DISS. ETH No. 28604

PATHOGEN PHYLOGENIES FOR PUBLIC HEALTH

SARAH ANN NADEAU

2022

DISS. ETH NO. 28604

PATHOGEN PHYLOGENIES FOR PUBLIC HEALTH

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

SARAH ANN NADEAU
M.Sc. Cornell University, New York, USA

born on 13 April 1994
citizen of the United States of America

accepted on the recommendation of

Prof. Dr. Tanja Stadler, examiner
Prof. Dr. Richard A. Neher, co-examiner
Prof. Dr. Jacques Fellay, co-examiner

2022

ABSTRACT

An organism's genome sequence is a rich source of information on its current characteristics and its evolutionary history. In this thesis, I refine and apply methods to extract information from pathogen genome sequences via phylogenetic reconstructions. I extend existing phylogeny-based models to new applications in genome-wide association studies (GWAS) and genomic epidemiology. First, I show that correlations in an infectious disease trait due to shared pathogen ancestry can reduce GWAS power. I extend a statistical model of evolution to estimate and correct for these correlations. Second, I apply a phylodynamic model to estimate the origin and early transmission patterns of the SARS-CoV-2 virus during the first European outbreaks of COVID-19. Third, I describe a data infrastructure we built to generate SARS-CoV-2 genome sequences from cases in Switzerland. Finally, I develop a phylogenetic and phylodynamic framework to perform a large-scale analysis on these data. In particular, I evaluate the effect of several major public health measures in Switzerland in 2020 on SARS-CoV-2 introduction and transmission dynamics. All together, this thesis aims to enhance our understanding of infectious diseases and how to combat them.

ZUSAMMENFASSUNG

Die Genomsequenz eines Organismus enthält vielfältige Informationen über seine Charakteristiken und Evolutionsgeschichte. In dieser Dissertation entwickle und verwende ich Methoden, die Informationen aus den Genomsequenzen von Pathogenen mithilfe phylogenetischer Bäume extrahieren. Ich erweitere bestehende phylogeniebasierte Modelle für neue Applikationen in genomweite Assoziationsstudien (GWAS) und in der genetischen Epidemiologie. Zuerst zeige ich, dass die aus der gemeinsamen Abstammung von Pathogenen resultierenden Korrelationen im Merkmal einer ansteckenden Krankheit die Power von GWAS reduzieren können. Ich erweitere ein statistisches Evolutionsmodell, um diese Korrelationen zu schätzen und dafür zu korrigieren. Zweitens benutze ich ein phylodynamisches Modell, um den Ursprung und die frühe Transmissionsgeschichte von SARS-CoV-2 Viren während den ersten europäischen Outbreaks von COVID-19 zu schätzen. Drittens beschreibe ich eine Dateninfrastruktur, die wir implementiert haben, um SARS-CoV-2 Genomsequenzen in der Schweiz zu generieren. Anschliessend entwickle ich ein phylogenetisches und phylodynamisches Framework, um umfangreiche Analysen dieser Daten durchführen zu können. Insbesondere evaluiere ich die Effekte von einigen wichtigen öffentlichen Gesundheitsmassnahmen in der Schweiz in 2020 auf die Dynamiken hinsichtlich der Einschleppungen und Übertragungen von SARS-CoV-2. Insgesamt hat diese Dissertation als Ziel, unser Verständnis von übertragbaren Krankheiten und ihrer Bekämpfung zu verbessern.

ACKNOWLEDGEMENTS

Thank you to the many individuals who contributed in countless ways to this thesis and the happiness of its author. The cEvo group past and present taught me how to ask good questions and offered critiques in the nicest possible way. Venelin Mitov, Chaoran Chen, and Tim Vaughan, in particular, were instrumental to the development of Chapters 2, 4 and 5 - I hope some of their brainpower rubbed off on me in the process! Thank you to Chaoran Chen and Adrian Lison, who vastly improved the German translation of the abstract. While supervising this thesis, Tanja Stadler provided an example of mentorship and leadership that I will take with me far beyond my PhD. Then, I am thankful for Gabriel, who signed on for this adventure and many more! Thank you also to my siblings, who make sure I never go through anything alone. Finally, I am grateful to my parents for showing me what the truly important things are in life.

CONTENTS

1

INTRODUCTION

One of the major promises of large-scale genome sequencing was to uncover the genetic basis of disease. Sequence-based insights were envisioned to yield improvements in diagnostics, therapeutics, and public health, ushering in a new era of precision genomic medicine. In this thesis, I extend and apply methods aiming to deliver on this promise for infectious diseases. In particular, I focus on methods that utilize pathogen phylogenetic reconstructions, which represent the evolutionary relationships between infecting strains. Applications include genome-wide association studies for understanding the genetic basis of infectious disease traits and phylogenetic and phylodynamic methods for quantifying infectious disease transmission dynamics. The following sections are devoted to introducing these concepts in greater detail.

1.1 GENOME SEQUENCING AND GENOMIC EPIDEMIOLOGY

The first human genome sequence was released in 2003. At the time, generating this ordered readout of the ≈3 billion DNA bases comprising the human genome cost an estimated 500 million to 1 billion U.S. dollars (Wetterstrand, 2019). Since then, the cost per megabase for DNA sequencing has dropped over 300,000 fold (Wetterstrand, 2019), thanks in large part to methodological advances in "next-generation" sequencing (EMBL-EBI, 2022). As of June 2022, over 1.7 billion whole-genome sequences are publicly available in the GenBank database (NCBI, 2022). Thus, the generation of large amounts of genome sequence data is no longer necessarily a primary limitation for research.

So, how has this abundance of genome sequence data advanced public health? This thesis falls into the broad field of "genomic" or "molecular" epidemiology, which is concerned with extracting epidemiological insights from genome sequence data - in other words, pushing the envelope on that question. The next sections outline two key methodological approaches in the field of genomic epidemiology with particular application to infectious disease. These methods form the basis for the following chapters.

1.2 GENOME-WIDE ASSOCIATION STUDIES

One of the first applications of large-scale genome sequencing was for genome-wide association studies (GWAS) (Uffelmann *et al.*, 2021). GWAS test for statistical associations between genetic variants and a phenotypic trait across a large cohort of genotypically and phenotypically diverse individuals. Depending on whether the phenotype is binary, like diseased versus healthy status, or continuous, like growth rate or viral load, a logistic or linear regression model is typically used to test the strength of association. The predictor variables in these models are typically the presence/absence or copy number of a genetic variant. GWAS are "genome-wide" in that they test thousands or millions of different genetic variants from an organism for association with the same trait, screening for those with the strongest associations. Since many variants are considered independently, correction for multiple testing is important to control for false discoveries. Even so, associated variants may not be truly causal due to linkage disequilibrium. In other words, causal and non-causal variants may be genomically linked and thus commonly inherited together. Therefore, functional annotation and/or validation in model organisms is necessary to determine which associated variant(s) are truly causal for a given phenotype (Albert and Kruglyak, 2015; Cano-Gamez and Trynka, 2020).

A flurry of GWAS were conducted in the late-2000s (Topol *et al.*, 2007), primarily focusing on non-infectious human disease. These early studies quickly revealed that susceptibility to most common human diseases like asthma and obesity is determined by a multitude of genetic variants, often of small individual effect (Vercelli, 2008; Bogardus, 2009). Due to the complex genetic underpinnings of these diseases, translation of GWAS results to clinical practice is challenging (Barton, 2007). However, a few breakthrough successes have been realized. For example, genetic risk prediction is now common for some types of cancer (Kuchenbaecker *et al.*, 2017). Attempts to utilize GWAS-identified genes and pathways as drug targets are ongoing, e.g. Li *et al.* (2022); Shu *et al.* (2018); Okada *et al.* (2014).

GWAS have also increasingly been applied to human genomes and infectious disease traits (Mozzi *et al.*, 2018). So far, these GWAS established that human genetic variation influences susceptibility to a wide variety of infectious diseases, including HIV-1, hepititis B and C, and malaria, among others (Chapman and Hill, 2012). However, such direct applications of GWAS to infectious disease traits ignore variation in pathogen genomes, which can also strongly influence infectious disease traits (Mitov and Stadler,

2018). In fact, GWAS using microbial genomes instead of human genomes, or "mGWAS", have identified pathogen genetic variants that influence a variety of infectious disease traits (San *et al.*, 2020). Several methods aim to incorporate information from both host and pathogen genomes in GWAS. For instance, genome-to-genome studies test for associations between host and pathogen genetic variants (Bartha *et al.*, 2013). Then, Wang *et al.* (2018) introduced a linear mixed-model approach to test for interaction and marginal effects of variants from both genomes on a trait. Methods like this to account for pathogen genetic effects in GWAS using host genomes are an area of ongoing development (Kwok *et al.*, 2021).

1.3 PHYLOGENETIC RECONSTRUCTIONS

A second, broad application of genome sequencing is for phylogenetic reconstruction. In this case, similarities and differences between related organisms are used to reconstruct their evolutionary relationships. This information is represented as a phylogenetic tree, where related individuals cluster closer together than more distantly related individuals and ancestral branches represent shared evolutionary history.

Phylogenetic reconstructions have been used for over 140 years, with initial applications focusing on systematics (Felsenstein, 2004). The first phylogenies were constructed on the basis of morphological characteristics. With the advent of molecular sequencing - first protein, then DNA sequencing - in the 1960s and 1970s, a rich new data source was available for reconstructing phylogenies (Barton, 2007). Concurrent methods developments introduced algorithmic methods for reconstructing phylogenies, including parsimony, distance matrix, and maximum likelihood methods (Felsenstein, 2004). These methods all make different assumptions to algorithmically generate a phylogenetic reconstruction of the evolutionary relationships between organisms. As opposed to GWAS, where linkage between sites makes it difficult to tease apart causal and non-causal genetic variants, these phylogenetic reconstruction methods assume complete linkage between sites. In other words, even sites far apart on the genome should be representative of the same evolutionary history. When this assumption is satisfied, phylogenetic reconstructions are useful summaries of the evolutionary relationships between organisms.

In addition, phylogenetic reconstructions are a key data structure for generating insights on population biology and evolution. For instance, a common question in evolutionary biology is whether certain traits are

correlated, for instance brain and body size, and if so, what selective forces constrain their evolution in this way. Basic statistical tests to assess such associations assume independence of samples. However, many traits are related in related species due to shared evolutionary history. Phylogenetic comparative methods apply evolutionary models to trait values at the tips of phylogenies, accounting for shared evolutionary history and enabling rigorous statistical hypothesis testing (Harvey *et al.*, 1998).

Another application of phylogenies is the study of their branch length distributions and topologies to generate insights on population dynamics, known as phylodynamics (Grenfell *et al.*, 2004). A dated phylogeny has branch lengths in time units. Short branch lengths then correspond to short observed speciation intervals and vice versa. For infectious diseases, "speciation" is transmission from one infected host to another, after which daughter lineages evolve independently in the two hosts. Stochastic models can be used to relate these branching times to underlying population dynamics. Two commonly used classes of models are derived from the coalescent process (Wakeley, 2009) and the birth-death process (Stadler, 2010). These processes provide mathematical frameworks for relating variation in branch lengths to population dynamics - population size, in the case of the coalescent, or birth and death rates, in the case of the birth-death model. The implementation of coalescent and birth-death models in Bayesian inference frameworks enables joint inference of a phylogeny and the population dynamics that generated it, representing another advance in methods for phylogenetic reconstruction (Suchard *et al.*, 2018; Bouckaert *et al.*, 2019). Development of these models is ongoing, extending their applicability to populations with more complex dynamics, e.g. Kühnert *et al.* (2016), and pathogens with more complex evolutionary processes, e.g. Müller *et al.* (2020).

1.4 PRACTICAL APPLICATIONS OF PATHOGEN PHYLOGENIES

In practice, phylogenetic reconstruction may be the focus of an analysis, or, increasingly, treated as a means to another end. For instance, statistics calculated from a phylogeny or parameters estimated while integrating over a multitude of plausible phylogenies may take center stage, while the actual phylogenetic reconstruction fades into the background as an intermediate data representation or even a nuisance parameter. Here, I will highlight prior applications of pathogen phylogenies in infectious disease research that set the stage for the advances presented in this thesis.

In GWAS, phylogenies are primarily used in microbial GWAS as a convenient representation for the uniquely strong population structure in clonally reproducing microorganisms. If unaccounted for, systematic differences in pathogen populations might yield spurious genetic correlations with a trait. Phylogenetic reconstructions can be used to adjust for these differences. For instance, phylogenies have been used to generate principle components or a kinship matrix describing microbial population structure. These can be included as covariates in linear association models (Naret *et al.*, 2018) or as a random effect in linear mixed models of association (Lees *et al.*, 2018) to control for population structure. These methods have been used to better understand infectious disease pathology. For example, Lees *et al.* (2019) identified two genes in Streptococcus pneumoniae associated with invasive propensity.

In public health, phylogenies are more often at the forefront of practical analyses. Outbreak phylogenies help establish whether cases are epidemiologically linked and where the outbreak may have originated (Armstrong *et al.*, 2019). Recent examples include identifying a household product as the source of a geographically wide-spread melioidosis outbreak in the U.S. (Gee *et al.*, 2022) and understanding the likely zoonotic origins of the new human pathogen SARS-CoV-2 (Wu *et al.*, 2020).

Phylodynamic methods aim to generate even more information than this from pathogen genome sequence data. As previously discussed, these methods jointly infer the phylogeny alongside population dynamics of interest, like the size of the infected population or transmission rates. Depending on the application, the phylogeny itself may be more or less interesting. For instance, Stadler *et al.* (2013) estimed the time-varying reproductive number of HIV-1 in the U.K. and hepititis C virus in Egypt using a birth-death phylodynamic framework and Dudas *et al.* (2018) used a structured coalescent framework to show that MERS-CoV in humans is driven by seasonal zoonotic spillover from camels with limited human-to-human transmission. More recently, phylodynamic methods have also been extensively applied to study the transmission dynamics of SARS-CoV-2 (Attwood *et al.*, 2022).

1.5 PUSHING THE ENVELOPE

In this thesis, I employ a variety of phylogeny-based methods in genomic epidemiology, extending and applying them to new infectious disease contexts. These methods include phylogenetic comparative methods, GWAS, the generation and linking of genome sequence data with metadata, and

phylogenetic and phylodynamic inference. Applications include HIV-1 in Switzerland and SARS-CoV-2 in Europe and in Switzerland. As a unifying theme, this thesis is primarily concerned with the generation of practical insights from pathogen genome sequence data.

This remainder of this thesis is organized as follows: Chapter 2 introduces a new approach to estimate and correct for pathogen effects prior to infectious disease GWAS using host genomes. Chapter 3 tests a specific hypothesis about the routes of introduction of SARS-CoV-2 into Europe at the onset of the COVID-19 pandemic. Chapter 4 outlines our efforts to generate SARS-CoV-2 genome sequences from the Swiss epidemic and link these sequences to relevant metadata. Chapter 5 probes how useful these genome sequence data are for evaluating public health measures in the first year of the Swiss COVID-19 epidemic. Finally, I summarize the advances presented in this thesis and highlight where I believe the biggest opporunities lie going forward for generating translational impact from pathogen genome sequence data.

BIBLIOGRAPHY

Albert, F. W. and Kruglyak, L. 2015. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4): 197–212.

Armstrong, G. L., MacCannell, D. R., Taylor, J., Carleton, H. A., Neuhaus, E. B., Bradbury, R. S., Posey, J. E., and Gwinn, M. 2019. Pathogen genomics in public health. *The New England Journal of Medicine*, 381(26): 2569–2580.

Attwood, S. W., Hill, S. C., Aanensen, D. M., Connor, T. R., and Pybus, O. G. 2022. Phylogenetic and phylodynamic approaches to understanding and combating the early sars-cov-2 pandemic. *Nature Reviews Genetics*, pages 1–16.

Bartha, I., Carlson, J. M., Brumme, C. J., McLaren, P. J., Brumme, Z. L., John, M., Haas, D. W., Martinez-Picado, J., Dalmau, J., López-Galíndez, C., *et al.* 2013. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *eLife*, 2: e01123.

Barton, N. H. 2007. *Evolution*. CSHL Press.

Bogardus, C. 2009. Missing heritability and GWAS utility. *Obesity*, 17(2): 209–210.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., *et al.* 2019. BEAST 2.5: An advanced software platform for bayesian evolutionary analysis. *PLOS Computational Biology*, 15(4): e1006650.

Cano-Gamez, E. and Trynka, G. 2020. From gwas to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, 11.

Chapman, S. J. and Hill, A. V. 2012. Human genetic susceptibility to infectious disease. *Nature Reviews Genetics*, 13: 175–188.

Dudas, G., Carvalho, L. M., Rambaut, A., and Bedford, T. 2018. Mers-cov spillover at the camel-human interface. *eLife*, 7.

EMBL-EBI 2022. What is next generation dna sequencing? | functional genomics ii. https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/. Accessed: 2022-7-5.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates Incorporated.

Gee, J. E., Bower, W. A., Kunkel, A., Petras, J., Gettings, J., Bye, M., Firestone, M., Elrod, M. G., Liu, L., Blaney, D. D., *et al.* 2022. Multistate outbreak of melioidosis associated with imported aromatherapy spray. *The New England Journal of Medicine*, 386(9): 861–868.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A., and Holmes, E. C. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303: 327–332.

Harvey, P. H., Harvey, P. P., Reader in Biology Department of Zoology Paul H Harvey, Pagel, M. D., and Pagel, M. D. 1998. *The Comparative Method in Evolutionary Biology*. Oxford University Press on Demand.

Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K.-A., Mooij, T. M., Roos-Blom, M.-J., Jervis, S., van Leeuwen, F. E., Milne, R. L., Andrieu, N., *et al.* 2017. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA*, 317(23): 2402–2416.

Kühnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. 2016. Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Molecular Biology and Evolution*, 33(8): 2102–2116.

Kwok, A. J., Mentzer, A., and Knight, J. C. 2021. Host genetics and infectious disease: new tools, insights and translational opportunities. *Nature Reviews Genetics*, 22: 137–153.

Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N., and Corander, J. 2018. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, 34(24): 4310–4312.

Lees, J. A., Ferwerda, B., Kremer, P. H. C., Wheeler, N. E., Serón, M. V., Croucher, N. J., Gladstone, R. A., Bootsma, H. J., Rots, N. Y., Wijmega-Monsuur, A. J., *et al.* 2019. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nature Communications*, 10(1): 1–14.

Li, M., Li, T., Xiao, X., Chen, J., Hu, Z., and Fang, Y. 2022. Phenotypes, mechanisms and therapeutics: insights from bipolar disorder GWAS findings. *Molecular Psychiatry*, pages 1–13.

Mitov, V. and Stadler, T. 2018. A practical guide to estimating the heritability of pathogen traits. *Molecular Biology and Evolution*, 35: 756–772.

Mozzi, A., Pontremoli, C., and Sironi, M. 2018. Genetic susceptibility to infectious diseases: Current status and future perspectives from genome-wide approaches. *Infection, Genetics and Evolution*, 66: 286–307.

Müller, N. F., Stolz, U., Dudas, G., Stadler, T., and Vaughan, T. G. 2020. Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 117(29): 17104–17111.

Naret, O., Chaturvedi, N., Bartha, I., Hammer, C., Fellay, J., and The Swiss HIV Cohort Study (SHCS) 2018. Correcting for population stratification reduces false positive and false negative results in joint analyses of host and pathogen genomes. *Frontiers in Genetics*, 9.

NCBI 2022. GenBank and WGS statistics. `https://www.ncbi.nlm.nih.gov/genbank/statistics/`. Accessed: 2022-6-14.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., *et al.* 2014. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488).

San, J. E., Baichoo, S., Kanzi, A., Moosa, Y., Lessells, R., Fonseca, V., Mogaka, J., Power, R., and de Oliveira, T. 2020. Current affairs of microbial genome-wide association studies: Approaches, bottlenecks and analytical pitfalls. *Frontiers in Microbiology*, 10: 3119.

Shu, L., Blencowe, M., and Yang, X. 2018. Translating GWAS findings to novel therapeutic targets for coronary artery disease. *Frontiers in Cardiovascular Medicine*, 0.

Stadler, T. 2010. Sampling-through-time in birth-death trees. *Journal of Theoretical Biology*, 267: 396–404.

Stadler, T., Kühnert, D., Bonhoeffer, S., and Drummond, A. J. 2013. 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: 228–233.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1): vey016.

Topol, E. J., Murray, S. S., and Frazer, K. A. 2007. The genomics gold rush. *JAMA*, 298(2): 218–221.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., and Lappalainen, T. 2021. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1): 1–21.

Vercelli, D. 2008. Discovering susceptibility genes for asthma and allergy. *Nature Reviews Immunology*, 8(3): 169–182.

Wakeley, J. 2009. *Coalescent Theory: An Introduction*. Roberts Publishers.

Wang, M., Roux, F., Bartoli, C., Huard-Chauveau, C., Meyer, C., Lee, H., Roby, D., McPeek, M. S., and Bergelson, J. 2018. Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(24): E5440–E5449.

Wetterstrand, K. A. 2019. The cost of sequencing a human genome. `https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost`. Accessed: 2022-6-14.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., *et al.* 2020. A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798): 265–269.

2

A PHYLOGENY-AWARE GWAS FRAMEWORK TO CORRECT FOR HERITABLE PATHOGEN EFFECTS ON INFECTIOUS DISEASE TRAITS

This chapter is submitted as:

Sarah Nadeau, Christian W. Thorball, Roger Kouyos, Huldrych F. Günthard, Jürg Böni, Sabine Yerly, Matthieu Perreau, Thomas Klimkait, Andri Rauch, Hans H. Hirsch, Matthias Cavassini, Pietro Vernazza, Enos Bernasconi, Jacques Fellay, Venelin Mitov[†], Tanja Stadler[†], and the Swiss HIV Cohort Study (SHCS). A phylogeny-aware GWAS framework to correct for heritable pathogen effects on infectious disease traits.

[†]equal contributions

ABSTRACT

Infectious diseases are particularly challenging for genome-wide association studies (GWAS) because genetic effects from two organisms (pathogen and host) can influence a trait. Traditional GWAS assume individual samples are independent observations. However, pathogen effects on a trait can be heritable from donor to recipient in transmission chains. Thus, residuals in GWAS association tests for host genetic effects may not be independent due to shared pathogen ancestry. We propose a new method to estimate and remove heritable pathogen effects on a trait based on the pathogen phylogeny prior to host GWAS, thus restoring independence of samples. In simulations, we show this additional step can increase GWAS power to detect truly associated host variants when pathogen effects are highly heritable, with strong phylogenetic correlations. We applied our framework to data from two different host-pathogen systems, HIV in humans and *X. arboricola* in *A. thaliana*. In both systems, the heritability and thus phylogenetic correlations turn out to be low enough such that qualitative results of GWAS do not change when accounting for the pathogen shared ancestry through a correction step. This means that previous GWAS results applied to these two systems should not be biased due to shared pathogen ancestry. In summary, our framework provides additional information on the evolutionary dynamics of traits in pathogen populations and may improve

GWAS if pathogen effects are highly phylogenetically correlated amongst individuals in a cohort.

2.1 INTRODUCTION

A key goal of genome-wide association studies (GWAS) is to understand the genetic basis of phenotypic variation among individuals. In a typical GWAS, millions of genetic variants from across an organism's genome are screened for statistical association with a trait of interest. Ideally, this procedure identifies variants that are located in, or are in linkage disequilibrium with, alleles that directly affect the trait. If GWAS finds a variant strongly associated with a disease trait, the gene product may be a good drug target (Okada *et al.*, 2014). Even if no single variant has a strong association, many small associations can be aggregated into a polygenic risk score to identify susceptible individuals (Dudbridge, 2013).

It is well-known that GWAS can be sensitive to confounding variables. Shared ancestry among individuals, especially between close relatives, can give rise to spurious genetic correlations with a trait. Corrections for these types of population structure in human GWAS cohorts are well-developed and widely accepted (Astle and Balding, 2009; Price *et al.*, 2006). More recently, analogous methods have been developed for microbial GWAS, where clonal reproduction exacerbates population structure (Power *et al.*, 2017). Phylogenetic methods to account for population structure in microbial GWAS include explicitly testing for lineage-specific effects as in Earle *et al.* (2016) and modified association tests that account for phylogenetic relationships amongst samples as in Collins and Didelot (2018). These approaches are designed to quantify genetic effects from one organism on a trait.

In the infectious disease context, genetic effects from two organisms - the host and the pathogen - may affect an infectious disease trait. GWAS using paired host-pathogen genotype data have previously been done to elucidate the marginal and interaction effects of host and pathogen genetic variants. Methods to account for microbial population structure when testing for marginal host associations or host-pathogen interaction effects include adding the microbial kinship matrix as a random effect in a linear mixed model as in Wang *et al.* (2018) and using principle components derived from either this matrix or the pathogen phylogeny as covariates in a linear model as in Naret *et al.* (2018). These methods focus on capturing and

accounting for correlations due to the pathogen phylogeny, without further investigating the nature of these correlations.

In this work, we draw from the field of phylogenetic comparative methods to propose a new two-step framework that corrects for pathogen population structure and thus satisfies the GWAS assumption of independent samples. The introduced framework relies on paired pathogen-host genotyping and is envisioned specifically for continuous-valued traits that are highly heritable from infection partner to infection partner. We hypothesized that our approach should improve GWAS power to identify host genetic variants broadly associated with disease traits.

In a first step, we fit an evolutionary model to trait data and the pathogen phylogeny. This first step provides an estimate of the correlation structure of the trait due to heritable pathogen effects. The estimate is used to remove pathogen effects on the trait. In the second step, the resulting corrected trait data is used in a GWAS with host genetic variants. The GWAS can be performed as normal under the assumption of independent samples. The main advantage of this two-step approach compared to the previously outlined methods to correct for pathogen population structure is that it generates additional information on the evolutionary dynamics of the trait in the pathogen population. The advances presented here are on the first step, while in the second step existing, highly optimized tools to perform GWAS association tests under a variety of models can be employed.

In the following, we describe the evolutionary model for heritable, continuous-valued infectious disease traits upon which our method is based. We derive a maximum likelihood estimate for the pathogen part of a trait under this model. We then describe a new infectious disease GWAS framework assessing associations of the trait with host genetic variants using the maximum likelihood estimates. In simulations, we show that this framework can improve GWAS power to detect host genetic variants that affect disease traits. Finally, we apply our framework to paired host-pathogen genotyping data from the Swiss HIV Cohort Study (SHCS) and a previously studied *Arabidosis thaliana-Xanthomonas arboricola* pathosystem. We show that associations with set-point viral load (spVL) and quantitative disease resistance (QDR) traits, respectively, are robust to a correction for pathogen effects.

NEW APPROACHES

A statistical model for heritable, continuous-valued infectious disease traits

Variation in infectious disease traits like viral load or infection severity can come from several sources. These include host genetic factors, pathogen genetic factors, interaction effects between the host and the pathogen, or non-genetic factors like healthcare quality or temperature. GWAS typically stratify samples or include covariates to correct for host genetic factors or non-genetic factors that may be correlated with a trait value. This leaves pathogen genetic factors as a remaining source of correlation, since close transmission partners may be infected with very similar pathogen strains. We aim to remove this pathogen-induced correlation in the trait data prior to performing GWAS on the host genomes.

Broad-sense pathogen heritability H^2 quantifies the fraction of total variance in a trait that is "inherited" from infection partner to infection partner, i.e., due to pathogen factors. To characterize H^2 and the heritable and non-heritable factors that determine infectious disease traits, we use a phylogenetic mixed model (PMM) (Housworth *et al.*, 2004). PMMs assume continuous traits are the sum of independent heritable and non-heritable parts. In the infectious disease GWAS case, we assume the heritable part comprises pathogen genetic factors and all other factors are non-heritable. The heritable pathogen part is modeled by a random process occurring in continuous time along the branches of the pathogen phylogeny, as in Figure 2.1A. The non-heritable part is modeled as Gaussian noise added to sampled individuals at the tips of the phylogeny.

PMMs have previously been applied to the study of infectious disease traits using two different types of random processes to model trait evolution. The Brownian Motion (BM) process assumes unbounded trait values, i.e. the trait can attain any value. The Ornstein-Uhlenbeck (OU) process assumes trait values fluctuate around an optimal value, i.e. extreme trait values are unlikely. Here, we assume the more flexible OU process as it encompasses a wider variety of evolutionary scenarios. For example, Mitov and Stadler (2018) and Bertels *et al.* (2018) previously showed the OU process has higher statistical support for HIV-1 spVL. This makes sense given that spVL is likely under stabilizing selection to maximize viral transmission potential (Fraser *et al.*, 2014). The full model is called the phylogenetic Ornstein-Uhlenbeck mixed model (POUMM) and is described in detail by Mitov and Stadler (2018). Here, we review the main points relevant to our method.

Under the POUMM, the trait z is the sum of heritable genetic effects g, i.e. due to the pathogen, and non-heritable "environmental" effects ϵ, i.e. host genetic effects and other environmental or interaction effects:

$$z = g + \epsilon \tag{2.1}$$

g is a pathogen trait that evolves along the phylogeny according to an OU process. The OU process is defined by a stochastic differential equation with two terms. The first term represents a deterministic pull towards an optimal trait value and the second term represents stochastic fluctuations modelled by Brownian motion (Butler and King, 2004):

$$\begin{aligned} dg(t) &= \alpha[\theta - g(t)]dt + \sigma dW_t \\ g(0) &= g_0 \end{aligned} \tag{2.2}$$

Here the parameter α represents selection strength towards an evolutionarily optimal value represented by parameter θ. The parameter σ measures the intensity of stochastic fluctuations in the evolutionary process. Finally, dW_t is the Wiener process underlying Brownian motion. The OU process is a Gaussian process, meaning that $g(t)$ is a Gaussian random variable. Assuming $g(t)$ starts at initial value g_0 at time $t = 0$ at the root of the phylogeny, we can write the expectation for $g(t)$ at time t:

$$E[g(t)] = g_0 e^{-\alpha t} + (1 - e^{-\alpha t})\theta \tag{2.3}$$

and the variance in $g(t)$ if we were to repeat the random evolutionary process many times (Butler and King, 2004):

$$Var[g(t)] = \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t}) \tag{2.4}$$

g evolves independently in descendent lineages after a divergence event in the phylogeny. The covariance between $g(t)$ in a lineage i at time t_i and another lineage j at time t_j, $Cov\big(g_i(t_i), g_j(t_j)\big)$, increases with the amount of time between t_0 and the divergence of the two lineages, $t_{0(ij)}$, and decreases with the total amount of time the lineages evolve independently, d_{ij} (Butler and King, 2004):

$$Cov\big(g_i(t_i), g_j(t_j)\big) = \frac{\sigma^2}{2\alpha}[e^{-\alpha d_{ij}}(1 - e^{-2\alpha t_{0(ij)}})] \tag{2.5}$$

Next, we recall that ϵ is the non-heritable part of the trait. ϵ is modeled as a Gaussian random variable that is time- and phylogeny-independent. The

expectation of ϵ is zero, meaning non-heritable effects are equally likely to raise or lower the trait from the pathogen-determined level. The parameter σ_ϵ^2 measures the between-host variance of the non-heritable effect.

$$\begin{aligned} E(\epsilon) &= 0 \\ Var(\epsilon) &= \sigma_\epsilon^2 \end{aligned} \tag{2.6}$$

Finally, broad-sense trait heritability can be calculated as the fraction of total trait variance that is heritable:

$$H_t^2 = \frac{Var[g(t)]}{Var[g(t)] + Var(\epsilon)} = \frac{\frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t})}{\frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t}) + \sigma_\epsilon^2} \tag{2.7}$$

Teasing apart pathogen and non-pathogen effects on a trait

Given the assumptions of the POUMM, we can estimate a heritable pathogen effect on a trait and a corresponding non-heritable, host and environmental effect. Here, we derive a maximum-likelihood estimate for these values for individuals in a GWAS cohort, given measured trait values and a pathogen phylogeny linking the infecting strains.

Let $\boldsymbol{g}(\boldsymbol{t})$ be a vector of g values, one for each individual in the cohort. $\boldsymbol{t}$ are the sampling times of each individual relative to the root of the phylogeny. To simplify notation, we omit the $\boldsymbol{t}$ from here on. $\boldsymbol{g}$ is a realization of a Gaussian random vector $\boldsymbol{G} \sim \mathcal{N}(\boldsymbol{\mu}_{OU}, \boldsymbol{\Sigma}_{OU})$. The expectation $\boldsymbol{\mu}_{OU}$ is defined by equation 2.3, the diagonal elements of the covariance matrix $\boldsymbol{\Sigma}_{OU}$ are defined by equation 2.4, and the off-diagonal elements of $\boldsymbol{\Sigma}_{OU}$ by equation 2.5. Similarly, let $\boldsymbol{\epsilon}$ be a vector of the non-heritable part of the trait for each individual. $\boldsymbol{\epsilon}$ is a realization of a Gaussian random vector $\boldsymbol{\mathcal{E}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{E}})$, where $\boldsymbol{\Sigma}_{\mathcal{E}}$ is a diagonal matrix with diagonal elements equal to σ_ϵ^2.

Considering that $\boldsymbol{G}$ and $\boldsymbol{\mathcal{E}}$ are independent random vectors and that their realizations $\boldsymbol{g}$ and $\boldsymbol{\epsilon}$ must sum together to equal the observed trait values $\boldsymbol{z}$, we can write the following proportionality for the joint probability density of $\boldsymbol{g}$ and $\boldsymbol{\epsilon}$:

$$f(\boldsymbol{g}, \boldsymbol{\epsilon}) \propto \mathcal{N}(\boldsymbol{g}; \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_G) \tag{2.8}$$

where the expected value of $\boldsymbol{g}$ and the covariance matrix $\boldsymbol{\Sigma}_G$ are defined as:

$$Exp(\boldsymbol{g}) = \boldsymbol{\mu}_G = \boldsymbol{\Sigma}_G\left(\boldsymbol{\Sigma}_{OU}^{-1}\boldsymbol{\mu}_{OU} + \boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\boldsymbol{z}\right) \tag{2.9}$$

$$\boldsymbol{\Sigma}_G = \left(\boldsymbol{\Sigma}_{OU}^{-1} + \boldsymbol{\Sigma}_{\mathcal{E}}^{-1}\right)^{-1} \tag{2.10}$$

Proof.

$$\begin{aligned} f(\boldsymbol{g},\ \boldsymbol{\epsilon}) &= f(\boldsymbol{g}|\ \boldsymbol{\epsilon}) \times f(\boldsymbol{\epsilon}) \\ &= f(\boldsymbol{g}) \times f(\boldsymbol{\epsilon}) \\ &= \mathcal{N}(\boldsymbol{g};\ \boldsymbol{\mu}_{OU}, \boldsymbol{\Sigma}_{OU}) \times \mathcal{N}(\boldsymbol{\epsilon};\ \mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{E}}) \\ &= \mathcal{N}(\boldsymbol{g};\ \boldsymbol{\mu}_{OU}, \boldsymbol{\Sigma}_{OU}) \times \mathcal{N}(\boldsymbol{z} - \boldsymbol{g};\ \mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{E}}) \\ &= \mathcal{N}(\boldsymbol{g};\ \boldsymbol{\mu}_{OU}, \boldsymbol{\Sigma}_{OU}) \times \mathcal{N}(\boldsymbol{g};\ \boldsymbol{z}, \boldsymbol{\Sigma}_{\mathcal{E}}) \end{aligned} \tag{2.11}$$

Equations 2.9 and 2.10 follow from eq. 2.11 and eq. 371, p. 42, section 8.1.8 "Product of Gaussian densities" in (Petersen and Pedersen, 2012). □

Importantly, equation 2.9 is the maximum likelihood estimate for $\boldsymbol{g}$, the pathogen effect on the trait, taking into account all available information - measured trait values, the pathogen phylogeny, and inferred POUMM parameters. This estimator is an inverse-variance weighted average of measured trait ($\boldsymbol{z}$) and information from the POUMM evolutionary model ($\boldsymbol{\mu}_{OU}$). In other words, $\boldsymbol{g}$ will be closer to the measured trait value if the trait is not very heritable. If the trait is highly heritable, $\boldsymbol{g}$ will be closer to the expected value under the POUMM, i.e. take more information from the phylogenetic relationships between infecting strains.

Given the estimator we just derived for $\boldsymbol{g}$, we can now estimate $\boldsymbol{\epsilon}$, the trait value *without* pathogen effects:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{z} - Exp(\boldsymbol{g}) \tag{2.12}$$

We will use this value to try to improve upon standard GWAS methods in infectious disease.

A POUMM-based GWAS framework for infectious disease

We propose to improve standard GWAS for infectious diseases by estimating and removing trait variability due to pathogen effects. Our new framework is as follows:

1. Sample paired host genotypes, pathogen genome sequences, and trait values from a cohort.

2. Construct a pathogen phylogeny using the pathogen genome sequences.

3. Estimate the parameters of the POUMM based on the trait values and the pathogen phylogeny. This can be done with the R package POUMM (Mitov and Stadler, 2017).

4. Generate maximum-likelihood estimates for the pathogen and corresponding non-pathogen effects on the trait using equations 2.9 and 2.12.

5. Perform GWAS with only the non-pathogen effects on the trait as the response variable.

2.2 RESULTS

Simulation study

To test the theoretical best-case performance of our method, we simulated data under the POUMM and applied our framework to the simulated data. We parameterized our simulation scheme with the time-scale and other parameters of an HIV-1 outbreak in mind, with spVL as the trait of interest.

We first simulated a phylogeny of 500 tips with exponentially distributed branch lengths and mean root-to-tip time of 0.14 substitutions per site per year as in Hodcroft *et al.* (2014). Then, we simulated pathogen trait values $\boldsymbol{g}$ along this phylogeny using the POUMM package in R (Mitov and Stadler, 2017). This part of the simulation is illustrated in Figure 2.1A. For the simulation, we considered a range of pathogen heritability parameter values H^2, from 15 to 75%, and a range of selection strength parameters values α, from 0.1 to 60 time^{-1}. The intensity of stochastic fluctuations parameter σ was determined based on H^2 and α (a re-arrangement of equation 2.4, equation given in Table 2.2). As shown in Figure 2.8, higher α values correspond to higher σ values to maintain constant H^2 under this parameterization. For each H^2 and α value considered in the simulation, we recorded the simulated pathogen part of the trait value for each tip in the phylogeny.

We paired each tip's simulated pathogen trait value with a simulated host trait value. Simulated hosts had 20 genome positions. We sampled alleles (0, 1, or 2) for each position from a binomial distribution with probability 0.13. 10 random positions had an effect size of 0.2 on the trait and 10 had an effect size of -0.2. This part of the simulation is illustrated in Figure

2.1B. Our parameterization produced roughly normally distributed host trait values centered at zero with variance equal to 25% of the total trait variance, which we constrained to 0.73 based on the variance in log spVL values measured by Mitov and Stadler (2018). We used 25% host heritability for spVL based on McLaren *et al.* (2015).

Finally, we sampled an additional random environmental effect for each tip from a normal distribution centered at zero, as illustrated in Figure 2.1C. The variance of this distribution was scaled based on the pathogen heritability of the trait, from zero (no effect) in the scenario with 75% pathogen heritability and 25% host heritability to 0.44 in the scenario with 15% pathogen heritability and 25% host heritability. Figure 2.9 provides a more detailed schematic of this simulation framework and Table 2.2 gives the value or expression for each parameter.

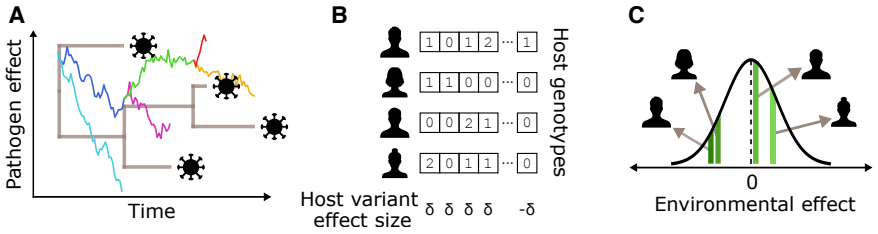


FIGURE 2.1: **A high-level schematic of our phylogenetic Ornstein-Uhlenbeck mixed model (POUMM)-based simulation framework in the context of HIV-1 set-point viral load (spVL).** (**A**) shows how the viral effects on spVL evolve along the viral phylogeny according to an Ornstein-Uhlenbeck process. (**B**) shows how human host genetic effects are the sum of independent effects from several causal variants. Each variant can be present in 0, 1, or 2 copies. Half the variants have a positive effect of size δ and half have a negative effect of size δ. (**C**) shows how other environmental effects are independently drawn from a Gaussian distribution centered at zero. These three effects sum to the trait value for each simulated individual.

Estimator accuracy

First, we evaluated how well our method estimated the additive host genetic effects from the simulated data. Additive host genetic effects represent an ideal (albeit unattainable) baseline for infectious disease GWAS. Figure 2.2A

shows that our method incorporating phylogenetic information can more accurately estimate these value compared to the trait value. To ensure a fair comparison, we scaled trait values to have the same mean, zero, as host genetic effects so as not to bias the root mean squared error (RMSE) by a constant factor. As shown in the supplemental material, we can calculate the expected RMSE using the scaled trait value across scenarios in our simulation scheme because the variance in the trait due to pathogen genetic effects and environmental effects is fixed. Thus, we expect the RMSE using the scaled trait value to be 0.74 across all simulation scenarios. By incorporating phylogenetic information, we can improve upon this error in scenarios where the trait is highly heritable, under low selection pressure, and with relatively moderate stochastic fluctuations compared to outbreak duration. Figure 2.3 gives some intuition for how this correction works by contrasting simulated scenarios with high and low heritability and low selection strength/ low stochastic fluctuations. Depending on these parameters, trait values are more or less phylogenetically correlated (see also Figure 2.4) and the phylogeny is more or less useful for accurately estimating the heritable pathogen and corresponding non-heritable, non-pathogen part of the trait values.

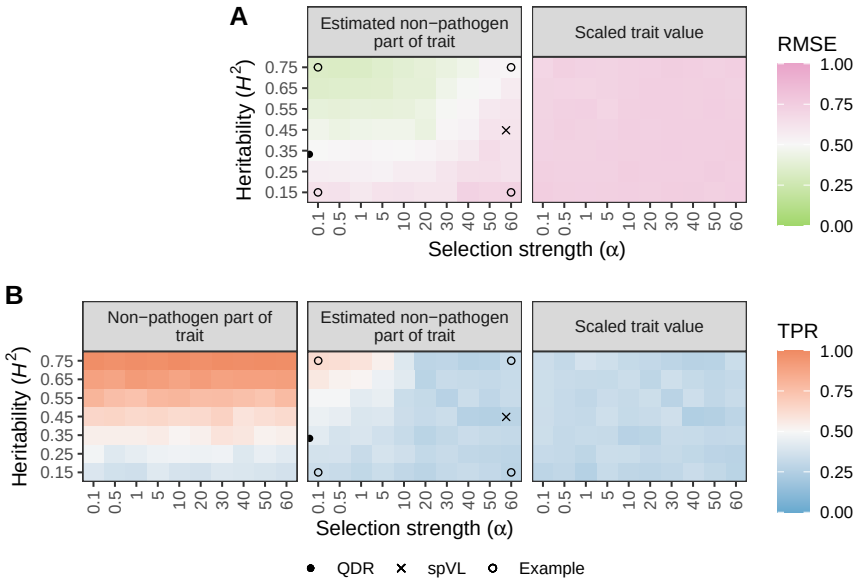


FIGURE 2.2: **Results from the simulation study.** We simulated host, pathogen, and environmental effects on a trait under the phylogenetic Ornstein-Uhlenbeck mixed model (POUMM) with different heritability (H^2; y-axis) and selection strength (α; x-axis) parameters. For each simulated dataset, we applied our method to estimate the non-pathogen effects and performed GWAS with these values. (**A**) shows the root mean squared error (RMSE) of our estimator (left) compared to un-corrected trait values, scaled by their mean (right) under each simulated evolutionary scenario. The RMSE is with reference to the true (simulated) host part of the trait values. Thus, more accurate estimates (lower RMSE) mean the trait value used for GWAS will be closer to the true host part of the trait value. (**B**) shows how genome-wide association study (GWAS) power can improve given the true, simulated non-pathogen effect on spVL (left) and using our estimate for this value (middle) compared to using the scaled trait value (right). Each tile's color corresponds to the average value across 20 simulated datasets of 500 samples. The points highlight specific heritability and selection strength values from the *A. thaliana*-*X. arboricola* quantitative disease resistance (QDR) analysis, HIV-1 spVL analysis, and four simulated scenarios that are presented in more detail in Figure 2.4.

Theoretical GWAS improvement

Next, we characterized the evolutionary scenarios under which our framework can actually improve GWAS power. We used the true positive rate (TPR) to evaluate the fraction of simulated causal host genetic variants we could recover as being significantly associated with the trait. We performed three different GWAS for each simulated dataset: the first represents an ideal in which we can exactly know and remove pathogen effects from trait values, the second is using our method to estimate this value and remove it, and the third represents a standard GWAS using the scaled trait value. Figure 2.2B shows that our framework can improve the TPR in simulated scenarios where selection strength < 10 time^{-1} and heritability $> 45\%$. If we were able to perfectly estimate and remove pathogen effects from a trait, the TPR would increase across all values of selection strength so long as the trait is more than marginally heritable. We estimate approximately 25% to be the heritability threshold above which GWAS power is negatively impacted by pathogen effects. In summary, we show that it is theoretically possible to improve GWAS power for heritable infectious disease traits by estimating and removing pathogen effects using information from the pathogen phylogeny.

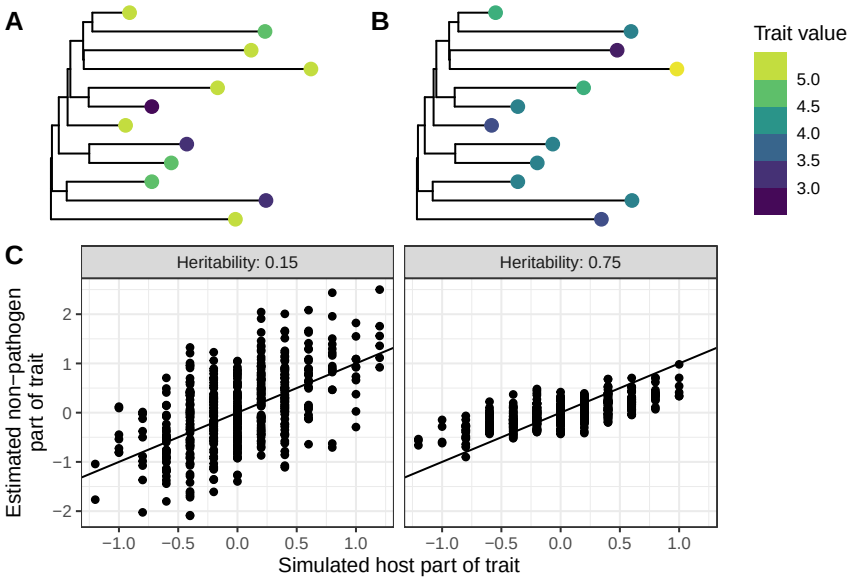


FIGURE 2.3: **Simulated data from two evolutionary scenarios where a phylogenetic correction to trait values improves genome-wide association study (GWAS) power (right side) and where it does not (left side).** These examples correspond to two of the unfilled points in Figure 2.2. (**A**) and (**B**) show total trait values for 12 randomly selected tips from the simulated phylogeny with pathogen heritability H^2 of 15 and 75%, respectively. Depending on the pathogen heritability, trait values are more or less correlated at clustered tips. (**C**) compares our method's estimate for the non-pathogen part of trait values (y-axis) with true simulated host trait values (x-axis) with pathogen heritability of 15 and 75%. The solid line is the y=x line. Selection strength α was fixed to 0.1 time $^{-1}$ for both scenarios and all other parameters were fixed as in the full simulation study.

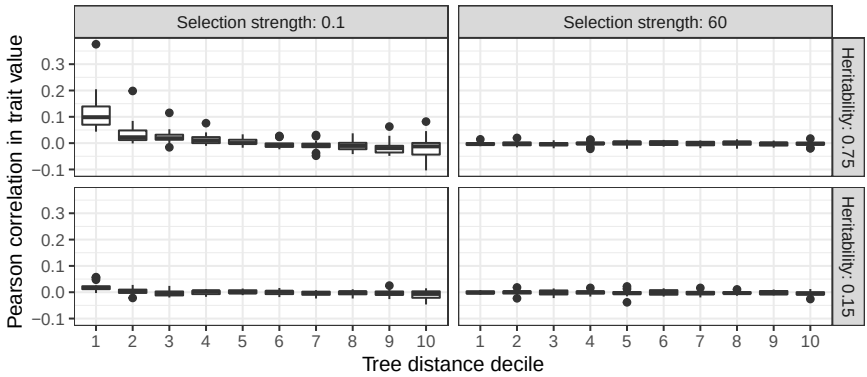


FIGURE 2.4: **Correlations between trait values in pairs of tips in four simulated scenarios.** These examples correspond to the four unfilled points in Figure 2.2. Correlations are calculated for pairs of tips binned by phylogenetic distance (into deciles) across the 20 replicate simulations for each of the four evolutionary scenarios. Trait values are only noticeably correlated for closely clustered tips under the scenario with high pathogen heritability H^2 and low selection strength α / low stochastic fluctuations σ (upper left facet).

Application to HIV-1 set-point viral load

We applied our framework to empirical data from two different host-pathogen systems with different experimental setups (Figure 2.5). First, we used data collected by the Swiss HIV Cohort Study (SHCS) from 1,493 individuals in Switzerland infected with HIV-1 subtype B between 1994 and 2018. The SHCS provided viral load measurements, *pol* gene sequences, and human genotype data for these individuals. We followed the method outlined above to estimate the pathogen and non-pathogen effects on spVL for the cohort from these data. Figure 2.10 shows the calculated (total) spVL values, which vary between approximately 1 and 6 log copies/mL in the cohort. We estimated spVL heritability in this cohort to be 45% (95% highest posterior density, HPD, 24 - 67%) and selection strength to be 58 time^{-1} (95% HPD 19 - 95) (Figure 2.11, Table 2.3). To put these values into the context of our simulation study, they are shown as points on Figure 2.2. The highest expected correlation in trait values between any two tips in the HIV-1 phylogeny under the POUMM was 0.45. However, Figure 2.12

shows that this trait is not obviously phylogenetically structured in the cohort in general, despite high heritability. Finally, Figure 2.13 shows that the estimated non-pathogen effects on spVL correlate quite strongly with total spVL.

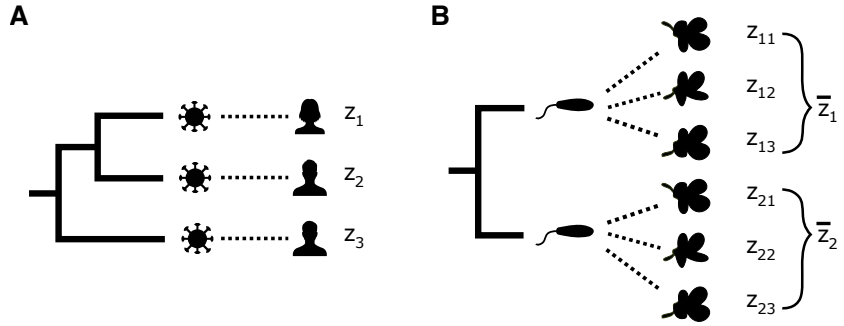


FIGURE 2.5: **A high-level schematic of the experimental setup for the two application datasets.** For (**A**) HIV-1 set-point viral load (spVL) in the Swiss HIV Cohort Study, data are paired viral and human genotypes and associated spVL measurements. We fit the phylogenetic Ornstein-Uhlenbeck mixed model (POUMM) to the viral phylogeny and spVL values associated with each infected individual (z_1, z_2, ..., z_{1493}). For (**B**) *A. thaliana*-*X. arboricola* quantitative disease resistance (QDR) from (Wang *et al.*, 2018), data are bacterial and plant genotypes with QDR measurements for all possible combinations of pathogen and host plant strains. We fit the POUMM to the bacterial phylogeny and mean QDR calculated for each pathogen strain across all the hosts plant types ($\bar{z}_1$, $\bar{z}_2$, ..., $\bar{z}_{22}$).

We compared our proposed GWAS framework with a more standard approach by performing two different GWAS on the same SHCS human genotypes. We retained 1,392 individuals of European ancestry for the GWAS. In the (i) "GWAS with standard trait value" we used the total trait value, calculated spVL values, as the GWAS response variable. In the (ii) "GWAS with estimated non-pathogen part of trait" we used our estimates for the non-pathogen effects on spVL. Figure 2.6A shows that results are qualitatively similar between the two GWAS. Q-Q plots show the distribution of p-values are very similar as well (Figure 2.14). Figure 2.6B shows how the strength of association changed for some variants in

the MHC and *CCR5* regions. Taking into account phylogenetic information slightly decreased association strength for most variants in the *CCR5* region. Association strength increased for some variants in the MHC, for example, SNP rs9265880 had the greatest increase in significance in the MHC region, from a p-value of 3.5×10^{-07} to 7.7×10^{-09}. However, the top-associated variants in the MHC and *CCR5* regions were consistent regardless of the GWAS response variable used (Table 2.4). Finally, Table 2.1 shows how our GWAS results compare for the two top-associated SNPs identified by McLaren *et al.* (2015), who performed the largest standard GWAS for HIV spVL to date. Effect sizes are smaller with a phylogenetic correction and p-values are slightly increased. We repeated the analysis using three different approximate maximum-likelihood phylogenies and these results were consistent (see Materials and Methods; Table 2.5). In summary, there are no clear patterns that point to new regions of association in the human genome with spVL when we take into account the pathogen phylogeny.

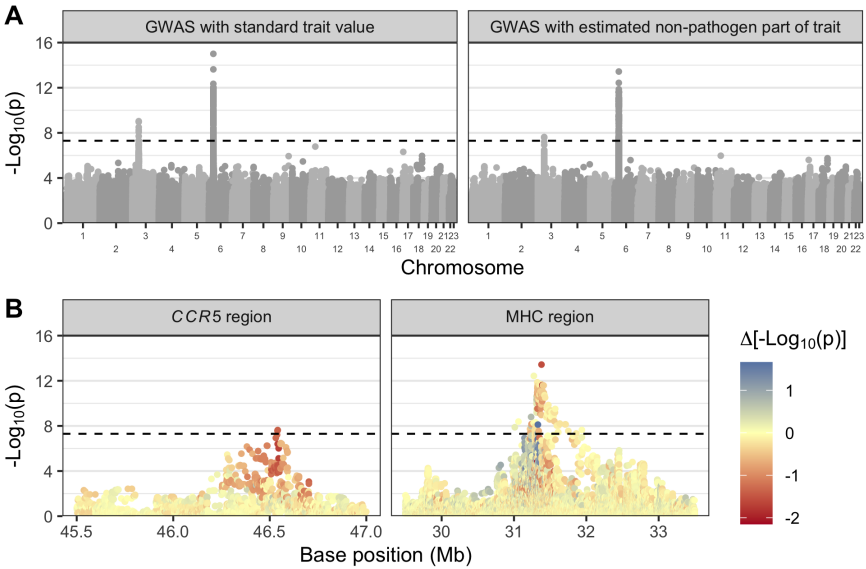


FIGURE 2.6: **Results from comparative genome-wide association studies (GWAS) on HIV-1 set-point viral load (spVL) data.** (**A**) shows association p-values for the same host variants from the Swiss HIV cohort in GWAS with two different response variables. On the left, we used unmodified (total) spVL values. On the right, we used our estimates for the non-pathogen effects on spVL. The alternating shades correspond to different chromosomes. (**B**) compares the strength of association for variants in the *CCR5* and MHC regions between the two GWAS (positions 45.4 - 47Mb on chromosome 3 and 29.5 - 33.5Mb on chromosome 6 for the *CCR5* and MHC, respectively). Base positions are with reference to genome build GRCh37. The color of each point represents the difference in -$\log_{10}$ p-value between the two GWAS. Red means taking into account phylogenetic information decreased the strength of association and blue means it increased it. The dashed lines show genome-wide significance at $p = 5 \times 10^{-8}$.

TABLE 2.1: Top association results from McLaren *et al.* (2015) compared to results from this study. Results from this study are for host variants from the SHCS in GWAS with two different response variables. "Standard trait value" means we used the unmodified (total) spVL value and "Estimated non-pathogen part of trait" means we used our estimates for the non-pathogen effects on spVL.

		McLaren et al.	Standard trait value		Estimated non-pathogen part of trait	
Region	Variant	p-value	Effect size	p-value	Effect size	p-value
MHC	rs594402612.0	$\times 10^{-83}$	-0.4	3.3×10^{-11}	-0.22	2.6×10^{-10}
CCR5	rs1015164	1.5×10^{-19}	0.15	7.5×10^{-7}	0.078	8.5×10^{-6}

Application to the A. thaliana-X. arboricola *pathosystem*

Next, we applied our method to data collected from the *A. thaliana-X. arboricola* pathosystem by Wang *et al.* (2018). Wang *et al.* (2018) performed a fully-crossed experiment in which they infected genetically diverse *A. thaliana* accessions with genetically diverse strains of the phytopathogenic bacteria *X. arboricola*. They scored quantitative disease resistance (QDR) on a scale of zero (resistant) to 4 (susceptible) for up to four infected leaves for three replicates of each *A. thaliana-X. arboricola* pairing. Our method requires a single trait value per pathogen strain, so we used mean QDR calculated for each pathogen strain across all the host *A. thaliana* types (Figure 2.5B). Figure 2.15A shows the inferred *X. arboricola* pathogen phylogeny annotated with the mean QDR trait value used for each strain. Mean QDR was generally low, varying between 0.11 for strain NL_P126 and 0.78 for strain FOR_F21. Fitting the POUMM yielded very low selection strength α and intensity of stochastic fluctuations σ parameter estimates (posterior mean 0.03 with 95% HPD 0.0 - 0.05 and 0.03 with 95% HPD 0.0 - 0.06, respectively; Table 2.6). These values deviated significantly from the respective priors (Figure 2.16). Heritability, on the other hand, was quite uncertain (posterior mean 0.33 with 95% HPD 0.0 - 0.77; Table 2.6). The posterior mean selection strength and heritability values are also shown in the context of the simulation study as points on Figure 2.2.

Given the posterior mean estimates for the POUMM parameters, expected correlation in trait values between tips were very low (maximum value 3.2×10^{-12} compared to maximum value of 0.45 in the HIV-1 spVL application). Thus, the phylogeny is not very informative for a trait value correction. Indeed, the estimated pathogen part of the QDR trait calculated by our method is simply a scaling of the total QDR trait value (Figure 2.17). We anyways selected 22 random host-pathogen strain pairings to perform a comparative GWAS analogous to that for HIV-1 spVL, where each host is infected with a single pathogen strain. In the first GWAS, we used the specific QDR measurement for each selected host-pathogen pairing. I.e., with reference to Figure 2.5, we selected z_{11} for the first sample, z_{23} for the second sample, and so on. In the second GWAS, we used our estimates for the non-pathogen effects on QDR for each pairing. Since our method did not utilize phylogenetic information in this case, the estimated non-pathogen part of the trait is simply the specific QDR for each selected host-pathogen pairing, minus mean QDR for the respective pathogen strain, calculated across all the host *A. thaliana* types. I.e., with reference to Figure 2.5, we used a scaled version of $z_{11} - \bar{z}_1$ for the first sample, $z_{23} - \bar{z}_2$ for the first sample, and so on. Figure 2.7 shows that results are qualitatively similar between the two GWAS, with a slight decrease in association strength for the top-associated variants. Q-Q plots show the distribution of p-values are also very similar (Figure 2.18). In the first, standard GWAS, one *A. thaliana* loci just exceeds the threshold for significant association after correction for multiple testing. In the second, corrected GWAS, no *A. thaliana* variants are significantly associated with QDR to *X. arboricola*.

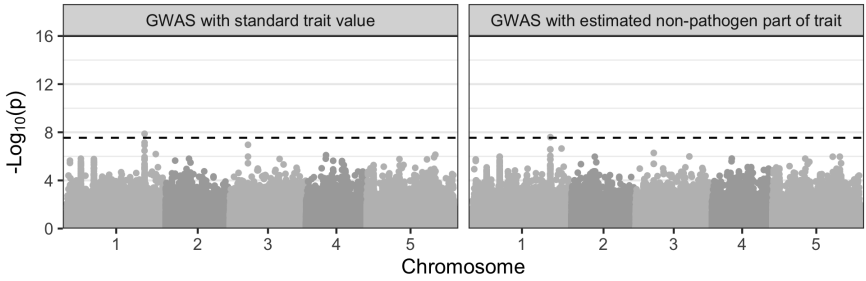


FIGURE 2.7: **Results from comparative genome-wide association studies (GWAS) on *A. thaliana* quantitative disease resistance (QDR) to *X. arboricola*.** The two facets show association p-values for the same host *A. thaliana* variants in GWAS with two different response variables. On the left, we used unmodified (total) QDR values for each of the 22 selected host-pathogen pairings on which these results are based. On the right, we used our estimates for the non-pathogen effects on QDR for these samples. In this case, estimated non-pathogen effects are the specific QDR for each selected host-pathogen pairing, minus mean QDR for the respective pathogen strain, calculated across all the host *A. thaliana* types. The alternating shades correspond to different chromosomes. The dashed lines show significance at significance level 0.05 with a Bonferroni correction for multiple testing.

2.3 DISCUSSION

In this paper, we presented a new phylogeny-aware GWAS framework to correct for heritable pathogen effects on infectious disease traits. By using information from the pathogen phylogeny, we show that it is possible to improve GWAS power to detect host genetic variants associated with a disease trait. This improved power is envisioned to contribute to a better understanding of which host factors are broadly protective against a disease versus which increase susceptibility or disease severity.

The main novelty of our approach is to estimate parameters governing the evolutionary dynamics of a trait in the pathogen population and use these estimates to correct infectious disease trait values prior to performing GWAS, thereby estimating and removing pathogen effects. In simulations, we show that when trait heritability due to shared pathogen ancestry amongst infection partners is greater than approximately 25%, GWAS power to detect host genetic variants associated with the same trait is

reduced. Our method can correct for this effect in certain evolutionary scenarios by using information from the full pathogen phylogeny. Based on our simulation results, our method is anticipated to be very useful for disease traits that are highly heritable from donor to recipient and maintain a high correlation between sampled individuals. In simulations, we showed this is the case when pathogen heritability is high, selection strength is low, and trait values are not subject to strong stochastic fluctuations. In summary, cohort-level, phylogenetically structured differences in the measured trait value are necessary for our approach to outperform state of the art methods.

We applied this model to two different host-pathogen systems where paired host and pathogen genetic data was generated alongside a measure of pathogen virulence. First, we fit the POUMM to set-point viral load data from individuals living with HIV in Switzerland. We estimated HIV-1 spVL heritability to be 45% (95% HPD 24 - 67%) in this cohort. Compared to previous studies, this estimate is at the higher end (see Mitov and Stadler (2018) and references therein). Also using the POUMM, Bertels *et al.* (2018) estimated a spVL heritability of 29% (N = 2014, CI 12 - 46%) from the same cohort and Blanquart *et al.* (2017) estimated 31% (N = 2028, CI 15 - 43%) from a pan-European cohort. We note that our sample size (N = 1493 individuals) is smaller than in these other studies. This might be because we restricted samples based on having *pol* gene sequences with at least 750 non-ambiguous bases. Our aim was to reconstruct a high-quality phylogeny, since the POUMM does not account for phylogenetic uncertainty and the POUMM parameter estimates are key to our downstream trait-correction method. Finally, the inferred selection strength parameter (α = 58 substitutions^{-1} with 95% HPD 19 - 95) was also higher compared to other studies (Table 2.8), though the wide uncertainty interval encompasses these prior estimates. So although our heritability and selection strength estimates are rather high compared to prior studies, the confidence intervals largely overlap the intervals of other studies and we note that estimating these parameters per se was not our primary focus.

For comparison, we also fit the POUMM to quantitative disease resistance measurements from *A. thaliana* infected with the phytopathogenic bacteria *X. arboricola*. We are able to compare inferred parameters across these two systems (in particular the selection strength parameter, which is in per-time units) because both the HIV-1 and *X. arboricola* phylogenies analyzed have branch lengths in units of substitutions. We estimated *X. arboricola* virulence heritability to be 33% (95% HPD 0 - 77%). Wang *et al.* (2018) originally estimated a QDR heritability of 44% in this dataset, falling within the wide

range of our estimate. We note that Wang *et al.* (2018) used a linear mixed model in which the experimental unit is QDR scored on individual leaves, whereas our estimate is based on much coarser binning of QDR scores into a mean score across all leaves on all host accessions and all replicates (N = 22). Furthermore, the QDR score trait values were not truly continuous (scores were measured on an integer scale from 0 to 4). Thus, these data partially violate the assumptions of the POUMM. We estimate very low selection strength for virulence in *X. arboricola*. As Wang *et al.* (2018) explain, *X. arboricola* strains with differing virulence can co-inhabit populations of *A. thaliana*. This might also point to low selection on *X. arboricola* virulence. Furthermore, expected correlation in virulence between related strains of *X. arboricola* was smaller than for HIV-1.

Given our estimates for trait heritability and selection strength on HIV-1 spVL and *A. thaliana* QDR to *X. arboricola*, our simulation results reveal that we cannot expect a significant improvement in GWAS power for these systems (Figure 2.2). Indeed, while certain pairs of samples in the HIV-1 cohort were expected to have phylogenetically correlated spvL values (maximum expected correlation between any two samples was 0.45), the overall effect on GWAS is small. For HIV-1 spVL, our phylogenetic correction slightly decreases p-values for variants in *CCR5* and slightly decreases some and increases other p-values for variants in the MHC (Figure 2.6B). Simulations show we shouldn't expect a net p-value decrease, but our simulations represent an ideal scenario since we simulate under the POUMM. For the empirical data, un-modeled evolutionary pressures like drug treatment and host-specific HLA alleles might cause the reduced p-values. However, the overall picture is consistent between the two GWAS (Figure 2.6A). For *A. thaliana* QDR to *X. arboricola*, the trait value correction does not utilize phylogenetic information because phylogenetic correlations between samples are too weak (maximum expected correlation between strains was 3.2×10^{-12}). We anyways corrected QDR trait values based on average QDR for each pathogen strain across the full range of host types. Results show slight decrease in p-values for the most-associated variants in this application as well, but the overall picture is consistent with previous GWAS results from Wang *et al.* (2018). That study found no significant *A. thaliana* variants associated with QDR using a linear mixed model jointly accounting for host genetic effects, pathogen genetic effects, and interaction effects. As with HIV-1 spVL, our results do not challenge this previous finding. Therefore, we conclude that GWAS for host determinants of HIV-1 subtype B spVL and

A. thaliana determinants of QDR to *X. arboricola* are robust to our correction for pathogen effects.

Our method has several limitations. When POUMM parameter estimates are highly uncertain, correcting trait values based on posterior mean or maximum likelihood parameter estimates neglects this uncertainty. Then, as in the *A. thaliana*-*X. arboricola* application, fitting the POUMM may reveal that expected phylogenetic correlations between samples are not strong enough to justify using our method to correct trait values in a GWAS. In this case, one may wish to use a linear mixed model as in Wang *et al.* (2018), where the pathogen effect is co-estimated as a random effect. The expected correlation structure estimated under the POUMM could be used for the covariance of the random effect, taking the phylogeny into account differently but still utilizing information from the evolutionary model. Finally, as we show here, our method is not anticipated to be useful in certain evolutionary scenarios. For instance, traits like antimicrobial resistance may be under strong selection pressure and be highly heritable. In these instances, our simulations do not point to a large improvement when adding our pre-processing step. In any case, such traits might violate the POUMM assumption that trait values vary as a random walk in continuous space if they are caused by few mutations of strong affect, meaning our approach would not apply. In this situation, one would rather account for antimicrobial resistance as a covariate in the GWAS association model.

The primary advantage of our approach is that it is complementary to previously developed methods for infectious disease GWAS. First, it provides additional information on the evolutionary dynamics of the trait in the pathogen population. Then, it is a convenient pre-processing step for GWAS because it simply produces a corrected response variable for GWAS association tests. In cases where a correction can be estimated and applied using our method, the corrected trait values are envisioned to be used in any of the previously developed GWAS models for the actual association testing (we used a linear model approach implemented in PLINK (Chang *et al.*, 2015), though a more advanced method would be to use a linear mixed model with host ancestry as a random effect). Further, additional model complexity can be added to the GWAS association tests. For instance, our method does not account for co-infection, which might add additional variance to trait values and decrease GWAS power. In this case, one could add co-infection status as a covariate in the GWAS association test to account for this variable.

Our method relies on the freely available R package POUMM (Mitov and Stadler, 2017), which scales to trees of up to 10,000 tips (Mitov and Stadler, 2019). All code for the simulations and HIV spVL analysis presented in this study is available on the project GitHub at https://github.com/cevo-public/POUMM-GWAS. Future applications of our method might investigate other clinically significant disease traits and outcomes that are affected by both host and pathogen genetic factors, for instance Hepatitis B Virus-related hepatocellular carcinoma (An *et al.*, 2018), Hepatitis C treatment success (Ansari *et al.*, 2017), and susceptibility to or severity of certain bacterial infections, e.g. Messina *et al.* (2016); Donnenberg *et al.* (2015). Transcriptomic data has also previously been modeled as an evolving phenotype using an Ornstein-Uhlenbeck model (Rohlfs *et al.*, 2014). Thus, one could also estimate pathogen effects on host gene expression.

In summary, we present a coherent infectious disease GWAS framework that takes the pathogen phylogeny into account when searching for host determinants of a disease trait. We further show that the pathogen phylogeny only has an impact on the GWAS outputs if heritability of the trait amongst infection partners is $> 25\%$. For the systems studied here, spVL in individuals living with HIV and QDR for *X. arboricola* infections in *A. thaliana*, the phylogenetic correction does not change GWAS results. Our findings indicate previously published GWAS results for these systems are not biased by shared evolutionary history amongst infecting pathogen strains.

2.4 MATERIALS AND METHODS

Simulation model

Whenever possible, we tried to parameterize our simulation model using empirical data on the spVL trait. We set the total variance in spVL to 0.73 log copies^2 mL^{-2} based on UK cohort data (Mitov and Stadler, 2018). Other studies have estimated slightly lower values though (Table 2.7). After allotting 25% of this variance to a host part of spVL h based on results by McLaren *et al.* (2015), we partitioned the remaining variance between a viral part g and an environmental part e in different ratios to assess estimator performance across a range of spVL heritabilities. h was simulated as the sum of contributions from 20 causal host genetic variants, 10 of which had an effect size of 0.2 log copies mL^{-1} and 10 of which had an effect size of -0.2 log copies mL^{-1}. Host genetic variants were generated from

a binomial distribution with probability p calculated such that h had the appropriate variance (see Table 2.2). We generated a random viral phylogeny with branch lengths on the same time scale as a previously inferred UK cohort HIV tree (Hodcroft *et al.*, 2014) using the R package ape (Paradis and Schliep, 2018). g was simulated by running an OU process along the phylogeny using the R package POUMM (Mitov and Stadler, 2017) and sampling values at the tips. For the OU parameters θ and g_0 we used 4.5 log copies mL^{-1} based on previous estimates of mean spVL (Table 2.7). This is similar to values previously inferred for HIV (Table 2.8). To assess our estimator's performance under a range of evolutionary scenarios, we co-varied the heritability H^2 and selection strength α parameters. The intensity of random fluctuations σ was determined based on these parameters (Table 2.2, Figure 2.8). Finally, the environmental part of spVL e was generated from a normal distribution with mean zero. For a full graphical model representation of the simulation scheme, see Figure 2.9.

We performed GWAS on the simulated data using a linear association model as implemented in the "lm" function in R. For each simulated dataset, we performed three association tests: (i) using the true (simulated) non-pathogen part of the trait (host + environmental parts), (ii) using the estimated non-pathogen part of the trait according to the method presented in this paper, and (iii) using the total trait value, scaled by its mean. We assessed the significance of each associations at a significance level of 0.05 with a Bonferroni correction for multiple testing. For our main results (Figure 2.2) we simulated 20 truly associated variants, as described above. To also check the false positive rate (FPR), we re-ran the simulations with an additional 80 non-associated variants. Across all the association tests in this second simulation setup (7 H^2 levels $\times$ 10 α levels $\times$ 100 variants $\times$ 20 replicates per scenario = 140,000 association tests), FPR was 0.0005 using the true (simulated) non-pathogen part of the trait, 0.0005 using the estimated non-pathogen part of the trait, and 0.0006 using the scaled total trait value. These rates are comparable to the expected FPR of 0.0005 at significance level 0.05 corrected for 100 tests. Given the stricter correction for multiple testing in this second simulation setup, the TPR decreased significantly across all three GWAS response variables used.

Swiss HIV-1 data

Human genotypes, viral load measurements, and HIV-1 *pol* gene sequences from HIV-1 positive individuals were all collected in the context of other

studies by the Swiss HIV Cohort Study (SHCS) (www.shcs.ch, (Schoeni-Affolter *et al.*, 2010; Scherrer *et al.*, 2021)). All participants were HIV-1–infected individuals 16 years or older and written informed consent was obtained from all cohort participants. The anonymized data were made available for this study after the study proposal was approved by the SHCS.

For phylogenetic inference, we retained sequences from 1,493 individuals with non-recombinant subtype B *pol* gene sequences of at least 750 characters and paired RNA measurements allowing for calculation of spVL, as well as 5 randomly chosen subtype A sequences as an outgroup. We used MUSCLE version 3.8.31 (Edgar, 2004) to align the *pol* sequences with –maxiters 3 and otherwise default settings. We trimmed the alignment to 1505 characters to standardize sequence lengths. We used IQ-TREE version 1.6.9 (Nguyen *et al.*, 2014) to construct an approximate maximum likelihood tree with -m GTR+F+R4 for a general time reversible substitution model with empirical base frequencies and four free substitution rate categories. Otherwise, we used the default IQ-TREE settings. After rooting the tree based on the subtype A samples, we removed the outgroup. Viral subtype was determined by the SHCS using the REGA HIV subtyping tool version 2.0 (de Oliveira *et al.*, 2005). We calculated spVL as the arithmetic mean of viral RNA measurements made prior to the start of antiretroviral treatment. For a comparison of several different filtering methods, see Figure 2.10.

For GWAS, we retained data from 1,392 of the 1,493 SHCS individuals with European ancestry who were not closely related to other individuals in the cohort (Table 2.9). These were 227 females and 1165 males. Ancestry was determined by plotting individuals along the three primary axes of genotypic variation from a combined dataset of SHCS samples and HapMap populations (Figure 2.19). Kinship was evaluated using PLINK version 2.3 (Chang *et al.*, 2015); we used the –king-cutoff option to exclude one from each pair of individuals with a kinship coefficient > 0.09375. Initial host genotyping quality control and imputation were done as in Thorball *et al.* (2021). Subsequent genotyping quality control was performed using PLINK version 2.3 (Chang *et al.*, 2015). We used the options –maf 0.01, –geno 0.01, and –hwe 0.00005 to remove variants with minor allele frequency less than 0.01, missing call rate greater than 0.05, or Hardy-Weinberg equilibrium exact test p-value less than $5x10^{-}5$. After quality filtering, approximately 6.2 million genetic variants from the 1,392 individuals were retained for GWAS (Table 2.10).

A. thaliana-X. arboricola *data*

A. thaliana and *X. arboricola* genotyping and quantitative disease resistance (QDR) measurements were generated by Wang *et al.* (2018) and are described in detail in that publication. Briefly, Wang *et al.* (2018) infected different *A. thaliana* host accessions with different *X. arboricola* pathogen strains in a fully-crossed experimental design. They infected up to 4 leaves on each of three biological replicates for each host-pathogen pairing. Then, they scored QDR for each leaf on a scale of 0 (resistant) to 4 (susceptible). We downloaded the genotype matrix with allele dosage of 33,610 SNPs for the 22 *X. arboricola* pathogen strains generated by Wang *et al.* (2018) from their supplemental material. We additionally downloaded a VCF file with allele dosage of 12,883,854 SNPs for the different *A. thaliana* accessions from the 1001 Genomes project (Alonso-Blanco *et al.*, 2016). QDR measurements were provided directly by the Wang *et al.* (2018) authors.

For phylogenetic inference, we used the "dist.gene" and "nj" functions from the ape package in R to construct a pairwise genetic distance matrix and then a neighbor-joining tree from the *X. arboricola* pathogen genotype matrix. The inferred tree topology (Figure 2.15) closely matches the hierarchical clustering presented in Wang *et al.* (2018), which was generated using the unweighted pair group method with arithmetic mean (UPGMA). Compared to UPGMA, the neighbor-joining method we used relaxes the assumptions of a strict molecular clock and sampling all at the same timepoint. For the trait value to fit the POUMM, we calculated mean QDR across all leaves infected on all hosts for each *X. arboricola* strain (see Figure 2.5B) We used PLINK version 2.0 to select bi-allelic variants from the VCF file using option –max-alleles 2. We then used options –maf 0.1 and –max-maf 0.9 to remove variants with minor allele frequencies less than 0.1 as in Wang *et al.* (2018). After filtering, approximately 1.1 million genetic variants from *A. thaliana* were retained for GWAS (Table 2.11).

POUMM parameter inference

We used the R package POUMM version 2.1.6 (Mitov and Stadler, 2017) to infer the POUMM parameters $g_0, \alpha, \theta, \sigma$, and σ_e from the HIV-1 and *X. arboricola* phylogenies and associated spVL and QDR trait values. The Bayesian inference method implemented in this package requires specification of a prior distribution for each parameter. For HIV-1 spVL, we used the same, broad prior distributions as in (Mitov and Stadler, 2018),

namely: $g_0 \sim \mathcal{N}(4.5,\ 3)$, $\alpha \sim Exp(0.02)$, $\theta \sim \mathcal{N}(4.5,\ 3)$, $H_t^2 \sim \mathcal{U}(0,\ 1)$, and $\sigma_e^2 \sim Exp(0.02)$. For *X. arboricola* QDR, we modified the g_0 and θ priors to match the empirical mean and standard deviation of QDR trait values in the dataset: $g_0 \sim \mathcal{N}(0.4,\ 0.2)$ and $\theta \sim \mathcal{N}(0.4,\ 0.2)$. We ran two MCMC chains for $4x10^6$ samples each with a target sample acceptance rate of 0.01 and a thinning interval of 1000 for both analyses. The first $2x10^5$ samples of each chain were used for automatic adjustment of the MCMC proposal distribution. Figures 2.11 and 2.16 show the posterior distributions for inferred parameters for HIV-1 spVL and *X. arboricola* QDR, respectively. Tables 2.3 and 2.6 give the posterior mean values used for subsequent calculations.

Phylogenetic trait correction

We estimated the pathogen and non-pathogen effects on HIV-1 spVL in humans and *X. arboricola* mean QDR in *A. thaliana* using the method described in this paper. For each individual, we estimated the pathogen part of the trait value using equation 2.9 and the corresponding non-pathogen part using equation 2.12. This is implemented in the function "POUMM:::gPOUMM" in the R package POUMM. In the HIV-1 case, each sample corresponds to one HIV-1 strain with one spVL value. In the *X. arboricola* case, each sample corresponds to one *X. arboricola* strain and the mean QDR score for that strain across all host types (see Figure 2.5). To calculate the expected correlation in trait values between tips in the pathogen phylogeny, we used the function "covVTipsGivenTreePOUMM" in the same package. For the POUMM parameters α, σ, θ, and σ_e, we used the posterior mean estimates generated as described above. All the code used to implement the method is available at `https://github.com/cevo-public/POUMM-GWAS`.

Association testing

We performed two comparative GWAS for each system, using the same host genotype data across the two GWAS. For the first "GWAS with standard trait value" we used the total (uncorrected) trait values ($\boldsymbol{z}$) as the response variable for association testing, replicating a standard GWAS set-up. For the second "GWAS with estimated non-pathogen part of trait" we replaced total trait values with the estimated non-pathogen component of the trait ($\hat{\boldsymbol{\epsilon}}$) as the response variable. Association testing was performed using a linear association model in PLINK version 2.3 and 2.0, respectively (Chang *et al.*, 2015) with the top 5 principle components of host genetic variation included

as covariates. For the HIV-1 spVL GWAS, we additionally included sex as a covariate. The sex and principle components covariates were included to reduce residual variance and control for confounding from host population structure, respectively.

Phylogenetic uncertainty

Our method assumes the phylogeny accurately reflects the evolutionary relationships between pathogen strains. Previously, Hodcroft *et al.* (2014) observed HIV spVL heritability estimates based on *pol* gene sequences were robust to including or not including resistance-associated codons. Our analysis includes these codons. For the HIV application, we additionally tested the sensitivity of the inference to phylogenetic uncertainty. We inferred the phylogeny again, this time using the IQ-TREE option -wt to output all locally optimal trees. We fit the POUMM to two randomly selected trees from this set and repeated the trait correction and association testing steps using these trees and the corresponding POUMM parameter estimates.

Data availability

The simulated data underlying this article can be re-generated using the code available on the project GitHub at `https://github.com/cevo-public/POUMM-GWAS`. The HIV pathogen genome sequences, clinical data, and human genotypes cannot be shared publicly due to the privacy of individuals who participated in the cohort study. The data may be shared on reasonable request to the Swiss HIV Cohort Study at `http://www.shcs.ch`. The *X. arboricola* pathogen genotypes are available in the supplemental material of Wang *et al.* (2018), the *A. thaliana* host genotypes are available at `https://1001genomes.org/`, and the *A. thaliana*-*X. arboricola* QDR measurements are available on request to the authors of Wang *et al.* (2018).

ACKNOWLEDGMENTS

This work was supported by ETH Zurich. We thank the patients who participate in the SHCS; the physicians and study nurses for excellent patient care; A. Scherrer, E. Mauro, and K. Kusejko from the SHCS Data Centre for data management; and D. Perraudin and M. Amstad for administrative assistance. We thank Joy Bergelson for sharing the *A. thaliana*-*X. arboricola*

QDR measurements. We also thank Michael Landis, who shared a LaTeX template for graphical model drawing.

The members of the SHCS are: Abela I, Aebi-Popp K, Anagnostopoulos A, Battegay M, Bernasconi E, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuffi A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H, Fux CA, Günthard HF (President of the SHCS), Hachfeld A, Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Huber M, Kahlert CR (Chairman of the Mother Child Substudy), Kaiser L, Keiser O, Klimkait T, Kouyos RD, Kovari H, Kusejko K (Head of Data Centre), Martinetti G, Martinez de Tejada B, Marzolini C, Metzner KJ, Müller N, Nemeth J, Nicca D, Paioni P, Pantaleo G, Perreau M, Rauch A (Chairman of the Scientific Board), Schmid P, Speck R, Stöckle M (Chairman of the Clinical and Laboratory Committee), Tarr P, Trkola A, Wandeler G, Yerly S.

The Swiss HIV Cohort Study is supported by the Swiss National Science Foundation (grant 201369), by SHCS project 858 and by the SHCS research foundation. Furthermore, the SHCS drug resistance database is supported by the Yvonne Jacob Foundation (to HFG). The data are gathered by the Five Swiss University Hospitals, two Cantonal Hospitals, 15 affiliated hospitals and 36 private physicians (listed in `http://www.shcs.ch/180-health-care-providers`).

BIBLIOGRAPHY

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., Cao, J., Chae, E., Dezwaan, T. M., Ding, W., *et al.* 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell*, 166(2): 481–491.

An, P., Xu, J., Yu, Y., and Winkler, C. A. 2018. Host and viral genetic variation in HBV-related hepatocellular carcinoma. *Frontiers in Genetics*, 9: 261.

Ansari, M. A., Pedergnana, V., Ip, C. L., Magri, A., Von Delft, A., Bonsall, D., Chaturvedi, N., Bartha, I., Smith, D., Nicholson, G., *et al.* 2017. Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nature Genetics*, 49(5): 666–673.

Astle, W. and Balding, D. J. 2009. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, 24(4): 451–471.

Bertels, F., Marzel, A., Leventhal, G., Mitov, V., Fellay, J., Günthard, H. F., Böni, J., Yerly, S., Klimkait, T., Aubert, V., *et al.* 2018. Dissecting HIV Virulence: Heritability of Setpoint Viral Load, CD41 T-Cell Decline, and Per-Parasite Pathogenicity. *Molecular Biology and Evolution*, 35(1): 27–37.

Blanquart, F., Wymant, C., Cornelissen, M., Gall, A., Bakker, M., Bezemer, D., Hall, M., Hillebregt, M., Ong, S. H., Albert, J., *et al.* 2017. Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe. *PLoS Biology*, 15(6): e2001855.

Bonhoeffer, S., Fraser, C., and Leventhal, G. E. 2015. High Heritability Is Compatible with the Broad Distribution of Set Point Viral Load in HIV Carriers. *PLoS Pathogens*, 11(2): e1004634.

Butler, M. A. and King, A. A. 2004. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American Naturalist*, 164(6): 683–695.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1): 7.

Collins, C. and Didelot, X. 2018. A Phylogenetic Method To Perform Genome-Wide Association Studies In Microbes That Accounts For Population Structure And Recombination. *PLoS Computational Biology*, 14(2): e1005958.

de Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E. J., Wensing, A. M. J., van de Vijver, D. A., *et al.* 2005. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, 21(19): 3797–3800.

Donnenberg, M. S., Hazen, T. H., Farag, T. H., Panchalingam, S., Antonio, M., Hossain, A., Mandomando, I., Ochieng, J. B., Ramamurthy, T., Tamboura, B., *et al.* 2015. Bacterial Factors Associated with Lethal Outcome of Enteropathogenic Escherichia coli Infection: Genomic Case-Control Studies. *PLOS Neglected Tropical Diseases*, 9(5): e0003791.

Dudbridge, F. 2013. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, 9(3).

Earle, S. G., Wu, C. H., Charlesworth, J., Stoesser, N., Gordon, N. C., Walker, T. M., Spencer, C. C., Iqbal, Z., Clifton, D. A., Hopkins, K. L., *et al.* 2016. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology*, 1(5).

Edgar, R. C. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5: 113.

Fraser, C., Lythgoe, K., Leventhal, G. E., Shirreff, G., Hollingsworth, T. D., Alizon, S., and Bonhoeffer, S. 2014. Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. *Science*, 343(6177): 1243727.

Hodcroft, E., Hadfield, J. D., Fearnhill, E., Phillips, A., Dunn, D., O'Shea, S., Pillay, D., Leigh Brown, A. J., and the UK HIV Drug Resistance Database and the UK CHIC Study 2014. The Contribution of Viral Genotype to Plasma Viral Set-Point in HIV Infection. *PLoS Pathogens*, 10(5): e1004112.

Höhna, S., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., and Huelsenbeck, J. P. 2014. Probabilistic Graphical Model Representation in Phylogenetics. *Systematic Biology*, 63(5): 753–771.

Housworth, E. A., Martins, E. P., and Lynch, M. 2004. The Phylogenetic Mixed Model. *The American Naturalist*, 163(1): 84–96.

McLaren, P. J., Coulonges, C., Bartha, I., Lenz, T. L., Deutsch, A. J., Bashirova, A., Buchbinder, S., Carrington, M. N., Cossarizza, A., Dalmau, J., *et al.* 2015. Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proceedings of the National Academy of Sciences of the United States of America*, 112(47): 14658–63.

Messina, J. A., Thaden, J. T., Sharma-Kuinkel, B. K., and Fowler, V. G. 2016. Impact of Bacterial and Human Genetic Variation on Staphylococcus aureus Infections. *PLOS Pathogens*, 12(1): e1005330.

Mitov, V. and Stadler, T. 2017. POUMM: An R-package for Bayesian Inference of Phylogenetic Heritability. *ArXiv*.

Mitov, V. and Stadler, T. 2018. A Practical Guide to Estimating the Heritability of Pathogen Traits. *Molecular Biology and Evolution*, 35(3): 756–772.

Mitov, V. and Stadler, T. 2019. Parallel likelihood calculation for phylogenetic comparative models: The SPLITT C++ library. *Methods in Ecology and Evolution*, 10(4): 493–506.

Naret, O., Chaturvedi, N., Bartha, I., Hammer, C., and Fellay, J. 2018. Correcting for Population Stratification Reduces False Positive and False Negative Results in Joint Analyses of Host and Pathogen Genomes. *Frontiers in Genetics*, 9: 266.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1): 268–274.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., *et al.* 2014. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488): 376–381.

Paradis, E. and Schliep, K. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35: 526–528.

Petersen, K. B. and Pedersen, M. S. 2012. *The Matrix Cookbook*. Technical University of Denmark.

Power, R. A., Parkhill, J., and de Oliveira, T. 2017. Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics*, 18(1): 41–50.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8): 904–909.

Rohlfs, R. V., Harrigan, P., and Nielsen, R. 2014. Modeling Gene Expression Evolution with an Extended Ornstein–Uhlenbeck Process Accounting for Within-Species Variation. *Molecular Biology and Evolution*, 31(1): 201–211.

Scherrer, A. U., Traytel, A., Braun, D. L., Calmy, A., Battegay, M., Cavassini, M., Furrer, H., Schmid, P., Bernasconi, E., Stoeckle, M., *et al.* 2021. Cohort Profile Update: The Swiss HIV Cohort Study (SHCS). *International Journal of Epidemiology*, 2021: 1–12.

Schoeni-Affolter, F., Ledergerber, B., Rickenbach, M., Rudin, C., Günthard, H. F., Telenti, A., Furrer, H., Yerly, S., and Francioli, P. 2010. Cohort profile: The Swiss HIV cohort study. *International Journal of Epidemiology*, 39(5): 1179–1189.

Thorball, C. W., Oudot-Mellakh, T., Ehsan, N., Hammer, C., Santoni, F. A., Niay, J., Costagliola, D., Goujard, C., Meyer, L., Wang, S. S., *et al.* 2021. Genetic variation near CXCL12 is associated with susceptibility to HIV-related non-Hodgkin lymphoma. *Haematologica*, 106(8): 2233–2241.

Wang, M., Roux, F., Bartoli, C., Huard-Chauveau, C., Meyer, C., Lee, H., Roby, D., McPeek, M. S., and Bergelson, J. 2018. Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(24): E5440–E5449.

2.5 SUPPLEMENTAL MATERIAL

Expected results from simulations

Here we show the root mean square error (RMSE) of the scaled trait value for each individual, $z_i - \bar{z}$ as an estimate for the host part of the trait for each individual, h_i, should be ≈ 0.74 in our simulation scheme. First we write the expression for the RMSE:

$$RMSE = \sqrt{\frac{\sum_i^N (z_i - \bar{z} - h_i)^2}{N}} \tag{2.13}$$

Note that under our simulation setup, $z_i - \bar{z}$ differs from h_i due to the individual pathogen effect g_i and environmental effect e_i. So the term inside the square root equals the combined variance of these two effects:

$$RMSE = \sqrt{\sigma_g^2 + \sigma_e^2} \tag{2.14}$$

We can calculate the variance due to these two effects because the total variance in spVL σ_z^2, and the fraction of the total variance due to host genetic effects, σ_h^2, are fixed parameters in our simulation scheme.

$$\begin{aligned} \sigma_h^2 + \sigma_g^2 + \sigma_e^2 &= \sigma_z^2 \\ 0.25 * \sigma_z^2 + \sigma_g^2 + \sigma_e^2 &= \sigma_z^2 \\ \sigma_g^2 + \sigma_e^2 &= 0.75 * \sigma_z^2 \\ \sigma_g^2 + \sigma_e^2 &= 0.75 * 0.73 \\ \sigma_g^2 + \sigma_e^2 &= 0.55 \end{aligned} \tag{2.15}$$

Therefore, we can expect the RMSE for $z_i - \bar{z}$ as an estimate for h_i to be around $\sqrt{0.55} \approx 0.74$.

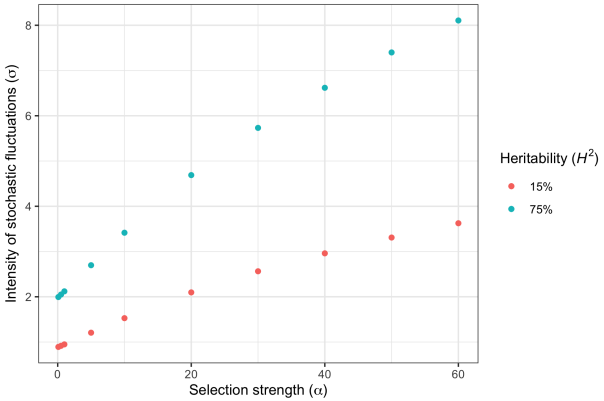


FIGURE 2.8: **Relationship between the selection strength parameter α and the intensity of stochastic evolutionary fluctuations parameter σ at two different heritability (H^2) values in the simulation scheme.** σ was determined as a function of α and H^2 under the POUMM (function given in Table 2.2).

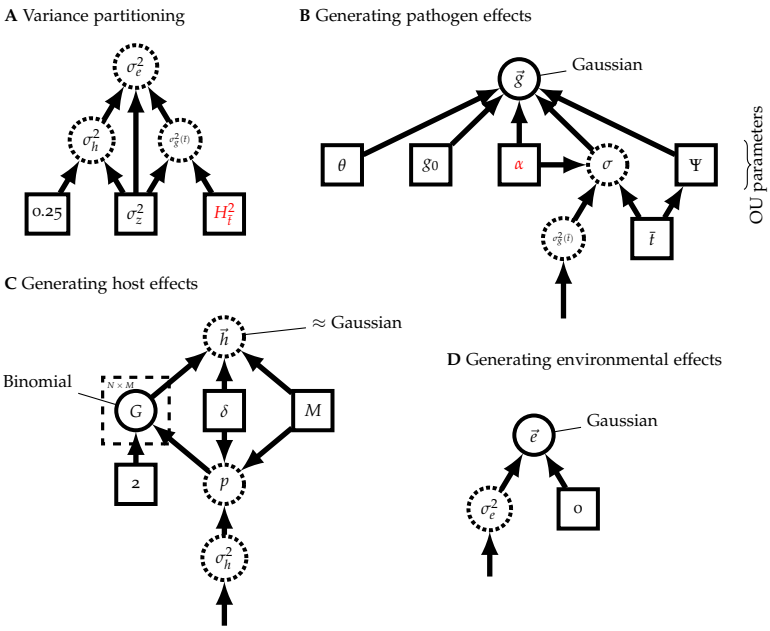


FIGURE 2.9: **A graphical model representation of our simulation scheme, following the recommendations in (Höhna *et al.*, 2014).** Variables in solid squares are constants, with parameters varied from simulation to simulation highlighted in red. Variables in solid circles are realizations of random variables and variables in dashed circles are a function of other variables. Arrows show dependencies and the dashed square represents repetition. (**A**) shows how the variance in the simulated environmental effect σ_e^2 is smaller if the master pathogen heritability value $H_{\vec{t}}^2$ is higher and vice-versa. (**B**) shows the OU parameters and the pathogen phylogeny, which generate the Gaussian-distributed pathogen effects. The OU parameters θ and g_0 are fixed, whereas σ is a deterministic function of the variance in the pathogen effect and the value of α. In other words, we use σ to maintain the desired pathogen heritability while varying α. (**C**) shows how host genotypes are drawn to generate host effects. The host genotype matrix G contains the number of copies (0, 1, or 2) for each of M causal variants with effect size δ. We assume half the variants have a positive effect and half have a negative effect. The allele frequency p for the causal variants set so that we achieve the desired variance in the host effects. (**D**) shows that the environmental effect is drawn from a Gaussian distribution with mean zero and variance as determined in part (A).

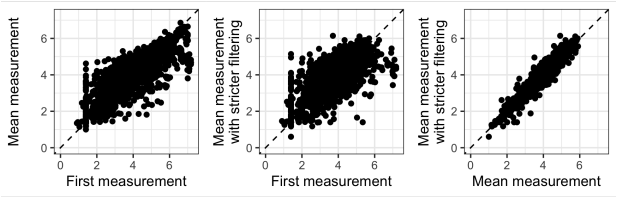


FIGURE 2.10: **A comparison of different ways to calculate spVL based on viral load measurements provided by the SHCS.** The stricter filtering excludes all measurements possibly < 6 months after infection and after treatment or AIDS, whereas the more lenient filtering excludes only measurements after treatment. We used the lenient filter, mean measurement values because these correlate well with the values from the stricter filter but allow us to retain many more individuals from the cohort for our study.

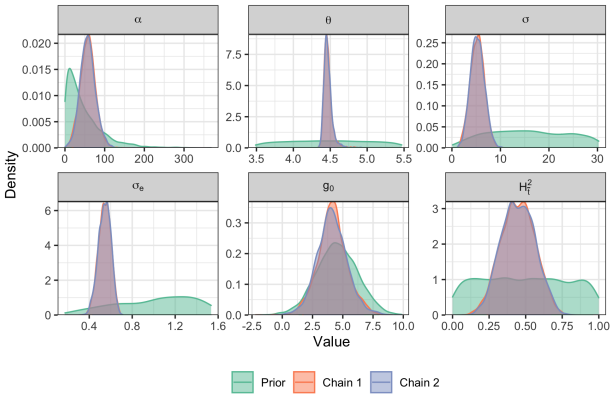


FIGURE 2.11: **Posterior distributions compared to the prior for POUMM parameter estimates based on HIV-1 spVL data from the SHCS.** We ran two different MCMC chains to ensure the estimates converged.

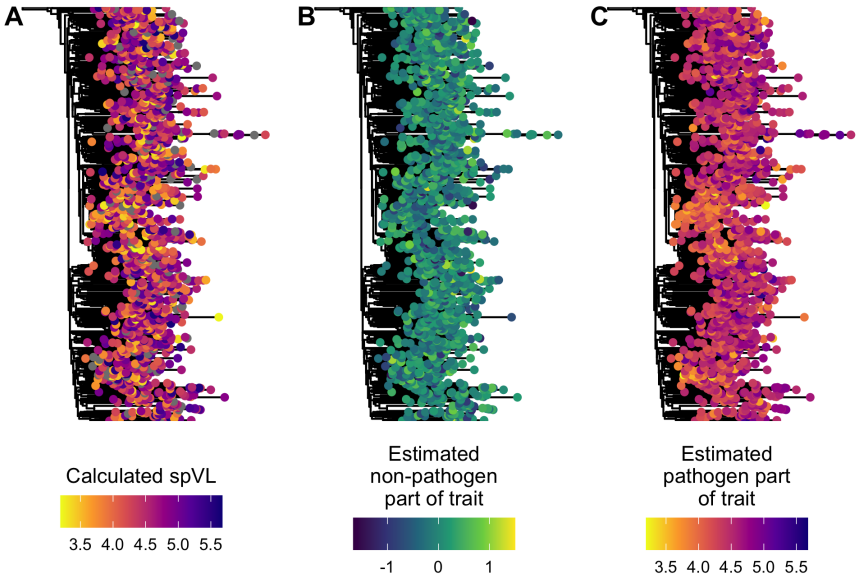


FIGURE 2.12: **Inferred HIV-1 *pol* gene phylogeny with tips colored by (A) calculated spVL, (B) estimated non-pathogen effects on spVL and (C) estimated pathogen effects on spVL.**

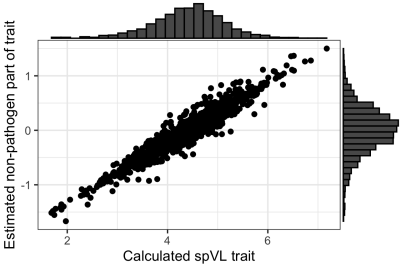


FIGURE 2.13: **A comparison of measured (calculated) spVL values versus our estimated non-pathogen effect on spVL for each SHCS cohort member used in the study.** The histograms show the marginal distribution of each value across the individuals.

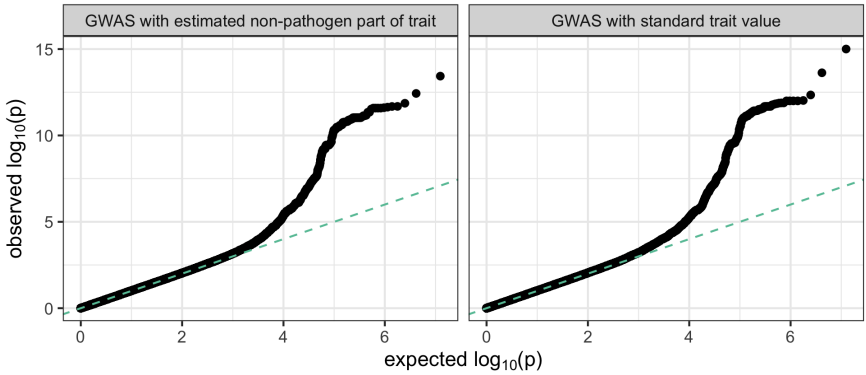


FIGURE 2.14: **Quartile-quartile plots from HIV-1 spVL association tests.** The dashed green line shows the y = x line.

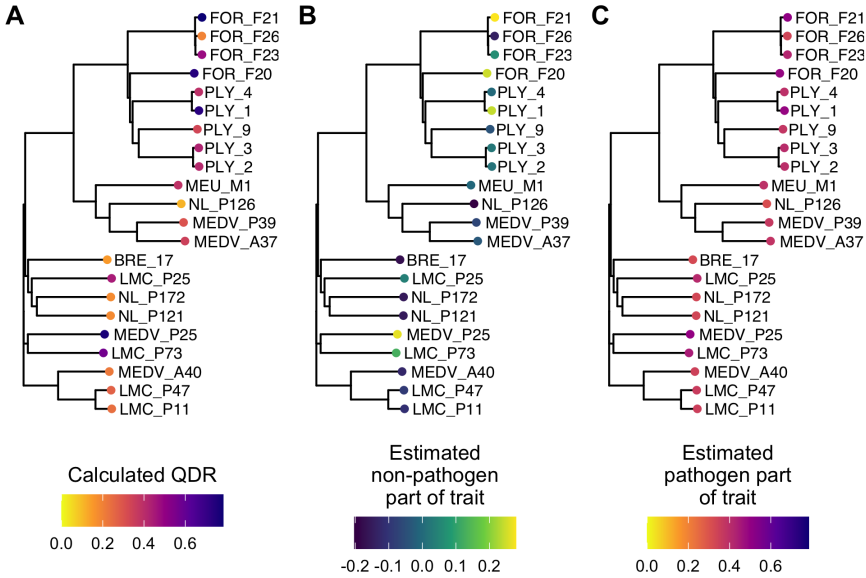


FIGURE 2.15: **Inferred *X. arboricola* phylogeny with tips colored by (A) calculated QDR, (B) estimated non-pathogen effects on QDR and (C) estimated pathogen effects on QDR.**

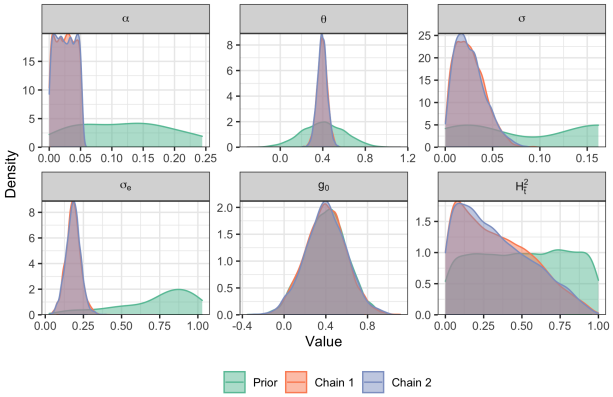


FIGURE 2.16: **Posterior distributions compared to the prior for POUMM parameter estimates based on *A. thaliana*-*X. arboricola* data.** We ran two different MCMC chains to ensure the estimates converged.

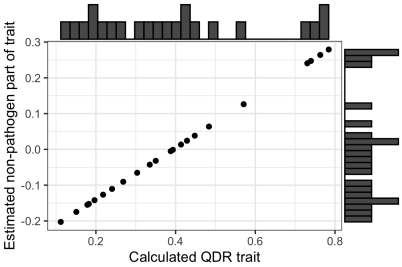


FIGURE 2.17: **A comparison of measured (calculated) mean QDR values across all *A. thaliana* accession pairings and replicates versus our estimated non-pathogen effect on mean QDR for each *X. arboricola* pathogen strain.** The histograms show the marginal distribution of each value across the strains. The Pearson correlation coefficient is 1.

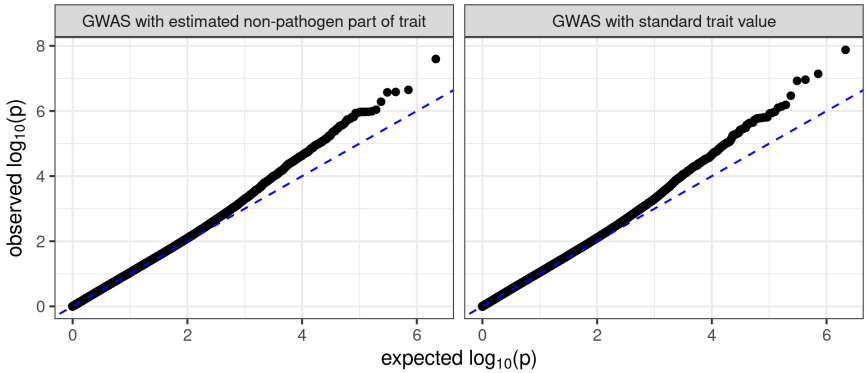


FIGURE 2.18: **Quartile-quartile plots from *A. thaliana*-*X. arboricola* QDR association tests.** The dashed blue line shows the y = x line.

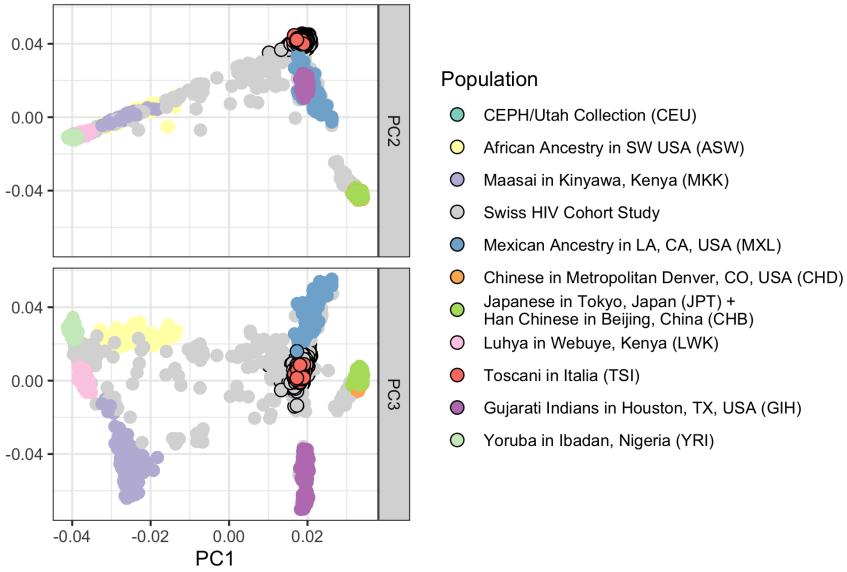


FIGURE 2.19: **SHCS individuals and HapMap3 individuals plotted along the top three principle components of genetic variation.** Points with black borders are within the thresholds used to select individuals of likely European ancestry.

TABLE 2.2: Simulation model parameters. For a full graphical model representation of the simulation scheme, including how these parameters are related, see Figure 2.9.

Variable	Expression	Definition
σ_z^2	0.73 log copies^2/mL^2	Total spVL variance
H_h^2	0.25	Host heritability of spVL
$H_{\bar{t}}^2$	varied	Pathogen heritability of spVL at $\bar{t}$
σ_h^2	$\sigma_h^2 = 0.25 * \sigma_z^2$	Variance in host part of spVL
$\sigma_g^2(\bar{t})$	$\sigma_g^2(\bar{t}) = H_{\bar{t}}^2 * \sigma_z^2$	Variance in pathogen part of spVL at $\bar{t}$
σ_e^2	$\sigma_e^2 = \sigma_z^2 - \sigma_g^2 - \sigma_h^2$	Variance in environmental part of spVL
$\bar{t}$	0.14 substitutions site^{-1} yr^{-1}	Mean root-tip time in pathogen phylogeny
$\boldsymbol{g}$	$\boldsymbol{g} \sim Norm(\boldsymbol{\mu}_{OU}, \boldsymbol{\Sigma}_{OU})$	Pathogen part of spVL for all individuals
θ	4.5 log copies/mL	Optimal spVL value
g_0	4.5 log copies/mL	g at the root of the phylogeny
α	varied	Selection strength of OU process
σ	$\sigma = \sqrt{\frac{2\alpha\sigma_g^2(\bar{t})}{1-exp(-2\alpha\bar{t})}}$	Time-unit standard deviation of OU process
Ψ	branch lengths $\sim Exp(\bar{t})$	Pathogen phylogeny
h_i	$h_i = \delta \sum_{j=1}^{j=M/2} G_{ij} - \delta \sum_{j=M/2}^{j=M} G_{ij}$	Host part of spVL for individual i
$G_{N\times M}$	$G_{ij} \sim Binom(2, p)\ \forall i \in 1...N, \forall j \in 1...M$	Host genotype matrix
p	$p = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{H_h^2\sigma_z^2}{2\delta^2 M}}$	Host variant allele frequency
δ	0.2	Host variant effect size
M	20	Number of causal host variants
e_i	$e_i \sim Norm(0, \sigma_e^2)$	Environmental part of spVL for individual i
N	500	Number of simulated samples

TABLE 2.3: POUMM parameter estimates for spVL based on SHCS data. HPD = Highest posterior density.

Parameter	Posterior mean	95% HPD
g_0	4.23	(1.72, 6.71)
θ	4.47	(4.37, 4.58)
σ	5.25	(2.37, 7.9)
α	57.65	(19.49, 95.2)
σ_ϵ	0.54	(0.43, 0.65)
$H^2_{\tilde{t}}$	0.45	(0.24, 0.67)

TABLE 2.4: Effect size and p-values from the top most strongly associated variants in the *CCR5* and MCR regions from each of the two GWAS performed in our study. "Standard" means the GWAS with standard spVL trait values and "Corrected" means the GWAS with the estimated non-pathogen part of the trait. Entries above the dividing line are the top-associated variants from the "Standard" GWAS and entries below the dividing line are the top-associated variants from the "Corrected" GWAS. Many entries overlap between the two.

Region	Position	Variant	Standard effect size	Standard p-value	Corrected effect size	Corrected p-value
CCR5	46531144	rs9845968	-0.16	5.6×10^{-9}	-0.083	1.2×10^{-7}
CCR5	46537849	rs867620	-0.16	3.2×10^{-9}	-0.085	6×10^{-8}
CCR5	46539864	rs11130092	-0.16	1.1×10^{-9}	-0.087	2.6×10^{-8}
CCR5	46540932	rs10865942	-0.16	8.4×10^{-9}	-0.081	4×10^{-7}
CCR5	46541147	rs7430431	-0.17	9.2×10^{-10}	-0.088	2.3×10^{-8}
MHC	31274380	rs9264942	-0.21	4.5×10^{-13}	-0.12	3.7×10^{-13}
MHC	31321919	rs1055821	-0.33	9.4×10^{-13}	-0.19	1.4×10^{-12}
MHC	31380034	rs112243036	-0.32	9.9×10^{-16}	-0.17	3.7×10^{-14}
MHC	31391401	rs4418214	-0.34	2.4×10^{-14}	-0.18	2.5×10^{-12}
MHC	31400137	rs138130755	-0.46	1×10^{-12}	-0.26	2.6×10^{-12}
MHC	31400705	rs138117378	-0.46	1×10^{-12}	-0.26	2.6×10^{-12}
MHC	31402358	rs148792134	-0.46	1×10^{-12}	-0.26	2.6×10^{-12}
MHC	31409677	rs140991764	-0.46	1×10^{-12}	-0.26	2.6×10^{-12}
CCR5	46531144	rs9845968	-0.16	5.6×10^{-9}	-0.083	1.2×10^{-7}
CCR5	46537849	rs867620	-0.16	3.2×10^{-9}	-0.085	6×10^{-8}
CCR5	46539864	rs11130092	-0.16	1.1×10^{-9}	-0.087	2.6×10^{-8}
CCR5	46541147	rs7430431	-0.17	9.2×10^{-10}	-0.088	2.3×10^{-8}
CCR5	46556835	rs6808142	0.15	8.3×10^{-8}	0.082	3.3×10^{-7}
MHC	31274380	rs9264942	-0.21	4.5×10^{-13}	-0.12	3.7×10^{-13}
MHC	31321919	rs1055821	-0.33	9.4×10^{-13}	-0.19	1.4×10^{-12}
MHC	31367874	rs111281598	-0.37	1.5×10^{-12}	-0.22	2.1×10^{-12}
MHC	31376266	rs73400361	-0.37	1.4×10^{-12}	-0.22	2.1×10^{-12}
MHC	31380034	rs112243036	-0.32	9.9×10^{-16}	-0.17	3.7×10^{-14}

TABLE 2.5: Results comparison using three different approximate maximum likelihood HIV-1 phylogenies (see Materials and Methods).

	Estimated non-pathogen part of trait main tree		Estimated non-pathogen part of trait tree 2		Estimated non-pathogen part of trait tree 3	
Variant	Effect size	p-value	Effect size	p-value	Effect size	p-value
rs59440261	-0.22	2.6×10^{-10}	-0.22	1.5×10^{-10}	-0.24	1.4×10^{-10}
rs1015164	0.078	8.5×10^{-6}	0.076	9.5×10^{-6}	0.083	9×10^{-6}

TABLE 2.6: POUMM parameter estimates for QDR based on *A. thaliana*-*X. arboricola* data. HPD = Highest posterior density.

Parameter	Posterior mean	95% HPD
g_0	0.40	(0.01, 0.78)
θ	0.39	(0.30, 0.49)
σ	0.03	(0.0, 0.06)
α	0.03	(0.0, 0.05)
σ_ϵ	0.18	(0.08, 0.27)
$H^2_{\bar{t}}$	0.33	(0.0, 0.77)

TABLE 2.7: Summary statistics for log spVL in previously sampled populations. $\bar{z}$ is average spVL (log copies/mL) and σ_z^2 is variance in measured spVL (log copies^2/mL^2). Values from (Blanquart *et al.*, 2017; Mitov and Stadler, 2018) are empirical; values from (Bonhoeffer *et al.*, 2015) were estimated by fitting a normal distribution to the data.

Measurement	Value	Reference
$\bar{z}$	≈ 4.5	(Mitov and Stadler, 2018)
$\bar{z}$	4.4	(Blanquart *et al.*, 2017)
$\bar{z}$	≈ 4.5	(Bonhoeffer *et al.*, 2015)
σ_z^2	0.73	(Mitov and Stadler, 2018)
σ_z^2	0.50	(Blanquart *et al.*, 2017)
σ_z^2	≈ 0.5	(Bonhoeffer *et al.*, 2015)

TABLE 2.8: POUMM parameter estimates for spVL from previous studies.

Parameter	Value (Uncertainty)	Reference	Notes
g_0	5.54 (4.04 - 7.25)	(Mitov and Stadler, 2018)	8,483 UK HIV cohort individuals, *pol* tree
θ	4.45 (4.41 - 4.49)	(Mitov and Stadler, 2018)	
θ	4.0 (1.6 - 4.))	(Bertels *et al.*, 2018)	3,036 SHCS individuals, *pol* tree
θ	4.1 (3.5 - 4.9)	(Blanquart *et al.*, 2017)	1,581 subtype B individuals from Europe, whole genome tree
α	28.78 (16.64 - 46.93)	(Mitov and Stadler, 2018)	
α	32.7 (0.03 - 57.6)	(Bertels *et al.*, 2018)	
α	7.6 (1.2 - 10)	(Blanquart *et al.*, 2017)	**limited α to ≤ 10
σ	2.97 (1.95 - 4.37)	(Mitov and Stadler, 2018)	
σ	1.3 (0.66 -1.87)	(Blanquart *et al.*, 2017)	
σ_e	0.77 (0.73, 0.8)	(Mitov and Stadler, 2018)	
σ_e	0.61 (0.54, 0.65)	(Blanquart *et al.*, 2017)	

TABLE 2.9: Number of samples for HIV-1 spVL GWAS after sequential filtering steps.

Sample filter	Number of samples remaining
Subtype B pol sequences	1516
With paired spVL measurement	1516
$>$ 750 characters in sequence	1493
Individual is of European ancestry	1396
Kinship coefficient $>$ 0.09375	1392

TABLE 2.10: Number of variants for HIV-1 spVL GWAS after sequential filtering steps.

Variant filter	Number of variants remaining
Raw data	76979521
Missing genotype rate > 0.05	11590002
Hardy-Weinburg exact test p-value $< 5 \times 10^{-5}$	11589246
Minor allele frequency < 0.01	6228626

TABLE 2.11: Number of variants for *A. thaliana* QDR GWAS after sequential filtering steps. The last entry lists variants without GWAS p-values because PLINK assessed the correlation between predictor variables (the variant and the top 5 principle components of host genetic variation are predictors) to be too strong. This did not occur in the HIV-1 spVL GWAS.

Variant filter	Number of variants remaining
Raw data	12883854
Bi-allelic variants	11769920
Minor allele frequency < 0.1	1743952
NA p-value (too-high covariate correlation)	1070541

3

THE ORIGIN AND EARLY SPREAD OF SARS-COV-2 IN EUROPE

This chapter is published as:

Sarah A. Nadeau, Timothy G. Vaughan, Jérémie Scire, Jana S. Huisman, and Tanja Stadler (2021). The origin and early spread of SARS-CoV-2 in Europe. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(9). https://doi.org/10.1073/PNAS.2012008118

ABSTRACT

The investigation of migratory patterns during the SARS-CoV-2 pandemic before border closures in Europe is a crucial first step towards an in-depth evaluation of border closure policies. Here we analyze viral genome sequences using a phylodynamic model with geographic structure to estimate the origin and spread of SARS-CoV-2 in Europe prior to border closures. Based on SARS-CoV-2 genomes, we reconstruct a partial transmission tree of the early pandemic, including inferences of the geographic location of ancestral lineages and the number of migration events into and between European regions. We find that the predominant lineage spreading in Europe has a most recent common ancestor in Italy and was probably seeded by a transmission event in either Hubei or Germany. We do not find evidence for preferential migration paths from Hubei into different European regions or from each European region to the others. Sustained local transmission is first evident in Italy and then shortly thereafter in the other European regions considered. Before the first border closures in Europe, we estimate that the rate of occurrence of new cases from within-country transmission was within the bounds of the estimated rate of new cases from migration. In summary, our analysis offers a view on the early state of the epidemic in Europe and on migration patterns of the virus before border closures. This information will enable further study of the necessity and timeliness of border closures.

3.1 INTRODUCTION

In response to the pandemic potential of the SARS-CoV-2 virus, many nations closed their borders in order to curb the virus' spread (Connor, 2020). These closures incurred high economic and social costs. To weigh the relative costs and benefits of border closures, it will be important to understand the efficacy of these policies. At the early stages of an outbreak, border closures can delay a pathogen's arrival, thereby giving countries additional time to prepare (WHO, 2020b). However, the success of this strategy depends on timely implementation and a good knowledge of where the pathogen is already circulating. To evaluate the efficacy of border closures in limiting the spread of SARS-CoV-2, it is important to reconstruct the timeline of the early international spread of the virus, before such policies were implemented.

In this analysis, we aim to estimate the early patterns of SARS-CoV-2 transmission into and across Europe. We also address the more specific question of where the predominant SARS-CoV-2 lineage circulating in Europe originated. We hope that by addressing these questions we can inform further analysis of the efficacy of border closures as a strategy to combat SARS-CoV-2.

The SARS-CoV-2 virus was identified as the cause of an epidemic in Wuhan, China in late 2019 (Wu *et al.*, 2020). The epidemic in Wuhan was reported to the WHO on 31 Dec. 2019 and within one month, SARS-CoV-2 was confirmed to have spread to 19 additional countries (WHO, 2020a). By the end of February 2020, the virus was detected in all WHO regions (WHO, 2020c). By late spring 2020, several lineages of the SARS-CoV-2 virus were circulating across the globe. The intermixing of these lineages in different countries and regions suggests that the virus has been transmitted across borders many times (Nextstrain, 2020d).

Here we focus on estimating the early introductions of SARS-CoV-2 into Europe and the virus' migration across European borders. Through national surveillance efforts, the first COVID-19 cases in Europe were detected in France on 24 Jan. 2020 and in Germany on 28 Jan. 2020 (Spiteri *et al.*, 2020; Robert Koch Institute, 2020). Of the 47 cases detected in Europe by 21 Feb. 2020, 14 were infected in China, 14 were linked to the initial cases in Germany, 7 were linked to the initial cases in France, and 12 were of unknown origin (Spiteri *et al.*, 2020). In addition to the unknown sources of transmission, some early introductions may not have been detected. This is especially probable given that a significant proportion of infected individu-

als are likely to be asymptomatic (Mizumoto *et al.*, 2020). In summary, it is difficult to draw firm conclusions about the source, number, and timing of SARS-CoV-2 introductions into Europe based on confirmed case data alone.

Viral genomes are an important secondary source of information on outbreak dynamics. If viruses acquire mutations on the same timescale as an outbreak, these mutations can provide information about past transmission events. Phylodynamic methods couple a model of viral evolution describing the mutational process to an epidemiological model describing the transmission process. By fitting the combined model to viral genomes sampled from a cohort of infected individuals, we can infer the evolutionary and epidemiological model parameters. Here we fit a phylodynamic model with geographic structure to SARS-CoV-2 genomes from Hubei, China and 19 European countries before the first borders were closed in these regions. We co-infer the transmission tree linking these sequences, the geographic location of ancestral lineages, migration rates of infected individuals between regions, the effective reproductive number, and the proportion of no-longer infectious cases sequenced in each region.

In addition to these inferences, we specifically focus on estimating the geographic origin of the predominant SARS-CoV-2 lineage in Europe. This lineage is defined by a characteristic amino acid substitution at position 314 in the ORF1b gene from proline to leucine and was provisionally named the "A2a" lineage by the Nextstrain team, later renamed to 20A. In the more dynamic, tree-based "pangolin" nomenclature suggested by (Rambaut *et al.*, 2020), this lineage corresponds to the "B.1" lineage described as "a large lineage that roughly corresponds to the large outbreak in Italy, and has since seeded many different countries" (Áine O'Toole, 2020). As of Apr. 1, 2020, two-thirds of the SARS-CoV-2 sequences collected in Europe belonged to this lineage and just 10% of sequences from the lineage were collected outside Europe (data from (GISAID, 2020), lineages assigned using (Nextstrain, 2020e)). Here, we use the name A2a to refer to the group of SARS-CoV-2 viruses defined by the ORF1b:P314L mutation.

The origin of the A2a lineage was initially controversial, with conflicting reports in the academic and media press (Bedford, 2020; Zehender *et al.*, 2020; Forster *et al.*, 2020; Mavian *et al.*, 2020a). Its characteristic ORF1b mutation was found in some of the earliest confirmed COVID-19 cases in Italy, Switzerland, Germany, Finland, Mexico, and Brazil in late February (Bedford, 2020; Zehender *et al.*, 2020). Intriguingly, a late-January sample from a cluster of infections in Bavaria, Germany linked to business travel from Shanghai, China (Böhmer *et al.*, 2020; Rothe *et al.*, 2020) shares a

mutation at site 614 in the S gene with the A2a lineage, but does not have the A2a lineage-defining ORF1b mutation. This German sample is part of a smaller clade that is closely related to the larger clade of A2a sequences and which was originally designated the "A2" lineage but was later included in the larger 19A (Nextstrain nomenclature) or B (pangolin nomenclature) lineage (Nextstrain, 2020d). As a result, it was hypothesized that a German transmission cluster may have seeded the larger European outbreak (Bedford, 2020; Zehender *et al.*, 2020; Forster *et al.*, 2020). However, it was quickly pointed out that incomplete and biased sampling must be taken into account before this hypothesis can be rigorously addressed (Zehender *et al.*, 2020; Mavian *et al.*, 2020a; Bedford, 2020).

Phylodynamic models with geographic structure aim to account for such biases. Firstly, parameter estimates are generated by integrating over a distribution of potential phylogenies, which acknowledges that we cannot reconstruct the true transmission tree with certainty. Secondly, sampling parameters are allowed to differ between regions, which acknowledges that testing and sequencing resources vary across regions. Here, we fit a phylodynamic model with geographic structure to full-length SARS-CoV-2 genomes collected before 8 Mar. 2020 to (i) estimate the early patterns of SARS-CoV-2 spread into and across Europe, (ii) weigh genomic evidence for competing hypotheses about the geographic origin of the predominant A2a lineage in Europe, (iii) report on the epidemiological parameters, and (iv) compare the rate of new cases arising from within-region transmission versus migration during the early epidemic.

3.2 RESULTS

Testing assumptions about source and sink locations

We assume that during the time span considered, the outbreak in Hubei and the different European outbreaks were only sources and not sinks for SARS-CoV-2 globally. The first assumption follows from the fact that Hubei is the location of the pandemic origin (see Materials and Methods for additional rationale). To test our second assumption that Europe was primarily a source and not a sink of infections before 8 Mar. 2020, we analyzed A2a sequences collected from different global regions on or before that date. We aggregated sequences into five demes: Africa, Asia Oceania, Europe, North America, and South Central America (Table 3.6), and then fit the multi-type birth-death model described in the Materials and Methods to these data.

The most recent common ancestor of the global set of A2a sequences was inferred to be in Europe with 95% posterior probability (Figure 3.7). The posterior distributions for the migration rates into Europe closely matched the prior, thus the data contains little information on these rates (Figure 3.8). However, in the analyzed dataset, zero introduction events were inferred from other parts of the world into Europe, while in total 24 migration events were inferred from Europe to other parts of the world (Table 3.8).

Inference results

SARS-CoV-2 transmission into and across Europe

For our main analysis we focused on estimating patterns of SARS-CoV-2 transmission into and across Europe. Based on the particular set of sequences analyzed, we infer that SARS-CoV-2 was introduced from Hubei into France, Germany, Italy and other European countries approximately 2-4 times each before 8 Mar. 2020 (Table 3.1). The largest number of estimated introductions was 18 from Italy to other European countries. Importantly, these estimates reflect only introductions occurring in the transmission history of the analyzed cases, not the full epidemic. In contrast, the inferred migration rate parameters should describe more general patterns of spread between regions. The sequence data were informative for inferring some, but not all, migration rates. We highlight here only the rates for which the data is the most informative; see Figure 3.3 for a full comparison of posterior and prior distributions. The highest migration rate was inferred to be from Italy into other European countries, with a median rate of 3.7/year. The lowest migration rate was from Italy to Germany, with a median rate of 0.43/year. We can translate these rates into the probability of an infected individual migrating using the fact that migration is modelled as a Poisson process. I.e., we infer it is 10 times more likely that an infected individual travelled from Italy to a country in the "other European" deme than to Germany. However, we note that the magnitude of the rates may be skewed by a bias towards genome sampling among recently returned travelers.

TABLE 3.1: Median inferred number of introductions from each source deme to each sink deme along the transmission tree linking analyzed cases. Hubei is assumed to be a source only. Values in brackets are the upper and lower bound of the 95% highest posterior density interval for these estimates.

Source / sink	France	Germany	Italy	other European
Hubei	3: [0, 6]	4: [1, 6]	2: [0, 6]	4: [0, 8]
France	-	0: [0, 1]	0: [0, 3]	2: [0, 4]
Germany	0: [0, 2]	-	1: [0, 3]	1: [0, 4]
Italy	6: [1, 9]	1: [0, 4]	-	18: [6, 34]
other European	2: [0, 6]	1: [0, 4]	1: [0, 4]	-

A2a lineage origin

The maximum clade credibility tree in Figure 3.1 summarizes the posterior sample of transmission trees linking analyzed sequences. The A2a lineage sequences form a clear clade with posterior probability of 1. The most recent common ancestor of the analyzed A2a sequences is estimated to be in Italy with 89% posterior probability. In contrast, the location of the most recent common ancestor between this clade and the A2 Shanghai-linked German sequence is less certain. This ancestor is inferred to have been in either Germany (45% posterior probability), Hubei (30%), or Italy (23%). It is very improbable that this ancestor was in France or another European country (2% posterior probability). Using a lower prior for migration rates (results shown in Figure 3.11), Hubei is more likely to be the location of this ancestor than Germany (62% posterior probability for Hubei, 16% for Germany).

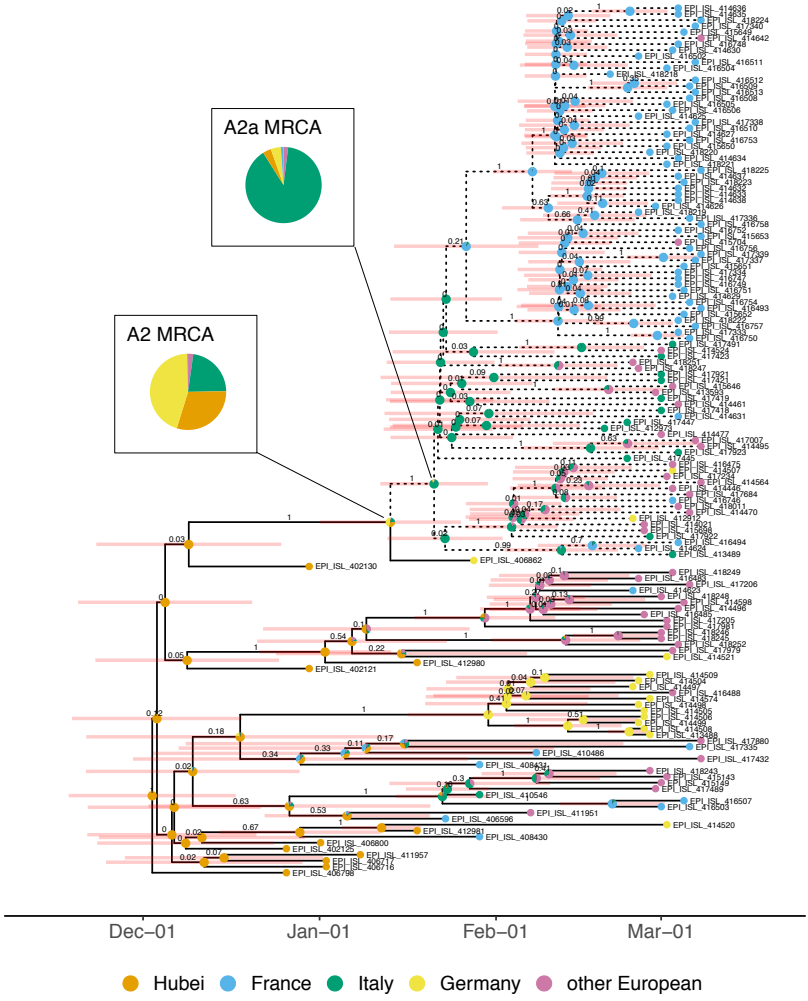


FIGURE 3.1: **Maximum clade credibility tree.** The clade of A2a sequences analyzed is highlighted with dashed branches. The values above the branches are the posterior clade probabilities and the pale red bars show the 95% highest posterior density interval for node ages. The pie charts at nodes show posterior probability for the ancestor being located in each deme (note that we assumed the root of the tree was in Hubei with probability 1). The deme for each tip is the deme in which the sequence was collected, irrespective of travel history. Tips are annotated with GISAID accession identifier.

Epidemiological parameters

Several epidemiologically relevant parameters were co-inferred along with the transmission tree. Firstly, we report on the reproductive number in the different demes, which varied from 1.2 to 1.9 in Hubei to 2.5 to 3.5 in France (Figure 3.4A). Secondly, we report on the prevalence of no-longer infectious cases in each deme as of the collection date of the last analyzed sequence. This quantity can be back-calculated from the estimated sampling proportion (prevalence = sequences analyzed / sampling proportion). We note that both the sampling proportion and prevalence estimates have large credible intervals (Figures 3.4B,C). Of the European demes analyzed, the outbreak in Germany was estimated to be smaller in early March (150 to 485 cumulative cases) than the outbreaks in France (709 to 2,185 cases) and other European countries (719 to 1,782 cases), while the outbreak in Italy was the largest (2,600 to 4,923 cases).

Comparing rates of migration and within-region transmission

Figure 3.2 compares the rate at which we estimate new cases to arise in each region from migration versus from within-region transmission. The estimated rates of new cases from migration and within-region transmission are represented here as point estimates 5 days before the date of case confirmation, which assumes a 5-day delay between infection and onward transmission or migration (the choice of 5 days is motivated by serial interval estimates for SARS-CoV-2 (Ali *et al.*, 2020)). We emphasize that we do not consider any non-European regions beyond Hubei; therefore, transmission from Hubei to a not-included location and then to Europe is considered to be migration directly from Hubei to Europe under our model.

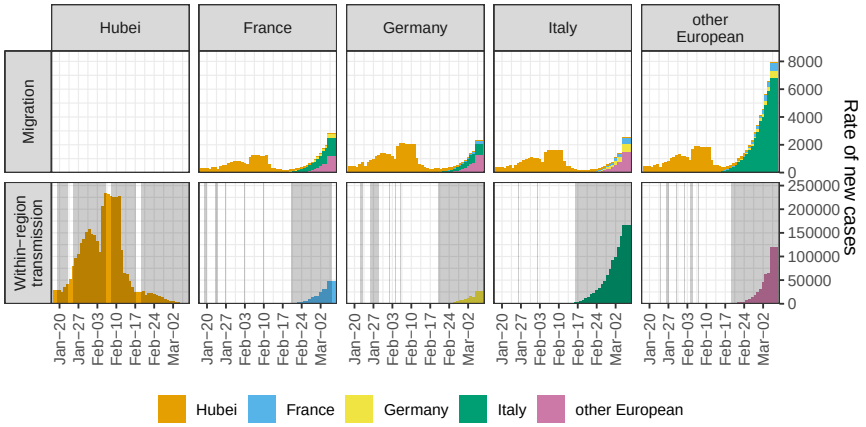


FIGURE 3.2: **Estimated rate of new cases arising from migration compared with the estimated rate of new cases arising from within-region transmission.** For each day, we multiplied the (smoothed) number of newly confirmed cases in each source region by the posterior sample of migration rates from source to sink. The median of these rates is show in the "Migration" row. We also multiplied the (smoothed) number of newly confirmed cases in each sink region by the posterior sample of transmission rates for the region. The median of these rates is shown in the "Within-region transmission" row. Grey shaded regions indicate dates on which new cases were reported in each region. Dates are lagged 5 days to account for a 5-day delay between infection and migration or onward transmission and daily case counts were smoothed by taking a rolling 7-day average. Case data comes from (for Systems Science and at Johns Hopkins University, 2020).

Beginning with the first day on which we have case data from Hubei, we estimate a substantial risk of infected individuals migrating from Hubei into European regions. Throughout late January to mid-February 2020, cases were sporadically detected in each European region, each of which is associated with a risk of subsequent within-region transmission. Sustained within-region transmission is first evident in Italy in mid-February. Shortly thereafter, sustained within-region transmission occurred in other European countries, in France, and in Germany. By 8 Mar. 2020, the estimated rate of occurrence of new cases from within-region transmission is within or exceeds the estimated bounds on the rate of new cases from migration for each region considered (Figure 3.9A). We obtain the same qualitative result in our sensitivity analysis using a very different prior on the migration

rate (Figure 3.9B). We note that the rates in Figure 3.2 are underestimates of the rates of new cases arising due to migration or transmission due to the underreporting in the confirmed case data. However, assuming that the amount of underreporting is comparable across regions, we can indeed compare the rates. Finally, we report support for a decrease in migration rates from the Hubei deme into European demes at the date of the lockdown of Wuhan (Figure 3.3). We infer that migration decreased by 40% (95% HPD 0% - 87%). Again, we note that the migration rate out of Hubei is not necessarily specific to Hubei, since we do not consider possible migration paths through other non-European locations.

3.3 DISCUSSION

We inferred the early spread of the SARS-CoV-2 virus into and across Europe as well as the geographic origin of the predominant A2a lineage spreading in Europe. To do this, we applied a previously published phylodynamic model to analyze publicly available viral genome sequences from the epidemic origin in Hubei, China and from the earliest detected and largest European outbreaks before 8 Mar. 2020. After performing Bayesian inference, we (i) report on inferred patterns of SARS-CoV-2 spread into and across Europe, (ii) compare posterior probabilities for several hypotheses on the origin of the A2a lineage, (iii) report on epidemiological parameters, and (iv) compare the timeline of new cases resulting from migration versus within-region transmission in Europe before borders were closed.

Genome sequence data indicates that prior to 8 Mar. 2020, SARS-CoV-2 was introduced from Hubei province into France, Germany, Italy and other European countries at least 2-4 times each (Table 3.1). These estimates, which are based on genome sequence data and thus do not rely on having line list data for individual migration cases, provide a complementary account of introduction events compared to line list data (Sun *et al.*, 2020) and phylogenetic inferences combining genome sequence and line list data (Gámbaro *et al.*, 2020; Díez-Fuertes *et al.*, 2020; du Plessis *et al.*, 2021; Munnink *et al.*, 2020; Stefanelli *et al.*, 2020; Walker *et al.*, 2020). The introduction events we report here are inferred to have occurred along the transmission tree specific to the analyzed sequence set and are not attributable to individual cases. In comparison, line list data (Spiteri *et al.*, 2020; Sun *et al.*, 2020) attributes introduction events to individual cases but cannot reconstruct previous, unobserved introductions. Since we analyze

only a fraction of all cases, we expect our estimates to be a lower bound on the true number of introductions.

Ideally, we want to go beyond counting migration events amongst the analyzed sequences and investigate general dynamics. To do this, we would interpret inferred migration rates as representing more general patterns of SARS-CoV-2 spread. However, the sequence data was only informative for inferring some of these rates (Figure 3.3). In regions with few lineages circulating during the period considered, there is little signal for the amount of outward migration. We observe information about the per-individual migration rate from Italy to other European countries (Figure 3.3). However, we do not find evidence for preferential migration paths from Hubei into different European regions or from each European region to the others, although we cannot exclude this possibility.

We estimate that the A2a viruses spreading in Europe by 8 Mar. 2020 had a common ancestor in Italy sometime between mid-January and early-February 2020 (Figure 3.1). In contrast, at the time of this paper's original submission Nextstrain placed this ancestor in the U.K. with 100% confidence (Nextstrain, 2020c). This Nextstrain result may have been an artefact of disproportionately high sequencing effort in the U.K. since biased sampling violates the assumptions of the "mugration" method employed (Sagulenko and Neher, 2017). We additionally report that the A2a lineage was most likely carried from Hubei to Italy or from Hubei to Italy via Germany. Both transmission routes have substantial posterior probability under our main model assumptions (Figure 3.1). Assuming a lower migration rate prior, transmission from Hubei to Italy instead of a route via Germany to Italy becomes the more likely scenario (Figure 3.11). Addressing the same question, recently-developed phylodynamic methods accounting for under-sampling and utilizing travel information from line list data have provided even stronger evidence for independent introductions from China into Germany and Italy instead of a route via Germany to Italy (Worobey *et al.*, 2020).

Although it is not the main focus of our analysis, we also report on epidemiological parameters of the early outbreaks considered. Estimates for the reproductive number fall roughly within the range of previous estimates (Liu *et al.*, 2020), though we mention a particular caveat with respect to the reproductive number in Hubei below. Unsurprisingly, prevalence estimates in early March generally exceed confirmed case counts by a factor of 1-3 (Figure 3.4). Our inferences of epidemiological parameters support the idea that the early reproductive number in different outbreaks is difficult to

estimate precisely, but not hugely variable, and that there is substantial under-reporting in line list data (Stringhini *et al.*, 2020).

Finally, we estimated the rate of new cases arising from migration compared with the rate of new cases arising from within-region transmission in the regions analyzed. The magnitudes of these rates are quite uncertain due to uncertainty in the inferred migration and transmission rates (Figure 3.9) and under-reporting in case counts, which we implicitly assume to be constant in time and between demes. However, the temporal trends suggested by these data are still compelling and robust towards different prior assumptions. We see that under sustained risk of case migration from abroad, isolated cases were confirmed throughout Europe beginning in late January 2020 but did not immediately cause large outbreaks. Shortly after the first evidence of sustained within-region transmission in Italy, outbreaks in the rest of Europe also took hold (Figure 3.2).

Our results based on the multi-type birth-death model take into account phylogenetic uncertainty and sampling biases between demes, which are two major concerns in genomic analyses of SARS-CoV-2 (Mavian *et al.*, 2020a). Indeed, wide confidence intervals around internal nodes in the maximum clade credibility tree and low clade support near the tips (Figure 3.1) indicate a high degree of phylogenetic uncertainty. Therefore, it is important that the parameter estimates we report result from integrating over a distribution of potential phylogenies with different geographic locations assigned to ancestral lineages. In comparison, some initial studies that estimated international SARS-CoV-2 spread constructed a median-joining network instead of a phylogeny to account for this uncertainty (Forster *et al.*, 2020; Skums *et al.*, 2020). In this approach, identical sequences are collapsed to single nodes and edges represent mutational differences. This disregards information from relative sampling times and means that ancestor-descendent relationships are highly dependent on the choice of the network root (Sánchez-Pacheco *et al.*, 2020; Kong *et al.*, 2016). Unaccounted-for sampling biases in these analyses may also yield spurious results for the geographic origin of lineages (Chookajorn, 2020; Mavian *et al.*, 2020b). Our analysis, which relies on a mechanistic model of migration and between-deme sampling differences, should be robust to such biases.

Despite the advantages of the multi-type birth-death model just mentioned, there are also several unique caveats to consider. The birth-death model assumes uniform-at-random sampling from the total infected population in each deme. However, particularly in the early stages of outbreaks, infected individuals were identified by health ministries via contact tracing

(Spiteri *et al.*, 2020). Non-random sampling may be one possible explanation for why we infer markedly different transmission rates in China when analyzing cases from within Hubei (as in this analysis) as opposed to cases exposed in Hubei but sequenced elsewhere (as in our previous analysis (Vaughan *et al.*, 2020)). Furthermore, the multi-type birth-death model assumes that parameters are constant through time and homogenous within demes. As a result, our inferences based on province-, country-, and continent-level demes are only coarse approximations of the true, heterogeneous epidemic dynamics occurring at a local level. Due to these limitations, we focus on estimating and interpreting particular events along the transmission tree of the analyzed sequences (e.g. Table 3.1, Figure 3.1) and advise caution when interpreting inferred migration rates (e.g. Figure 3.3).

We expect that our results will be useful in parameterizing more specialized models aimed to understand the efficacy of border closures as a means to fight pandemic disease. So far, such analyses have primarily used line list data and information on travel networks to estimate SARS-CoV-2 migration patterns (Linka *et al.*, 2020; Chinazzi *et al.*, 2020; Wells *et al.*, 2020). Here we present independent estimates of migration patterns based on genome sequence data. By combining case count data and our estimates for migration and transmission rates, we provide a timeline of early SARS-CoV-2 introduction and spread before border closures were implemented. Despite migration risk from outside Europe being on the same order of magnitude as later migration risk from Italy, we only observe sustained outbreaks in other European regions after the onset of sustained within-region transmission in Italy. Finally, before the first border closures in Europe, we estimate the risk of new cases arising from within-region transmission to be within or exceeding the estimated range for the risk of new migration cases.

3.4 MATERIALS AND METHODS

Model

We fit a simplified version of the multi-type birth-death model described in (Scire *et al.*, 2020). Under this model, beginning with a single infected host in a single geographic region (deme), the virus can be transmitted from one host to another (a birth event), die out due to host recovery or death (a death event), be sequenced (a sampling event, assumed to

correspond to a death event), or migrate from one deme to another (a migration event). The birth, death, and sampling processes are assumed to occur at deme-specific rates that are constant through time. Importantly, this model aims to capture heterogeneity in epidemiological parameters (birth and death rates) and sequencing effort (sampling proportion) among demes. Additionally, there is a unique migration rate from each deme to each other deme. All migration rates are assumed to be constant through time except for migration out of Hubei. In our main analysis, migration out of Hubei is assumed to be constant before and after the date of lockdown on 23 Jan. 2020 and is assumed to decrease by a constant factor at the date of lockdown. This factor is a parameter of the model and is also inferred based on the genome sequence data. Finally, we used a version of the model parameterized in terms of the effective reproductive number, which allows us to additionally infer this epidemiologically relevant quantity for each deme.

Dataset

We analyzed SARS-CoV-2 genome sequences from five different demes: Hubei province in China, France, Germany, Italy, and a composite deme of other European countries ("other European"). All sequences were accessed from GISAID (GISAID, 2020) and a full table of sequence identifiers is available in the online supporting information at `https://doi.org/10.1073/pnas.201200811`. To represent the pandemic origin, we randomly chose 10 sequences from Hubei collected on or before the lockdown of Wuhan city on 23 Jan. 2020. To investigate the earliest outbreaks in Europe, we considered all available sequences collected in France, Germany, and Italy on or before the lockdown of the Lombardy region of Italy on 8 Mar. 2020. These countries had the first detected (France and Germany) and the largest (Italy) early outbreaks in Europe (WHO, 2020a; Spiteri *et al.*, 2020). By limiting sampling to before regional lockdowns and border closures went into effect, we hope to (i) satisfy model assumptions that epidemiological and migration parameters are constant through time, and (ii) get a picture of the early, unimpeded spread of SARS-CoV-2 within Europe. To represent the pool of SARS-CoV-2 circulating in other European countries during this time, we randomly down-sampled sequences from other countries to the cumulative number of confirmed COVID-19 deaths in each country by 8 Mar. 2020 plus one (Table 3.4). We used this quantity as a proxy value roughly proportional to the outbreak size in each country. Table 3.2

characterizes the sequences analyzed from each deme for the main analysis. As a sensitivity analysis, we repeated the analysis while down-sampling based on confirmed death data from 28 Mar. 2020, considering that deaths occur with a delay after transmission. This yielded a slightly larger sequence set for analysis. For this analysis, we also did not consider a change in migration rates out of Hubei at 23 Jan. 2020 (results in supplement).

TABLE 3.2: Analyzed sequence information. Location is the location of sample collection, as recorded in the Nextstrain metadata (Nextstrain, 2020e). Date is the date of sample collection, as given on GISAID (GISAID, 2020). No. = number; Seq. = sequence.

Deme	No. seqs.	Locations represented	First seq. date	Last seq. date
Hubei	10	Hubei province, China	26.12.2019	18.01.2020
France	66	France	23.01.2020	08.03.2020
Germany	15	Germany	28.01.2020	03.03.2020
Italy	13	Italy	29.01.2020	04.03.2020
Other European	41	Spain (18), Netherlands (4), United Kingdom (4), Switzerland (3), Belgium (1), Czech Republic (1), Denmark (1), Finland (1), Iceland (1), Ireland (1), Luxembourg (1), Norway (1), Poland (1), Portugal (1), Slovakia (1), Sweden (1)	07.02.2020	08.03.2020

Alignment generation

We prepared a sequence alignment from data publicly available on GISAID (GISAID, 2020) on 1 Apr. 2020 using the Nextstrain pipeline for SARS-CoV-2 (Nextstrain, 2020e). Short sequences (< 25,000 bases), sequences without fully specified collection dates, and sequences in the Nextstrain exclude list (Nextstrain, 2020a) (duplicate sequences from the same case, or with suspicious amounts of nucleotide divergence) were excluded. We aligned selected sequences to reference genome GenBank accession MN908947. To eliminate sites identified by the Nextstrain team as prone to sequencing

errors (Nextstrain, 2020b), we masked the first 130 and final 50 sites from the alignment, as well as sites 18,529, 29,849, 29,851, and 29,853.

Testing assumptions about source and sink locations

We assume that during the time span considered, (i) once a strain was in Europe, the strain could have been transmitted from Europe to other global regions, but subsequent re-introductions of this strain did not occur. Similarly, we assume (ii) strains were not re-introduced into Hubei. These assumptions allow us to ignore sequences from outside of Hubei and Europe. To justify assumption (ii), we argue there was not sufficient time between the pandemic origin in Hubei and Jan. 23, 2020 for a significant amount SARS-CoV-2 export, transmission outside-Hubei, and subsequent re-introduction into Hubei. Furthermore, confirmed case data shows that Hubei province was the epicenter of the SARS-CoV-2 pandemic until this time, with comparatively less transmission occurring outside of the province than within it (WHO, 2020a). To justify assumption (i), we tested whether there was evidence for significant migration into European demes by running a separate analysis on A2a SARS-CoV-2 sampled from all global regions (results in supplement).

Parameter inference

For inferences, we used the implementation of the multi-type birth-death model in the bdmm package (Scire *et al.*, 2020; Kühnert *et al.*, 2016) in the BEAST2 software (Bouckaert *et al.*, 2019). Since this is a parameter-rich model, we fixed some parameters to improve the identifiability of others. The values for fixed parameters, priors for estimated parameters, and the rationale behind these decisions are given in Table 3.3. We ran four MCMC chains to approximate the posterior distribution of the model parameters. The first 10% of samples from each chain were discarded as burn-in before samples from the chains were pooled. We used Tracer (Rambaut *et al.*, 2018) to assess the convergence and confirm that ESS was > 200 for all parameters.

TABLE 3.3: Values and priors for the parameters of the multi-type birth-death model. Confirmed case data for Hubei came from Statista (Statista, 2020), for Germany, France, and Italy from the World Health Organization (WHO, 2020a), and for other European countries from the European Center for Disease Control (ECDC, 2020). The number of analyzed sequences divided by the number of confirmed cases provides an upper bound to the sampling proportion since confirmed cases are only a fraction of total cases. No. = number; approx. = approximately; IQR = inter-quartile range, LogN = Lognormal, Unif = Uniform.

Parameter	Value or Prior	Rationale
Nucleotide substitution model	HKY + Γ	Unequal transition/transversion rates, unequal base frequencies, rate heterogeneity among sites
Clock rate	0.0008	Approx. 24 mutations/year (Nextstrain, 2020e)
Death rate	36.5 $year^{-1}$	Period between infection and becoming un-infectious assumed exponentially distributed with a mean of 10 days
Sampling start time	23 Dec. 2019	Just before date of first sample
Sampling end time (Hubei only)	23 Jan. 2019	Only included sequences collected until lockdown
Location of origin	Hubei	Putative pandemic origin
Reproductive number	LogN(0.8, 0.5)	Median 2.2, 95% IQR 0.8 - 5.9
Migration rates	LogN(0, 1)	Median time until travel is 1 year, 95% IQR 51 days - 7.1 years
Migration rate decrease from Hubei at lockdown	Unif(0, 1)	Migration out of the Hubei deme is expected to decrease after lockdown
Time of origin	LogN(-1, 0.2)	Median 26 Oct., 95% IQR 22 Aug. - 8 Dec. 2019
Sampling proportion		Upper bounds based on confirmed cases:
Hubei	Unif(0, 0.15)	10/66 cases on 18 Jan. 2020
France	Unif(0, 0.093)	66/706 cases on 8 Mar. 2020
Germany	Unif(0, 0.10)	15/157 cases on 3 Mar. 2020
Italy	Unif(0, 0.005)	13/2,502 cases on 4 Mar. 2020
other European	Unif(0, 0.057)	41/712 cases on 8 Mar. 2020

Comparing rates of migration and within-region transmission

To weigh the significance of cases from migration versus within-region transmission during the early epidemic, we compare the rate at which new cases migrate into a region (= per-individual migration rate x case count in source region) to the rate at which new cases arise from within-region transmission (= transmission rate x case count in sink region). When signal in the sequence data is low, e.g. for some migration rates, our prior assumptions determine the magnitude of these rates. To assess the sensitivity of our main conclusions to the prior, we additionally analyzed the same sequences using a lower migration rate prior (Figure 3.9B). We note that the migration and transmission rates are assumed to be constant through time for this analysis, with the exception of the decrease in migration out of Hubei at 23 Jan. 2020. Thus, the temporal trends depend largely on the confirmed case data, which we take from the Johns Hopkins Center for Systems Science and Engineering (for Systems Science and at Johns Hopkins University, 2020).

ACKNOWLEDGMENTS

The authors would like to thank Louis du Plessis for helpful feedback on the original manuscript. S.N, T.V., J.S., J.H., and T.S. thank ETH Zürich for funding. SN and TS are supported by the Swiss National Science Foundation (grant number 31CA30$_1$96267).

BIBLIOGRAPHY

Ali, S. T., Wang, L., Lau, E. H., Xu, X. K., Du, Z., Wu, Y., Leung, G. M., and Cowling, B. J. 2020. Serial interval of sars-cov-2 was shortened over time by nonpharmaceutical interventions. *Science*, 369: 1106–1109.

Bedford, T. 2020. Trevor bedford on twitter: "a follow up to yesterday's thread on the possible connection between the bavarian cluster and the italian covid19 epidemic. https://t.co/rkqjrclwg7 1/5" / twitter.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., *et al.* 2019. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, 15: e1006650.

Böhmer, M. M., Buchholz, U., Corman, V. M., Hoch, M., Katz, K., Marosevic, D. V., Böhm, S., Woudenberg, T., Ackermann, N., Konrad, R., *et al.* 2020. Investigation of a covid-19 outbreak in germany resulting from a single travel-associated primary case: a case series. *The Lancet Infectious Diseases*, 20: 920–928.

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., y Piontti, A. P., Rossi, L., Sun, K., Viboud, C., *et al.* 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (2019-ncov) outbreak. *medRxiv*, page 2020.02.09.20021261.

Chookajorn, T. 2020. Evolving covid-19 conundrum and its impact. *Proceedings of the National Academy of Sciences of the United States of America*, 117: 12520–12521.

Connor, P. 2020. 91amid covid-19. `https://www.pewresearch.org/fact-tank/2020/04/01/more-than-nine-in-ten-people-worldwide-live-in-countries-with-travel-re`

du Plessis, L., McCrone, J. T., Zarebski, A. E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T. R., *et al.* 2021. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*, 371(6530): 706–712.

Díez-Fuertes, F., Iglesias-Caballero, M., Monzón, S., Jiménez, P., Varona, S., Cuesta, I., Ángel Zaballos, Thomson, M., Jiménez, M., Pérez, J. G., *et al.* 2020. Phylodynamics of sars-cov-2 transmission in spain. *bioRxiv*, page 2020.04.20.050039.

ECDC 2020. Download the daily number of new reported cases of covid-19 by country worldwide. `https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide`.

for Systems Science, C. and at Johns Hopkins University, E. C. 2020. Covid-19 data repository. `https://github.com/CSSEGISandData/COVID-19`.

Forster, P., Forster, L., Renfrew, C., and Forster, M. 2020. Phylogenetic network analysis of sars-cov-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America*, page 202004999.

GISAID 2020. gisaid.org.

Gámbaro, F., Behillil, S., Baidaliuk, A., Donati, F., Albert, M., Alexandru, A., Vanpeene, M., Bizard, M., Brisebarre, A., Barbet, M., *et al.* 2020. Introductions and early spread of sars-cov-2 in france, 24 january to 23 march 2020. *Eurosurveillance*, 25: 2001200.

Kong, S., Sánchez-Pacheco, S. J., and Murphy, R. W. 2016. On the use of median-joining networks in evolutionary biology. *Cladistics*, 32: 691–699.

Kühnert, D., Stadler, T., Vaughan, T. G., and Drummond, A. J. 2016. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Molecular Biology and Evolution*, 33: 2102–2116.

Linka, K., Peirlinck, M., and Kuhl, E. 2020. The reproduction number of covid-19 and its correlation with public health interventions. *Computational Mechanics*, 66: 1.

Liu, Y., Gayle, A. A., Wilder-Smith, A., and Rocklöv, J. 2020. The reproductive number of covid-19 is higher compared to sars coronavirus. *Journal of Travel Medicine*, 27.

Mavian, C., Marini, S., Manes, C., Capua, I., Prosperi, M., and Salemi, M. 2020a. Regaining perspective on sars-cov-2 molecular tracing and its implications. *medRxiv*, 21: 1–9.

Mavian, C., Pond, S. K., Marini, S., Magalis, B. R., Vandamme, A.-M., Dellicour, S., Scarpino, S. V., Houldcroft, C., Villabona-Arenas, J., Paisie, T. K., *et al.* 2020b. Sampling bias and incorrect rooting make phylogenetic network tracing of sars-cov-2 infections unreliable. *Proceedings of the National Academy of Sciences of the United States of America*, 117: 12522–12523.

Mizumoto, K., Kagaya, K., Zarebski, A., and Chowell, G. 2020. Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board the diamond princess cruise ship, yokohama, japan, 2020. *Eurosurveillance*, 25: 2000180.

Munnink, B. B. O., Nieuwenhuijse, D. F., Stein, M., Áine O'Toole, Haverkate, M., Mollers, M., Kamga, S. K., Schapendonk, C., Pronk, M., Lexmond, P., *et al.* 2020. Rapid sars-cov-2 whole-genome sequencing and analysis for informed public health decision-making in the netherlands. *Nature Medicine*, 26: 1405–1410.

Nextstrain 2020a. ncov/exclude.txt at master · nextstrain/ncov · github. `https://github.com/nextstrain/ncov/blob/master/defaults/exclude.txt`.

Nextstrain 2020b. ncov/parameters.yaml at master · nextstrain/ncov · github. `https://github.com/nextstrain/ncov/blob/master/defaults/parameters.yaml`.

Nextstrain 2020c. Nextstrain / ncov / europe. `https://nextstrain.org/ncov/europe?c=country&label=clade:20A`.

Nextstrain 2020d. Nextstrain / ncov / global. `https://nextstrain.org/ncov/global?c=region`.

Nextstrain 2020e. nextstrain/ncov: Nextstrain build for novel coronavirus (ncov). `https://github.com/nextstrain/ncov`.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. 2018. Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic Biology*, 67: 901–904.

Rambaut, A., Holmes, E. C., Hill, V., OToole, A., McCrone, J., Ruis, C., du Plessis, L., and Pybus, O. G. 2020. A dynamic nomenclature proposal for sars-cov-2 to assist genomic epidemiology. *bioRxiv*, page 2020.04.17.046086.

Robert Koch Institute 2020. Beschreibung des bisherigen ausbruchsgeschehens mit dem neuartigen coronavirus sars-cov-2 in deutschland (stand: 12. februar 2020). https://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2020/07/Art_02.html.

Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., Zimmer, T., Thiel, V., Janke, C., Guggemos, W., *et al.* 2020. Transmission of 2019-ncov infection from an asymptomatic contact in germany. *New England Journal of Medicine*, 382: 970–971.

Sagulenko, P. and Neher, R. 2017. Inference of transition between discrete characters and 'mugration' models — treetime 0.7.6 documentation. https://treetime.readthedocs.io/en/latest/tutorials/mugration.html.

Scire, J., Barido-Sottani, J., Kühnert, D., Vaughan, T. G., and Stadler, T. 2020. Improved multi-type birth-death phylodynamic inference in beast 2. *bioRxiv*, page 2020.01.06.895532.

Skums, P., Kirpich, A., Baykal, P. I., Zelikovsky, A., and Chowell, G. 2020. Global transmission network of sars-cov-2: from outbreak to pandemic. *medRxiv*, page 2020.03.22.20041145.

Spiteri, G., Fielding, J., Diercke, M., Campese, C., Enouf, V., Gaymard, A., Bella, A., Sognamiglio, P., Moros, M. J. S., Riutort, A. N., *et al.* 2020. First cases of coronavirus disease 2019 (covid-19) in the who european region, 24 january to 21 february 2020. *Eurosurveillance*, 25: 2000178.

Statista 2020. Coronavirus (covid-19) cumulative cases by day worldwide 2020. https://www.statista.com/statistics/1103040/cumulative-coronavirus-covid19-cases-number-worldwide-by-day/.

Stefanelli, P., Faggioni, G., Presti, A. L., Fiore, S., Marchi, A., Benedetti, E., Fabiani, C., Anselmo, A., Ciammaruconi, A., Fortunato, A., *et al.* 2020. Whole genome and phylogenetic analysis of two sarscov-2 strains isolated in italy in january and february 2020: Additional clues on multiple introductions and further circulation in europe. *Eurosurveillance*, 25.

Stringhini, S., Wisniak, A., Piumatti, G., Azman, A. S., Lauer, S. A., Baysson, H., Ridder, D. D., Petrovic, D., Schrempft, S., Marcus, K., *et al.* 2020. Seroprevalence of anti-sars-cov-2 igg antibodies in geneva, switzerland (serocov-pop): a population-based study. *The Lancet*, 396: 313–319.

Sun, K., Chen, J., Viboud, C., Viboud, C., Sun, K., and Chen, J. 2020. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. *The Lancet Digital Health*, 2: e201–e208.

Sánchez-Pacheco, S. J., Kong, S., Pulido-Santacruz, P., Murphy, R. W., and Kubatko, L. 2020. Median-joining network analysis of sars-cov-2 genomes is neither phylogenetic nor evolutionary. *Proceedings of the National Academy of Sciences of the United States of America*, 117: 12518–12519.

Vaughan, T. G., Nadeau, S., Scire, J., and Stadler, T. 2020. virological post on basic reproductive number. `http://virological.org/t/phylodynamic-analyses-of-outbreaks-in-china-italy-washington-state-usa-a439/2`.

Walker, A., Houwaart, T., Wienemann, T., Vasconcelos, M. K., Strelow, D., Senff, T., Hülse, L., Adams, O., Andree, M., Hauka, S., *et al.* 2020. Genetic structure of sars-cov-2 reflects clonal superspreading and multiple independent introduction events, north-rhine westphalia, germany, february and march 2020. *Eurosurveillance*, 25: 2000746.

Wells, C. R., Sah, P., Moghadas, S. M., Pandey, A., Shoukat, A., Wang, Y., Wang, Z., Meyers, L. A., Singer, B. H., and Galvani, A. P. 2020. Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proceedings of the National Academy of Sciences of the United States of America*, 117: 7504–7509.

WHO 2020a. Covid-19 situation reports. `https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports`.

WHO 2020b. Updated who recommendations for international traffic in relation to covid-19 outbreak. `https://www.who.int/news-room/articles-detail/updated-who-recommendations-for-international-traffic-in-relation-to-cov`

WHO 2020c. Who coronavirus disease (covid-19) dashboard. `https://covid19.who.int/`.

Worobey, M., Pekar, J., Larsen, B. B., Nelson, M. I., Hill, V., Joy, J. B., Rambaut, A., Suchard, M. A., Wertheim, J. O., and Lemey, P. 2020. The emergence of sars-cov-2 in europe and north america. *Science*.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y., *et al.* 2020. A new coronavirus associated with human respiratory disease in china. *Nature 2020 579:7798*, 579: 265–269.

Zehender, G., Lai, A., Bergna, A., Meroni, L., Riva, A., Balotta, C., Tarkowski, M., Gabrieli, A., Bernacchia, D., Rusconi, S., *et al.* 2020. Genomic characterization and phylogenetic analysis of sars-cov-2 in italy. *Journal of Medical Virology*.

Áine O'Toole 2020. hcov-2019/lineages: Resources for calling and describing the circulating lineages of sars-cov-2. `https://github.com/hCoV-2019/lineages`.

3.5 SUPPLEMENTAL MATERIAL

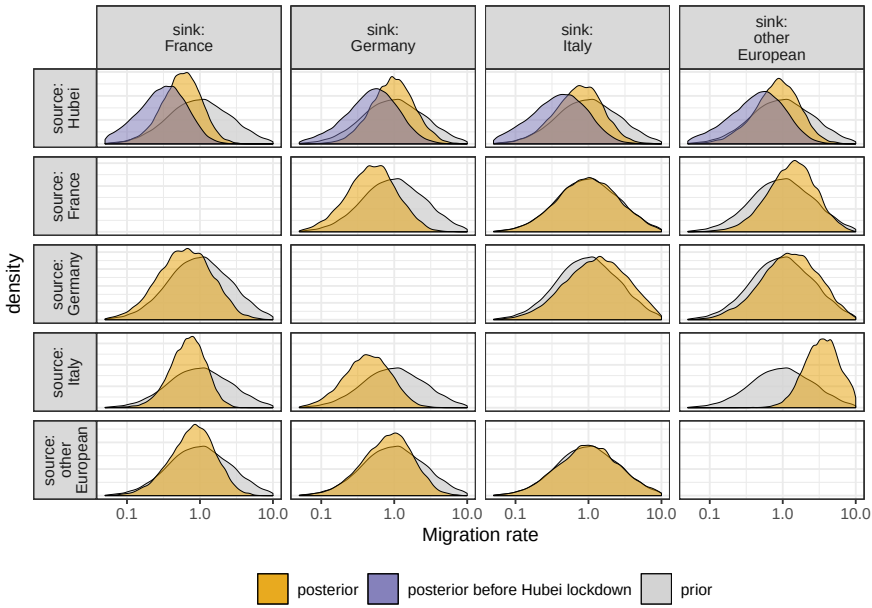


FIGURE 3.3: **Posterior distributions for migration rates (yellow for rates out of Hubei before 23 Jan. 2020 and rates between non-Hubei demes, purple for rates out of Hubei after 23 Jan. 2020) compared to the prior distribution (grey).**

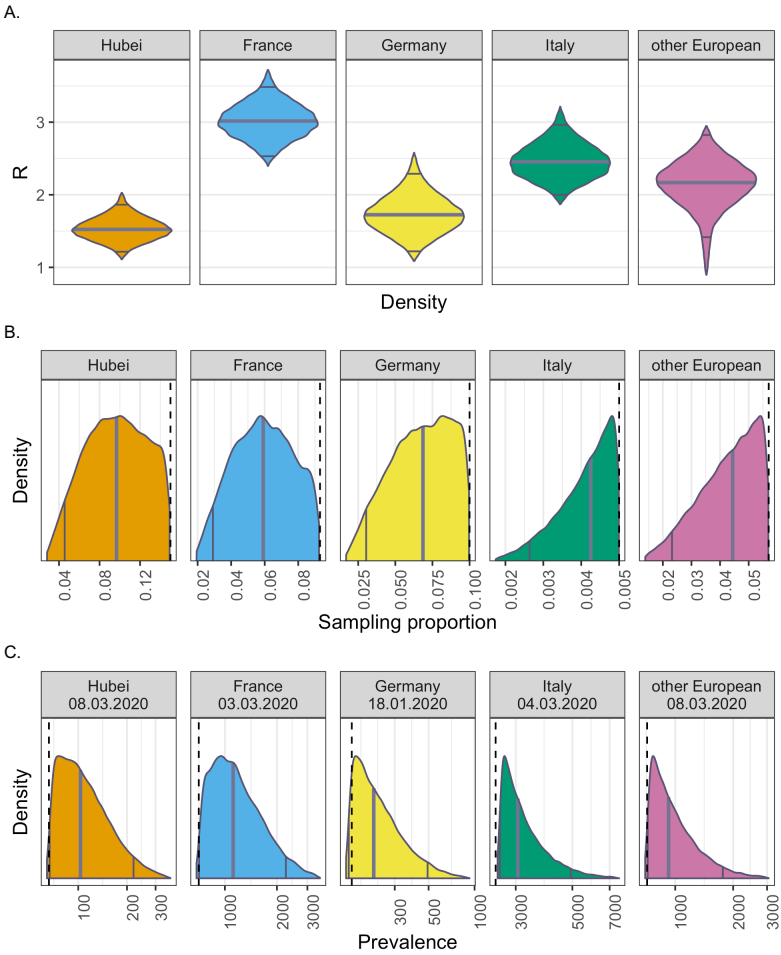


FIGURE 3.4: **Posterior distributions for inferred parameters in the main analysis.** (**A**) reproductive number, (**B**) sampling proportion, and (**C**) the number of no-longer infectious cases on the date of the last analyzed sequence (dates in facet titles) for each region. The dashed lines in (B) show the upper bound of the uniform sampling proportion prior for each region and in (C) the number of confirmed cases on the date of the last analyzed sample. The solid grey lines show the 95% highest posterior density interval and the median for each posterior.

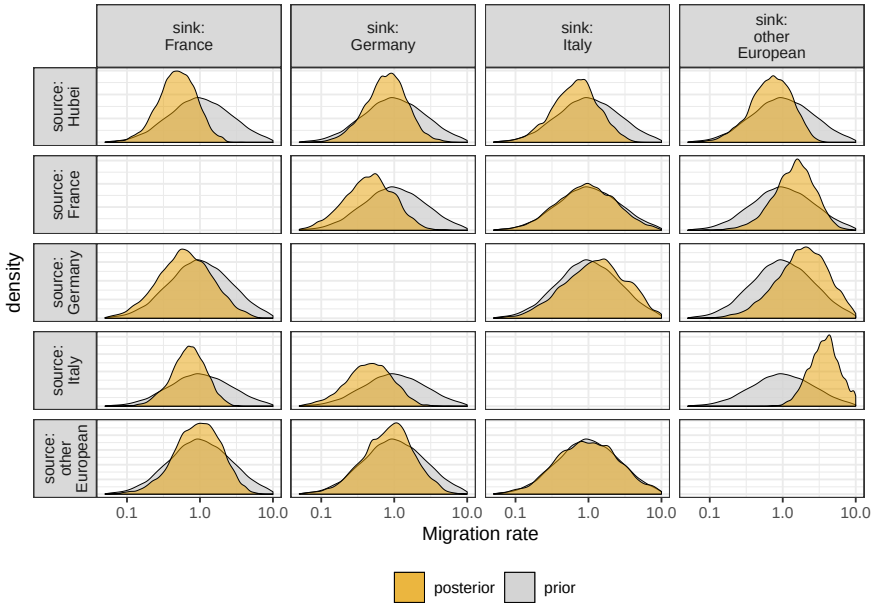


FIGURE 3.5: **Posterior distributions (yellow) for migration rates compared to the prior distribution (grey) in the analysis with down-sampling based on death data from 28 Mar. 2020.**

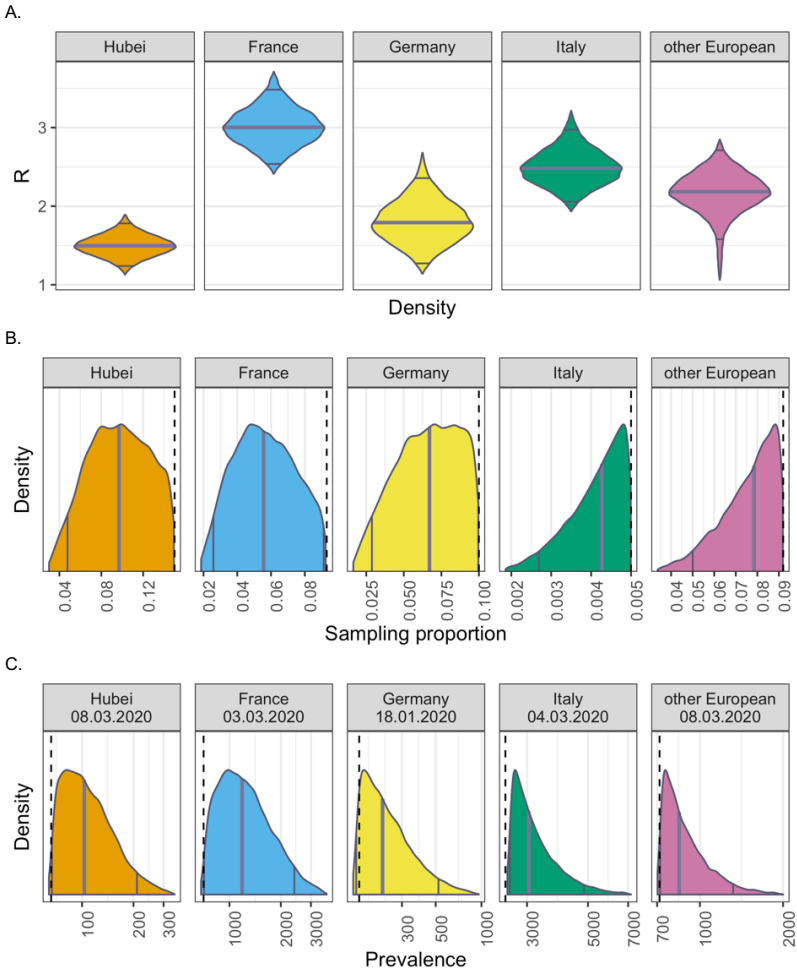


FIGURE 3.6: **Posterior distributions for inferred parameters in the analysis with down-sampling based on death data from 28 Mar. 2020.** (**A**) reproductive number, (**B**) sampling proportion, and (**C**) the number of no-longer infectious cases on the date of the last analyzed sequence (dates in facet titles) for each region. The dashed lines in (B) show the upper bound of the uniform sampling proportion prior for each region and in (C) the number of confirmed cases on the date of the last analyzed sample. The solid grey lines show the 95% highest posterior density interval and the median for each posterior.

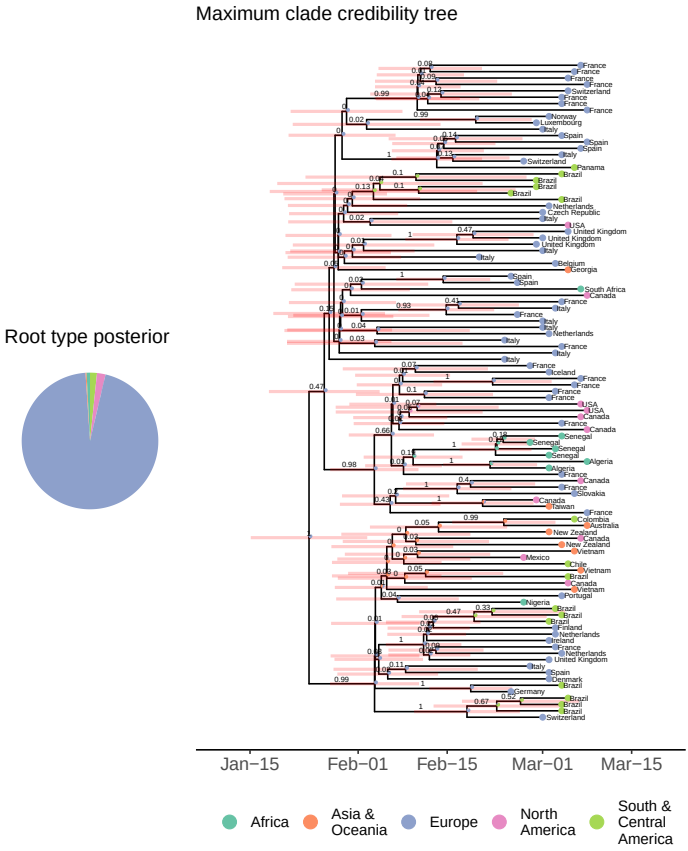


FIGURE 3.7: **Ancestral location inference results for the sensitivity analysis with sequences from the A2a clade from all global regions.** The pie chart on the left shows the posterior probabilities for the location of the most recent common ancestor of the analyzed sequences. The tree on the right is the maximum clade credibility tree. The values above the branches are posterior clade probabilities. The pie charts at nodes show the posterior probability of the node being in each deme. The deme for each tip is the region in which the sequence was collected (irrespective of travel history). Tips are annotated with the country in which the sequence was collected.

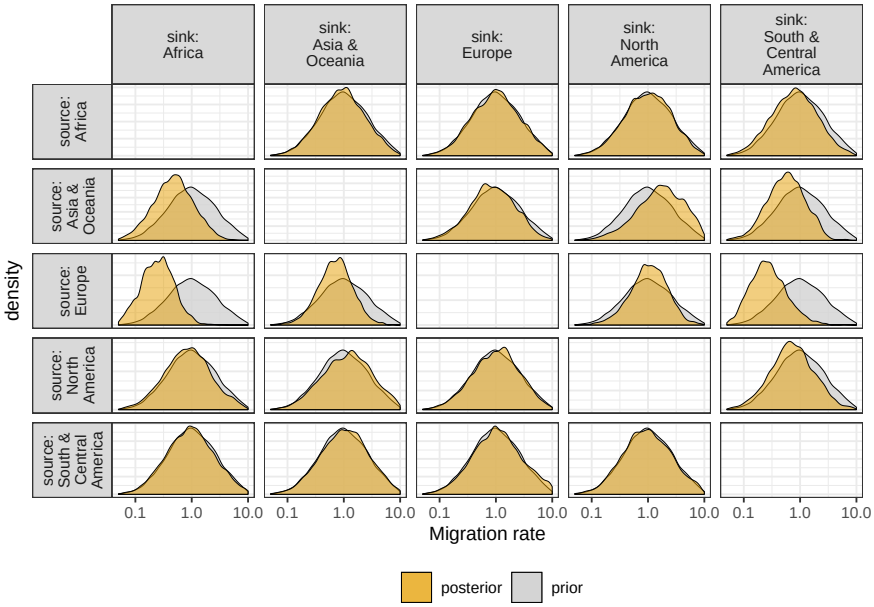


FIGURE 3.8: **Posterior distributions (yellow) for migration rates compared to the prior distribution (grey) for the analysis with sequences from the A2a clade from all global regions.**

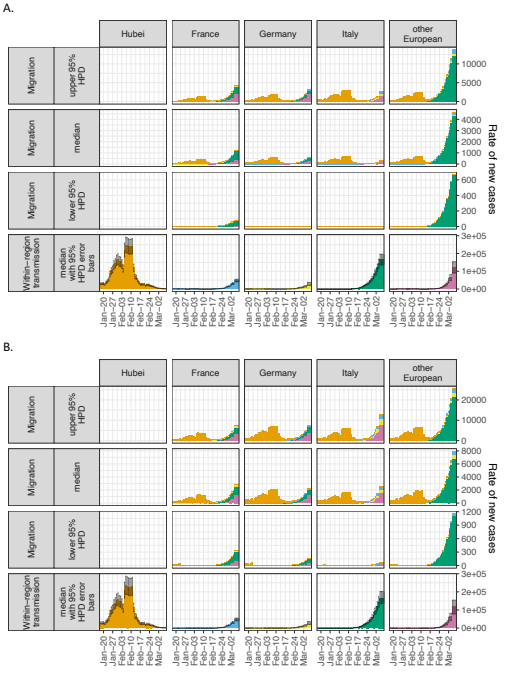


FIGURE 3.9: **Estimated rate of new cases from migration compared with the rate of new cases from within-region transmission (A) using the migration rate prior as in the main analysis and (B) using a lower migration rate prior: Lognormal(-2.29, 1.25).** Under the lower prior, median time until travel is 10 years and the 95% IQR is 1.3 days - 78 years. For each day, we multiplied the (smoothed) number of newly confirmed cases in each source region by the posterior sample of migration rates from source to sink. The upper and lower bounds of the 95% HPD interval and the median of these rates are show in the top three rows. We also multiplied the (smoothed) number of newly confirmed cases in each sink region by the posterior sample of transmission rates for the region. The median of these rates is shown in the bottom row, with error bars showing the upper and lower bounds of the 95% HPD interval. Dates are lagged 5 days to account for a 5-day delay between infection and migration or onward transmission and daily case counts were smoothed by taking a rolling 7-day average. Case data comes from (for Systems Science and at Johns Hopkins University, 2020). IQR = inter-quartile range; HPD = highest posterior density.

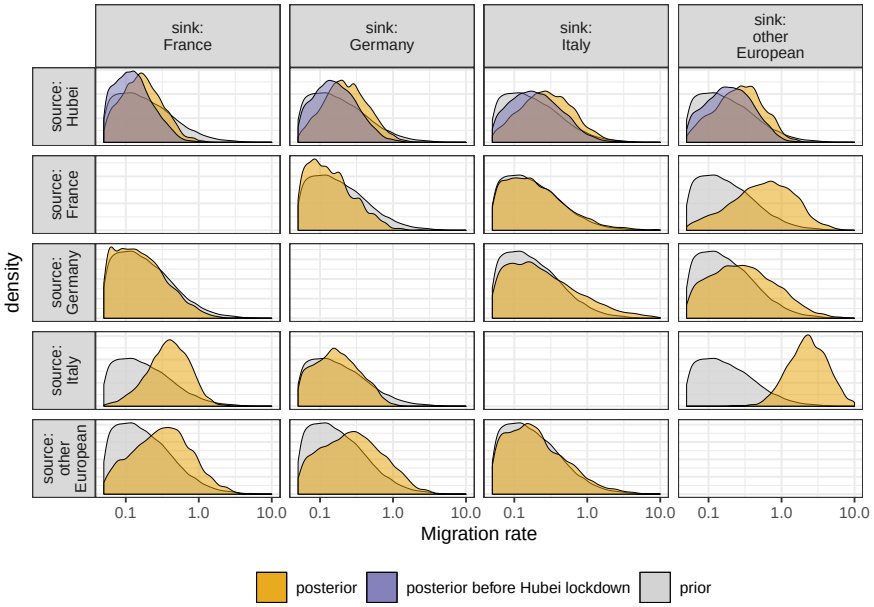


FIGURE 3.10: **Posterior distributions (yellow for rates out of Hubei before 23 Jan. 2020 and rates between non-Hubei demes, purple for rates out of Hubei after 23 Jan. 2020) compared to the prior distribution (grey) using the same sequences as in the main analysis with a lower migration rate prior: Lognormal(-2.29, 1.25).**

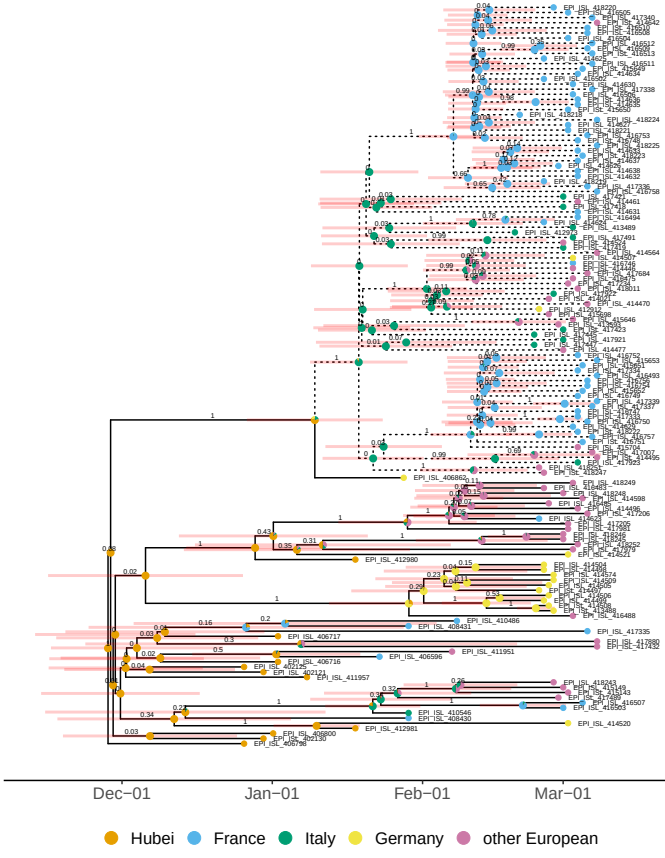


FIGURE 3.11: **Maximum clade credibility tree using the same sequences as in the main analysis with a lower migration rate prior: Lognormal(-2.29, 1.25).** The clade of A2a sequences analyzed is highlighted with dashed branches. The values above the branches are the posterior clade probabilities and the pale red bars show the 95% highest posterior density interval for node ages. The pie charts at internal nodes show the posterior probability that the ancestor was located in each deme (note that we assumed the root of the tree was in Hubei with probability 1). The deme for each tip is the deme in which the sequence was collected, irrespective of travel history. Tips are annotated with GISAID accession identifiers.

TABLE 3.4: Sequence information for the "other European" deme in the main analysis. Case and death data are from the Johns Hopkins Center for Systems Science and Engineering (https://github.com/CSSEGISandData/COVID-19). The number of sequences collected refers to the number of genome sequences available on GISAID (https://www.gisaid.org/). No. = number, Seqs. = sequences.

Country	No. seqs. included	No. confirmed deaths as of 8 Mar. 2020	No. confirmed cases as of 8 Mar. 2020	No. seqs. before down-sampling
Spain	18	17	673	26
Netherlands	4	3	265	107
United Kingdom	4	3	273	172
Switzerland	3	2	337	49
Belgium	1	0	200	46
Czech Republic	1	0	31	1
Denmark	1	0	35	9
Finland	1	0	23	13
Iceland	1	0	50	54
Ireland	1	0	19	9
Luxembourg	1	0	3	1
Norway	1	0	176	8
Poland	1	0	11	1
Portugal	1	0	30	19
Slovakia	1	0	3	4
Sweden	1	0	203	1

TABLE 3.5: Sequence information for the "other European" deme in the sensitivity analysis with down-sampling based on death data from 28 Mar. 2020. Case and death data are from the Johns Hopkins Center for Systems Science and Engineering (https://github.com/CSSEGISandData/COVID-19). The number of sequences collected refers to the number of genome sequences available on GISAID (https://www.gisaid.org/). No. = number, Seqs. = sequences.

Country	No. seqs. included	No. confirmed deaths as of 28 Mar. 2020	No. confirmed cases as of 28 Mar. 2020	No. seqs. before down-sampling
Spain	26	5982	73235	26
United Kingdom	11	1021	17312	172
Netherlands	7	640	9819	107
Belgium	5	353	9134	46
Switzerland	4	264	14076	49
Denmark	2	65	2366	9
Portugal	2	100	5170	19
Czech Republic	1	11	2631	1
Finland	1	9	1167	13
Iceland	1	2	963	54
Ireland	1	36	2415	9
Luxembourg	1	18	1831	1
Norway	1	23	4015	8
Poland	1	18	1638	1
Slovakia	1	0	292	4
Sweden	1	105	3447	1

TABLE 3.6: Sequence information for the sensitivity analysis with sequences from the A2a clade from all global regions. Location is the location of sample collection, as recorded in the Nextstrain metadata (previously available at https://github.com/nextstrain/ncov, as of publication the metadata is available for download from GISAID, https://www.gisaid.org/). Date is the date of sample collection, as given on GISAID. No. = number; Seq. = sequence.

Deme	No. seqs.	Locations	First seq. date	Last seq. date
Africa	8	Nigeria (1), South Africa (1), Senegal (4), Algeria (2)	27.02.2020	08.03.2020
Asia & Oceania	8	Taiwan (1), Vietnam (3), Georgia (1), New Zealand (2), Australia (1)	02.03.2020	08.03.2020
Europe	60	Belgium (1), Czech Republic (1), Denmark (1), Finland (1), France (20), Germany (1), Iceland (1), Ireland (1), Italy (12), Luxembourg (1), Netherlands (4), Norway (1), Portugal (1), Slovakia (1), Spain (6), Switzerland (3), United Kingdom (4)	20.02.2020	08.03.2020
North America	11	Mexico (1), USA (3), Canada (7)	27.02.2020	08.03.2020
South & Central America	16	Brazil (13), Chile (1), Panama (1), Colombia (1)	25.02.2020	06.03.2020

TABLE 3.7: Sequence information for the European deme in the sensitivity analysis with sequences from the A2a clade from all global regions. Confirmed case and death data are from the Johns Hopkins Center for Systems Science and Engineering (https://github.com/CSSEGISandData/COVID-19). The number of sequences collected refers to the number of genome sequences available on GISAID (https://www.gisaid.org/). No. = number, Seqs. = sequences.

Country	No. seqs. included	No. confirmed deaths as of 8 Mar. 2020	No. confirmed cases as of 8 Mar. 2020	No. A2a seqs. before down-sampling
Belgium	1	0	200	40
Czech Republic 1	0	31	1	
Denmark	1	0	35	9
Finland	1	0	23	8
France	20	19	1126	58
Germany	1	0	1040	2
Iceland	1	0	50	53
Ireland	1	0	19	8
Italy	12	366	7375	12
Luxembourg	1	0	3	1
Netherlands	4	3	265	64
Norway	1	0	176	1
Portugal	1	0	30	19
Slovakia	1	0	3	3
Spain	6	17	673	6
Switzerland	3	2	337	48
United Kingdom	4	3	273	101

TABLE 3.8: Median inferred number of introductions from each deme to each other deme amongst the analyzed cases in the analysis with sequences from the A2a clade from all global regions. Values in brackets are the upper and lower bound of the 95% highest posterior density interval.

Source / sink	Africa	Asia & Oceania	Europe	North America	South & Central America
Africa	-	0: [0, 1]	0: [0, 1]	0: [0, 1]	0: [0, 1]
Asia & Oceania	1: [0, 3]	-	0: [0, 2]	3: [0, 8]	3: [0, 6]
Europe	3: [0, 5]	5: [0, 9]	-	8: [2, 14]	8: [3, 13]
North America	0: [0, 2]	0: [0, 2]	0: [0, 1]	-	0: [0, 3]
South & Central America	0: [0, 1]	0: [0, 1]	0: [0, 1]	0: [0, 1]	-

4

ADVANCING GENOMIC EPIDEMIOLOGY BY ADDRESSING THE BIOINFORMATICS BOTTLENECK: CHALLENGES, DESIGN PRINCIPLES, AND A SWISS EXAMPLE

This chapter is published as:

Chaoran Chen*, Sarah Nadeau*, Ivan Topolsky, Niko Beerenwinkel, and Tanja Stadler. (2022). Advancing genomic epidemiology by addressing the bioinformatics bottleneck: Challenges, design principles, and a Swiss example. *Epidemics*, *39*, 100576. https://doi.org/10.1016/j.epidem.2022.100576

*equal contributions

Author contributions

For this chapter, I was fortunate to work closely with Chaoran Chen. Here, I describe my primary contributions. These were: populating the relational database with sequences and linked metadata, sequence submission to GISAID, troubleshooting changes or errors in sequences and metadata, and developing database views for billing and mutation reports. We shared the conceptualization and writing of the manuscript.

ABSTRACT

The SARS-CoV-2 pandemic led to a huge increase in global pathogen genome sequencing efforts, and the resulting data are becoming increasingly important to detect variants of concern, monitor outbreaks, and quantify transmission dynamics. However, this rapid up-scaling in data generation brought with it many IT infrastructure challenges. In this paper, we report about developing an improved system for genomic epidemiology. We (i) highlight key challenges that were exacerbated by the pandemic situation, (ii) provide data infrastructure design principles to address them, and (iii) give an implementation example developed by the Swiss SARS-CoV-2 Sequencing Consortium (S3C) in response to the COVID-19 pandemic. Finally, we discuss remaining challenges to data infrastructure for genomic epi-

demiology. Improving these infrastructures will help better detect, monitor, and respond to future public health threats.

4.1 INTRODUCTION

An increasingly important tool to help fight pathogenic diseases is genomic epidemiology. The analysis of pathogen genome sequences allows us to learn about pathogen evolution and epidemic or endemic transmission dynamics (Kraemer *et al.*, 2019; Grenfell *et al.*, 2004). However, the SARS-CoV-2 pandemic has highlighted a growing disparity between global sequencing data generation capacities and analysis capacities (Black *et al.*, 2020). As Hodcroft *et al.* (2021) underscores, we seem to be drowning in data rather than swimming in information.

Genome sequence data are becoming increasingly important for epidemic response, as highlighted during the SARS-CoV-2 pandemic. In December 2019, when an unknown respiratory disease was identified in Wuhan, China, the first whole genome sequence from the causal virus helped classify the new human pathogen SARS-CoV-2 (Wu *et al.*, 2020) and establish its likely origins (Andersen *et al.*, 2020). Then, comparison of mutational differences in genomes collected from different regions helped distinguish imported cases from community transmission (Worobey *et al.*, 2020). Next, genome surveillance efforts identified more transmissible variants of concern, e.g. the alpha variant (World Health Organization, 2021) in the UK in late 2020 (Volz *et al.*, 2021). Finally, phylogenetic and phylodynamic methods use genome sequence data to quantify epidemic dynamics, including the reproductive number, transmission routes, effects of public health measures, and the role of super-spreading (Nadeau *et al.*, 2021; du Plessis *et al.*, 2021; Miller *et al.*, 2020). Thus, pathogen genome sequence data is instrumental for disease detection, outbreak tracking, and quantifying transmission dynamics.

The wealth and geographic distribution of available genomic data underlying these and other analyses indicates many groups around the world have developed their own infrastructures for genomic epidemiology. So far, several large national initiatives have published descriptions of their technical infrastructures. In particular, Nicholls *et al.* (2021); Matthews *et al.* (2018); Egli *et al.* (2019) describe UK-, Canadian- and Swiss-specific infrastructures that enable linking of genome sequence data with associated metadata and integrate data from multiple regional contributors. Other examples are available as code bases, for instance that of the Spanish SARS-CoV-2 Sequencing Consortium (Spanish SARS-CoV-2 sequencing consortium, 2022).

Despite these successes, developing a data infrastructure for genome-based surveillance and genomic epidemiology remains a challenge (Black *et al.*, 2020; Bernasconi *et al.*, 2021). In the COVID-19 pandemic, bioinformatics capacity has proven to be a key bottleneck in pandemic response (Hodcroft *et al.*, 2021). This is particularly true in countries without a well-supported national initiative, or in the period before such an initiative is established. As the US-focused report Committee on Data Needs to Monitor Evolution of SARS-CoV-2 *et al.* (2020) highlights, a key priority for pandemic preparedness is to improve upon existing systems to integrate clinical and genomic data and better coordinate between different public health stakeholders. In this paper, we share lessons learned in the Swiss SARS-CoV-2 Sequencing Consortium (S3C) pertaining to three challenges that were particularly exacerbated by the COVID-19 pandemic: unstable data sources, rapid development of new tools, and the need for timely reporting. We outline design principles to address these challenges and describe our implementation of a relational database and containerized microservices as an example. Finally, we highlight remaining challenges in data management for genomic epidemiology.

The S3C began generating and analyzing SARS-CoV-2 genome sequences in March 2020. The Consortium started as a partnership between two academic groups, an associated academic sequencing facility, and a large Swiss medical diagnostics company (S3C, 2021). Since then, S3C has partnered with three core sequencing facilities in Switzerland to sequence over 44,000 samples from companies, hospitals, and research institutions. These data are made available on GISAID (Elbe and Buckland-Merrett, 2017) and the European Nucleotide Archive. To meet the demands of a growing genomic surveillance program in Switzerland, S3C benefited from early data infrastructure design choices that enabled rapid extension to new data sources, types, and users.

In the following sections we describe major implementation challenges for data infrastructure in light of the pandemic and outline design principles to address them. In particular, we discuss S3C's implementation of a relational database and microservices-based approach as an example fulfilling these design criteria using open source tools. Finally, we consider remaining challenges in data infrastructure for genomic epidemiology that must be met to improve future public health response to pathogenic diseases.

4.2 UNSTABLE DATA SOURCES

Emerging public health threats bring great uncertainties, including in data availability and formats. The basic data necessary for genomic surveillance are pathogen genome sequences and minimal patient metadata, e.g., sample collection date and location. Coupling these data and analyzing them in aggregate allows public health officials to track transmission and monitor key mutations. However, the format of these data may shift over the course of an outbreak, and new data may become available. For example, accommodating genomic restructuring by the pathogen itself (e.g., by insertion, deletion, recombination, or reassortment), annotating samples with the presence or absence of newly discovered key mutations, and newly available or re-formatted metadata all represent shifts in the basic data required for effective genomic surveillance. Furthermore, it might not be possible to define a fixed and sensible file format for data exchange in the early stages of outbreak response due to time pressure.

Recommendation: ensure clean data

Unreliable and shifting source data can quickly lead to messy data with, for example, missing values and different spellings of the same entity. Ideally, infrastructure developers will work with data submitters to develop a standardized data dictionary with clearly defined permitted values for each variable. However, it is also essential to strictly validate data upon import as a double-check. It should also be anticipated that changes and corrections to the data will be necessary over time. Therefore, data should be maintained in a non-redundant form so that changes to one attribute can be easily made without the danger of causing inconsistencies. Data relations should be tracked so that the effect of changes to one attribute on others are easy to identify. Data types should be strictly enforced so that changes to data formats are rapidly detected and mistakes are not incorporated. Finally, it should be easy to define custom data types and add attributes as new data is made available.

Example: relational database

Relational database management systems provide a good way to fulfill these design criteria. In a relational database management system, data are stored in a collection of tables, also known as the "relational format". Each

table is independent from the others, but they may be linked (related) via shared keys, i.e. information common to two or more tables. This allows us to formulate complex queries by joining different tables together.

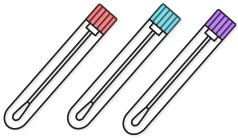
A relational database approach helps keep data clean in the face of unstable data sources. Each table's columns have fixed data types and it is possible to define custom types with a limited set of allowed values. Foreign keys, CHECK constraints and triggers allow definitions of arbitrarily complex validations. Invalid entries are rejected upon import so we know when corrections are necessary. This is especially important in the S3C, since we accept partially human-edited Excel files and non-documented output data from PCR machines as input. Non-redundancy between tables makes it easier to correct mistakes in these data when they arise. Finally, new and corrected data is simultaneously available to all database users.

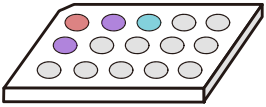
Several relational database management systems are available. The S3C uses PostgreSQL[1], which is freely available and open-source. In our implementation, we have three core database tables, one each for tests (samples), plates of RNA extracts, and SARS-CoV-2 genome sequences (Figure 4.1). The test table contains sample metadata from the originating laboratory, the plate table tracks where each plate was sent for sequencing and when, and the sequence table stores the assembled SARS-CoV-2 whole-genome sequence and associated quality control statistics. Finally, a mapping table links the respective keys from each table. These tables represent the core of our database, though we have added other tables through time to accommodate new data. For example, we store the identifiers assigned by public databases and additional sample metadata provided by the Swiss Federal Office for Public Health (FOPH).

4.3 NEW TOOLS

State-of-the art computational tools are also likely to change or are even being newly developed over the course of a public health response. This is exemplified in the COVID-19 pandemic by evolving nomenclature systems. Lineage assignment tools were frequently updated to keep up with nomenclature changes as new lineages arose. For example, the popular pangolin software for assigning SARS-CoV-2 genome sequences to global lineages has 75 releases since its development in April 2020 (O'Toole *et al.*, 2021).

1 https://www.postgresql.org/





test

Sample No.	City
●	Basel
●	Basel
●	Zurich

plate

Plate No.	Seq. lab
1	GFB
2	GFB
3	FGCZ
4	H2030

sequence

Seq. No.	Sequence
1	TCTTTG
2	ACTTCG
3	TCTTCG
4	NNTTCG

test_plate_sequence_mapping

Sample No.	Plate No.	Well	Seq. No.
●	1	A1	1
●	1	A2	2
●	1	A3	3
●	1	B1	4

FIGURE 4.1: **An illustration of how three key entities – tests, plates, and sequences – are stored in database tables and the mapping table that links the information from each.**

Recommendation: modular analysis workflows

Analysis workflows should be modular, rather than monolithic pipelines. It should be easy to update one component or swap it out for a different tool without having to re-run a full suite of analysis programs on the entire cohort. This modular structure allows individual components to be adapted or re-used for other pathogens or other projects. For use cases where software version tracking is especially important, workflow and software versions can be stored alongside the data in the database.

Example: containerized microservices

A microservices approach separates different tasks performed by different tools into loosely-coupled programs that operate autonomously, each performing a single, well-defined task. For the S3C, we implemented a growing set of microservices that import, export, and process data by adding or extracting data from the database (Figure 4.2). The microservices each

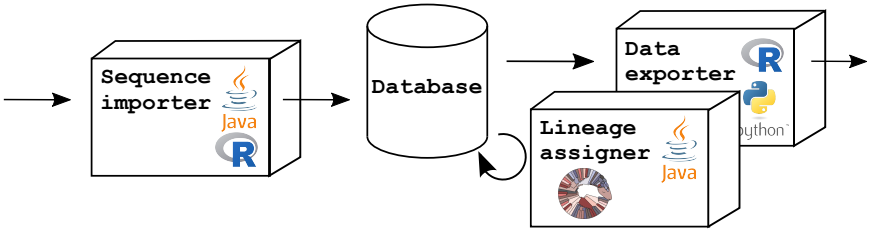


FIGURE 4.2: **Containerized microservices operate autonomously to add or extract data from the database.**

have their own code base, and, depending on the task, they are written in different languages.

We used a containerization technology to deploy these microservices. This packages software applications together with their dependencies into single units, called containers. For example, a Pango lineage assigner requires the pangolin tool (O'Toole *et al.*, 2021), a Nextclade importer needs Nextclade (Aksamentov *et al.*, 2021), and the metadata importer has to mount a network folder. The services can be written in different programming languages, perhaps even different versions of the same language to accommodate different dependencies.

Most services act only upon missing data. For example, we have a Nextclade importer service that runs the Nextclade program and imports resulting quality scores and mutations. This service queries the database every ten minutes and looks for entries in the sequence table where Nextclade quality scores were previously unpopulated. Other services avoid redundancy by maintaining a database table that stores a state, e.g. filenames which have already been processed and should not be re-imported. For example, our metadata importer service operates in this way.

The containerized microservices allow fast adoption of new or updated tools. Since they are packaged and deployed independently, they can be started or stopped without impacting other services. The containerization further serves to isolate each tool and remove dependency conflicts between tools. Finally, since services only act upon missing data or when a state is changed, we avoid redundant computation. Another complementary approach to achieving analysis modularity would be to use scientific workflow systems, such as Snakemake (Mölder *et al.*, 2021) or Nextflow (Di Tommaso *et al.*, 2017). These systems can be used together with containerization tech-

nologies and further simplify tracking of component software versions and workflow revisions used to generate output files.

4.4 TIMELY REPORTING

Timely reporting is crucial for an evidence-based public health response. Turn-around times for SARS-CoV-2 sequences to be made available on GISAID vary from a few days to a few weeks post-sampling, or more. Sample transport logistics, sequencing capacities, bioinformatics analysis, and report preparation all contribute to this turn-around time. Here, we focus on how to ensure rapid final reporting, as this is the aspect data managers have the most influence on.

Recommendation: Multiple levels of querying

A data management system needs to support rapid, ad-hoc querying in addition to generation of regular, stable reports. The prior is necessary for early outbreak detection and detection of new variants of concern, while the latter is essential for longer-term monitoring. Ideally, the system should be able to expose an application programming interface (API) for safe public data sharing.

Example: Database queries

Relational database systems support querying in several ways, fulfilling the above design criteria. One way to interact with data in a relational database is by directly using structured query language (SQL), which is a high-level and declarative language specifically designed for efficient querying. In SQL, the user describes (declares) what data should be added or retrieved, but not exactly how. The language then works behind-the-scenes to optimize the necessary computations and return the desired information (Figure 4.3). SQL is widely used by data analysts and does not require prior programming experience. Graphical user interfaces, for example DataGrip[2], allow users to manually add or modify data and submit queries. For those who are programmers, popular languages like R and python have packages like dplyr and pandas that enable reading data from a database directly into data frames.

2 https://www.jetbrains.com/datagrip/

```
select t.sample_no, t.city, p.plate, p.well
from
   test t
   join plate p on t.sample_no = p.sample_no
   join sequence s on p.plate = s.plate and p.well = s.well
   join mutation m on s.sequence_id = m.sequence_id
where
   m.mutation = 'S:N501Y';
```

Sample No.	City	Plate	Well
100	Basel	1	A1
101	Zurich	1	A2
101	Zurich	1	B1

FIGURE 4.3: **A SQL query that finds the samples with the S:N501Y mutation.**

For recurring queries, for instance for regular reporting, the database enables easy aggregation and reporting using "views". These are derived tables that aggregate data from existing tables according to a query. For reporting purposes, we created a number of views, for instance a billing view that contains the number of sequenced and submitted samples per week and a surveillance view that aggregates per-sample lineage assignment and mutation information for the Swiss FOPH. These views are automatically updated with the correction or addition of data. We also have a microservice that exports the mutation information view on a daily basis to a drop-point for the Swiss FOPH.

Finally, for monitoring purposes, a relational database can also serve as the back-end to dashboards or websites. We offer two public-facing websites to interact with sequencing and case data stored in our database. One is a dashboard focused on Swiss case data[3] and the other enables monitoring of global SARS-CoV-2 variants[4] (Chen *et al.*, 2021).

4.5 DISCUSSION

The COVID-19 pandemic has underscored both the utility of genomic epidemiology for public health response and remaining challenges in supporting related data infrastructure. Here we highlighted three challenges

3 https://ibz-shiny.ethz.ch/covidDashboard/?_inputs_&tab=%22ts%22
4 https://cov-spectrum.org

that were exacerbated by the rapidly changing pandemic situation: unstable data sources, rapid development of new tools, and the need for timely reporting. Then, we outlined general design principles to address these challenges. As an example, we describe the S3C's implementation of a relational database and containerized microservices.

These design choices directly enabled genome-based outbreak detection, monitoring, and public health response in the Swiss SARS-CoV-2 epidemic. Even before a new variant could be reliably called by lineage classification tools, we could quickly query Swiss data for mutations characterizing variants of concern. This enabled us to detect the first instances of the Beta, Gamma, and Delta variants in Switzerland. Our database also enabled us to quickly develop two public-facing websites for epidemic monitoring. Finally, we collaborate with the Swiss FOPH as members of the Swiss National COVID-19 Science Task Force[5] to link genome sequences to patient metadata. Lineage assignment and mutation data are passed back to the FOPH to support the health authorities in their pandemic response.

Many labs around the world have developed a data infrastructure for genomic epidemiology over the course of the COVID-19 pandemic. In fact, there are over 4000 unique submitting labs in the GISAID EpiCoV database as of January 2022. Unfortunately, a paucity of published examples makes it difficult to compare the strengths and weaknesses of various implementations in light of the challenges outlined by Black *et al.* (2020); Bernasconi *et al.* (2021) and highlighted here. The largest pathogen genome sequencing consortium in the world is that of COG-UK. Like S3C, they use a relational database. On top of it, they developed an API and a web interface for the collaborators to submit and retrieve data (Nicholls *et al.*, 2021). In comparison, we did not define a fixed metadata or sequence data format but adapted to the data provided by collaborators. Our aim was to reduce overhead for our collaborators. However, as data inputs stabilize, a future improvement would be to develop a more robust procedure for defining formats and updating data. An improved technical interface for data upload and correction by sequence submitters like that of COG-UK would also help.

There are also larger outstanding challenges to developing data infrastructures for genomic epidemiology. First, genome sequencing efforts are highly skewed towards high-income countries. In an interconnected world, local variants and fast epidemic spread are of global concern no matter where they arise. Expanding the technical and personnel resources for genome se-

5 https://sciencetaskforce.ch

quencing and data management in low and middle-income countries would enable a better, more coordinated public health response. Second, mistakes are common - from sequencing errors introducing spurious mutations, to sample contamination, to metadata errors. SARS-CoV-2 sequences and their metadata are regularly modified or deleted from public repositories. While some amount of mistakes are inevitable, better tools for tracking of changes to sequence data and their metadata would make correcting mistakes easier and promote reproducible science and transparency. Finally, we need robust infrastructures for safe linking of patient metadata with genome data. It can be a challenge to establish standardized, anonymized identifiers at the relevant scale for national sequencing projects, particularly in countries with decentralized health care services. Strong partnerships with government health ministries will help here, with metadata like vaccination and hospitalization status being provided to ensure actionable results for public health response.

In conclusion, generating pathogen genome sequence data and linking it to case-level metadata facilitates a rapid, evidence-based public health response to evolving infectious pathogens. Effective and timely generation of these data in rapidly changing situations relies on robust and agile data infrastructures, and improvements in the area should be a priority for pandemic preparedness.

CODE AVAILABILITY

Our code is openly available under the LGPL-license on GitHub at `https://github.com/cevo-public/harvester-database-and-automation`.

FUNDING

TS, SN and CC are supported by the Swiss National Science Foundation (grant number 31CA30_196267). NB and IT are supported by the SIB Swiss Institute of Bioinformatics.

BIBLIOGRAPHY

Aksamentov, I., Roemer, C., Hodcroft, E. B., and Neher, R. A. 2021. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software*, 6(67): 3773.

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. 2020. The proximal origin of sars-cov-2. *Nature Medicine 2020 26:4*, 26: 450–452.

Bernasconi, A., Canakoglu, A., Masseroli, M., Pinoli, P., and Ceri, S. 2021. A review on viral data sources and search systems for perspective mitigation of COVID-19. *Briefings in Bioinformatics*, 22(2): 664–675.

Black, A., MacCannell, D. R., Sibley, T. R., and Bedford, T. 2020. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nature Medicine*, 26(6): 832–841.

Chen, C., Nadeau, S., Yared, M., Voinov, P., Xie, N., Roemer, C., and Stadler, T. 2021. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics*, 38(6): 1735–1737.

Committee on Data Needs to Monitor Evolution of SARS-CoV-2, Board on Health Sciences Policy, Health and Medicine Division, Board on Life Sciences, Division on Earth and Life Studies, and National Academies of Sciences, Engineering, and Medicine 2020. *Genomic Epidemiology Data Infrastructure Needs for SARS-CoV-2: Modernizing Pandemic Response Strategies*. National Academies Press, Washington, D.C. Pages: 25879.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. 2017. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4): 316–319.

du Plessis, L., McCrone, J. T., Zarebski, A. E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T. R., *et al.* 2021. Establishment and lineage dynamics of the sars-cov-2 epidemic in the uk. *Science*, 371(6530): 708–712.

Egli, A., Blanc, D. S., Greub, G., Keller, P. M., Lazarevic, V., Lebrand, A., Leib, S., Neher, R. A., Perreten, V., Ramette, A., *et al.* 2019. Improving

the quality and workflow of bacterial genome sequencing and analysis: paving the way for a Switzerland-wide molecular epidemiological surveillance platform. *Swiss Medical Weekly*, (49). Publisher: EMH Media.

Elbe, S. and Buckland-Merrett, G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(1): 33–46.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., and Holmes, E. C. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656): 327–332.

Hodcroft, E. B., Maio, N. D., Lanfear, R., MacCannell, D. R., Minh, B. Q., Schmidt, H. A., Stamatakis, A., Goldman, N., and Dessimoz, C. 2021. Want to track pandemic variants faster? fix the bioinformatics bottleneck. *Nature 2021 591:7848*, 591: 30–33.

Kraemer, M. U. G., Cummings, D. A. T., Funk, S., Reiner, R. C., Faria, N. R., Pybus, O. G., and Cauchemez, S. 2019. Reconstruction and prediction of viral disease epidemics. *Epidemiology and Infection*, 147: e34.

Matthews, T. C., Bristow, F. R., Griffiths, E. J., Petkau, A., Adam, J., Dooley, D., Kruczkiewicz, P., Curatcha, J., Cabral, J., Fornika, D., *et al.* 2018. The integrated rapid infectious disease analysis (irida) platform. *bioRxiv*.

Miller, D., Martin, M. A., Harel, N., Tirosh, O., Kustin, T., Meir, M., Sorek, N., Gefen-halevi, S., Amit, S., Vorontsov, O., *et al.* 2020. Full genome viral sequences inform patterns of sars-cov-2 spread into and within israel. *Nature Communications*.

Mölder, F., Jablonski, K., Letcher, B., Hall, M., Tomkins-Tinch, C., Sochat, V., Forster, J., Lee, S., Twardziok, S., Kanitz, A., *et al.* 2021. Sustainable data analysis with snakemake [version 2; peer review: 2 approved]. *F1000Research*, 10(33).

Nadeau, S. A., Vaughan, T. G., Scire, J., Huisman, J. S., and Stadler, T. 2021. The origin and early spread of sars-cov-2 in europe. *Proceedings of the National Academy of Sciences of the United States of America*, 118.

Nicholls, S. M., Poplawski, R., Bull, M. J., Underwood, A., Chapman, M., Abu-Dahab, K., Taylor, B., Colquhoun, R. M., Rowe, W. P. M., Jackson, B., *et al.* 2021. CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. *Genome Biology*, 22(1): 196.

O'Toole, , Scher, E., Underwood, A., Jackson, B., Hill, V., McCrone, J. T., Colquhoun, R., Ruis, C., Abu-Dahab, K., Taylor, B., *et al.* 2021. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution*, 7(2). veab064.

S3C 2021. Swiss SARS-CoV-2 Sequencing Consortium (S3C). https://bsse.ethz.ch/cevo/research/sars-cov-2/swiss-sars-cov-2-sequencing-consortium.html.

Spanish SARS-CoV-2 sequencing consortium 2022. FISABIO-NGS / SARS-CoV2-Mapping. https://gitlab.com/fisabio-ngs/sars-cov2-mapping.

Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., Áine O'Toole, *et al.* 2021. Assessing transmissibility of sars-cov-2 lineage b.1.1.7 in england. *Nature 2021 593:7858*, 593: 266–269.

World Health Organization 2021. Tracking SARS-CoV-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/.

Worobey, M., Pekar, J., Larsen, B. B., Nelson, M. I., Hill, V., Joy, J. B., Rambaut, A., Suchard, M. A., Wertheim, J. O., and Lemey, P. 2020. The emergence of sars-cov-2 in europe and north america. *Science*, 370(6516): 564–570.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y., *et al.* 2020. A new coronavirus associated with human respiratory disease in china. *Nature*, 579: 265–269.

5

SWISS PUBLIC HEALTH MEASURES ASSOCIATED WITH REDUCED SARS-COV-2 TRANSMISSION USING GENOME DATA

This chapter is submitted as:

Sarah A. Nadeau, Timothy G. Vaughan, Christiane Beckmann, Ivan Topolsky, Chaoran Chen, Emma Hodcroft, Tobias Schär, Ina Nissen, Natascha Santacroce, Elodie Burcklen, Pedro Ferreira, Kim Philipp Jablonski, Susana Posada-Céspedes, Vincenzo Capece, Sophie Seidel, Noemi Santamaria de Souza, Julia M. Martinez-Gomez, Phil Cheng, Philipp P. Bosshard, Mitchell P. Levesque, Verena Kufner, Stefan Schmutz, Maryam Zaheri, Michael Huber, Alexandra Trkola, Samuel Cordey, Florian Laubscher, Ana Rita Gonçalves, Sébastien Aeby, Trestan Pillonel, Damien Jacot, Claire Bertelli, Gilbert Greub, Karoline Leuzinger, Madlen Stange, Alfredo Mari, Tim Roloff, Helena Seth-Smith, Hans H. Hirsch, Adrian Egli, Maurice Redondo, Olivier Kobel, Christoph Noppen, Louis du Plessis, Niko Beerenwinkel, Richard A. Neher, Christian Beisel, and Tanja Stadler. Swiss public health measures associated with reduced SARS-CoV-2 transmission using genome data.

ABSTRACT

Genome sequences from evolving infectious pathogens allow quantification of case introductions and local transmission dynamics. We sequenced 11,357 SARS-CoV-2 genomes from Switzerland in 2020 - the 6th largest effort globally. Using a representative subset of these data, we estimated viral introductions to Switzerland and their persistence over the course of 2020. We contrast these estimates with simple null models representing the absence of certain public health measures. We show that Switzerland's border closures de-coupled case introductions from incidence in neighboring countries. Under a simple model, we estimate a 94% reduction in introductions during Switzerland's strictest border closures. Furthermore, the Swiss 2020 partial lockdown roughly halved the time for sampled introductions to die out. Finally, we quantified local transmission dynamics once introductions into Switzerland occurred, using a novel phylodynamic model. We find that transmission slowed 35 - 63% upon outbreak detection

in summer 2020, but not in fall. This finding may indicate successful contact tracing over summer before overburdening in fall. The study highlights the added value of genome sequencing data for understanding transmission dynamics.

5.1 INTRODUCTION

SARS-CoV-2 genomes were collected at an unprecedented scale in 2020 (Munnink *et al.*, 2021) and have been extensively used to characterize transmission dynamics, in particular because genetic data contains information on the epidemiological relationships between cases. These genomic data enable the reconstruction of introductions and downstream transmission chains in the absence of contact tracing data (Kraemer *et al.*, 2019). Where contact tracing data is available, this approach has been verified and has additionally helped with linking unassigned individuals to known transmission chains (Rockett *et al.*, 2020; Douglas *et al.*, 2021).

Several methods have been successfully used to reconstruct transmission dynamics at the onset of the COVID-19 pandemic using genetic data. Phylogenetic approaches reconstruct pathogen phylogenies and calculate relevant statistics from them without fitting any further explicit models. For example, phylogenetic reconstructions were used to show that reduced lineage size and diversity coincided with national lockdowns during the early Irish and English epidemics (Mallon *et al.*, 2020; du Plessis *et al.*, 2021). In Switzerland, Stange *et al.* (2021) linked regional super-spreading events to a dominant lineage in the city of Basel using a phylogenetic reconstruction. Phylodynamic studies, on the other hand, assume the phylogeny arises from an underlying model of transmission between hosts, possibly including additional complexities like migration of hosts between regions. This assumption enables estimation of population-level transmission dynamics from pathogen genome data. For example, Miller *et al.* (2020); Geoghegan *et al.* (2020); Müller *et al.* (2021) showed that public health measures reduced SARS-CoV-2 transmission rates in Israel, New Zealand, and Washington State, USA.

New models and careful considerations of potential biases are required to quantify the effects of different public health measures in different regions. Here, we developed an analysis framework to quantify the association between the implementation and lifting of major public health interventions, such as border closures, lockdown measures, and contact tracing - three front-line tools in the fight against COVID-19 in 2020 - on transmission

dynamics. Our framework uses a two-step process that carefully combines phylogenetic and phylodynamic methods to address potential sampling biases and phylogenetic uncertainty. Within the Swiss SARS-CoV-2 Sequencing Consortium (S3C; (S3C, 2021)) we sequenced 11,357 Swiss SARS-CoV-2 genomes until 1 December 2020. After combining these genomes with additional data available on GISAID (Bogner *et al.*, 2006) and down-sampling to control for biases in sampling efforts over time and among geographic regions, we were left with 5,520 Swiss SARS-CoV-2 genomes, representing up to 5% of weekly confirmed cases in Switzerland. We use these genomes to characterize transmission dynamics in Switzerland until the emergence and widespread dissemination of more transmissible variants of concern, starting in December 2020 (World Health Organization, 2020). Our framework allows us to identify a clear effect of border closures and the spring 2020 partial lockdown on the rate of new introductions to Switzerland and their persistence. Furthermore, we were able to quantify the degree to which local transmission slowed upon outbreak detection. We find that this effect was strongest during summer 2020, when cases were low and contact tracing efforts likely more effective. To demonstrate the broader applicability of our analysis framework, we additionally analyzed data from New Zealand, where quarantine measures were stricter and local transmission was extremely limited throughout 2020. In New Zealand, we quantify a stronger transmission slowdown after outbreak detection, consistent with contact tracing there being highly effective.

5.2 RESULTS

Introductions and their persistence shed light on the effects of border closure and lockdown

First, we identified putatively independent introductions of SARS-CoV-2 into Switzerland and estimated their persistence. To do this, we selected SARS-CoV-2 genome sequences corresponding to up to 5% of confirmed cases each week, stratified to be geographically representative when possible (Figure 5.4). We divided these sequences by Pango lineage, as these lineages should represent monophyletic clades in the global SARS-CoV-2 phylogeny (Rambaut *et al.*, 2020). Because of the hierarchical nature of Pango lineages, we aggregated lineages dominated by Swiss sequences into their respective parent lineages, allowing us to assume each analyzed lineage originated outside Switzerland (Table 5.1). To provide global context,

we additionally selected the most genetically similar sequences from abroad for each lineage. We then constructed an approximate maximum-likelihood phylogeny for each such lineage of Swiss and genetically similar foreign sequences. We subsequently identified putatively independent introductions into Switzerland from these phylogenies, while allowing for a fixed number of export events. Importantly, we identified two plausible sets of introductions into Switzerland resulting from two different assumptions about the ordering of transmission events at polytomies with both Swiss and non-Swiss descendants. The set of "few" introductions was generated assuming the majority of polytomic lineages are from within-Switzerland transmission, whereas the set of "many" introductions was generated assuming the majority are new introductions. Sensitivity analyses show these two polytomy assumptions capture most of the uncertainty in the size and number of introductions amongst analyzed sequences (Supplementary text S1; Figure 5.5). Using additional data on which cases were from managed isolation and quarantine facilities in New Zealand versus identified in the community, we show that, as expected, the "many introductions" polytomy assumption is more realistic when the probability of infection abroad is high compared to the probability of locally acquired infection (Supplementary text S2). Throughout, we report uncertainty based on the difference between the few and many introductions sets.

We estimate that the analyzed sequences originate from between 557 (few) and 2284 (many) introductions into Switzerland. These introductions are roughly power law-distributed in size (Figure 5.6), with the 10 largest introductions accounting for 16 to 30% of sampled genomes. Introductions that yielded more than one sampled Swiss case in our dataset tended to be geographically constrained. Between 64% (few) and 92% (many) of sampled transmission chains (introductions with >1 sample) were sampled in only 1-2 of the 26 Swiss cantons (Figure 5.7A). As expected, larger introductions were sampled in more cantons (Figure 5.7B; Pearson's R between introduction size and number of cantons is 0.86 for many introductions, 0.75 for few introductions). From a down-sampling analysis, we observe that if we were to include more sequences, we would identify more introductions (Figure 5.5C). Therefore, the analyzed genomes do not represent all introductions into Switzerland but, given the samples are spatio-temporally representative, are a representative subset of introductions. Due to incomplete sampling, each sampled introduction contains only a subset of all cases in the full transmission chain.

Since we sampled sequences proportionally to confirmed cases through time (Figure 5.4A; R^2 between number of confirmed cases and number of analyzed Swiss sequences each week 0.72), we can assume that trends through time in the number and persistence of introductions are representative of the underlying dynamics. Figure 1A shows the number of newly sampled introductions identified each week from our dataset, which peaked the week of 15 March under both polytomy assumptions. Switzerland closed its external borders to Italy 13 March 2020 and with the rest of the world shortly thereafter (Bradley, 2020). To disentangle the effect of the border closures versus local control measures, we back-calculated the expected number of total (both sampled and unsampled) introductions each week under a birth-death skyline model (Stadler *et al.*, 2013). This calculation corrects for the probability that an introduction went extinct or remained unsampled each week until the end of the sampling period, given estimates of the sampling proportion and the time-varying effective reproductive number Re in Switzerland. Then, we develop a simple null model that assumes that prior to 13 March 2020, total introductions are a linear function of case counts in Switzerland's largest neighboring countries (Italy, France, Germany, and Austria). Here we are assuming incidence in travelers to Switzerland follows incidence in the general community in these countries. Figure 1C shows this model fit to total introduction estimates generated based on each polytomy assumption and model projections (dashed lines) from 13 March through the partial re-opening of Switzerland's European borders on 15 June 2020 (Bradley, 2020). In the following, we report uncertainty based on the 95% HPD upper and lower bound estimates for Re used to estimate total introductions. Uncertainty in travel patterns is discussed later. Compared to the null model, we estimate a reduction of 7,000 (few introductions; uncertainty 4,500 - 11,000) or 79,000 case introductions (many introductions; uncertainty 41,000 - 130,000). Despite the high uncertainty in the absolute number of introductions averted depending on the polytomy assumption and the precise value of Re in Switzerland, we estimate a consistent percentage-wise reduction of 94.1% (few introductions; uncertainty 85.9 - 97.8%) or 94.2% (many introductions; uncertainty 86.2 - 97.9%). We note that total European case counts peaked later than in Switzerland's neighboring countries while our analysis only considers neighboring countries. Thus, the period of high import pressure may have extended longer than we assume, depending on where most introductions were coming from (Figure 5.8). However, our focus on neighboring countries is supported by travel statistics. For instance, neighboring countries comprise 99% of

cross-border working permits granted by Switzerland for the first quarter of 2020 (approximately 330,000 individuals). These countries also account for 36% of registered arrivals at Swiss hotels in January and February 2020 (approximately 450,000 individuals) (FSO, 2020). Thus, we assume introduction dynamics are largely driven by these neighboring countries. However, our estimates of the precise reduction in imported cases depend strongly on this assumption.

New introductions cannot sustain an epidemic unless they persist in the local population. Our analysis suggests several introductions were quite persistent in Switzerland, including one that may have persisted across our entire sampling period (Figure 5.9). On average, introductions persisted 5 days (many introductions; standard deviation 16 days) to 34 days (few introductions; standard deviation 53 days) from the oldest to the most-recent sample of each introduced lineage in our dataset. Lineage persistence until last sampling was lower during the partial lockdown (17 March - 27 April; Figures 1B and D) compared to summer 2020. While only 0.5 - 8% of introductions in April were sampled for at least 60 days, this fraction increased to 12 - 52% in September, just before a large fall wave in Switzerland. We also developed a simple null model to assess whether the spring 2020 lockdown measures and associated behavioral changes affected the persistence of introduced lineages. Here, our null model is that persistence, measured as the time until introductions circulating each day are last sampled, does not change through time. We assume this delay distribution always equals the median persistence calculated over the spring period (until 15 June). Figure 1D contrasts this null model assumption with empirical persistence calculated from each day under each polytomy assumption. The distribution does indeed vary through time, deviating from the null model. We estimate median persistence of introductions at the start of the lockdown is less than or around the median calculated over the whole spring and rises to above this null model threshold in the post-lockdown period. Quantitatively, introductions persisted roughly twice as long until last being sampled at a post-lockdown peak around 10 June compared to at the lockdown start (Figure 1D). We note that under the few introductions assumption, persistence estimates are upper-bounded by the end of our sampling period, so the increase in persistence may also be an underestimate (Figure 1D).

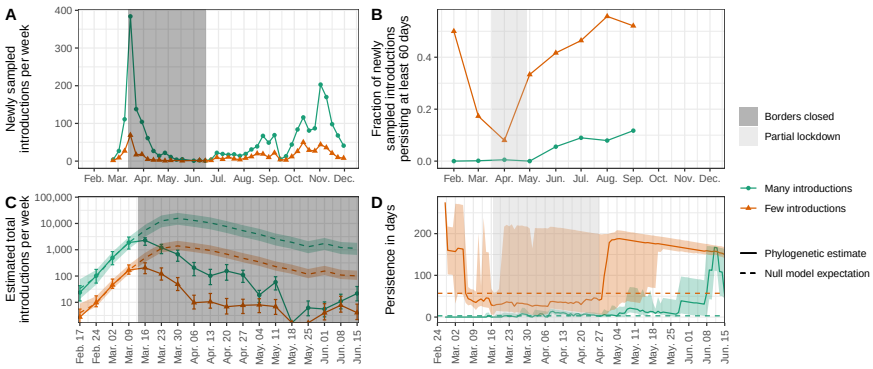


FIGURE 5.1: **Genome-based estimates of SARS-CoV-2 introductions into Switzerland and their persistence.** (**A**) shows the number of newly sampled introductions identified each week and (**B**) shows the fraction of newly sampled introductions each month that persist for at least 60 days from the oldest to the most-recent sample. This persistence measure is only defined until September because we only considers sequences obtained until 1 December 2020. Orange and green correspond to estimates generated under the few and many introductions polytomy assumptions, respectively. (**C**) and (**D**) focus on dynamics around the Swiss border closure and partial lockdown periods, which are highlighted with shaded rectangles. (C) shows estimated total introductions (solid lines) compared to a null model (dashed lines) where total introductions are a linear function of case numbers in Switzerland's neighboring countries. The null model is fit to the points prior to the border closure, values after that are projections. Uncertainty bounds for total introductions (error bars) and null model predictions (colored shaded areas) are based on the 95% upper and lower HPD bounds for Re when estimating total introductions. Uncertainty in travel patterns is not shown, see Figure 5.8. (D) shows the distribution of ongoing persistence for introductions circulating each day (solid lines), compared to a null model (dashed lines) where persistence is constant through time (equal to the median calculated until 15 June). Solid lines are median time to last sampling amongst introductions newly sampled or still ongoing each day. The shaded areas show the interquartile range of this persistence distribution.

Phylodynamic model indicates summer introductions slowed after detection

Next, we investigated local transmission dynamics once SARS-CoV-2 lineages were introduced to Switzerland in more detail. To do this, we quantified time-varying transmission dynamics in Switzerland in a Bayesian phylodynamic framework. As a base model, we used the birth-death model with serial sampling originally described in (Stadler, 2010). We modified the model to condition on the previously identified few or many introductions sets, i.e., sequences from each introduction have an independent origin. In a nutshell, the model assumes that once lineages are introduced, they are (i) transmitted between hosts, according to a time-varying transmission rate which is the same across all introductions; (ii) die out upon recovery/death of the host, according to a constant becoming-uninfectious rate; and (iii) yield genome samples with a time-varying sampling proportion which is the same across all introductions. We assume individuals who test positive adhere to self-isolation regulations, so sampling corresponds to a death event for the viral lineage. Under this parameterization, Re is a function of the transmission rate, becoming-uninfectious rate, and sampling proportion.

We developed a novel extension to this methodology by adding a transmission rate "damping" factor, as shown in Figure 2. The transmission rate is allowed to decrease by a multiplicative damping factor two days after an introduction is first sampled. We use a spike-and-slab prior on this factor to include the possibility of no transmission slowdown. We allow this damping factor to vary between spring, summer, and fall 2020 - periods characterized by very different case numbers and testing regimes in Switzerland (Figure 3A; (of Public Health, 2020)). Using this model, we aim to test whether contact tracing efforts in Switzerland slowed transmission once introductions were detected. We reason that test-trace-isolate can only slow transmission from shortly after the first case of an introduction tests positive but not beforehand, as beforehand the introduction was circulating cryptically. The two-day delay aims to account for the time between an individual giving a sample (i.e., being swabbed) and having their contacts notified. Specifically, this delay consists of the time to RT-PCR results, which was generally below 24 hours in Swiss diagnostic laboratories (Marquis *et al.*, 2021), plus the time for contact tracers to reach contacts or an individual to receive and input their positive test code to the SwissCovid contact tracing app. We fit the phylodynamic model in several configurations: conditioning on either the many or few introductions set, using a bounded or an unbounded sam-

pling proportion prior (see Supplementary text S3), and with or without a transmission damping factor.

Across these model configurations, we recover roughly the same trends in Re as estimates based on confirmed case numbers beginning with the first analyzed sequence from 27 February (Figure 5.10B). Compared to confirmed case-based estimates, we estimate a sharper decline in Re coinciding with lockdown measures. Depending on the polytomy assumption, we estimate Re was 2.2 (many introductions; 95% HPD 1.5 - 2.9) or 3.5 (few introductions; 95% HPD 2.9 - 4.2) in the week of 9 March. Re fell to 0.3 (many introductions; 95% HPD 0.2 - 0.4) or 0.4 (few introductions; 95% HPD 0.2 - 0.6) in the week of 16 March 2020 (posterior median estimates with no damping factor and an unbounded sampling proportion prior). With a bounded sampling proportion, peak Re estimates are slightly higher (Figure 5.10). Results in fall 2020 are highly dependent on the sampling proportion prior, where Re estimates better match confirmed case-based estimates when the sampling proportion is treated as a fitting parameter (i.e., with an unbounded prior, resulting in unrealistic estimates of the sampling proportion; see Figure 5.10A).

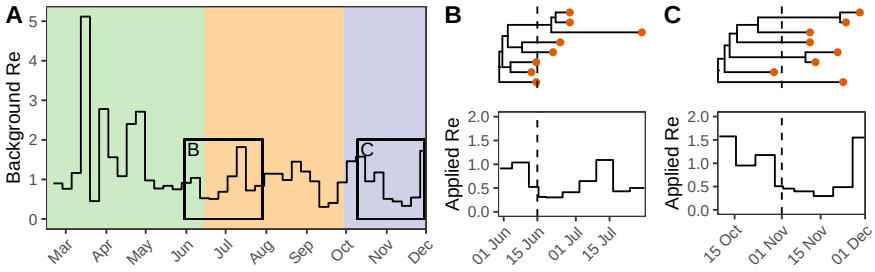


FIGURE 5.2: **Illustration of how transmission rate damping is modeled.** (**A**) shows a background Swiss-wide time-varying effective reproductive number Re before any damping. Here we show the median posterior result from the model applied to the many introductions data as an illustration. In each of the colored areas (green = spring, orange = summer, and purple = fall), a different damping factor is proposed. The black boxes in (A) highlight the spread of two real introductions (**B**) and (**C**) generated under the many introductions polytomy assumption. The genome data sampled from these introductions are shown as red dots in (B) and (C). The appropriate damping factor on Re is applied to each introduction 2 days after the first genome sample (dashed lines). We used 0.6 for the summer damping factor and 0.9 for fall for this illustration. The likelihood of the genome sequence data at the tips of the phylogenies is calculated given the "applied" Re specific to each introduction (B and C, bottom).

From the model fit with a damping factor, we estimate a 35% (few introductions; 95% HPD 29 - 41) - 63% (many introductions; 95% HPD 56 - 70) slowdown in transmission after introductions are first sampled in summer 2020 (posterior median estimates with an unbounded sampling proportion prior). In comparison, there is little support for a slowdown effect upon the first sampling during fall 2020 (Figure 3). These results are qualitatively robust to bounding the sampling proportion prior (Figure 5.11). In contrast, damping factor estimates in spring 2020 are inconsistent, depending on the polytomy assumption. Low genomic diversity in SARS-CoV-2 during this period causes high phylogenetic uncertainty ((Morel *et al.*, 2021); see also the differences in several selected introductions in Figure 5.12). This results in quite different estimates for the damping factor depending on the polytomy assumption used. In summary, we report a summer 2020 "slowdown" dynamic in SARS-CoV-2 transmission in Switzerland, where transmission slows after the first genome in a new introduction is sampled. This slowdown is not observed in fall 2020.

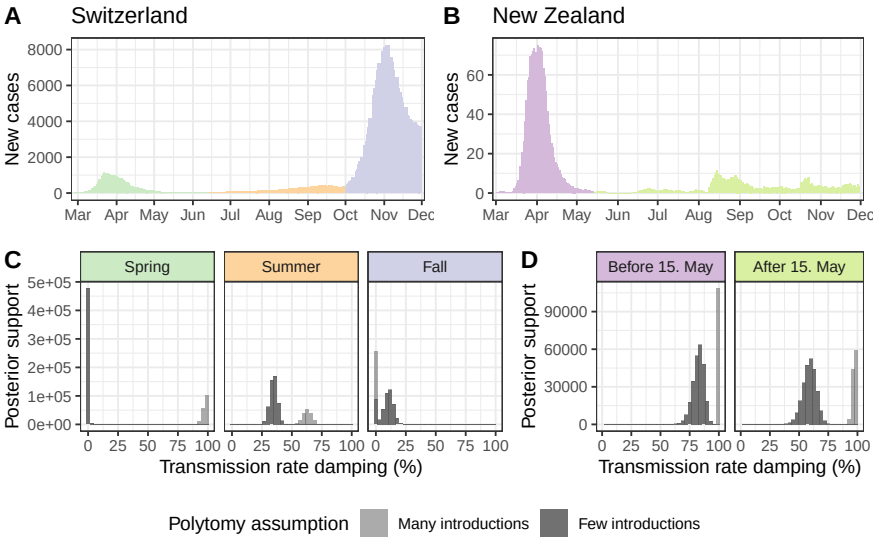


FIGURE 5.3: **Phylodynamic estimates for the transmission damping factor in Switzerland and New Zealand compared to case numbers.** Case numbers in (**A**) Switzerland and (**B**) New Zealand during 2020 are shown as a 7-day rolling average of daily new confirmed cases (ECDC, 2020). (**C**) and (**D**) show estimates for if and how much transmission rates were dampened after introductions were sampled during different time periods in (C) Switzerland and (D) New Zealand. The inference was done twice, once conditioning on introductions identified assuming many introductions (light gray) and once assuming few introductions (dark gray). Thus, the difference between estimates in light and dark gray are due to phylogenetic uncertainty. Results shown are from the model with an unbounded sampling proportion prior, results with a bounded sampling proportion prior are similar (Figure 5.11).

New Zealand data shows slowdown effect is not Switzerland-specific

While Switzerland is centrally located in Europe and well-connected to other countries, especially those in the (normally) barrier-free Schengen zone, New Zealand is a relatively isolated island nation. Additionally, New Zealand aimed to eradicate SARS-CoV-2 throughout 2020 using strong measures, such as keeping its borders closed and enforcing strict quarantine-on-arrival (New Zealand Government, 2020), while Switzerland partially

reopened its borders to Europe on 15 June. We applied the same analysis framework for introduction estimation and phylodynamic inference to SARS-CoV-2 sequences from New Zealand as a comparison to our Switzerland-specific results. For the phylodynamic analysis, we estimated independent damping factors before and after an epidemic breakpoint in mid-May 2020 when local transmission was briefly eradicated (Geoghegan *et al.*, 2020, 2021). Case numbers were subsequently held at lower levels through December 2020 (Figure 3B, (Geoghegan *et al.*, 2021)).

From the model fit with a damping factor, we estimate transmission damping in New Zealand before and after 15 May to be comparable with or stronger than in Switzerland during summer and fall 2020 (Figure 3D), regardless of the polytomy assumption used. Thus, the existence of a transmission damping effect is not specific to Switzerland. From the model fit without a damping factor, our estimates for the sampling proportion and Re are inconsistent. In particular, the sampling proportion is estimated to be unrealistically high when conditioning on the many introductions data set. However, including the damping factor in the model reconciles estimates based on each polytomy assumption, yielding more realistic estimates for the sampling proportion and pre-damping Re (Figure 5.13).

5.3 DISCUSSION

We quantify the change in cross-border and local transmission dynamics with the introduction or lifting of major public health measures in Switzerland based on genome sequence data. First, we quantify the reduction in case introductions during the period of Switzerland's strictest border closures. Travel from Italy was tightly restricted beginning 13 March and with the rest of the world beginning 16 March 2020. These measures were partially lifted on 15 June, when Switzerland re-opened to European countries in the Schengen zone (Bradley, 2020). We used phylogenetic estimates for the number and timing of viral introductions into Switzerland to show that newly sampled introductions peaked during the week of 15 March, coinciding with the implementation of border closures. Due to many identical or near-identical SARS-CoV-2 lineages circulating widely in Europe during spring 2020, the total number of introductions to Switzerland is highly uncertain. We considered two extreme cases, encompassing most of the phylogenetic uncertainty in the size and number of introductions. We additionally corrected these estimates based on the time-varying probability that an introduction went unsampled. After disentangling the effect of

border closures and local control measures in this way, we show that border closures de-coupled introduction dynamics from case counts in neighboring countries. Compared to a simple null model assuming that the incidence in travelers corresponds to the incidence in Switzerland's neighboring countries, we quantify a 94% reduction in case imports from 13 March - 15 June. While the de-coupling of case introductions and incidence in neighboring countries is clear, our estimates for precisely how many and what fraction of introductions were averted are subject to several strong assumptions, namely that incidence in travelers is the same as the average in the different source populations, and that the majority of imported cases would have come from Switzerland's neighboring countries. Finally, we note that the fraction of polytomic lineages that were independent introductions likely decreased throughout spring 2020 as local incidence rose, travel declined, and the probability of locally acquired infection rose (Supplemental text S2). We expect the truth to lie somewhere between the estimates generated under our two polytomy assumptions.

Second, we quantify the reduction in local transmission during Switzerland's partial lockdown in spring 2020 compared to the pre- and post-lockdown time period. A suite of lockdown measures, including closure of schools, non-essential shops, restaurants, and entertainment and leisure establishments was introduced on 17 March 2020. Many non-essential shops re-opened on 27 April, before schools and most other shops reopened on 11 May (The Swiss Federal Council, 2020). We estimate that sampled introductions circulating on 17 March persisted only about half as long until last sampling as in mid-June. We also estimate that only 0.5 - 8% of newly sampled introductions in April persisted more than 60 days until last being sampled, compared to 12 - 52% in September. These findings agree with previous findings (Ladoy *et al.*, 2021), which demonstrated a reduction in the number of transmission clusters and the risk of transmission within clusters in the Canton of Vaud, Switzerland after the implementation of lockdown measures. Finally, we obtained genome-based estimates for the time-varying effective reproductive number Re in 2020 from our phylodynamic model. We estimate that Re dropped from 2.2 - 3.5 the week of 9 March to 0.3 - 0.4 the week of 16 March, coinciding with lockdown measures. Two models fit to hospitalization and death (Lemaitre *et al.*, 2020) and confirmed case (Huisman *et al.*, 2021) data in Switzerland gave similar or slightly lower pre-lockdown Re estimates of 2.1 - 3.8 and 1.6 - 1.9, respectively, and similar or slightly higher post-lockdown Re estimates of 0.3 - 0.6 and 0.6 - 0.8 after 29 March, respectively. Our phylodynamic estimates, which account for an

influx of introduced cases, suggest a sharper reduction in Re coinciding with the Swiss lockdown than these estimates based on epidemiological data. This could be due to accounting for imported cases or the case-count smoothing used by Lemaitre *et al.* (2020); Huisman *et al.* (2021).

Finally, we quantified a summertime "slowdown" dynamic in Switzerland in which introductions initially spread faster, then slowed 35 - 63%. This dynamic was not observable in fall 2020 in Switzerland. A plausible explanation of this dynamic is a successful test-trace-isolate implementation that roughly halved transmissions once an introduction was identified during summer 2020 in Switzerland. We cannot make a statement about the relative speed of transmission chains pre- and post- first sampling in spring 2020. This is because many lineages are ambiguous as to whether they were imported and died out quickly, or resulted in extensive, ongoing local transmission. Therefore, conditioning the birth-death phylodynamic model on few or many introductions during this period yields very different results. For the damping factor analysis, we make the strong assumption that transmission in all lineages descending from an introduction slows simultaneously 2 days after the first genome sample belonging to the introduction is collected. This may be justified if efficient informal backward contact tracing occurred or if individuals in sister lineages were identified around the same time but their samples were not sequenced or not included in our analysis. Then, there are other possible explanatory factors at play. First, travelers returning to Switzerland during summer 2020 have been implicated in transmitting more than non-travelers (Hodcroft *et al.*, 2021). Thus, a passive transmission slowdown might have happened as introduced lineages moved into the non-traveler population. We would expect travelers in fall to have similar contact networks as those in summer, but we do not quantify a transmission slowdown in Switzerland in fall. This coincides with high case numbers during a fall wave, when Swiss contact tracing was reported to be overburdened (SWI, 2020). Second, contacts of positive cases are likely tested more intensely, potentially yielding "bursts" of samples around the first detected cases that subsequently disappear. If so, we can still interpret the slowdown dynamic as evidence that test-trace-isolate implementation was working, but it is difficult to determine precisely by how much transmission actually slowed.

International comparisons also lend perspective to the transmission slowdown effect we quantify from Swiss genome data. Using the same analysis framework, we quantified a significant slowdown effect in New Zealand during two different time periods. Thus, this slowdown effect is not unique

to Switzerland in summer 2020. Importantly, Douglas *et al.* (2021) showed - using genome sequence data - that New Zealand contact tracing was highly effective in identifying SARS-CoV-2 infection clusters. Then, Fetzer and Graeber (2021) exploited an accidental, partial breakdown of English contact tracing to show that normal contact tracing in early fall 2020 reduced transmissions by 63% in the 6 weeks following a positive case. This measure is within the range of our estimates for a transmission slowdown in Switzerland in summer 2020.

Together, our results quantify the reduction of case importation and local transmission in Switzerland during the spring 2020 partial lockdown and partial border closure periods. Further, we provide genome-based quantification of a summertime transmission slowdown in Switzerland that may be linked to successful contact tracing efforts. This slowdown is not observed in fall when contact tracing efforts were overwhelmed in Switzerland but is observed in data from New Zealand in 2020. We have shown that our inference framework is straightforward to apply to different datasets and produces quantitative estimates that we envision can help policy-makers weigh general and specific measures against the respective burdens they impose.

5.4 MATERIALS AND METHODS

Genomic surveillance by the Swiss SARS-CoV-2 Sequencing Consortium in 2020

Altogether 11,357 SARS-CoV-2 genome sequences sampled in Switzerland during 2020 were generated by the Swiss SARS-CoV-2 Sequencing Consortium (S3C, 2021). This sequencing effort represents the majority (79%) of Swiss SARS-CoV-2 genome sequences collected in 2020 and represents the 6th largest contribution of SARS-CoV-2 sequences globally in 2020 (Table 5.2, based on data available on GISAID as of June 2022 (`https://www.gisaid.org/`; (Bogner *et al.*, 2006)). Here, we briefly describe how these samples were generated.

RNA extracts from qPCR-positive patient nasal or oropharyngeal swabs were provided by Viollier AG, a Swiss medical diagnostics company. RNA was extracted using either the Abbott m2000sp or Seegene STARMag 96x4 Universal Cartridge kits. Extracts were then transferred to the Genomics Facility Basel or the Functional Genomics Center Zurich for whole-genome sequencing. Both centers used the ARTIC v3 primer scheme (ARTIC Network, 2020) to generate tiled, approximately 400bp-long amplicons. Li-

brary preparation was done with the New England Biolabs (NEB) library preparation kit. Libraries were sequenced on Illumina MiSeq or NovaSeq machines, resulting in 2 × 251 basepair reads. Bioinformatics processing was performed using V-pipe (Posada-Céspedes *et al.*, 2021), including read trimming and filtering with PRINSEQ (Schmieder and Edwards, 2011), alignment to GenBank accession MN908947 (Wu *et al.*, 2020) with bwa (Li and Durbin, 2009), and consensus base calling. Positions with <5x coverage were masked, positions with >5% and >2 reads supporting a minor base were called with IUPAC ambiguity codes, and positions with >50% reads supporting a deletion were called as a deletion. We rejected samples with <20,000 non-N bases. The consensus sequences are available in the Global Initiative on Sharing Avian Influenza Data (GISAID) repository (Bogner *et al.*, 2006)) under submitting lab "Department of Biosystems Science and Engineering, ETH Zürich".

Dataset construction and sampling procedure

From all sequences available on GISAID (accessed 31 May 2021), we filtered the collection date to on or before 1 December 2020, removed non-human sequences, and sequences <27,000 bases long. We also filtered sequences flagged by the Nextclade tool (Aksamentov *et al.*, 2021) for suspiciously clustered SNPs (QC SNP clusters status metric not "good"; >= 6 mutations in 100 bases), too many private mutations (QC private mutations status metric not "good"; >= 10 mutations from the nearest tree node), or overall bad quality (Nextclade QC overall status "bad"). We aligned sequences to the reference genome MN908947.3 using MAFFT (38). Finally, we followed the Nextstrain pipeline's recommendation to mask the first 100 and last 50 sites of the alignment (Nextstrain, 2020b) since the start and end of SARS-CoV-2 sequences are prone to sequencing errors (De Maio *et al.*, 2020).

From all available Swiss sequences, we sampled up to 5% of confirmed case counts in each Swiss canton each week until 1 December 2020. Confirmed case data was provided by the Swiss Federal Office of Public Health (now available on `https://www.covid19.admin.ch`) (Figure 5.4). At the time of data access, cases were only attributed at the cantonal level beginning in mid-May. Before then, we sampled randomly from across Switzerland. Where not enough sequences were available from a canton in a week, we used all available sequences. To reduce the size of the alignments for phylogenetic analysis, we divided the focal Swiss set into Pango lineages

(Rambaut *et al.*, 2020), similar to Müller *et al.* (2021). Lineages composed of >50% Swiss sequences were aggregated into their parent lineage(s) until <= 50% were Swiss. This aims to ensure that each analyzed lineage originated outside of Switzerland. Table 5.1 lists the analyzed aggregated lineages and the number of sequences per lineage.

We then added the most genetically similar sequences from abroad to each lineage alignment to add a global context. This aims to help distinguish between SARS-CoV-2 variants unique to Switzerland (likely within-Switzerland transmission) and variants also circulating abroad (possibly recent introductions or exports). We considered all non-Swiss sequences from each lineage available on GISAID that pass the quality filtering steps detailed above and applied the Nextstrain priority script (Nextstrain, 2020b) to rank these sequences by their genetic similarity to Swiss sequences in each lineage alignment. Briefly, the priority script ranks a set of foreign context sequences by the Hamming distance to their nearest neighbor within a set of focal sequences. Context sequences are further penalized for having high numbers of masked positions or for being more distant neighbors of the same focal sequence. We selected twice as many context sequences as focal Swiss sequences for each analyzed lineage alignment. Our results are based on a final set of 5,520 focal sequences from Switzerland and 11,009 genetically similar sequences from abroad, which were divided into 148 lineage alignments (Table 5.1).

Phylogenetic analysis

We estimated an approximate maximum likelihood phylogeny for each lineage alignment using IQ-TREE (Nguyen *et al.*, 2014) under an HKY substitution model (Hasegawa *et al.*, 1985) with empirical base frequencies and four gamma rate categories to account for site-to-site heterogeneity (Yang, 1994). We added one of the earliest collected SARS-CoV-2 genomes Wuhan/WH01/2019 (GISAID strain EPI_ISL_406798, GenBank accession MT019529.1) as an outgroup for rooting to each alignment and estimated branch lengths in calendar time units using least-squares dating (LSD) (To *et al.*, 2016) implemented in IQ-TREE. We used a strict molecular clock and a minimum mutation rate of 8×10^{-4} substitutions per site per year (s/s/y), based on estimates by Nextstrain (45). We constrained the most-recent common ancestor to be between 15 November and 24 December 2019, also based on estimates by Nextstrain (Nextstrain, 2020a), and set the minimum branch length to zero. Sequences that violated the strict

clock assumption (Z-score threshold > 3) were removed and near-zero length branches ($< 1.7 \times 10^{-5}$ substitutions per site) were collapsed into polytomies, reflecting the fact that the sequence data alone is not sufficient to resolve the ordering of these transmission events. Given the root date constraints, the mutation rate conformed to the lower bound of 8×10^{-4} with extremely narrow confidence intervals. After removal of sequences violating the strict clock assumption, 5,452 sequences remained across all lineage trees.

Identifying introductions

We identified putative Swiss transmission chains (collections of two or more genome sequences resulting from within-Switzerland transmissions) from each lineage tree while allowing for a fixed number of export events. We used the following criteria applied on a recursive tip-to-root tree traversal: at least two Swiss sequences are part of a clade in the tree and the subtree spanned by these Swiss sequences is monophyletic upon removing (a) up to three export events where (b) only one export event may occur along each internal branch. Exports are clades containing non-Swiss sequences. We chose a conservative value for (b) while still allowing some exports and note that the number of inferred transmission chains is robust to different values for (a) given (b) (Figure 5.5A). We assume the identified transmission chains and remaining singleton Swiss sequences each represent an independent introduction into Switzerland.

We repeated this procedure twice for each lineage tree, making different assumptions upon reaching a polytomy where non-Swiss descendent(s) of the polytomy would cause the proposed introduction to violate criterion (a). First, we split all Swiss clades descending from the polytomy into independent introductions. The second time, we aggregated descendent Swiss clades, going in descending size order, into a single introduction. If in doing this we reached criterion (a), we continued aggregating descendants into a second introduction, and so on. The above procedures are heuristic, but analogous to the ACCTRAN (accelerated transformations) and DELTRAN (delayed transformations) methods for assigning character transformations when multiple scenarios are equally parsimonious (Miyakawa and Narushima, 2004). In summary, we identify introductions twice, generating estimates that represent two plausible sets of many and few introductions at polytomies, where sequence data is not informative about the order of the branching events.

Uncertainty in identifying introductions

We evaluated the effect of several variables on the number and size of identified introductions, as discussed in Supplementary text S1. We found that our two different polytomy assumptions are sufficient to capture most of the uncertainty in the number and size of introductions due to the specific heuristic criteria used to identify introductions from a phylogenetic tree (Figure 5.5A). As expected, increasing the ratio of foreign context to focal Swiss sequences analyzed identifies more, smaller introductions compared to a lower ratio. However, our two different polytomy assumptions at a 2:1 ratio are again sufficient to capture most of this uncertainty (Figure 5.5B).

Quantifying the reduction of introductions during the time of border closures

Prior to fitting our null model for introductions through time, we back-calculated the total number of introductions each week expected under a birth-death skyline model, as described in the section "Phylodynamic analysis" below. Under this model, one can calculate the probability x(t) that a new introduction at time t would have no sampled descendants by 1 December 2020. This formula is given in Stadler *et al.* (2013). We used weekly time bins, taking the median and 95% HPD upper and lower bounds for Re from our phylogenetic analysis (see below), a constant sampling proportion of 5% based on our known sampling scheme, and a constant become-uninfectious rate of 36.5 per year, which corresponds to an average of 10 days to becoming uninfectious (roughly in line with estimates provided by the Swiss Federal Office of Public Health (Swiss Federal Office of Public Health, 2020)). We divided the number of sampled introductions each week by 1 - x(t), the probability an introduction at the start of the week would yield a sampled descendant by 1 December 2020. This yields an estimate for the total number of introductions each week (both sampled and unsampled), while accounting for varying local transmission dynamics.

Then, we assumed a simple null model in which introductions are a linear function of case counts in Switzerland's largest neighboring countries: Italy, France, Germany, and Austria. We used a 7-day rolling average of case count data from the European Centre for Disease Prevention and Control (ECDC) (ECDC, 2020). Further, we considered up to 18 days delay between the actual introduction event and an introduction being sampled. This is based on the 8-day lag from importation to first local transmission estimated by du Plessis *et al.* (2021) in the U.K. and a 10-day infectious period. We back-

calculated total introductions as described above for each plausible delay value using either the median or 95% upper or lower HPD Re estimate from our phylodynamic analysis (see below). We fit the model independently to each of these weekly estimates up to 13 March. We selected the delay yielding the best model fit (lowest root mean squared error using the median Re estimate) for each set of few or many introductions. These were 4 and 5 days, respectively. Finally, we projected introductions after 13 March using the fitted model coefficients and ongoing case counts in the surrounding countries. We did not fit the model to data after border closures were partially lifted because travel behavior was still affected by risk of infection, risk of new restrictions being introduced, and ongoing stay-at-home guidance. This is apparent in data collected by the Swiss Tourism Federation, which demonstrates a marked drop in overnight stays by foreign residents in Switzerland from approximately 6.3 million in the winter season November 2019 - April 2020 to 3.1 million in the summer season May - October 2020 (STV-FST, 2020). As a sensitivity analysis, we also fit the model using confirmed cases in all non-Swiss European countries as defined in the ECDC's case count data (ECDC, 2020) (Figure 5.8).

Quantifying the reduction of persistence of introductions during the lockdown

We developed a second simple null model to test whether the Swiss partial lockdown from 17 March to 27 April 2020 coincided with a change in the persistence of introductions. This null model assumes that in the absence of measures, introductions circulating on any given day persist equally long. In other words, introductions die out (are no longer sampled) according to a delay distribution that is constant through time. For each date, we calculated the time from that date to the last sample for each introduction persisting on that date. Singleton introductions are trivially assumed to persist for 1 day. Then, we report the median and interquartile range of this delay distribution from each date.

Phylodynamic analysis

After identifying introductions, we performed phylodynamic inference on them using the BDSKY (birth-death skyline) method (Stadler *et al.*, 2013) in BEAST2 (Bouckaert *et al.*, 2019). To avoid model mis-specification due to the more transmissible alpha variant, we analyzed data only until 1 December 2020. We also pruned introductions to only include genomes generated

by the S3C, as these were explicitly surveillance samples. This left 4,136 genome sequences for phylodynamic analysis. The phylodynamic inference relies on two main models: a nucleotide substitution model describing an evolutionary process and a population dynamics model describing a transmission and sampling process. For the nucleotide substitution model, we assumed an HKY (Hasegawa *et al.*, 1985) model with four Gamma rate categories to account for site-to-site rate heterogeneity (Yang, 1994). We used the default priors for kappa and the scale factor of the Gamma distribution. We assumed a strict clock with the clock rate fixed to 8×10^{-4} s/s/y, as estimated by (Nextstrain, 2020a).

For the population dynamics model, we used BDSKY (Stadler *et al.*, 2013). In BDSKY, the identified introductions are the result of a birth-death with sampling process parameterized by an effective reproductive number, a becoming-uninfectious rate, and a sampling proportion. As in (Müller *et al.*, 2021), we inferred these population dynamical parameters jointly from the different introductions. More concretely, each introduction is assumed to result from an independent birth-death process having its own origin time, but sharing all other parameters with the processes associated with the other introductions. We applied a uniform prior on the time of origin for each introduction, between 15 February and the oldest sample in the introduction. This constrains introductions to have an origin no earlier than 15 February, excluding the possibility of introductions and subsequent local transmission before the date the first confirmed Swiss case was reported infected abroad in Italy (Keystone-SDA, 2020). After 15 February, our prior expectation is a uniform rate of introductions through time. We fixed the become-uninfectious rate to 36.5 per year, as above. We allowed Re to vary week-to-week, with an Ornstein-Uhlenbeck smoothing prior applied to the logarithm of this parameter. The stationary distribution is LogNormal(0.8, 0.5) and we applied an Exp(1) hyperprior on the relaxation parameter of the process. This prior constrains Re to a wide range of reasonable values (95% range 0.8 - 5.9) and penalizes large changes in Re from week-to-week. Finally, we allowed the sampling proportion to vary when Swiss testing or genome sampling regimes changed significantly (Table 5.3). For our main analysis, we applied a broad LogUniform(10^{-4}, 1) prior on the sampling proportion, since we do not know how many individuals were truly infected. Alternatively, we also tried a LogUniform(10^{-4}, 0.05) prior since we upper-bounded our sampling to 5% of confirmed cases each week (Supplementary text S3).

Finally, we added an additional transmission damping factor to the model. This factor is a multiplicative damping of Re applied to each introduction from 2 days after the oldest to the most-recent sampling date in the introduction. Since we hypothesized contact tracing was not functioning as well during periods of high case numbers, we estimated a separate damping factor for each of three periods: before 15 June 2020 (spring), 15 June to 30 September 2020 (summer), and 30 September to 1 December 2020 (fall). We used the same uninformative spike and slab prior for the damping factor in each period, with an inclusion probability of 0.5 and a uniform prior between 0 and 1, if included.

For each phylodynamic model configuration (bounded and unbounded sampling proportion prior, with and without the contact tracing damping factor) and set of introductions (many and few), we ran five independent MCMC chains. We discarded the first 10% of each chain as burn-in and combined the remaining samples across the five chains. We evaluated the effective sample size (ESS) using Tracer (Rambaut *et al.*, 2018) and verified that the ESS was at least 100 for all inferred parameters.

New Zealand analysis

Genome sequence selection was done as for the Swiss analysis, except that we down-sampled available sequences from GISAID to 40% of confirmed case counts each week rather than 5% and we used national case count numbers rather than stratified by region. Phylogenetic analysis was performed as for the Swiss data. The phylodynamic analysis was also the same, except that we assumed a constant sampling proportion through time and for the bounded sampling proportion prior we used a LogUniform(10^{-4}, 0.4) prior to match the down-sampling scheme.

ACKNOWLEDGMENTS

We gratefully acknowledge the authors from the originating laboratories responsible for obtaining the specimens and the submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based. Joep de Ligt, David Winter, and Jing Wang kindly provided additional information on the source of analyzed New Zealand sequences. A full acknowledgements table of the contributing groups, including the identifiers for all GISAID data used in this study, is available on the project GitHub repository at `https://github.`

com/SarahNadeau/cov-swiss-phylo. We thank Jana Huisman for valuable discussions on the manuscript.

FUNDING

This work was supported by: Swiss National Science Foundation grant 31CA30_196267 (TS) and ETH Zurich.

DATA AND MATERIALS AVAILABILITY

All genome sequence data used in the analysis is available on GISAID (gisaid.org). Data generated by the Swiss SARS-CoV-2 Sequencing Consortium is available on GISAID (submitting lab: Department of Biosystems Science and Engineering, ETH Zürich). The code used to generate figures and values for the manuscript is available at https://github.com/SarahNadeau/cov-swiss-phylo. The phylogenetic analysis code is at https://github.com/cevo-public/Grapevine-SARS-CoV-2-Introduction-Analysis and the phylodynamic analysis code, including BEAST2 XML files, is at https://github.com/tgvaughan/TransmissionChainAnalyses.

BIBLIOGRAPHY

2020. Contact tracing not working properly, writes paper. https://web.archive.org/web/20211020105928/https://www.swissinfo.ch/eng/contact-tracing-not-working-properly--writes-paper/46090060.

Aksamentov, I., Roemer, C., Hodcroft, E. B., and Neher, R. A. 2021. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software*, 6(67): 3773.

ARTIC Network 2020. artic-ncov2019/primer_schemes/nCoV-2019/V3.

Bogner, P., Capua, I., Cox, N. J., and Lipman, D. J. 2006. A global initiative on sharing avian flu data [1]. *Nature*, 442: 981.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., *et al.* 2019. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, 15: e1006650.

Bradley, S. 2020. Switzerland re-opens its european borders - swi swissinfo.ch. https://web.archive.org/web/20220527140231/https://www.swissinfo.ch/eng/covid-19_what-s-happening-at-swiss-borders-and-airports-/45727184.

De Maio, N., Walker, C., Borges, R., Weilguny, L., Slodkowicz, G., and Goldman, N. 2020. Issues with sars-cov-2 sequencing data. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.

Douglas, J., Mendes, F. K., Bouckaert, R., Xie, D., Jiménez-Silva, C. L., Swanepoel, C., de Ligt, J., Ren, X., Storey, M., Hadfield, J., *et al.* 2021. Phylodynamics reveals the role of human travel and contact tracing in controlling the first wave of covid-19 in four island nations. *Virus Evolution*.

du Plessis, L., McCrone, J. T., Zarebski, A. E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T. R., *et al.* 2021. Establishment and lineage dynamics of the sars-cov-2 epidemic in the uk. *Science*, 371: 706–712.

ECDC 2020. Download the daily number of new reported cases of covid-19 by country worldwide. https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide.

Fetzer, T. and Graeber, T. 2021. Measuring the scientific effectiveness of contact tracing: Evidence from a natural experiment. *Proceedings of the National Academy of Sciences of the United States of America*, 118.

FSO 2020. Cross-border commuters statistics in 2nd quarter 2020. https://www.bfs.admin.ch/bfs/en/home/statistics/work-income.gnpdetail.2020-0507.html.

Geoghegan, J. L., Ren, X., Storey, M., Hadfield, J., Jelley, L., Jefferies, S., Sherwood, J., Paine, S., Huang, S., Douglas, J., *et al.* 2020. Genomic epidemiology reveals transmission patterns and dynamics of sars-cov-2 in aotearoa new zealand. *Nature Communications*, 11: 1–14.

Geoghegan, J. L., Douglas, J., Ren, X., Storey, M., Hadfield, J., Silander, O. K., Freed, N. E., Jelley, L., Jefferies, S., Sherwood, J., *et al.* 2021. Use of genomics to track coronavirus disease outbreaks, new zealand. *Emerging Infectious Diseases*, 27: 1317–1322.

Hasegawa, M., Kishino, H., and aki Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of Molecular Evolution*, 22: 160–174.

Hodcroft, E. B., Zuber, M., Nadeau, S., Vaughan, T. G., Crawford, K. H. D., Althaus, C. L., Reichmuth, M., Bowen, J. E., Walls, A. C., Corti, D., *et al.* 2020. Emergence and spread of a sars-cov-2 variant through europe in the summer of 2020. *medRxiv*, page 2020.10.25.20219063.

Hodcroft, E. B., Zuber, M., Nadeau, S., Vaughan, T. G., Crawford, K. H., Althaus, C. L., Reichmuth, M. L., Bowen, J. E., Walls, A. C., Corti, D., *et al.* 2021. Spread of a sars-cov-2 variant through europe in the summer of 2020. *Nature*, pages 1–6.

Huisman, J. S., Scire, J., Angst, D. C., Neher, R. A., Bonhoeffer, S., and Stadler, T. 2021. Estimation and worldwide monitoring of the effective reproductive number of sars-cov-2. *medRxiv*.

Keystone-SDA 2020. Switzerland confirms first coronavirus case - swi swissinfo.ch. https://www.swissinfo.ch/eng/politics/covid-19_switzerland-confirms-first-coronavirus-case/45579278.

Kraemer, M. U., Cummings, D. A., Funk, S., Reiner, R. C., Faria, N. R., Pybus, O. G., and Cauchemez, S. 2019. Reconstruction and prediction of viral disease epidemics. *Epidemiology and Infection*, 147.

Ladoy, A., Opota, O., Carron, P. N., Guessous, I., Vuilleumier, S., Joost, S., and Greub, G. 2021. Size and duration of covid-19 clusters go along with a high sars-cov-2 viral load: A spatio-temporal investigation in vaud state, switzerland. *Science of The Total Environment*, 787: 147483.

Lemaitre, J. C., Perez-Saez, J., Azman, A. S., Rinaldo, A., and Fellay, J. 2020. Assessing the impact of non-pharmaceutical interventions on sars-cov-2 transmission in switzerland. *Swiss Medical Weekly 2020 :21*, 150.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25: 1754–1760.

Mallon, P., Crispie, F., Gonzalez, G., Tinago, W., Leon, A. G., McCabe, M., de Barra, E., Yousif, O., Lambert, J., Walsh, C., *et al.* 2020. Whole-genome sequencing of sars-cov-2 in the republic of ireland during waves 1 and 2 of the pandemic. *medRxiv*, pages 1–13.

Marquis, B., Opota, O., Jaton, K., and Greub, G. 2021. Impact of different sars-cov-2 assays on laboratory turnaround time. *Journal of Medical Microbiology*, 70.

Miller, D., Martin, M. A., Harel, N., Tirosh, O., Kustin, T., Meir, M., Sorek, N., Gefen-halevi, S., Amit, S., Vorontsov, O., *et al.* 2020. Full genome viral sequences inform patterns of sars-cov-2 spread into and within israel. *Nature Communications*.

Miyakawa, K. and Narushima, H. 2004. Lattice-theoretic properties of mpr-posets in phylogeny. *Discrete Applied Mathematics*, 134: 169–192.

Morel, B., Barbera, P., Czech, L., Bettisworth, B., Hübner, L., Lutteropp, S., Serdari, D., Kostaki, E. G., Mamais, I., Kozlov, A. M., *et al.* 2021. Phylogenetic analysis of sars-cov-2 data is difficult. *Molecular Biology and Evolution*, 38: 1777–1791.

Munnink, B. B. O., Worp, N., Nieuwenhuijse, D. F., Sikkema, R. S., Haagmans, B., Fouchier, R. A., and Koopmans, M. 2021. The next phase of sars-cov-2 surveillance: real-time molecular epidemiology. *Nature Medicine*, 27: 1518–1524.

Müller, N. F., Wagner, C., Frazar, C. D., Roychoudhury, P., Lee, J., Moncla, L. H., Pelle, B., Richardson, M., Ryke, E., Xie, H., *et al.* 2021. Viral genomes reveal patterns of the sars-cov-2 outbreak in washington state. *Science Translational Medicine*, 13: 202.

New Zealand Government 2020. History of the covid-19 alert system | unite against covid-19. `https://web.archive.org/web/20211020111429/https://covid19.govt.nz/alert-levels-and-updates/history-of-the-covid-19-alert-system/`.

Nextstrain 2020a. Nextstrain / ncov / gisaid / global / 6m. `https://nextstrain.org/ncov/gisaid/global/6m?l=clock`.

Nextstrain 2020b. nextstrain/ncov: Nextstrain build for novel coronavirus (ncov). `https://github.com/nextstrain/ncov`.

Nguyen, L.-T., Schmidt, H. A., Haeseler, A. V., and Minh, B. Q. 2014. Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32: 268–274.

of Public Health, S. F. O. 2020. Covid-19 switzerland | coronavirus | dashboard. `https://web.archive.org/web/20211104165544/https://www.covid19.admin.ch/en/epidemiologic/case`.

Posada-Céspedes, S., Seifert, D., Topolsky, I., Jablonski, K. P., Metzner, K. J., and Beerenwinkel, N. 2021. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. *Bioinformatics*, 37: 1673–1680.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. 2018. Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic Biology*, 67: 901–904.

Rambaut, A., Holmes, E. C., Áine O'Toole, Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., and Pybus, O. G. 2020. A dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. *Nature Microbiology*, 5: 1403–1407.

Rockett, R. J., Arnott, A., Lam, C., Sadsad, R., Timms, V., Gray, K.-A., Eden, J.-S., Chang, S., Gall, M., Draper, J., *et al.* 2020. Revealing covid-19 transmission in australia by sars-cov-2 genome sequencing and agent-based modeling. *Nature Medicine 2020 26:9*, 26: 1398–1404.

S3C 2021. Swiss SARS-CoV-2 Sequencing Consortium (S3C). https://bsse.ethz.ch/cevo/research/sars-cov-2/swiss-sars-cov-2-sequencing-consortium.html.

Schmieder, R. and Edwards, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27: 863–864.

Stadler, T. 2010. Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, 267: 396–404.

Stadler, T., Kühnert, D., Bonhoeffer, S., and Drummond, A. J. 2013. 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: 228–233.

Stange, M., Mari, A., Roloff, T., Seth-Smith, H. M., Schweitzer, M., Brunner, M., Leuzinger, K., Søgaard, K. K., Gensch, A., Tschudin-Sutter, S., *et al.* 2021. Sars-cov-2 outbreak in a tri-national urban area is dominated by a b.1 lineage variant linked to a mass gathering event. *PLOS Pathogens*, 17: e1009374.

STV-FST 2020. Schweizer tourismus in zahlen. https://www.stv-fst.ch/de/stiz.

Swiss Federal Office of Public Health 2020. Frequently asked questions (faqs). https://www.bag.admin.ch/bag/en/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov/haeufig-gestellte-fragen.html.

The Swiss Federal Council 2020. Federal council to gradually ease measures against the new coronavirus. https://www.admin.ch/gov/en/start/documentation/media-releases.msg-id-78818.html.

To, T. H., Jung, M., Lycett, S., and Gascuel, O. 2016. Fast dating using least-squares criteria and algorithms. *Systematic Biology*, 65: 82–97.

World Health Organization 2020. Tracking sars-cov-2 variants. https://web.archive.org/web/20211104163350/https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y., *et al.* 2020. A new coronavirus associated with human respiratory disease in china. *Nature 2020 579:7798*, 579: 265–269.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39: 306–314.

5.5 SUPPLEMENTAL MATERIAL

Supplementary text

S1: Sensitivity analyses for identifying introductions

Here we describe different sensitivity analyses we performed for the definition of an introduction.

Criteria for identifying introductions. First, we assessed the sensitivity of identified introductions to the precise heuristic definition of an introduction. We began with the same lineage-specific phylogenetic trees generated for the main analysis. Then, we re-identified introductions from these trees while varying (a) the maximum number of export events allowed from each introduction and (b) the maximum number of consecutive export events allowed to occur along each single internal branch. Figure 5.5A shows that increasing (a) yields fewer, larger introductions. Increasing (b) for each level of (a) has a negligible effect. However, the greatest differences come from the different assumptions about how to resolve polytomies before applying these heuristics. Increasing (a) from 1 to 4 yields approximately 25% fewer introductions, while resolving polytomies such that Swiss descendants cluster together yields approximately 75% fewer introductions (Figure 5.5A). We chose to present results using an introduction definition based on (a) a maximum of three exported lineages and (b) a maximum one consecutive export on each internal branch. This allows for some exports from Swiss introductions but not arbitrarily many. We rely on our different polytomy assumptions to capture most of the uncertainty in the number and size of introductions.

Ratio of foreign context to focal Swiss sequences analyzed. Next, we assessed the sensitivity of identified introductions to the sequence set analyzed. We re-sampled sequences to analyze three times, each time taking a number of Swiss sequences corresponding to up to 5% of confirmed cases each week. Then we sampled foreign context sequences at a 1:1, 2:1, and 3:1 ratio to the Swiss sequences. Figure 5.5B shows that as we add foreign context sequences, we identify more numerous, smaller introductions. However, the greatest differences come from the different assumptions about how to resolve polytomies (few vs. many introductions), not the ratio of foreign context to focal Swiss sequences. Therefore, we chose to present results using the 2:1 ratio to balance speed (smaller dataset = faster tree search convergence) and information content (larger dataset = more introductions represented).

Number of focal Swiss sequences analyzed. Finally, we assessed whether the number of identified introductions saturates as we add more Swiss sequences. To do this, we sub-sampled the Swiss genome sequences used in our main analysis to 20, 40, 60, and 80% of the full set of sequences analyzed, pruning the not-included Swiss sequences from the phylogenetic trees generated for the main analysis. Then, we calculated the number of introductions we would have identified on the pruned trees. We performed the random sub-sampling and pruning 50 times for each sub-sampling level. Figure 5.5C shows that as we approach the number of Swiss sequences used in the main analysis, we do not reach saturation. Therefore, if we were to include even more sequences, we would identify more introductions.

S2: New Zealand validation data

For New Zealand, the sequence submitters provided additional information on which samples were from cases in managed isolation and quarantine (MIQ) facilities versus the broader community. This allows us to partially evaluate our introduction identification methods. 117 of the 1234 analyzed focal sequences in the New Zealand analysis originated from MIQ facilities. 63 (54%) of these were singletons under the "many introductions" polytomy assumption versus 37 (32%) under the "few introductions" assumption. 44 (38%) or 37 (32%) were plausible within-MIQ outbreaks. These were identified as introductions with cases all from a single region and all MIQ. They may represent groups of individuals quarantining together or infected in the same source location. These outbreaks included, on average, 3 samples spanning 5 days (many introductions) or 2 samples spanning 9 days (few introductions). The remaining 10 (9%) or 43 (37%) of MIQ sequences were in introductions including community cases or including cases in multiple MIQ facilities in different regions, which we deem unrealistic. These results support that the "many introductions" polytomy assumption is more realistic when the probability of infection abroad is high compared to the probability of locally acquired infection.

S3: Sensitivity analyses for phylodynamic modeling

Here we describe a sensitivity analysis and some example intermediate outputs from our phylodynamic analysis.

Sampling proportion prior. We repeated our analyses using two different priors on the sampling proportion. The first, unbounded prior was LogUniform(10^{-4}, 1). This broad prior allows the sampling proportion to

assume any value. The second, bounded prior was LogUniform(10^{-4}, 0.05). This narrower prior is motivated by our 5% down-sampling based on confirmed case numbers. Figure 5.10A shows that in Switzerland, the estimated sampling proportion in late fall 2020 varies greatly depending on the prior. The rise in prevalence of lineage B.1.177 during this period (Hodcroft *et al.*, 2020), representing a drop in SARS-CoV-2 diversity in Switzerland, might explain why the inference under the broader sampling prior estimates a proportion corresponding to fewer individuals than we know were infected during this time. Figure 5.10B shows that the effective reproductive number estimates in fall 2020 for Switzerland more closely match estimates based on confirmed case data when the sampling proportion is treated as a fitting parameter, i.e., under the first, broad prior. Therefore, we report results under this prior in the main text. In Figure 5.11A, we show that the damping factor results are qualitatively similar between the two sampling proportion priors. For the New Zealand analysis, Re estimates are not affected by bounding the sampling proportion or not (Figure 5.13).

Logged trees. Finally, we visually inspected phylogenetic trees for a few introductions. These trees were sampled and logged by the Markov chains in the phylodynamic analyses. Note that the damping factor results are jointly inferred from all the branching events across introductions in each time period. For each set of model assumptions and each month, we inspected maximum clade credibility summary trees for the 50th and 95th percentile largest introductions that were first sampled that month and eventually yielded >2 samples. Figure 5.12 shows as an example summary trees for these introductions from one of the MCMC chains in the phylodynamic analysis for Switzerland with damping factors and an unbounded sampling proportion prior.

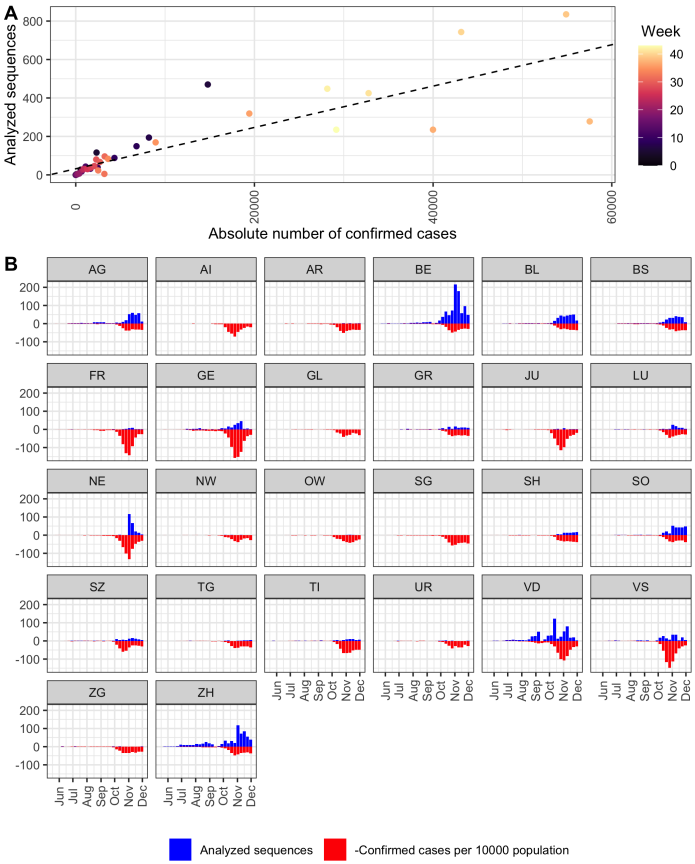


FIGURE 5.4: **Number of analyzed sequences compared to confirmed case counts each week for (A) all of Switzerland and (B) stratified by canton after the week of 18 May 2020, when case count data is also stratified by canton.** The best-fit line in (A) has an R^2 value of 0.72. Week 0 corresponds to the start of sampling with the first sequence from Switzerland collected on 24 February 2020. Facet names in (B) are standard abbreviations for Swiss cantons.

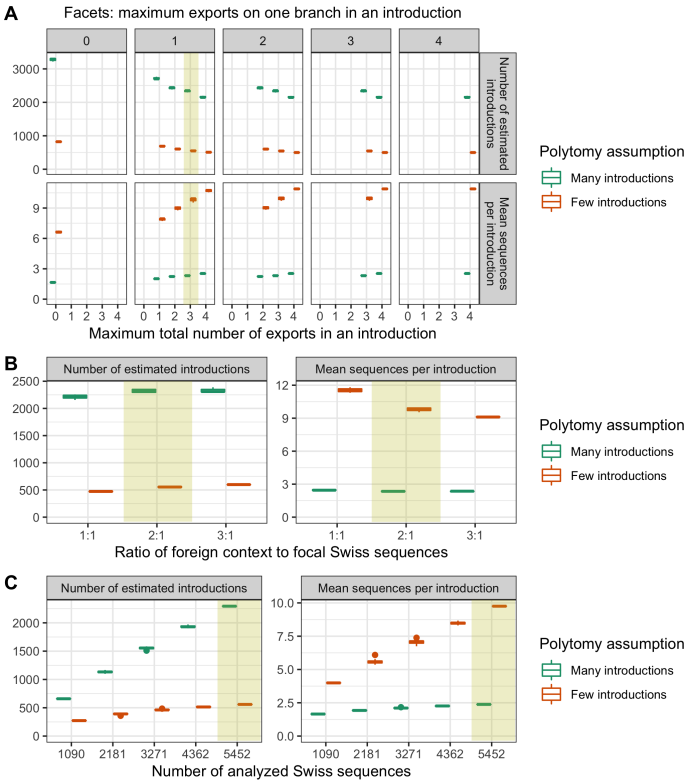


FIGURE 5.5: **Sensitivity analyses showing how summary statistics on the number and size of identified Swiss introductions change depending on the definition of an introduction.** See Supplementary text S1 for details of the sensitivity analyses. (**A**) shows sensitivity to the heuristic thresholds used to define an introduction based on the lineage phylogenies, (**B**) shows sensitivity to the ratio of foreign context to focal sequences analyzed, and (**C**) shows sensitivity to the number of focal sequences analyzed. All statistics were generated under two different polytomy assumptions giving rise to either few or many introductions. Boxplots in (A) and (B) are for 3 randomly drawn datasets, boxplots in (C) are for 50 random sub-samples from the same dataset. Shaded yellow rectangles highlight values used for the main analysis.

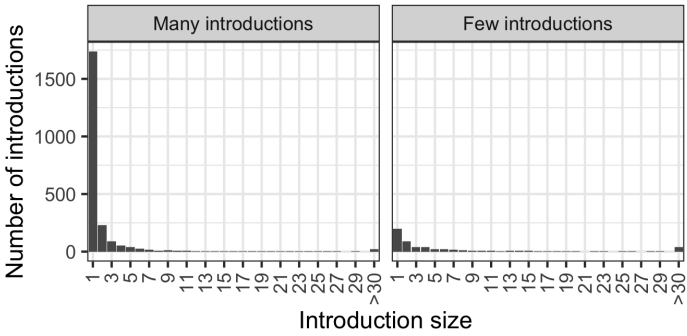


FIGURE 5.6: **Size distribution of estimated Swiss introductions.**

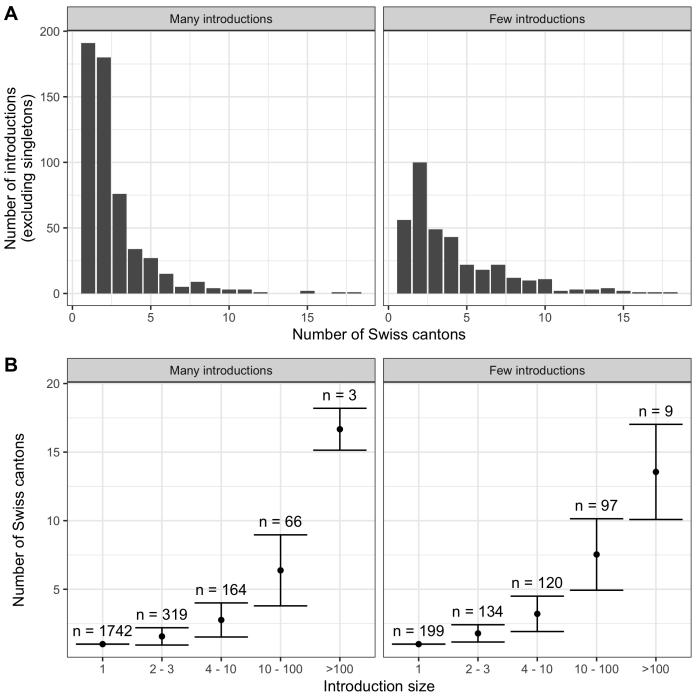


FIGURE 5.7: **Geographic distribution of Swiss introductions estimated under each polytomy assumption.** (**A**) shows that most transmission chains were sampled in only one or two cantons. (**B**) shows that larger introductions were sampled in more Swiss cantons. Points show the mean number of cantons and error bars show the standard deviation in the number of cantons. Labels given the number of introductions in each size bin under each polytomy assumption.

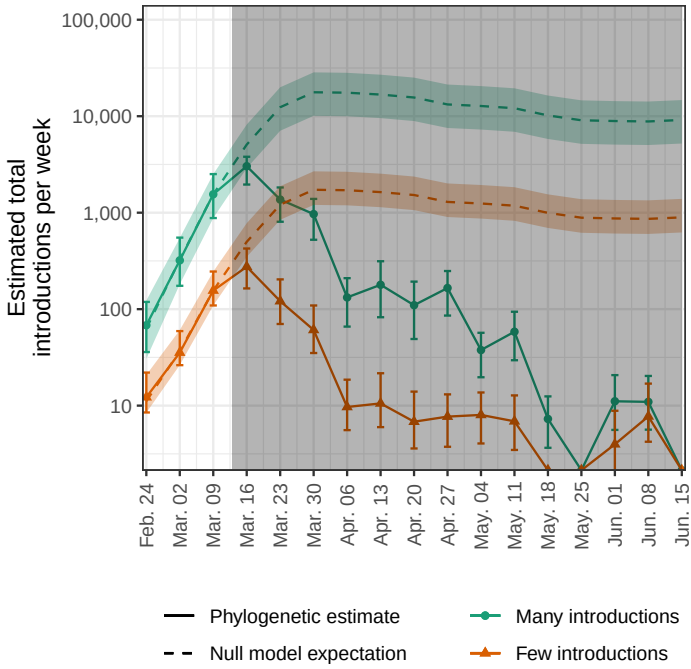


FIGURE 5.8: **Sensitivity analysis for modeling Swiss introductions.** Estimated total introductions i.e., introductions scaled to account for time-varying sampling proportion (solid lines) are compared to a null model (dashed lines) where total introductions are a linear function of case numbers in all non-Swiss European countries, as defined in the European Centre for Disease Control (ECDC)'s case count data (ECDC, 2020). The null model is fit to the points prior to the border closure on 13 March (highlighted with shaded rectangle), values after that are a model prediction. Uncertainty bounds for total introductions (error bars) and null model predictions (colored shaded areas) are based on the 95% upper and lower HPD bounds for Re when estimating total introductions. The orange and green colors correspond to estimates generated under our few and many introductions polytomy assumptions, respectively.

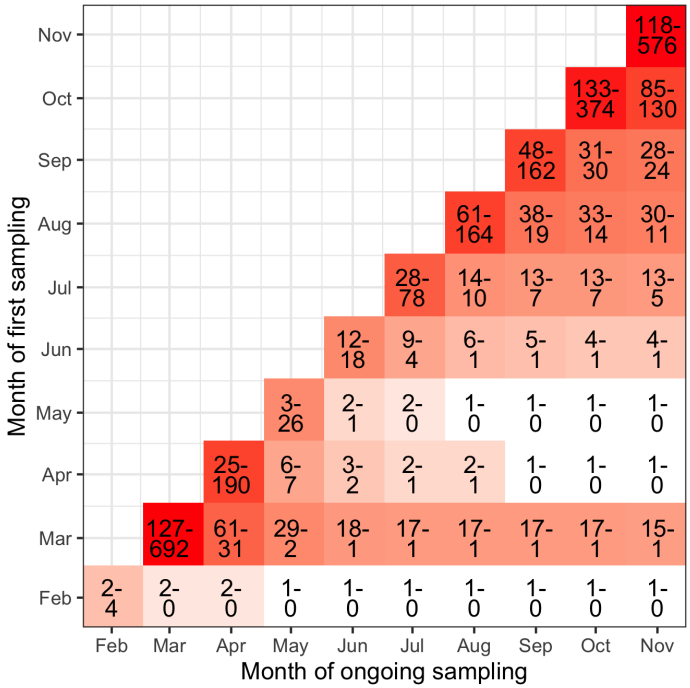


FIGURE 5.9: **Heatmap of the number of newly sampled introductions in Switzerland each month (diagonal entries) and the number continuing to persist into each following month (off-diagonal entries).** Introductions are counted once in the month they are first sampled ("Month of first sampling") and one every following month ("Month of ongoing sampling") until the date of the latest sample. Estimates were generated under two different polytomy assumptions giving rise to either few or many introductions. Ranges are: few-many.

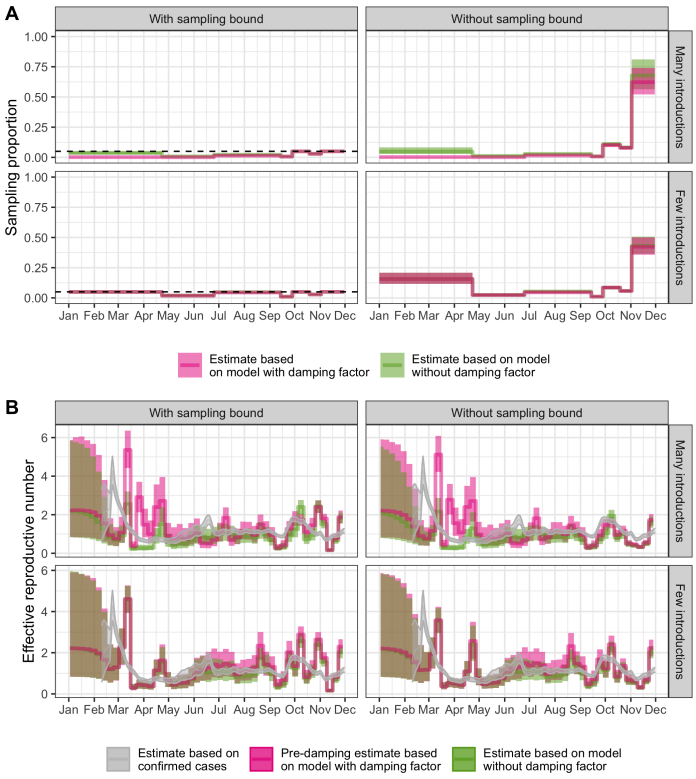


FIGURE 5.10: **Phylodynamic estimates for (A) the sampling proportion and (B) the time-varying effective reproductive number Re in Switzerland.** The dashed line in (A) shows the sampling proportion prior's upper bound, if applicable. Re estimates in (B) are overlaid with estimates based on confirmed case count data (Huisman *et al.*, 2021) in gray. Additionally, Re estimates from the models with a damping factor (pink) are the "baseline" Re before introduction-specific damping (i.e., before application of a damping factor once introductions are older than 2-days post sampling).

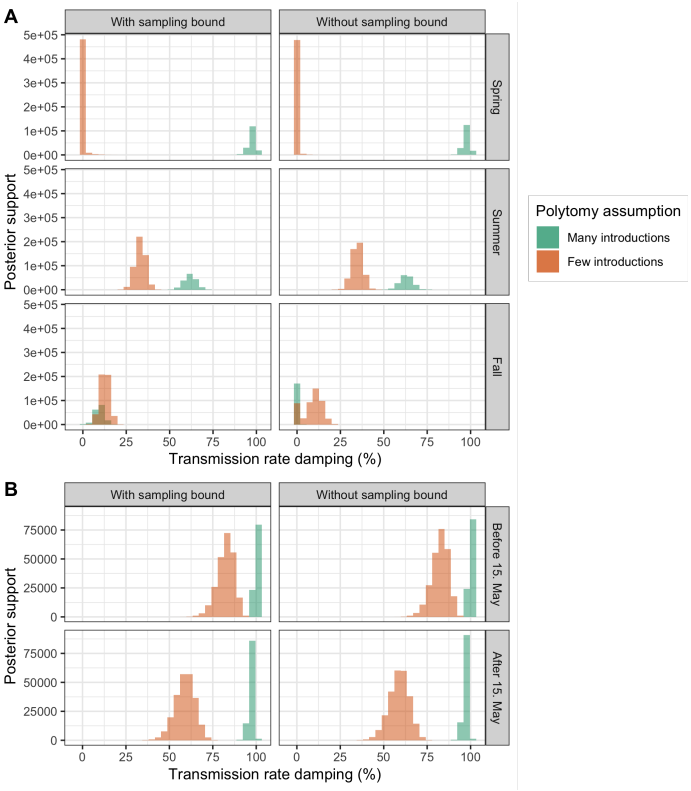


FIGURE 5.11: **Phylodynamic estimates for the damping factor in (A) Switzerland and (B) New Zealand in different time periods, conditioned on introductions estimated under two different polytomy assumptions giving rise to either few or many introductions.**

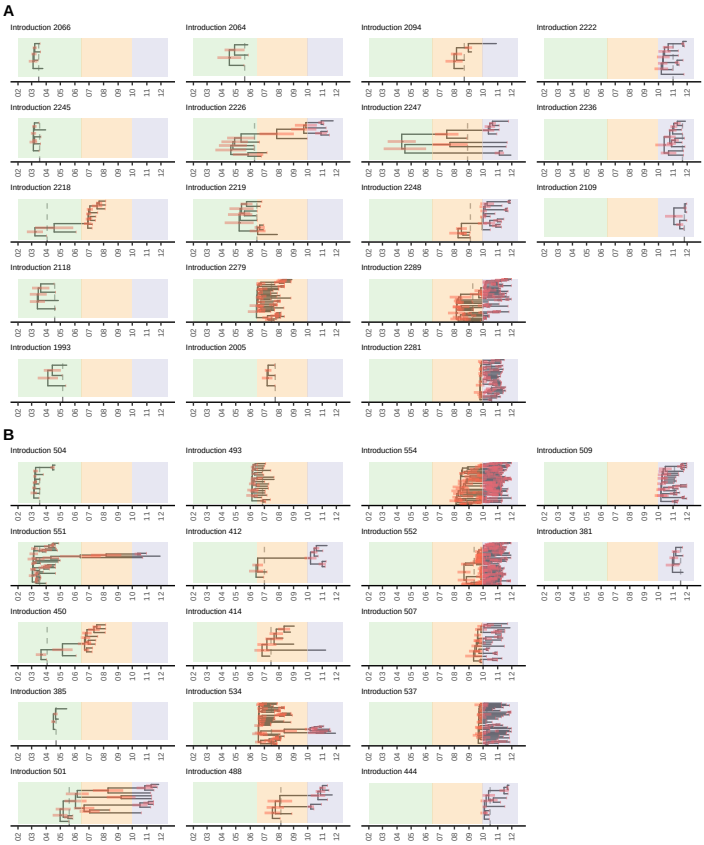


FIGURE 5.12: **A selection of maximum clade credibility summary trees from one of the MCMC chains in the phylodynamic analysis for Switzerland with damping factors and an unbounded sampling proportion prior.** Here we show the 50th and 95th percentile introduction by size each month Mar - Nov 2020. Months are abbreviated by their number. (**A**) shows trees from the analysis conditioned on many introductions and (**B**) conditioned on few introductions. The three different color regions represent the spring (green), summer (orange) and fall (blue) periods. Vertical dashed lines show when the damping factor applies for each introduction - two days after the first sample date. Red bars show the 95% highest posterior density uncertainty in node dates.

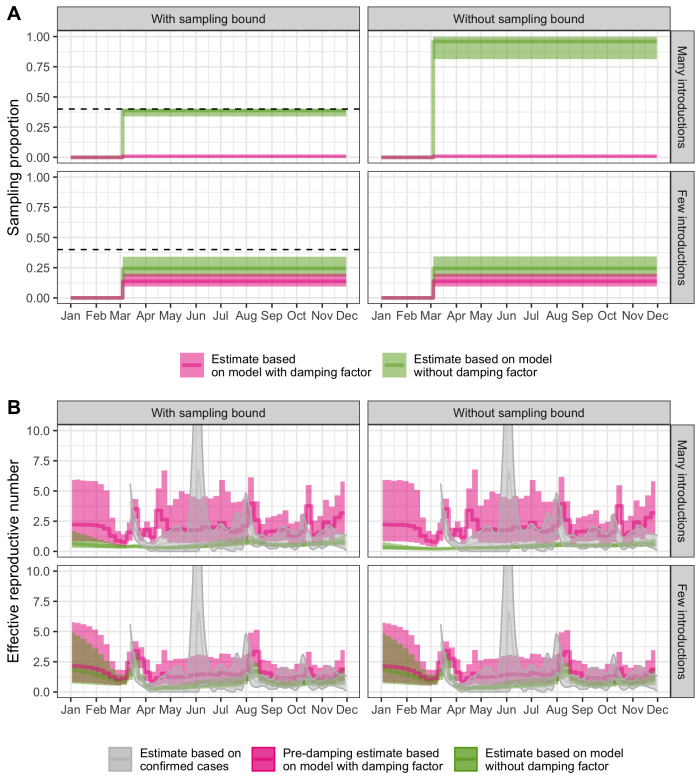


FIGURE 5.13: **Phylodynamic estimates for (A) the sampling proportion and (B) the time-varying effective reproductive number Re in New Zealand.** The dashed line in (A) shows the sampling proportion priors upper bound, if applicable. Re estimates in (B) are overlaid with estimates based on confirmed case count data (28) in gray. Additionally, Re estimates from the models with a damping factor (pink) are the "baseline" Re before introduction-specific damping (i.e. before application of a damping factor once introductions are older than 2-days post sampling).

TABLE 5.1: Summary of Pango lineages analyzed. If more than 50% of the samples from a lineage in the full, quality-filtered dataset were Swiss, we aggregated them into the parent lineage. The percentage of Swiss samples in the final, aggregated lineage sets are given in column "% lineage Swiss". Lineage aliases were also aggregated with their extended-form names. A separate phylogeny was constructed for each lineage analyzed.

Lineage analyzed	No. Swiss samples analyzed	Lineages aggregated	% lineage Swiss
B.1.160	1347	B.1.160, B.1.160.10, B.1.160.11, B.1.160.12, B.1.160.14, B.1.160.15, B.1.160.16, AB, B.1.160.19, B.1.160.20, B.1.160.22, B.1.160.26, B.1.160.29, B.1.160.30, B.1.160.31, B.1.160.32, B.1.160.9, B.1.160.16.1, AB.1	19.00
B.1.177	1260	B.1.177, B.1.177.23, B.1.177.28, B.1.177.43, B.1.177.44, B.1.177.71	4.80
B.1	930	B.1, B.1.214.2	2.20
B.1.1	655	B.1.1, B.1.1.144, B.1.1.327, B.1.1.39, AQ, B.1.1.524	3.10
B.1.221	176	B.1.221	8.10
B.1.1.70	108	B.1.1.70, AP	15.00
B.1.416.1	105	B.1.416.1	45.00
B.1.258	101	B.1.258	4.40
B.1.367	60	B.1.367	10.00
B.1.236	59	B.1.236	33.00
B.1.1.1.35	53	B.1.1.1.35, C.35	13.00
B.1.36.1	47	B.1.36.1	35.00
B.1.128	31	B.1.128	3.30
B.1.93	31	B.1.93	3.30
B.1.1.277	27	B.1.1.277, K	5.80

B.1.1.47	24	B.1.1.47	32.00
B.1.1.269	19	B.1.1.269	5.00
B.1.1.1.36	16	B.1.1.1.36, C.36, B.1.1.1.36.2, C.36.2	9.30
B.1.1.10	16	B.1.1.10, L	2.70
B.1.1.7	16	B.1.1.7, Q	0.85
B.1.1.1	15	B.1.1.1, C, B.1.1.1.5, C.5	0.86
B.1.1.189	15	B.1.1.189	12.00
B.1.146	15	B.1.146	30.00
B.1.1.232	14	B.1.1.232, AK	3.10
B	11	B	0.44
B.1.1.153	11	B.1.1.153	6.50
B.1.1.305	11	B.1.1.305, AF, B.1.1.305.1, AF.1	8.40
B.1.1.372	11	B.1.1.372	0.95
B.1.177.75	11	B.1.177.75	12.00
B.1.177.77	11	B.1.177.77	6.10
B.1.1.200.1	10	B.1.1.200.1, AN.1	33.00
B.1.147	10	B.1.147	0.84
B.1.177.81	10	B.1.177.81	1.80
B.1.1.37	9	B.1.1.37	0.42
B.1.177.33	8	B.1.177.33	4.50
B.1.36	8	B.1.36	0.80
B.1.509	8	B.1.509	2.30
B.1.1.433	7	B.1.1.433	7.80
B.1.1.521	7	B.1.1.521	19.00
B.1.36.17	7	B.1.36.17	1.20
B.1.8	7	B.1.8	1.80
B.1.91	7	B.1.91	1.60
B.1.177.51	6	B.1.177.51	20.00
B.1.258.17	6	B.1.258.17	1.40
B.1.467	6	B.1.467	33.00
B.1.1.242	5	B.1.1.242	35.00

B.1.1.58	5	B.1.1.58	14.00
B.1.177.83	5	B.1.177.83	7.80
B.1.177.85	5	B.1.177.85	11.00
B.1.535	5	B.1.535	0.59
B.40	5	B.40	0.23
B.1.1.218	4	B.1.1.218	5.20
B.1.1.241	4	B.1.1.241, AH	4.20
B.1.1.428	4	B.1.1.428	50.00
B.1.1.464	4	B.1.1.464, AW	1.20
B.1.258.14	4	B.1.258.14	11.00
B.1.356	4	B.1.356	0.82
A	3	A	0.15
B.1.1.170	3	B.1.1.170	2.60
B.1.1.231.1	3	B.1.1.231.1, AL.1	0.34
B.1.1.297	3	B.1.1.297, AG	1.90
B.1.1.317	3	B.1.1.317, AS	2.10
B.1.1.371	3	B.1.1.371	6.20
B.1.177.52	3	B.1.177.52, Y	2.80
B.1.177.53	3	B.1.177.53, W	3.60
B.1.389	3	B.1.389	1.50
B.1.474	3	B.1.474	14.00
B.1.480	3	B.1.480	4.30
B.1.9.5	3	B.1.9.5	11.00
B.11	3	B.11	1.80
B.3	3	B.3	0.37
A.2	2	A.2	0.22
B.1.1.219	2	B.1.1.219	1.60
B.1.1.243	2	B.1.1.243	4.20
B.1.1.33	2	B.1.1.33, N	0.11
B.1.1.44	2	B.1.1.44	0.58
B.1.1.50	2	B.1.1.50	1.20
B.1.160.28	2	B.1.160.28	1.40

B.1.177.15	2	B.1.177.15, AA	0.21
B.1.177.32	2	B.1.177.32	1.10
B.1.177.53.1	2	B.1.177.53.1, W.1	7.70
B.1.177.55	2	B.1.177.55	0.87
B.1.177.60	2	B.1.177.60, U	2.50
B.1.177.62	2	B.1.177.62	6.20
B.1.177.80	2	B.1.177.80	17.00
B.1.177.82	2	B.1.177.82	0.63
B.1.177.86	2	B.1.177.86	2.10
B.1.218	2	B.1.218	6.50
B.1.408	2	B.1.408	3.50
B.1.416	2	B.1.416	0.94
B.1.523	2	B.1.523	0.88
B.1.9.4	2	B.1.9.4	12.00
B.28	2	B.28	0.57
B.4	2	B.4	0.54
B.58	2	B.58	2.20
B.59	2	B.59	1.30
A.5	1	A.5	0.21
B.1.1.1.30	1	B.1.1.1.30, C.30	0.19
B.1.1.142	1	B.1.1.142	6.00
B.1.1.145	1	B.1.1.145	4.50
B.1.1.198	1	B.1.1.198	0.18
B.1.1.221	1	B.1.1.221	1.20
B.1.1.266	1	B.1.1.266	4.90
B.1.1.28	1	B.1.1.28, P	0.07
B.1.1.294	1	B.1.1.294, M	0.28
B.1.1.294.2	1	B.1.1.294.2, M.2	50.00
B.1.1.315	1	B.1.1.315, AD	1.40
B.1.1.331	1	B.1.1.331	2.40
B.1.1.336	1	B.1.1.336	7.10
B.1.1.355	1	B.1.1.355	2.50

B.1.1.369	1	B.1.1.369	0.05
B.1.1.406	1	B.1.1.406	3.10
B.1.1.409	1	B.1.1.409	0.83
B.1.1.519	1	B.1.1.519	2.60
B.1.1.71	1	B.1.1.71	1.30
B.1.12	1	B.1.12	0.89
B.1.127	1	B.1.127	0.53
B.1.149	1	B.1.149	2.70
B.1.177.31	1	B.1.177.31	50.00
B.1.177.50.1	1	B.1.177.50.1, Z.1	0.50
B.1.177.53.3	1	B.1.177.53.3, W.3	0.65
B.1.177.6	1	B.1.177.6	0.19
B.1.177.7	1	B.1.177.7	0.03
B.1.177.72	1	B.1.177.72	1.80
B.1.2	1	B.1.2	0.01
B.1.213	1	B.1.213	4.20
B.1.220	1	B.1.220	1.20
B.1.221.1	1	B.1.221.1	0.34
B.1.229	1	B.1.229	1.10
B.1.258.4	1	B.1.258.4	0.46
B.1.258.7	1	B.1.258.7	0.27
B.1.258.9	1	B.1.258.9	0.65
B.1.36.22	1	B.1.36.22	0.24
B.1.36.24	1	B.1.36.24	4.50
B.1.36.35	1	B.1.36.35	2.10
B.1.397	1	B.1.397	0.86
B.1.398	1	B.1.398	1.40
B.1.400	1	B.1.400	0.10
B.1.406	1	B.1.406	1.90
B.1.415	1	B.1.415	1.40
B.1.513	1	B.1.513	1.40
B.1.520	1	B.1.520	0.10

B.1.540	1	B.1.540	1.60
B.1.88.1	1	B.1.88.1	0.60
B.39	1	B.39	0.26
B.55	1	B.55	0.55
B.6	1	B.6	0.14
None	1	None	1.00

TABLE 5.2: Top 20 largest SARS-CoV-2 sequencing data contributors to GISAID in 2020 by submitting lab.

Submitting lab	Countries represented (ISO codes)	Number of sequences
Wellcome Sanger Institute for the COVID-19 Genomics UK (COG-UK) Consortium	GBR	96441
COVID-19 Genomics UK (COG-UK) Consortium	GBR	71371
Albertsen Lab, Department of Chemistry and Bioscience, Aalborg University, Denmark	DNK	27936
Houston Methodist Hospital	USA	27409
Pathogen Genomics Center, National Institute of Infectious Diseases	JPN; MMR	19708
Department of Biosystems Science and Engineering, ETH Zürich	CHE	11357
MDU-PHL	AUS; TLS	10459
TGen North	USA	9491
Wyoming Public Health Laboratory	USA	9172
Aalborg University	DNK	8439
SeqCOVID-SPAIN consortium/IBV(CSIC)	ESP	8279
Chan-Zuckerberg Biohub	USA	7803
BCCDC Public Health Laboratory	CAN	7646
Laboratoire de santé publique du Québec	CAN	6914
Andersen lab at Scripps Research	JOR; MEX; USA	6258
Utah Public Health Laboratory	USA	5925
MEPHI, Aix Marseille University	FRA	5617
Respiratory Virus Unit, Microbiology Services Colindale, Public Health England	GBR; UKR	5142
deCODE genetics	ISL	5005
Erasmus Medical Center	BEL; BHR; LUX; NLD; SUR	4594

TABLE 5.3: Sampling proportion change-points for the phylodynamic analysis on Swiss data. The sampling proportion was modeled as a piecewise-constant function in time, with the following change-points motivated by major shifts in the testing regime or genome sequencing intensity in Switzerland.

Start date	Description
23 April 2020	All symptomatic individuals can get tested
25 June 2020	Government pays for tests for symptomatic individuals
14 September 2020	Genome sequencing << 5% of confirmed cases
28 September 2020	Number of tests conducted and % positivity dramatically increase, genome sequencing also increases
19 October 2020	Genome sequencing << 5% of confirmed cases again
11 November 2020	Genome sequencing increases again

6

SUMMARY

Ten years after the release of the first human genome in 2003, the overall impact of genome sequencing on diagnostics, therapeutics, and public health was rated as modest (Nature, 2010). Now, approximately twenty years later, the successes and failures of the genomic revolution are still being tallied. In this thesis, I refined and applied phylogeny-based methods to learn about infectious disease pathology and transmission dynamics. In this final section, I will try to put these projects into perspective. I will highlight where we were able to push the envelope to generate new public-health relevant information and where we stumbled on remaining hurdles. Lowering or eliminating these hurdles will be key to generating translational impact from pathogen genome sequencing going forward.

In the human GWAS field, a current focus is on generating high-quality phenotype data from increasingly large cohorts in order to increase power to detect rare human genetic variants or variants of small effect (Uffelmann *et al.*, 2021). In Chapter 2, we took a different approach by combining host and pathogen sequencing to improve host GWAS for infectious disease. Namely, we estimated and then removed correlations due to shared pathogen ancestry from trait values prior to GWAS using host genomes. In our two applications, to HIV-1 set-point viral load in humans and quantitative disease resistance to *X. arboricola* in *A. thaliana*, we find that host GWAS is robust to a correction for pathogen effects, supporting prior GWAS results. I envision this work's impact will be twofold. First, we publicized the phylogenetic Ornstein-Uhlenbeck mixed model to the human and plant GWAS fields. I anticipate this model will be useful in understanding the genetic basis and evolutionary dynamics of important traits in other host-pathogen systems. Second, we highlighted that host GWAS can miss true host genomic associations if pathogen effects are significantly heritable from infection partner to infection partner. I anticipate future host GWAS in infectious disease will be aware of this problem and use either our method or another of the methods highlighted in our discussion to account for pathogen effects to ensure GWAS results are robust.

In phylogenetic and phylodynamic applications to infectious diseases, novel insights are being driven by more comprehensive sampling, incorporation of different data types, and more bespoke models (Featherstone *et al.*,

2022). In Chapters 3 and 5, we capitalized on global sequencing efforts, case count data, and the flexible BEAST2 phylodynamic inference framework to study SARS-CoV-2 transmission during the early COVID-19 pandemic. First, we showed that border closures in Europe came too late to delay the onset of local transmission at the start of the pandemic. Then, we generated estimates, albeit with large uncertainty, for the effect of border closures on limiting case introductions, the effect of lockdown on reducing the effective reproductive number and the persistence of introduced viral lineages, and, finally, the potential effect of contact tracing on reducing onward transmission in the 2020 Swiss epidemic. The first project's impact derived from its timeliness - as a publicly-funded scientist, I was proud we could contribute to the ongoing public discourse in 2020 about travel restrictions to combat COVID-19. The second project progressed much slower, as we poured time and energy into refining our sampling scheme, improving data integration, and extending our phylodynamic model to analyze large genomic datasets and answer specific hypotheses. In parallel, other groups worldwide hacked away at similar challenges for real-time genomic epidemiology (Attwood *et al.*, 2022). While our primary goal was always to inform the public and government discourse on appropriate pandemic control measures, due to the slower progress of this project I think its most significant impact is in the lessons learned for the next emerging infectious disease threat.

One lesson, which pertains to all the chapters of this thesis, is that there is huge room for improvement in standardizing, linking, and sharing genome sequence data. In chapter 4, we describe how we were successful in generating large amounts of genome sequence data for SARS-CoV-2 in Switzerland thanks to collaborations with private diagnostics companies. We were also able to link these data to case-based data thanks to a collaboration with the Swiss Federal Office of Public Health. However, there were large delays in generating and sharing these data. Even as we were analyzing incoming data, we were still working to solidify and extend collaborations, develop robust procedures for materials and data transfer, and build a flexible infrastructure for data processing and storage, largely from scratch. Furthermore, we are only able to publicly share the sampling date and coarse geographic location for each sequence, as we lack a legal framework for sharing other data like patient age, sex, and vaccination status. This seems to be a common problem, as for Chapter 2 it was difficult to find appropriate datasets to test our method, which requires linked pathogen, host, and phenotype data. Thankfully, initiatives like BeYond-COVID[1] are working to ensure progress

1 https://by-covid.eu/about/

made during the pandemic in generating, linking, and sharing pathogen genome sequence data is not lost. Funding agencies are being made aware of this challenge and are called to support solutions (Committee on Data Needs to Monitor Evolution of SARS-CoV-2 *et al.*, 2020). A particularly promising approach, in my opinion, is the expansion of scientific staff roles at academic research institutions to aid in knowledge retention and facilitate longer-term collaborations across academia and public health.

A second lesson that I anticipate will continue to occupy us over the coming years is the importance of validation for phylogenetic and phylodynamic studies. As demonstrated in Chapters 3 and 5, pathogen genome sequences contain valuable information on transmission dynamics like case imports versus local transmission. This information is extremely important for inferring transmission histories and evaluating public health measures, especially in the absence of linked data on case exposure. However, observational studies such as ours must be carefully interpreted. We did our best to consider sampling biases, identifiability issues, and the inadequacy of our models to incorporate all the heterogeneity in real-life transmission dynamics, but these factors still lead us to draw cautious conclusions. Efforts to integrate additional case-based data in phylodynamic models (Andréoletti *et al.*, 2022; Lemey *et al.*, 2020), the involvement of modelers early on in genome sampling scheme design, and the availability of seroprevalence information and other epidemiological data to inform priors should help further mitigate these limitations. Beyond these efforts, I think it would be particularly interesting to explore more creative validation strategies. We focused on evaluating non-pharmaceutical interventions to combat COVID-19. Why not implement trials for these interventions analogous to clinical trial validation for the safety and efficacy of vaccines? One could imagine case/control studies for travel restrictions or mock contact-tracing experiments, for example, that include monitoring of the physical and mental health of participants. These follow-up data could complement first-line observational studies performed in the midst of the pandemic public health emergency.

Abraham Maslow said "I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail." In this thesis, I focused on a powerful tool for generating epidemiological insights - pathogen genome sequencing coupled with phylogenetic reconstruction. I applied this tool to identifying host genetic risk factors for infectious disease, quantifying epidemic transmission dynamics, and evaluating pandemic control measures. However, the biological complexities of infectious diseases mean

that genome sequencing is not a panacea for combatting them. Thus, I think it will be important moving forward to reach back into the public health toolbox to combine phylogeny-based approaches with other methods and other data sources. Coordinated serological and pathogen genome sampling schemes, integration of travel history or contact tracing data in phylogenetic reconstruction, and side-by-side modeling contrasting phylodynamic inferences with other mathematical epidemiolgical modeling approaches are examples of how genome-based and non-genome-based data and methods can complement one another. In conclusion, I see a growing role for pathogen phylogenies in public health - with key advancements coming from better integration with other data and methods.

BIBLIOGRAPHY

Andréoletti, J., Zwaans, A., Warnock, R. C. M., Aguirre-Fernández, G., Barido-Sottani, J., Gupta, A., Stadler, T., and Manceau, M. 2022. The occurrence Birth-Death process for combined-evidence analysis in macroevolution and epidemiology. *Systematic Biology*.

Attwood, S. W., Hill, S. C., Aanensen, D. M., Connor, T. R., and Pybus, O. G. 2022. Phylogenetic and phylodynamic approaches to understanding and combating the early sars-cov-2 pandemic. *Nature Reviews Genetics*, pages 1–16.

Committee on Data Needs to Monitor Evolution of SARS-CoV-2, Board on Health Sciences Policy, Health and Medicine Division, Board on Life Sciences, Division on Earth and Life Studies, and National Academies of Sciences, Engineering, and Medicine 2020. *Genomic Epidemiology Data Infrastructure Needs for SARS-CoV-2: Modernizing Pandemic Response Strategies*. National Academies Press, Washington, D.C. Pages: 25879.

Featherstone, L. A., Zhang, J. M., Vaughan, T. G., and Duchene, S. 2022. Epidemiological inference from pathogen genomes: A review of phylodynamic models and applications. *Virus Evolution*, 8.

Lemey, P., Hong, S. L., Hill, V., Baele, G., Poletto, C., Colizza, V., O'Toole, Á., McCrone, J. T., Andersen, K. G., Worobey, M., *et al.* 2020. Accommodating individual travel history and unsampled diversity in bayesian phylogeographic inference of SARS-CoV-2. *Nature Communications*, 11(1): 1–14.

Nature 2010. The human genome at ten. *Nature*, 464: 649–650.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., and Lappalainen, T. 2021. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1): 1–21.

CURRICULUM VITAE

PERSONAL DATA

Name	Sarah Ann Nadeau
Date of Birth	April 13, 1994
Place of Birth	Rochester, New York, USA
Citizen of	the United States of America

EDUCATION

2017 – 2018	Cornell University, Ithaca, New York, USA *Final degree:* M.Sc. in Biological Engineering
2012 – 2016.	Cornell University, Ithaca, New York, USA *Final degree:* B.Sc. in Biological Engineering; B.Sc. in Plant Science
2008 – 2012.	West Irondequoit High School, Rochester, New York, USA *Final degree:* High School Diploma

EMPLOYMENT

Dec 2021 – Apr 2022	Bioinformatics & Data Science Intern; ORISE Fellow *U.S. Centers for Disease Control and Prevention*, Atlanta, Georgia, USA
Jun 2018 – Aug 2018	Bioinformatics Intern *Prime Discoveries Inc.*, New York, New York, USA
Jun 2018 – Aug 2018	Fermentation Lab Operations Intern *Prime Discoveries Inc.*, New York, New York, USA

Jun 2017 – Aug 2017	Bioinformatics Intern *Amyris Biotechnologies*, Emeryville, California, USA
Jun 2016 – Dec 2016	Soy Breeding Co-op *Monsanto Company*, Huxley, Iowa, USA

PUBLICATIONS

ARTICLES IN PEER-REVIEWED JOURNALS

Chen, C., Nadeau, S., Topolsky, I., Beerenwinkel, N., and Stadler, T. 2022a. Advancing genomic epidemiology by addressing the bioinformatics bottleneck: Challenges, design principles, and a Swiss example. *Epidemics*, 39: 100576.

Chen, C., Nadeau, S., Yared, M., Voinov, P., Xie, N., Roemer, C., and Stadler, T. 2022b. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics*, 38: 1735–1737.

Yermanos, A., Hong, K. L., Agrafiotis, A., Han, J., Nadeau, S., Valenzuela, C., Azizoglu, A., Ehling, R., Gao, B., Spahr, M., et al. 2022. DeepSARS: simultaneous diagnostic detection and genomic surveillance of SARS-CoV-2. *BMC Genomics*, 23: 289.

Nadeau, S. A., Vaughan, T. G., Scire, J., Huisman, J. S., and Stadler, T. 2021b. The origin and early spread of SARS-CoV-2 in Europe. *Proceedings of the National Academy of Sciences of the United States of America*, 118.

Klaus, J., Meli, M. L., Willi, B., Nadeau, S., Beisel, C., Stadler, T., Egberink, H., Zhao, S., Lutz, H., Riond, B., et al. 2021. Detection and genome sequencing of SARS-CoV-2 in a domestic cat with respiratory signs in Switzerland. *Viruses*, 13.

Hodcroft, E. B., Zuber, M., Nadeau, S., Vaughan, T. G., Crawford, K. H., Althaus, C. L., Reichmuth, M. L., Bowen, J. E., Walls, A. C., Corti, D., et al. 2021. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature*, 595: 707–712.

Chen, C., Nadeau, S. A., Topolsky, I., Manceau, M., Huisman, J. S., Jablonski, K. P., Fuhrmann, L., Dreifuss, D., Jahn, K., Beckmann, C., et al. 2021. Quantification of the spread of SARS-CoV-2 variant B.1.1.7 in Switzerland.

Epidemics, 37: 100480.

Cabecinhas, A. R. G., Roloff, T., Stange, M., Bertelli, C., Huber, M., Ramette, A., Chen, C., Nadeau, S., Gerth, Y., Yerly, S., et al. 2021. SARS-CoV-2 N501Y introductions and transmissions in Switzerland from beginning of October 2020 to February 2021—implementation of Swiss-wide diagnostic screening and whole genome sequencing. *Microorganisms*, 9.

Scire, J., Nadeau, S., Vaughan, T., Brupbacher, G., Fuchs, S., Sommer, J., Koch, K. N., Misteli, R., Mundorff, L., Götz, T., et al. 2020. Reproductive number of the covid-19 epidemic in Switzerland with a focus on the cantons of Basel-Stadt and Basel-Landschaft. *Swiss Medical Weekly*, 150: w20271.

Nadeau, S. A., Roco, C. A., Debenport, S. J., Anderson, T. R., Hofmeister, K. L., Walter, M. T., and Shapleigh, J. P. 2019. Metagenomic analysis reveals distinct patterns of denitrification gene abundance across soil moisture, nitrate gradients. *Environmental Microbiology*, 21: 1255–1266.

Truhlar, A. M., Rahm, B. G., Brooks, R. A., Nadeau, S. A., Makarsky, E. T., and Walter, M. T. 2016. Greenhouse gas emissions from septic systems in New York State. *Journal of Environmental Quality*, 45: 1153–1160.

PREPRINT ARTICLES

Nadeau, S. A., Vaughan, T. G., Beckmann, C., Topolsky, I., Chen, C., Hodcroft, E., Schär, T., Nissen, I., Santacroce, N., Burcklen, E., et al. 2021c. Swiss public health measures associated with reduced SARS-CoV-2 transmission using genome data. https://doi.org/10.1101/2021.11.11.21266107

Nadeau, S., Thorball, C. W., Kouyos, R., Günthard, H. F., Böni, J., Yerly, S., Perreau, M., Klimkait, T., Rauch, A., Hirsch, H. H., et al. 2021a. A phylogeny-aware GWAS framework to correct for heritable pathogen effects on infectious disease traits. https://doi.org/10.1101/2021.11.22.21266687

Jahn, K., Dreifuss, D., Topolsky, I., Kull, A., Ganesanandamoorthy, P., Fernandez- Cassi, X., Bänziger, C., Devaux, A. J., Stachler, E., Caduff, L., et al. 2021. Detection and surveillance of SARS-CoV-2 genomic variants in wastewater. https://doi.org/10.1101/2021.01.08.21249379