Phylodynamic Patterns in Pathogen Ecology and Evolution

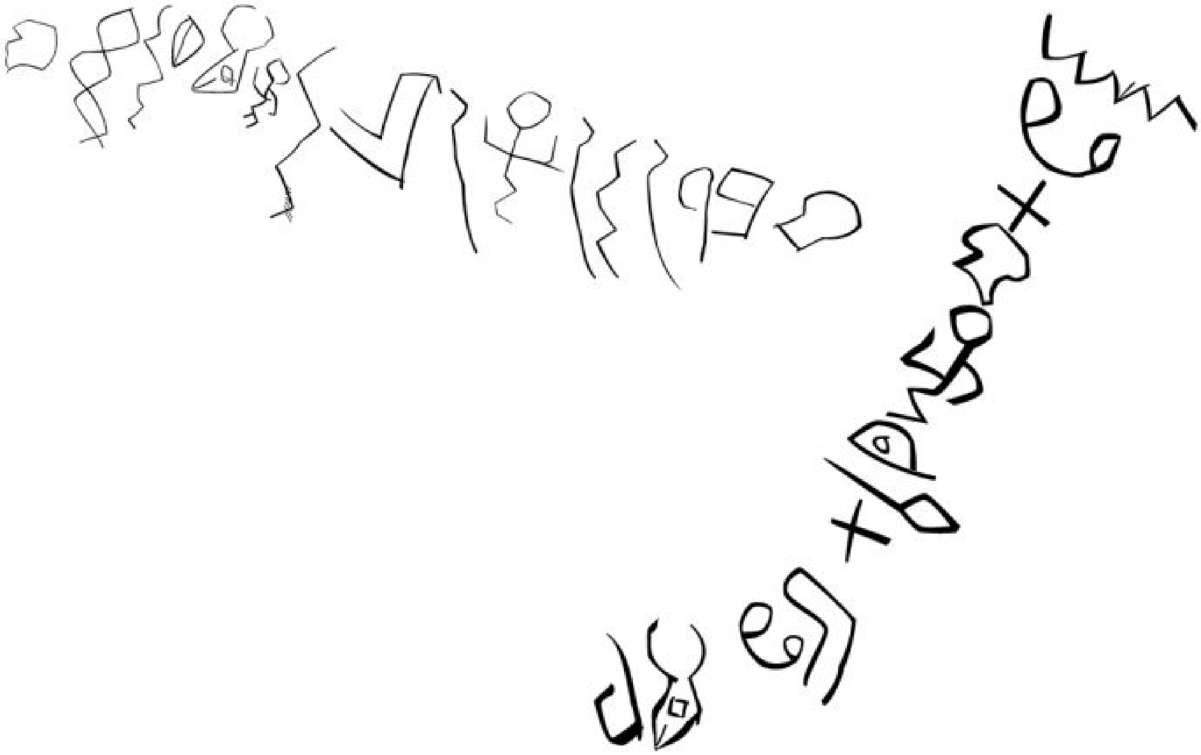
by

Daniel Zinder

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2015

Doctoral Committee:

Professor Mercedes Pascual, Co-Chair
Assistant Professor Stephen A. Smith, Co-Chair
Associate Professor Aaron A. King
Assistant Professor Adam S. Lauring
Associate Professor Patricia J. Wittkopp



A common ancestor, proto-Sinaitic (or proto-Canaanite) script (16th-17th century BC). One of the earliest known relatives of the common ancestor of alphabetic scripts, including amongst many others English, Hebrew, and Bengali. Principles of inheritance and its correlation with phenotypic similarity, have long been used outside the context of biology in philological research and are likely to have inspired early ideas of evolution in biology.

Acknowledgments

I have been very fortunate to conduct research with highly skilled and insightful colleagues and with inspirational mentors. I would first like to thank my dissertation committee: Aaron King, Stephen Smith, Adam Lauring, Patricia Wittkopp and my advisor Mercedes Pascual. Their guidance and instruction have profoundly shaped and improved my dissertation. They have likewise influenced my approach to research, and I am grateful for their stimulating questions and insightful suggestions over the past several years. It has been the greatest privilege to have the opportunity to learn science from one of the most inspiring people in the field, Mercedes. I have been very fortunate to work with someone who combines an incredible talent for research and science, together with radiating enthusiasm and an enticing personality. Our lab, has been my home away from home for the last several years, and I am thankful to Mercedes for generating such a stimulating environment for her students. I am also deeply grateful for her support, patience and encouragement. I would also like to give special thanks to my coauthors within our lab: Bob Woods, Ed Baskerville, Trevor Bedford, Manojit Roy, Molly Rorick (and Maria Riolo). None of the chapters here would have existed without their teamwork and help, and I am thankful to be able to work with such professional, ethical and gifted scientists.

Many thanks are appropriate for specific chapters. Mercedes originally posed the idea of incorporating evolution into the study of 'strain theory' and studying how

phylogenetic patterns may correspond to its different regimes (Chapter 2). She also forged the collaboration with Sunetra Gupta, who commented and improved on the paper considerably. Trevor was first to suggest that the results should be framed within the context of two basic components of evolution: mutation and competition, and is also responsible for giving me access to his well-written code.

Chapter 3 was inspired by this same major theme, but Bob Woods has been the motivation behind many of the types of dynamics we eventually decided to choose. My knowledge of rotavirus comes directly from his guidance and literature references. Although not in this text, a stronger connection to data in this chapter, will be achieved with the aid of Maria who has been a recent and incredible addition to our lab as a postdoc.

My exposure to phylogenetic inference (Chapter 4 and 5) has been facilitated by Trevor Bedford. Trevor has a canny talent for neat science generated by few magical lines of code in Mathematica or Java. In Chapter 4, originally the intent was to study evolutionary patterns of influenza through changing weather, a project which I may yet re-approach. In its current form, an analysis of the patterns of migration was stimulated by ideas inspired from both Robert Woods and my committee advisor Stephen Smith. Bob has also been the big driver behind our project on the rotavirus population described in Chapter 5. The machinery for inference has been generated with the help of numerous lines of elegant code and theoretical support by Ed Baskerville. Ed has a talent for computational science which exceeds anything I have seen.

Molly and Yael represent collaborations on the part of my work which was not included in this dissertation, yet are part of my PhD nonetheless. I'm grateful to Yael for the many methods and tools she has taught me, and for the initiation into the field of theoretical disease ecology. Molly has introduced me to the molecular world of *var* genes, and I hope our collaboration will continue and flourish. Similarly, I would like to thank Prof. Oliver He and Dr. Andrew Hodges for a great and highly valuable experience during my rotation project.

I would also like to thank all my friends and colleagues in the lab throughout the years. I'm listing Manojit here first for both guiding me scientifically and encouraging my bottom up understanding of knowledge from basic principles, but also as a true friend in times of need. Not listed before, I'm grateful for many stimulating conversations with my lab-mates Qixin He, Xiangjun Du, Richard Zinck, Bobby Reiner, Tony Golubski, Andrew Strayer, Andres Baeza, Kevin Baker, Pamela Martinez and Mauricio (Oscar) Santos Vega, all of whom I have learned a lot from, and hope to continue and do so in years to come.

I am also grateful for the support of people and staff in the Bioinformatics, PIBS and Ecology departments, who have made arranging research, travel and collaboration much easier, including Julia Eussen, Sonja Botes, Katia Kitchen and Michelle Melis. I would also like to thank Margit Burmeister and Dan Burns for their help and academic guidance throughout my graduation process. I am thankful for my core financial support through Mercedes' appointment at the Howard Hughes Medical Institute, and to Rackham and the Bioinformatics department for providing additional funding.

Finally, I would like to thank my family, with a mix of English, Hebrew and some Bengali (which I do not speak): Mom, אבא היקר, Michael and דפנה, my dearest grandmother who helped raise me and brought together a fantastic family שתי מלחמות עולם לא עצרו אותה והכוחות שלה איתנו עד היום. মিষ্টি সুন্দরী Sudeshna who is always সবসময় ঠিক, but more importantly, very dear to me.

Table of Contents

List of tables

List of figures

Chapter 1. Introduction

1.1 Background

Phylodynamics is the term used for the interaction of pathogen evolution with the population dynamics of infectious diseases, in which immunological processes play a major role (B. T. Grenfell 2004). The study of phylodynamics has been so far most profound in the context of RNA viruses such as HIV, dengue, and influenza, but is of importance to many other pathogens. Unique to RNA viruses is the growing availability of full, short (kilo-bases) and sequentially sampled sequence data on timescales ranging from days to decades. The rapid evolution of RNA viruses makes the timescales of evolutionary, epidemiological and immunological processes comparable. As such, these three processes should be considered together in modeling and in decision making. Such phylodynamic considerations may involve a single patient as is the case of HIV where viruses evolve within a single chronic infection (Price et al. 1997), or whole populations and even global health concerns, as demonstrated by recent epidemics such as those of MERS corona virus, in which the identity of the zoonotic source of the disease in bats and camels has been of major importance (Annan et al. 2013; Reusken et al. 2013).

Besides theory, the study of phylodynamics further involves inference based on phylogenetic trees together with epidemiological data. These two data sources are used together to understand transmission processes and infection dynamics. One of the classically studied diseases in epidemiology has been measles, which also offers a good baseline model in the context of phylodynamics. The *Measles virus* is a single stranded RNA virus. Phylogenetic inference has identified its likely split from the cattle *rinderpest* virus as early as the 5th century (Furuse, Suzuki, and Oshitani 2010) with a large degree of uncertainty (Wertheim and Kosakovsky Pond 2011) because of the possible effect of complex spatial processes and purifying selection. Today, humans are the only known reservoir of the measles virus, which has a single serotype, meaning that following infection with any measles virus immunity is equal for all strains (Rota et al. 2011). Immunity to measles is strong and lifelong, and the virus has a failed to evolve to escape both vaccination and acquired immunity. As a consequence of this immunity, the phylogenetic tree of the measles surface glycoprotein hemagglutinin is thought to represent neutral processes involving changes in the population size of the virus with dispersal and epidemiological processes only, in the absence of directional evolution. With a constant population size, a neutral phylogenetic tree has a constant rate of coalescence or continuous branching through time (Figure 1.1A).

Recently, new conceptual models have contributed to the unification of epidemiological and population dynamics. This progress allows for example for the calculation of the expected rate of coalescence given a compartmental model describing epidemiological processes (Volz 2011; T. Stadler and Bonhoeffer 2013). A simple example of a relationship between epidemiological and coalescence processes is evident in the phylogenetic tree of HIV, which in contrast to measles shows deep coalescence of recent samples reflecting the epidemic growth of the virus (B. T.

Grenfell 2004) (Figure 1.1B). With exponential growth, more recent samples of the virus in a larger population are less likely to have immediate kin that was sampled. In contrast with the population level tree of HIV, within the human host at a certain time point of infection the phylogenetic tree of the virus is spindly, the consequence of multiple selective sweeps of the virus as it evades immunity (Price et al. 1997). HIV-1, the more widely distributed HIV type, is thought to have crossed the species barrier to humans from simian hosts on at least three times giving rise to the three groups of the virus M, N and O (Thomson, Pérez-Álvarez, and Nájera 2002).

Similar to HIV, the *dengue virus* was also introduced to the human population on multiple occasions, giving rise to at least four distinct serotypes (Messina et al. 2014) (Figure 1.1C). Typically protection against one viral serotype provides some level of *cross-immunity* against another. However, interaction between the different dengue serotypes (Kliks et al. 1989) has an unusual feature termed *antibody dependent enhancement,* in which exposure to one serotype can lead to more severe disease in a subsequent infection by a different serotype. Neutralizing antibodies against the first infection are not fully efficient against the second infection. As a consequence, macrophages clearing the infection are compromised, and the virus replicates within them leading to severe hemorrhagic disease.

In spite of significant progress made into the phylodynamics of these and several other viral pathogens many challenges remain. First in organisms which despite their importance have not been a the focus of sufficient research, such as rotavirus, and in addition in pathogens such as influenza where the full scope of their complex global dynamics and the role of immunity in them has not been fully established.

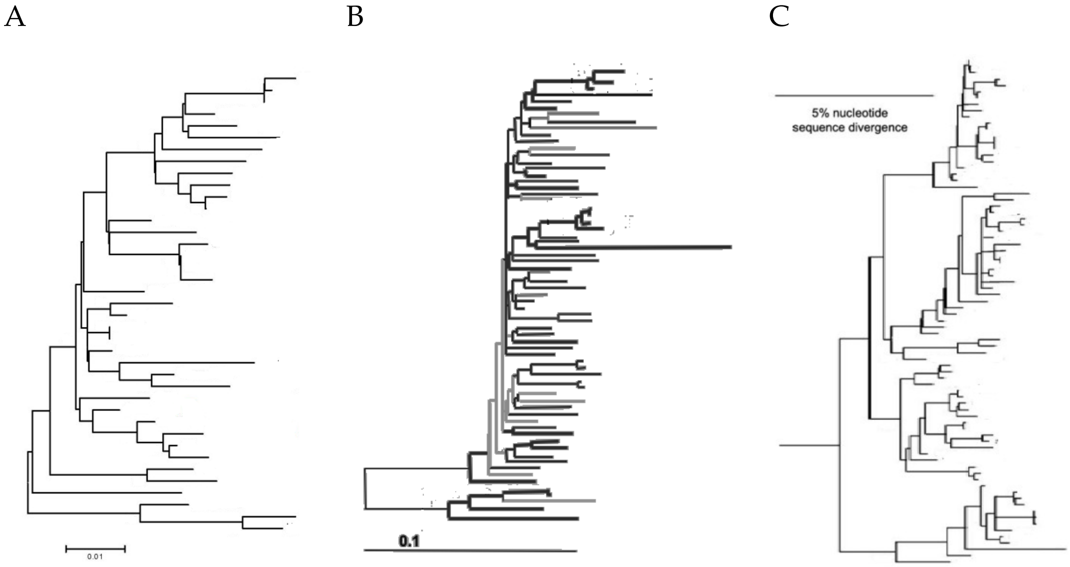


Figure 1.1 Phylogenetic trees of RNA viruses
A Phylogenetic tree of *Measles virus* hemagglutinin (adapted from (Schierup et al. 2005))
B Phylogenetic tree of HIV-1B pol protein in Europe (adapted from (Salemi et al. 2008))
C Phylogenetic tree of *Dengue virus* (adapted from (Weaver and Vasilakis 2009)

1.2 Phylodynamic patterns in rotavirus and influenza

Rotavirus A (RVA) and influenza are the main study organisms in this thesis, and were chosen for being amongst the ten leading causes of morbidity and mortality from infectious diseases worldwide. In addition, the high prevalence of both viruses, makes the likelihood of non-trivial interactions with the immune system, involving more than fully naïve hosts, much more likely. Influenza A/H3N2 is amongst the best sampled RNA viruses, providing one of the study organisms most used for the development of phylodynamic methods. In contrast, despite its importance as the major cause of diarrheal disease mortality and morbidity in infants and children, relatively little phylodynamic work has been conducted in rotavirus, work that is in great need following the current onset of a global vaccination campaign. Both viruses share many similarities and some major differences, which are further discussed below and in the respective chapters.

Influenza is single stranded RNA virus encoding 11 proteins (Bouvier and Palese 2008) and is classified into types A, B and C, with type A being the most pathogenic and responsible for an access mortality estimated at 1-5 million people since 1968 (WHO 2013). The major targets of humoral immunity against influenza A are its envelope glycoproteins, hemagglutinin (HA) and neuraminidase (NA). Variations among these proteins form the basis of classification of influenza A into subtypes H1N1, H2N2 and H3N2 etc. Multiple zoonotic introductions of influenza A subtypes to the human population have taken place, with H3N2, H1N1 (pre- and post-pandemic) being the most prevalent subtypes whose continuous endemic circulation has lasted decades. Influenza A (H3N2) emerged in 1968. Since then, it has continually circulated in the human population, exhibiting rapid geographic spread and turnover rates. Globally, human influenza A (H3N2) populations diverged only to within 2-8 years in their past (Figure 1.2) (Andrew Rambaut et al. 2008b; Bedford et al. 2010a; Bedford, Cobey, and Pascual 2011b).

Circulating influenza A lineages represent recent invasions from zoonotic reservoirs, while the animal source of influenza B is unknown (Osterhaus et al. 2000). In humans, influenza subtypes A/H3N2, A/H1N1, and two lineages of influenza B, display different rates of antigenic evolution. Correspondingly, the faster evolving H3N2, has the lowest genetic diversity, followed by H1N1 and the most genetically diverse influenza B lineages. This higher genetic diversity with a lower antigenic mutation rate, corresponds to a transition from dynamics driven by positive selection, to more neutral dynamics of evolution (Figure 1.2) (Ferguson, Galvani, and Bush 2003b; Zinder et al. 2013; Bedford et al. 2014).

Like influenza A, rotavirus circulates extensively in the human population, infecting most children below the age of five and causing over 400,000 deaths annually (Tate et al. 2012). The rotavirus genome contains eleven double stranded RNA genome segments encoding 12 proteins (Estes and Kapikian 2013) compared to 11 in influenza A (Bouvier and Palese 2008). The main classification of RVA is based on the two outer surface proteins, VP7 which encodes the G protein, and VP4 which encodes the P protein (Jelle Matthijnssens et al. 2008). Circulating rotavirus lineages represent recent invasions from zoonotic reservoirs (Osterhaus et al. 2000). At a segment level, each segment type (i.e. G1, G12 etc...) corresponds to a monophyletic cluster (Figure 1.3), and is thought to reflect the cross-species transmission of the serotype into humans from an animal source. There is a certain degree of antigenic change within individual G types (i.e. within G1) through mutation (Arista et al. 2006; S. M. McDonald et al. 2011).

Evolutionary patterns of RVA differ considerably from that of influenza. In contrast to influenza, rotavirus genotypes frequently *reassort*. However, in spite of the high levels of confection and of *reassortment* not all genotypes are as common (Jelle Matthijnssens and Van Ranst 2012). Serotypes of rotavirus, exhibit extensive variability in both time and space. And in contrast to the within-subtype dynamics of influenza A, rotavirus serotypes do not necessarily fully replace their predecessors, but rather may become locally dominant for several years, before becoming infrequent (M. H. Afrad et al. 2013; De Grazia et al. 2014; Hasing and Trueba 2009) and re-emerging years or decades later (S. M. McDonald et al. 2012). Thus, at the global level, rotavirus presents a diverse set of coexisting genotypes whose local representation and abundance changes locally in a dynamic fashion.

An additional main distinction between rotavirus and influenza A is found in their population dynamics and epidemic nature. Whereas influenza A exhibits epidemics when a new cluster emerges and leads other types to extinction, rotavirus can display sustained levels of infection (M. H. Afrad et al. 2013) despite dynamic changes in serotypes (Figure 1.4).

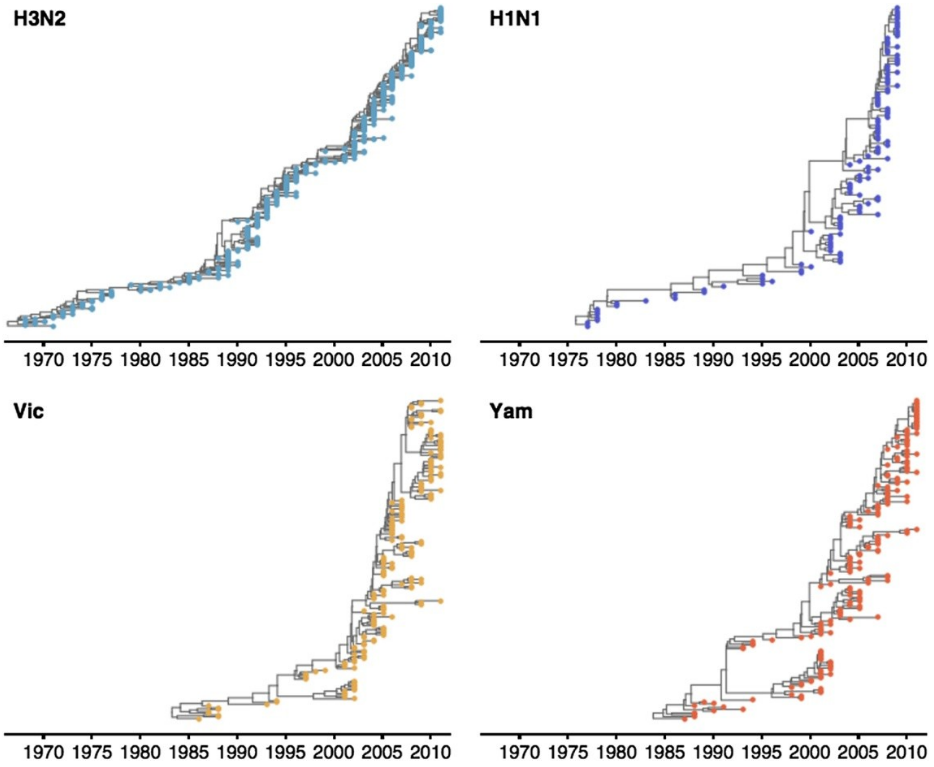


Figure 1.2 Time-resolved phylogenetic trees of A/H3N2, A/H1N1, B/Vic and B/Yam viruses

The maximium-clade credibility (MCC) tree is shown for each virus. These trees show genealogical relationships, so that branches are measured in terms of years rather than substitutions. From: Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, McCauley JW, Russell CA, Smith DJ, Rambaut A: **Integrating influenza antigenic dynamics with molecular evolution**. *Elife* 2014, **3**:e01914.

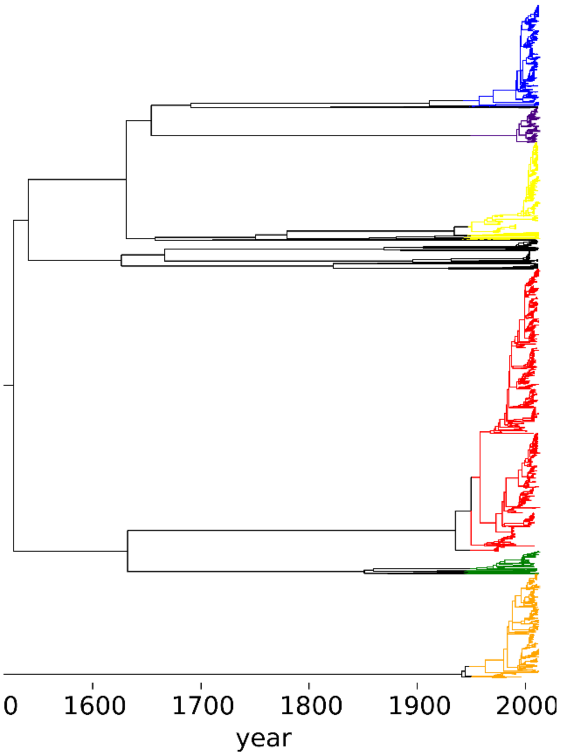


Figure 1.3 Time-resolved phylogenetic tree of rotavirus A, VP7 segment (G protein) in humans

A single representative phylogeny of VP7 from the posterior distribution of trees, generated from a sample of 1911 annotated GenBank sequences, and color-coded by serotype (see Chapter 3 for further details on this figure).

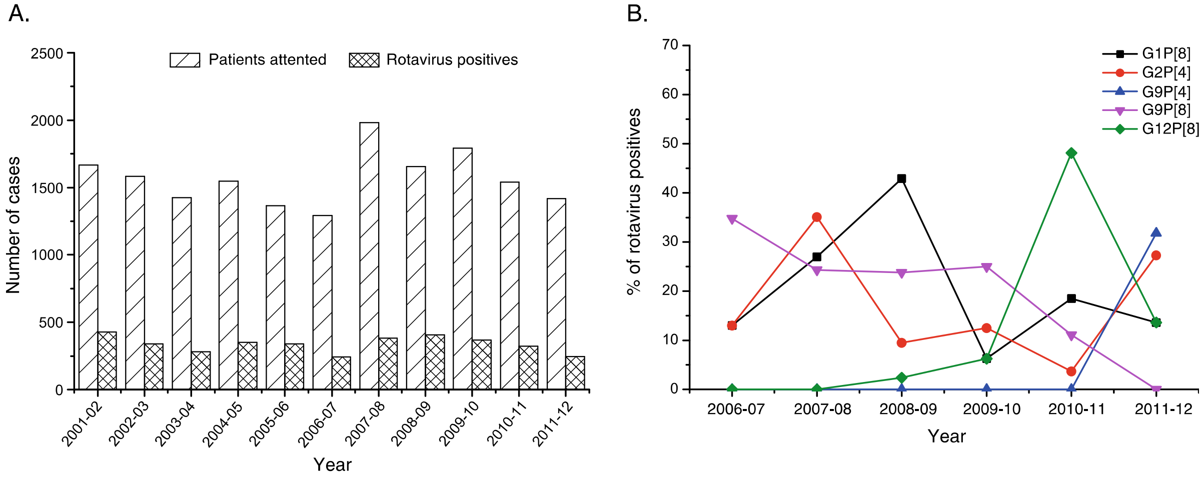


Figure 1.4 Prevalence and genotype distribution of rotaviruses in Matlab, Bangladesh
Prevalence and genotype distribution of rotaviruses in Matlab, Bangladesh. **A** Number of diarrhea patients attended and number of cases tested rotavirus positive, June 2001-May 2012 **B** Temporal changes in the distribution of major rotavirus genotypes, June 2006-May 2012. From: Afrad, Mokibul H., Zahid Hassan, Saiada Farjana, Sayra Moni, Subarna Barua, Sumon K. Das, Abu Syed Golam Faruque, Tasnim Azim, and Mustafizur Rahman. Changing profile of rotavirus genotypes in Bangladesh, 2006–2012 BMC infectious diseases 13, no. 1 (2013): 320.

1.3 Phylodynamic patterns of multiple serotypes

Strain formation and co-existence has practical implications with respect to understanding the importance of host population immunity in limiting the invasion potential of pathogens such as avian or swine influenza, and for understanding the intensity and mechanisms of interaction between circulating strains, such as influenza subtypes. Phylodynamic work has so far mainly concentrated on the emergence of serotypes through cross-species barriers and on the replacement of serotypes through directional selection. Much less, yet valuable research (Bedford et al. 2014), often with a more epidemiological perspective (J. R. Gog and Grenfell 2002; Simonsen, Reichert, and Miller 2004) has been conducted on the interaction and co-existence of multiple serotypes, while additional research has been done into patterns of immunity to different serotypes within the population (Fonville et al. 2014; Gladstone et al. 2011a). Several immune mediated patterns of interaction and co-existence between pathogen strains have been suggested in the absence of evolution (S. Gupta 1998; J. R. Gog and Grenfell 2002). While ecological forces of interaction between species and their relationship with evolutionary processes have mainly been suggested in other organisms (Nosil, Harmon, and Seehausen 2009; Davis et al. 2014). The first suggestion for relationships between phylogenetic trees and species interaction was posited by Darwin himself in an important example of such an interaction.

In one of his important hypothesis Darwin suggested (Darwin and Britannica 1872) a relation between phylogenetic and ecological similarity. Darwin specifically hypothesized that closely related species tend to possess similar niches and hence perform similarly under the same environmental conditions. This first hypothesis, led to a subsequent hypothesis termed *Darwin's naturalization hypothesis*, which suggests that introduced species are more successful in communities in which their close relatives are absent. This enhanced success will be the case when the close relatives of the invading species compete for the same resources. Under the alternative outcome, an invading species will occupy the same niches in the environment as existing species, and to do so, they will be similar, and hence more closely related to the existing species occupying these similar niches. Work on testing these alternative hypotheses has led to conclusions favoring both alternatives in a case and scale dependent manner (three of many examples can be found in (Bezeng et al. 2015; Schaefer et al. 2011; Lebrija-Trejos et al. 2010).

The immune system of the host at the population level can be thought of as a niche, or resource axis, which pathogens use and compete for. The depletion of susceptible hosts through the generation of immunity following an epidemic, makes the generation of immune niches a dynamical and emergent process. The effect of such dynamics on the generation of serotypes and strains has been explored in several studies stemming from what has been termed 'strain theory' (S. Gupta 1998; S. Cobey and Lipsitch 2012; Caroline O. Buckee et al. 2011). Recker and Wikramaratna et al. (Recker et al. 2007; Wikramaratna, S. P et al. 2012) suggested for example that influenza A/H3N2 patterns of antigenic waves of dominance, in which viruses circulating at a given time point are replaced by a different antigenic type, could be driven by the opening in a frequency-dependent manner of unexploited immune niches in the human population, rather than by the generation of new antigenic types. In their model, circulating strains are selected from an existing pool of limited variation of the virus whose possibilities are continuously present at low abundance (given that the model does not incorporate the possibility of stochastic extinction). However, their work did not consider evolution and extinction in an explicit manner. .

Alternative non-immune based evolutionary and ecological processes such as geographical isolation and host tropism can also generate discrete pathogen strains (Bourhy et al. 2008; Holmes and Zhang 2015; Parrish, Murcia, and Holmes 2015), where reduced cross-immunity could be thought of as a by-product of these processes. Immune and speciation processes can work in tandem to maintain strain structure through *epistasis* (Heiman et al. 2008). Tree patterns generated by these two processes are not guaranteed to be different once strains are formed, and in both cases strains correspond to deep independent branches. It is possible that immune neutral and immune -mediated strain structuring mechanisms could be differentiated by identifying immune and non-immune related signatures of selection and of recombination, however further work is necessary in this field.

The interplay between population dynamics with transmission and with the generation of immunity is much less understood when processes of recombination between strains are included. With recombination, phylogenetic trees become reticulated networks, a tree representation in this case is either an approximation representing a unit of the genome, such as a single protein, or an aggregate, intended for instance to represent non-hybridizing species. Recombination involves the swap of genome parts, while viral reassortment involves the swapping of entire viral genome segments. Sexual reproduction is a major source of evolution and evolutionary change in the context of pathogen *antigenic* dynamics. A distinction is required between sexual reproduction, which combines existing antigenic variation in new forms, and one that involves the creation of antigenic novelty. The important role of recombination in generating antigenic novelty and crossing species barriers has been the main focus of phylodynamic studies. Examples in the context of influenza can be found in the work of (Lindstrom et al. 1999; Simonsen et al. 2007; Ince et al. 2013; Martha I. Nelson et al. 2008; Neverov et al. 2014; G. J. D. Smith et al. 2009).

1.4 Overview of the thesis

My thesis is divided into two sections, the first relating to *theory* and the second to *inference*. These two sections relate to two approaches, which have been used to answer different questions in the field of phylodynamics. In the first two research chapters, I focus on the qualitative features necessary to generate alternative dynamics of stable co-existence and of replacement in viral pathogens, considering both competitive interactions and different mechanisms of evolution through mutation and through recombination. To investigate this extension of existing theory, I use agent-based models and examine how simple sets of dominant processes structure pathogen diversity and generate alternative phylodynamic patterns. In the *inference* part of the thesis, I investigate the ways in which the rotavirus and influenza populations migrate and evolve through mutation. To do so, I use a Bayesian inference approach which I developed (*SeasMig*) and is based on stochastic mapping and ancestral state reconstruction of seasonal migration and mutation processes described in detail in the appendix.

Chapter 2 specifically considers the minimum set of mechanisms required for generating spindly A/H3N2 influenza like trees, and extends the theory by considering competition for hosts in the context of explicit evolution through mutation. Explaining the low genetic diversity observed in A/H3N2 influenza has proven a difficult problem (Koelle et al. 2006; Wolf et al. 2006; Ferguson, Galvani, and Bush 2003), which has generated several alternatives hypotheses and theoretical models. With a model which samples phylogenetic trees by tracking transmission events, I find the minimum conditions in terms of competition and mutation under which trees are spindly and flu like. The chapter also describes the type of trees corresponding to immune-mediated strain structure together with the corresponding qualitative levels of competition, mutation and selection forces observed. Alternative regimes of antigenic neutrality and continuous diversification, corresponding to low and high antigenic mutation rates, respectively, are also described.

In Chapter 3, I expand on the work done in the previous chapter to consider key phenomena characteristic of many pathogens such as rotavirus: the introduction of antigenic novelty from an animal source, reassortment between strains, and an additional component of immunity, which is independent of the infecting serotype and depends only on the number of previous infections. Consideration of reassortment together with antigenic novelty, results in antagonistic phenomena driving both the sharing of antigenic novelty and the partitioning of available antigenic alleles between serotypes. In such a way, I describe a phylodynamic regime that differs from those previously identified with theory focused exclusively on mutation rather than reassortment. This work underscores the important interplay of immigration and reassortment , which motivates in part the more empirical study of Chapter 5. Both chapters that follow give a special place to the subject of migration.

Chapter 4, an inference chapter, attempts to explain how A/H3N2 influenza migrates in light of seasonal incidence patterns. This works attempts to capture the importance of the 'pull' of influenza migrants through necessary conditions at the destination, compared to the prevalence and 'availability' of migrating viruses at the source.

In Chapter 5 tools used for the mapping of stochastic processes along trees, are used to distinguish between three different mechanisms for rotavirus adaptation in response to routine vaccination. These mechanisms of adaptation include migration, mutation and changes in the frequency of circulating serotypes. This method is applied to the question of what are the main mechanisms behind genetic changes in rotavirus in the US after the introduction of vaccination.

Chapter 6, presents a short conclusion and some future directions.

Section I - Theory

Chapter 2. The role of competition and mutation in shaping antigenic and genetic diversity in H3N2 influenza

2.1 Abstract

Influenza A (H3N2) offers a well-studied, yet not fully understood, disease in terms of the interactions between pathogen population dynamics, epidemiology and genetics. A major open question is why the virus population is globally dominated by a single and very recently diverged (2-8 years) lineage. Classically, this has been modeled by limiting the generation of new successful antigenic variants, such that only a small subset of progeny acquire the necessary mutations to evade host immunity. An alternative approach was recently suggested by Recker et al. in which a limited number of antigenic variants are continuously generated, but most of these are suppressed by pre-existing host population immunity. Here we develop a framework spanning the regimes described above to explore the impact of rates of mutation and levels of competition on phylodynamic patterns. We find that the evolutionary dynamics of the subtype H3N2 influenza is most easily generated within this framework when it is mutation limited as well as being under strong immune selection at a number of epitope regions of limited diversity.

2.2 Introduction

Influenza viruses are classified into types A-C, among which influenza A is the most pathogenic. These viruses cause between a quarter to half a million deaths worldwide (WHO 2009) and tens of thousands of deaths in the US during annual epidemics (C.D.C 2010). The economic burden of seasonal influenza in the US is estimated at more than ten billion dollars in healthcare costs alone (Walsh and Maher 2011).

The major targets of humoral immunity against influenza A are its envelope glycoproteins, hemagglutinin (HA) and neuraminidase (NA); these form the basis of its crude classification into subtypes H1N1, H2N2 and H3N2 etc. Since its emergence in 1968, influenza A (H3N2) has continually circulated in the human population. The phylogeny of its HA protein (Figure 2.1) shows a distinctive 'cactus-like' shape with a narrow, usually single-trunked, tree (Fitch et al. 1997; Bush, Bender, et al. 1999). The 'narrowness' of the tree is derived from the fact that contemporaneous H3 proteins share a single common ancestor 2-8 years in the past (Andrew Rambaut et al. 2008a; Bedford, Cobey, and Pascual 2011a). This short time is unique to H3N2 given its global spread and its high prevalence and incidence (Bedford, Cobey, and Pascual 2011a).

The classical view of influenza evolution is one of antigenic drift (Gerhard and Webster 1978; Hay 2001; Nakajima et al. 2005) in which antigenic change continually and gradually accumulates in the virus through the influence of selection by way of changes to the HA and NA proteins. By itself, the 'cactus-like' structure of the A/H3N2 phylogenetic tree suggests the presence of adaptive evolution (Bedford, Cobey, and Pascual 2011a) and several studies have provided evidence for positive selection (Bush, Bender, et al. 1999; Fitch et al. 1991; Suzuki 2008; Wolf et al. 2006; Lindstrom et al. 1999). However, it is difficult to explain the limited standing diversity of influenza (Tria et al. 2005; Ferguson, Galvani, and Bush 2003b), and the empirical evidence for discontinuous antigenic change (Koelle et al. 2006; D. J. Smith 2004), under a general antigenic drift framework. Multiple epidemiological hypotheses have been advanced to reconcile these observations with a process of continual antigenic divergence including short-lived strain-transcending immunity (Tria et al. 2005; Ferguson, Galvani, and Bush 2003b; Andreasen and Sasaki 2006), epochal or punctuated evolution (Wolf et al. 2006; Koelle et al. 2006), trait-space reduction (Kryazhimskiy et al. 2007) and canalized evolution (Bedford, Rambaut, and Pascual 2012a).

A competing hypothesis advanced by Recker et al. (Recker et al. 2007) eschews the paradigm of antigenic drift, instead considering that, owing to functional constraints on the defining epitopes, the virus population is limited phenotypically to a restricted set of antigenic types. Antigenic types replace each other with waves of dominance resulting from frequency-dependent immune mediated selection as “niches” in antigenic space are dynamically generated and are exploited by the existing virus population. In its original implementation, this model assumes that all antigenic types remain present in the population in low frequencies, as an approximation to the idea that they can be generated by mutation from preexisting strains at a sufficient rate as not to limit the emergence of a type favored by selection. Thus, the model describes in practice a case where influenza outbreaks are caused by host immune selection in a manner that is not limited by the rate of antigenic mutations. Although patterns of turnover are consistent with those observed for H3N2, it is not clear whether the characteristic phylogenetic trees can be generated by this model.

Here, we have attempted to resolve this question using a large-scale individual-based simulation of epidemiological and evolutionary dynamics that allows the complete phylogenetic tracking of a virus population characterized by defined repertoires of polymorphic epitopes. Our model is based on the multi-locus structure employed by (Recker et al. 2007) with host immunity operating at an epitope-specific level. When contacted by a virus, a host's risk of infection is determined by the number of alleles/epitopes recognized by its immune system. We also introduce the possibility of a long-lived strain-transcending component to the model. Thus, competition between strains is determined both by the number of shared epitopes and a variable level of generalized immunity. Our model differs in this regard from that of Recker et al. (Recker et al. 2007) which does not permit full cross-protection except in the case of having experienced the exact same combination of epitopes, a feature that implicitly accounts for the effect of highly variable epitopes unique to each strain.

This model structure allows us to make inferences about the roles of mutation and competition in a more general context. Models of antigenic dynamics tend to polarize between those in which the availability of antigenic types dictate the dynamics (Wolf et al. 2006; Koelle et al. 2006; Russell et al. 2008; Kuiken 2006), and those where host immune-mediated selection is the only driver (S. Gupta 1998; C.O. Buckee et al. 2008). We refer to the latter regime, where antigenic change is constrained by host population immunity, as selection limited, whereas the former, in which the availability of antigenic mutations pose the rate limiting step, is described as mutation limited. The approach we take in this paper offers a tool for locating influenza on this continuum and would easily generalize to other antigenically diverse infectious diseases.

2.3 Results

We use an individual-based SIR model, explicitly tracking the chains of infection of viral lineages as well as the antigenic phenotype of every virus in the population (Figure 2.2A). Thus, our model explicitly tracks viral genealogy rather than conducting phylogenetic inference, and therefore does not include any genotype to phenotype map.

A host population of constant size N is simulated with an equal birth and death rate of μ. Infected hosts randomly contact other hosts in the population at rate β. The risk of infection is based on crossimmunity interactions based on their previous unordered infection history (May and Nowak 1994; Martin A. Nowak and May 1994). Following infection with a virus displaying a specific antigenic phenotype, hosts acquire partial protection against reinfection with viruses sharing epitopes, and full protection against the exact same strain. There is no super-infection in the model; while infected, hosts are protected from co-infection with other strains. Hosts recover from infection at a constant rate ν. (see Methods for details of the model).

2.3.1 Antigenic-Phenotype Replacement and Single Phenotype Dominance in an Evolution-Free Framework

To explore the epidemiological dynamics of our model in isolation, we implemented a parameterization lacking mutation, in which extinction was preempted by maintaining at least one carrier for each antigenic-phenotype. Different colors (Figure

2.3) represent the prevalence of different antigenic-phenotypes. Here, the antigenic repertoire is derived from combinations of variants at 5 distinct epitopes (see Methods for full description of epidemiological parameters).

A mutation-free model can result in different alternative dynamics based on model parameters ranging from stable coexistence of completely discordant antigenic-phenotypes to the successive replacement of strains through chaotic or cyclic behavior (S. Gupta 1998). Not surprisingly, our model implementation with no explicit evolution also generates these waves of replacement (Figure 2.3), suggestive of H3N2 influenza as proposed by Recker et al. (Recker et al. 2007).

2.3.2. Phylodynamic Patterns across Different Mutation Rates

We examined the effects of mutation at different rates on the resulting phylodynamic patterns of the virus by seeding the population with a single strain and tracking antigenic and evolutionary changes. We measure diversity π as the average time separating two randomly selected contemporaneous viruses since their divergence from a common ancestor. Because branch lengths in our genealogies are measured in years, the resulting diversity is also measured in years.

In the absence of antigenic mutation, only a single strain persists, experiencing transient oscillatory dynamics between near extinction, and endemic equilibrium conditions (Figure 2.4A). As all viral traits are equal, there are no selective forces and the observed phylogeny and coalescence rates can be directly related to prevalence and incidence (Volz 2011). This yields random coalescence within contemporaneous viral lineages and a large associated pairwise genetic diversity (π=30±12 years; mean±std across 5 simulations).

For low mutation rates (Figure 2.4B) the introduction of new mutations is the critical determinant of strain dynamics. Each new variant outcompetes the one that came before, resulting in a spindly phylogenetic tree and therefore low diversity (π=5.7±0.8 years). Temporally adjacent strains are antigenically similar, rather than discordant, forcing strong competitive exclusion and single strain dominance (ε=0.93±0.02), where ε is the proportion of infections caused by the most common strain. An increase in mutation rate (Figure 2.4C) leads to deeper branches with a corresponding increase in phylogenetic diversity (π=220±100 years) and more pronounced antigenic divergence. Here, the population dynamics are ruled by the endemic or cyclic behavior of discordant antigenic sets. The emergence of new intermediate antigenic types is suppressed by competition from the two prevalent strains (S. Gupta 1998).

At a relatively high mutation rate (Figure 2.4D), we approach population dynamics similar in appearance to those of the mutation-free model (Figure 2.3). Diversity is high (π=120±30 years), with deep yet occasionally coalescing branches. In general, a threshold exists at which mutation overwhelms selection resulting in a population drifting away from the fittest genotype (M.A. Nowak 1992). For sufficiently high rates of antigenic mutation, all antigenic types reach near equal frequency in the population (Figure 2.4E). On a population scale, the high mutation rate weakens frequency-dependent selection and results in the breakage of antigenic strain structure; antigenic types do not cluster across the genealogy. The loss of selection forces breaks down phylogenetic structure and leads to a reduction in the depth of the branches (π=120±70 years) compared to the one observed for discordant antigenic types.

2.3.3 Measuring Selection Strength and Direction

By comparing fixation versus extinction of antigenic mutations, using a quantity related to the McDonald-Kreitman (MK) index (J. H. McDonald and Kreitman 1991; Bush, Fitch, et al. 1999; Bhatt, Katzourakis, and Pybus 2010) we estimated the strength and direction of selection on antigenic mutations in our model (see Methods). Here we calculate an MK related (MKR) index as the ratio of the per-year rate of antigenic mutation on the trunk to the per-year rate of antigenic mutation on the side branches. If antigenic mutations are advantageous for long-term virus persistence, an MKR ratio above 1 is expected. In this case, individuals exhibiting these antigenic mutations will be more likely to fix in the population and contribute to substitutions on the trunk of the phylogeny. Similarly, if antigenic mutations are deleterious to the long-term success of the virus, an MKR index of less than 1 is expected. This is because mutant individuals will tend to be lost from the population and side branches will show an excess rate of substitution.

We find that, when rare, antigenic mutations show highly increased rates of fixation (MKR=19±11), and therefore evidence of strong positive selection (Figure 2.4B). Hence, we find that strong positive selection results in both a spindly tree and an overabundance of antigenic mutations of the trunk of the phylogeny. An increase in the mutation rate leads to the emergence of antigenically discordant types, and the suppression of other antigenic mutants (Figure 2.4C); here, we find strong negative selection mediated by host immunity with an MKR index of 0.47±0.22. At a still higher mutation rate (Figure 2.4D), we observe a balance of positive and negative selection resulting in MKR=1.1±0.3. With saturating mutation rates (Figure 2.4E) we further lose the signature of selection (MKR=1.0±0.1) on phylodynamic patterns.

2.3.4 Relationship between Cross-immunity, Mutation and Selection

Our model contains a cross-immunity parameter σ which allows us to explore a range of immune selection regimes: when σ=1, we have full cross-protection (as might arise if each epitope elicited a very strong immune response) and when σ=0, cross-protection between strains is only high if they share their entire variable repertoire.

In general, stronger cross-immunity results in lower prevalence as hosts fail to be re-infected (Figure S2.1). We find that, for most of the parameter space, genetic (genealogical) diversity π, increases with weaker cross-immunity and with more rapid mutation (Figure 2.5A). The (mostly) monotonic relationship between competition and diversity is broken at the threshold of limiting similarity (Abrams P 1983) where, regardless of epitope differences, two strains suffer full cross-protection. This scenario, shown as a band on the right-hand side of Figure 2.5A where σ=1, results in the disappearance of selective effects and greater levels of genetic diversity. Here, diversity rebounds to its neutral expectation due to random coalescence. Exceptions to the monotonic pattern of diversity with competition can also be found for intermediate mutation rates.

The relationship between mutation, cross-immunity, and the MKR index is less straight-forward (Figure 2.5B). Here, the highest levels of positive selection are present when cross-immunity is strong (σ=0.8-0.9), and mutation is weak ($\xi=10^{-5}$).

When mutation rate is limiting ($\xi<5\times10^{-5}$) then antigenic mutations are favored by natural selection (MKR>1). However, when mutation rates are higher ($5\times10^{-5}< \xi <10^{-3}$), negative selection by-and-large predominates. The strongest negative selection occurs in a region of moderate cross-immunity (σ=0.6) corresponding to previously observed discordant dynamics (Figure 2.4C).

There is also a clear relationship between diversity and selection as measured by the MKR index (Figure 2.5C). We observe a strong negative correlation between MKR and levels of diversity (ρ=-0.86; Pearson's correlation). If we separate results into a regime of positive selection (MKR>1) and a regime of negative selection (MKR<1), we observe similar results within each regime. Stronger positive selection coincides with a decrease in genetic diversity (ρ=-0.85 when MKR>1), and stronger negative selection tends further increase diversity through the persistence of discordant strains and associated deep branches (ρ=-0.28 when MKR<1). As expected from population genetic theory (Volz 2011), increases in viral prevalence also coincide with increases in viral diversity, however, the correlation is weaker under positive selection (ρ=0.74 when MKR>1) than under negative selection (ρ=0.87 when MKR<1) and cannot be trivially dissociated from the effects of selection.

Two additional strain diversity measurements based on the ecological dynamics are shown in Figure S2.2, the Shannon diversity index and the level of single strain dominance (Methods). Similar to genetic diversity, positive selection is correlated with an increase in single strain dominance (ρ=0.88 when MKR>1) and a decrease in Shannon diversity (ρ =-0.88 when MKR>1). Negative selection decreases Shannon diversity (ρ=0.39 when MKR<1) and increases single strain dominance (ρ =-0.23 when

MKR<1). While negative selection lowers the number of circulating strains, it increases genetic diversity π through the existence of deep non coalescing branches.

2.3.5 H3N2-like Characteristics

The patterns described so far suggest that the dynamics of H3N2 influenza within this framework correspond to a regime in which host immune mediated selection is strong and the antigenic mutation rate is low. We now extend the model in order to examine other characteristics relevant to H3N2 in a more detailed epidemiological setting that includes seasonality and a basic global population structure.

In this analysis we include three demes representing the northern hemisphere, the southern hemisphere and the tropics. Northern and southern hemisphere demes experience an opposing seasonal modulation (with a 14% amplitude and six months phase difference) while tropical regions experience two weaker seasons annually (Chew et al. 1998) (see Figure S2.3 and Methods). In addition the southern hemisphere population is reduced in comparison to northern hemisphere and tropical populations (Methods).

In this model we use an antigenic repertoire with 4 epitopes differing in the number of alternative variants per epitope. A typical tree for this configuration together with the corresponding diversity skyline is depicted in Figure 2.6A. We observe 13±6 antigenic clusters that come to dominate the virus population over the course of the 40 year simulation (Figure S2.4-A) with an average duration of 4±2 years. Clusters are defined based on cumulative changes in two or more epitopes based on (Huang and Yang 2011) (Methods). The turnover of virus strains results in a characteristic spindly phylogenetic tree and low standing genetic diversity (π=5.7±0.1 years). Over the course of the 40-year timespan, genetic diversity experiences a boom and bust pattern

(Figure 2.6A) with a 10%-90% range of 3-9.5 years measured by combining diversity skylines of five repeated simulations. The repeat of exact antigenic types is uncommon in the model (Figure S2.5) while epitopes with more restricted variability (2-3 variants) frequently reemerge (Figure S2.6).

Average yearly incidence in the northern and southern hemisphere demes is (5.7%±0.1, 5.8%±0.1) respectively (Figure 2.6B), while incidence in the tropics is slightly lower 5.5%±0.1. Annual epidemics are generated almost regularly yet display a high level of variability in peak size in the northern hemisphere (CoV=1.1±0.1, coefficient of variation) and lower variability in the tropics (CoV=0.7±0.1). The interquartile range in peak weekly cases ranges (IQR≈200-800 cases per 100000) in the northern hemisphere and (IQR≈400-1100) in the tropics.

In this model the tropics or lower and mixed seasonality populations exhibit a greater role (68%±9) in establishing the trunk of the influenza tree (Figure 2.6B). The southern hemisphere experiences a smaller (12%±3) part in establishing the trunk of the tree in comparison to the northern hemisphere (20%±8). In addition we find that antigenic variants are more likely to reach significant prevalence in the tropics earlier making the tropics "antigenically ahead" (Figure S2.4-B). Antigenic variants reach 5% of their total deme prevalence 2±1.5 months earlier in the tropics compared to the northern hemisphere and 3±2 months earlier in the tropics compared to the southern hemisphere ($p<0.001$ for the combined results). Similarly antigenic variants decline (reach 95% of their total prevalence) earlier (1.7±0.3 month, $p<0.0005$) in the tropics compared to the northern hemisphere, yet not significantly earlier or later than the southern hemisphere.

We find that strong competition and high R0 values generates more regular annual epidemic peaks while maintaining low genetic diversity. In addition we find that within antigenic cluster evolution also contributes to maintain low genetic diversity. An increase (0.005 compared to 0.001 of contacts) in the strength of the metapopulation coupling slightly improves the epidemiology by decreasing the likelihood of long periods without annual epidemics.

To establish whether low genetic diversity can be maintained when the number of epitopes or variants per epitope is increased we repeated the same parameterization with double the number of epitopes and with twice the number of variants per epitope and by keeping either the per-epitope or overall mutation rate. The results are summarized in Table S1.1. When doubling the number of epitopes, the model can attain similar results with respect to genetic diversity π and overall incidence, when the total mutation rate across epitopes is maintained and the cross-immunity decrease per–epitope change is halved. In contrast, it is not clear whether a model that includes an increase in the number of variants per epitope can maintain low genetic diversity levels and maintain similar or higher incidence levels.

2.4. Discussion

Herein, we implemented an individual-based model that allowed us to track both the ecological and evolutionary dynamics of a pathogen population, in which cross-immunity is orchestrated by a finite set of antigenic loci of limited variability (Recker et al. 2007). We used this model to compare phylodynamic patterns under a regime governed primarily by limitation on the introduction of antigenic mutations (mutation limited), to a regime determined by the availability of antigenic niches (selection limited), and under varying strengths of competition between strains. We use this framework to determine the conditions under which a limited antigenic repertoire could explain the observed phylodynamic patterns of H3N2 influenza.

Explicit modeling of evolution, through the introduction of antigenic mutation at different rates, allows us to consider phylogenetic trees in addition to epidemiological dynamics. Resulting phylodynamic patterns range from successive strain turnover, to discordant antigenic sets, to dynamics resembling those of a model lacking explicit evolution and finally to the collapse of antigenic structure. Each of these can be explained by the interplay of selection and mutation, as measured here through the MKR index, and by considering different strengths of immunity generating competition between strains.

The dynamics of our individual-based model are generally in good agreement with the epidemic behavior of influenza A (Figure 2.6). Like observed epidemiological patterns (Koelle et al. 2006; Russell et al. 2008; Fleming and Ayres 1988; Monto and Sullivan 1993; Karpova, Marinich, I. G, and Krainova, T. I 2006; šAltytė Benth and Hofoss 2008; Sumi et al. 2011), annual temperate climate epidemics occur almost regularly with substantial year-to-year variation in incidence (CoV=1.1±0.1 compared to (CoV=1.0±0.2) in literature survey. Observed temperate climate annual attack rates of influenza A (H3N2) are slightly higher, approximately 8% from 1976 to 1981 (Monto and Sullivan 1993) compared to 5.8%±0.1 in simulation, while peak epidemic weakly cases are higher in the simulation (IQR≈200-800 cases per 100000) in comparison to (IQR≈130-380, IQR≈80-240) in (Fleming and Ayres 1988) and (šAltytė Benth and Hofoss 2008) respectively. The tropics exhibit lower and weaker seasonality (Figure 2.6, Figure S2.3) with slightly lower yearly attack rates (5.5%±0.1) and substantially lower prevalence (Figure 2.6). In agreement with antigenic cartography (D. J. Smith 2004; Du et al. 2012) 13±6 clusters dominate the global world population

(Figure S2.4-A) with an average duration of 4±2 years, exhibiting mostly the dominance of 1-2 clusters globally. With respect to individual epitope changes we find the model reproduces the observation of the tropics being "antigenically ahead" (Russell et al. 2008), giving rise to antigenic changes 2.5±1.5 month ahead of the northern and southern hemisphere (Figure S2.4-B) and showing decline in antigenic variants 1.7±0.2 month earlier than the northern hemisphere. In agreement with observed phylodynamic patterns (Bedford et al. 2010b) the tropics metapopulation has a higher proportion in establishing the trunk (68%±9) of the phylogeny followed by the northern (20%±8) and the southern (12%±3) population. The higher contribution of East and South-East Asia as the origin of H3N2 globally circulating lineages is hypothesized to originate from lower and mixed seasonality in these regions and is consistent with our model (Russell et al. 2008). The key difference between the hemispheres in the model being, lower population size in the southern hemisphere with proportionally lower contact rate between the meta-populations. Refinement of the epidemiological model, such as the inclusion of an exposed period, can further improve the comparison to empirical data. In particular, the above properties were obtained with a basic reproduction number of R0≈3.24, on the upper bounds of current estimates for seasonal influenza. This value can possibly be decreased by considering such an extension.

We find that a model with 4 epitopes and a low but variable number of variants per site, an antigenic mutation rate of $\approx 10^{-5}$ per day and a reduction of cross-immunity of 13% per epitope results in phylodynamic patterns broadly consistent with those seen in H3N2 influenza (Figure 2.6). When doubling the number of epitopes, the model maintained similar results with respect to genetic diversity π and overall incidence, when the total mutation rate across epitopes was maintained and the cross-immunity was modified to a 6.5% per epitope change. These parameters are quite comparable to parameters used in other models of influenza evolution. For example, Koelle et al. [18] use 5 epitopes with mutations of either large or small antigenic effect. Small mutations reduce cross-immunity by 7% and occur at a rate of ~5×10^{-4} per day, while large mutations reduce cross-immunity by 20% and occur at a rate of ~10^{-5} per day. In the model of Bedford et al. (Bedford, Rambaut, and Pascual 2012a) mutations reduce cross-immunity by between 1% and 11% (95% bounds), but occur at a faster rate of 10^{-4} per day. Ferguson et al. [17] find that a model with 12 codons, each with 20 amino acid variants, in which mutations occur at a rate of 3×10^{-5} per day and reduce cross-immunity by ~7% gives restricted diversity without short-term strain-transcending immunity, and 1.2x10^{-4} per day, when transient immunity is included. From this, it seems clear that models involving a slow influx of antigenic mutants of around 10^{-5} per day are generally capable of producing influenza-like patterns of restricted diversity.

Increasing host population size in the model results in an increase in viral genetic diversity, as more opportunities for antigenic mutation arise within the larger host population. Thus, scaling competitive interactions between strains, and/or antigenic mutation rate, is required to maintain limitations on the effective exploration of antigenic space. In addition, other epidemiological phenomena, besides low antigenic mutation rates, may also contribute to limit the rate at which novel antigenic phenotypes emerge in the influenza population. These may be provided by population structure and the seasonality of transmission (Russell et al. 2008; Creanza, Schwarz, and Cohen 2010), as well as by short-term strain-transcending immunity, which was found capable of limiting genetic and antigenic diversity in a similar model with a much larger antigenic space (Ferguson, Galvani, and Bush 2003b) and in a limited diversity antigenic model (Minayev and Ferguson 2009). However, a global metapopulation structure is not expected to be the dominant cause behind the low standing genetic diversity of influenza. Influenza B exhibits similar epidemiological dynamics, and lower prevalence, yet it exhibits much higher genetic diversity through the co-circulation of multiple lineages (Ferguson, Galvani, and Bush 2003b; Yamashita et al. 1988). Also, a more complex metapopulation structure with multiple patches can either increase genetic diversity by facilitating the coexistence of viruses at different weakly coupled patches, or decrease genetic diversity through the generation of population bottlenecks. The role of variation in viral fitness is an important consideration in future studies, particularly in light of recent observations linking binding properties of HA with antigenic escape (Hensley et al. 2009). The empirical finding of a non-trivial relationship between virus fitness in susceptible individuals and immune evasion was suggested as a possible alternative mechanism for generating positive selection pressure on antigenic sites and for limiting antigenic diversity (Hensley et al. 2009).

Future work should investigate quantitative patterns and statistical approaches for discriminating among the different models and associated hypotheses that currently exist in the literature and for inferring the relative importance of the mechanisms they represent, keeping in mind that the models are not necessarily mutually exclusive. At the same time, empirical advances on the molecular basis of immune evasion and recognition, on the genotype-to-phenotype map, and on epitope identification and population serology, will allow a better evaluation of the models' assumptions, including the representation of serological space.

In common with the Recker et al. (Recker et al. 2007) model and in contrast with other phylodynamic models (Ferguson, Galvani, and Bush 2003b; Koelle et al. 2006; Bedford, Rambaut, and Pascual 2012a), we find here that antigenic epitopes are frequently recycled (Figure S2.6). Importantly, this does not mean that such recycling is observed for the antigenic types (epitope repertoires) themselves, since the same antigenic type only re-emerges at long intervals (Figure S2.5) and rarely in the course of 40 simulated years. It's possible that such reemergence could explain the antigenic cross-reactivity between sera from around the 1918 H1N1 pandemic and viruses emerging in the 2009 H1N1 pandemic (Hancock et al. 2009; Itoh et al. 2009; Bandaranayake et al. 2010; Chi et al. 2010). However, antigenic stasis of the swine lineage leading to the 2009 pandemic could also explain these observations. Much further work on epitope identification and population-wide serological surveys is necessary to establish the validity of this model's prediction on the re-cycling of constituent low diversity epitope variants (Figure S2.6). Nevertheless, several empirical observations are becoming available that are consistent with such recycling and the subject is discussed in detail in the companion paper (Wikramaratna, S. P et al. 2012). For example, an antigenic analyses performed on H2N2 influenza, a number of monoclonal antibodies raised against a 1957 strain were shown to cross-react strongly with a strain isolated in 1964, yet not with the 1963 strain (Tsuchiya et al. 2001). In Reichert et al. (Reichert et al. 2010) the hemagglutinin of both novel pandemic H1N1 and pre-1940 H1N1 lack specific glycosylation sites on the globular head of HA1. These reverse glycosylation patterns were suggested to possibly shield antigenic sites for a timescale of decades and in so doing, to effectively contribute to their recycling and to the age distribution of cases. In Bui et al. (Bui et al. 2007), several protective antigenic and T cell H3 epitopes show temporal variability across drift variants, with two of these specifically exhibiting a decrease and increase in conservancy consistent with epitope "recycling". Post translational and conformational changes may hinder the validity of this analysis especially for epitope 1 which acquired two surrounding glycosylation sites. In Wang et al. (T. T. Wang et al. 2010), mice Anti-H3 mAbs were shown to neutralize H3 viruses that span 40 years, as measured by immunofluorescence against MDCK cells (Table 2, REF). All three mAbs (Figure 2.4, REF) displayed variability in their ability to neutralize H3 viruses for lower concentrations (<15μg/ml of 7A7 and <25μg/ml for the other two) in plaque reduction assays. For example for mAb 7A7 neutralization was better for HK68, than diminished for BJ92 and then increases for PAN99 and BRIS07. This pattern could also be due to secondary effects of amino acid differences outside the actual epitope as well through structural effects, but effectively behaves as epitope recycling over substantial durations of many years.

In conclusion, within our framework, the rate of antigenic mutation was found to strongly influence whether selection was positive or negative, and hence, the topology of the tree and associated diversity of the virus. Strong positive selection is generated by effective competition under low mutation rates, and results in spindly trees with low genetic diversity. In this regime, antigenic mutations often fix in the virus population, lowering genetic diversity, as consistent with H3N2. An increase in mutation rate across a broad spectrum in competition strength, leads to negative selection and generates antigenic divergence. This can potentially result in the coexistence of discordant antigenic types repressing the emergence of antigenic hybrids, through strong negative selection on antigenic change, with each discordant antigenic type maintaining a deep phylogenetic branch. Although not strictly mirroring the assumptions about development of the Recker et al model, our framework strongly implies that limitations on antigenic architecture alone are unlikely to reliably reproduce "skinny' trees and some restrictions on mutation rate and/or other considerations such as fitness differences are likely to play a role. It is important to note that this exercise does not also privilege other hypotheses concerning diversity restriction in influenza as these also are strongly sensitive to mutation rate. Overall, it emphasizes that phylogenetic patterns do not serve as a discriminatory tool between these by no means mutually exclusive hypotheses. However, they can provide a basis to exclude specific hypotheses and offer a means by which the contributions of mutation and selection can be assessed. Needless to say, the latter has important implications for the updating of vaccines against influenza. Under a mutation-limited regime, a hypothetical vaccine should be effective until a new antigenic variant is introduced to the population through migration or mutation. Alternatively, when a limited number of alternative but conserved epitopes are continuously circulating with their abundance determined through competitive interactions and immune mediated selection, a vaccine against one of them may lead to rapid strain replacement, while a vaccine against all of them may result in effective intervention.

2.5. Methods

2.5.1 Cross Immunity

We assume that a strain's antigenic attributes are determined by a set of separate epitopes and that each epitope contains a discrete number of alternative variants. Thus, a strain's antigenic properties are defined by an *n*-tuple with k_i variants per epitope giving a possible number of $\prod_{i=1}^{n} k_i$ antigenically distinct strains. Hosts acquire immunity to viral epitopes following infection. Fully naïve hosts are always infected following contact at rate β. The risk of reinfection with the exact same strain is always zero. The chance of reinfection with a different strain is based on the similarity to previously encountered strains, measured through the fraction of previously encountered epitopes: $\text{risk of infection} = (1-f)\cdot(1-\sigma)\cdot n_{epitopes}$ and is at most 100%. Where *f* is the fraction of previously encountered epitopes and σ is the strength of crossimmunity. Lower σ values correspond to weaker competition between strains. A form of generalized immunity is attained for σ > 0.8 in the five epitope case, relating to a reduced risk of reinfection following previous exposure to any strain (Figure 2.2B).

2.5.2 Mutations

Phenotype changes are driven by mutation events. Mutations change the antigenic properties of a strain but do not influence the shape of the genealogical tree directly; the tree shape will be determined by selection, epidemiological dynamics and the stochastic processes favoring a specific isolate and its offsprings, implicitly rather than explicitly. Mutations involve changes in a single epitope; this makes some phenotypic changes more attainable in comparison to others, even for high mutation rates. The antigenic mutation rate ξ gives the per-day probability for a virus to mutate in a single epitope site.

2.5.3 Genealogy Tracking and Related Diversity Quantities

The genealogy of the virus population was tracked directly throughout the simulation (Figure 2.2A). Constant random sampling of viruses was performed periodically. Genealogical pairwise diversity (π) was measured by averaging the time unit distance on the tree between random contemporaneous sample pairs (Figure 2.2C). This quantity relates to pairwise genetic diversity, as measured on an accurately reconstructed phylogenetic tree. Diversity measures are limited in our simulation to a maximum total of twice the total running time of the simulation which amounts to 240 years. Although it's clear that some parameter ranges would show diversity greater than 240 years, they will saturate at this threshold (Figure 2.5C).

The MKR index was calculated by dividing the observed rate of occurrence of phenotypic mutations on the trunk of the tree, by the per-year mutation rate on the side branches (Figure 2.2D). This allows us to estimate the importance of antigenic mutations on the likelihood of fixation of a given viral linage.

Single strain dominance was calculated based on the quantity ε from (Recker et al. 2007) and is calculated using the following formula:

$$\varepsilon = \frac{1}{P}\sum_{i=1}^{P} \frac{y_{\max}^{i} - y_{sub}^{i}}{y_{\max}^{i}} \qquad (1)$$

where $y_{\max}$ and y_{sub} are the prevalence of the most and 2nd most prevalent strains. The normalized difference between the two is averaged across P epidemic peaks.

2.5.4 Increasing Mutation Rate Parameters

For figures 3, 4 and S7, R0 was set to ≈2.4. Antigenic diversity was limited to 2 variants per epitope to attain tractable results for a wide range of mutation and cross-immunity parameters, and for the same reason no metapopulation structure was established.

Parameter	Value
population size N	40,000,000
contact rate β	0.6 [1/day]
recovery rate ν	0.25 [1/day]
birth/death rate μ	1/25 [1/year]
Epitopes	5
variants per epitope	2x2x2x2x2
cross-immunity σ	0.825

2.5.5 Mutation-Competition Simulation Parameters

For this set of simulations (Figure 2.5) R0 was set to ≈3 and no metapopulation structure was assumed. Higher R0 and longer duration of infection reduce the effect of critical community size over the range of parameters analyzed given the population size determined by computer resource use.

Parameter	Value
population size N	50,000,000
contact rate β	0.6 [1/day]
recovery rate ν	0.2 [1/day]
birth/death rate μ	1/30 [1/year]
epitopes	5
variants per epitope	2x2x2x2x2

2.5.6 H3N2 Simulation Parameters

This single simulation parameterization is intended to test whether H3N2-like phylogenetic trees can result from a model with a restricted set of antigenic phenotypes using an alternative epitope configuration and a basic metapopulation structure which includes seasonality. The basic reproduction number was set to R0≈3.24. In this case three demes were assumed, representing the northern hemisphere, the southern hemisphere and the tropics. Four epitopes with a variable number of variants per epitope were used. Contact rate was attenuated sinusoidally for southern and northern hemispheres to establish seasonal patterns. Tropical climate seasonality was sinusoidally modulated include two seasons (Chew et al. 1998) of weaker amplitude (Cécile Viboud, Alonso, and Simonsen 2006).

Parameter	Value
population size N North/Tropics/South	16M/16M/10M
contact rate β_0	0.6 [1/day]
Between deme contact	0.005 [1/day]
recovery rate ν	0.185[1/day]
birth/death rate μ	1/30 [1/year]
epitopes	4
variants per epitope	5x4x3x2
Temperate Climate	14% Jan/July
Seasonality North/South	Six Month Phase
Tropics Seasonality	7% Dec/June
Mutation rate ξ	0.000008 [1/day]
cross-immunity σ	0.87

2.5.7 Antigenic Cluster Transitions

It is not clear how changes in individual epitopes relate to the antigenic clusters as proposed by Smith et al. (D. J. Smith 2004) or Du et al. (Du et al. 2012). A new antigenic variant differing by 2 or more epitopes from any previous cluster strain was grouped in a new cluster in agreement with Huang et al. (Huang and Yang 2011). This does not affect model dynamics, and relates only to the coloring of clusters in Figure 2.6 and Figure S2.4-A.

2.6. Acknowledgments

We thank Pamela Martinez, Oliver Pybus and Adam Lauring for comments and discussion. MP was supported by a Centennial Fellowship of the James S. McDonnell Foundation (http://www.jsmf.org/), SG, by the European Research Council (ERC Advanced Grant – Diversity, http://erc.europa.eu/advanced-grants), and TB, by the Howard Hughes Medical Institute, a Long-Term Fellowship from the European Molecular Biology Organization and a Newton International Fellowship from the Royal Society. MP is an investigator of the Howard Hughes Medical Institute and SG is a Royal Society Wolfson Research Fellow (royalsociety.org/grants/schemes/wolfson-research-merit/).

2.7. Summary

Influenza A (H3N2) has circulated in the human population since 1968 causing considerable annual morbidity and mortality worldwide. Despite the rapid evolution of the hemagglutinin (HA) protein and strong diversifying selection, the global virus population is characterized by a low standing diversity, evident in the serial replacement of antigenic types and in the 'cactus-like' structure of its genealogical tree. Elucidating the mechanisms behind these puzzling patterns is key to understanding the evolution of seasonal (H3N2) influenza. One recent epidemiological model proposes a restricted set of antigenic types whose waves of dominance result from frequency-dependent immune selection. Here we develop a model of limited

antigenic diversity that explicitly incorporates mutational processes, and use it to address, first, whether this type of antigenic space is capable of generating the characteristic phylogeny of HA sequences, and second, whether the dynamics of (H3N2) influenza are primarily limited by the arrival of mutations or by the opening of antigenic niches. We conclude that a limited antigenic space can explain the observed phylogenetic patterns and that a limited mutation rate is a key property underlying the dynamics of (H3N2) influenza. Our study provides a general framework for assessing the relative roles of selection and mutation in a variety of infectious disease systems.

2.8 Figures

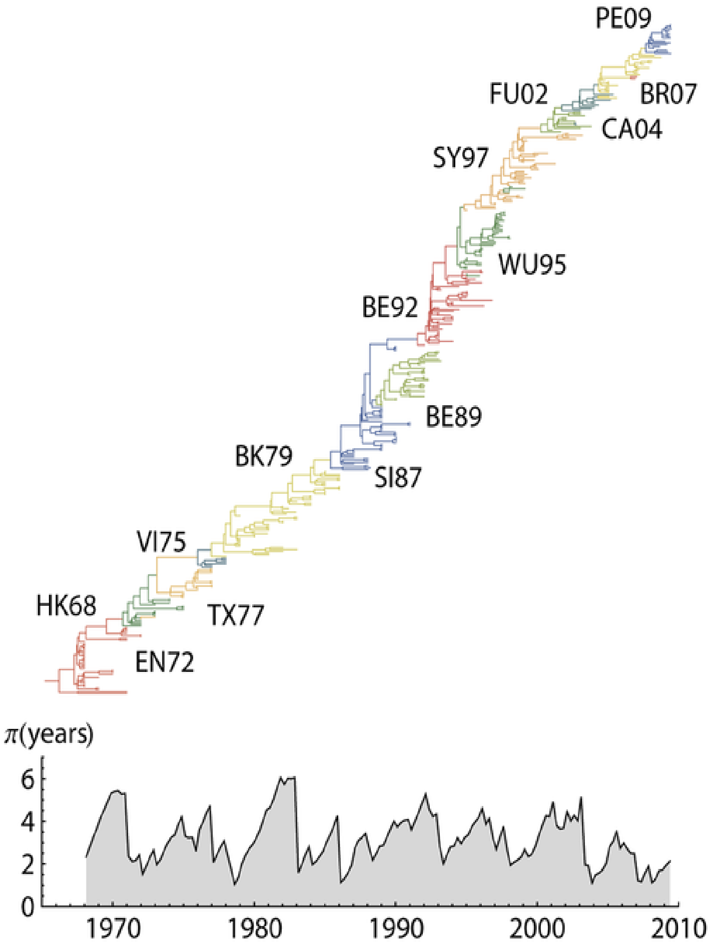


Figure 2.1. Phylogenetic tree reconstruction of H3 depicting major antigenic clusters and including the associated pairwise diversity.

Phylogenetic tree with highest posterior likelihood was reconstructed using 377 representative sequences sampled between 1968-2009. Colors represent estimated antigenic clusters (Hong-Kong 1968 – Perth 2009). Approximately half of the samples include an established cluster annotation (D. J. Smith 2004) and three additional clusters relating to: California 2004, Brisbane 2007, and Perth 2009. Additional sequences were sampled uniformly overtime on a bi-annual scale. Phylogenetic tree was reconstructed using Bayesian MCMC analysis (A. J. Drummond et al. 2012; P. Lemey et al. 2009) and includes state reconstruction for unannotated sequences and ancestral sequences. Diversity skyline was calculated for the same representative tree. Branches with colors differing from their main neighboring cluster represent uncertainty in the reconstruction, rather than actual cluster changes.

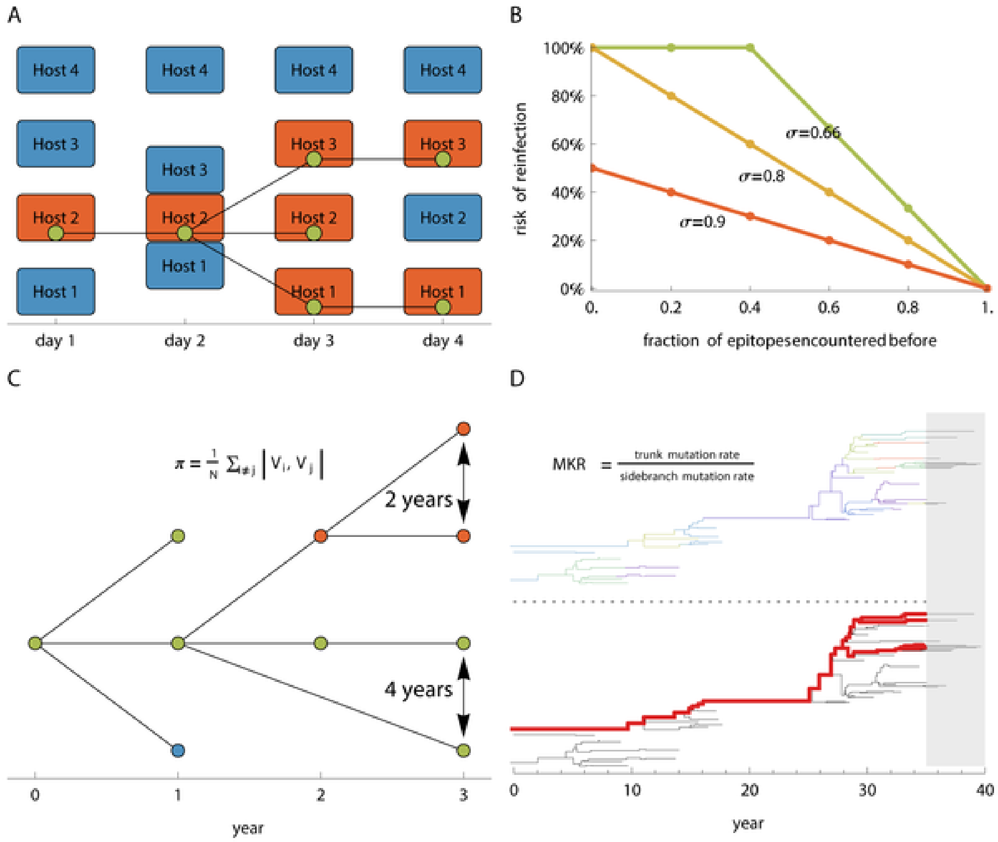


Figure 2.2. Methods

(A) Virus genealogy is tracked at the inter-host level. The genealogy is periodically sampled and the resulting tree is used for analysis. **(B)** Hosts acquire immunity to viral epitopes following infection. Fully naïve hosts are always infected following contact at rate β. The risk of reinfection is based on the similarity to previously encountered strains, as measured through the number of previously encountered epitopes: $\text{risk of infection} = (1-f)\cdot(1-\sigma)\cdot n_{epitopes}$ and is at most 100%. Where f is the fraction of previously encountered epitopes and σ is the strength of crossimmunity. Lower σ values correspond to weaker competition between strains. A form of generalized immunity is attained for σ > 0.8 in the five epitope case, relating to a reduced risk of reinfection following previous exposure to any strain. **(C)** Mean pairwise genealogical diversity π is measured by averaging the pairwise distance in years between random contemporaneous samples on the genealogical tree. **(D)** The MK related index is calculated as the ratio of the antigenic mutation rate on the trunk of the genealogy (red) versus the antigenic mutation rate on the sidebranches (black). The trunk of the genealogy was determined by tracing back viral lineages that survived until the end of the simulation and excluding the last 5 years. Antigenic changes are represented by color changes on tree branches (top-tree). The rate of antigenic change on the sidebranches is calculated as the number of antigenic changes on the sidebranches divided by the total length of the side branches in years. The rate of antigenic change on the trunk is calculated as the number of antigenic changes on the trunk divided by the total length of the trunk in years.

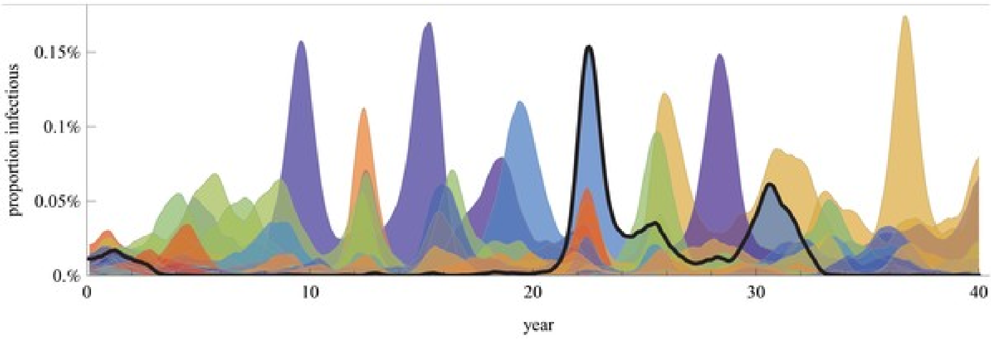


Figure 2.3. Changes in the proportions of hosts that are infectious with different strains within a 2 variants per epitope, 5 epitope system in an "evolutionary free" framework
for all of the possible 32 strains the existence of at least one carrier was assured and no antigenic mutations were introduced. The superimposed time series were smoothed and ordered back to front by peak prevalence, maintaining the least prevalent strain in the front. The 3th highest peaking strain was outlined as an example. Single strain dominance was calculated based on the quantity ε from (Recker et al. 2007). Major peaks of incidence are generally associated with one or two dominant antigenic-phenotypes with ε=0.36±0.07 (mean ± standard-deviation across 5 simulations) and a myriad of lower prevalence ones. Antigenic-phenotypes reemerge with alternating frequency. This simulation includes a single homogeneously mixed host population of 40M hosts, contact rate β=0.6 and a 4 day recovery rate. Each epitope unencountered by the host contributes to a 17.5% increase in the risk of infection with a different strain (see methods for full description of epidemiological parameters).

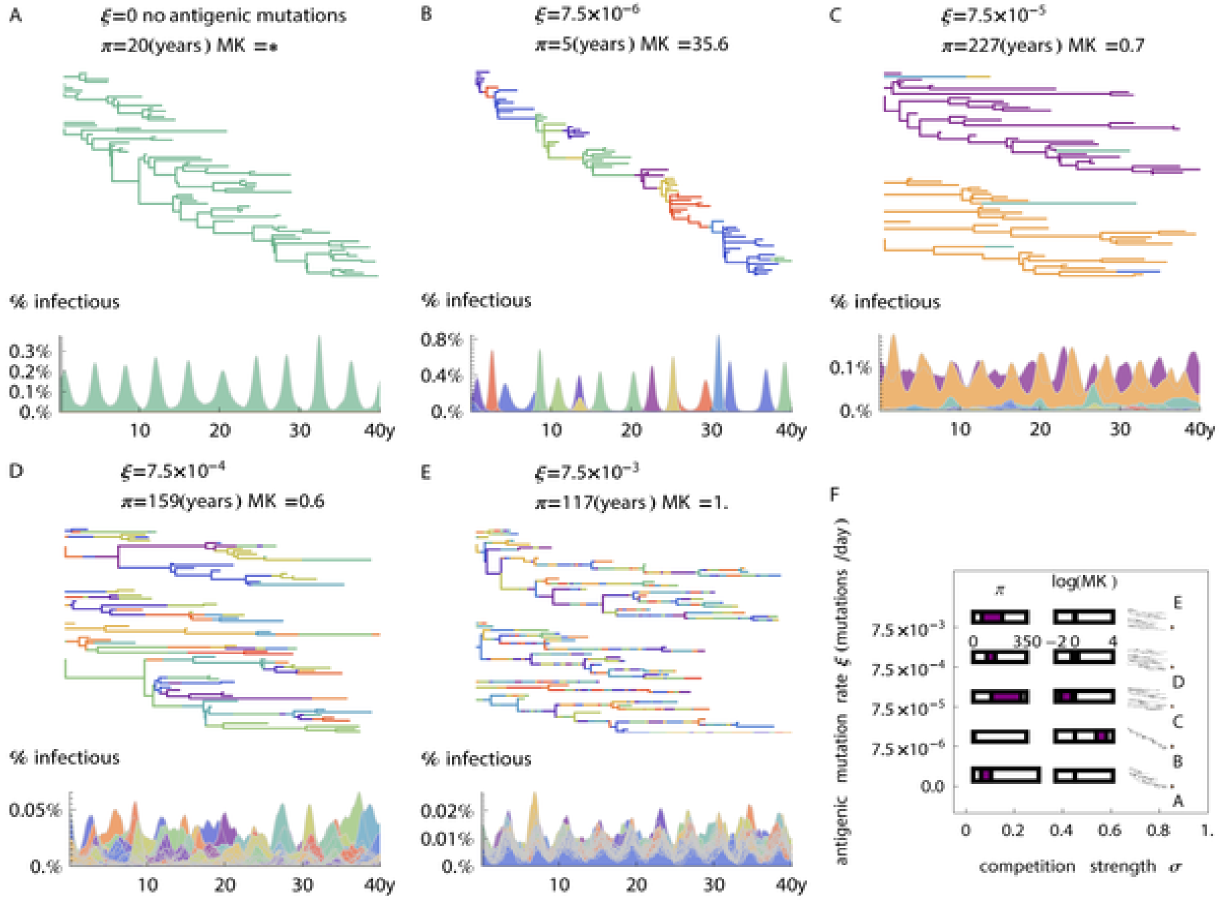


Figure 2.4. Changes in the proportions of hosts that are infectious with different strains and the related phylogenetic behavior with increasing mutation rate.

Phylogenetic trees are based on samples of directly measured virus genealogy in the simulation, and only the last 40 years are visualized in the figure (for the complete genealogy over the whole time period see Figure S2.7). Diversity π is calculated as the mean distance, measured in years, for the coalescence of random pairs of contemporaneous samples in a tree. The MK related index (MKR) is calculated as the ratio of the antigenic mutation rate on the trunk (fixed) versus the antigenic mutation rate on the sidebranches. **(A)** Model with no mutation, a single antigenic type persists under neutral evolution. When there are no antigenic mutations a genealogical tree which follows neutral viral evolution exists. Genetic diversity for this tree relates to population dynamics only – to incidence and the prevalence. **(B)** Model with low mutation rate of ξ=7.5×10^{-6} antigenic-mutations per day. Successive strain replacement with higher epidemic peaks is observed. Rare antigenic mutations are advantageous and are more likely to fix and have viable offsprings, consequently lowering genetic diversity. **(C)** The introduction of a higher mutation rate ξ=7.5×10^{-5} leads to antigenic and genetic divergence. Dynamics are ruled by endemic or cyclic behavior of discordant antigenic strains. Mutations are more likely to be deleterious, facing competition from the two prevalent strains. Phylogenetic patterns include two deep branches representing each strain and a low rate of coalescence between strains. **(D)** For a mutation rate ξ=7.5×10^{-4} epidemiological behavior resembles the evolution free framework (Figure 2.3). Phylogenetic patterns exhibit high genetic diversity and weak negative selection pressure. **(E)** Loss of strain structure due to high mutation rate ξ=7.5×10^{-3}. At this high mutation rate the antigenic traits are no longer heritable and each linage displays a constantly varying antigenic phenotype. No selection forces are measured and genetic diversity is expected to be determined by random coalescence. **(F)** Summary statistics and typical trees for varying mutation rates and fixed crossimmunity (filled area within rectangles indicates 1σ confidence interval for 5 repeated runs). Simulation parameters are the same as those described in Figure 2.3, but include the possible extinction of strains, and mutations to individual epitopes at a specified rate ξ (see methods for full description of epidemiological parameters).

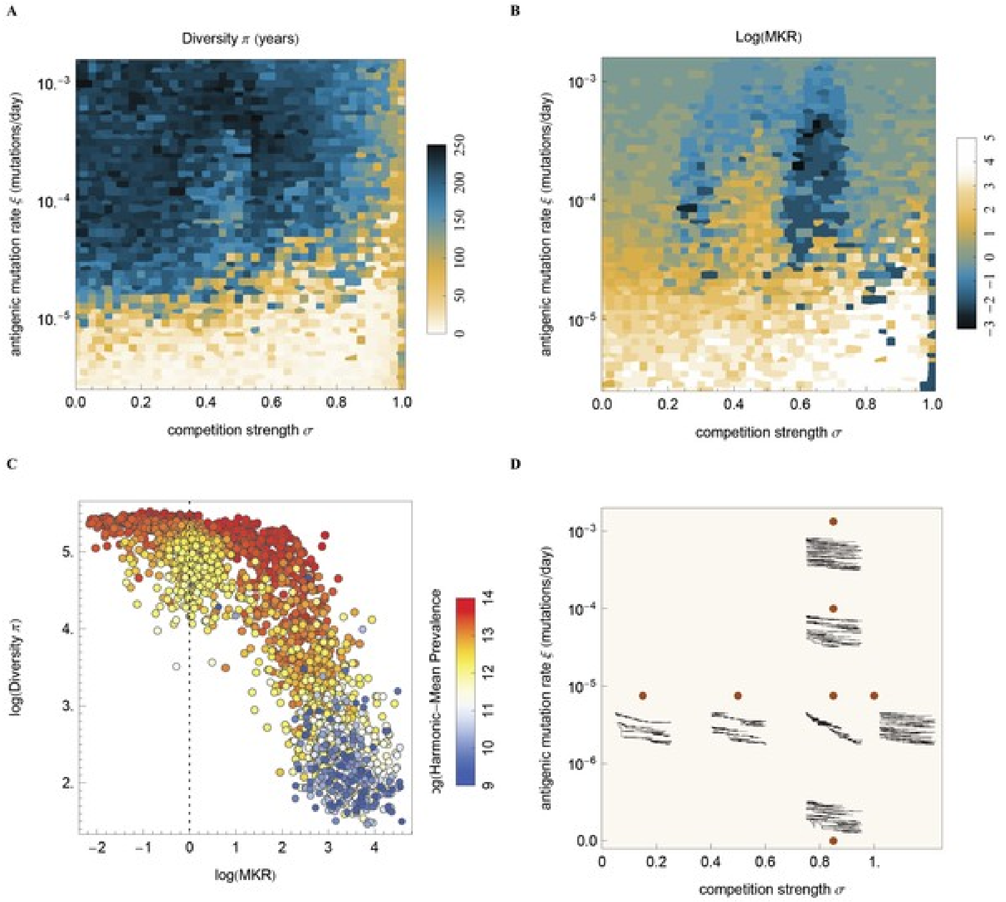


Figure 2.5. Changes in genetic diversity and the McDonald-Kreitman related index (MKR) for varying strengths of strain competition and antigenic mutation rates.

(A) Mean pairwise genetic diversity π is measured as the mean distance, measured in years, for the coalescence of random pairs of contemporaneous samples in a tree. Diversity measurement is capped by twice the simulation run length which amounts to 240 years **(B)** The MKR is measured as the ratio between the trunk antigenic mutation rate (fixed) and the sidebranches antigenic mutation rate. Evidence of positive selection is observed when the MKR index is significantly above one, and negative selection when it is significantly below one. Areas of strong positive selection are associated with lower genetic diversity as a small subset of the population contributes to long term viral evolution. Strong negative selection is associated with disruptive selection maintained by existing strains. **(C)** Diversity for varying strengths and directions of selection as measured by the MKR index. Diversity decreases with stronger positive selection ρ=-0.85 (Pearson's correlation right of dotted line), and increases for stronger negative selection ρ=-0.28. The harmonic mean of the prevalence is also strongly correlated with genetic diversity ρ=0.76 (heat map) **(D)** Typical trees for varying strengths of strain competition and antigenic mutation rate. Effective competition combined with a limited availability of antigenic mutations results in narrower trees with lower pairwise diversity. This figure was parameterized to use R0=3 and a population size of 50M to limit stochastic extinctions for a large parameter range (see methods for full description of epidemiological parameters).

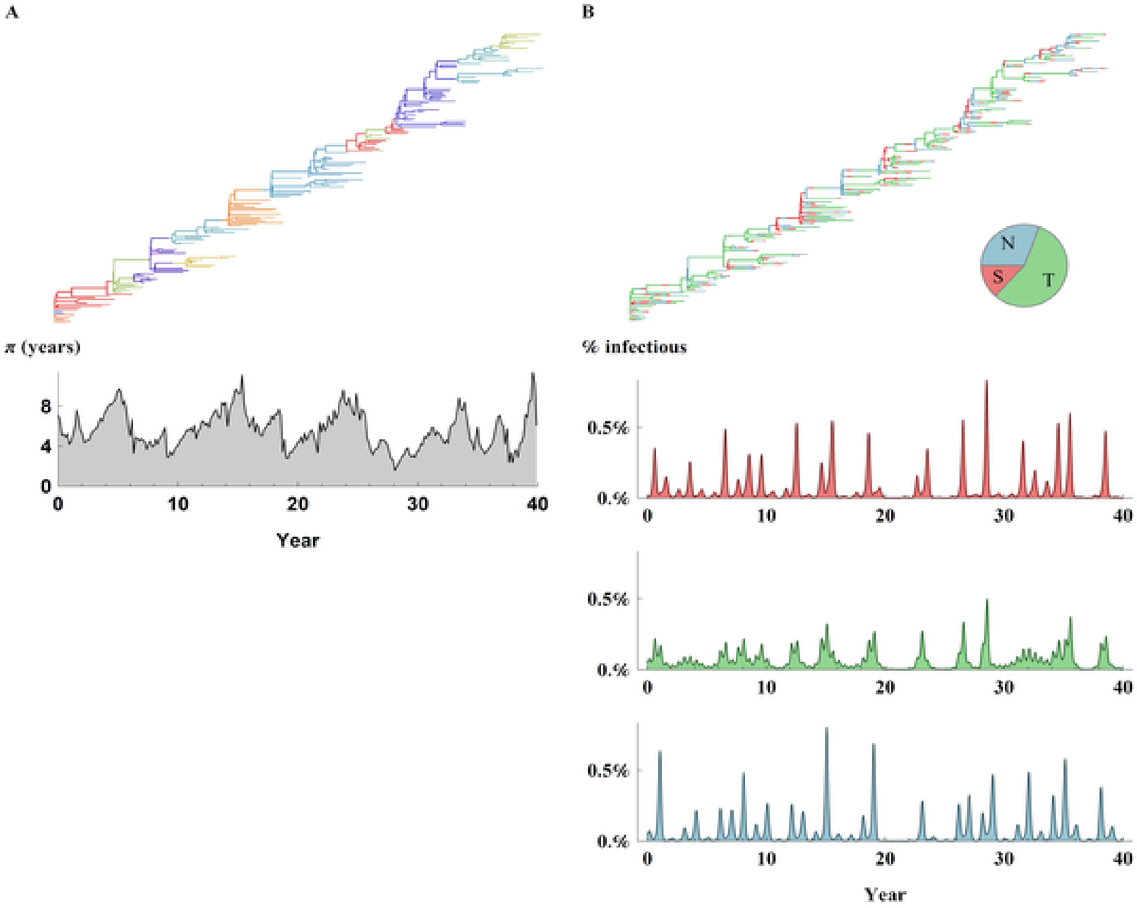


Figure 2.6. H3N2 like characteristics.

Using a model with an alternative epitope configuration and three global demes representing the northern hemisphere N, southern hemisphere S, and the tropics T. **(A)** Top - typical tree colored by antigenic clusters (see Methods). On average 13±6 clusters lasting 4±2 years come to dominate the virus population over the course of the 40 year simulation. Bottom - Genealogical diversity (π) displaying "boom & bust" patterns (10%-90% range of 3-9.5 years) associated with H3 diversity with an average of 5.7±0.1 years **(B)** Top - typical tree colored by deme. The tropics metapopulation has a higher proportion in establishing the trunk (68%±9) of the phylogeny followed by the northern (20%±8) and the southern (12%±3) population. Bottom – proportion of hosts infected in the northern hemisphere, tropics and southern hemisphere. Epitope configuration for this figure was 5×4×3×2 variants per epitope, 4 epitope system. Average yearly incidence in the northern and southern hemisphere demes is (5.7%±0.1, 5.8%±0.1) respectively, while incidence in the tropics is slightly lower 5.5%±0.1. Annual epidemics are generated almost regularly yet display a high level of variability in peak size in the northern hemisphere (CoV=1.1±0.1, coefficient of variation) and lower variability in the tropics (CoV=0.7±0.1). In this simulation 42M hosts were divided to three demes with the tropics and the south having 16M hosts and the southern hemisphere having a lower 10M population. Annual seasonal patterns were established for the temperate demes, and biannual weaker seasonality in the tropics (see methods for full description of epidemiological parameters).

2.9 Supporting Information

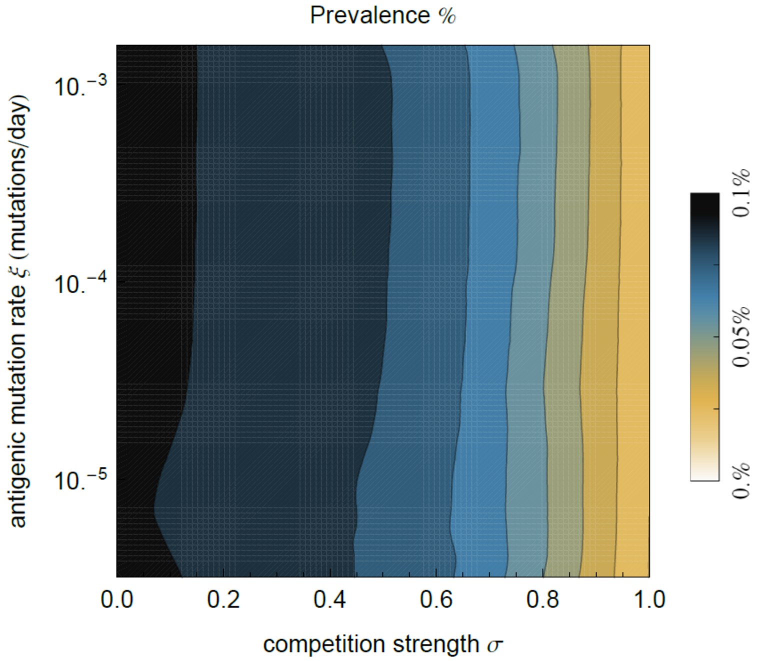


Figure S2.1. Changes in prevalence for carrying strengths of strain competition and antigenic mutation rates.

Prevalence increases as crossimmunity between strains decreases, enabling multiple infections. When crossimmunity is 1, all strains are antigenically equal, and one lifetime infection is possible. When no crossimmunity is present, each antigenic-type can independently infect a host once.

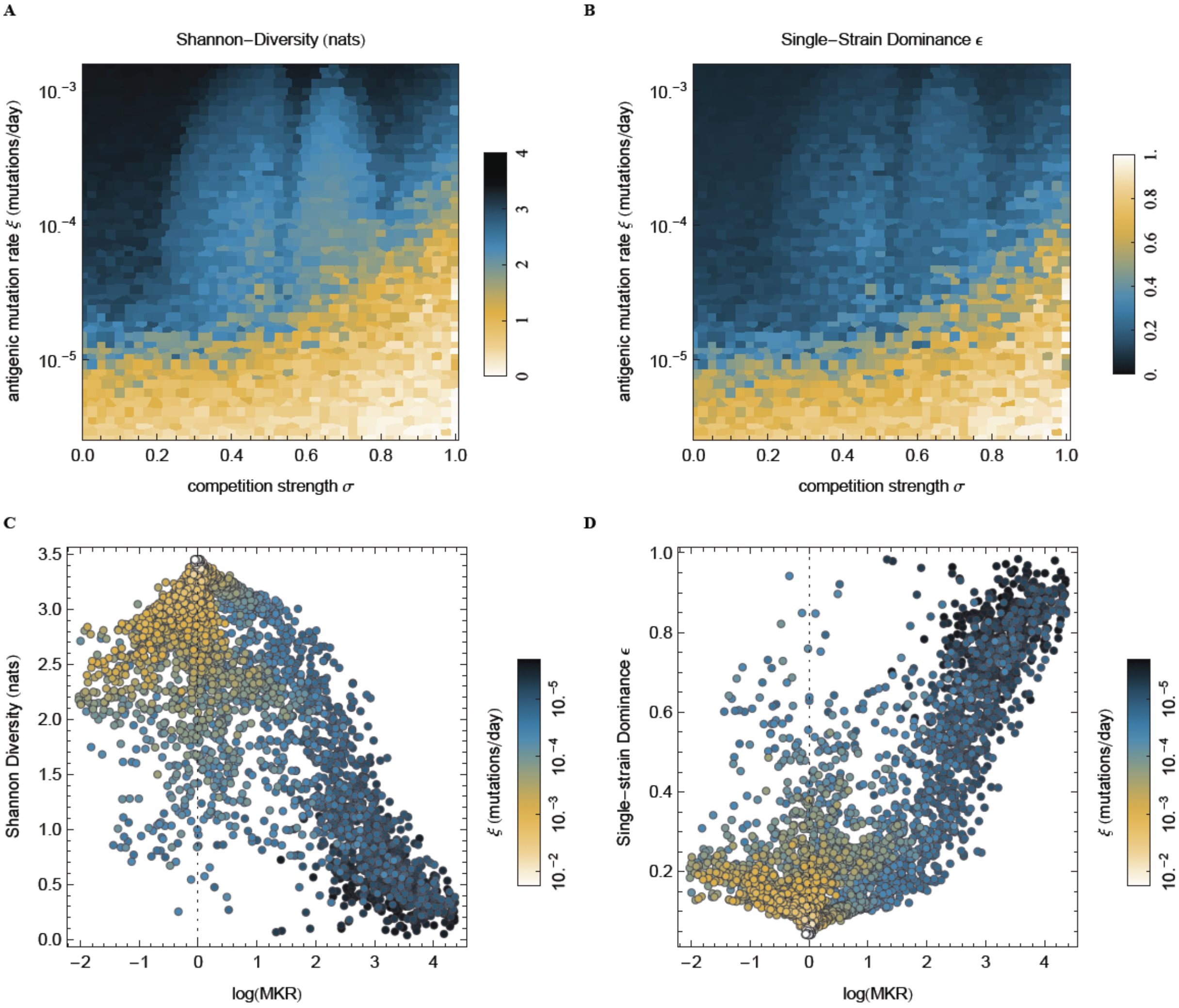


Figure S2.2. Changes in antigenic diversity and the McDonald-Kreitman related index (MKR) for varying strengths of strain competition and antigenic mutation rates.

(A) Mean Shannon diversity was measured for 40 years of simulation. Shannon diversity ranges from zero when a single circulating antigenic variant is present at each time point, to approximately 3.5nats when all the possible antigenic variants are continuously present. For a large range of the parameter space, stronger competition and lower mutation rates decrease Shanon diversity as fewer circulating antigenic types co-exist. **(B)** Single strain dominance based on the quantity ε from (Recker et al. 2007) (see Methods). With stronger competition and lower mutation rates epidemics are contain a larger fraction of a single antigenic type **(C)** Shannon diversity decreases with stronger positive selection (ρ =-0.88 when MKR>1) and with stronger negative selection (ρ=0.39 when MKR<1). Positive selection roughly corresponds to lower mutation rates ($\xi<10^{-3}$), while negative selection corresponds to higher mutation rates ($\xi>10^{-4}$) **(D)** Single strain dominance increases with stronger positive selection (ρ=0.88 when

MKR>1) and an increases for stronger negative selection (ρ =-0.23 when MKR<1). (see methods for full description of epidemiological parameters).

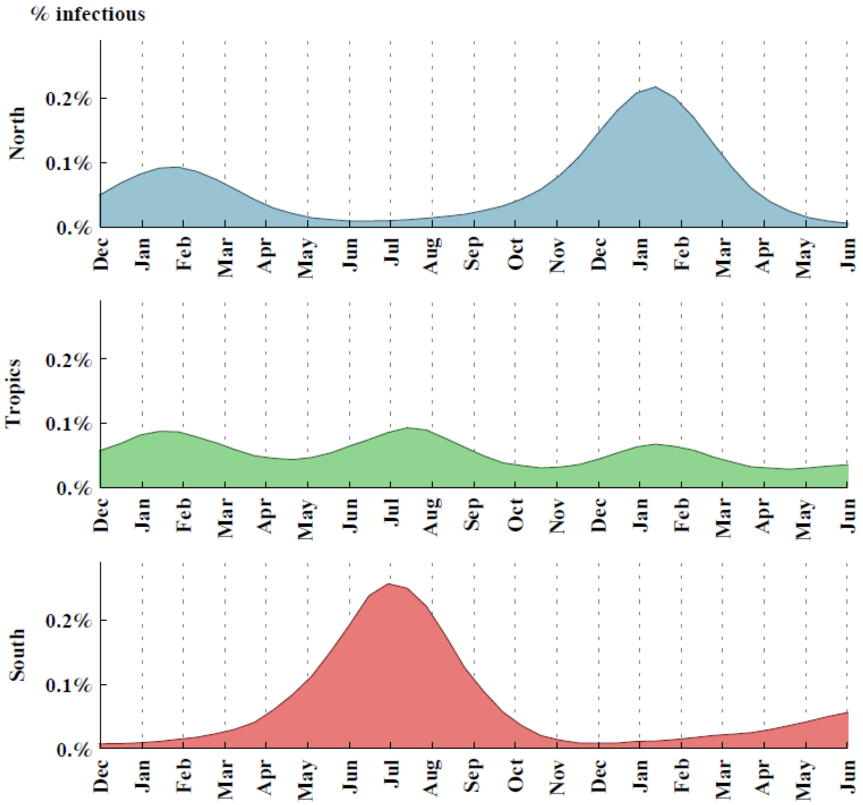


Figure S2.3. Seasonal patterns.

For simulations including metapopulation sinusoidal seasonal forcing was used (see Methods). Contact rate was modulated sinusoidally with 14% amplitude in temperate demes, and lower biannual seasonal cycles of weaker (7%) amplitude in the tropics. The observed seasonal patterns in the simulation include annual peaks centered around Jan-Feb in the northern hemisphere, July in the southern hemisphere and weaker peaks centered around late July and mid January in the tropics.

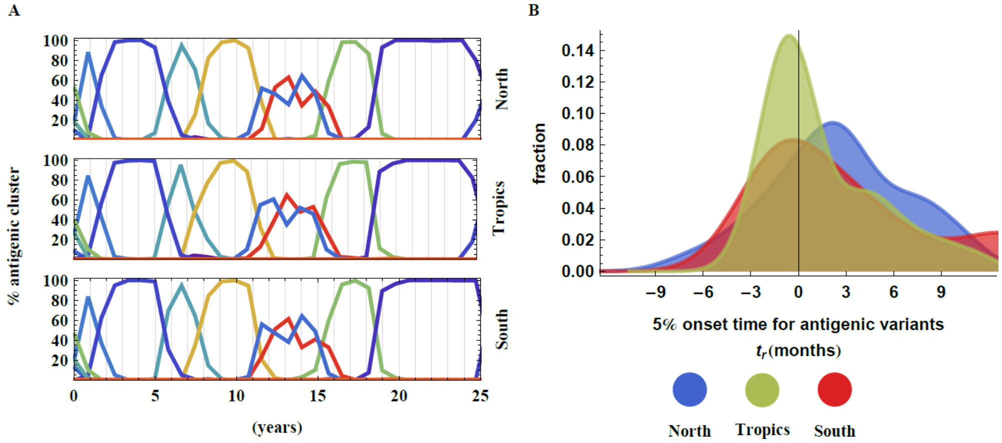


Figure S2.4 Dynamic changes in the percentage of antigenic clusters and the dominance of antigenic variants in different metapopulation demes.

(A) Changes in the percentage of the population infected with a specific antigenic cluster variant for the northern hemisphere, the southern hemisphere and the tropics. On average we observe 13±6 antigenic clusters that come to dominate the virus population over the course of the 40 year simulation with an average duration of 4±2 years. One or two clusters usually dominate the deme population. Clusters are defined based on a threshold set when the cumulative change of two or more epitopes between any previous cluster antigenic-type is reached, based on (Huang and Yang 2011) (see Methods). Clusters are only used for coloring of strains and figures and do not affect the model dynamics. A time window of 25 years was selected for comparison with (Du et al. 2012) **(B)** Measurement of the onset time for all antigenic types (prior to cluster subdivision). Onset time was measured as the point where prevalence was estimated to reach 5% of its overall deme prevalence. Antigenic variants are more likely to reach significant prevalence in the tropics: 2±1.5 months earlier in the tropics compared to the northern hemisphere and 3±2 months earlier in the tropics compared to the southern hemisphere (p<0.001 for the combined results).

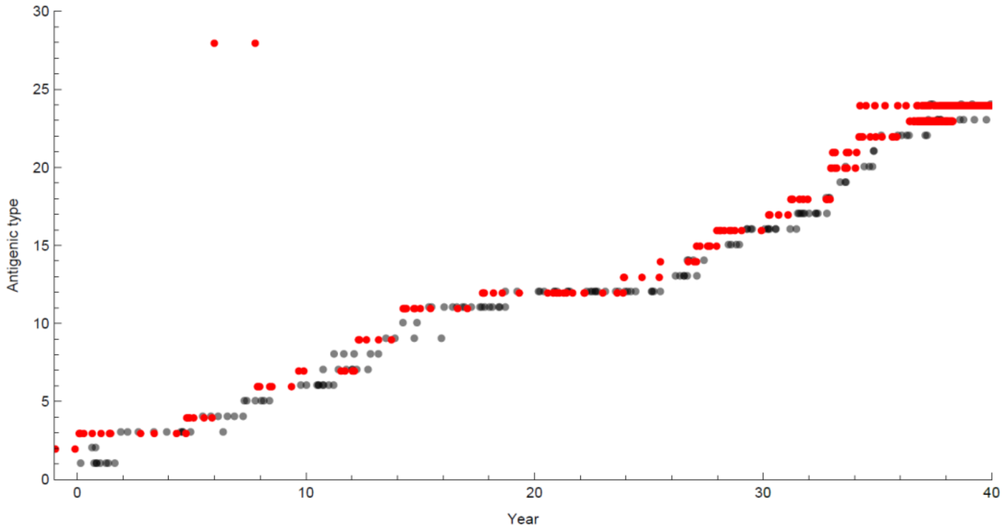


Figure S2.5. Antigenic types across 40 years of simulated years.
A sample of antigenic types that emerge in the simulation are sampled and numbered sequentially. Red – Antigenic variants sampled from the trunk of the tree (fixed). Black – Antigenic variants sampled from side branches of the phylogenetic tree.

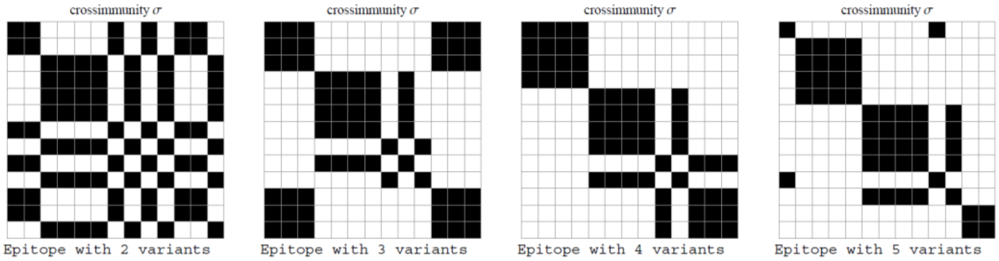


Figure S2.6 Crossimmunity patterns for individual epitopes.
The thirteen most prevalent antigenic types from a span of 40 years of simulation were sampled (Figure 2.6) and ordered by year of introduction. Individual epitopes were compared between the strains. Epitopes with lower variability (2-3 variants per epitope) show a larger degree of reemergence while epitopes with higher variability (4-5 variants per epitope) show a lower degree of epitope reemergence.

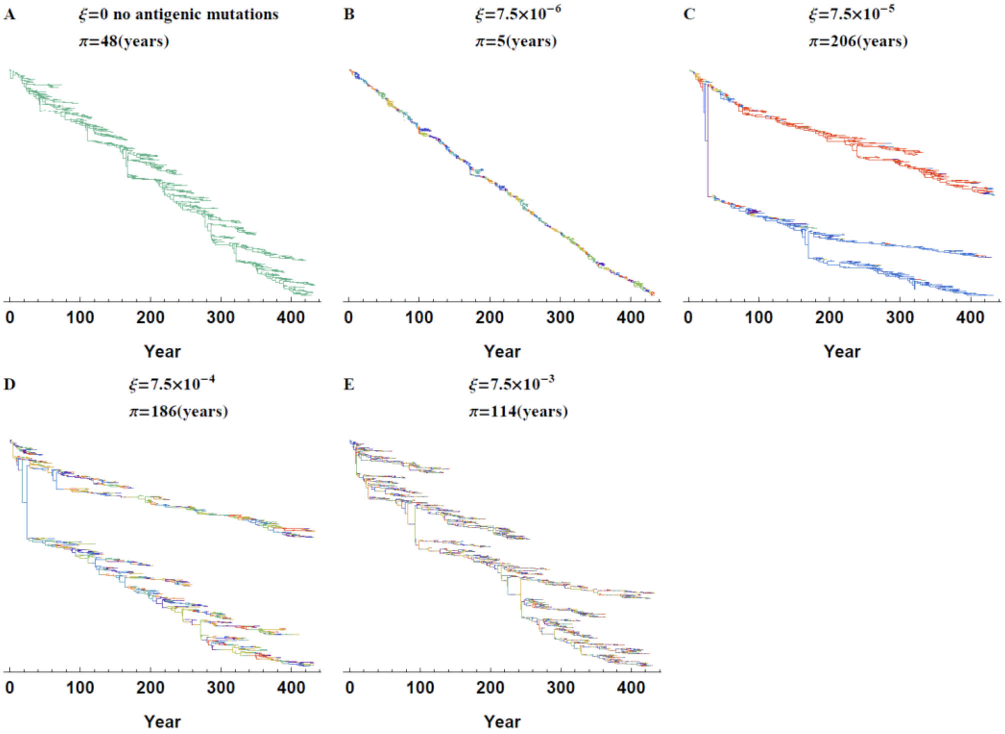


Figure S2.7 Full phylogenetic trees following the initial introduction of a virus with an increasing mutation rate.

Phylogenetic trees are based on samples of directly measured virus genealogy in the simulation (434 years). Figure 2.4 in the main body of the paper shows the last 40 years of a simulation with the same parameters (see caption of that figure for details). For these longer sampling windows, extinction was prevented by maintaining at least 50 infected individuals.

Table **S1.1. Model results for doubling the number of epitopes or variants per epitope**

Epitope Configuration	Total Mutation Rate Across Epitopes (1/day)	Immunity Reduction Per Epitope Change (1-σ)	Mean Genetic Diversity π	Incidence Increase / Decrease
5x4x3x2	8x10^{-6}	13%	5.7±0.1	Baseline
5x5x4x4x3x3x2x2	8x10^{-6}	13%	10±4	↑
5x5x4x4x3x3x2x2	8x10^{-6}	6.5%	6±1	≈
5x5x4x4x3x3x2x2	1.6x10^{-5}	6.5%	8±5	≈
10x8x6x4	8 x10^{-6}	13%	16±15	↑
10x8x6x4	8 x10^{-6}	6.5%	6±1	↓
10x8x6x4	1.6 x10^{-5}	6.5%	10±5	↓

Chapter 3. Antagonism between reassortment and antigenic niche partitioning in a model of strain dynamics motivated by rotavirus

3.1 Abstract

Several theories have been put forward to explain the co-existence of multiple pathogen strains. A major body of work postulates that the patterns of immunity in the host population can drive pathogens to differentiate into, and maintain groups with, reduced immune cross-reactivity. However, these dynamics were not fully explored under conditions characteristic of pathogens such as rotavirus, which include the repeated introduction of antigenic novelty through zoonosis or immigration, high rates of genome segment transfer in the form of reassortment, and strong non-specific (heterotypic) immunity. Here we use a model motivated by rotavirus to investigate serotype population structure under these conditions. Our work extends previous work on strain and serotypes' community structure to viral and bacterial pathogens that share these rotavirus-like characteristics, and has implications for understanding responses to interventions such as vaccination or mass drug administration. It introduces an additional regime to phylodynamics investigations, distinct from those previously described, namely the successive

replacement of antigenic clusters and the coexistence of fully non-overlapping antigenic repertoires.

3.2 Introduction

Understanding how pathogen variation is maintained and structured is a major goal of disease ecology. Ultimately, such research requires the synthesis of a wide set of fields covering evolutionary, epidemiological, immunological, and ecological dynamics (Bryan T. Grenfell et al. 2004). A major body of work on strain theory posits that the combined immunity of the host population can drive pathogens to differentiate into groups with reduced immune cross-reactivity (Sarah Cobey 2014; Sunetra Gupta et al. 1996) while intermediate hybrids are suppressed by the immunity of the population to the established parental strains. This organization into discrete strains is based on much broader phenomena described in theoretical ecology termed niche differentiation (Armstrong and McGehee 1980; Hutchinson 1959; Armstrong and McGehee 1980; Strong 1983; Leibold 1995) and to limiting similarity (Abrams P 1983). However in contrast with more recent work (Scheffer and van Nes 2006; Zinder et al. 2013) in both strain theory and in these models the role evolution through mutation and through invasion was not dynamically considered.

Non-immune based, evolutionary and ecological processes can also generate discrete pathogen strains. Reduced cross-immunity in these cases could be thought of as a by-product of these non-immune processes. For instance, different subtypes of influenza A (i.e A/H1N1, A/H3N2, A/H5N7) have reduced cross-immunity, each subtype represents a separate spillover of influenza from an avian reservoir, with a possible intermediate host (Parrish, Murcia, and Holmes 2015). Similarly, geographical isolation and host tropism play a role in the sub-division of viruses such as dog rabies and hantaviruses (Bourhy et al. 2008; Holmes and Zhang 2015) into 'phylogroups', although the immune implications of this division are less clear. In addition tissue tropism can also be involved in the differentiation of pathogen strains into groups, displaying different pathophysiology, as is the case in enteroviruses (Whitton, Cornell, and Feuer 2005). In theory, speciation processes can work in tandem with immune mediated ones to generate and maintain pathogen strains. Both immune and non-immune based mechanisms of co-existance are likely to play a role in shaping diversity patterns observed in rotavirus.

Rotavirus A (RVA) is a non-enveloped virus, with 11 double stranded RNA genome segments. The outermost capsid proteins VP4 (P types) and VP7 (G types) are often used for classification into genotypes (i.e. G1P[8], G3P[8], G9P[11] etc...) and are through to play a role in the generation of humoral immunity (J. Matthijnssens et al. 2008). There is evidence of increased protection against infections by the same genotype (Ward, Clark, and Offit 2010; Offit 1996; Guerra et al. 2015) yet a large component of immunity is heterotypic (Crawford et al. 2011; Velazquez et al. 1993; Leshem et al. 2014a). At a segment level, each segment type (i.e. G1, G12 etc...) corresponds to a monophyletic cluster, and is thought to reflect the cross-species transmission into humans from an animal source (Figure 3.1). There is a certain degree of antigenic change within individual G types (e.g. within G1) through mutation (Arista et al. 2006; S. M. McDonald et al. 2011). Multiple G and P types have been described, and in contrast with influenza, serotypes coexist and undergo frequent reassortment (S. M. McDonald et al. 2009). Several of the discovered G types are much more abundant than others, and in addition a small fraction of possible genotypes of

RVA are much more frequent (S. M. McDonald et al. 2009; J. Matthijnssens et al. 2008; Crawford et al. 2011). This has been attributed to a balance between preferred genome constellations and reassortment (S. M. McDonald et al. 2009), and there is evidence that these constellations have an improved fitness unrelated to immunity (Heiman et al. 2008).

The frequency of circulating rotavirus serotypes varies over time and across global geographic regions (Santos and Hoshino 2005a). On a local level circulating serotypes are often partially or fully replaced after being common in a region for several years (S. M. McDonald et al. 2012; M. Afrad et al. 2013; De Grazia et al. 2014), yet those can re-emerge at later periods. These replacement dynamics, taken together with changes in serotype prevalence following vaccination (Guerra et al. 2015; Zinder, Woods, and Pascual 2014) (see also Chapter 5) suggest selective pressures generated by host immunity can drive changes in circulating serotypes. In the US, there is little evidence that changes in RVA arise through mutation. It is rather immigration of existing and newly introduced serotypes that appears to dominate the population dynamics when seen through the sequence diversity of a single antigenic segment (Chapter 5) (Zinder, Woods, and Pascual 2014).

Here we use a model motivated by rotavirus to investigate how serotype structure is maintained, under different levels of specific (homotypic) and generalized (or heterotypic) immunity and varying rates of reassortment between strains. In contrast with previous work we consider antigenic novelty that can arise through the introduction or invasion of new segments, and reassortment rates that vary over a wider parameter range. Our work extends the dynamic regimes previously identified by strain theory and phylodynamic models, to include invasion by new segments that sweep through the population under sufficient reassortment. An improved understanding of how strain communities are generated and maintained provides a basis to also understand their expected response to various interventions.

3.3 Methods

Model description We use an individual-based SIR model based on (Bedford, Rambaut, and Pascual 2012b; Zinder et al. 2013) that explicitly tracks the chains of infection by viral lineages as well as the antigenic phenotype of every virus in the population. This model explicitly tracks the genealogy of each viral segment and therefore does not require the specification of a genotype to phenotype map. A virus in our model is composed of three antigenic segments (e.g. A, B, C), each segment allele (e.g. A_3) representing a unique serotype, with a combination such serotypes representing a viral strain or constellation of antigenic types (e.g. $A_1B_2C_2$). The model was modified to include reassortment and the genealogical tracking of each viral segment independently. Immunity and infectivity are different from the cited models and are further described next.

Immunity A host's risk of infection follows an exponential decline based on the number of infections and the number of unique antigenic segments to which that the host has previously been exposed:

$$e^{-\sigma_{\text{gen}} \cdot n_i - \sigma_{\text{spec}} \cdot \frac{n_s}{N_s}} \qquad (1)$$

where σ_{gen} is the level of heterotypic immunity, σ_{spec} is the level of homeotypic immunity, n_{i} is the number of infections a host has already experienced, n_s is the number of segments the host has been exposed to previously, and N_s is the number of antigenic segments of the virus has (N_s=3 for all the simulations reported here). Immunity is gained upon recovery from an infection.

Infectivity Only symptomatic infections transmit in our model. The first infection is symptomatic with probability x_0. The chances of a host transmitting an infection declines exponentially with each infection at a rate ξ according to

$$x_0 e^{-\xi \cdot n_i} \qquad (2)$$

New introductions: Antigenically novel segments are introduced at a rate i_I in the context of an existing background present in the population. That is, an introduction brings into the population one new segment in the background of an existing constellation. These new segments are derived from an initial parent segment at the same loci sampled at the beginning of the simulation, to reflect their distant coalescence, possibly in a source animal or region.

Model Parameters Model parameters are listed in Table 3.1

Co-infections Infected hosts are exposed to additional new infections. A co-infection may involve multiple circulating serotypes, and repeated infection with the same serotype.

Reassortment When a host infected with multiple serotypes transmits an infection, a single infecting serotype is generated and selected from the existing infecting serotypes. For the infecting serotype, a random parent serotype is selected, and with probability ρ each segment in this serotype is replaced by a random segment allele at the same loci, from all the infecting serotypes (including the parent strain).

Host Sampling A fraction ($2 \cdot 10^{-4}$) of infected hosts is sampled daily in the simulation, starting at the end of an initialization period lasting 50 years. When a host is sampled all infecting strains and segments are recorded. In addition the immune history of a fraction ($4 \cdot 10^{-6}$) of all hosts (infected and non-infected) is sampled.

3.4 Results

Epidemiological dynamics with varying degrees of specific immunity. To calculate how variable, or stable, are the infection levels with varying parameters, we plot the coefficient of variation for incidence. The coefficient of variation (CV) was calculated for incidence as the standard deviation divided by the mean, when sampled during two year long time intervals. The CV is plotted as a function of the strength of specific immunity and the rate of reassortment (Figure 3.2A). For the specified level of generalized immunity and contact rate, higher specific immunity leads to more epidemic dynamics (Figure 3.3A, B). Also, more stable infection levels are established with a higher contact rate (Figure 3.2B).

Phylodynamic patterns in the absence of reassortment

Previously a dynamical regime in which strains co-exist with minimal cross-immunity, was described in the context of a limited pool of variant antigenic types. Here, we find the presence of such a regime, in the context of an open antigenic pool, with novelty introduced through zoonosis from a source of unlimited antigenic novelty (Figure 3.3A, C). The lack of (or lower) segment overlap (Figure 3.3C) is representative of immune-mediated niche partitioning as established by previous studies. In this model, we assume that antigenic novelty is generated by introductions from a single source, and newly introduced segments coalesce at this source before host sampling is initiated. Segments are introduced to the system on the background of existing strains. A lack of overlap between common strains on short time scales (Figure 3.3C) indicates the extinction of background strains carrying these new segments. A strain with a newly introduced segment has a clear competitive advantage over the background strain, and will compete more strongly with the parental strain, than with antigenically distinct strains due to immunologic similarity (Figure 3.4).

Epidemic replacement dynamics, similar to those of influenza, occur with higher specific immunity. Epidemiologically, this replacement regime is similar the one for e spindly A/H3N2 influenza like trees (Figure S3.1). Because new segments are simulated to be generated through zoonosis or through immigration, segments have a deep coalescence in the source animal or region (Figure 3.3B, Figure S3.1). In contrast with the regime in which strains coexist, segments are shared between subsequent strains (Figure 3.3D).

Phylodynamic patterns of coexistence with reassortment in the context of antigenic novelty. With the introduction of reassortment (Figure 3.5) strain diversity increases (Figure S3.2 A,B). Before, the lack of reassortment prevented the sweeping of a new advantageous segments across multiple existing backgrounds. With reassortment new segments which confer competitive advantage can travel across multiple backgrounds (Figure 3.6). With an intermediate reassortment rate (Figure 3.4A), the signature of niche partitioning is reduced (Figure 3.5C), and with a higher reassortment rate

(Figure 3.5B, D) it is lost. These different levels of strain structure with reassortment are also evident when linkage disequilibrium (D') between pairs of segments is calculated (Figure 3.7A). Finally, reassortment rates have a limited impact on the diversity of segments when specific immunity is low, and reduce segment diversity on shorter durations of time when specific immunity is high (Figure S3.2, C).

The effect of reassortment on serotype community stability

We measured the time it takes for serotype communities sampled in different years to diverge (Figure 3.7B). With increasing specific immunity, the population is less stable. This is in contrast with the dynamics observed for limited antigenic diversity, where stronger specific immunity stabilized niche partitioning (Zinder et al. 2013). Increased reassortment rates also reduce the stability of the serotype population.

3.5 Limitations

Our work does not consider other, non-immune interactions between segment types, such as epistasis (Sanjuan, Moya, and Elena 2004; Heiman et al. 2008) which can lead to conserved genome constellations in rotavirus. Epistasis can also reinforce, or be the consequence of, immune-mediated mechanisms that have created or maintain serotype structure.

The model used, assumes a simple population structure which includes a homogenously mixing homogenous population of hosts. Work on immune mediated strain structure has shown some robustness to a certain level of population spatial structure (Caroline O. Buckee et al. 2011). Further work, on the role of population structure on the maintenance and formation of serotypes is necessary but goes beyond the scope of this manuscript.

In addition, a small population size of 5M hosts was used. With a growing effective population size genetic and antigenic diversity is expected to increase (Zinder et al. 2013). However, we do not expect changes in the trends observed with changes in specific immunity and reassortment rates.

We used a simple SIR model, which is symmetric with respect to immunity generated against each segment, and against all strains. We used this simplifying assumption for model comprehension and simplicity. Also, some models have considered different a fitness for different rotavirus serotypes (Pitzer et al. 2009). Similarly we assumed all strains carry the same fitness for model comprehension and simplicity. Work on the implication of different fitness levels for different alleles has resulted in interesting observations in models of malaria (Artzy-Randrup et al. 2012b).

3.6 Discussion

The strain structure and temporal stability of pathogen communities in face of invasion is strongly influenced by the level of reassortment they exhibit. When a novel antigenic allele at a given segment is introduced into a population, it can disturb the existing serotype structure. In our model, when reassortment rates are low, strain communities maintain a signature of niche partitioning as the novel allele largely replaces the same genetic background upon which it arose (Figure 3.3A, C, Figure 3.4). In previous work (Caroline O. Buckee et al. 2011; Sunetra Gupta et al. 1996; Recker et al. 2007; Zinder et al. 2013) strong specific immunity was necessary to maintain immune mediated strain structure. In addition in (Caroline O. Buckee et al. 2011; Recker et al. 2007) in previous work, immunity which is heterotypic was not considered, and the immunity structure favored the gain of protection based on a single shared segment between a challenging strain and previous infections. In this work, we use an immunity structure that lends itself more readily to observations of cross-protection in rotavirus (Figure S3.3) (Velázquez et al. 1996; Pitzer et al. 2009; Gladstone et al. 2011a). In our results, in the absence of reassortment, the selective advantage of a newly introduced antigenic allele, in competition with the parent background on which it was introduced, maintains the signature of niche partitioning at low levels of specific immunity.

With increasing reassortment rates, a newly introduced allele provides strains with a competitive advantage, and can sweep across multiple genetic backgrounds before significant levels of host immunity build against it, thus disrupting the viral strain structure to different degrees (Figure 3.5, Figure 3.6). Previous work had demonstrated a certain resilience of immune-mediated strain structure to reassortment and recombination (Caroline O. Buckee et al. 2011). However, this work did not consider the introduction of antigenic novelty from outside the local transmission system and considers only a limited pool of antigenic alleles.

In the population genetics literature, sexual reproduction, including reassortment, is hypothesized to increase the rate of adaptation (Fisher 1930; Muller 1932) and reduces what has been termed clonal interference. For example, in viral pathogens reassortment has been shown to facilitate the dissemination of antiviral drug resistance in seasonal A/H3N2 influenza (Simonsen et al. 2007). Despite newly introduced alleles having a fitness advantage that allows them to travel across multiple backgrounds, both incidence (Figure S3.4) and incidence variability (Figure 3.1) largely remain the same for this regime in our model. This is in keeping with observational data of rotavirus, where yearly incidence is quire predictable, but serotype turnover can be substantial (M. Afrad et al. 2013). This behavior contrasts with the epidemiological dynamics seen in the replacement regime of successive antigenic clusters characteristic of H3N2 influenza, where each replacement underlies an increase in attack rates (Koelle et al. 2006). The rate of stochastic extinction does increase however for higher specific immunity in our simulations, as evident in the failed runs represented by missing dots. It is possible that this is the result of the impact of generalized immunity in reducing secondary infections.

Although incidence remains relativly stable with reassortment, with increasing reassortment rates, the strain population structure becomes less stable (Figure 3.7B). This effect is amplified with stronger specific immunity, the consequence of an increased selective coefficient for newly introduced segments.

Our results are relevant to other pathogens and dynamical systems in the context of other recombining pathogens which generate antigenic novelty and experience reassortment or recombination, including malaria, influenza, and pneumococcus among others, and should be considered in the interpretation of data. The degree to which strains are maintained by immune-mediated niche partitioning in highly recombining and diverse pathogens such as malaria remains unclear (Artzy-Randrup et al. 2012a). The statistical detection of partial niche partitioning will be more difficult in the context of a sweeping regime, and this needs to be considered in analyses of strain structure. In a broader context, the generation of antigenic novelty is equivalent to the generation of new resource axes, in ecological models of species coexistence (Scheffer and van Nes 2006).

Targeted interventions, identifying shared commonly recognized antigens, may identify transient sweeping alleles as being most common. These alleles have no inherent fitness differences compared to other alleles, with the exception of a transiently higher host susceptibility. As such, targeted vaccination against these alleles may not offer higher efficacy, and it will be necessary to survey antigenic diversity for continued periods of time to identify the true extent of antigenic variation. In contrast general non-immune targeted interventions may provide a mechanism to reduce pathogen diversity by reducing co-infection and reassortment, which in turn may benefit the acquisition of immunity.

3.7 Tables

Table 3.1. Model Parameters

Description		**Value**	**reference**
contact rate of symptomatic infection*	β	*1.0 day^{-1} or 3.514 day^{-1}*	(Pitzer et al. 2009)
duration of infection	$\frac{1}{\rho}$	*7.0 days*	(Pitzer et al. 2009)
homotypic immunity	σ_{ge}	*0.2-1.6*	(Velázquez et al. 1996; Pitzer et al. 2009; Gladstone et al. 2011a) *
heterotypic immunity	σ_{sp}	*0-1.6*	
mutation rate	μ	*0*	
introduction rate	i_I	*0-32 year^{-1}*	
probability of first infection to be symptomatic*	x_0	*0.47*	
reduction in infectivity with subsequent infections	ξ	*0.62*	
population size	N	*5,000,000*	
number of viral segment types	N_S	*3*	

* also see Figure S3.3

3.8 Figures

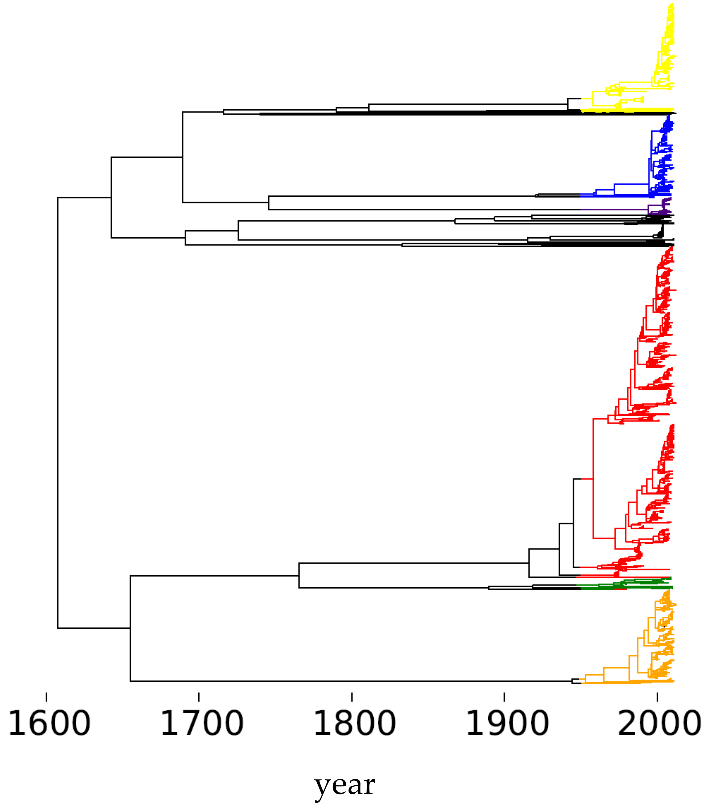


Figure 3.1. Phylogeny of rotavirus A VP7 segment (G type) color coded by the main segment serotypes (Woods RJ et. al 2015, unpublished).

1911 Annotated VP7 sequences were collected from GenBank and a phylogenetic tree was reconstructed using BEAST. Each G type (in order of top to bottom): G2 (yellow), G9 (blue), G3 (purple), G1 (red), G4 (green) , G2 (orange) and other (black), is thought to represent an introduction from an animal source.

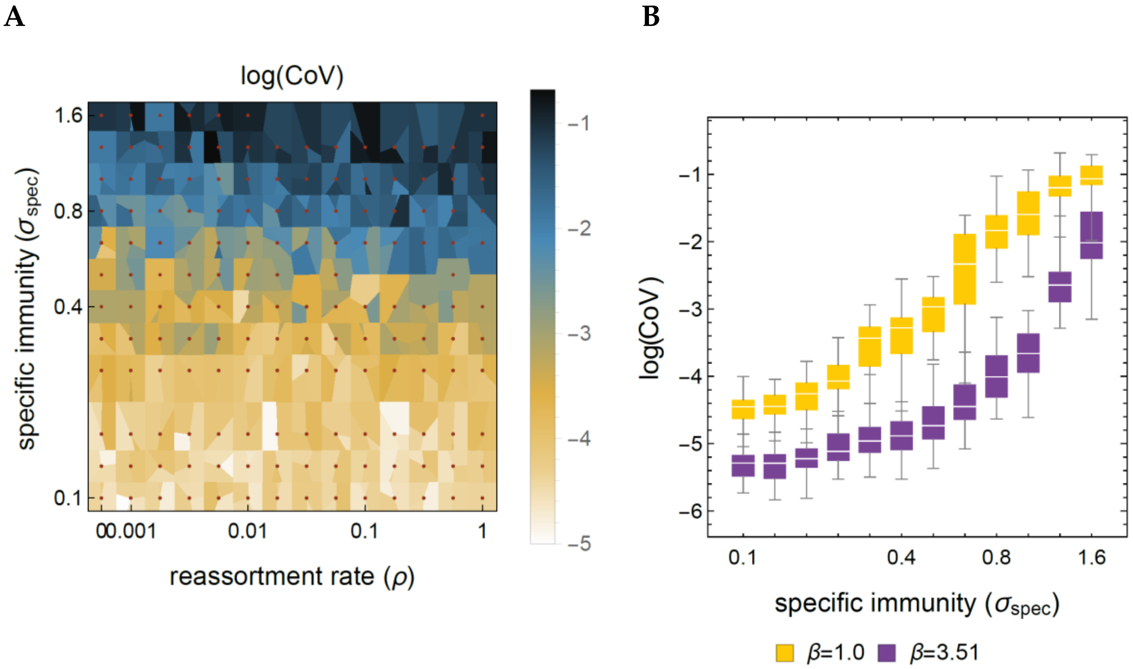
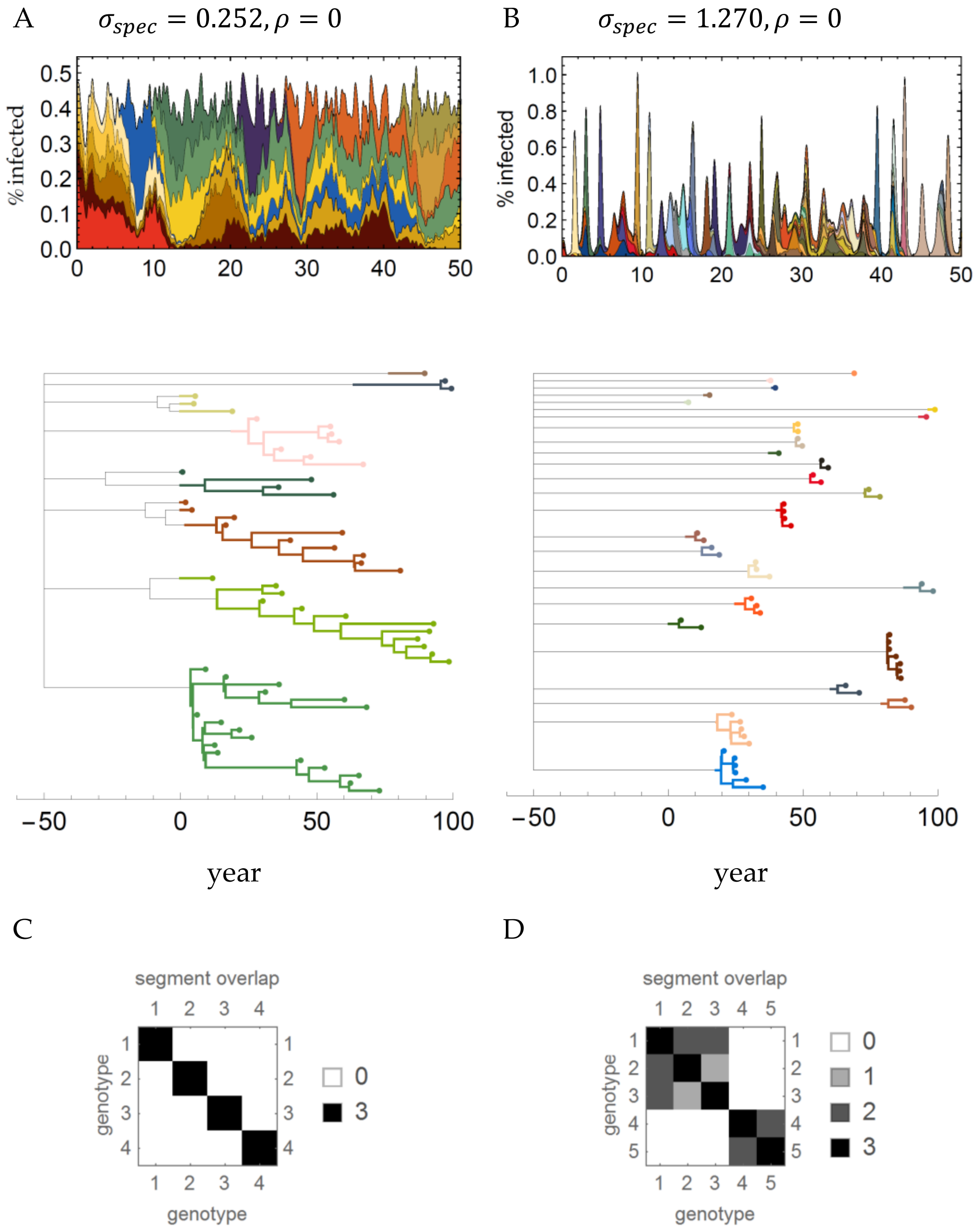


Figure 3.2. Incidence variability with varying specific immunity, reassortment, and contact rates

A. The biannual incidence coefficient of variation was calculated and is plotted as a function of the strength of specific immunity and of the reassortment rate. The reassortment rate is the probability at which a segment is chosen at random rather than from the parent strain in cases where the infecting host is co-infected. Dots represent parameters or which simulations were run. For this figure the contact rate is $R_0 = 7.0$, generalized immunity $\sigma_{gen} = 0.4$, introduction of new antigenic segments at a rate of i_I=8.0[$\frac{1}{year}$], while lower and higher introduction rates were considered in a sensitivity



$\sigma_{spec} = 0.252, \rho = 0$

$\sigma_{spec} = 1.270, \rho = 0$

Figure 3.3. Serotype co-existance and replacement dynamics, with antigenic novelty driven by the introduction of new segments, in the absence of reassortment

A-B Prevalence, color coded by serotype, together with the phylogenetic tree of a single segment for a longer time period. **C-D** Serotypes were sampled from a short time window (approx. year 22-28) ordered by their prevalence, until 80% of prevalence was covered for the purpose of capturing the main circulating serotypes. A matrix was constructed plotting the number of shared segments between these common serotypes. The matrix is ordered with minimum bandwidth.

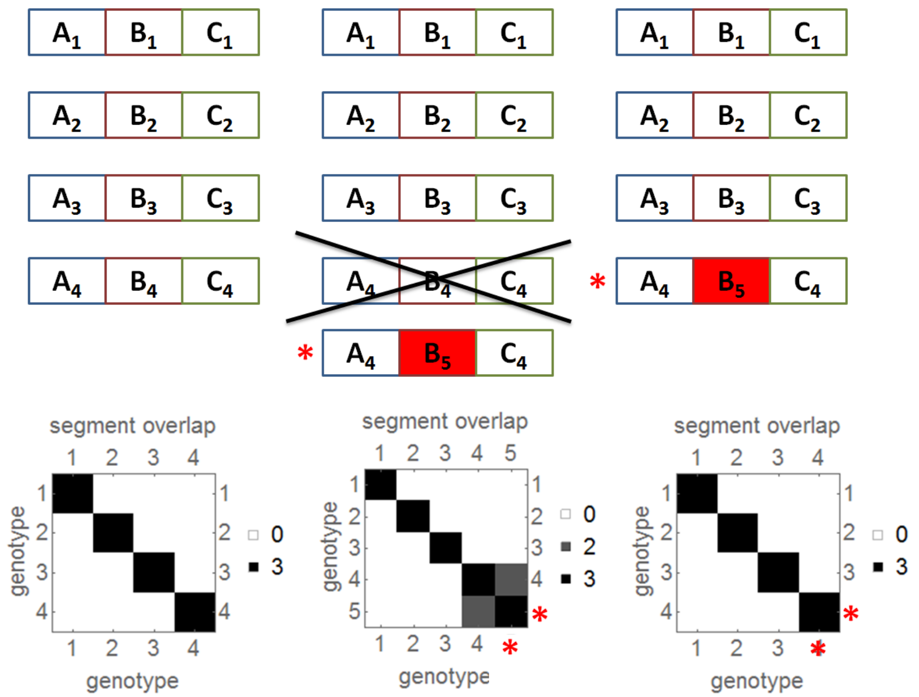


Figure 3.4. The introduction of a novel antigenic segment in the absence of reassortment (left) With the absence of antigenic novelty immune mediated strain structure is maintained and strains do not share segment alleles. **(center)** A new introduction of the antigenic segment B5, leads to competition between the parent strain which shared two alleles with the new strain, but suffers from a competitive disadvantage as the population is susceptible to the new allele. The remaining serotypes are disturbed to a lesser extent because they only share heterosubtypic immunity with the newly introduced serotype. **(right)** A replacement of the background strain by a strain with the new allele segment follows, generating an immune mediated non-overlapping strain structure.

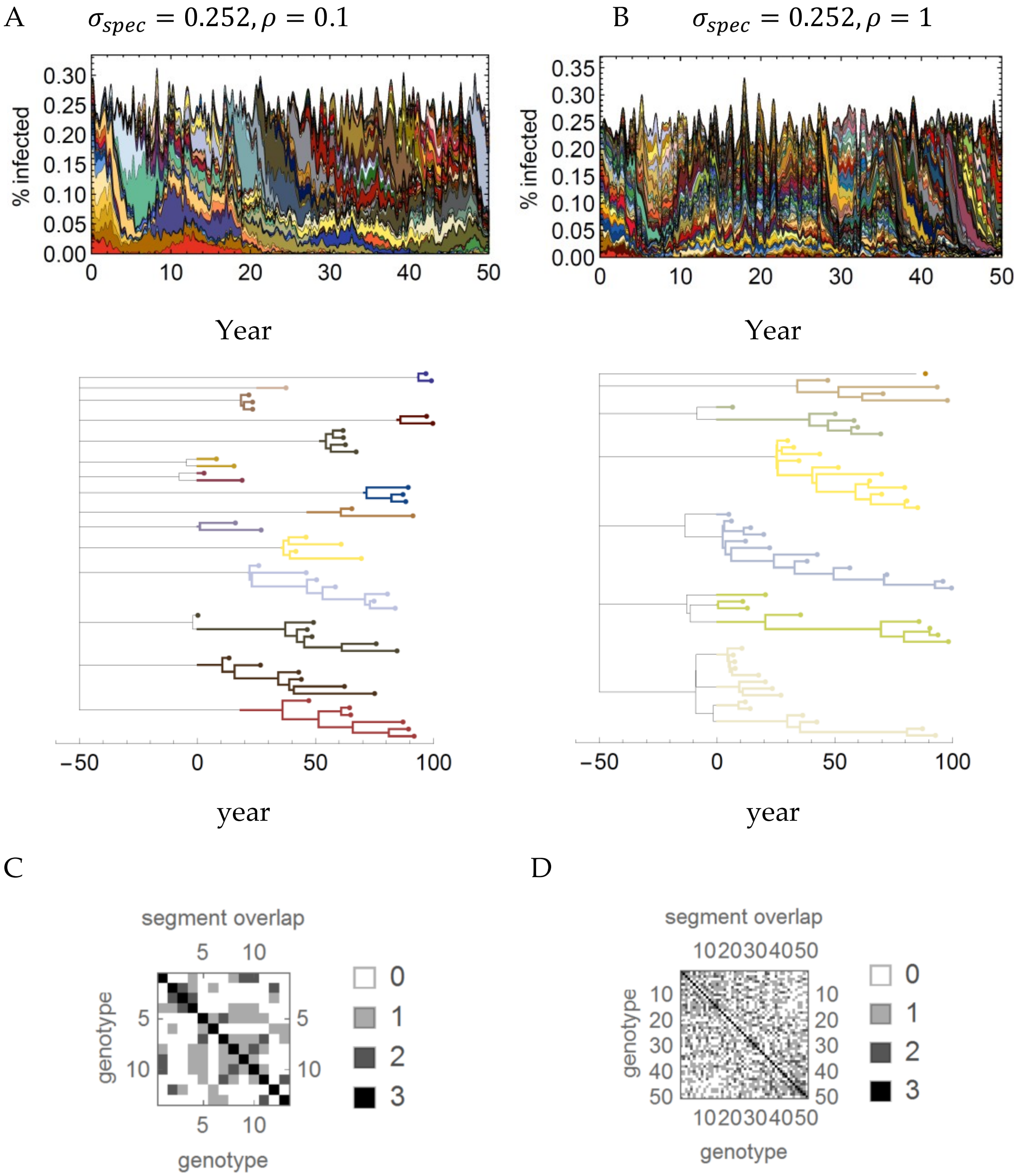


Figure 3.5. Serotype co-existance and replacement dynamics, with antigenic novelty driven by the introduction of new segments, in the presence of reassortment

A-B Prevalence, color coded by serotype, together with the phylogenetic tree of a single segment for a longer time period. **C-D** Serotypes were sampled from a short time window (approx. year 22-28) ordered by their prevalence, until 80% of prevalence was covered. A matrix was constructed plotting the number of shared segments between these common serotypes. The matrix is ordered as to minimze its bandwidth.

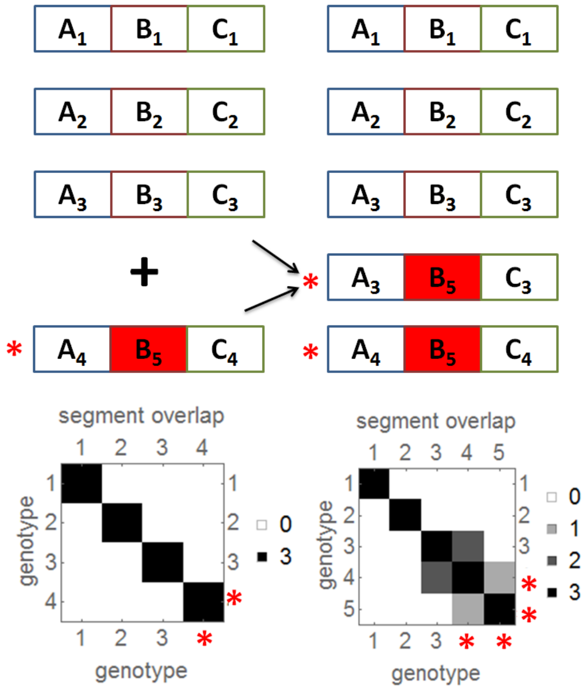


Figure 3.6. An introduction of a novel antigenic segment in the presence of reassortment **(left)** With the introduction of a new allele on a background strain, reassortment of the new allele on alternative backgrounds is possible. **(right)** A reassortant strain (A_3,B_5,C_3) has a temporary advantage and circulate, until sufficient immunity against the new segment allele is generated.

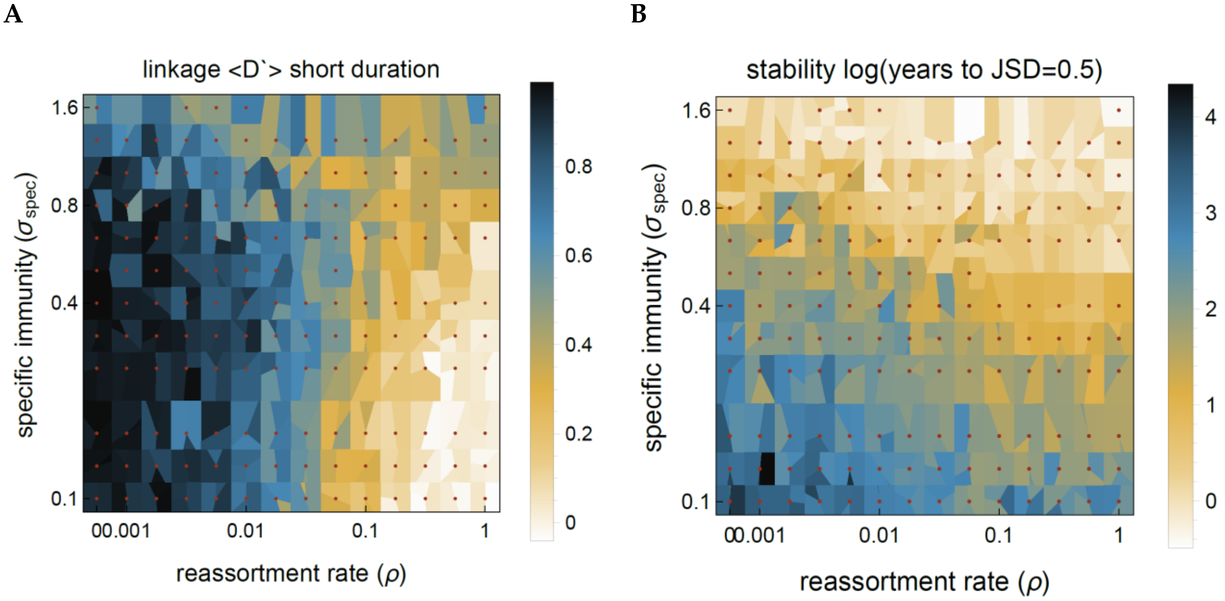
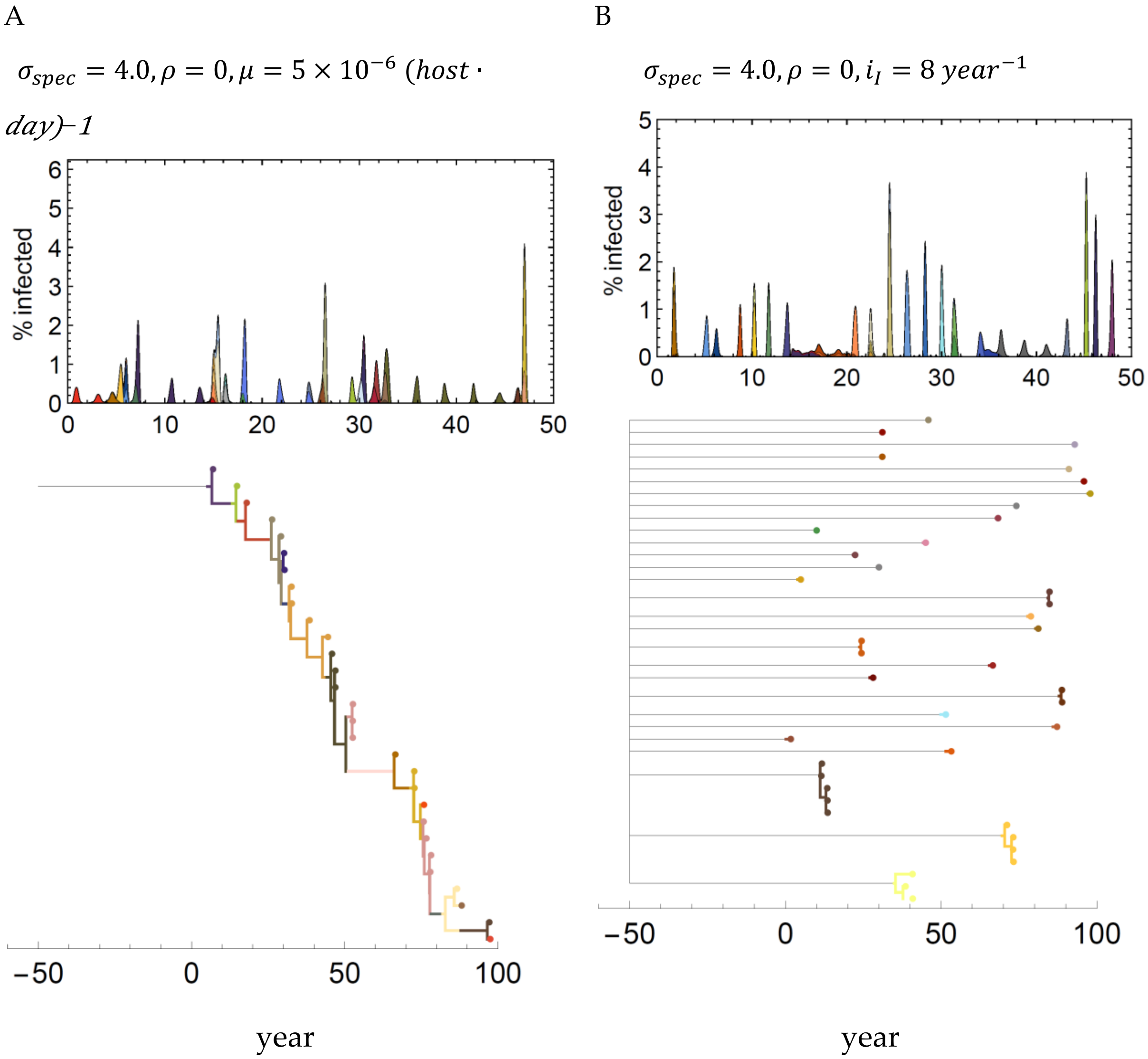


Figure 3.7. Segment serotype linkage, and community stability with varying specific immunity and reassortment rates

A Average linkage disequilibrium <D'> between randomly sampled pairs of segments present together in the same serotype. Allele frequencies used were based on the frequencies of segments during short windows of time (6.25 years) **B** The duation of time it takes for serotype communities sampled in different years to diverge. The divergence metric used was Jensen-Shannon-Divergence (JSD) and the time to JSD=0.5 was estimated. JSD reflects the disimilarity between two distributions of serotypes, taking into account that differences in the presence and absence of more common serotypes is more informative (Lin 1991).

3.9 Supplement



A

$\sigma_{spec} = 4.0, \rho = 0, \mu = 5 \times 10^{-6}\ (host \cdot$

$day)–1$

B

$\sigma_{spec} = 4.0, \rho = 0, i_I = 8\ year^{-1}$

Figure S3.1. Phylodynamic patterns of replacement
A Antigenic novelty is driven by mutation **B** Antigenic novelty is driven by zoonosis

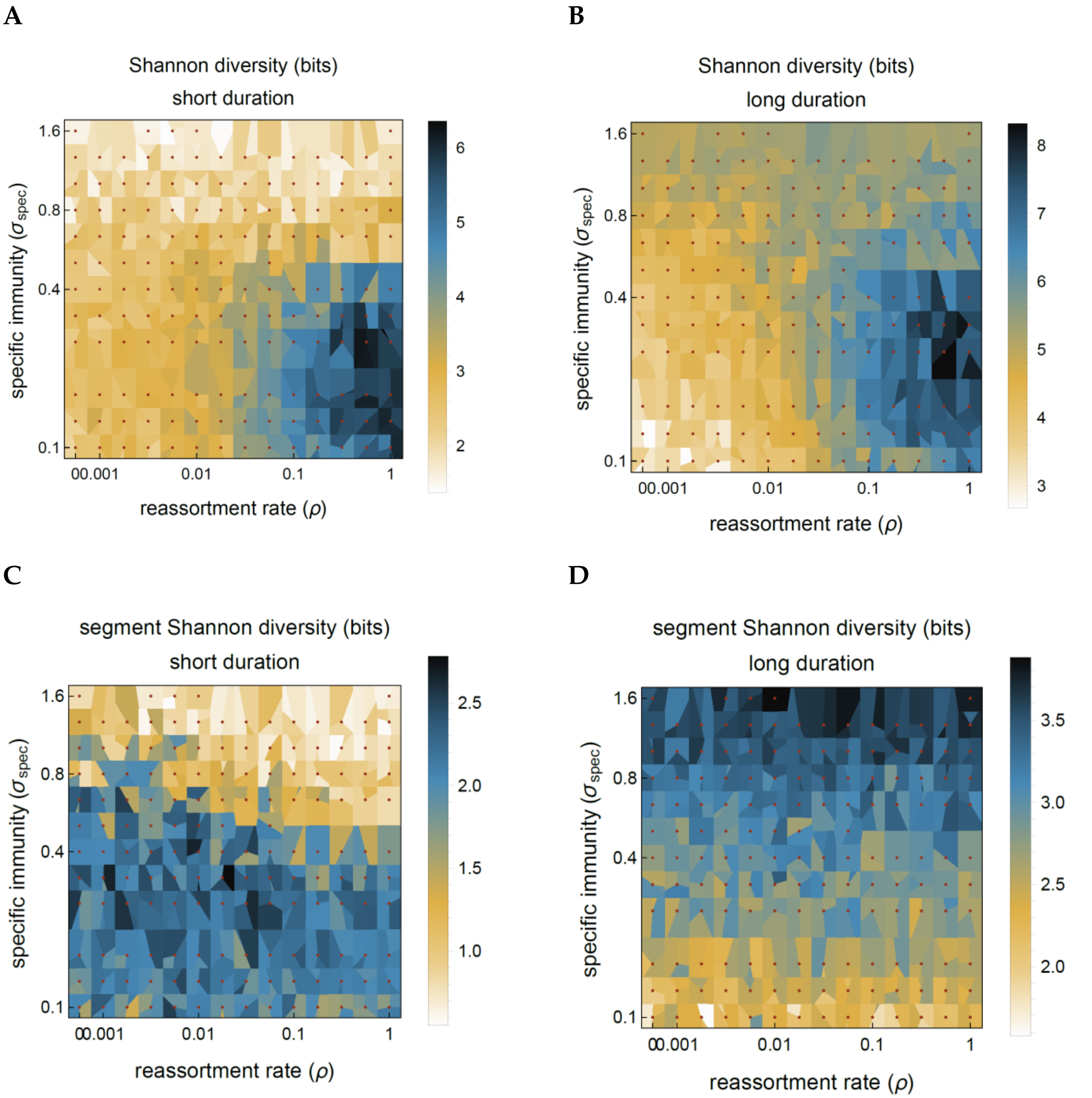


Figure S3.2. Segment and serotype diversity with varying specific immunity and reassortment rates

A Shannon diversity of strains sampled during short (1.5 year) windows and averaged **B** Shannon diversity of strains from a long (50 year) time period **C** The average Shannon diversity of a single segment sampled during short (1.5 year) windows and averaged **D** The average Shannon diversity of a single segment sampled during from a long (50 year) time period.

A.   B.

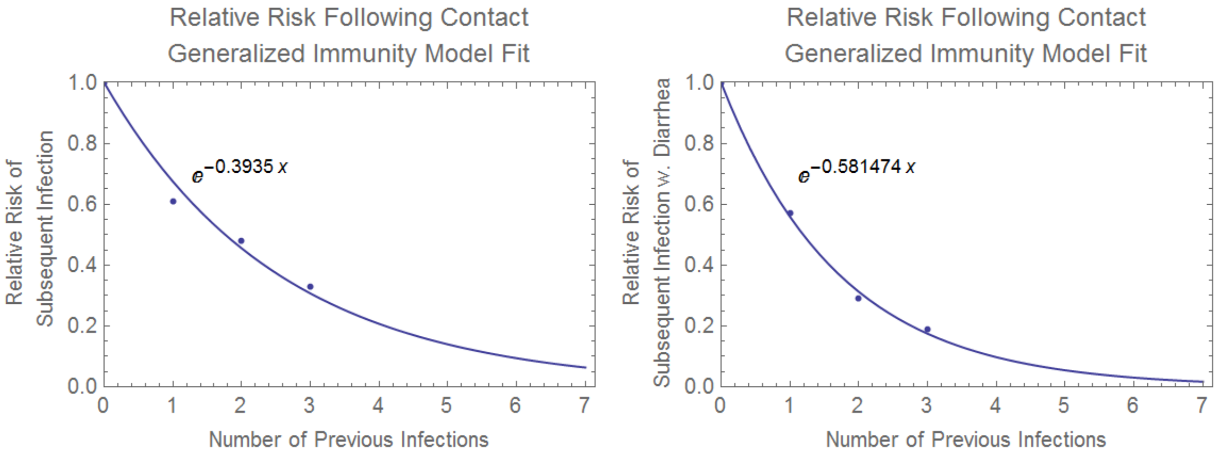


Figure S3.3. Estimated bounds of generalized immunity

Exponential model fit based on data from Gladstone et al. (Gladstone et al. 2011b). The decline in relative risk can be used as an upper bound on the expected amount of generalized immunity in the model. **A.** Relative risk based on the recurrence of symptomatic and asymptomatic infections **B.** Relative risk based on the recurrence of symptomatic infections (with Diarrhea).

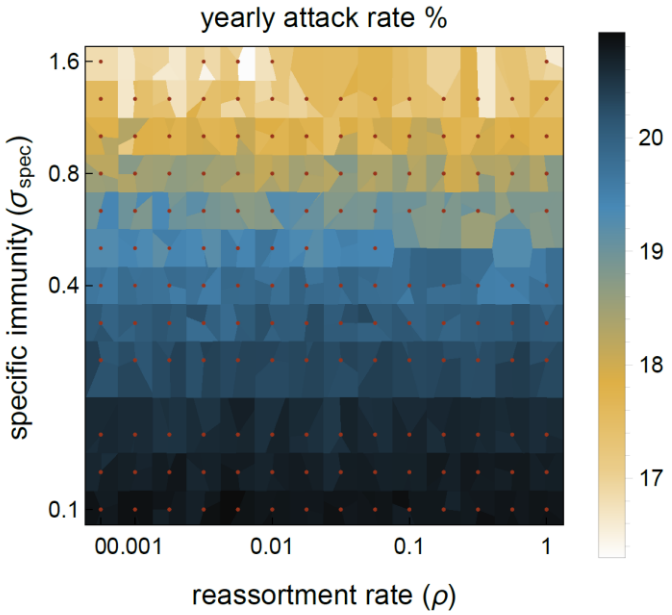


Figure S3.4. Yearly attack rate with varying specific immunity and reassortment rates

Section II - Inference

Chapter 4. Seasonality in the migration and establishment of H3N2 Influenza with epidemic growth and decline

4.1 Abstract

Background Influenza A/H3N2 has been circulating in humans since 1968, causing considerable morbidity and mortality. Although H3N2 incidence is highly seasonal, how such seasonality contributes to global phylogeographic migration dynamics has not yet been established. In this study, we incorporate time-varying migration rates in a Bayesian MCMC framework. We focus on migration within China and between China and North-America as case studies, then expand the analysis to global communities.

Results Incorporating seasonally varying migration rates improves the modeling of migration in our regional case studies, and also in a global context. In our global model, windows of increased immigration map to the seasonal timing of epidemic spread, while windows of increased emigration map to epidemic decline. Seasonal patterns also correlate with the probability that local lineages go extinct and fail to contribute to long term viral evolution, as measured through the trunk of the

phylogeny. However, the fraction of the trunk in each community was found to be better determined by its overall human population size.

Conclusions Seasonal migration and rapid turnover within regions is sustained by the invasion of 'fertile epidemic grounds' at the end of older epidemics. Thus, the current emphasis on connectivity, including air-travel, should be complemented with a better understanding of the conditions and timing required for successful establishment. Models which account for migration seasonality will improve our understanding of the seasonal drivers of influenza, enhance epidemiological predictions, and ameliorate vaccine updating by identifying strains that not only escape immunity but also have the seasonal opportunity to establish and spread. Further work is also needed on additional conditions that contribute to the persistence and long term evolution of influenza within the human population, such as spatial heterogeneity with respect to climate and seasonality.

4.2 Background

Seasonal influenza causes considerable morbidity and mortality, and presents a complex problem due to the intimate relationship between its evolution and epidemiology. The WHO estimates influenza A causes between a quarter to half a million deaths worldwide annually (WHO 2009) with yearly epidemics in the US resulting in tens of thousands of deaths (CDC, n.d.). The economic burden of seasonal influenza in the US is estimated in billions of dollars in health care costs (Walsh and Maher 2011; Simonsen 1999). Influenza A is classified into subtypes (*e.g.* H1N1, H2N2, H3N2) based on its envelope glycoproteins hemagglutinin and neuraminidase, the two major targets of humoral immunity. Multiple zoonotic introductions of influenza A subtypes to the human population have taken place, with H3N2 and H1N1 being the most prevalent subtypes whose continuous endemic circulation has lasted decades.

Sequences sampled since the introduction of H3N2 into the human population in 1968 serve as primary data for phylodynamic inference that seeks to understand joint epidemiological and evolutionary dynamics. H3N2 exhibits rapid geographic spread and turnover rates. On a time scale of several years, all previously circulating lineages are globally replaced by new ones, sharing a single and more recent common progenitor (2-8 years) in the past (Fitch et al. 1997; Bush, Bender, et al. 1999; Zinder et al. 2013). The nature of this swift global turnover remains an open question in terms of both its geographical path and its underlying mechanisms.

It has been proposed that the evolution of the virus is predominantly maintained by a reservoir in the tropics, where annual epidemics experience less severe bottlenecks, which increases the likelihood of local persistence (Andrew Rambaut et al. 2008a). Further research (Bedford et al. 2010b; Bahl et al. 2011) has suggested that the ancestry of global influenza lineages are found mainly in East and Southeast Asia (SEA) rather than the tropics in general, where a network of temporally overlapping epidemics with limited local persistence (Russell et al. 2008) maintains continuous circulation. An alternative hypothesis suggests that a global metapopulation exists in which temperate lineages frequently revisit the tropics at the end of a seasonal epidemic (Bedford et al. 2010b; Bahl et al. 2011). The much lower contribution of South America (SA) and other subtropical and tropical communities to the long-term evolution of the virus has been attributed to demographics and air travel connectivity (Bedford et al. 2010a; Philippe Lemey et al. 2014).

Epidemiological and molecular (phylogenetic) studies of influenza use different sources of primary data and their findings are not yet in complete concordance. Both methodologies show support for the common occurrence of strong seasonal epidemics followed by deep troughs limiting viral diversity, for the existence of multiple viral introductions during a season, and for the lack of sustained viral persistence between epidemics (Andrew Rambaut et al. 2008a; Martha I. Nelson et al. 2007; M. I. Nelson et al. 2006; C Viboud et al. 2004; Cécile Viboud et al. 2013; Martha I. Nelson and Holmes 2007). In contrast, several epidemiological observations, such as spatially structured diffusion patterns (J. Gog et al. 2011; Alonso et al. 2007) and hierarchical spread driven by population size and distance, in gravity models (Balcan et al. 2009), have not been evident from phylogenetic methods (Cécile Viboud et al. 2013; Cécile Viboud et al. 2006; Eggo, Cauchemez, and Ferguson 2011).

Recently, migration seasonality has been incorporated into phylogenetic analysis, in Bahl et al. (Bahl et al. 2011) where alternative seasonal migration patterns from and to the tropics were considered, and in Bielejec et al. (Bielejec et al. 2014) where support for seasonal, rather than constant, H3N2 global migration patterns was established. In Lemey et al. (Philippe Lemey et al. 2014), migration rates were assumed to be constant throughout the year, but alternative variables, used as surrogates for measuring the effect of incidence seasonality were tested as predictors of these invariable rates. Specifically, the following features were considered: the overlap in incidence between the source and destination community, the source incidence product with the destination growth rate, and the relative timings of peak incidence. However, these were not informative in predicting these migration rates.

It follows that we still lack an understanding of the relationship between changing incidence throughout the year and the level and timing of immigration and emigration. Specifically, we are interested in whether the timing of migration events with phylogeographic consequence is mainly the outcome of an increased introduction effort from the source (propagule pressure), or is determined by the receptiveness (or the effective R0) of the destination community to introductions. Additionally, we seek quantitative support for the fundamental hypothesis that viral persistence relates to reduced seasonal bottlenecks in incidence.

By using a model that can infer migration rates that differ from season to season, we find clear seasonal migration patterns between and within different global communities. We identify epidemic incidence and growth as predictors of these patterns. Our approach has similarities and differences with (Bielejec et al. 2014) on which we expand upon in the Method's section. Notably, we are able to integrate over alternative partitions of the year when measuring the seasonality of migration, and to explicitly incorporate the timing of migration events in addition to migration rates through stochastic mapping.

4.3 Results and Discussion

Seasonal Migration to and from, and Local Persistence within, North America

For this analysis, we are interested in measuring the local persistence of temperate climate North American (NA) H3N2 lineages, and in inferring the seasonal timing of introductions to and from the global community to NA. For this purpose, we partition globally sampled sequences based on their country of collection. Sequences from the

US and Canada are designated as NA. A representative sample of “other” global sequences (OT) is randomly sampled in equal proportion from every month in multiple geographic regions. The frequency of sequence sampling (Table 4.1, Figure 4.S1) in NA exhibits a winter seasonal pattern, while representative global sequences reflect our equal proportion sampling (Figure 4.1A). Phylogenetic trees from both the neuraminidase and hemagglutinin proteins are reconstructed based on nucleotide data and sampling time alone. The likelihood of a specific realization of a migration model is marginalized across this posterior distribution of phylogenetic trees in an additional step (Pagel, Meade, and Barker 2004), in which tree likelihood is calculated based on the tree topology, sequence collection locations, and the specified model parameters.

Bayesian variable selection is used to decide whether there is sufficient support for migration between communities, and whether seasonal migration rates, measured in terms of migrations per lineage per unit time, are supported. A seasonal migration model has a higher marginal likelihood than a non-seasonal one (Table 4.2), and supports migration that is seasonal both to (BF=70) and from NA (BF>150). We used this best supported model in further inference.

For each migration model parameterization (Figure 4.S2) and each tree (see Methods), stochastic mapping is used to sample the internal state of branches and the timing of migration events. Each of these stochastic mappings results in a fully geographically annotated tree sample, including the timing of migration events within branches. Using these data, we can explicitly sample alternative phylogenetic histories from the posterior distribution of trees and model parameters.

Figure 4.1C shows the mean number of migration events to and from NA in different months. The number of migration events is summed across all lineages, and reported as a yearly rate, during an average month of the year, together with its Bayesian credible interval (referred to as 90% CI). Immigration to NA (Figure 4.1C) grows during late summer and declines before the end of winter. In contrast, emigration from NA is highest during winter months and during spring. Such emigration from NA during spring could be suggestive of the seeding of epidemics in tropical or southern hemisphere SA (Bedford et al. 2010b).

For this and following figures we report the average number of stochastically mapped migration events per unit time in different months, instead of the directly inferred migration rates on a per lineage basis. The number of events per unit time differs from phylogenetic migration rates which take the perspective of a single lineage. The difference is apparent if one considers that a constant migration rate from a specific location measured in terms of a single lineage will necessarily mean many more migration events during an epidemic. We nevertheless report the support (BFs) for seasonal migration, which is used in model selection, with respect to per-lineage phylogenetic migration rates, the unit for which they were originally inferred.

When inferring migration rates we divide the year in two, with constant migration rates inferred individually in each partition. Alternative partitions of the year are weighted according to their likelihood using the MCMC. Each of these partitions and migration model parameterizations is followed by ancestral state reconstruction and subsequent stochastic mapping. The number of migration events per month is counted, and is an estimate of the actual number of migration events expected in the different months, that were captured in our dataset. This can lead to observed positive migration rates during periods with low or zero incidence, such as is the case in NA during its summer incidence trough.

The duration for which tips have been in NA, as traced across the multiple trees is presented in Figure 4.1B. Sampled NA sequences are inferred to have arrived to NA, 3.8 months (median) before their sampling, with 8% (2-16%) of tips (median and 90% CI) persisting locally for more than one year. Local persistence times for tips are not equivalent to the distance from the trunk (*e.g.* (Bedford et al. 2010b; Russell et al. 2008)). Phylogenetic reconstruction of locations includes stochastic mapping in addition to ancestral state reconstruction (such as in *e.g.*(Philippe Lemey et al. 2014)) to resolve branches and uses time variable migration rates which were better supported compared to constant ones.

Seasonal Migration to and from, and Local Persistence within North China and South China

China is a key source location for H3N2 influenza [79, 97]. Here we measure how differing seasonality in northern and southern China relates to H3N2 migration within China and to and from the global community. For this purpose, we use the number of sequences collected in different months to establish broad seasonal patterns in Chinese provinces and several individual cities. For each province, an approximate seasonal pattern is established based on the number of samples in two-months bins, and a clustering algorithm is used to partition the provinces by their incidence seasonality (Figure 4.S3). This process results in the partitioning of Chinese provinces and associated sequences into a northern (NC) and southern (SC) cluster (Figure 4.2B). H3 and N2 sequence sampling (Table 4.3, Figure 4.S4) in the northern cluster peaks in winter, while the southern cluster sustains two annual peaks (Figure 4.2A). A representative sample of global sequences was included and designated as (OT). A geographically based partition of China into two was suggested in Du et al. (Du et al. 2012) and an alternative partition into three seasonality based clusters was reported in Yu et al. (Yu, et al. 2013). Although, we find a similar partition to Yu et al. when a division into three clusters is explicitly specified to the algorithm, this partition was not supported over a division into two clusters by our clustering algorithm.

Models that allow for seasonal migration rates between SC and NC and the global community (Table 4.2) and include seasonality based on the destination community (Appendix A 2.7) have the highest marginal likelihood (-561.0 to -561.8) compared to non-seasonal models (marginal likelihood=-587.0, -588.4 without and with variable selection, respectively). Migration seasonality based on the source community has an intermediate marginal likelihood (-568.6). This suggests that both conditions at the source and at the destination community contribute in determining migration rates between the global community, SC and NC. We used a model that includes variable selection for migration and for seasonality between each pair of communities in further analysis (Methods, Appendix A 2.17).

We identify strong support (BF>250 both directions) for migration between SC and NC, and between SC and the global community (BF>250 both directions). There is support for a model without migration between the global community and NC

(BF=14), and no indication whether migration between NC and the global community (BF=1.9) is present during the sampled years (Figure 4.S5). Figure 4.2D shows the mean number of migration events for different months (black) and their 90% CI across the posterior distribution of model parameters.

Migration events (Figure 4.2D) from NC to SC peak in January following a peak in incidence in NC and when SC experiences on average a rise in incidence. Migration of SC lineages to NC peaks in October, during epidemic decline in SC and prior to the full onset of the winter epidemic in NC. These patterns are somewhat similar to emigration patterns from NA which remains high during April and May following a decline in incidence. They suggest a potential role for both the propagule pressure (push from the source) and favorable conditions at the destination (fertile ground) in determining effective migration rates.

Local persistence is expected to change with variation in yearly seasonal patterns. We measure the time for which tip ancestry persist locally in SC and NC. Sequences sampled in SC are expected to have emigrated to SC 4.3 months (median) prior to their sampling, where we infer 10% (4-20%) of tips (median and 90% CI) to be locally persistent for more than a year (Figure 4.2C). NC sequences are inferred to have been in that location for 2 months (median) with 0% (0-2%) of tips (median and 90% CI) locally persistent for more than a year. Since NC lineages are most often inferred to be derived from recent immigration events, and migration to NC is mainly from SC, we conclude that SC serves as the primary source for NC H3N2 influenza consistent with (Du et al. 2012).

Our ability to correctly reconstruct the seasonal timing of migration events depends on the unbiased inference of phylogenetic trees, and the reconstruction of the state of nodes along these branches. Sparse trees reduce our inference power, as branches become uninformative with respect to the underlying migration processes. In addition, the inference of the seasonal timing of migration events is sensitive to the sequence sampling scheme.

Incidence Seasonality and Global Persistence

To increase the spatial scope of our analyses, we partition both the neuraminidase and hemagglutinin sequences into seven global communities (demes) and an additional representative sample of unclassified sequences (OT) from multiple geographic locations. Sequences were down-sampled for computational efficiency, maintaining broad seasonal signals (Figure 4.3A, Table 4.4, Figure 4.S6). A limited (disproportionally higher) number of samples from trough periods were kept to maintain a representation of sequence diversity during troughs. Comparison of alternative, non-seasonal and seasonal models of migration, supported seasonal ones (BF>33) (Table 4.2). The best supported model incorporates seasonal migration rates with two partitions of the year, as well as variable selection for the inclusion of migration, and seasonality of this migration, between every pair of demes.

Stochastic mapping is used to infer the state of trunk lineages (2001-2009.5) of both the neuraminidase and hemagglutinin proteins (Figure 4.S7) taken from the posterior distribution of trees and model parameters. A single stochastically mapped hemagglutinin and neuraminidase tree sample is included in (Figure 4.S8, S9). We define the trunk of the phylogeny as all the ancestral lineages of the most recent tip samples, discarding lineages which are too young (at most 2.5 years prior to the last tip time).

To show the relation between global persistence and seasonal incidence patterns in different locations, we first generated seasonal incidence profiles from weekly surveillance data (FluNet/WHO who.in/flunet) in each of the seven populations. The yearly incidence of H3N2 changes with age (Waalen et al. 2012), and is expected to have some variation across the populations (e.g. such was the case in A/H1N1pdm09 (Kerkhove et al. 2013)). However, the mean estimates of the reproduction number are not expected to vary widely (Chowell et al. 2010), and in the absence of available data, we use a simplifying assumption considering yearly attack rates to be similar in each of the seven focal communities. As such, we normalize the incidence seasonality profile based on each community's population size (Figure 4.3B). We use these profiles to calculate the harmonic mean (HM) of the estimated percent of the total global incidence in each community, in different months, across twelve months. The HM has been used extensively in population genetics when calculating the effects of fluctuations in population size on the effective population size (Caballero 1994) because it captures the increased risk of allelic extinction at low population sizes, *i.e.* the role of population bottlenecks.

We find correlation (N=14, adj. r^2=0.41) between the percent of the global trunk inferred to be in a community, and the HM of incidence seasonality. However, when using surveillance data, population size (human population of the countries in a community) was found to be a better predictor (adj. r^2=0.46, p=0.01) of the percent of the global trunk in the different demes and was selected for as the only predictor in multiple linear regression model ranking (Figure 4.4, Table 4.5) when considered together with the HM and with population density (human population divided by land area).

Although in agreement with the general trend (Figure 4.4), SA contributes less than expected to the global H3N2 trunk, while NA contributes more. Also, a substantial proportion of the neuraminidase (25%) and hemagglutinin (21%) trunks are inferred to be outside of the seven global communities we sampled (OT).

In other studies, global migration rates were found to be highly correlated with global air travel connectivity (Philippe Lemey et al. 2014). In the case of constant migration rates, in the absence of viral phenotypic evolution, long-term persistence theoretically corresponds to the stationary distribution of the migration rate matrix. However, with the exception of successfully predicting the reduced contribution of SA lineages (Figure 4.4) (Bedford et al. 2010b), this connectivity failed to explain the relative role of different populations to long term viral evolution and therefore, persistence. Seasonal migration rates offer no such single stationary distribution, but may result instead in a periodic cycle between alternative attractor configurations throughout the year (in the absence of phenotypic evolution). As such they have the potential to better describe a global metapopulatoin structure in which the trunk of the phylogenetic tree travels throughout the year.

Since the same predictor values are used to infer both the neuraminidase and the hemagglutinin trunk proportions, and we are limited to seven predictor values, our results are sensitive to variation in these predictors such as the underlying seasonal patterns or the estimates of population sizes.

Estimates, of the percent of the global trunk in each community, are dependent on the sampling and inclusion of near trunk lineages in the inference, the discovery of which is dependent on the sampling effort and on incidence itself, both of which vary from location to location, and with time. Our results largely correspond to the Bayesian credible intervals reported in (Philippe Lemey et al. 2014), in which sampling considerations were taken into account.

An important factor which was not considered in our analysis is the role of spatial patterns of incidence on viral persistence and long term evolution. Increased or reduced spatial correlation with respect to seasonality across large geographic regions, may account for additional unexplained variability between the regions with respect to the amount of long term viral evolution that they sustain. This is the case as fadeouts in incidence were observed in individual countries in, *e.g.* SEA (Russell et al. 2008), while our analysis of seasonal incidence patterns aggregates these countries together. In the future, better availability of data from extensive year-round sequence sampling may allow more detailed partitioning of the world population and could help mitigate some of these limitations, as well as improve our estimation of the contribution of different mechanisms in maintaining long-term viral evolution.

Incidence Seasonality and Global Migration

Multiple mechanisms contribute to the number of effective migrations, that is, to migration events that are not lost to rapid extinction and are able to achieve a sufficient population size to be picked up in a sequencing study that is deposited in the database. Contact between the source and destination community may depend on the pathogen abundance at the source, and is also modulated by factors such as air travel. Furthermore, once at its destination, an invading pathogen experiences different seasonal transmission rates and host susceptibility levels. This is further complicated by the dependence of the invasion on antigenic or other traits of the invading pathogen (Philippe Lemey et al. 2014; Catford, Jansson, and Nilsson 2009).

Here we wish to establish how two basic features of seasonal epidemics, namely incidence and growth, correlate with the number of emigration and immigration events observed along the H3N2 phylogeny. An understanding of these associations provides a basis for identifying the processes contributing to the seasonality of global H3N2 migration.

As before, we use surveillance data from FluNet/WHO to obtain a broad representation of the seasonal patterns in each community (Figure 4.3B). Whereas in the above analyses the more likely partitions of the year were selected by the MCMC, here, each partition of the year is centered along a different month in order to prevent the confounding of the identification of a likely partition with our measurements of incidence and growth during the same six month period. For each partition, we use our MCMC framework to infer migration rates between each pair of communities. This is done jointly for tree pairs from the sampled distribution of hemagglutinin and neuraminidase trees. We use stochastic mapping to map emigration and immigration events along branches of the phylogeny (Figure 4.S10).

For each of seven global communities and six alternative partitions of the year into two equal parts, we measure the fraction of immigration events to, or emigration events from, the community (N=42) during the corresponding six months. This is repeated across multiple samples from the posterior distribution of rate parameters. We then apply multiple linear regression models to identify possible correlations between incidence and growth to median immigration and emigration. The fraction of positive growth in a six month window is calculated as the fraction of months in that period that show an increase in incidence compared to the previous month based on yearly incidence profiles from surveillance data.

Immigration is significantly correlated (adj. r^2=0.69) with both absolute incidence (adj. r^2=0.22, p=$2 \cdot 10^{-6}$) and positive growth (adj. r^2=0.52, p=$6 \cdot 10^{-9}$) (Figure 4.5C, D). There is also correlation (adj. r^2=0.13, p=0.01) between epidemic decline and increased emigration (Figure 4.5B). We find however no support for correlation between the fraction of the yearly incidence within a single community during a six month long period, and the fraction of emigration during that same period (Figure 4.5A). The global migration patterns largely correspond to the two case studies, where the absolute incidence as well as growth in incidence, are indicative of higher immigration, while epidemic decline is associated with increased emigration (Figure 4.5).

These results correspond to a case where sufficient propagule pressure is achieved at varying levels of incidence at the source. Thus, the amount of effective emigration does not correlate directly with incidence. This may be the case when sufficient contact between the global communities is reached before and after peak incidence levels. In contrast, the results suggest a much greater role for conditions suitable for growth (fertile grounds) at the destination, in determining the probability that an immigration event will lead to successful establishment and be counted as an effective immigration event.

4.4 Limitations

Bayesian inference and the estimation of marginal likelihoods are dependent on the choice of priors. These serve as baseline assumptions for the model parameters, and sufficient data to the contrary will pull the estimated parameters away from these assumptions. We used constant-population coalescent process priors because they are well established in the context of Influenza A phylogentic inference (Bedford et al. 2010b; Bahl et al. 2011; P. Lemey et al. 2009; Lu, Lycett, and Brown 2014; G. J. Smith et al. 2009). These priors are simpler to implement, and had little influence on phylogeographic inference in previous reports (P. Lemey et al. 2009). Therefore alternative tree priors, such as ones involving birth death processes (Tanja Stadler 2009), were not used.

In addition, the inference of trees based on combined sequence data and geography, rather than sequential consideration of these two aspects as implemented here, should be more accurate by allowing the more thorough exploration of tree space. Although in general this scheme will sample tree space more exhaustively, it will incur a computational effort which will at the least amount to multiplying the computational effort by the number of models evaluated. As most of the information in our analysis (magnitude of 10,000s vs. 1,000s log likelihood units) relating to tree topology is contained in nucleotide data compared to geography, we expect the tree samples to be sufficiently representative, and the combined inference of trees and migration processes to be equivalent to a sequential one. Some of these considerations are listed in (Pagel, Meade, and Barker 2004).

The availability of sequence samples is limited, and we assume that a sufficiently representative sample of sequences is available from each community and across time. Sufficient sampling is required in order to generate dense trees which will contain information about the correct seasonal timing of migration events. Differences in sampling and sequence availability may bias the amount of migration inferred to take place between two communities. When analyzing the results, we refer only to the seasonality of migration and not to the total amount of migration throughout the year. Although we expect this quantity to be less sensitive to sampling, it may be that an increase in the sampling of one community vs. the others will bias the timing of immigration to weigh more towards the timing relevant to this community. We acknowledge that this is a limitation of our analysis which cannot be completely avoided at the moment as the number of available samples globally is highly skewed towards specific countries, and homogenous sampling reduces the power of our analysis considerably. In the future with the increase in worldwide sampling cover, alternative sampling scenarios could be considered. In agent based simulations (Zinder et al. 2013) we evaluated similar sampling scenarios as the ones used here, and found the results of the inference largely consistent with the underlying migration processes (Appendix A 3.3).

4.5 Conclusions

Our results show clear support for seasonal variation in migration rates. We used models incorporating this variation to estimate patterns of global seasonal migration and of persistence.

H3N2 persistence is short, on the scale of several months, with only a small fraction of the lineages persisting for over a year (Figure 4.1B, Fig 2C, Figure 4.S11). Seasonal patterns also correlate with the probability that local lineages go extinct and fail to contribute to long term viral evolution. However, the probability that a region will contribute to long term viral evolution as a part of the trunk of the phylogenic tree was found to be better determined by its overall human population size. In general, this short local persistence indicates a massive replacement of circulating lineages on both annual and sub-annual timescales, much shorter than those characteristic of global turnover, typically around several years.

This rapid replacement is mediated by migration which is by itself highly seasonal in nature (Table 4.2). In particular, the likelihood of successful immigration increases during periods of the year that coincide with epidemic growth and higher incidence, a pattern suggestive of a 'fertile ground' hypothesis, where incoming viruses survive and spread more effectively during this upward season (Figure 4.5C,D). Emigration, on the other hand, coincides with periods of epidemic decline (Figure 4.5B), which suggests a 'tail-to-beginning' migration pattern between overlapping epidemics. These overlapping seasons would correspond to the major epidemics of the two hemispheres which are known to exhibit opposing seasonality. This pattern could also indicate the reintroduction of lineages to the tropics at the end of temperate climate epidemics as was suggested by Bedford et al. in (Bedford et al. 2010b).

Surprisingly, the dominant factors behind influenza's incidence seasonality remain a subject of debate, with emphasis on either environmental factors influencing transmission or host susceptibility (Shaman et al. 2010; Tamerius et al. 2013; Harper 1961; Cecile Viboud et al. 2004; Cannell et al. 2006). With respect to migration, focus has been largely given to factors such as connectivity through air travel, and not to other aspects of effective migration related to conditions at the source and destination. The phylogenetic tree of the virus, if correctly reconstructed, can provide multiple natural experiments involving the state of the environment and the phylogenetic outcome. Incorporating seasonality into the reconstruction of environmental conditions appears essential.

4.6 Methods

SeasMig We implemented in Java a tool (http://bitbucket.org/pascualgroup/seasmig). A detailed description of SeasMig in the context of this manuscript is available in Appendix A. Using SeasMig alternative migration models parameters could be inferred and compared by their marginal likelihood including seasonal, epochal, and non-seasonal phylogeographic migration models. An empirical distribution of trees in nexus format (e.g. generated using BEAST (A. Drummond and Rambaut 2007)) is given as input. Our tool uses an MCMC to sample from the posterior distribution of model parameters and stochastically mapped migration events along branches and trunk lineages (Figure 4.6). Multiple MCMC chains run in parallel and perform chain swaps in a lockstep manner.

Stochastic mapping Stochastic mapping is an additional step following the calculation of tree likelihood and ancestral state reconstruction at the nodes of a tree. This mapping allows us to generate a stochastic realization of the state of branches along the tree, in addition to the state of internal nodes, and in so doing, provides samples of migration and mutation events, and their timing along the tree that lead to the observed tip states. Stochastic mapping of both sequence (nucleotide) and character (*e.g.* geographic) annotations is available in *SeasMig*, together with the option of incorporating seasonal migration models. Stochastic mapping is implemented directly in our code based on (J. P. Bollback 2006). Improved performance could be achieved using (Minin and Suchard 2008). Briefly, a given type of event, migration or mutation, is assumed to behave as a Poisson process along a branch. As such, the timing of the next event follows an exponential distribution with a mean equal to one over the total rate of emigration (mutation from) a location (a base). Once the timing of the next event is determined, the event is chosen based on its relative probability compared to other emigration events. Branch reconstructions that span different seasons were performed by sampling the state (here the location) within the first season's boundaries using the initial migration matrix, and by continuing the stochastic mapping forward using the second matrix.

Tree sampling For each of the inferences, we use a sample of trees taken from the empirical posterior distribution of trees generated by BEAST based on sequence data and sampling times alone. Sampled sequences are aligned using MUSCLE (Edgar 2004), sequences with low alignment quality were manually removed. We perform phylogenetic tree reconstruction of coding region sequences with high coverage using

BEAST 1.7.4. A differential codon location evolutionary model is used (HKY1+2) (Shapiro, Rambaut, and Drummond 2006). Models with a biological meaning, which account for variation in the evolutionary rate in different codon positions, such as the HKY1+2, were found to be better supported than standard nucleotide substitution models such as general time reversible with gamma distributed rate heterogeneity and a proportion of invariant sites (GTR+Γ+I) in most RNA viruses in (Shapiro, Rambaut, and Drummond 2006) while requiring less parameters. In each analysis 2000 trees from the stationary distribution of four independent chains sampled every 10000 steps, are combined. BEAST XML, tree and log files are available through Dryad (doi:10.5061/dryad.t120k). BEAST tree inference was carried out on computational resources and services provided by Advanced Research Computing at the University of Michigan.

Sequence sampling We sampled sequences from the NCBI flu database. All, but at most n samples chosen randomly per k consecutive months per community are used. This sampling scheme is intended to reduce the number of overall samples for computational reasons, by decreasing the number of samples taken in more recent years as more sequences were generated, while maintaining available data during the seasonal troughs. For the sampling of "other" global sequences (OT), equal proportion sampling in every month for every geographic location was used with the objective of capturing a sufficient representation of the underlying global genetic diversity. The number of sequences used in each analysis is included in Table 4.1, 4.3 and 4.4 and in Figures 4.S1, 4.S4 and 4.S6. Sequences were sampled between 1999 and 2013; their accession numbers, dates, and geographic classification, for the different figures are included in Additional File 3. In simulation studies, the seasonal pattern of effective migration events was better reconstructed when tip (sequence) sampling was proportional to incidence (Appendix A 3.3.2.2), as opposed to when sampling was uniform across time and in different populations (Appendix A 3.3.2.1).

Seasonal migration model We generate a seasonal migration model (Figure 4.6) by using two different constant migration rate matrices (Q_A and Q_B) for two parts of the year labeled as "season A" and "season B" respectively. To estimate the transition probabilities between two geographic locations at different times, we calculate the respective transition probability matrices P for the individual constant rate periods through matrix exponentiation (Bielejec et al. 2014). For the complete time interval, the individual transition probability matrices are multiplied accordingly. Given a tree topology, we integrate unknown internal node states over the tree efficiently by caching conditional probabilities of individual node states as described in (Joseph Felsenstein 1985). Our method has similarities with [157] but is parameterized differently as detailed in the next paragraph. This parameterization allows a smooth transition from a seasonal to a non-seasonal model which is well suited for variable selection. Our approach is also different because it considers alternative partitions of the year which are either sampled using the MCMC in proportion to their likelihood (in Figure 4.1, 4.2,4. 4) or are integrated upon with equal probability (in Figure 4.5). The latter approach allows us to consider correlates of migration seasonality realted to incidence independent of the choice of the likely partitions of the year.

Seasonal migration model parameterization and priors We parameterize migration rates for the two partitions of the year as $r_{from,\ to} \cdot (1+\sigma_{from,\ to})$ and $r_{from,\ to} \cdot (1-\sigma_{from,\ to})$. This is done with the purpose of measuring the inferred rates, and their seasonality separately and allowing for separate indicators, used in variable selection, for the inclusion or exclusion of seasonality, and for the inclusion and exclusion of any migration from one community to another. Migration rates $r_{from,\ to}$ are drawn from an exponential prior, giving a non-diminishing probability to high migration rates. The *'seasonal scaling'* $\sigma_{from,\ to}$ parameter gives the relative increase (and decrease) in contribution of migration in one season compared to the average migration rate. This parameter is sampled from a uniform U(-1,1) prior.

Metropolis-Coupled MCMC (MC3) We use an MC3 algorithm to sample model parameters and to from the sample of trees. Metropolis-coupled Markov-chain Monte Carlo, or MC3, is an MCMC algorithm that allows sampling from analytically intractable distributions, and builds on standard MCMC by improving mixing (Baskerville et al. 2013). Such distributions include the distribution of tree likelihoods given a mutational or a migration model. In particular, MC3 includes multiple MCMC chains: a cold chain samples from the target distribution, while hot chains sample from a flattened likelihood surface exploring more of the parameter space. MC3 algorithms explore and swap proposals with heated chains that continue to sample parameters from the prior distribution and from flattened likelihood surfaces. MC3 offers a relatively robust method for integrating marginal likelihoods (Friel and Pettitt 2008).

Variable selection To assess whether the inclusion of migration between different communities is informative, and to establish if rates are seasonal, we implemented Bayesian variable selection (Kuo and Mallick 1998) in an MC3 framework. Indicator variables $I_{Rfrom,\ to}$ and $I_{Sfrom,\ to}$ are added, giving the full parameterization of a single seasonal migration matrix cell as $I_{Rfrom,\ to} \cdot r_{from,\ to} \cdot (1+I_{S\ from,\ to} \cdot \sigma_{from,\ to})$ and $I_{Rfrom,\ to} \cdot r_{from,\ to} \cdot (1-I_{S\ from,\ to} \cdot \sigma_{from,\ to})$ for the other part of the year. In this case symmetric non-informative priors are used for the indicators. Reported Bayesian support for migration between two communities (BF) is the ratio of cases for samples for which an indicator variable $I_{Rfrom,\ to}$ is 1 vs. 0. Similarly this ratio is used to show support for seasonal migration between two communities.

Combined likelihood based on the hemagglutinin and neuraminidase proteins We use a conservative approach to combine the information present in both protein trees with respect to model likelihood. The combined protein tree log-likelihood is weighed down by half, to account for the possible lack of independence in the information contained in the two trees with respect to migration rates and seasonality. This choice does not affect the maximum likelihood model but has the effect of widening confidence intervals when the two trees provide independent data, while providing the correct confidence interval when the two proteins are in complete linkage.

4.7 Availability of supporting data

The datasets supporting the results of this article are available online. The GenBank accession numbers of sequences used in this study and their geographic classification are available in Additional file 3. BEAST XML, tree and log files are available through the Dryad data repository under doi:10.5061/dryad.t120k.

4.8 Authors contributions

Author list: Daniel Zinder, Trevor Bedford, Edward Baskerville, Robert J Woods, Manojit Roy, Mercedes Pascual. Author contributions: DZ: conceived the project. DZ, TB, EB and RW added methodological contributions DZ, TB, RW, MR and MP analyzed the results and contributed to drafting and writing of the paper

4. 9 Acknowledgments

We thank Stephen Smith for comments and discussion. MP is an investigator of the Howard Hugues Medical Institue.

4.10 Tables

Table 4.1. Number of Sampled Sequences in Partition into NA and OT

Partition	Hemagglutinin (H3)	Neuraminidase (N2)	Total
NA	909	438	1346
OT	975	556	1531
Total	1884	994	2878

NA – North America (USA, Canada), OT – Representative of global community

Table 4.2. Marginal Likelihood of Alternative Migration Models

Migration Seasonality	+dof	NA/OT	NC/SC/OT	Global
none	0	-747.7	-587.0	-2246.3
none, variable selection for the presence of any migration between two communities			-588.4	-2241.2
origin based	n+1		-568.6	-2193.5
destination based	n+1		-561.2	-2207.7
origin and destination based	2n+1		-561.1	-2207.6
specific origin and destination based	n^2-n+1		-561.0	-2200.6
specific origin and destination based migration seasonality, variable selection for any migration and for seasonality between each pair of communities		-726.6	-561.8	-2192.6

dof - degrees of freedom, NA/OT – North-America and the Global Community, NC/SC/OT – North-China, South China and the Global Community Global – 7+1 global communities

Table 4.3. Number of Sampled Sequences in Partition into NC, SC and OT

Partition	**Hemagglutinin (H3)**	**Neuraminidase (N2)**	**Total**
SC	528	56	**584**
NC	150	1	**151**
OT	1302	270	**1572**
Total	**1980**	**327**	**2307**

SC – South China, **NC** – North China, **OT –** Representative of global community

Table 4.4. The number of sequences used in seven global communities

Partition	Hemagglutinin (H3)	Neuraminidase (N2)	Total
CH	320	151	**471**
EU	134	69	**203**
JA	180	83	**263**
AUS	172	314	**486**
SA	86	5	**91**
SEA	309	429	**738**
NA	288	413	**701**
OT	236	188	**424**
Total	**1725**	**1652**	**3377**

CH – China, EU – Europe, JA – Japan, AUS – Oceania, SA – South-America, SEA – South East Asia, NA – North America, OT – Representative of unclassified sequences from multiple geographic locations

Table 4.5. Ranking of alternative linear regression predictors for the percent of the global H3N2 trunk in different communities

Model	Adjusted R^2	R^2	AIC	BIC
{Pop}	0.464411	0.506	226.15	228.06
{HM, Pop}	0.421031	0.510	228.02	230.57
{Pop, Density}	0.419179	0.508	228.06	230.62
{HM}	0.414195	0.459	227.40	229.32
{HM, Pop , Density}	0.367029	0.513	229.93	233.13
{HM, Density}	0.365295	0.463	229.30	231.86
{}	0.	0.	234.00	235.29
{Density}	-0.0538263	0.027	235.62	237.53

HM – Harmonic mean of the estimated percent of the total global incidence in each community, in different months, across twelve months **Pop** – Population size of a community, as total of the countries included in a community, **Density** – Population density of a community, as the total area divided by the total population size of countries within the community.

4.11 Figures

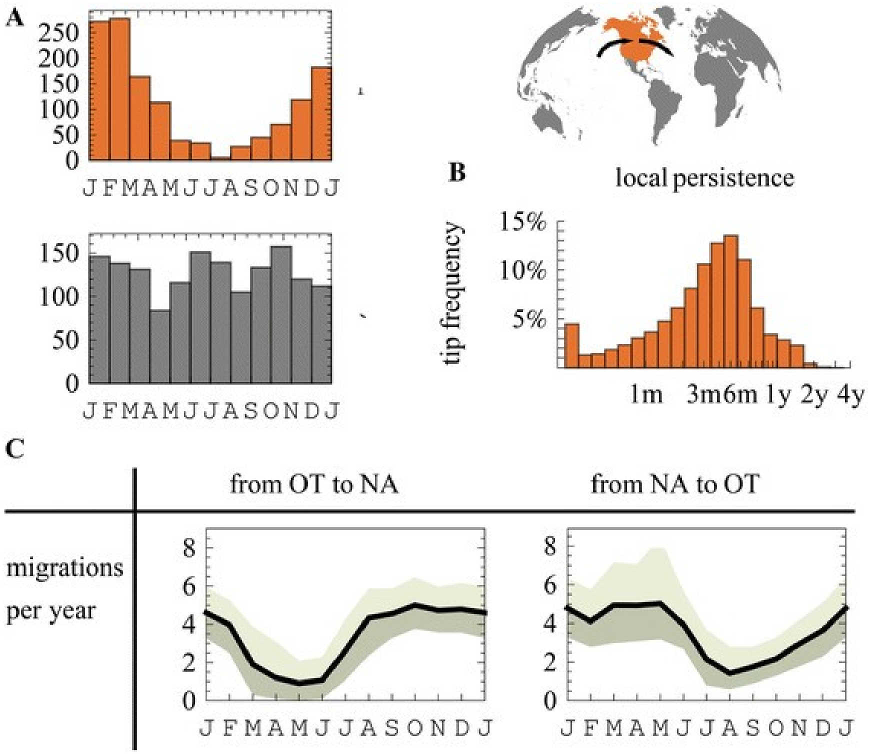


Figure 4.1. Seasonal Migration to and from North-America.
(A) The monthly sampling distribution of included publically available human H3N2 hemagglutinin and neuraminidase sequences in North-America (NA) and a representative sample of other (OT) global communities. **(B)** The inferred distribution of time for which sampled NA sequences have remained in NA following a migration event. **(C)** The inferred mean (90% CI) number of migration events to and from NA in different months.

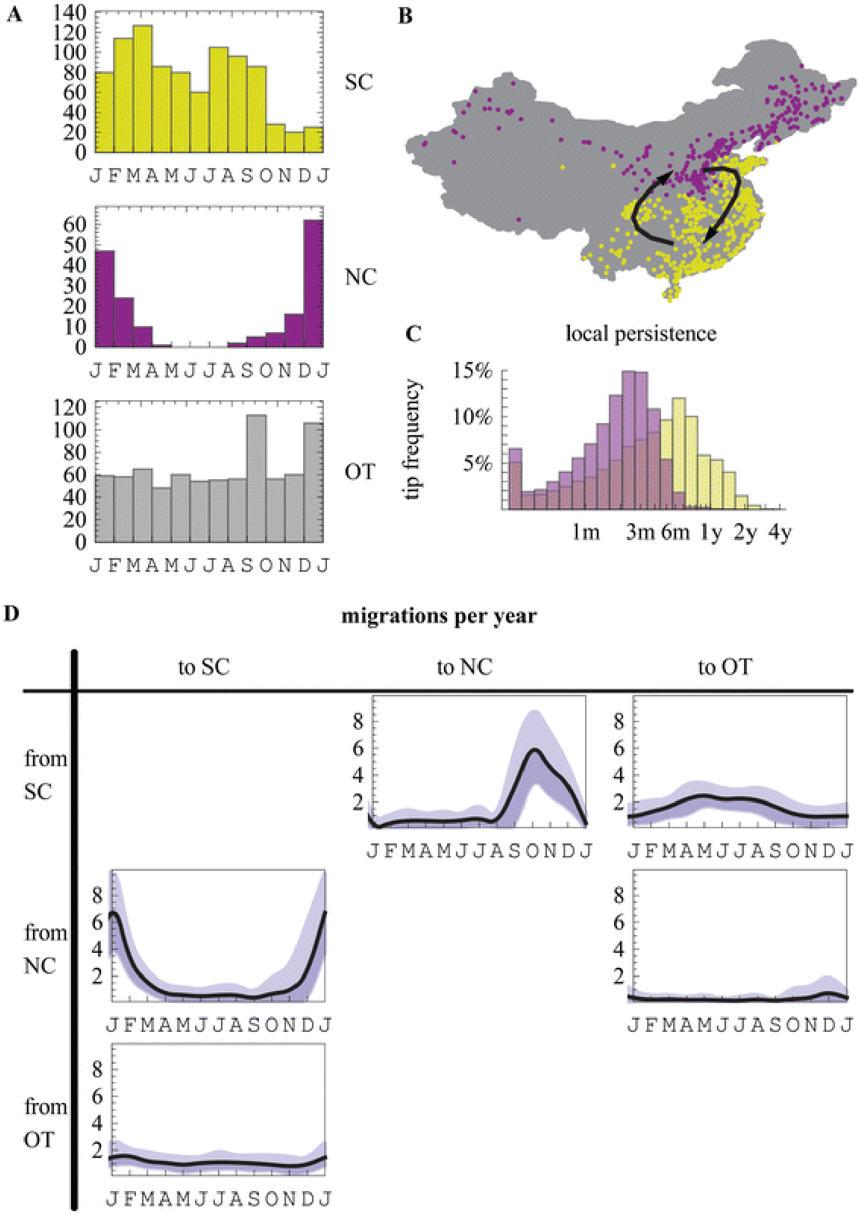


Figure 4.2. Seasonal Migration between North-China and South-China.

(A) Provinces in China were partitioned based on broad incidence seasonality patterns, inferred using the bi-monthly distribution of sequence samples (Figure 4.S2). The partitioning of the provinces divided China into a northern (NC) and a southern (SC) seasonality clusters. SC maintains two seasonal peaks while NC displays a single season during winter months. **(B)** The location of cities in the provinces included in the analysis colored by their seasonal partitioning. **(C)** In yellow (purple), the inferred distribution of time for which sampled SC (NC) sequences have remained in SC (NC) following a migration event. **(D)** The inferred mean (90% CI) number of migration events between NC, SC and the other (OT) global communities.

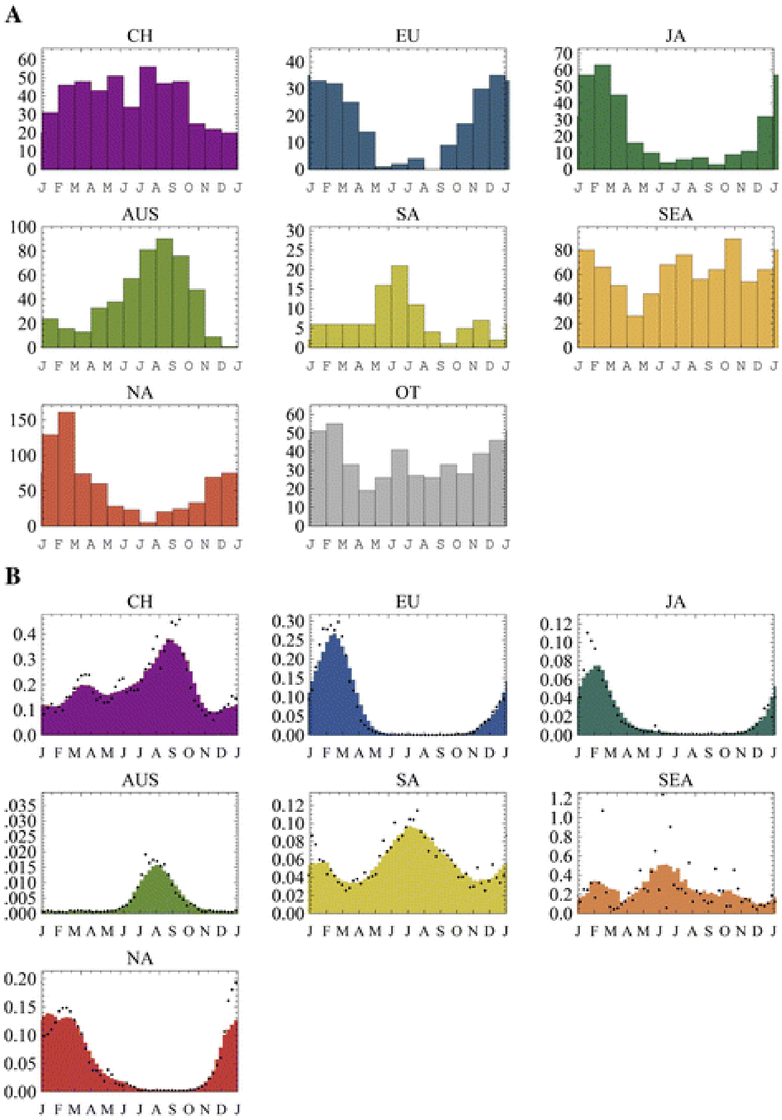


Figure 4.3 Sequence Sampling and Incidence Profiles for Seven Global Communities
(A) The monthly sampling distribution of global (CH – China, EU – Europe, JA – Japan, AUS – Oceania, SA – South-America, SEA – South East Asia, NA – North-America) publically available human H3N2 hemagglutinin and neuraminidase sequence samples used for the inference of phylogenetic trees and migration seasonality (Table 4, Figure 4.S4). A representative collection of sequences sampled from other parts of the world is designated as other (OT). **(B)** Surveillance data (WHO FluNet, 2000 week 1 to 2012 week 52) was aggregated on a weekly basis and smoothed (8 weeks moving average) to obtain broad seasonal incidence profiles in seven global communities. Within each global community, country level surveillance counts were normalized (divided by the total number of counts in the country), and added up in proportion to the country's population size. Each community was weighted based on its population size to approximate its relative contribution to worldwide incidence.

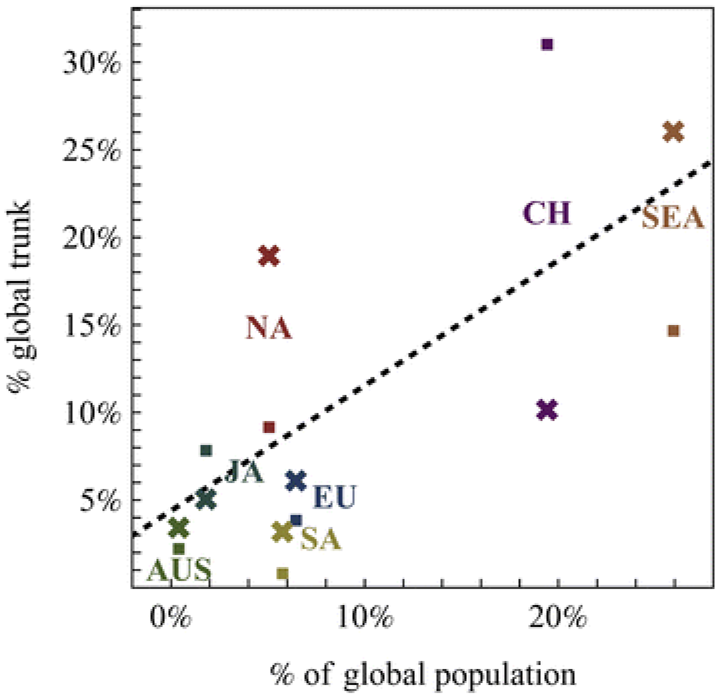


Figure 4.4. Global Seasonality Patterns and Trunk Proportions

The percent of the global hemagglutinin (·) and neuraminidase (×) trunks inferred to be in the different global demes vs. the population size of the deme (CH – China, EU – Europe, JA – Japan, AUS – Oceania, SA – South-America, SEA – South East Asia, NA – North-America). For each protein the median proportion of the phylogenetic tree trunk in each deme (Figure 4.S2) was taken from the distribution of stochastically mapped trunk states given the sampled empirical posterior distribution of seasonal migration model parameters.

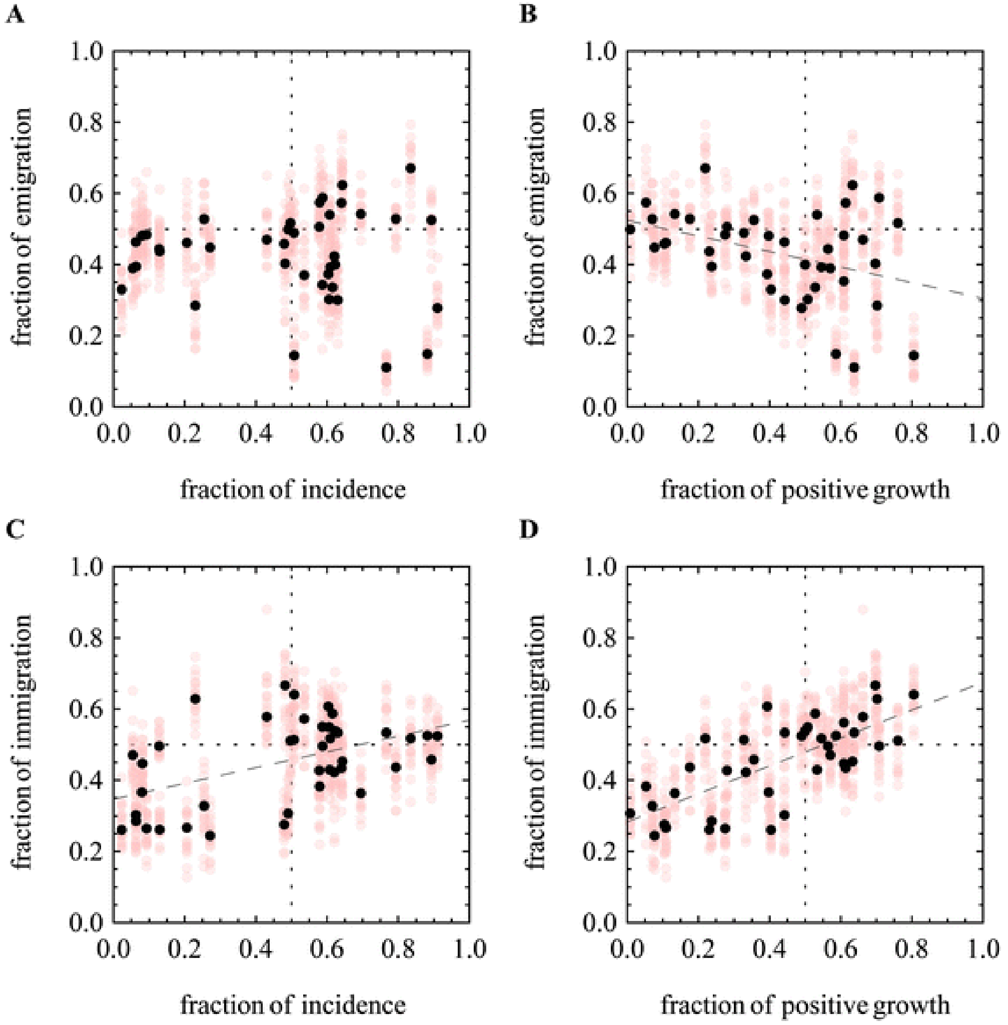


Figure 4.5. Correlation between Growth, Incidence, Immigration and Emigration

Seasonal incidence profiles (Figure 4.3B) in each community, are used to identify gross periods of growth, decline, and to estimate the monthly % of the yearly incidence during six month periods, averaged across multiple years. Alternative seasonal migration models, partitioning the year into two, and centered on consecutive months (Jan-May) were parameterized using the MCMC. The number of stochastically mapped migration events between each pair of locations was counted in each of the six month long partitions. **(A)** Medians (black) and samples (gray) of the fraction of the total emigration events from a location during a six month long period, vs. the fraction of the yearly incidence in the corresponding six month period in the source location. **(B)** The fraction of the total emigration events from a location during a six month long period, vs. the fraction of the yearly positive growth in the corresponding six month period in the source location. **(C)** The fraction of the total immigration events to a location during a six month long period, vs. the fraction of the yearly incidence in the corresponding six month period in the destination location. **(D)** The fraction of the total immigration events to a location during a six month long period, vs. the fraction of the yearly positive growth in the corresponding six month period in the destination location.

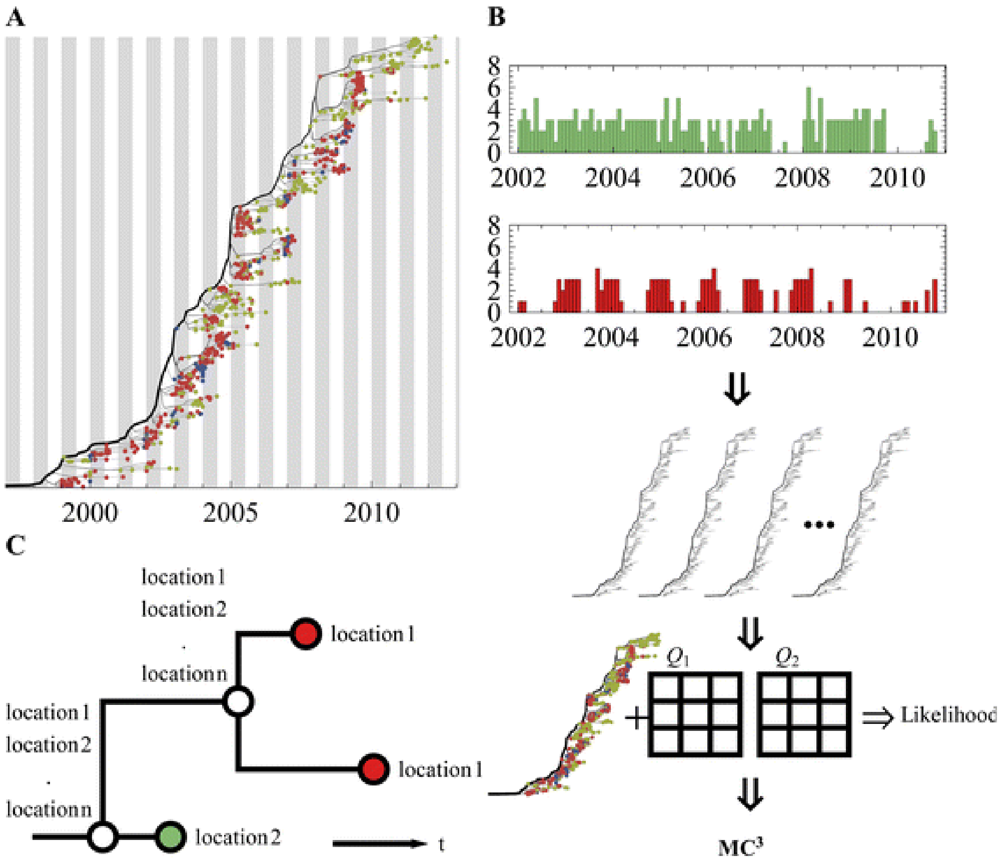


Figure 4.6. Methods

(A) Phylogenetic tree of H3 reconstructed based on nucleotide data, for this tree sequences were partitioned into two communities for illustration purpose (red, green). Our method allows for different migration rate matrices (Q1, Q2) to be calculated for different month of the year *e.g.* Jan-Jun (gray background), and Jul-Dec (white background). **(B)** Inference pipeline: 1. Date annotated sequences are sampled from a global pool, and are partitioned based on their collection location. 2. Multiple possible phylogenetic trees are reconstructed based on nucleotide data. This can be done for several protein segments. 3. Likelihood is calculated for: a given migration model, known tip location, and a single tree topology. 4. An MCMC is used to sample from migration model parameters and trees based on their likelihood. 5. Counts and the timing of stochastically mapped migration events along the tree branches and its trunk are sampled. **(C)** In order to calculate a single tree and model parameters likelihood, transition probabilities are generated based on the matrix exponent of the rate matrices. Transition probabilities are multiplied across the phylogenetic tree, and integrated over internal unknown nodes. For transitions spanning across seasonal borders the transition matrices are accordingly multiplied. An additional step, called stochastic mapping, is used to sample from the timing of migration events along the branches of tree, in addition to the nodes.

Chapter 5. Global metapopulation dynamics of rotavirus allow response to vaccination

5.1 Abstract

Genetically diverse and rapidly evolving pathogens present a challenge to vaccination efforts. A global vaccination campaign is underway for rotavirus, the leading cause of diarrheal deaths in children. The implications of the genetic diversity in the rotavirus population for this campaign are currently unknown. Here we use publicly available sequences of the antigenic surface protein VP7 to fit a Bayesian phylogenetic model of global migration and sequence evolution. We show that the rotavirus population in the USA has already evolved to become genetically less like the vaccine following vaccine introduction in 2006. Ancestral state reconstruction was used to identify migration of strains into the USA rather than mutation as the major cause for the genetic shift away from the vaccine. This analysis yields a model of global rotavirus population dynamics in which the response to intense local selection to escape host immunity is achieved through migration from a diverse global metapopulation. The migration-driven dynamics we describe for rotavirus present an alternative to the established model of seasonal influenza, in which mutation plays a central role. This result indicates that a polyvalent vaccine strategy or an adaptive strategy, where the vaccine is updated in response to pathogen evolution, will likely be superior to a monovalent one. Further survey of existing genetic variation is critically needed to

improve predictions of future responses to vaccination, and to guide development of the next generations of vaccines to be more robust to evolutionary change.

5.2 Introduction

Rotavirus is a globally prevalent virus that infects virtually every child before the age of five, resulting in over 400,000 deaths annually (Tate et al. 2012). It is a double stranded RNA virus with eleven genome segments encoding 12 proteins (Estes and Kapikian 2013). Standard typing is based on the two outer surface proteins, VP7 which encodes the G protein, and VP4 which encodes the P protein (Jelle Matthijnssens et al. 2008). Although six genotypes result in over 90% of the cases, serotypes exhibit extensive variability in both time and space. Infection with rotavirus leads to partial immunity and children frequently may have two or more infections, but seldom have more than three (Crawford et al. 2011). While immunity is complex, involving innate immunity, cell-mediated immunity and humoral immunity, protection is thought to be greater against infections by the same serotype than those by different serotypes (Ward, Clark, and Offit 2010; Offit 1996). Specifically, neutralizing antibodies to the outer surface protein VP7 have been shown to correlate with protection. These neutralizing antibodies to VP7 have been mapped to three regions of this protein, or epitopes, in a serotype specific manner (Aoki et al. 2009a) .

Although rotavirus shares certain biological and epidemiological features with the more thoroughly studied influenza A virus, it also exhibits important differences. Both are acute viral infections that circulate extensively in the human population, leading to hundreds of thousands of deaths a year (Tate et al. 2012; Nair et al. 2011; Dawood et al. 2012). Infection with either virus leads to partial immunity allowing reinfection. Extensive evolutionary and epidemiological investigation has shown that positive selection in the defined epitopes leads to 'antigenitc drift' in influenza. This process results in the rapid replacement of the global human influenza A population every 2-8 years and phylogenetic trees with a 'spindly' appearance, consistent with positive selection (Andrew Rambaut et al. 2008b; Bedford et al. 2010a; Bedford, Cobey, and Pascual 2011b). Rotavirus serotypes may become dominant in a region for 1 to 3 years, then become infrequent (M. H. Afrad et al. 2013; De Grazia et al. 2014; Hasing and Trueba 2009) and re-emerge years or decades later (S. M. McDonald et al. 2012). This rapid local turnover in serotypes, and emergence and spread of new serotypes, is suggestive of strong selective pressure imposed by host immunity. However standard analysis fail to reveal evidence for positive selection on the RNA sequence and the phylogenetic trees of rotavirus are distinctly different from those of influenza, with a large amount of global genetic diversity maintained through time. This contrasting evolutionary pattern raises the question of whether global population dynamics differ in an important way from those of influenza.

Current rotavirus vaccines, RV5 and RV1, became available in the United States (USA) in 2006 and 2008 respectively, and the World Health Organization (WHO) recommended vaccination for all countries in 2009 (Zinser 2009). RV5 is a live pentavalent human-bovine reassortant vaccine, which contains a bovine rotavirus backbone into which a human rotavirus VP7 gene (G1, G2, G3, or G4) or VP4 gene (P[8]) has been reassorted. RV1 is a human origin monovalent attenuated G1P[8] virus. Although both vaccines are highly effective against severe disease or hospitalization in developed countries, including the USA (Rha et al. 2014), they are less protective against mild infection, and less effective in developing countries in Africa and Asia (E. A. S. Nelson and Glass 2010; Benjamin A. Lopman et al. 2012).

Both vaccines have been used broadly in the USA, and by August 2014, 69 countries had introduced rotavirus vaccination through their national programs.

The large and diverse population of the virus, combined with an imperfect vaccine, has raised concern about vaccine escape (Ben A Lopman et al. 2012; Zeller et al. 2012). Studies that have sought to measure the effect of widespread vaccination on strain makeup have been difficult to interpret in light of the natural fluctuations that occur even in the absence of vaccination. This study takes advantage of the introduction of vaccination in the USA to measure the population-wide response of the virus to this broad perturbation of the system, in order to address the relative importance of the major processes behind rotavirus global population dynamics.

5.3 Results

From all rotavirus sequences available in GenBank up to October 2013, a total of 1911 VP7 sequences of human origin were identified with an annotated country of origin, year of collection, and host species. These VP7 sequences represent 48 countries and have collection dates between 1971 and 2012, with the greatest number of sequences from the six years prior to the introduction of the vaccine in 2006, and the six years after (Figure 5.2). The most common serotype was G1 but all common serotypes were represented.

The introduction of vaccines, RV5 and RV1, into the USA created a well-defined, population wide, antigenic challenge. Twenty-nine amino acid positions in VP7 whose mutation allows escape from neutralizing antibodies have been identified and mapped to three epitopes, 7-1a, 7-1b and 7-2 (Aoki et al. 2009b; Green et al. 1988). The change in the makeup of the USA rotavirus population with respect to this vaccine challenge, was measured as the average number of mismatches between sequences and their closest match among the vaccine strains at these 29 amino acid positions. Thus a perfect match to the vaccine corresponds to a value of 0, and a mismatch at every position, to a value of 29. The sequences collected in the USA after 2007 differed by an average of 5.4 amino acids from their closest match in the vaccine. This differs significantly from the average of all USA samples after 1971, which were 2.8 amino acids (p<0.00001 based on bootstrap resampling with replacement).

To gain a clearer picture of the change in VP7 sequences over time and geographic space, a total of 1270 phylogenetic trees were constructed that represent a sample from the range of possible evolutionary histories supported by the data, along with corresponding individual stochastic realizations of the possible mutation and migration events (Methods) (Figure 5.1). In the 2 years prior to vaccine introduction, 2004 to 2006, the average distance to the vaccine of strains predicted to be in the USA, was 4.2 amino acids. After vaccine introduction into the USA in 2006, the average distance to the vaccine increased to a peak of 11.0 in 2009.

The use of genetic distance allows us to partition the total genetic shift in the population among different evolutionary mechanisms. The average genetic distance of strains predicted to be in the USA in the time period prior to vaccine introduction (2004-2006) was compared to the time period following vaccine introduction (2007-2012). This difference is 2.6 amino acids (95% Bayesian credibility interval (CI) 1.62 - 3.62). Three different evolutionary mechanisms can account for this shift in the average distance to the vaccine in USA strains: (1) mutations that occurred in the USA after the introduction of the vaccine in 2006 (2) migration of strains into the USA after 2006, or (3) a shift in the relative frequency of strains already in the USA by 2006. The effect of migration was measured by comparing the observed changes to the change that is measured when all migration events to the USA after 2006 are removed from the trees (Methods). Similarly, the effect of mutation was identified by ignoring all mutations that occurred in the USA after 2006. The remaining change in the mean was due to changes in the frequencies of sequences predicted to have entered the USA prior to 2006. The contribution of each of these three processes was calculated for each of the sampled states from the Bayesian posterior.

Mutations occurring in the USA after 2006 contributed a negligible amount to the total change in the distance from the vaccine from before (2004-2006) to after (2007-2012) vaccination. The increase in the average distance to the vaccine was the result of both migration of new sequences into the USA and the shifting frequencies of those already in the USA (Figure 5.2a). The model could not identify which of these two mechanisms was most important. This uncertainty reflects the variation from one stochastic run to another in whether a sequence had entered the USA precisely before or after 2006. Figure 4.2b shows the distribution of the time between the predicted entrance to the USA and its sampling. On average, 50% of the strains entered the USA within 2.5 years of their collection and 90% within 5 years. There is no evidence that the sequences from 2007 to 2012 were on average more recent migrants than those collected from 2004 to 2006 (Figure 5.2b). Thus, all USA strains were relatively recent migrants.

Further partitioning the change in the distance from the vaccine in the USA following vaccination according to serotype, reveals that the majority of the change was due to an increase in G9 in the USA, with G9 accounting for 2.0 of the total 2.6 amino acid shift (95% CI 1.2 - 2.9). Of the sixteen G9 sequences in our dataset from the USA after 2006, there were at least four separate introductions from other regions (Figure 5.4b). Importantly, there was also a significant change in the makeup of G1, with the average G1 sequence predicted to be in the US becoming less like the vaccine by 0.7

amino acids (95% CI 0.3-1.1). Thus, the response to the vaccine includes both serotype replacement and changes within serotypes.

5.5 Discussion

This approach provides new insight into the global population dynamics of rotavirus. In accord with numerous studies, we see that there can be a rapid shift in local serotype prevalence(Hasing and Trueba 2009; M. H. Afrad et al. 2013; De Grazia et al. 2014). However, this analysis also shows that the shift is directional, insofar as it increases the genetic distance from the vaccine. Additionally, the change is seen in multiple independent G9 migrants to the USA. Finally, the two best-represented serotypes in the USA in the dataset (G9 and G1) both show this movement away from the vaccine. Taken together, these lines of evidence suggest that the shift in the USA is driven by the vaccine.

Despite the indication of strong selection imposed by local host immunity, there is no signal of positive selection on the VP7 gene (Figure 5.7). The analysis offers a solution to this apparent contradiction. The genetic variation that allows a local response to the selective force of local immunity arrives through migration. Thus, on a time scale of 5 years, much of the global diversity is locally available to respond to local selective pressure.

Based on these results, we propose that the global dynamics of rotavirus are characterized by three major elements: (1) strong local selective pressure that drives replacement of dominant serotypes, (2) genetic variation that allows for local replacement of existing serotypes and is introduced predominantly through migration from regional and global pools and (3) the geographic asynchrony in serotype replacement that maintains large standing global diversity levels. These migration-driven global metapopulation dynamics differ significantly from those of influenza A, in which mutation plays a central role(Koel et al. 2013). This model should be considered for other pathogens that have high levels of diversity but lack the characteristic signatures of positive selection. This model can reconcile the appearance of strong immune mediated selection, based on rapid local serotype replacement, with the lack of a signature of positive selection in the gene sequence, or the structure of the phylogeny.

The conclusions are limited to the data available, which have been unevenly sampled in time and space. In an effort to minimize the sampling bias, we included as much data as possible (Methods). The phylogenetic approach itself limits the bias introduced by oversampling local sequences, as very similar sequences will rapidly coalesce reducing the contribution that each adds to the estimate of the mean distance to the vaccine. Additional sampling is, in general, expected to increase the number of migration events observed. Thus, the migration rates are likely underestimated.

Vaccine driven evolution suggests the vaccine will become less effective over time, as strains that are less affected by the vaccine increase in frequency relative to the more susceptible strains. The degree to which vaccine effectiveness will wane depends on the amount of cross immunity between current vaccines and non-included strains. Recent analyses suggest that current vaccines have significant protection against partially or fully heterotypic strains(Dóró et al. 2014; Leshem et al. 2014b). Thus, we would expect the vaccines to retain a significant amount of efficacy despite the evolution in the rotavirus population. However, the differential spread of strains following widespread vaccination is strong evidence that the protection is not equal across all genotypes. Well-designed, quantitative studies of vaccine effectiveness over time are needed. In this setting, a polyvalent strategy is expected to be superior to a monovalent strategy.

Over the time frame that we have examined, the genetic change in the USA was made possible by the introduction of alleles from other geographic regions, and not from novel mutations. Although our findings do not rule out the possibility of sequence evolution over longer time periods, they indicate that extensive surveys of the available global diversity will contain critical information on the genetic material available to respond to vaccination pressures, which can also guide development of vaccines more robust to evolutionary change. Thus, there is a pressing need for increased sampling from around the world, and specifically in Africa, where extensive diversity is thought to exist and vaccination appears to be less effective (Santos and Hoshino 2005b; E. A. S. Nelson and Glass 2010).

5.6 Materials and Methods

Download of data from Genbank. On October 9, 2013 a GenBank search was conducted using the search term “rotavirus” and limiting the results to sequences with a length between 500 and 4000 base pairs (bp). This resulted in 25,369 total sequences. The sequences were then assigned to a segment as follows. A subset of about 500 sequences from each segment that were clearly annotated to a specific segment were used to create a consensus sequence for each segment. The 11 consensus sequences were used to create a local reference file. Using local blastn, all 25,369 sequences were compared to the reference file, with an adjusted word size of 9 (Camacho et al. 2009). Sequences assigned to a segment with low e-score were effectively removed in subsequent steps, as was confirmed by visual inspection of alignments. The data were further screened for the following criteria: (1) able to assign the year at which they were collected, (2) able to assign a country of origin, (3) able to assign a host species from which the sequence was isolated. Sequences were removed if they were clearly marked as being from tissue culture with "tc", "TC", "Vaccine" was in the strain name, or the id contained the name of a known vaccine or common used laboratory strain. Only VP7 sequences of greater than 962bp were used. This search resulted in 1911 VP7 sequences isolated from human hosts that were included in the phylogenetic analysis (Figure 5.5 and Figure 5.6). Each sequence was assigned a G-type using rotaC (Maes et al. 2009).

Phylogenetic analysis. The resulting 1911 human origin VP7 sequences were aligned with MUSCLE (Edgar 2004) and visually inspected to ensure appropriate alignment. A Bayesian phylogenetic model was estimated using BEAST2 (Bouckaert et al. 2014). RNA sequence evolution was modeled with a constant clock and an HKY base-substitution model (Hasegawa, Kishino, and Yano 1985) for which codon substitution rates and kappa values were estimated separately for each codon position. Additionally, the defined 29 amino acids, which make up the defined epitopes, were partitioned to allow them to evolve at a different rate. The population was modeled as a coalescent with exponential population growth. A discrete migration model was fit using the six WHO geographic regions, with the exception of the USA, which was assigned its own geographic region. In addition to the migration rate, a rate indicator was estimated for each pair of regions to determine if any migration between these regions was supported. Four independent runs in BEAST2, for a total of over 100 million MCMC steps, were performed. The first half of each chain was removed as "burn in". The remaining chains were sampled at even intervals, resulting in 1270 trees and parameter samples from the posterior. Results were visualized in Tracer for convergence.

Ancestral state reconstruction and stochastic mapping. In order to test specific hypotheses regarding the geographic location of mutational changes, a stochastic realization of ancestral states was performed for each tree and associated set of parameters. Ancestral state reconstruction was performed for each tree from the posterior output from BEAST2, using the parameters (i.e. mutation and migration parameters) associated with that MCMC sample. Ancestral state reconstruction was performed using the standard approach, of using Felsenstein's tree-pruning algorithm to calculate probability distribution of ancestral states at each internal node (J Felsenstein 1981). Stochastic mapping is an additional step, following ancestral state reconstruction, which produces possible realizations of the individual mutation and migration events as they may have occurred across a phylogeny. Stochastic mapping was performed using established methods whereby events are mapped onto branches of the tree as a stochastic realization of a Poisson process (for both migration and mutations), fulfilling the condition that the state of the nodes at the beginning and end of the branch correspond to those chosen at the previous step (Nielsen 2002; J. P. Bollback 2006). The analysis was implemented in a Java code package SeasMig (http://bitbucket.com/pascualgroup/seasmig)

Partitioning genetic change following vaccination among evolutionary mechanisms. In order to identify the evolutionary mechanisms that accounted for the changing in genetic distance to the vaccine over time, each of the 1270 phylogenetic trees having migration events and mutation events mapped were analyzed. The post-vaccine introduction time period was taken to be 2007 to 2012, and was compared to a pre-vaccine reference time period from 2004 to 2006. This reference time period was chosen because it corresponds to the period just prior to vaccine introduction, and a period with increased sampling, both in the US and the rest of the world (Figure 5.1B). To quantify the role of migration in the shift in the genetic distance to the vaccine following vaccine introduction, we compared the average distance to the vaccine (measured in number of amino acids, out of the 29 amino acids that make up the epitopes) of all branches in the tree that were predicted to be in the USA from the pre-vaccine period, 4.19 aa, to the post-vaccine period, 6.79 aa, which is a difference of 2.60 aa. We then repeated this process, but ignored all sequences that were predicted to have entered the USA after 2006, yielding a difference of 1.22. Thus migration accounted for 1.38 aa of the change. Similarly, to quantify the role of mutation in contributing to the shift in the genetic distance to the vaccine following vaccine introduction we compared the observed difference form the pre-vaccine to post vaccine periods in the USA (2.60 aa), to the difference if every mutation that was predicted to have occurred in the USA after 2006 was reverted to its pre-vaccine state (2.59aa), revealing virtually no change due to mutations occurring in the USA.

Finally, by removing the effects of migration and mutation, the remaining genetic change in the USA was due to the shifting frequency of strains that were already in the USA in 2006, or 1.21 aa.

Positive Selection Analysis. A maximum likelihood tree was constructed using an HKY with gamma+invariant sites model, in MEGA 5.2.2(Tamura et al. 2011), and HyPhy (implemented in MEGA) was used to estimate dn and ds for each codon position. Not a single codon position had a significantly positive dn-ds. 266 of 324 had a significantly negative dn-ds value (Figure 5.7).

5.7 Author contribution

Author list: Robert J Woods, Daniel Zinder, Mercedes Pascual

Author contribution: RJW, DZ, and MMP were involved in experimental design, analysis and writing.

My main contribution to this work, was a joint contribution in establishing the questions, and designing and running the methods which relate to partitioning the mechanisms which contribute to a change in the distance from the vaccine. I was involved in drafting the paper from conception, specifically parts of the methods section.

5.8 Figures

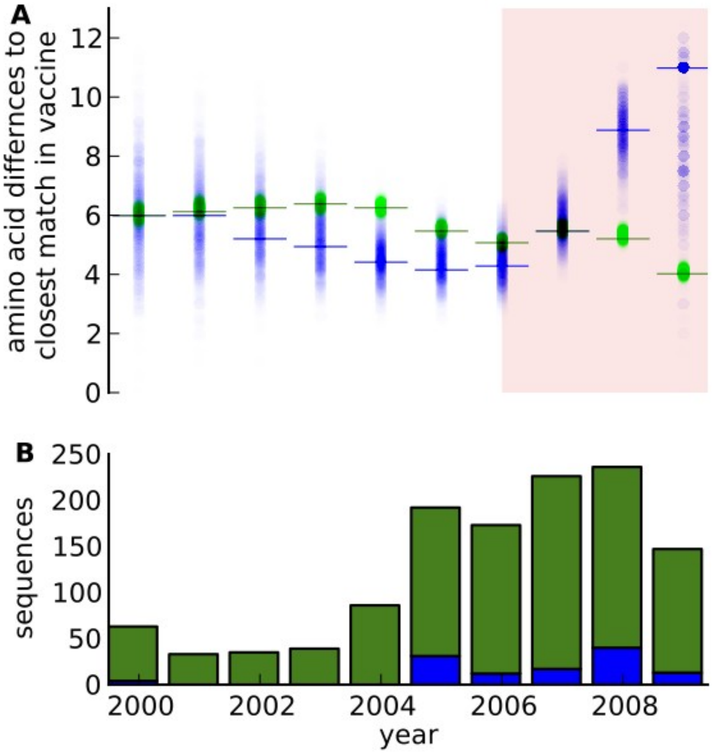


Figure 5.1. Change in amino-acid match with vaccine with time
A. Mean number of amino acid mismatches (out of 29) between the epitope sequences of the vaccine strains and those of sequences predicted to be in the USA (blue) and in the rest of the world (green), from individual MCMC samples (circles) and from the mode of the samples from the posterior distribution (lines). Pink shading indicates the time of vaccination in the USA. **B**. Number of sequences included in the analysis, from the USA (blue) and the rest of the world (green).

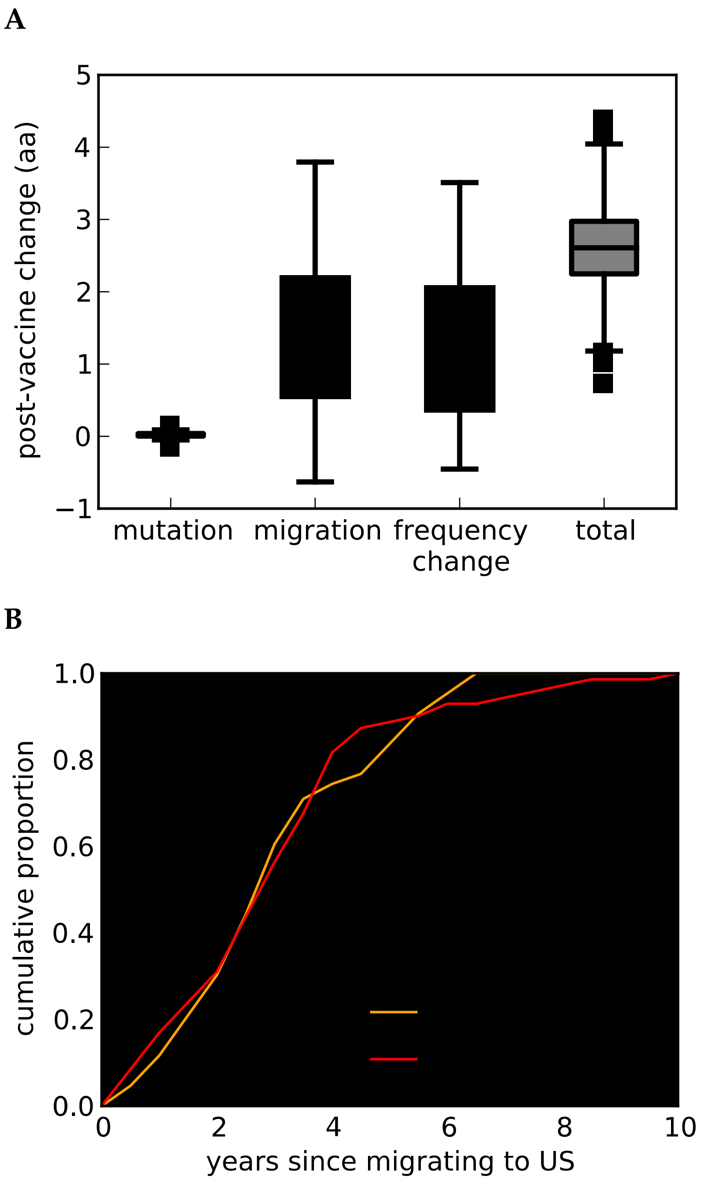


Figure 5.2. Partitioning the epitope change following vaccination in the USA among evolutionary mechanisms

A. Shift in the mean number of amino acid mismatches in the USA from before (2004-2006) to after (2007-2012) vaccination, partitioned into the fractions due respectively to mutation, migration, and shifting prevalence of strains already present in the USA in 2006. **B**. The predicted distribution of time since arrival to the USA of strains sampled before vaccination (2004-2006) and after vaccination (2007-2012). Shading indicates the 95% credibility intervals.

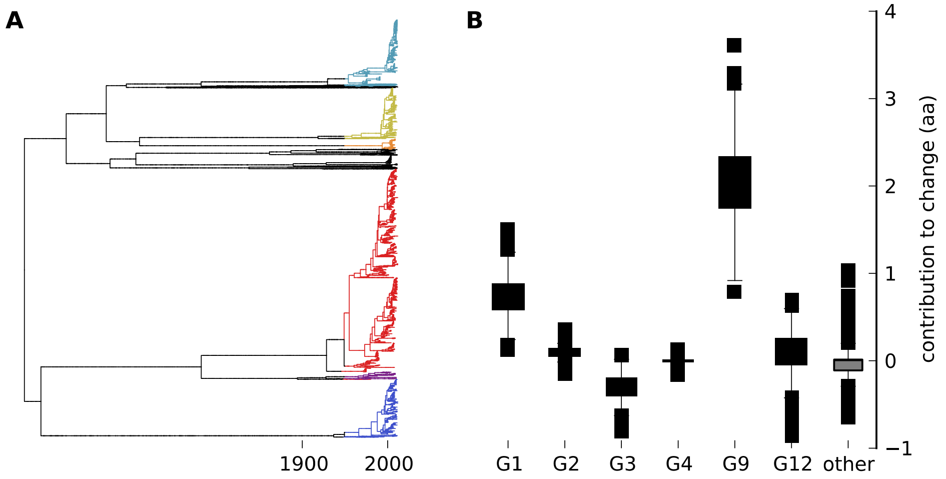


Figure 5.3. The contribution of common serotypes to the increased distance to the vaccine in the USA following vaccine introduction

A. A single representative phylogeny from the posterior distribution of trees, color-coded by serotype as indicated in panel B. **B**. The amount, in amino acids, that each serotype contributed to the total shift in the vaccine from the time prior to vaccination (2004-2006) to the period just after vaccination (2007-2012).

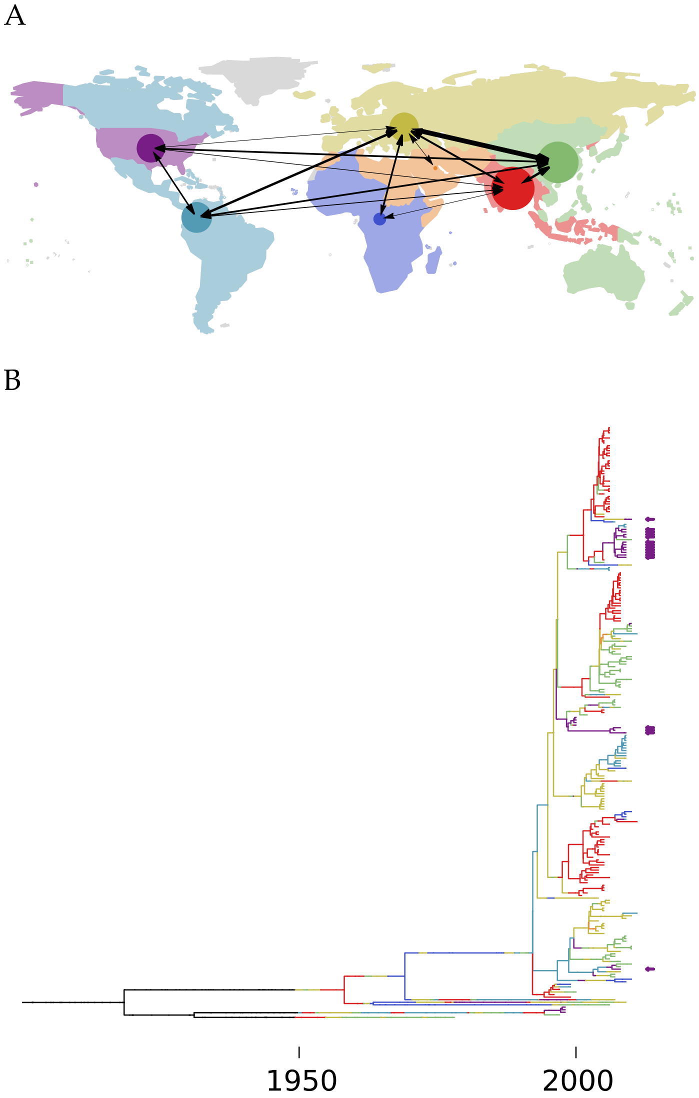


Figure 5.4. Global migration of rotavirus

A. Global migration rates, indicated with arrow thickness proportional rate. The area of the circles is proportional to the number of samples from each geographic region. **B.** The G9 portion of a single tree from the Bayesian phylogenetic analysis, with stochastic mapping of location, color coded by the geographical regions of panel A. Tips representing samples from the USA after 2006 are indicated with arrows, and represent four distinct introductions into the USA.

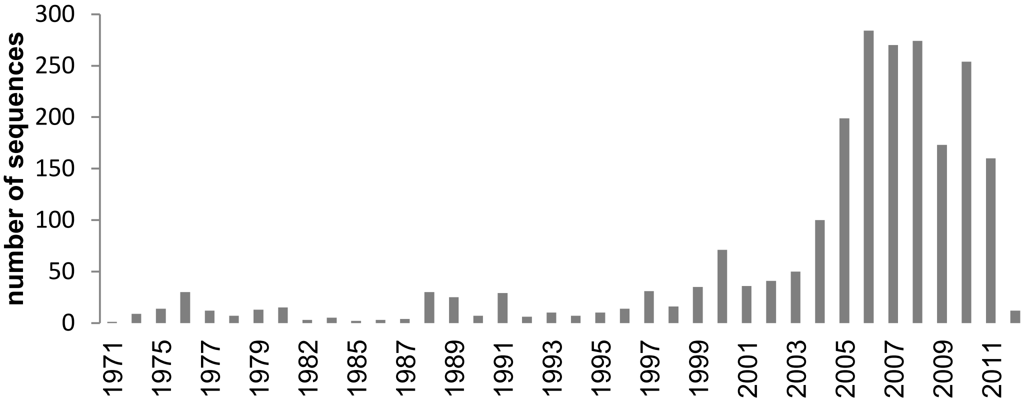


Figure 5.5. Distribution of collection dates of the 1911 sequence used in the Bayesian phylogenetic analysis.

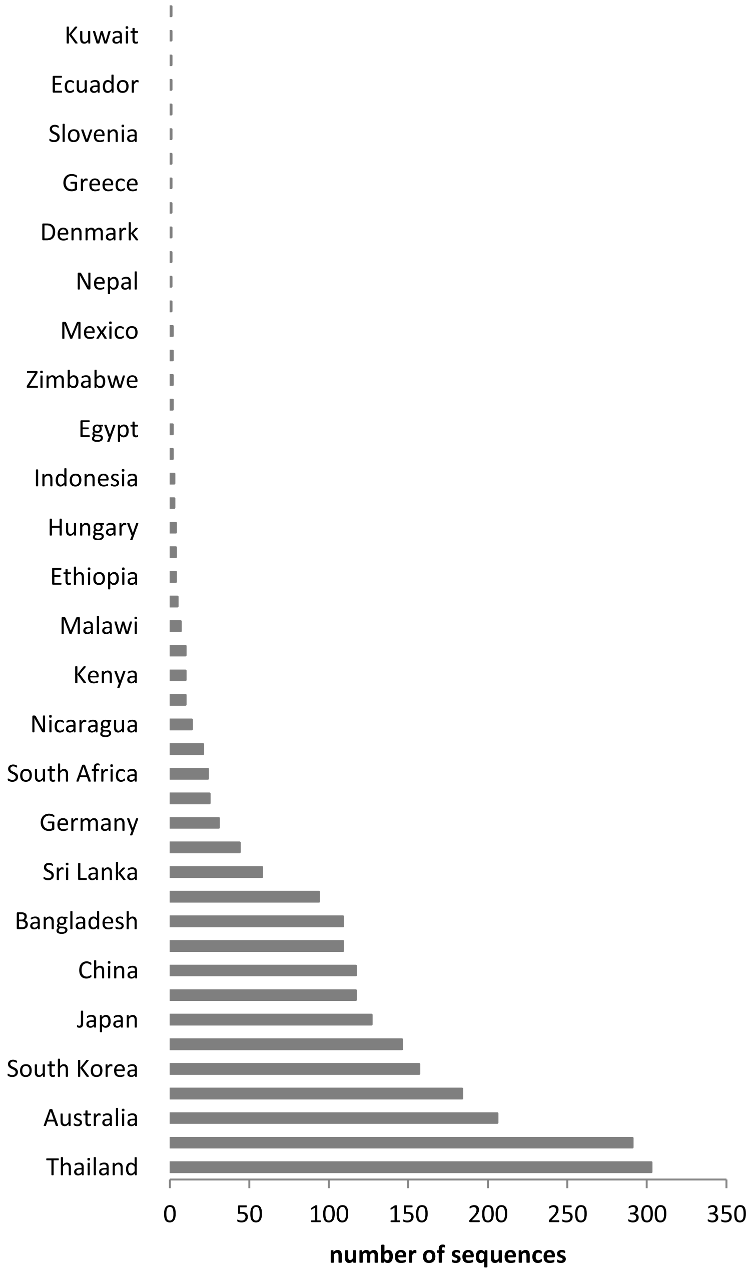


Figure 5.6. Distribution of country of origin of the 1911 sequences used in the Bayesian phylogenetic analysis.

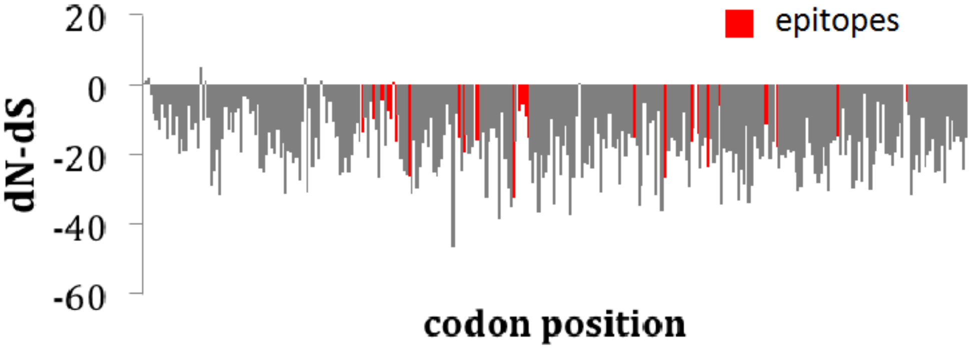


Figure 5.7. Selection analysis by codon position in the VP7 gene
(see supplemental methods for details). Codon position, in order from 1 to 324 is indicated on the x-axis; the dn – ds for each codon position in VP7 is indicted on the y-axis. Negative selection was identified in 266 of 324 sites (p<0.05). Positive selection was identified in no sites. The 29 codons corresponding to amino acids mapped to epitopes of neutralizing antibodies are highlighted in red.

5.9 Tables

Table 5.1. Migration rates, per year, estimated from the Bayesian phylogenetic model between each pair of the WHO geographic regions. The upper and lower bounds of the 95% credibility intervals are indicated.

Location 1	Location 2	median	lower	upper
USA	Africa	0.0000	0.0000	0.0029
USA	Americas	0.0500	0.0193	0.1093
USA	SE Asia	0.0234	0.0086	0.0545
USA	Europe	0.0169	0.0000	0.0704
USA	Eastern Mediterranean	0.0000	0.0000	0.0000
USA	Western Pacific	0.0538	0.0223	0.1168
Africa	Americas	0.0000	0.0000	0.0206
Africa	SE Asia	0.0049	0.0000	0.0343
Africa	Europe	0.0530	0.0192	0.1240
Africa	Eastern Mediterranean	0.0000	0.0000	0.0073
Africa	Western Pacific	0.0000	0.0000	0.0000
Americas	SE Asia	0.0279	0.0088	0.0646
Americas	Europe	0.0744	0.0306	0.1581
Americas	Eastern Mediterranean	0.0000	0.0000	0.0000
Americas	Western Pacific	0.0581	0.0245	0.1202
SE Asia	Europe	0.0633	0.0259	0.1228
SE Asia	Eastern Mediterranean	0.0000	0.0000	0.0000
SE Asia	Western Pacific	0.0715	0.0333	0.1369
Europe	Eastern Mediterranean	0.0109	0.0022	0.0335
Europe	Western Pacific	0.1457	0.0666	0.2688
Eastern Mediterranean	Western Pacific	0.0000	0.0000	0.0000

Chapter 6. Conclusions

The rate and mechanisms of viral evolution should be considered when attempting to understand patterns of antigenic and genetic diversity. Ecological models of competition, replacement and coexistence of viral strains should be developed to include and consider evolutionary processes generating diversity, including mutation migration, zoonosis and reassortment. Although models of species co-existence and competition are at the core of theoretical ecology (e.g. Hubbell 2001; Armstrong and McGehee 1980; MacArthur and Levins 1967; Scheffer and van Nes 2006; Chesson 2000) only some have considered explicit modeling of evolution and, more importantly, have not done so in a dynamical and stochastic context. An example of a model in theoretical disease ecology which did not explicitly model evolution is 'strain theory' (S. Gupta 1998; Recker et al. 2007). My thesis is a contribution to the further development of this interface between ecology and evolution, from the perspective of two major viral pathogens.

In the first research chapter, I explore the phylogenetic consequence of different regimes described by 'strain theory' and show additional dynamic regimes, such as ones involving continuous antigenic diversification, which are generated when the rate of evolutionary change through mutation is explicitly considered. I place influenza in this context as 'mutation limited' meaning that limits on the exploration of antigenic space drives phylodynamics, rather than the opening of a given 'niche' in the resource space constituted by the immune status of individuals in the population, or availability of resources in the form of susceptible hosts. This finding has implications in the context of intervention and vaccines. Under a mutation limited regime, a hypothetical vaccine based on the antigens for which natural immunity is commonly generated against, will only be effective until a new antigenic variant is introduced to the population through migration or mutation. The alternative hypothesis suggested by Recker et al. (Recker et al. 2007) would imply that seasonal influenza is selection limited, with a response to vaccination from an existing antigenic pool. As a result, a sufficient representation of this pool in a vaccine may provide long term protection. Although in the first research chapter I suggest and find support for the possibility of a limited antigenic pool in influenza, the exploration of the amount of antigenic variability that A/H3N2 can generate, should be explored from a more molecular and biological approach.

The study of theoretical models for strain or species co-existence should also consider evolution through recombination. Previous work in the context of strain theory has shown strain structure to be mostly robust to the presence of recombination (Caroline O. Buckee et al. 2011). However, this previous model considered immunity which is strong and specific, increasing the ratio of inter vs. intra specific competition and promoting co-existence (Armstrong and McGehee 1980). As was demonstrated before in the context of influenza (Ferguson, Galvani, and Bush 2003a), in which generalized immunity was shown to reduce antigenic and genetic diversity, both specific and generalized cross-immunity between strains influences phylodynamic patterns. In the second research chapter, I show how generalized immunity and reassortment can influence strain structure in a model of rotavirus. In contrast with previous work (Caroline O. Buckee et al. 2011; Zinder et al. 2013), with a sufficient level of generalized immunity, immune mediated strain structure was maintained even with low levels of cross-immunity between strains. With the addition of an unlimited antigenic pool, this structure was perturbed and strongly disrupted when reassortment was present, generating a new dynamic regime which involves both the signature of immune-mediated *niche differentiation* and of the sweeping of immunologically advantageous alleles across multiple backgrounds through reassortment and *frequency- dependent selection*.

As was the case in the modeling of rotavirus and influenza, phylodynamic methods in general could benefit from the inclusion of improved models of population immunity, a difficult task given the inherent memory-less nature of many phenotypic evolution models used in phylogenetic inference such as CTMC models. In this chapter, I also refer to the possibility that targeting a commonly shared alleles as a hypothetical vaccine antigen candidate may prove less effective, if such an allele is the result of a population sweep. This will be the case as the shared allele has a temporary fitness advantage because of reduced population immunity, rather than being the consequence of stable fitness differences between alleles.

To elucidate the different roles of different evolutionary mechanisms involved in viral strain evolution, including mutation, migration and changes in the abundance of locally circulating strains, I developed a tool (*SeasMig*) which offers stochastic mapping of seasonal migration and mutation processes (Appendix A) in a Bayesian manner. I use this tool in the two inference chapters and gain additional insights into the phylodynamics of both influenza and rotavirus.

Specifically, in the third research chapter, I use *SeasMig* to explore patterns of seasonality in H3N2 migration. Although influenza A/H3N2 is highly seasonal, the drivers of this seasonality are largely unknown with some pertaining to the direct role of weather on transmission dynamics e.g. (Shaman et al. 2010) and others to the effect of seasonal changes (e.g. sunlight) on host susceptibility (Cannell et al. 2006). Regardless of the specific mechanism of incidence seasonality in influenza, with the exception of three studies (Bahl et al. 2011; Bielejec et al. 2014; Philippe Lemey et al. 2014), the implications of such seasonality on phylogenetic patterns has largely been ignored or has remained unknown. In this work, I identify the phylogenetic consequence of the seasonality of influenza incidence. I identify the roles of 'fertile epidemic grounds' and 'end to beginning' transmission in global phylogeographic migration patterns. This work has implications in the context of intervention, and it suggests that the correct timing in the season should be considered when evaluating vaccine strain candidates. This work could benefit from improved sampling and surveillance, which can promote better prediction of influenza viruses migration dynamics.

In the fourth research chapter, I use *SeasMig* to explore the mechanisms behind changes in the US rotavirus population following vaccination. Doing so, I identify migration rather than mutation or changes in the prevalence of existing serotypes as the main driver of the population dynamics in the US. Although several studies have surveyed the rotavirus population following vaccination, and suggested selection pressures imposed by vaccination (Guerra et al. 2015; Dóró et al. 2014), this has not been done in the US, and the different mechanisms by which such a change would occur have not been fully explored. Implications from this study suggest that a polyvalent vaccine strategy against rotavirus may be superior for longer time periods, and that continuous surveillance of vaccine efficacy and circulating strains of rotavirus are necessary in the US and across the world to identify sources of antigenic novelty.

Despite the growing availability of viral sequence data, global sequence sampling is at the best adequate and very often poor. Sampling of rotavirus and influenza sequences is missing from critical areas of human habitation, which may harbor great diversity, including India and Africa. For reassortment studies sequences including the full genome are less common, and available sampling is often interspersed through space and time. Certain schemes have been devised to consider the influence of biased sampling on the analysis of phylogeographic migration patterns and of diversity. These include the use of structured coalescence models of migration, which are somewhat more robust to sampling (P. Beerli and Felsenstein 1999; Stack et al. 2010) and the inclusion of an observation model in birth-death processes (Tanja Stadler 2009). However, further work remains in the field, and ultimately the quality of inference will depend on the availability of data.

Finally, I would like to suggest future research directions in the context of phylogenetic inference. As the number of possible phylogenetic trees grows super exponentially with the number of taxa., traversal of tree space through MCMC methods is extremely slow and limited to very few thousand taxa when Bayesian methods are employed. Several developments have been made through the use of parallel and distributed processing units such as GPUs (Ayres et al. 2012; Suchard and Rambaut 2009). Algorithmic improvements to phylogenetic inference methods have been made through the use of alternative likelihood methods such as approximate Bayesian likelihood (Ratmann et al. 2012). In general, to improve performance, both inference and MCMC mixing could benefit from the inclusion of pseudo- approximate- and synthetic- likelihoods (J. Wang 2001; Ratmann et al. 2012; Wood 2010). Using such approaches to promote mixing of chains will have the benefit of maintaining the use of an 'exact likelihood', while somewhat improving convergence speeds (unpublished *bitbucket.org/dzinder/strangeheat*). Furthermore, the traversal of tree and parameter space, which may have many valleys and terraces (Sanderson, McMahon, and Steel 2011) could benefit from the inclusion of recombination between trees during MCMC inference (unpublished *bitbucket.org/dzinder/tremcomb*).

I would like to thank Mercedes and my committee again for
the patience and time dedicated to help me accomplishing this work

DZ May 2015

Appendix A - Bayesian inference and stochastic mapping of seasonal migration and mutation processes from phylogenetic tree distributions (*SeasMig*)

Supplement for: Seasonality in the migration and establishment of H3N2 Influenza with epidemic growth and decline (Zinder et al.)

In this supplement we describe in further detail the general Bayesian modeling approach, the mathematical details of the model, and the computational techniques used to perform inference and model selection. We demonstrate the method for several simulated scenarios in the context of seasonal migration.

Note: Sections 2.10-2.13 of this text partially overlap with (Baskerville, et al. 2011; Baskerville et al. 2013) *supplementary-text and* (Baskerville et al. 2013), *with author's (co-author Ed Baskerville) permission.*

A.1 Inference of migration processes from tree distributions

A variety of tools has been developed to generate phylogenetic trees from sequence data. Some are Bayesian in nature and provide a distribution of possible trees (A. J. Drummond et al. 2012; Huelsenbeck and Ronquist 2001; Pagel, Meade, and Barker 2004). These trees are sampled according to their likelihood and according to the given prior probability for the parameters used when estimating this likelihood. A sample from this distribution is referred to as a sample from the empirical posterior distribution of trees. Other phylogenetic tree reconstruction tools capable of handling larger datasets e.g. (J. Felsenstein 1981; Swofford 2003) often provide a single tree output or distributions which are based on randomizations of the data. It is very common for tree reconstruction to be the most computationally expensive part of phylogenetic analysis. This is the case since the number of possible trees grows super-exponentially in relation to the number of tips (or taxa).

In some cases it is possible to separate the inference of tree topology based on nucleotide data, from additional steps relating to the inference of phylogeography and phenotypic traits. This is the case when the contribution of these traits to the combined sequence and trait based tree likelihood is sufficiently small. In this case, the tree distribution is mostly defined based on sequence data and it can be further refined and used for trait based analysis (Pagel, Meade, and Barker 2004). If all the trees based on nucleotide data were enumerated, a step which is computationally infeasible for more than a few taxa, than an additional step involving inference based on traits will be mathematically equivalent to joint inference based on nucleotide and trait data, as long as the evolutionary models for traits and for individual nucleotide substitutions are independent. It is not easy to know exactly how many tree samples based on nucleotide data are required for trait based inference, but it is necessary that the sample should be large enough to capture the major different tree topologies.

In addition, given a distribution of trees, it is often useful to generate stochastic realizations of possible mutation and migration events as they occurred along the branches of the tree (J. Bollback 2006; Huelsenbeck, Nielsen, and Bollback 2003; Minin and Suchard 2008). This can be done in a Bayesian manner, where for each tree topology coupled with a corresponding mutational and trait evolutionary model sampled form the empirical posterior distribution, a stochastic realization is generated. Finally, the different genes, or proteins of an organism, may have alternative evolutionary histories. If the underlying migration processes is assumed to be the same, we can use information from multiple proteins together when inferring migration processes.

A.2 Implementation

2.1 **SeasMig**

We implemented in Java a tool (http://bitbucket.org/pascualgroup/seasmig) for migration model inference. This tool can also perform stochastic mapping, based on an initial distribution of trees and geographic annotations. Alternative migration model parameters can be inferred and compared by their marginal likelihoods including seasonal and epochal phylogeographic migration models. An empirical distribution of trees in nexus format is given as input. Our tool uses an MCMC to sample from a posterior distribution of model parameters and stochastically mapped migration events along branches and trunk lineages.

2.2 **Bayesian Inference**

In a Bayesian framework, both the data and the model parameters are assumed to be stochastic. Rather than finding the set of parameters that maximizes the likelihood of a particular observation, we estimate the distribution of the model parameters that can lead to the observed data. The probability of observing a specific set of model parameters θ conditioned on observing the data D is known as the *posterior* probability and can be written according to Bayes' rule:

$$Pr(\theta|D) = \frac{Pr(\theta)Pr(D|\theta)}{Pr(D)} \tag{1}$$

Pr(θ) denotes he prior probability of observing a specific set of parameters, while Pr(D|θ) denotes the likelihood of observing the data D given the model parameters θ. The probability of observing the data Pr(D) without the context of a model (or models) is most often unknown. As such samples from the posterior distribution are known in probability often only with relation to other samples.

The probability of observing the data Pr(D|M) in a context of a specific model M, used to fit the data, can be calculated by summing up the probability of observing the specific model parameters (*prior*) multiplied by the probability of observing the data conditioned on the model parameters (*likelihood*), across all the parameter values θ:

$$\mathrm{Pr(D|M)} = \sum_{\theta} \mathrm{Pr(}\,\theta\mathrm{)Pr(D|}\theta\mathrm{)} \tag{2}$$

or in a more general continuous notation:

$$\mathrm{Pr(D|M)} = \int_{\theta} \mathrm{f(}\theta\mathrm{)f(D|}\theta\mathrm{)d}\theta \tag{3}$$

Where Pr(D|M) is referred to as the *marginal likelihood*, f(θ) is the *prior distribution* and f(D|θ) is the *likelihood function*.

2.3 Non-Seasonal Migration Model

We assume that discrete geographic location diffuses along branches of the tree following a continuous time Markov chain (CTMC) process. In this case, non-seasonal migration processes are characterized by a single rate matrix Q:

$$Q = \begin{pmatrix} -\sum_{i\neq 1}^{n} r_{1i} & r_{12} & \cdots & r_{1n} \\ r_{21} & -\sum_{i\neq 2}^{n} r_{2i} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & -\sum_{i\neq n}^{n} r_{ni} \end{pmatrix} \qquad (4)$$

where r_{ij} represents the migration rate between location i to location j.

2.4 Non-Seasonal Migration Model Parameterization and Priors

Rates are assumed to have an exponential prior $r_{ij} \sim \exp(\lambda)$ with a rate hyperprior parameter λ which is shared across all the rates and is itself exponentially distributed with unit mean $\lambda \sim \exp(1)$.

Note: the rate hyper prior was added at a later stage and is not included in non-seasonal analysis in the main body of the text .

2.5 Matrix Exponentiation

Matrix exponentiation is used to convert migration rates, to probability distributions, which concern the state of nodes along the tree. We first focus on processes along individual branches of the phylogenetic tree.

Given a branch connecting parent node x to child node y, $x \rightarrow y$, of length t. We assign node x a vector $\bar{p}_x$ which defines its probability of being at state i:

$$Pr(x_{state} = i) = \begin{cases} p_{x1} & i = 1 \\ \vdots & \vdots \\ p_{xn} & i = n \end{cases} \tag{5}$$

We assume, in the simplest case, that states along branches behave as homogeneous Poisson processes with a rate matrix Q, as such the state distribution of node y can be written as:

$$\bar{p}_y = \bar{p}_x \cdot \mathrm{P} \tag{6}$$

Where P is the transition probability matrix and can be calculated as follows:

$$P = \exp{(Qt)} \tag{7}$$

The matrix exponent can be defined by the Taylor expansion of the exponent function:

$$P = I + \frac{(Qt)}{1!} + \frac{(Qt)^2}{2!} + \frac{(Qt)^3}{3!} + \cdots \tag{8}$$

Multiple alternative algorithms are implemented for matrix exponentiation. Several algorithms were either imported (JBLAS) or directly implemented in the code including the Taylor Series, the Padé approximate (Higham 2009), and Eigen-Decomposition. For matrices of rank 3 or less, and for specific cases of matrices of rank 4 (HKY (P. Beerli and Felsenstein 1999), JC69 (Jukes and Cantor 1969)), analytic solutions exist and were implemented. All matrix exponentiation algorithms were cross-validated within the package.

2.6 Two Seasonal Migration Model

A seasonal migration model which is a variation of (Bielejec et al. 2014), was established by using two different migration rate matrices Q_A and Q_B for two parts (seasons) of the year:

$$Q_A = \begin{pmatrix} -\sum_{i\neq 1}^{n} r_{A1i} & r_{A12} & \cdots & r_{A1n} \\ r_{A21} & -\sum_{i\neq 2}^{n} r_{A2i} & \cdots & r_{A2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{An1} & r_{An2} & \cdots & -\sum_{i\neq n}^{n} r_{ni} \end{pmatrix}, Q_B = \cdots \qquad (9)$$

The exact partitioning of the year is defined by the start ϕ and end ε of season A (without loss of generality). Where: $\phi \in [0,1)$ and $\varepsilon \in [\phi, 1)$.

For example, for $\phi = 0$ and $\varepsilon = 0.25$ the rate matrix Q_A applies to all branch parts within January-March, while rate matrix Q_B applies to all branch parts within April-December. To estimate the transition probabilities between two states at different times, the respective transition probability matrices P_A, P_B are calculated for the individual year parts through matrix exponentiation. For instance, for the same partitioning of the year, given a branch $x \rightarrow y$ spanning from year $t_x = 2$ to year $t_y = 2.75$, the state distribution of node y can be calculated as:

$$\bar{p}_b = \bar{p}_a \cdot \mathrm{P} \qquad (10)$$

Where P is the transition probability matrix:

$$P = \exp(Q_A \cdot \Delta t_A) \cdot \exp(Q_B \cdot \Delta t_B) \quad (11)$$

$\Delta t_A = 2.25 - 2 = 0.25$ is the fraction of the branch within season A, and $\Delta t_B = 2.75 - 2.25 = 0.5$ is the fraction of the branch within season B.

2.7 Two Seasonal Migration Model Parameterization and Priors

2.7.1 Migration Seasonality Based on a Specific Source and Destination

Migration rates for the two partitions of the year were parameterized as follows:

$$r_{Aij} = r_{ij} \cdot (1 + \sigma_{ij}), r_{Bij} = r_{ij} \cdot (1 - \sigma_{ij}) \quad (12)$$

Where r_{ij} is referred to as the mean migration rate, and σ_{ij} are referred to as the *seasonal scaling* parameters.

As is the case in the non-seasonal model, mean rates are assumed to have an exponential prior $r_{ij} \sim \exp(\lambda)$ with a rate hyper prior parameter λ which is shared across all the rates and is itself exponentially distributed with unit mean $\lambda \sim \exp(1)$. The *seasonal scaling* parameters ($\sigma_{12} .. \sigma_{n-1,n}$) are assumed to have a uniform prior $\sigma_{ij} \sim U(-1,1)$. The scaling parameter was used instead of two separate rates, to separate the inference of mean migration rates, from the inference of the seasonality of migration.

2.7.2 Migration Seasonality Based on Source

For source based migration seasonality rates (Equation 9) are parameterized in the following way:

$$r_{Aij} = r_{ij} \cdot (1 + \sigma_{Si}), r_{Bij} = r_{ij} \cdot (1 - \sigma_{Si}) \tag{13}$$

where σ_{Si} are the source based *seasonal scaling* parameters. Mean rates are assumed to have an exponential prior $r_{ij} \sim \exp(\lambda)$ with a rate hyper prior parameter λ which is shared across all the rates and is itself exponentially distributed with unit mean $\lambda \sim \exp(1)$. The *seasonal scaling* parameters $(\sigma_{S1}..\sigma_{Sn})$ are assumed to have a uniform prior $\sigma_{Si} \sim U(-1,1)$

2.7.3 Migration Seasonality Based on Destination

For destination based migration seasonality rates (Equation 9) are parameterized in the following way:

$$r_{Aij} = r_{ij} \cdot (1 + \sigma_{Dj}), r_{Bij} = r_{ij} \cdot (1 - \sigma_{Dj}) \tag{14}$$

where σ_{Dj} are the destination based seasonal scaling parameters. Mean rates are assumed to have an exponential prior $r_{ij} \sim \exp(\lambda)$ with a rate hyper prior parameter λ which is shared across all the rates and is itself exponentially distributed with unit mean $\lambda \sim \exp(1)$. The seasonal scaling parameters $(\sigma_{D1}..\sigma_{Dn})$ are assumed to have a uniform prior $\sigma_{Dj} \sim U(-1,1)$.

2.7.4 Migration Seasonality Based on Source and on Destination

For destination based migration seasonality rates (Equation 9) are parameterized in the following way:

$$r_{Aij} = r_{ij} \cdot (1 + \sigma_{Si}) \cdot (1 + \sigma_{Dj}), r_{Bij} = r_{ij} \cdot (1 - \sigma_{Si}) \cdot (1 - \sigma_{Dj}) \tag{15}$$

where σ_{Si} and σ_{Dj} are the source and destination based seasonal scaling parameters respectively. Mean rates are assumed to have an exponential prior r_{ij}~exp (λ) with a rate hyper prior parameter λ which is shared across all the rates and is itself exponentially distributed with unit mean $\lambda\sim$ exp(1). The *seasonal scaling* parameters $(\sigma_{S1}..\sigma_{Sn}, \sigma_{D1}..\sigma_{Dn})$ are assumed to have a uniform prior or $\sigma_{Sj}\sim U(-1,1)$ and $\sigma_{Di}\sim U(-1,1)$

2.8 Tree Likelihood Calculation

Given a tree, a specific and parameterized trait evolutionary (substitution) model, and the state of traits on the tips of the tree, a tree likelihood can be calculated (J. Felsenstein 1981).

In general, this likelihood can be calculated by integrating (enumerating and summing up) the likelihood of all possible internal node states. This is done efficiently by calculating and storing the likelihood of sub-trees, recursively progressing from the tips towards the trunk of the tree (J. Felsenstein 1981).

The transition probability matrix is defined according to Equation 7 for a non-seasonal model and according to Equation 11 for a two seasonal model. The transition probability matrix is used to calculate the likelihood of node states along individual branches of the tree.

The prior assumption $\bar{\pi}$ about the state of the root of the tree usually follows either an equal probability of being at each state, an empirical estimate of being at a given state, or the stationary distribution of the substitution model:

$$\bar{\pi} = \bar{p}_0 \cdot \lim_{t\to\infty} \exp\,(Qt) \qquad (16)$$

where $\bar{p}_0$ is the initial state of the system and assumed to be an equal probability of being in each state The value of $\bar{p}_0$ is only relevant if isolated populations exist, and stationary conditions depend on their populace.

Since there is no such stationary distribution for a seasonal model, we used the stationary distribution of the corresponding seasonal migration matrix at the root node time, this assumes some convergence to the stationary distribution within each season. Alternative estimates can be derived. The inference is not sensitive to the specific root prior assumptions in this case.

2.9 **Stochastic Mapping**

Stochastic mapping is an additional step following the calculation of tree likelihood and ancestral state reconstruction at the nodes of a tree. This mapping allows us to generate a stochastic realization of the state of branches along the tree, in addition to the state of internal nodes, and in so doing, provides samples of migration and mutation events, and their timing along the tree that lead to the observed tip states. Stochastic mapping of both sequence (nucleotide) and character (*e.g.* geographic) annotations is available in *SeasMig*, together with the option of incorporating seasonal migration models. Stochastic mapping is implemented directly in our code based on (J. Bollback 2006). Improved performance could be achieved using (Minin and Suchard 2008).

A given type of event, migration or mutation, is assumed to behave as a Poisson process along a branch with a rate matrix Q:

$$Q=\begin{pmatrix} -\sum_{i\neq 1}^{n} r_{1i} & r_{12} & \cdots & r_{1n} \\ r_{21} & -\sum_{i\neq 2}^{n} r_{2i} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & -\sum_{i\neq n}^{n} r_{ni} \end{pmatrix} \qquad (17)$$

As such, the timing of the next event given the present state follows an exponential distribution with the rate parameter $\lambda_x = \sum_{i\neq x}^{n} r_{xi}$, where x is the present character state.

Once the timing of the next event is determined, it is chosen based on its relative probability compared to other transition (emigration) events:

$$Pr(y = i) = \begin{cases} \frac{r_{x1}}{\lambda_\mathrm{x}} & i = 1 \\ \vdots & \vdots \\ \frac{r_{xn}}{\lambda_\mathrm{n}} & i = n \end{cases} \tag{18}$$

Given a branch connecting parent node x to child node y, defined to span from time t_x to time t_y, and with ancestrally reconstructed states s_x and s_y respectively. Stochastic events are generated starting from s_x, repeatedly until the state of node y is correctly reconstructed. That is, until an event prior to the timing of node y results in the state s_y, and an additional event time generated is timed to be beyond t_y.

Branch reconstructions that span across seasons were performed by stochastically reconstructing the state between the seasons' boundaries using the initial migration matrix, and by continuing the stochastic mapping forward using the second seasonal matrix and so forth. This is process is reinitialized from node x, until the state of the node y is correctly mapped. The validity of these processes relies on the memory less nature of the Poisson process.

2.10 **Markov-Chain Monte Carlo (MCMC)**

Markov-chain Monte Carlo, or MCMC, is an algorithm that allows sampling from analytically intractable distributions. Such distributions include the distribution of tree likelihoods given a mutational or a migration model.

The general idea of an MCMC method is to set up a sequence of dependent samples $\theta_1, \theta_1, \ldots$ that is guaranteed to converge to a target distribution, in this case the posterior distribution of our model. In the Metropolis-Hastings algorithm, a change is proposed to the current state, drawn from a *proposal distribution* over possible changes $q(\theta \rightarrow \theta^*)$. This change is either rejected, in which case the current sample is repeated, or the proposed change is accepted as the new sample. The Metropolis-Hastings acceptance probability (Metropolis et al. 1953; Hastings 1970):

$$r(\theta \rightarrow \theta^{*}) = min\left\{1, \frac{f(\theta^{*})q(\theta \rightarrow \theta^{*})}{f(\theta)q(\theta \rightarrow \theta^{*})}\right\} \tag{19}$$

guarantees that the sequence of samples will converge to the *posterior distribution*.

2.11 Metropolis-coupled MCMC (MC3)

We use (http://github.com/edbaskerville/mc3kit) for MCMC functionality (Baskerville, et al. 2011; Baskerville et al. 2013). Additional functions for sampling and evaluating tree likelihoods were implemented.

Although the Metropolis-Hastings algorithm is guaranteed to converge to the target distribution at some point, local maxima in the likelihood surface can cause a chain to become stuck for long periods of time. One approach to avoiding this problem, known as “Metropolis coupling”, involves running multiple chains in parallel. One chain, the “cold chain”, explores the target distribution, while the other chains, “hot chains”, explore low-likelihood configurations more freely. Periodically, swaps are proposed between chains, allowing good configurations discovered on hot chains to propagate toward to the cold chain.

Rather than exploring the target distribution $f(\theta|D) \propto f(\theta)f(D|f(\theta|D)$, heated chains explore

$$f_{\tau}(\theta|D) \propto f(\theta)[f_{\tau}(D|\theta)]^{\tau}, \tau \in [0,1] \tag{20}$$

Where τ is a heating parameter. We use uneven spaced values of τ (Friel and Pettitt 2008), with the hottest chain exploring the prior ($\tau = 0$) and the coldest chain exploring the posterior ($\tau = 1$).

Swap moves are standard Metropolis-Hastings proposals, but rather than considering a change to a single chain, they consider a change to the joint distribution of two chains. The acceptance proability is thus the ratio of the joint distribution after and before the move:

$$\begin{aligned} r\left(\left(\theta_i, \theta_j\right) \rightarrow \left(\theta_j, \theta_i\right)\right) &= \frac{f\left(\theta_j\right)\left[f\left(D|\theta_j\right)\right]^{\tau_i} f(\theta_i)[f(D|\theta_i)]^{\tau_j}}{f\left(\theta_j\right)[f(D|\theta_i)]^{\tau_i} f\left(\theta_j\right)\left[f\left(D|\theta_j\right)\right]^{\tau_j}} \\ &= \left[\frac{f(D|\theta_i)}{f(D|\theta_j)}\right]^{\tau_j - \tau_i} \end{aligned} \tag{21}$$

Where θ_i, θ_j are the configurations that begin in chains i and j, and τ_i, τ_j are the heat parameters of the two chains.

The use of multiple heated chains has the side effect of drastically improving estimates of marginal likelihoods for model selection, as described in the next section.

2.12 **Marginal Likelihood Estimation**

Enumeration across all possible model parameters is computationally costly and grows exponentially with the number of model parameters. We would like to use MCMC to estimate the marginal likelihood for the sake of comparison among different models. Marginal likelihood estimates derived from a single chain, such as the harmonic mean estimator of Raftery (Kass and Raftery 1995), converge very slowly, because MCMC fails to sample sufficiently from low-likelihood areas. However, it is possible to use the information gathered about low-likelihood areas in heated chains using a technique called thermodynamic integration (Lartillot and Philippe 2006; Peter Beerli and Palczewski 2010), or path sampling (Calderhead and Girolami 2009).

Assuming a continuum of heated chains, the thermodynamic estimator of the log-marginal likelihood is:

$$log\hat{\mathcal{L}}(M) = \int_0^1 \frac{1}{m}\sum_{i=1}^{m} \pi\left(\theta_{i,\tau}\right) log\, \mathcal{L}\left(\theta_{i,\tau}\right) d\tau \qquad (22)$$

where m is the number of samples in the MCMC output, and $\theta_{i,\tau}$ is a single sample from the output in a chain with heat parameter τ (Peter Beerli and Palczewski 2010). With a finite number of chains, we use the trapezoid rule to numerically integrate this integral (Figure 1), using uneven spacing of heats to improve the estimate (Friel and Pettitt 2008).

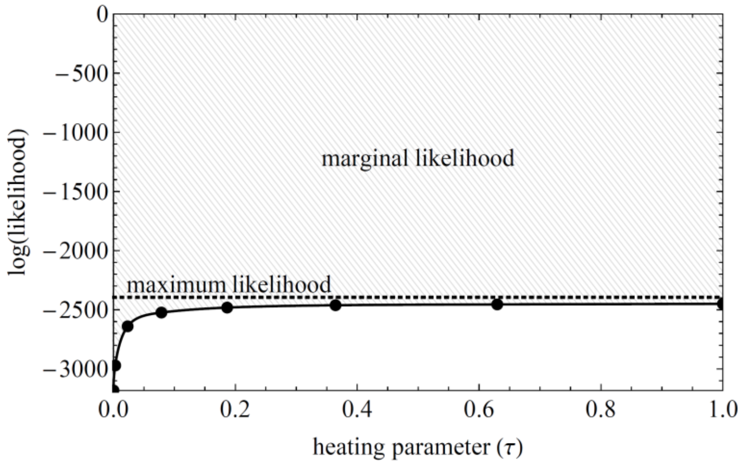


Figure A.1 Thermodynamic integration of the marginal likelihood
The mean likelihoods of each chain (black dots) are interpolated and used to estimate the marginal likelihood (gray area) (Friel and Pettitt 2008). The maximum likelihood (dotted line) is asymptotically approaches the mean likelihood as $\tau \rightarrow \infty$.

2.13 Model Selection via Marginal Likelihood

The Bayesian framework provides a natural way to make probabilistic inferences based on a particular model. However, we also want to be able to choose between different models by quantifying their relative goodness of fit. One approach to Bayesian model selection can be framed directly in terms of Bayes' rule, mirroring the process for estimating the posterior distribution over parameters for a single model.

Consider two models, M_1 and M_2, to which we assign prior weight $Pr(M_1)$ and $Pr(M_2)$. After the data has been observed, we can calculate the posterior probability of the models using Bayes' rule:

$$\begin{aligned} \mathrm{Pr}\,(M_1|D) &= \frac{\mathrm{Pr}\,(M_1)\mathrm{Pr}\,(D|M_1)}{\mathrm{Pr}\,(D)} \\ \mathrm{Pr}\,(M_2|D) &= \frac{\mathrm{Pr}\,(M_2)\mathrm{Pr}\,(D|M_2)}{\mathrm{Pr}\,(D)} \end{aligned} \qquad (23)$$

Where the denominator is equal to the probability of observing the data unconditional of the particular model at play, $\Pr(D) = \Pr(M_1)\Pr(D|M_1) + \Pr(M_2)\,P(D|M_2)$. The probabilities $\Pr(D|M_1) = \int_{\theta_1} f(\theta_1)f(D|\theta_1)d\theta_1$ and $\Pr(D|M_2) = \int_{\theta_2} f(\theta_2)f(D|\theta_2)d\theta_2$ are the marginal likelihoods of the two models, corresponding to Equation 3. If we give the two models equal prior weight, then the relative posterior weight of the two models is simply given by the marginal likelihoods. This reasoning extends naturally to any number of models.

The ratio of the marginal likelihoods is often called the Bayes factor (Jeffreys 1935; Jeffreys 1961; Kass and Raftery 1995), and is equal to the posterior odds ratio of the two models, assuming equal prior weight:

$$B_{12} = \frac{\Pr\,(D|M_1)}{\Pr\,(D|M_2)} \tag{24}$$

The Bayes factor provides a convenient way to compare models: if B_{12}=10, then we consider support for model M_1 to be ten times stronger than model M_2. In AIC-based selection, the Bayes factor is analogous to a ratio of Akaike weights (Burnham and Anderson 2002).

The marginal likelihood of a a model is the likelihood averaged over the prior distribution. That is, it is the likelihood one would expect by randomly sampling parameters from the prior distribution:

$$f(D|M) = \int_{\theta} f(\theta)f(D|\theta)d\theta \tag{25}$$

This value serves as a useful measure of model fit because it directly incorporates the dependence of the likelihood on uncertainty in parameter values, implicitly penalizing extra degrees of freedom (Bolker 2008). If an additional parameter improves the maximum likelihood but decreases the average likelihood, the model suffers from over fitting relative to the simpler model.

2.14 Convergence

Methods for estimating model convergence were not directly implemented within our package. Such tools include (A. Rambaut and Drummond 2003) which can be used to estimate the number of effective number of samples from an MCMC chain. This is necessary since MCMC chains include auto correlated samples.

2.15 Variable Selection

To assess whether the inclusion of migration between different communities is informative, and to establish if rates are seasonal, Bayesian variable selection (O'Hara and Sillanpää 2009) was implemented.

Our implementation is based on (Kuo and Mallick 1998) but differs in that it is implemented within an MC3 framework. Indicator variables which can take a value of either 0 or 1 prefix parameters of interest. Bayes factors for the inclusion of a specific parameter are calculated as:

$$BF = \frac{\Pr\,(D|M, I = 1)}{\Pr\,(D|M, I = 0)} \tag{26}$$

and represent the ratio of the marginal likelihoods of the two models, with and without the variable of interest parameterized. Symmetric non-informative priors were used for the indicators. Bayes factors are estimated as the ratio of the number of posterior samples of the cold chain in which the indicator was 1 compared to 0. The use of an MC3 framework reduces the probability of variables getting stuck in a specific configuration (on or off) as heated chains continue to sample from the prior and flattened likelihood distributions. In theory, it may be possible to use thermodynamic integration to obtain better estimates of Bayes factors.

2.16 Non-Seasonal Migration Model Parameterization with Variable Selection

Rates (Equation 4) are parameterized in the following way $r_{ij} = \tilde{r}_{ij} \cdot \mathrm{I}_{\mathrm{ij}}$ where $\tilde{r}_{ij} \sim \exp\,(\lambda)$ have an exponential prior with a rate hyper prior parameter λ which is shared across all the rates and is itself exponentially distributed with unit mean $\lambda \sim \exp(1)$. The indicators I_{ij} are drawn from an equal probability prior distribution.

Note: a rate hyper prior was added at a later stage and is not included in non-seasonal analysis in the main body of the text.

2.17 Two-Seasonal Migration Model Parameterization with Variable Selection

Rates (Equation 9) are parameterized in the following way:

$$r_{Aij} = r_{ij} \cdot I_{Rij} \cdot \left(1 + \sigma_{ij} \cdot I_{\sigma ij}\right), r_{Bij} = r_{ij} \cdot I_{Rij} \cdot \left(1 - \sigma_{ij} \cdot I_{\sigma ij}\right) \qquad (27)$$

where r_{ij} is referred to as the mean migration rate, and σ_{ij} are referred to as the *seasonal scaling* parameters. As is the case in the non-seasonal model, mean rates are assumed to have an exponential prior $r_{ij} \sim \exp\,(\lambda)$ with a rate hyper prior parameter λ which is shared across all the rates and is itself exponentially distributed with unit mean $\lambda \sim \exp(1)$. The *seasonal scaling* parameters $(\sigma_{12} .. \sigma_{n-1,n})$ are assumed to have a uniform prior $\sigma_{ij} \sim U(-1,1)$. The *seasonal scaling* indicators $\mathrm{I}_{\sigma\mathrm{ij}}$, and the rate indicators $\mathrm{I}_{\mathrm{Rij}}$ are drawn from an equal probability prior distribution.

2.18 Combining the likelihood of multiple protein trees

A conservative approach was used to combine the information present in multiple protein trees with respect to the model likelihood. The combined protein tree log-likelihood is averaged across the multiple protein trees, to account for the possible lack of independence in the information contained in the two trees with respect to migration rates and seasonality. This choice does not affect the maximum likelihood model parameter choice but has the effect of widening confidence intervals when the multiple protein trees provide independent data, while providing the correct confidence interval when the proteins are in complete linkage and have the exact same evolutionary history. Tree weights can be specified as configuration parameters.

A.3 Results

3.1 **Inference of non-seasonal and seasonal migration rates**

In this analysis we infer seasonal and non-seasonal migration rates from a single tree topology and stochastically generated tip locations based on a known input migration model. A single hemagglutinin tree topology with 2859 tips was used for this analysis. Tip collection dates span from 1981-2009. Non-seasonal and two-seasonal migration models without variable-selection are used in this section.

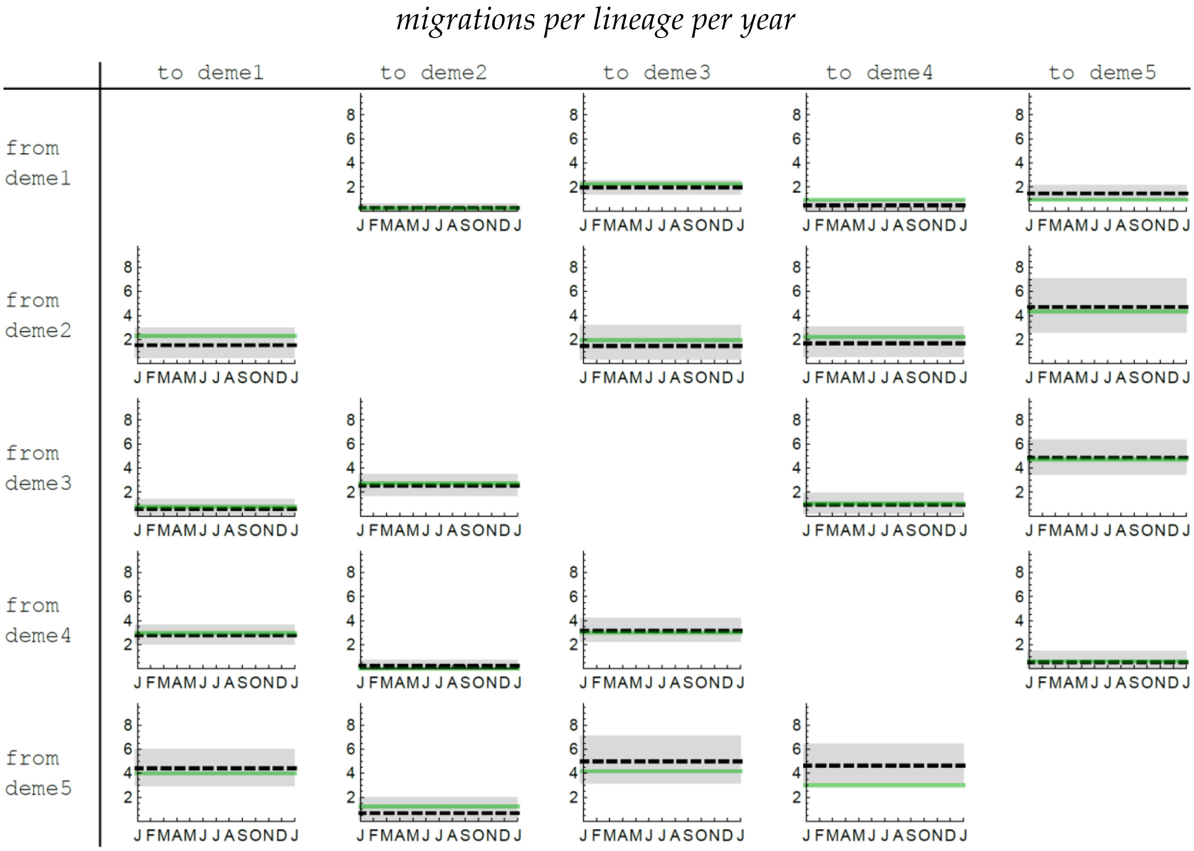


Figure A.2 Inferred migration rates given a non-seasonal tip location generating model Inferred median (dashed) and 90% Bayesian credible intervals (gray) (CI) for migration rates between five locations. Tip locations were generated stochastically using an input non-seasonal migration model (green line). A non-seasonal migration rate model was used for inference.

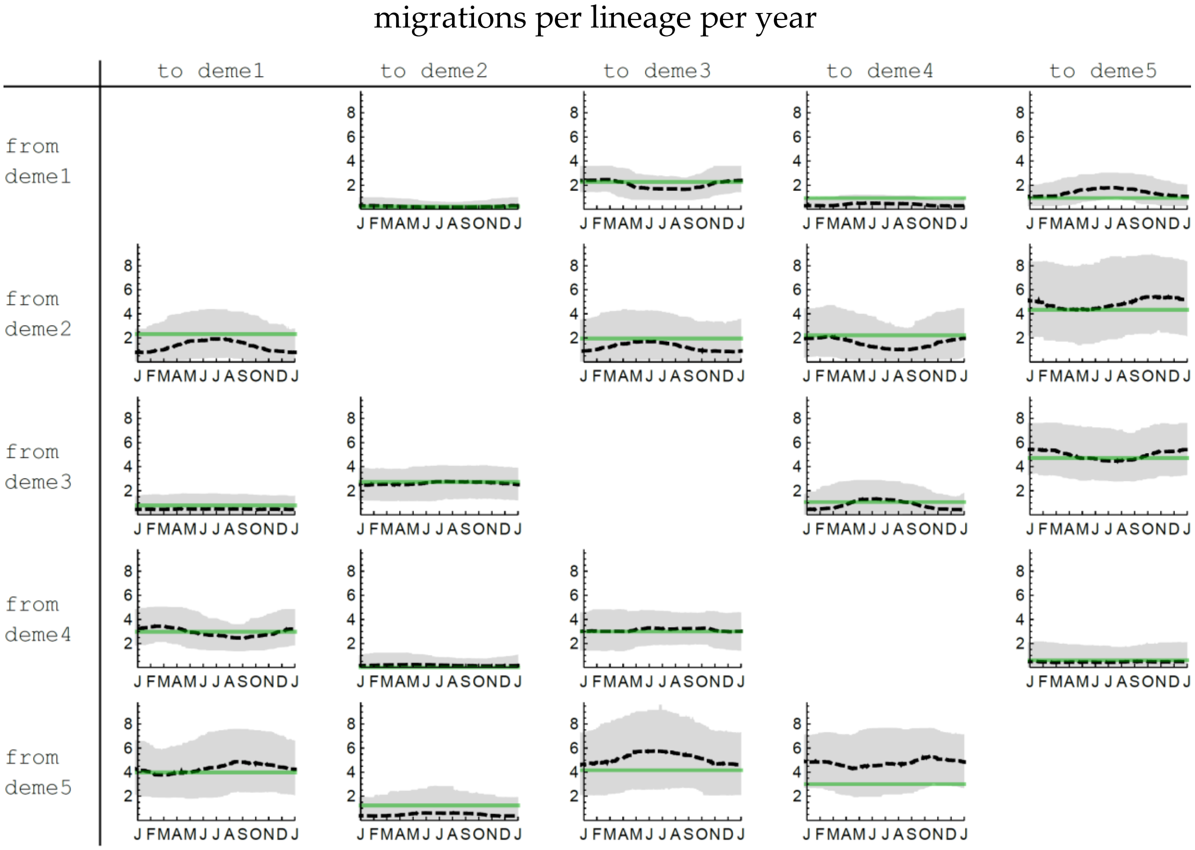


Figure A.3 Inferred migration rates given a non-seasonal tip location generating model Inferred median (dashed) and 90% CI (gray) migration rates between five locations. Tip locations were generated stochastically using an input non-seasonal migration model (green line). A two-seasonal migration rate model was used for inference.

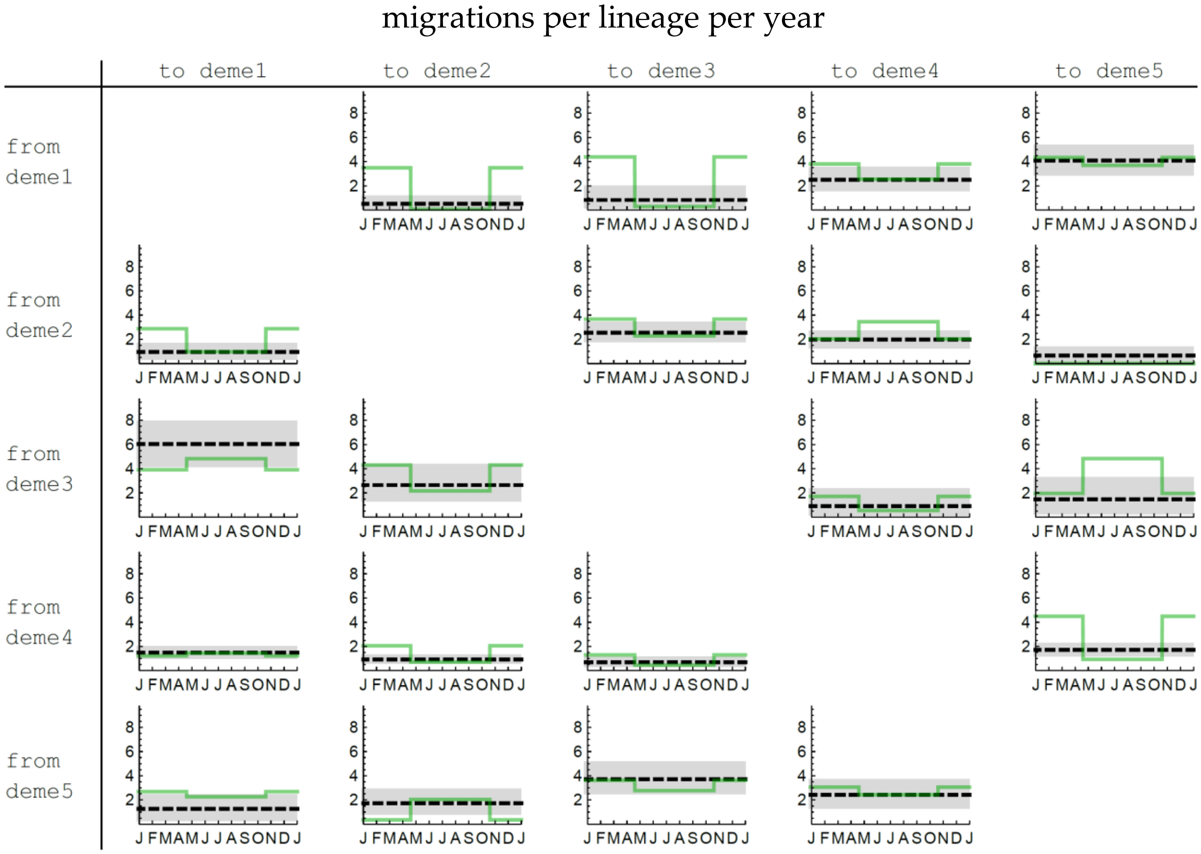


Figure A.4 Inferred migration rates given a two-seasonal tip location generating model Inferred median (dashed) and 90% CI (gray) migration rates between five locations. Tip locations were generated stochastically using an input two-seasonal migration model (green line). A non-seasonal migration rate model was used for inference.

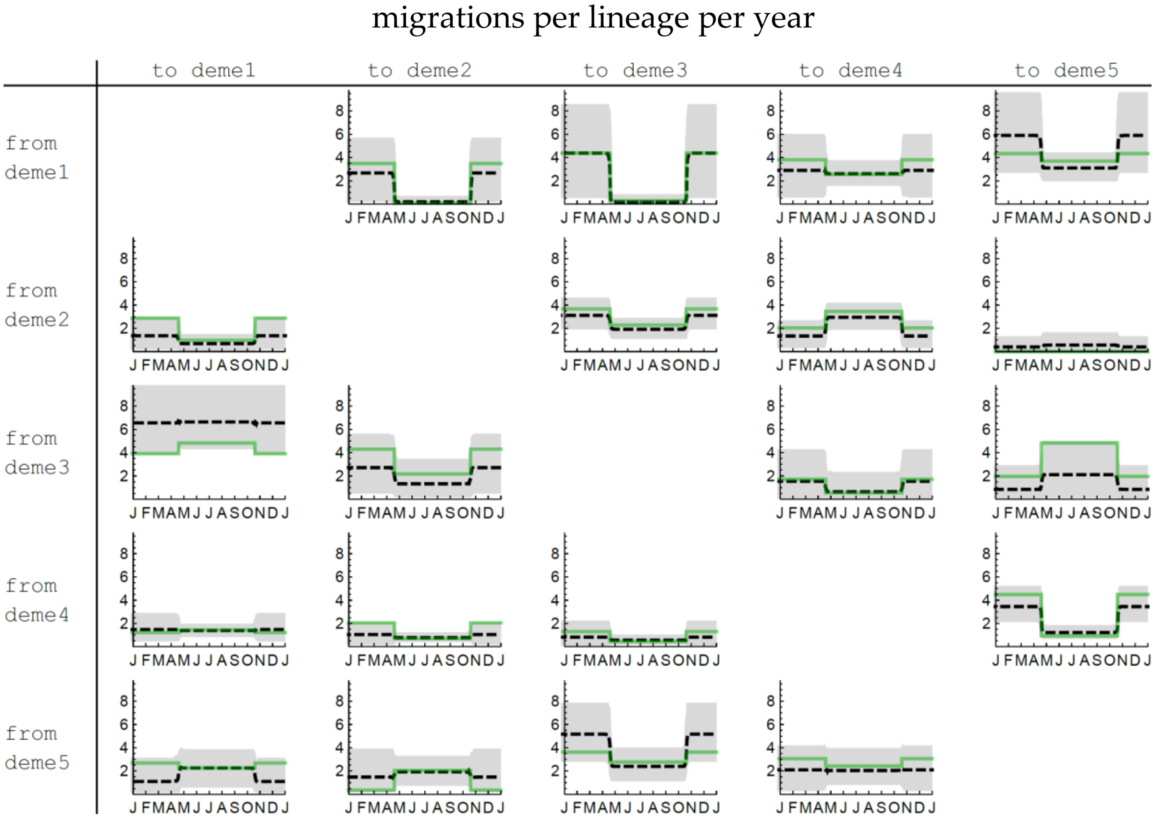


Figure A.5 Inferred migration rates given a two-seasonal tip location generating model Inferred median (dashed) and 90% CI (gray) migration rates between five locations. Tip locations were generated stochastically using an input two-seasonal migration model (green line). A two-seasonal migration rate model was used for inference.

When using the input migration model type (seasonal vs. two-seasonal), migration rates are mostly, but not always within the 90% Bayesian credible intervals.

3.2 Marginal likelihood of non-seasonal and seasonal migration models

We compare the marginal likelihood of alternative non-seasonal and two-seasonal migration models. For this tree and the specified input models (Table 1). The correct migration model (seasonal vs. non-seasonal) is supported for 3, 5 and 8 demes based on the marginal likelihood.

Table A.1. Marginal likelihood of seasonal and non-seasonal migration rate models

# demes	tip generating model	inference model	marginal likelihood
3	non-seasonal	non-seasonal	-2370.4
3	non-seasonal	two-seasonal	-2374.2
3	two-seasonal	non-seasonal	-2491.3
3	two-seasonal	two-seasonal	-2478.9
5	non-seasonal	non-seasonal	-3899.3
5	non-seasonal	two-seasonal	-3914.7
5	two-seasonal	non-seasonal	-4016.4
5	two-seasonal	two-seasonal	-3979.7
8	non-seasonal	non-seasonal	-5055.2
8	non-seasonal	two-seasonal	-5069.6
8	two-seasonal	non-seasonal	-5494.4
8	two-seasonal	two-seasonal	-5430.4

3.3 Inference of seasonal migration from simulated trees

3.3.1 Five populations, single protein tree

We used an agent based simulation (Zinder et al. 2013) to simulate migration between different populations with random population size, associations and seasonal incidence patterns (Table 2, Table 3, Figure 6). In this model the number of infected contacts between *deme i* and *deme j* was drawn from a Poisson distribution with mean:

$$contacts_{i\rightarrow j} = \beta_0 \cdot I_i(t) \cdot \frac{S_j(t)}{N_j} \cdot c_{i\rightarrow j} \cdot s_i(t) \qquad (28)$$

where β_0 is the contact rate, $I_i(t)$ the number of infected at the source deme, $S_j(t)$ the number of susceptible at the destination deme, N_j the population size of the destination deme, $c_{i\rightarrow j}$ is the fraction of contacts between the demes (Table 3) as part of within deme contact, and $s_i(t)$ is the seasonality in the contact rate at the source deme (an alternative could be at the destination deme contact seasonality).

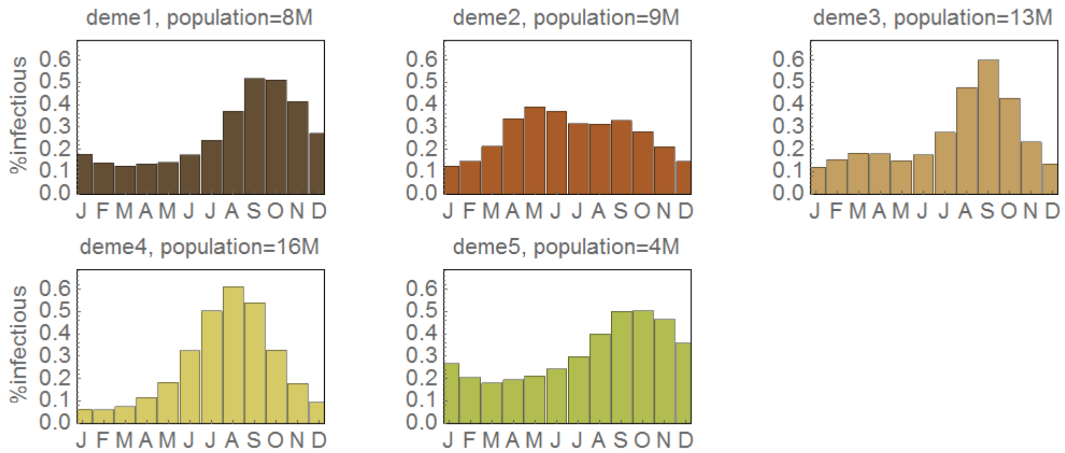
Infection is further determined by the immune history of the host and the cross-immunity with the infecting strain. As such none of these parameters can be directly associated with migration rates on a per lineage basis. A single tree with ~3000 tips tracking the genealogy of the simulated virus was sampled proportional to the prevalence.

Table A.2. Agent based five deme population parameterization

Parameter	Value
contact rate β0	0.6 [1/day]
recovery rate ν	0.2[1/day]
birth/death rate μ	1/30 [1/year]
epitopes	4
variants per epitope	5x4x3x2
epitope mutation rate ξ	0.000008 [1/day]
cross-immunity σ	0.87

Table A.3. Random associations ($c_{i \to j}$) between five simulated populations as a fraction of within deme contact

from/to	deme1	deme2	deme3	deme4	deme5
deme1		0.007	0.039	0.039	0.041
deme2	0.045		0.005	0.005	0.031
deme3	0.014	0.032			0.013
deme4	0.007	0.029	0.033	0.033	0.014
deme5	0.042	0.040	0.011	0.011	


deme1, population=8M

deme2, population=9M

deme3, population=13M

deme4, population=16M

deme5, population=4M

Figure A.6 Prevalence seasonality in simulated population
Simulation of 5 populations with seasonal incidence patterns and random associations.

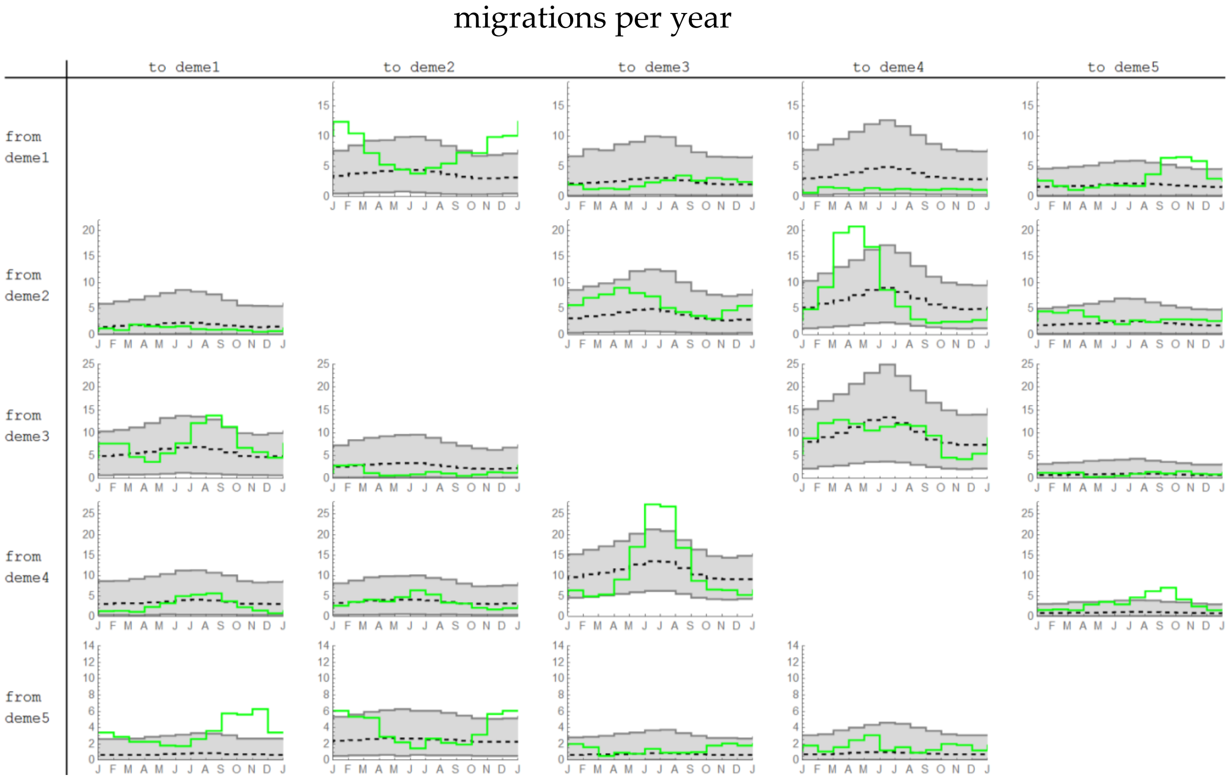


Figure A.7 Inferred migration events seasonality using a constant migration model Inferred median (dashed) and 90% CI (gray) stochastically mapped migrations between five locations. Samples of the migration events on the simulation tree (green line). A model with constant migration rates is used (marginal likelihood = -4593.7).

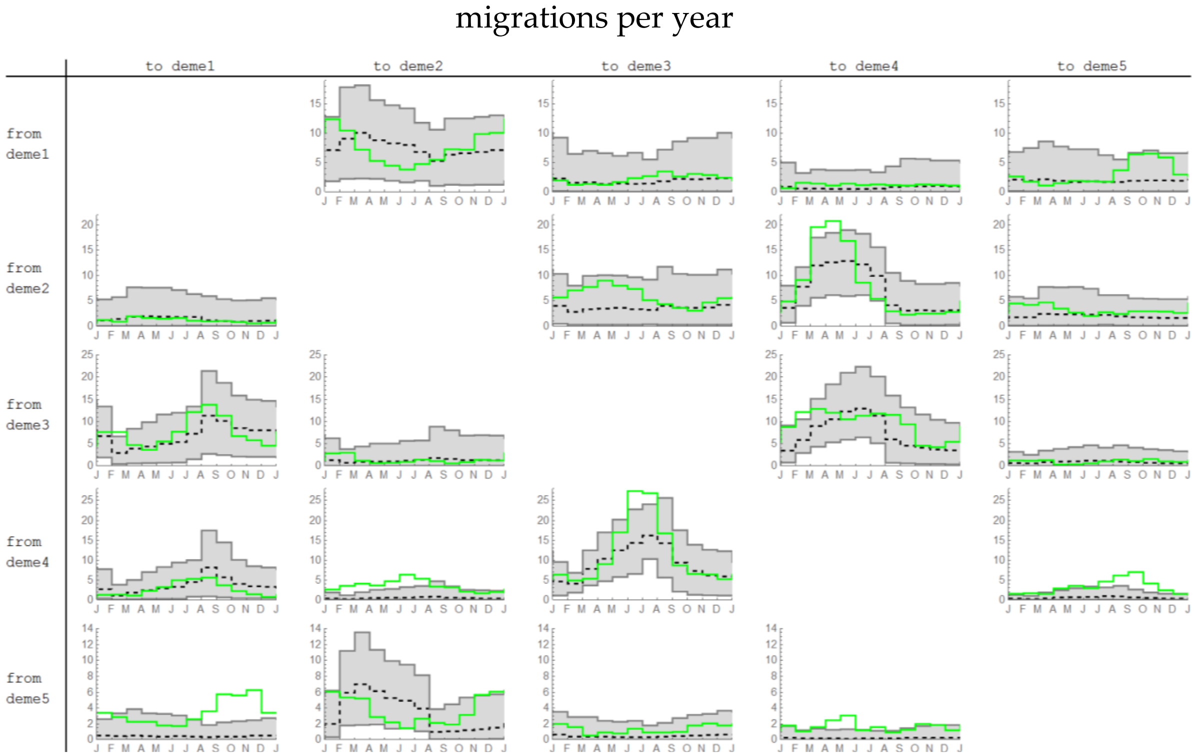


Figure A.8 Inferred migration events seasonality using a two-seasonal migration model Inferred median (dashed) and 90% CI (gray) stochastically mapped migrations between five locations. Samples of the migration events on the simulation tree (green line). A two seasonal migration model is used (marginal likelihood = -4510.2)

3.3.2 Three populations, two protein trees

We simulated (Zinder et al. 2013) migration between three different seasonal populations with a specified population size. Contact seasonality (Figure 9, observed seasonality) and a associations (Table 4) were randomly parameterized. A limited number of tip samples were used to sample the transmission tree as specified by Table 5 intended to approximate the sampling profile of the main text, and Table 6 representing uniform sampling over time. Each simulation was repeated twice to attain to alternative evolutionary histories (proteins) driven by the same migration process. Future simulations will include the direct simulation of segmented genome viral evolution.

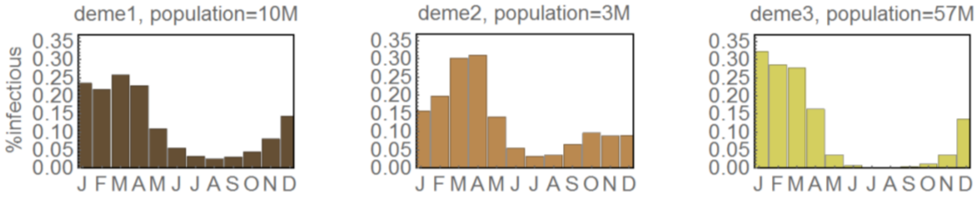


Figure A.9 Prevalence seasonality in simulated population
Simulation of three populations with seasonal incidence patterns and random associations.

Table A.4. Random associations ($c_{i \to j}$) between three simulated populations as a fraction of within deme contact

from/to	deme1	deme2	deme3
deme1		0.007	0.039
deme2	0.045		0.005
deme3	0.014	0.032	

3.3.2.1 Uniform sampling over time

In this simulation, we sampled approximately the same number of tips over time irrespective of population size and of seasonal incidence patterns.

Table A.5. Number of tips samples from two simulated protein trees with alternative evolutionary histories and the same underlying migration processes. Tips were sampled stochastically, with approximately the same number of tips sampled over time and in each population.

	protein A	protein B	Total
deme1	287	278	565
deme2	252	273	525
deme3	290	297	587
Total	829	848	1677

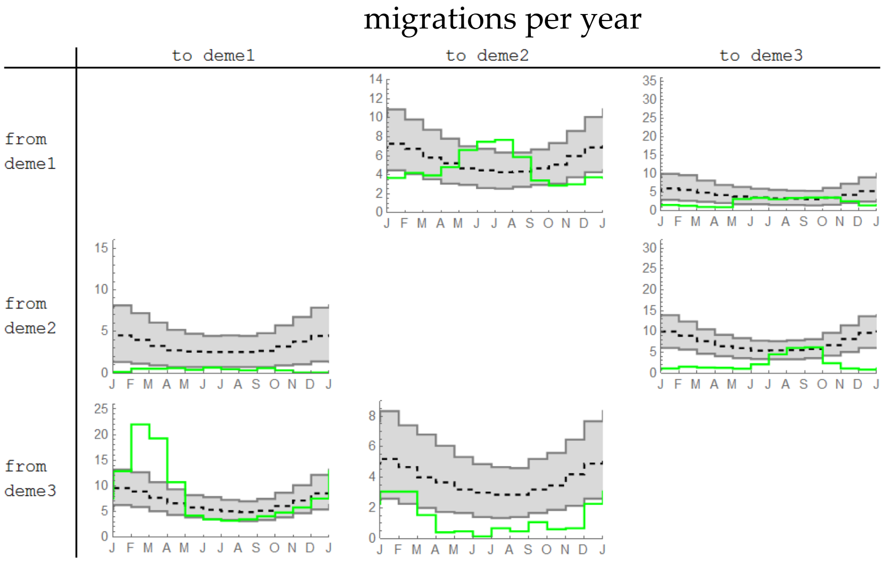


Figure A.10 Inferred migration events seasonality using a two-seasonal migration rate model and uniform tip sampling

Inferred median (dashed) and 90% CI (gray) stochastically mapped migrations between five locations. Samples of the migration events on the simulation tree (green line). A two seasonal migration model is used (marginal likelihood = -906.9)

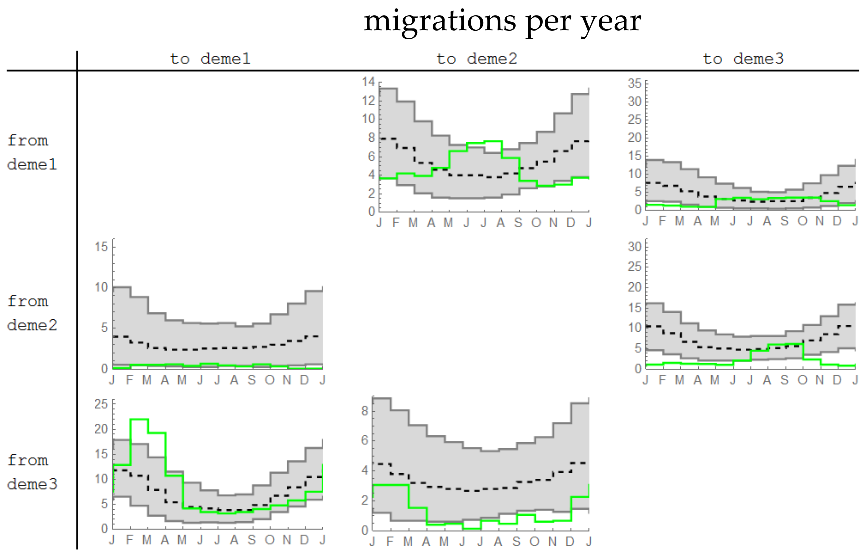


Figure A.11 Inferred migration events seasonality using a two-seasonal migration rate model and uniform tip sampling

Inferred median (dashed) and 90% CI (gray) stochastically mapped migrations between five locations. Samples of the migration events on the simulation tree (green line). A two seasonal migration model is used (marginal likelihood = -906.9)

3.3.2.2 Proportional sampling

In this simulation tips were sampled proportional to incidence. The number of tips (Table 5) sampled is intended to approximate the available hemagglutinin and neuraminidase sequences which were used in the main text for a similar inference.

Table A.6. Number of tips samples from two simulated protein trees

with alternative evolutionary histories and the same underlying migration processes. Tips were sampled stochastically, proportional to prevalence, with an approximate number of tips specified in each population.

	protein A	**protein B**	**Total**
deme1	514	47	561
deme2	138	2	140
deme3	1369	323	1692
Total	2021	372	2393

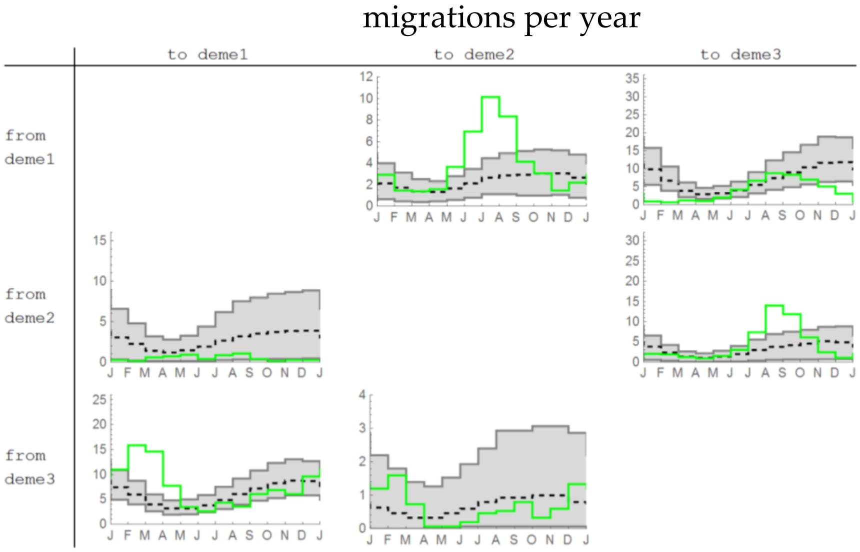


Figure A.12 Inferred migration events seasonality using a constant migration rate model Inferred median (dashed) and 90% CI (gray) stochastically mapped migrations between five locations. Samples of the migration events on the simulation tree (green line). A model with constant migration rates is used (marginal likelihood = -899.3).

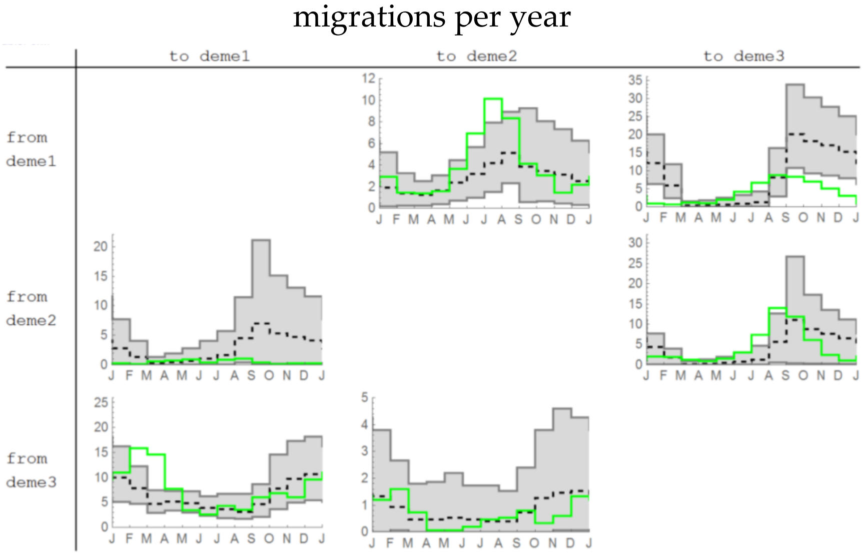


Figure A.13 Inferred migration events seasonality using a two-seasonal migration rate model Inferred median (dashed) and 90% CI (gray) stochastically mapped migrations between five locations. Samples of the migration events on the simulation tree (green line). A two seasonal migration model is used (marginal likelihood = -862.7)

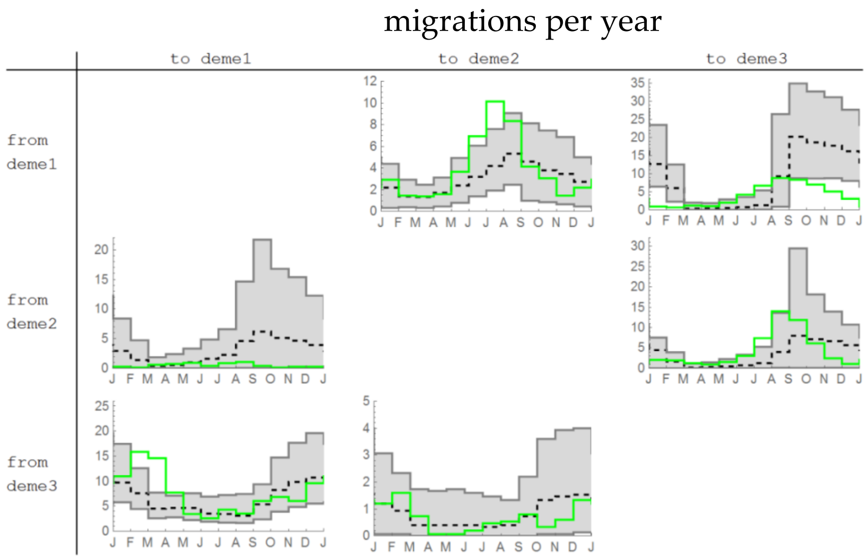


Figure A.14 Inferred migration events seasonality using a two-seasonal migration rate model with variable selection for the inclusion of migration between deme pairs and for the inclusion of seasonal migration between deme pairs Inferred median (dashed) and 90% CI (gray) stochastically mapped migrations between five locations. Samples of the migration events on the simulation tree (green line). A two seasonal migration model is used (marginal likelihood = -866.5)

References

Abrams P. 1983. *The Theory of Limiting Similarity*. Vol. 14. Annual Review of Ecology and Systematics. Palo Alto, CA, ETATS-UNIS: Annual Reviews.

Afrad, Mokibul Hassan, Zahid Hassan, Saiada Farjana, Sayra Moni, Subarna Barua, Sumon Kumar Das, Abu Syed Golam Faruque, Tasnim Azim, and Mustafizur Rahman. 2013. "Changing Profile of Rotavirus Genotypes in Bangladesh, 2006-2012." *BMC Infectious Diseases* 13 (1). BMC Infectious Diseases: 320. doi:10.1186/1471-2334-13-320.

Afrad, Mokibul, Zahid Hassan, Saiada Farjana, Sayra Moni, Subarna Barua, Sumon Das, Abu Faruque, Tasnim Azim, and Mustafizur Rahman. 2013. "Changing Profile of Rotavirus Genotypes in Bangladesh, 2006--2012." *BMC Infectious Diseases* 13 (1): 320.

Alonso, Wladimir J., Cécile Viboud, Lone Simonsen, Eduardo W. Hirano, Luciane Z. Daufenbach, and Mark A. Miller. 2007. "Seasonality of Influenza in Brazil: A Traveling Wave from the Amazon to the Subtropics." *American Journal of Epidemiology* 165 (12): 1434–42.

Andreasen, V., and A. Sasaki. 2006. "Shaping the Phylogenetic Tree of Influenza by Cross-Immunity." *Theoretical Population Biology* 70 (2): 164–73.

Annan, A., H.J. Baldwin, V.M. Corman, S.M. Klose, M. Owusu, E.E. Nkrumah, E.K. Badu, et al. 2013. "Human Betacoronavirus 2c EMC/2012-Related Viruses in Bats, Ghana and Europe." *Emerging Infectious Diseases* 19 (3): 456–59. doi:10.3201/eid1903.121503.

Aoki, Scott T, Ethan C Settembre, Shane D Trask, Harry B Greenberg, Stephen C Harrison, and Philip R Dormitzer. 2009. "Structure of Rotavirus Outer-Layer Protein VP7 Bound with a Neutralizing Fab." *Science (New York, N.Y.)* 324 (5933): 1444–47. doi:10.1126/science.1170481.

Arista, S., G. M. Giammanco, S. De Grazia, S. Ramirez, C. Lo Biundo, C. Colomba, A. Cascio, and V. Martella. 2006. "Heterogeneity and Temporal Dynamics of Evolution of G1 Human Rotaviruses in a Settled Population." *Journal of Virology* 80 (21): 10724–33. doi:10.1128/JVI.00340-06.

Armstrong, Robert A., and Richard McGehee. 1980. "Competitive Exclusion." *The American Naturalist* 115 (2): 151–70. doi:10.2307/2460592.

Artzy-Randrup, Yael, Mary M. Rorick, Karen Day, Donald Chen, Andrew P. Dobson, and Mercedes Pascual. 2012a. "Population Structuring of Multi-Copy, Antigen-Encoding Genes in Plasmodium Falciparum." *eLife* 1. http://elife.elifesciences.org/content/1/e00093.short.

Artzy-Randrup, Yael, Mary M Rorick, Karen Day, Donald Chen, Andrew P Dobson, and Mercedes Pascual. 2012b. “Population Structuring of Multi-Copy, Antigen-Encoding Genes in *Plasmodium Falciparum*.” *eLife* 1 (December). doi:10.7554/eLife.00093.

Ayres, Daniel L, Aaron Darling, Derrick J Zwickl, Peter Beerli, Mark T Holder, Paul O Lewis, John P Huelsenbeck, et al. 2012. “BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics.” *Systematic Biology* 61 (1): 170–73. doi:10.1093/sysbio/syr100.

Bahl, Justin, Martha I. Nelson, Kwok H. Chan, Rubing Chen, Dhanasekaran Vijaykrishna, Rebecca A. Halpin, Timothy B. Stockwell, Xudong Lin, David E. Wentworth, and Elodie Ghedin. 2011. “Temporally Structured Metapopulation Dynamics and Persistence of Influenza A H3N2 Virus in Humans.” *Proceedings of the National Academy of Sciences* 108 (48): 19359–64.

Balcan, Duygu, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J. Ramasco, and Alessandro Vespignani. 2009. “Multiscale Mobility Networks and the Spatial Spreading of Infectious Diseases.” *Proceedings of the National Academy of Sciences* 106 (51): 21484–89. doi:10.1073/pnas.0906910106.

Bandaranayake, Don, Q. Sue Huang, Ange Bissielo, Tim Wood, Graham Mackereth, Michael G. Baker, Richard Beasley, et al. 2010. “Risk Factors and Immunity in a Nationally Representative Population Following the 2009 Influenza A(H1N1) Pandemic.” Edited by Benjamin J. Cowling. *PLoS ONE* 5 (10): e13211. doi:10.1371/journal.pone.0013211.

Baskerville, Edward B., Trevor Bedford, Robert C. Reiner, and Mercedes Pascual. 2013. “Nonparametric Bayesian Grouping Methods for Spatial Time-Series Data.” *arXiv Preprint arXiv:1306.5202*.

Baskerville, Edward B., Andy P. Dobson, Trevor Bedford, Stefano Allesina, T. Michael Anderson, and Mercedes Pascual,. 2011. “Spatial Guilds in the Serengeti Food Web Revealed by a Bayesian Group Model.” *PLoS Comput Biol* 7 (12): e1002321. doi:10.1371/journal.pcbi.1002321.

Bedford, Trevor, Sarah Cobey, Peter Beerli, and Mercedes Pascual. 2010a. “Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2).” *PLoS Pathogens* 6 (5): e1000918. doi:10.1371/journal.ppat.1000918.

Bedford, Trevor, Sarah Cobey, and Mercedes Pascual. 2011a. “Strength and Tempo of Selection Revealed in Viral Gene Genealogies.” *BMC Evolutionary Biology* 11 (1): 220.

Bedford, Trevor, Andrew Rambaut, and Mercedes Pascual. 2012a. “Canalization of the Evolutionary Trajectory of the Human Influenza Virus.” *BMC Biology* 10 (1): 38. doi:10.1186/1741-7007-10-38.

Bedford, Trevor, Marc A. Suchard, Philippe Lemey, Gytis Dudas, Victoria Gregory, Alan J. Hay, John W. McCauley, Colin A. Russell, Derek J. Smith, and Andrew Rambaut. 2014. “Integrating Influenza Antigenic Dynamics with Molecular Evolution.” *Elife* 3: e01914.

Beerli, Peter, and Michal Palczewski. 2010. “Unified Framework to Evaluate Panmixia and Migration Direction Among Multiple Sampling Locations.” *Genetics* 185 (1): 313–26. doi:10.1534/genetics.109.112532.

Beerli, P., and J. Felsenstein. 1999. “Maximum-Likelihood Estimation of Migration Rates and Effective Population Numbers in Two Populations Using a Coalescent Approach.” *Genetics* 152 (2): 763.

Bezeng, Bezeng Simeon, T. Jonathan Davies, Kowiyou Yessoufou, Olivier Maurin, and Michelle Van der Bank. 2015. “Revisiting Darwin’s Naturalization Conundrum: Explaining Invasion Success of Non-native Trees and Shrubs in Southern Africa.” *Journal of Ecology*.

Bhatt, S., A. Katzourakis, and O.G. Pybus. 2010. “Detecting Natural Selection in RNA Virus Populations Using Sequence Summary Statistics.” *Infection, Genetics and Evolution* 10 (3): 421–30.

Bielejec, Filip, Philippe Lemey, Guy Baele, Andrew Rambaut, and Marc A. Suchard. 2014. “Inferring Heterogeneous Evolutionary Processes Through Time: From Sequence Substitution to Phylogeography.” *Systematic Biology* 63 (4): 493–504. doi:10.1093/sysbio/syu015.

Bolker, Benjamin M. 2008. *Ecological Models and Data in R*. Princeton University Press.

Bollback, Jonathan. 2006. “SIMMAP: Stochastic Character Mapping of Discrete Traits on Phylogenies.” *BMC Bioinformatics* 7 (1): 88.

Bollback, Jonathan P. 2006. “SIMMAP: Stochastic Character Mapping of Discrete Traits on Phylogenies.” *BMC Bioinformatics* 7 (January): 88. doi:10.1186/1471-2105-7-88.

Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc a Suchard, Andrew Rambaut, and Alexei J Drummond. 2014. “BEAST 2: A Software Platform for Bayesian Evolutionary Analysis.” *PLoS Computational Biology* 10 (4): e1003537. doi:10.1371/journal.pcbi.1003537.

Bourhy, H., J.-M. Reynes, E. J. Dunham, L. Dacheux, F. Larrous, V. T. Q. Huong, G. Xu, J. Yan, M. E. G. Miranda, and E. C. Holmes. 2008. “The Origin and Phylogeography of Dog Rabies Virus.” *Journal of General Virology* 89 (11): 2673–81. doi:10.1099/vir.0.2008/003913-0.

Bouvier, Nicole M., and Peter Palese. 2008. “The Biology of Influenza Viruses.” *Vaccine* 26 (September): D49–53. doi:10.1016/j.vaccine.2008.07.039.

Buckee, Caroline O., Mario Recker, Eleanor R. Watkins, and Sunetra Gupta. 2011. “Role of Stochastic Processes in Maintaining Discrete Strain Structure in Antigenically Diverse Pathogen Populations.” *Proceedings of the National Academy of Sciences* 108 (37): 15504–9. doi:10.1073/pnas.1102445108.

Buckee, C.O., K.A. Jolley, M. Recker, B. Penman, P. Kriz, S. Gupta, and M.C.J. Maiden. 2008. “Role of Selection in the Emergence of Lineages and the Evolution of Virulence in Neisseria Meningitidis.” *Proceedings of the National Academy of Sciences* 105 (39): 15082.

Bui, Huynh-Hoa, Bjoern Peters, Erika Assarsson, Innocent Mbawuike, and Alessandro Sette. 2007. “Ab and T Cell Epitopes of Influenza A Virus, Knowledge and Opportunities.”

Proceedings of the National Academy of Sciences 104 (1): 246–51. doi:10.1073/pnas.0609330104.

Burnham, Kenneth P, and David R Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

Bush, R. M., C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch. 1999. "Predicting the Evolution of Human Influenza A." *Science* 286: 1921–25. doi:10.1126/science.286.5446.1921.

Bush, R. M., W. M. Fitch, C. A. Bender, and N. J. Cox. 1999. "Positive Selection on the H3 Hemagglutinin Gene of Human Influenza Virus A." *Mol. Biol. Evol.* 16: 1457–65.

Caballero, Armando. 1994. "Developments in the Prediction of Effective Population Size." *Heredity* 73 (6): 657–79.

Calderhead, Ben, and Mark Girolami. 2009. "Estimating Bayes Factors via Thermodynamic Integration and Population MCMC." *Computational Statistics & Data Analysis* 53 (12): 4028–45. doi:10.1016/j.csda.2009.07.025.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (January): 421. doi:10.1186/1471-2105-10-421.

Cannell, J. J., R. Vieth, J. C. Umhau, M. F. Holick, W. B. Grant, S. Madronich, C. F. Garland, and E. Giovannucci. 2006. "Epidemic Influenza and Vitamin D." *Epidemiology and Infection* 134 (06): 1129. doi:10.1017/S0950268806007175.

Catford, Jane A., Roland Jansson, and Christer Nilsson. 2009. "Reducing Redundancy in Invasion Ecology by Integrating Hypotheses into a Single Theoretical Framework." *Diversity and Distributions* 15 (1): 22–40. doi:10.1111/j.1472-4642.2008.00521.x.

C.D.C. 2010. "Estimates of Deaths Associated with Seasonal Influenza—United States, 1976–2007."

CDC. n.d. "Flu Activity, Reports & Surveillance Methods in the United States." http://www.cdc.gov/flu/weekly/fluactivity.htm.

Chesson, P. 2000. "Mechanisms of Maintenance of Species Diversity." *Annual Review of Ecology and Systematics*, 343–66.

Chew, F. T., S. Doraisingham, A. E. Ling, G. Kumarasinghe, B. W. Lee, and others. 1998. "Seasonal Trends of Viral Respiratory Tract Infections in the Tropics." *Epidemiology and Infection* 121 (1): 121–28.

Chi, C.-Y., C.-C. Liu, C.-C. Lin, H.-C. Wang, Y.-T. Cheng, C.-M. Chang, and J.-R. Wang. 2010. "Preexisting Antibody Response against 2009 Pandemic Influenza H1N1 Viruses in the Taiwanese Population." *Clinical and Vaccine Immunology* 17 (12): 1958–62. doi:10.1128/CVI.00212-10.

Chowell, Gerardo, Cécile Viboud, Lone Simonsen, Mark Miller, and Wladimir J. Alonso. 2010. "The Reproduction Number of Seasonal Influenza Epidemics in Brazil, 1996–2006." *Proceedings of the Royal Society B: Biological Sciences*, rspb20091897.

Cobey, Sarah. 2014. "Pathogen Evolution and the Immunological Niche: Evolution of Pathogen Diversity." *Annals of the New York Academy of Sciences* 1320 (1): 1–15. doi:10.1111/nyas.12493.

Cobey, S., and M. Lipsitch. 2012. “Niche and Neutral Effects of Acquired Immunity Permit Coexistence of Pneumococcal Serotypes.” *Science* 335 (6074): 1376–80. doi:10.1126/science.1215947.

Crawford, Sue E, Jacob John, Sudhir Babji, Mary K Estes, D Ph, and Gagandeep Kang. 2011. “Protective Effect of Natural Rotavirus Infection in an Indian Birth Cohort,” 337–46.

Creanza, N., J.S. Schwarz, and J.E. Cohen. 2010. “Intraseasonal Dynamics and Dominant Sequences in H3N2 Influenza.” *PloS One* 5 (1): e8544.

Darwin, Charles, and Encyclopaedia Britannica. 1872. “The Origin of Species by Means of Natural Selection.”

Davis, C. C., H. Schaefer, Z. Xi, D. A. Baum, M. J. Donoghue, and L. J. Harmon. 2014. “Long-Term Morphological Stasis Maintained by a Plant-Pollinator Mutualism.” *Proceedings of the National Academy of Sciences*, April. doi:10.1073/pnas.1403157111.

Dawood, Fatimah S, a Danielle Iuliano, Carrie Reed, Martin I Meltzer, David K Shay, Po-Yung Cheng, Don Bandaranayake, et al. 2012. “Estimated Global Mortality Associated with the First 12 Months of 2009 Pandemic Influenza A H1N1 Virus Circulation: A Modelling Study.” *The Lancet. Infectious Diseases* 12 (9). Elsevier Ltd: 687–95. doi:10.1016/S1473-3099(12)70121-4.

De Grazia, Simona, Floriana Bonura, Claudia Colomba, Antonio Cascio, Francesca Di Bernardo, Antonina Collura, Diane M Terranova, Vito Martella, and Giovanni M Giammanco. 2014. “Data Mining from a 27-Years Rotavirus Surveillance in Palermo, Italy.” *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, March. Elsevier B.V., 1–8. doi:10.1016/j.meegid.2014.03.001.

Dóró, Renáta, Brigitta László, Vito Martella, Eyal Leshem, Jon Gentsch, Umesh Parashar, and Krisztián Bányai. 2014. “Review of Global Rotavirus Strain Prevalence Data from Six Years Post Vaccine Licensure Surveillance: Is There Evidence of Strain Selection from Vaccine Pressure?” *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, no. September (September). doi:10.1016/j.meegid.2014.08.017.

Drummond, Alexei J., Marc A. Suchard, Dong Xie, and Andrew Rambaut. 2012. “Bayesian Phylogenetics with BEAUti and the BEAST 1.7.” *Molecular Biology and Evolution*, February. doi:10.1093/molbev/mss075.

Drummond, Alexei, and Andrew Rambaut. 2007. “BEAST: Bayesian Evolutionary Analysis by Sampling Trees.” *BMC Evolutionary Biology* 7 (1): 214.

Du, Xiangjun, Libo Dong, Yu Lan, Yousong Peng, Aiping Wu, Ye Zhang, Weijuan Huang, et al. 2012. “Mapping of H3N2 Influenza Antigenic Evolution in China Reveals a Strategy for Vaccine Strain Recommendation.” *Nat Commun* 3 (February): 709. doi:10.1038/ncomms1710.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. doi:10.1093/nar/gkh340.

Eggo, Rosalind M., Simon Cauchemez, and Neil M. Ferguson. 2011. "Spatial Dynamics of the 1918 Influenza Pandemic in England, Wales and the United States." *Journal of The Royal Society Interface* 8 (55): 233–43. doi:10.1098/rsif.2010.0216.

Estes, Mary K., and Albert Z. Kapikian. 2013. "Rotaviruses." In *Fields Virology*, edited by Peter Howley and David M Knipe, Sixth.

Felsenstein, J. 1981. "Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach." *J Mol Evol* 17: 368–76.

Felsenstein, J. 1981. "Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach." *Journal of Molecular Evolution* 17 (6): 368–76.

Felsenstein, Joseph. 1985. "Phylogenies from Gene Frequencies: A Statistical Problem." *Systematic Biology* 34 (3): 300–311.

Ferguson, Neil M., Alison P. Galvani, and Robin M. Bush. 2003a. "Ecological and Immunological Determinants of Influenza Evolution." *Nature* 422 (6930): 428–33.

———. 2003b. "Ecological and Immunological Determinants of Influenza Evolution." *Nature* 422 (6930): 428–33. doi:10.1038/nature01509.

Fisher, Ronald A. 1930. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Fitch, W. M., R. M. Bush, C. A. Bender, and N. J. Cox. 1997. "Long Term Trends in the Evolution of H(3) HA1 Human Influenza Type A." *Proc. Natl Acad. Sci. USA* 94: 7712–18. doi:10.1073/pnas.94.15.7712.

Fitch, W. M., J. M. Leiter, X. Q. Li, and P. Palese. 1991. "Positive Darwinian Evolution in Human Influenza A Viruses." *Proc Natl Acad Sci U S A* 88 (10): 4270–74.

Fleming, D. M., and J. G. Ayres. 1988. "Diagnosis and Patterns of Incidence of Influenza, Influenza-like Illness and the Common Cold in General Practice." *The Journal of the Royal College of General Practitioners* 38 (309): 159.

Fonville, J. M., S. H. Wilks, S. L. James, A. Fox, M. Ventresca, M. Aban, L. Xue, et al. 2014. "Antibody Landscapes after Influenza Virus Infection or Vaccination." *Science* 346 (6212): 996–1000.

Friel, Nial, and Anthony N. Pettitt. 2008. "Marginal Likelihood Estimation via Power Posteriors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (3): 589–607.

Furuse, Yuki, Akira Suzuki, and Hitoshi Oshitani. 2010. "Origin of Measles Virus: Divergence from Rinderpest Virus between the 11th and 12th Centuries." *Virology Journal* 7 (1): 52.

Gerhard, W., and RG Webster. 1978. "Antigenic Drift in Influenza A Viruses. I. Selection and Characterization of Antigenic Variants of A/PR/8/34 [HON1] Influenza Virus with Monoclonal Antibodies." *The Journal of Experimental Medicine* 148 (2): 383.

Gladstone, Beryl P., Sasirekha Ramani, Indrani Mukhopadhya, Jayaprakash Muliyil, Rajiv Sarkar, Andrea M. Rehman, Shabbar Jaffar, Miren Iturriza Gomara, James J. Gray, and David WG Brown. 2011a. "Protective Effect of Natural Rotavirus Infection in an Indian Birth Cohort." *New England Journal of Medicine* 365 (4): 337–46.

Gog, J., S. Ballesteros, C. Viboud, L. Simonsen, O. Bjornstad, L. Edelman, and B. T. Grenfell. 2011. "Spatiotemporal Patterns of the 2009 Pandemic Influenza in the United States."

In *Epidemics3–3rd Int. Conf. on Infectious Disease Dynamics 2011, Boston, MA, 30 November–2 December 2011*.

Gog, J. R, and B. T Grenfell. 2002. "Dynamics and Selection of Many-Strain Pathogens." *Proceedings of the National Academy of Sciences of the United States of America* 99 (26): 17209.

Green, K Y, Johnna F Sears, Koki Taniguchi, K Midthun, Y Hoshino, M Gorziglia, K Nishikawa, S Urasawa, A Z Kapikian, and Robert M Chanock. 1988. "Prediction of Human Rotavirus Serotype by Nucleotide Sequence Analysis of the VP7 Protein Gene." *Journal of Virology* 62 (5): 1819–23.

Grenfell, Bryan T., Oliver G. Pybus, Julia R. Gog, James LN Wood, Janet M. Daly, Jenny A. Mumford, and Edward C. Holmes. 2004. "Unifying the Epidemiological and Evolutionary Dynamics of Pathogens." *Science* 303 (5656): 327–32.

Grenfell, B. T. 2004. "Unifying the Epidemiological and Evolutionary Dynamics of Pathogens." *Science* 303 (5656): 327–32. doi:10.1126/science.1090727.

Guerra, Sylvia F.S., Alexandre C. Linhares, Joana D'Arc P. Mascarenhas, Alessilva Oliveira, Maria Cleonice A. Justino, Luana S. Soares, Elza Caroline Müller, et al. 2015. "Rotavirus Strain Surveillance for Three Years Following the Introduction of Rotavirus Vaccine into Belém, Brazil: RV Strain Surveillance in Belem, Brazil." *Journal of Medical Virology*, April, n/a – n/a. doi:10.1002/jmv.24183.

Gupta, S. 1998. "Chaos, Persistence, and Evolution of Strain Structure in Antigenically Diverse Infectious Agents." *Science* 280 (5365): 912–15. doi:10.1126/science.280.5365.912.

Gupta, Sunetra, Martin C.J. Maiden, Ian M. Feavers, Sean Nee, Robert M. May, and Roy M. Anderson. 1996. "The Maintenance of Strain Structure in Populations of Recombining Infectious Agents." *Nat Med* 2 (4): 437–42. doi:10.1038/nm0496-437.

Hancock, K., V. Veguilla, X. Lu, W. Zhong, E.N. Butler, H. Sun, F. Liu, et al. 2009. "Cross-Reactive Antibody Responses to the 2009 Pandemic H1N1 Influenza Virus." *New England Journal of Medicine* 361 (20): 1945–52.

Harper, G. J. 1961. "Airborne Micro-Organisms: Survival Tests with Four Viruses." *J Hyg* 59: 479–86.

Hasegawa, Masami, H Kishino, and T Yano. 1985. "Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA." *Journal of Molecular Evolution*, 160–74.

Hasing, ME, and Gabriel Trueba. 2009. "Rapid Changes in Rotaviral Genotypes in Ecuador." *Journal of Medical Virology* 81: 2109–13. doi:10.1002/jmv.

Hastings, W. 1970. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika* 57: 97–109.

Hay, Alan J. 2001. "The Evolution of Human Influenza Viruses." *Philosophical Transactions of the Royal Society B: Biological Sciences* 356 (1416): 1861–70. doi:10.1098/rstb.2001.0999.

Heiman, Erica M., Sarah M. McDonald, Mario Barro, Zenobia F. Taraporewala, Tamara Bar-Magen, and John T. Patton. 2008. "Group A Human Rotavirus Genomics: Evidence That Gene Constellations Are Influenced by Viral Protein Interactions." *Journal of Virology* 82 (22): 11106–16.

Hensley, S. E., S. R. Das, A. L. Bailey, L. M. Schmidt, H. D. Hickman, A. Jayaraman, K. Viswanathan, et al. 2009. “Hemagglutinin Receptor Binding Avidity Drives Influenza A Virus Antigenic Drift.” *Science* 326 (5953): 734–36. doi:10.1126/science.1178258.

Higham, Nicholas J. 2009. “The Scaling and Squaring Method for the Matrix Exponential Revisited.” *SIAM Review* 51 (4): 747–64.

Holmes, Edward C, and Yong-Zhen Zhang. 2015. “The Evolution and Emergence of Hantaviruses.” *Current Opinion in Virology* 10 (February): 27–33. doi:10.1016/j.coviro.2014.12.007.

Huang, J. W., and J. M. Yang. 2011. “Changed Epitopes Drive the Antigenic Drift for Influenza A (H3N2) Viruses.” *BMC Bioinformatics* 12 (Suppl 1): S31.

Hubbell, Stephen P. 2001. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Vol. 32. Princeton University Press. https://books.google.com/books?hl=en&lr=&id=EIQpFBu84NoC&oi=fnd&pg=PR3&dq=%22a+geographical+context.+I+de%EF%AC%81ne+biodiversity+to+be%22+%22and+abundance+of+species+and+their+causes.%22+%22array+of+ecological+communities,+ranging+from%22+%22in+many+ways,+including+species+richness,+the+degree%22+&ots=cQIPG2OAt1&sig=14hFeEz9U6enVI_sEcOPVTA4WIw.

Huelsenbeck, J. P., R. Nielsen, and J. P. Bollback. 2003. “Stochastic Mapping of Morphological Characters.” *Syst Biol* 52: 131–58.

Huelsenbeck, J. P., and F. Ronquist. 2001. “MrBayes: Bayesian Inference of Phylogenetic Trees.” *Bioinformatics* 17: 754–55.

Hutchinson, G. Evelyn. 1959. “Homage to Santa Rosalia or Why Are There so Many Kinds of Animals?” *American Naturalist*, 145–59.

Ince, W. L., A. Gueye-Mbaye, J. R. Bennink, and J. W. Yewdell. 2013. “Reassortment Complements Spontaneous Mutation in Influenza A Virus NP and M1 Genes To Accelerate Adaptation to a New Host.” *Journal of Virology* 87 (8): 4330–38. doi:10.1128/JVI.02749-12.

Itoh, Yasushi, Kyoko Shinya, Maki Kiso, Tokiko Watanabe, Yoshihiro Sakoda, Masato Hatta, Yukiko Muramoto, et al. 2009. “In Vitro and in Vivo Characterization of New Swine-Origin H1N1 Influenza Viruses.” *Nature* 460 (7258): 1021–25. doi:10.1038/nature08260.

Jeffreys, Harold. 1935. “Some Tests of Significance, Treated by the Theory of Probability.” *Mathematical Proceedings of the Cambridge Philosophical Society* 31 (02): 203–22. doi:10.1017/S030500410001330X.

Jukes, T., and C. Cantor. 1969. “Evolution of Protein Molecules.” *Mammalian Protein Metabolism*, 21–132.

Karpova, L, S, Marinich, I. G, and Krainova, T. I. 2006. “Etiology of influenza according to the data of serological diagnostics in hospital patients in St. Petersburg at the period of 1968-2000.” *Zhurnal Mikrobiologii Epidemiologii i Immunobiologii*, 41–45.

Kass, Robert E., and Adrian E. Raftery. 1995. “Bayes Factors.” *Journal of the American Statistical Association* 90 (430): 773–95. doi:10.1080/01621459.1995.10476572.

Kerkhove, Maria D., Siddhivinayak Hirve, Artemis Koukounari, and Anthony W. Mounts. 2013. “Estimating Age-specific Cumulative Incidence for the 2009 Influenza

Pandemic: A Meta-analysis of A (H1N1) pdm09 Serological Studies from 19 Countries." *Influenza and Other Respiratory Viruses* 7 (5): 872–86.

Kliks, Srisakul C., Amanda Nisalak, Walter E. Brandt, Larry Wahl, and Donald S. Burke. 1989. "Antibody-Dependent Enhancement of Dengue Virus Growth in Human Monocytes as a Risk Factor for Dengue Hemorrhagic Fever." DTIC Document.

Koel, Björn F, David F Burke, Theo M Bestebroer, Stefan van der Vliet, Gerben C M Zondag, Gaby Vervaet, Eugene Skepner, et al. 2013. "Substitutions near the Receptor Binding Site Determine Major Antigenic Change during Influenza Virus Evolution." *Science (New York, N.Y.)* 342 (6161): 976–79. doi:10.1126/science.1244730.

Koelle, K., S. Cobey, B. Grenfell, and M. Pascual. 2006. "Epochal Evolution Shapes the Phylodynamics of Interpandemic Influenza A (H3N2) in Humans." *Science* 314 (5807): 1898–1903. doi:10.1126/science.1132745.

Kryazhimskiy, Sergey, Ulf Dieckmann, Simon A. Levin, and Jonathan Dushoff. 2007. "On State-Space Reduction in Multi-Strain Pathogen Models, with an Application to Antigenic Drift in Influenza A." *PLoS Computational Biology* 3: e159. doi:10.1371/journal.pcbi.0030159.

Kuiken, T. 2006. "Host Species Barriers to Influenza Virus Infections." *Science* 312 (5772): 394–97. doi:10.1126/science.1122818.

Kuo, Lynn, and Bani Mallick. 1998. "Variable Selection for Regression Models." *Sankhyā: The Indian Journal of Statistics, Series B*, 65–81.

Lartillot, Nicolas, and Hervé Philippe. 2006. "Computing Bayes Factors Using Thermodynamic Integration." *Systematic Biology* 55 (2): 195–207. doi:10.1080/10635150500433722.

Lebrija-Trejos, Edwin, Eduardo A. Pérez-García, Jorge A. Meave, Frans Bongers, and Lourens Poorter. 2010. "Functional Traits and Environmental Filtering Drive Community Assembly in a Species-Rich Tropical System." *Ecology* 91 (2): 386–98.

Leibold, Matthew A. 1995. "The Niche Concept Revisited: Mechanistic Models and Community Context." *Ecology* 76 (5): 1371–82.

Lemey, Philippe, Andrew Rambaut, Trevor Bedford, Nuno Faria, Filip Bielejec, Guy Baele, Colin A. Russell, Derek J. Smith, Oliver G. Pybus, and Dirk Brockmann. 2014. "Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2." *PLoS Pathogens* 10 (2): e1003932.

Lemey, P., A. Rambaut, A. J Drummond, and M. A Suchard. 2009. "Bayesian Phylogeography Finds Its Roots." *PLoS Computational Biology* 5 (9): e1000520.

Leshem, Eyal, Ben Lopman, Roger Glass, Jon Gentsch, Krisztián Bányai, Umesh Parashar, and Manish Patel. 2014a. "Distribution of Rotavirus Strains and Strain-Specific Effectiveness of the Rotavirus Vaccine after Its Introduction: A Systematic Review and Meta-Analysis." *The Lancet Infectious Diseases* 14 (9): 847–56.

Lindstrom, S. E, Y. Hiromoto, H. Nishimura, T. Saito, R. Nerome, and K. Nerome. 1999. "Comparative Analysis of Evolutionary Mechanisms of the Hemagglutinin and Three Internal Protein Genes of Influenza B Virus: Multiple Cocirculating Lineages and Frequent Reassortment of the NP, M, and NS Genes." *Journal of Virology* 73 (5): 4413.

Lin, Jianhua. 1991. “Divergence Measures Based on the Shannon Entropy.” *Information Theory, IEEE Transactions on* 37 (1): 145–51.

Lopman, Ben A, Daniel C Payne, Jacqueline E Tate, Manish M Patel, Margaret M Cortese, and Umesh D Parashar. 2012. “Post-Licensure Experience with Rotavirus Vaccination in High and Middle Income Countries; 2006 to 2011.” *Current Opinion in Virology* 2 (4). Elsevier B.V.: 434–42. doi:10.1016/j.coviro.2012.05.002.

Lopman, Benjamin A., Virginia E. Pitzer, Rajiv Sarkar, Beryl Gladstone, Manish Patel, John Glasser, Manoj Gambhir, et al. 2012. “Understanding Reduced Rotavirus Vaccine Efficacy in Low Socio-Economic Settings.” Edited by Cécile Viboud. *PloS One* 7 (8). Public Library of Science: e41720. doi:10.1371/journal.pone.0041720.

Lu, Lu, Samantha J. Lycett, and Andrew J. Leigh Brown. 2014. “Reassortment Patterns of Avian Influenza Virus Internal Segments among Different Subtypes.” *BMC Evolutionary Biology* 14 (1): 16.

MacArthur, Robert, and Richard Levins. 1967. “The Limiting Similarity, Convergence, and Divergence of Coexisting Species.” *American Naturalist*, 377–85.

Maes, Piet, Jelle Matthijnssens, Mustafizur Rahman, and Marc Van Ranst. 2009. “RotaC: A Web-Based Tool for the Complete Genome Classification of Group A Rotaviruses.” *BMC Microbiology* 9 (January): 238. doi:10.1186/1471-2180-9-238.

Matthijnssens, J., M. Ciarlet, E. Heiman, I. Arijs, T. Delbeke, S. M. McDonald, E. A. Palombo, et al. 2008. “Full Genome-Based Classification of Rotaviruses Reveals a Common Origin between Human Wa-Like and Porcine Rotavirus Strains and Human DS-1-Like and Bovine Rotavirus Strains.” *Journal of Virology* 82 (7): 3204–19. doi:10.1128/JVI.02257-07.

Matthijnssens, Jelle, Max Ciarlet, Erica Heiman, Ingrid Arijs, Thomas Delbeke, Sarah M McDonald, Enzo A Palombo, et al. 2008. “Full Genome-Based Classification of Rotaviruses Reveals a Common Origin between Human Wa-Like and Porcine Rotavirus Strains and Human DS-1-like and Bovine Rotavirus Strains.” *Journal of Virology* 82 (7): 3204–19. doi:10.1128/JVI.02257-07.

Matthijnssens, Jelle, and Marc Van Ranst. 2012. “Genotype Constellation and Evolution of Group A Rotaviruses Infecting Humans.” *Current Opinion in Virology* 2 (4): 426–33.

May, Robert M., and Martin A. Nowak. 1994. “Superinfection, Metapopulation Dynamics, and the Evolution of Diversity.” *Journal of Theoretical Biology* 170 (1): 95–114. doi:10.1006/jtbi.1994.1171.

McDonald, John H., and Martin Kreitman. 1991. “Adaptive Protein Evolution at the Adh Locus in Drosophila.” *Nature* 351 (6328): 652–54. doi:10.1038/351652a0.

McDonald, Sarah M., Kristin Davis, John K. McAllen, David J. Spiro, and John T. Patton. 2011. “Intra-Genotypic Diversity of Archival G4P [8] Human Rotaviruses from Washington, DC.” *Infection, Genetics and Evolution* 11 (7): 1586–94.

McDonald, Sarah M., Jelle Matthijnssens, John K. McAllen, Erin Hine, Larry Overton, Shiliang Wang, Philippe Lemey, et al. 2009. “Evolutionary Dynamics of Human Rotaviruses: Balancing Reassortment with Preferred Genome Constellations.” Edited by Colin Parrish. *PLoS Pathogens* 5 (10): e1000634. doi:10.1371/journal.ppat.1000634.

McDonald, Sarah M., Allison O. McKell, Christine M. Rippinger, John K. McAllen, Asmik Akopov, Ewen F. Kirkness, Daniel C. Payne, Kathryn M. Edwards, James D. Chappell, and John T. Patton. 2012. "Diversity and Relationships of Cocirculating Modern Human Rotaviruses Revealed Using Large-Scale Comparative Genomics." *Journal of Virology* 86 (17): 9148–62. doi:10.1128/JVI.01105-12.

Messina, Jane P., Oliver J. Brady, Thomas W. Scott, Chenting Zou, David M. Pigott, Kirsten A. Duda, Samir Bhatt, Leah Katzelnick, Rosalind E. Howes, and Katherine E. Battle. 2014. "Global Spread of Dengue Virus Types: Mapping the 70 Year History." *Trends in Microbiology* 22 (3): 138–46.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1953. "Equations of State Calculations by Fast Computing Machines." *Journal of Chemistry and Physics* 21: 1087–92.

Minayev, Pavlo, and Neil Ferguson. 2009. "Improving the Realism of Deterministic Multi-Strain Models: Implications for Modelling Influenza A." *Journal of The Royal Society Interface* 6 (35): 509–18. doi:10.1098/rsif.2008.0333.

Minin, Vladimir N., and Marc A. Suchard. 2008. "Fast, Accurate and Simulation-Free Stochastic Mapping." *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1512): 3985–95.

Monto, AS, and KM Sullivan. 1993. "Acute Respiratory Illness in the Community. Frequency of Illness and the Agents Involved." *Epidemiology and Infection* 110 (01): 145–60.

Muller, Hermann Joseph. 1932. "Some Genetic Aspects of Sex." *American Naturalist*, 118–38.

Nair, Harish, W Abdullah Brooks, Mark Katz, Anna Roca, James a Berkley, Shabir a Madhi, James Mark Simmerman, et al. 2011. "Global Burden of Respiratory Infections due to Seasonal Influenza in Young Children: A Systematic Review and Meta-Analysis." *Lancet* 378 (9807). Elsevier Ltd: 1917–30. doi:10.1016/S0140-6736(11)61051-9.

Nakajima, K., E. Nobusawa, A. Nagy, and S. Nakajima. 2005. "Accumulation of Amino Acid Substitutions Promotes Irreversible Structural Changes in the Hemagglutinin of Human Influenza AH3 Virus during Evolution." *J Virol* 79 (10): 6472–77.

Nelson, E Anthony S, and Roger I Glass. 2010. "Rotavirus: Realising the Potential of a Promising Vaccine." *Lancet* 376 (9741): 568–70. doi:10.1016/S0140-6736(10)60896-3.

Nelson, Martha I., and Edward C. Holmes. 2007. "The Evolution of Epidemic Influenza." *Nature Reviews Genetics* 8 (3): 196–205.

Nelson, Martha I., Lone Simonsen, Cecile Viboud, Mark A. Miller, and Edward C. Holmes. 2007. "Phylogenetic Analysis Reveals the Global Migration of Seasonal Influenza A Viruses." *PLoS Pathogens* 3 (9): e131. doi:10.1371/journal.ppat.0030131.

Nelson, Martha I., Cécile Viboud, Lone Simonsen, Ryan T. Bennett, Sara B. Griesemer, Kirsten St. George, Jill Taylor, et al. 2008. "Multiple Reassortment Events in the Evolutionary History of H1N1 Influenza A Virus Since 1918." Edited by Yoshihiro Kawaoka. *PLoS Pathogens* 4 (2): e1000012. doi:10.1371/journal.ppat.1000012.

Nelson, M. I., L. Simonsen, C. Viboud, M. A. Miller, J. Taylor, K. St. George, S. B. Griesemer, et al. 2006. "Stochastic Processes Are Key Determinants of the Short-Term Evolution of Influenza A Virus." *PLoS Path* in press.

Neverov, Alexey D., Ksenia V. Lezhnina, Alexey S. Kondrashov, and Georgii A. Bazykin. 2014. “Intrasubtype Reassortments Cause Adaptive Amino Acid Replacements in H3N2 Influenza Genes.” *PLoS Genetics* 10 (1): e1004037.

Nielsen, Rasmus. 2002. “Mapping Mutations on Phylogenies.” *Systematic Biology* 51 (5): 729–39. doi:10.1080/10635150290102393.

Nosil, Patrik, Luke J. Harmon, and Ole Seehausen. 2009. “Ecological Explanations for (incomplete) Speciation.” *Trends in Ecology & Evolution* 24 (3): 145–56. doi:10.1016/j.tree.2008.10.011.

Nowak, M.A. 1992. “What Is a Quasispecies?” *Trends in Ecology & Evolution* 7 (4): 118–21.

Nowak, Martin A., and Robert M. May. 1994. “Superinfection and the Evolution of Parasite Virulence.” *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255 (1342): 81–89. doi:10.1098/rspb.1994.0012.

Offit, P a. 1996. “Host Factors Associated with Protection against Rotavirus Disease: The Skies Are Clearing.” *The Journal of Infectious Diseases* 174 Suppl (September): S59–64.

O’Hara, Robert B., and Mikko J. Sillanpää. 2009. “A Review of Bayesian Variable Selection Methods: What, How and Which.” *Bayesian Analysis* 4 (1): 85–117.

Osterhaus, ADME, G. F. Rimmelzwaan, B. E. E. Martina, T. M. Bestebroer, and R. A. M. Fouchier. 2000. “Influenza B Virus in Seals.” *Science* 288 (5468): 1051–53.

Pagel, Mark, Andrew Meade, and Daniel Barker. 2004. “Bayesian Estimation of Ancestral Character States on Phylogenies.” *Systematic Biology* 53 (5): 673–84. doi:10.1080/10635150490522232.

Parrish, Colin R., Pablo R. Murcia, and Edward C. Holmes. 2015. “Influenza Virus Reservoirs and Intermediate Hosts: Dogs, Horses, and New Possibilities for Influenza Virus Exposure of Humans: FIG 1.” Edited by S. Schultz-Cherry. *Journal of Virology* 89 (6): 2990–94. doi:10.1128/JVI.03146-14.

Pitzer, V. E., C. Viboud, L. Simonsen, C. Steiner, C. A. Panozzo, W. J. Alonso, M. A. Miller, et al. 2009. “Demographic Variability, Vaccination, and the Spatiotemporal Dynamics of Rotavirus Epidemics.” *Science* 325 (5938): 290–94. doi:10.1126/science.1172330.

Price, David A., Philip JR Goulder, Paul Klenerman, Andrew K. Sewell, Philippa J. Easterbrook, Maxine Troop, Charles RM Bangham, and Rodney E. Phillips. 1997. “Positive Selection of HIV-1 Cytotoxic T Lymphocyte Escape Variants during Primary Infection.” *Proceedings of the National Academy of Sciences* 94 (5): 1890–95.

Rambaut, A., and A. J. Drummond. 2003. “Tracer [computer Program].” http://beast.bio.ed.ac.uk/tracer.

Rambaut, Andrew, Oliver G. Pybus, Martha I. Nelson, Cecile Viboud, Jeffery K. Taubenberger, and Edward C. Holmes. 2008a. “The Genomic and Epidemiological Dynamics of Human Influenza A Virus.” *Nature* 453 (7195): 615–19. doi:10.1038/nature06945.

Rambaut, Andrew, Oliver G Pybus, Martha I Nelson, Cecile Viboud, Jeffery K Taubenberger, and Edward C Holmes. 2008b. “The Genomic and Epidemiological Dynamics of Human Influenza A Virus.” *Nature* 453 (7195): 615–19. doi:10.1038/nature06945.

Ratmann, Oliver, Gé Donker, Adam Meijer, Christophe Fraser, and Katia Koelle. 2012. "Phylodynamic Inference and Model Assessment with Approximate Bayesian Computation: Influenza as a Case Study." Edited by Sergei L. Kosakovsky Pond. *PLoS Computational Biology* 8 (12): e1002835. doi:10.1371/journal.pcbi.1002835.

Recker, M., O. G Pybus, S. Nee, and S. Gupta. 2007. "The Generation of Influenza Outbreaks by a Network of Host Immune Responses against a Limited Set of Antigenic Types." *Proceedings of the National Academy of Sciences* 104 (18): 7711.

Reichert, Tom, Gerardo Chowell, Hiroshi Nishiura, Ronald Christensen, and Jonathan McCullers. 2010. "Does Glycosylation as a Modifier of Original Antigenic Sin Explain the Case Age Distribution and Unusual Toxicity in Pandemic Novel H1N1 Influenza?" *BMC Infectious Diseases* 10 (1): 5.

Reusken, Chantal BEM, Bart L Haagmans, Marcel A Müller, Carlos Gutierrez, Gert-Jan Godeke, Benjamin Meyer, Doreen Muth, et al. 2013. "Middle East Respiratory Syndrome Coronavirus Neutralising Serum Antibodies in Dromedary Camels: A Comparative Serological Study." *The Lancet Infectious Diseases* 13 (10): 859–66. doi:10.1016/S1473-3099(13)70164-6.

Rha, Brian, Jacqueline E Tate, Daniel C Payne, Margaret M Cortese, Benjamin a Lopman, Aaron T Curns, and Umesh D Parashar. 2014. "Effectiveness and Impact of Rotavirus Vaccines in the United States - 2006-2012." *Expert Review of Vaccines* 13 (3): 365–76. doi:10.1586/14760584.2014.877846.

Rota, Paul A., Kevin Brown, Annette Mankertz, Sabine Santibanez, Sergey Shulga, Claude P. Muller, Judith M. Hübschen, et al. 2011. "Global Distribution of Measles Genotypes and Measles Molecular Epidemiology." *Journal of Infectious Diseases* 204 (suppl 1): S514–23. doi:10.1093/infdis/jir118.

Russell, Colin A., Terry C. Jones, Ian G. Barr, Nancy J. Cox, Rebecca J. Garten, Vicky Gregory, Ian D. Gust, et al. 2008. "The Global Circulation of Seasonal Influenza A (H3N2) Viruses." *Science* 320 (5874): 340–46. doi:10.1126/science.1154137.

Salemi, Marco, Tulio de Oliveira, Massimo Ciccozzi, Giovanni Rezza, and Maureen M. Goodenow. 2008. "High-Resolution Molecular Epidemiology and Evolutionary History of HIV-1 Subtypes in Albania." Edited by Jason Stajich. *PLoS ONE* 3 (1): e1390. doi:10.1371/journal.pone.0001390.

šAltytė Benth, J., and D. Hofoss. 2008. "Modelling and Prediction of Weekly Incidence of Influenza A Specimens in England and Wales." *Epidemiology and Infection* 136 (12): 1658. doi:10.1017/S0950268808000307.

Sanderson, Michael J., Michelle M. McMahon, and Mike Steel. 2011. "Terraces in Phylogenetic Tree Space." *Science* 333 (6041): 448–50. doi:10.1126/science.1206357.

Sanjuan, R., A. Moya, and S. F. Elena. 2004. "The Contribution of Epistasis to the Architecture of Fitness in an RNA Virus." *Proc Natl Acad Sci U S A* 101 (43): 15376–79.

Santos, Norma, and Yasutaka Hoshino. 2005a. "Global Distribution of Rotavirus Serotypes/genotypes and Its Implication for the Development and Implementation of an Effective Rotavirus Vaccine." *Reviews in Medical Virology* 15 (1): 29–56.

———. 2005b. “Global Distribution of Rotavirus Serotypes/genotypes and Its Implication for the Development and Implementation of an Effective Rotavirus Vaccine.” *Reviews in Medical Virology*, no. October 2004: 29–56.

Schaefer, Hanno, Olivier J. Hardy, Luís Silva, Timothy G. Barraclough, and Vincent Savolainen. 2011. “Testing Darwin’s Naturalization Hypothesis in the Azores: Predicting Invasiveness in the Azores.” *Ecology Letters* 14 (4): 389–96. doi:10.1111/j.1461-0248.2011.01600.x.

Scheffer, Marten, and Egbert H. van Nes. 2006. “Self-Organized Similarity, the Evolutionary Emergence of Groups of Similar Species.” *Proceedings of the National Academy of Sciences* 103 (16): 6230–35.

Schierup, Mikkel H., Carl H. Mordhorst, Claude P. Muller, and Laurids S. Christensen. 2005. “Evidence of Recombination among Early-Vaccination Era Measles Virus Strains.” *BMC Evolutionary Biology* 5 (1): 52.

Shaman, J., V. E. Pitzer, C. Viboud, B. T. Grenfell, and M. Lipsitch. 2010. “Absolute Humidity and the Seasonal Onset of Influenza in the Continental United.”

Shapiro, Beth, Andrew Rambaut, and Alexei J. Drummond. 2006. “Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences.” *Molecular Biology and Evolution* 23 (1): 7–9.

Simonsen, Lone. 1999. “The Global Impact of Influenza on Morbidity and Mortality.” *Vaccine* 17, Supplement 1 (0): S3–10. doi:10.1016/S0264-410X(99)00099-7.

Simonsen, Lone, Thomas A Reichert, and Mark A Miller. 2004. “The Virtues of Antigenic Sin: Consequences of Pandemic Recycling on Influenza-Associated Mortality.” *Options for the Control of Influenza V. Proceedings of the International Conference on Options for the Control of Influenza V* 1263 (0): 791–94. doi:10.1016/j.ics.2004.01.029.

Simonsen, Lone, Cécile Viboud, Bryan T. Grenfell, Jonathan Dushoff, Lance Jennings, Marita Smit, Catherine Macken, Mami Hata, Julia Gog, and Mark A. Miller. 2007. “The Genesis and Spread of Reassortment Human Influenza A/H3N2 Viruses Conferring Adamantane Resistance.” *Molecular Biology and Evolution* 24 (8): 1811–20.

Smith, D. J. 2004. “Mapping the Antigenic and Genetic Evolution of Influenza Virus.” *Science* 305 (July): 371–76. doi:10.1126/science.1097211.

Smith, Gavin JD, Justin Bahl, Dhanasekaran Vijaykrishna, Jinxia Zhang, Leo LM Poon, Honglin Chen, Robert G. Webster, JS Malik Peiris, and Yi Guan. 2009. “Dating the Emergence of Pandemic Influenza Viruses.” *Proceedings of the National Academy of Sciences* 106 (28): 11709–12.

Smith, Gavin J. D., Dhanasekaran Vijaykrishna, Justin Bahl, Samantha J. Lycett, Michael Worobey, Oliver G. Pybus, Siu Kit Ma, et al. 2009. “Origins and Evolutionary Genomics of the 2009 Swine-Origin H1N1 Influenza A Epidemic.” *Nature* 459 (7250): 1122–25. doi:10.1038/nature08182.

Stack, J. C., J. D. Welch, M. J. Ferrari, B. U. Shapiro, and B. T. Grenfell. 2010. “Protocols for Sampling Viral Sequences to Study Epidemic Dynamics.” *Journal of The Royal Society Interface* 7 (48): 1119–27. doi:10.1098/rsif.2009.0530.

Stadler, Tanja. 2009. "On Incomplete Sampling under Birth–death Models and Connections to the Sampling-Based Coalescent." *Journal of Theoretical Biology* 261 (1): 58–66. doi:10.1016/j.jtbi.2009.07.018.

Stadler, T., and S. Bonhoeffer. 2013. "Uncovering Epidemiological Dynamics in Heterogeneous Host Populations Using Phylogenetic Methods." *Philosophical Transactions of the Royal Society B: Biological Sciences* 368 (1614): 20120198–20120198. doi:10.1098/rstb.2012.0198.

Strong, Donald R. 1983. "Natural Variability and the Manifold Mechanisms of Ecological Communities." *American Naturalist*, 636–60.

Suchard, Marc A., and Andrew Rambaut. 2009. "Many-Core Algorithms for Statistical Phylogenetics." *Bioinformatics* 25 (11): 1370–76. doi:10.1093/bioinformatics/btp244.

Sumi, Ayako, Ken-ichi Kamo, Norio Ohtomo, Keiji Mise, and Nobumichi Kobayashi. 2011. "Time Series Analysis of Incidence Data of Influenza in Japan." *Journal of Epidemiology* 21 (1): 21–29. doi:10.2188/jea.JE20090162.

Suzuki, Y. 2008. "Positive Selection Operates Continuously on Hemagglutinin during Evolution of H3N2 Human Influenza A Virus." *Gene* 427 (1-2): 111–16.

Swofford, David L. 2003. "{PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4.}."

Tamerius, James D., Jeffrey Shaman, Wladmir J. Alonso, Kimberly Bloom-Feshbach, Christopher K. Uejio, Andrew Comrie, and Cécile Viboud. 2013. "Environmental Predictors of Seasonal Influenza Epidemics across Temperate and Tropical Climates." *PLoS Pathogens* 9 (3): e1003194.

Tamura, Koichiro, Daniel Peterson, Nicholas Peterson, Glen Stecher, Masatoshi Nei, and Sudhir Kumar. 2011. "MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods." *Molecular Biology and Evolution* 28 (10): 2731–39. doi:10.1093/molbev/msr121.

Tate, Jacqueline E, Cynthia Boschi-pinto, A Duncan Steele, Anthony H Burton, Umesh D Parashar, Jazmin Duque, and Who-coordinated Global. 2012. "2008 Estimate of Worldwide Rotavirus-Associated Mortality in Children Younger than 5 Years before the Introduction of Universal Rotavirus Vaccination Programmes: A Systematic Review and Meta-Analysis." *The Lancet Infectious Diseases* 12 (2). Elsevier Ltd: 136–41. doi:10.1016/S1473-3099(11)70253-5.

Thomson, Michael M., Lucía Pérez-Álvarez, and Rafael Nájera. 2002. "Molecular Epidemiology of HIV-1 Genetic Forms and Its Significance for Vaccine Development and Therapy." *The Lancet Infectious Diseases* 2 (8): 461–71.

Tria, Francesca, Michael Laessig, Luca Peliti, and Silvio Franz. 2005. "A Minimal Stochastic Model for Influenza Evolution." http://arxiv.org/abs/q-bio/0505035.pdf.

Tsuchiya, E., K. Sugawara, S. Hongo, Y. Matsuzaki, Y. Muraki, Z. N. Li, and K. Nakamura. 2001. "Antigenic Structure of the Haemagglutinin of Human Influenza A/H2N2 Virus." *Journal of General Virology* 82 (10): 2475–84.

Velazquez, F. Raul, JUAN J. CALVA, M. Lourdes Guerrero, DAVID MASS, ROGER I. GLASS, LARRY K. PICKERING, and GUILLERMO M. RUIZ-PALACIOS. 1993. "Cohort Study of Rotavirus Serotype Patterns in Symptomatic and Asymptomatic

Infections in Mexican Children." *The Pediatric Infectious Disease Journal* 12 (1): 54–61.

Velázquez, F. Raúl, David O. Matson, Juan J. Calva, M. Lourdes Guerrero, Ardythe L. Morrow, Shelly Carter-Campbell, Roger I. Glass, Mary K. Estes, Larry K. Pickering, and Guillermo M. Ruiz-Palacios. 1996. "Rotavirus Infection in Infants as Protection against Subsequent Infections." *New England Journal of Medicine* 335 (14): 1022–28. doi:10.1056/NEJM199610033351404.

Viboud, C, PY Boëlle, K Pakdaman, F Carrat, AJ Valleron, and A Flahault. 2004. "Influenza Epidemics in the United States, France, and Australia, 1972-1997." *Emerging Infectious Diseases* 10 (1): 32–39.

Viboud, Cécile, Wladimir J. Alonso, and Lone Simonsen. 2006. "Influenza in Tropical Regions." *PLoS Medicine* 3 (4): e89. doi:10.1371/journal.pmed.0030089.

Viboud, Cécile, Ottar N. Bjørnstad, David L. Smith, Lone Simonsen, Mark A. Miller, and Bryan T. Grenfell. 2006. "Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza." *Science* 312 (5772): 447–51.

Viboud, Cécile, Martha I. Nelson, Yi Tan, and Edward C. Holmes. 2013. "Contrasting the Epidemiological and Evolutionary Dynamics of Influenza Spatial Transmission." *Philosophical Transactions of the Royal Society B: Biological Sciences* 368 (1614). doi:10.1098/rstb.2012.0199.

Viboud, Cecile, Khashayar Pakdaman, Pierre-Yves Boelle, Mark L. Wilson, Monica F. Myers, Alain-Jacques Valleron, and Antoine Flahault. 2004. "Association of Influenza Epidemics with Global Climate Variability." *European Journal of Epidemiology* 19 (11): 1055–59.

Volz, E. M. 2011. "Complex Population Dynamics and the Coalescent under Neutrality." *Genetics*, October. doi:10.1534/genetics.111.134627.

Waalen, K., A. Kilander, S. G. Dudman, R. Ramos-Ocao, and O. Hungnes. 2012. "Age-Dependent Prevalence of Antibodies Cross-Reactive to the Influenza A (H3N2) Variant Virus in Sera Collected in Norway in 2011." *Euro Surveill* 17 (19): 20170.

Walsh, J.A., and C. Maher. 2011. "Economic Implications of Influenza and Influenza Vaccine." *Influenza Vaccines for the Future*, 425–40.

Wang, Jinliang. 2001. "A Pseudo-Likelihood Method for Estimating Effective Population Size from Temporally Spaced Samples." *Genetical Research* 78 (03): 243–57.

Wang, Taia T., Gene S. Tan, Rong Hai, Natalie Pica, Erin Petersen, Thomas M. Moran, and Peter Palese. 2010. "Broadly Protective Monoclonal Antibodies against H3 Influenza Viruses Following Sequential Immunization with Different Hemagglutinins." Edited by Ron A. M. Fouchier. *PLoS Pathogens* 6 (2): e1000796. doi:10.1371/journal.ppat.1000796.

Ward, Richard L, H Fred Clark, and Paul a Offit. 2010. "Influence of Potential Protective Mechanisms on the Development of Live Rotavirus Vaccines." *The Journal of Infectious Diseases* 202 Suppl (Suppl 1): S72–79. doi:10.1086/653549.

Weaver, Scott C., and Nikos Vasilakis. 2009. "Molecular Evolution of Dengue Viruses: Contributions of Phylogenetics to Understanding the History and Epidemiology of the

Preeminent Arboviral Disease." *Infection, Genetics and Evolution* 9 (4): 523–40. doi:10.1016/j.meegid.2009.02.003.

Wertheim, J. O., and S. L. Kosakovsky Pond. 2011. "Purifying Selection Can Obscure the Ancient Age of Viral Lineages." *Molecular Biology and Evolution* 28 (12): 3355–65. doi:10.1093/molbev/msr170.

Whitton, J. Lindsay, Christopher T. Cornell, and Ralph Feuer. 2005. "Host and Virus Determinants of Picornavirus Pathogenesis and Tropism." *Nature Reviews Microbiology* 3 (10): 765–76.

WHO. 2013. "Pandemic Influenza Risk Management: WHO Interim Guidance." *Geneva: WHO*.

WHO, 2009. 2009. "WHO | Statements 2009." http://www.who.int/mediacentre/news/statements/2009/en/.

Wikramaratna, S. P, Michi Sandeman, Mario Recker, and Sunetra Gupta. 2012. "The Antigenic Evolution of Influenza: Drift or Thrift?" *Proceedings of the Royal Society B: Biological Sciences* In press.

Wolf, Y. I, C. Viboud, E. C Holmes, E. V Koonin, and D. J Lipman. 2006. "Long Intervals of Stasis Punctuated by Bursts of Positive Selection in the Seasonal Evolution of Influenza A Virus." *Biology Direct* 1 (1): 34.

Wood, Simon N. 2010. "Statistical Inference for Noisy Nonlinear Ecological Dynamic Systems." *Nature* 466 (7310): 1102–4. doi:10.1038/nature09319.

Yamashita, Makoto, Mark Krystal, Walter M. Fitch, and Peter Palese. 1988. "Influenza B Virus Evolution: Co-Circulating Lineages and Comparison of Evolutionary Pattern with Those of Influenza A and C Viruses." *Virology* 163 (1): 112–22. doi:10.1016/0042-6822(88)90238-3.

Yu, Hongjie, Wladimir J. Alonso, Luzhao Feng, Yi Tan, Yuelong Shu, Weizhong Yang, and Cécile Viboud,. 2013. "Characterization of Regional Influenza Seasonality Patterns in China and Implications for Vaccination Strategies: Spatio-Temporal Modeling of Surveillance Data." *PLoS Med* 10 (11): e1001552. doi:10.1371/journal.pmed.1001552.

Zeller, Mark, John T Patton, Elisabeth Heylen, Sarah De Coster, Max Ciarlet, Marc Van Ranst, and Jelle Matthijnssens. 2012. "Genetic Analyses Reveal Differences in the VP7 and VP4 Antigenic Epitopes between Human Rotaviruses Circulating in Belgium and Rotaviruses in Rotarix and RotaTeq." *Journal of Clinical Microbiology* 50 (3): 966–76. doi:10.1128/JCM.05590-11.

Zinder, Daniel, Trevor Bedford, Sunetra Gupta, and Mercedes Pascual. 2013. "The Roles of Competition and Mutation in Shaping Antigenic and Genetic Diversity in Influenza." *PLoS Pathog* 9 (1): e1003104. doi:10.1371/journal.ppat.1003104.

Zinder, Daniel, Robert J Woods, and Mercedes Pascual. 2014. "Early Signs of Post-Vaccination Change in the USA Rotavirus Population through Mutation, Migration and Shifting Prevalence." In . Sacramento, California: Ecological Society of America. http://eco.confex.com/eco/2014/webprogram/Paper49281.html.

Zinser, E. R. 2009. "Meeting of the Strategic Advisory Group of Experts on Immunization, October 2009 - Conclusions and Recommendations." *Relevé Épidémiologique Hebdomadaire / Section D'hygiène Du Secrétariat de La Société Des Nations = Weekly*

Epidemiological Record / Health Section of the Secretariat of the League of Nations 84 (50): 517–32.