10. We see similar patterns to those seen with the complete dataset, with Karachi contributing the largest number of exportations to other regions, and a clear cycle of transmission in the South, from Karachi to Centre Pakistan and the South Corridor and back in the reverse order (supplementary Fig. A.7). The Central Corridor, which has supported most transmission since the 2019 outbreak (figure 2.6), is shown by the AFP data alone to have imported virus primarily from the North Corridor over the time course of the included data but the recently circulating cluster is linked back to transmission in the southern regions. The reduced dominance of Karachi on supplementary figure A.7, when compared with the complete dataset presented in figure 2.5, again highlights the effect of high levels of environmental sampling in Karachi on inferences made from polio surveillance data.

Appendix B

Supplementary information for chapter 3

B.1 Number of phylogenies retained

For several simulated phylogenies, standard errors or maximum likelihood estimates could not be estimated due to the non-convex (ridged) nature of the likelihood surface. In these cases, the estimated relative migration rates from these phylogenies were dropped from model comparison. Figure B.1 shows the number of phylogenies retained from the initial 100 simulations under the "low sampling rate" and "transmission difference" scenarios when calculating the results. Where a value could be estimated with only one of the methods, the phylogeny was dropped from calculations for both methods. Figure B.1 presents the number retained where two-location, 100 tip phylogenies were simulated, under both equal transmission rates (A) and unequal transmission rates (B).

We see a clear divide on figure B.1 between the upper left and bottom right corners of the plot. Under the unequal transmission scenario represented in this plot where transmission is faster in population 2 we see a greater number of phylogenies dropped where sampling in 1 is high and the rate of leaving 1 is low (top left of the plot). In these phylogenies population 2 is over-represented compared to the same sampling scenario without a transmission rate difference. The likelihood surfaces for the relative and base migration rates on these dropped phylogenies have a wide, flat maximum likelihood region, making 95% confidence interval estimation not possible with the hessian method used here. Phylogenies retained under the medium (Fig. B.2A) and high (Fig. B.2B) sampling rate scenarios are shown in figure B.2.

Figure B.1: Number of simulations retained after removing those where a maximum likelihood value or standard error for the relative migration rate could not be estimated with both the corrected and uncorrected method. Panel A shows the numbers retained for the low transmission rate scenario and panel B the transmission difference scenario.

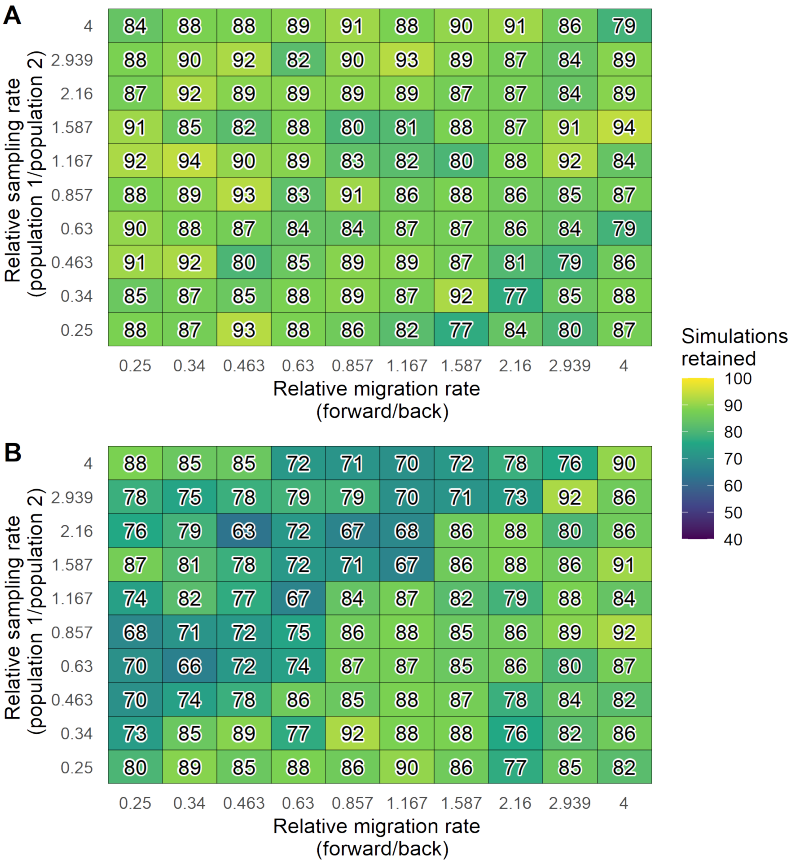
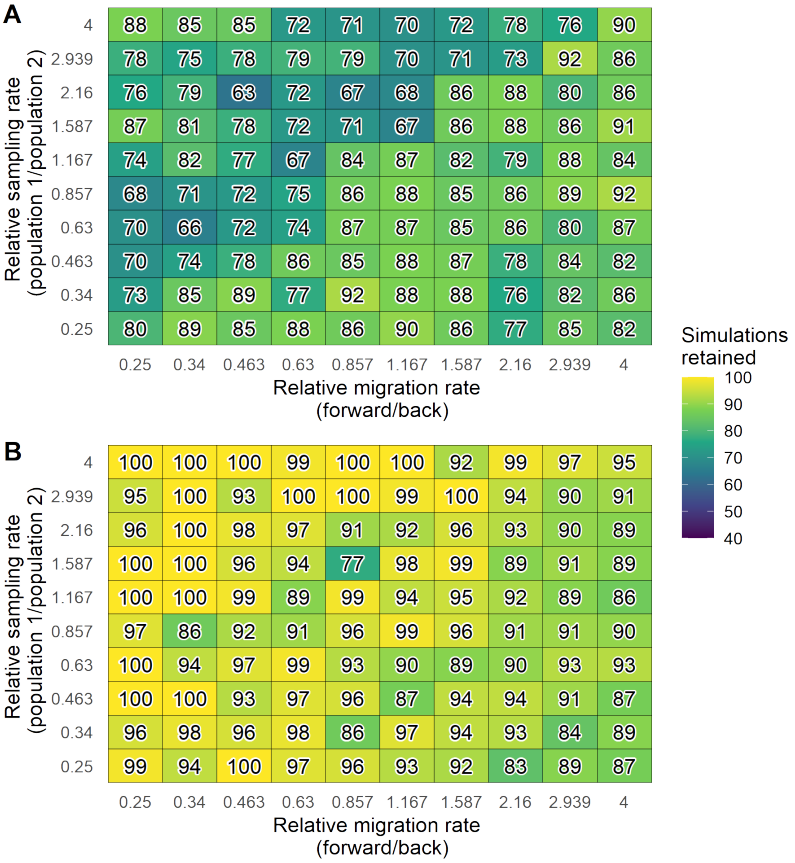


Figure B.2: Number of simulations retained after removing those where a maximum likelihood value or standard error for the relative migration rate could not be estimated with both the corrected and uncorrected method. Panel A shows the numbers retained for the medium rate scenario and panel B the high transmission rate scenario.



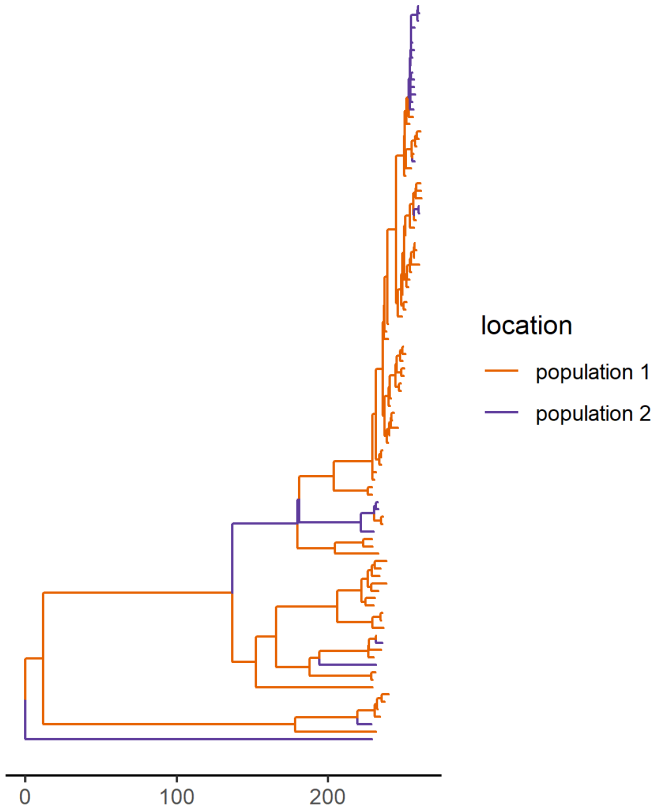


Figure B.3: Illustrative example of a phylogeny sampled under the high sampling rate scenario. Here we see the tips are sampled close in time and the root is more recent than in the low sampling rate example shown in figure 3.1.

B.2 High sampling rate phylogeny

Under the high sampling rate scenario presented in the main text, the corrected and uncorrected methods performed equivalently, with the correction no longer providing improved results (Fig. 3.8) An example phylogeny under the high sampling rate scenario is shown in supplementary figure B.3.

Appendix C

Supplementary information for chapter 5

Initial analysis was carried out with fewer beta break points to assess the viability of the method. Breaks were set such that β_1 represented transmission from the estimated introduction date (30 days before the tenth reported case) to the introduction of the first restrictions [200]. β_2 covered the time from the introduction of the first restrictions to the first long term curfew in each city. β_3 represents the curfew period, ending where all curfew restrictions were lifted nationally and β_4 represents the time period after the lifting of restrictions to the end of the prediction period at the start of September.

Restrictions implemented vary widely during each of the time periods due to the introduction and removal of certain restrictions at different times by location and there is potential for lag between the introduction of restrictions and the effect of these restrictions leading to meaningful changesin transmission rates. Due to this the main analysis incorporates arbitrary break points in β representing calendar months rather than attempting to link break points and NPI events. The results of this initial analysis are presented here as findings differ slightly from the final analysis and estimates of R0 from these runs were used to set starting values in the final runs.

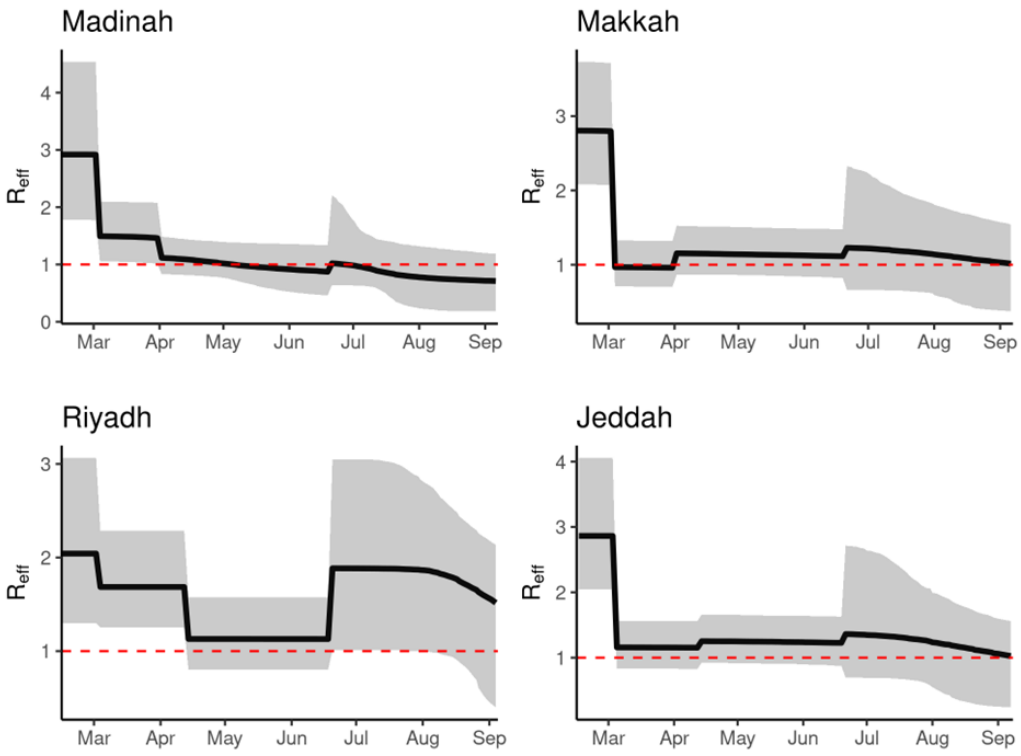


Figure C.1: Effective reproduction number estimates (black line) and 95% credible interval (grey shaded area) over time from the PhyDyn model for 4 cities of interest. The red horizontal line represents an Reff of 1, the number required for virus replacement. These estimates have three step changes set to align with changes in NPIs

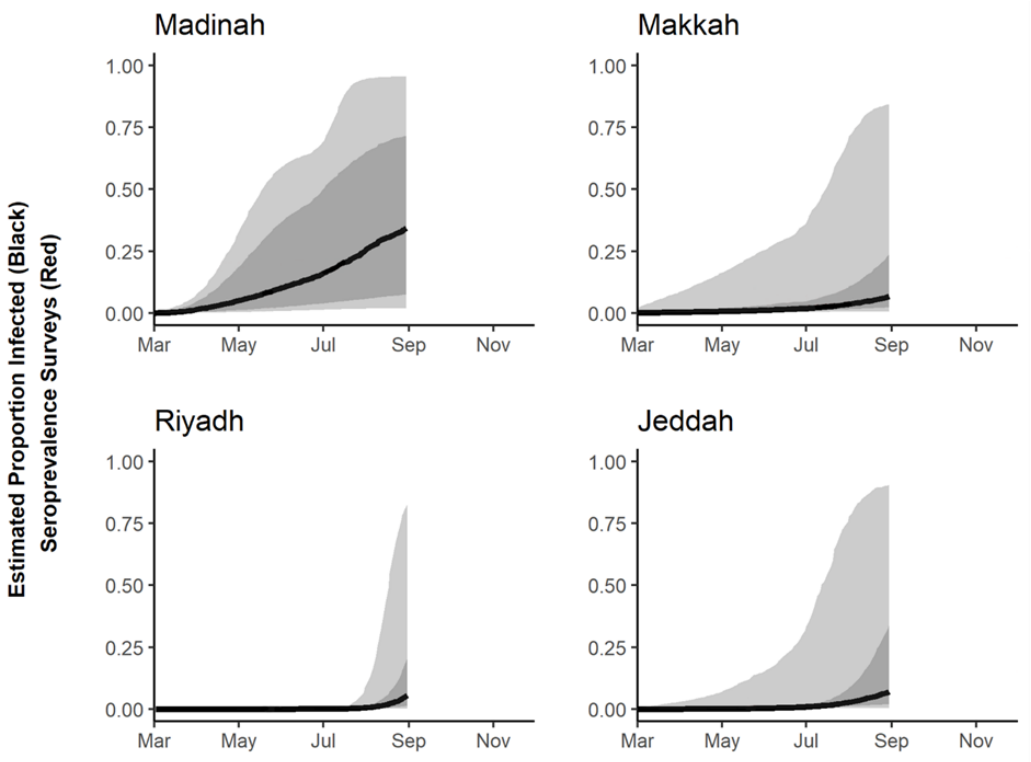


Figure C.2: Estimates of the cumulative proportion of the population infected (black line), 95% credible interval (light grey shaded area) and interquartile range (dark grey shaded area) in the model over time for each city.

SUPPLEMENTAL TABLE

Data Availability

GISAID Identifier: EPI_SET_230619tc
doi: 10.55876/gis8.230619tc

All genome sequences and associated metadata in this dataset are published in GISAID's EpiCoV database. To view the contributors of each individual sequence with details such as accession number, Virus name, Collection date, Originating Lab and Submitting Lab and the list of Authors, visit 10.55876/gis8.230619tc

Data Snapshot

- EPI_SET_230619tc is composed of 96,699 individual genome sequences.
- The collection dates range from 2019-12-24 to 2020-09-14;
- Data were collected in 117 countries and territories;
- All sequences in this dataset are compared relative to hCoV-19/Wuhan/WIV04/2019 (WIV04), the official reference sequence employed by GISAID (EPI_ISL_402124). Learn more at https://gisaid.org/WIV04.