Viral phylodynamics of host adaptation and antigenic evolution

Sidney M. Bell

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Trevor Bedford, Chair

Julie Overbaugh

Jesse Bloom

Program Authorized to Offer Degree:
Molecular and Cellular Biology

University of Washington

Abstract

Viral phylodynamics of host adaptation and antigenic evolution

Sidney M. Bell

Chair of the Supervisory Committee:
Dr. Trevor Bedford
Molecular and Cellular Biology

RNA viruses evolve quickly, on a comparable timescale to viral spread. Thus, each sampled case of an outbreak often carries a unique viral genomic signature. By comparing these genomes, phylogenies can reconstruct the course of viral evolution. These reconstructions are embedded with information about viral epidemiological patterns, which can be extracted via statistical models of population dynamics termed 'phylodynamics'. Here, I apply phylodynamic methods to two host/virus systems to investigate how viruses enter new populations, adapt to their hosts, and move through populations. I apply discrete trait analysis to lentiviral genomes to characterize the natural history of cross-species transmission (CST) among primates. I find that there have been at least 13 interlineage recombination events among lentiviruses, and identify 14 ancient CST events. This reveals a far more extensive history of lentiviral CST than previously recognizd, emphasizing the importance of ongoing surveillance. Next, I use a phylogeny-based model of antigenic change to quantify the extent and impact of antigenic evolution in dengue virus. I identify at least 12 distinct antigenic phenotypes of dengue virus, suggesting unrecognized, ongoing antigenic evolution; this presents an important consideration for vaccine design. I also find that antigenic novelty is a strong driver of dengue population turnover, providing context for epidemic preparedness efforts. Together, these findings demonstrate the power of phylodynamics to broaden our understanding of how viruses evolve and move through populations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work would not be possible without the thoughtful insights and scientific guidance from my colleagues and collaborators. The work presented in Chapter 2 was assisted by Michael Emerman, Janet Young, Vladimir Minin, Chris Whidden, Duncan Ralph, Anna Wald, and the Emerman lab. The work presented in Chapters 3 and 4 was assisted by Leah Katzelnick (a coauthor on the related manuscript), Richard Neher, Molly OhAinle, David Shaw, Paul Edlefsen, and Michal Juraska. The members of the Bedford lab have been an invaluable source of helpful advice and kind assistance for all of this work. I thank each of my colleagues for their contributions to my research and to my development as a scientist.

I must sincerely thank my advisor, Dr. Trevor Bedford, for his kind and careful mentorship. He both taught me how to do science and serves as a model for the kind of scientist I strive to be.

Finally, I must thank my family and friends. Their love, encouragement and support has been instrumental in helping me grow personally and professionally throughout my doctoral work.

This work was financially supported by the NSF Graduate Research Fellowship Program (DGE-1256082), NIH Interdisciplinary Training Grant (T32CA080416), Vassar Fellowships for Graduate Study, and the ARCS Scholars Program.

Chapter 1

INTRODUCTION TO VIRUS EVOLUTION AND PHYLODYNAMICS

1.1 Virus evolution is shaped by host/virus relationships

As obligate parasites, viruses are intimately dependent upon their hosts for survival. The unique nature of the host/virus relationship has a profound impact on viral fitness and evolutionary dynamics [48]. Broadly, viral fitness is comprised of its ability to replicate within a host and transmit between hosts. Evolutionary conflict between the virus and the host immune system influences both of these dynamics: immunologically evasive viruses are able to replicate more extensively within a host, and are able to access a greater number of susceptible hosts. These viral evolutionary imperatives to replicate, transmit, and evade are interrelated, and may either compete or complement each other.

For example, influenza virus evolution is largely driven by antigenic fitness [16]. Influenza strains that are antigenically distinct from previously circulating strains are better able to infect susceptible hosts. This generally favors rapid diversification of external influenza proteins, although receptor binding must be maintained to preserve replicative fitness. Contrastingly, dengue virus evolution is primarily driven by replicative fitness [41]. Dengue infects primates (including humans), but is spread via mosquitos. Thus, in order to maintain an unbroken chain of transmission, dengue virions must be able to infect and replicate within two vastly different host environments. Maintaining this delicate balance generally favors conservation of most dengue proteins. As I show in Chapter 3, dengue does evolve antigenically, but much more slowly than influenza [77, 14]. This may be due to the dual replicative constraints placed on dengue viral evolution.

1.2 A framework for investigating host adaptation

For viruses with RNA genomes (*e.g.*, influenza, dengue, and HIV), their high intrinsic mutation rate puts viral evolution and spread on comparable timescales. This embeds the selective pressures exerted by the host environment within the viral phylogeny [10]. These viral phylogenies then become powerful tools for reconstructing patterns of viral population dynamics and spread via statistical methods collectively referred to as 'phylodynamics' [40]. Phylodynamic methods rely upon phylogenies of virus evolution, which describe how sampled viruses are related to one another. In phylogenetic trees, leaves represent sampled viruses; internal nodes represent hypothesized ancestral viruses; and branches are typically scaled to genetic divergence. The topology — or 'shape' — of the phylogeny reflects the order of 'relatedness' and sequential divergence between viruses on the phylogeny.

Figure 1.1. Comparison of dengue and influenza phylogenies **A** shows a phylogeny of dengue virus sequences (serotype 2). **B** shows a phylogeny of influenza virus sequences (H3N2 lineage). The strikingly different topologies of these two trees reflects the different selective pressures that have shaped the evolution of these two viruses. Both trees were generated via Nextstrain [43].

Returning to our previous example, the contrasting selective pressures presented by the host environments of dengue and influenza viruses are evident in the corresponding viral phylogenies (Figure 1.1) [98, 10]. Influenza must rapidly diversify antigenically to evade

the host's immune system. As a result, viruses that are phenotypically similar to ancestral strains tend to have fewer progeny, while antigenically divergent strains are more likely to seed future viral diversity. This results in a spindly, 'ladder-like' tree topology [10, 11]. In contrast, dengue virus does antigenically evolve, but is much more tightly constrained by the dual host system in which it replicates. Here, we see that within each serotype, ancestral phenotypes persist in circulation for much longer. This gives the dengue phylogeny a 'bushier' tree topology [10].

1.3 Summary of presented analyses

Phylodynamic methods provide a more formalized view of how different selective pressures have shaped viral evolution by applying well-established statistical models of how traits evolve over time. In doing so, they can provide crucial evolutionary, ecological and epidemiological context to viral outbreaks through descriptive and/or explanatory analyses. In the studies reported here, I leverage phylodynamic methods to analyze how viruses infect new host species, evade host immunity, and spread through host populations. I characterize two host/virus systems: simian immunodeficiency viruses in primates, and dengue virus in a hyperendemic human population.

1.3.1 Characterizing the natural history of viral cross-species transmission

In Chapter 2, I present a descriptive analysis contextualizing the HIV pandemic within the broader scope of viral cross-species transmission. HIV's origins and evolution have been shaped by unique evolutionary forces and epidemiological dynamics. While influenza and dengue cause relatively short infections and transmit rapidly, HIV's extraordinary ability to evade the immune system results in chronic infection, enabling much slower transmission (as a function of time). This presents an interesting evolutionary strategy, wherein high antigenic fitness is used to compensate for lower transmissability. And yet, despite this relatively low transmission rate, HIV has been acquired from primates via at least sixteen independent cross-species tranmissions (CSTs). This is particularly remarkable when one considers the

many challenges inherent to viral replication in an entirely foreign host evironment (collectively referred to as the 'species barrier').

The immediate history of HIV is well-established [93]. HIV is comprised of two different viruses, HIV-1 and HIV-2; HIV-1, the cause of the global pandemic, was acquired via four independent transmissions from apes, and HIV-2 was acquired via many transmissions from Old World primates. However, HIV-1 and -2 represent only two of over 45 distinct immunodeficiency viruses (SIVs) known to infect primates. Although we understand the most immediate origins that generated the global pandemic, we lack crucial ecological context around how often these viruses switch hosts and what enables a lentiviral cross-species transmission to take hold in a population.

Here, I take a phylodynamics-based approach to characterize the extent and patterns of cross-species transmission that have shaped primate lentiviral evolution and seeded the HIV pandemic. To do so, I curated a dataset of publicly available SIV sequences, isolated from 24 primate host species; identified breakpoints of interlineage recombination across the viral genome; and reconstructed the evolutionary history of each genomic segment. I then modeled the host state of each sampled and ancestral virus (tip and internal node, respectively) as a discrete trait and inferred pairwise transmission rates within a Bayesian framework.

I identify 14 novel, well-supported, ancient cross-species transmission events. I also examine the origins of SIV*cpz* (the predecessor of HIV-1) in greater detail than previous studies, and find that there are still large portions of the genome with unknown origins. This places the HIV outbreak into ecological context, alongside many other cross-species transmission events (for a detailed discussion, see Chapter 2.3.2 — 2.3.3). Interestingly, we observe that lentiviral lineages vary widely in their ability to infect new host species: SIV*col* (from colobus monkeys) is evolutionarily isolated, while SIV*agm*s (from African green monkeys) frequently move between host subspecies. Although the vast timescale limits direct causal inference across the phylogeny, these cross-species transmissions are more common among pairs of closely related host species, which is likely a result of ecological circumstance and differences in innate immune factors. Overall, using phylodynamics to frame the HIV pandemic in this

way helps us understand the broader context in which it arose and more accurately evaluate risk of future cross-species transmission.

1.3.2 Quantifying the extent and impact of antigenic evolution

As illustrated in Chapter 2, phylodynamics is an effective tool for retrospective, descriptive analyses; however, these methods are also powerful tools for explanatory, prospective analyses. In Chapters 3 and 4, I apply phylodynamics to understand how dengue virus adapts to evade human immunity and spread through populations.

Dengue virus exists as four genetically distinct clades canonically termed 'serotypes'. As discussed above, dengue evolution is relatively slow and putatively driven by replicative fitness. However, antigenic relationships between dengue virus clades dictate case outcomes and epidemic severity. This is mediated by antibody-dependent enhancement (ADE), wherein circulating cross-reactive, nonneutralizing antibodies augment heterotypic secondary infection severity. Because the serotypes are so clearly distinct, and because dengue evolves so slowly overall, it has long been assumed that all viruses within a serotype are antigenically uniform (*i.e.*, they look very similar to the host immune system) [96]. Interestingly, recent analyses suggest that antigenic heterogeneity may exist within each serotype, but its source, extent and impact remain unclear [56].

In Chapter 3, I leverage both genomic and functional assay data to quantify and characterize the extent of dengue antigenic evolution. To do so, I adapt a statistical model (originally developed for analyzing influenza's rapid antigenic evolution [77]) to characterize dengue's slower, subtler antigenic dynamics. In this model, each branch in the dengue phylogeny is treated as essentially one 'step' in the overall course of evolution. The model then maps antigenic divergence to underlying genetic evolution by quanitfying how much antigenic change has occurred along each branch of the phylogeny. This provides a granular view of the evolutionary paths that have shaped dengue antigenic phenotypes over time. I find moderate antigenic diversity within each serotype, and identify 12 antigenically distinct clades. This represents a major departure from the null hypothesis that there are only four

extant antigenic phenotypes of dengue.

However, when compared to pathogens that antigenically evolve very rapidly (such as influenza), the breadth of dengue antigenic diversity is very modest. In Chapter 4, I investigate the impact of antigenic heterogeneity on real-world DENV population dynamics. To do so, I hypothesize that antigenic distance from standing population immunity ('antigenic novelty') results in higher viral fitness and thus drives clade growth. To test this hypothesis, I analyze dengue population dynamics across Southeast Asia over a thirty year time period using a model adapted from [73]. At each quarterly timepoint, I quantify standing population immunity across the region, and compare this to which dengue clades are most successful over a subsequent five-year window. I find that antigenic novelty is able to explain most of the observed variation in the dengue population composition, although this appears to be driven by coarser serotype-level antigenic differences.

This strongly suggests that although dengue evolution is tightly constrained by replicative fitness, antigenic fitness ultimately drives dengue population dynamics. Furthermore, the existance of large, ancient antigenic divergence between serotypes and smaller, more recent antigenic divergence within serotypes suggests that dengue antigenic evolution is gradual but ongoing. These results provide a more nuanced understanding of dengue antigenic evolution, with important ramifications for vaccine design and epidemic preparedness. More broadly, this also carries significant implications for our view of antigenic evolution in even slowly-evolving pathogens.

Chapter 2

SIV EVOLUTION AND CROSS-SPECIES TRANSMISSION

This chapter is adapted from a previously published manuscript, which I coauthored with Dr. Trevor Bedford [13, 12].

2.1 Introduction to simian immunodeficiency viruses

As demonstrated by the recent epidemics of EBOV and MERS, and by the global HIV pandemic, viral cross-species transmissions (CST) can be devastating [76, 81]. As such, understanding the propensity and ability of viral pathogens to cross the species barrier is of vital public health importance. Of particular interest are transmissions that not only 'spillover' into a single individual of a new host species, but that result in a virus actually establishing a sustained chain of transmission and becoming endemic in the new host population ('host switching') [69].

HIV is the product of not just one successful host switch, but a long chain of host switch events [7, 93]. There are two human immunodeficiency viruses, HIV-2 and HIV-1. HIV-2 arose from multiple cross-species transmissions of SIV*smm* (simian immunodeficiency virus, sooty mangabey) from sooty mangabeys to humans [21, 33, 47]. HIV-1 is the result of four independent cross-species transmissions from chimpanzees and gorillas. Specifically, SIV*cpz* was transmitted directly from chimpanzees to humans twice; one of these transmissions generated HIV-1 group M, which is the primary cause of the human pandemic [32]. SIV*cpz* was also transmitted once to gorillas, generating SIV*gor* [97], which was in turn transmitted twice to humans [25].

Looking further back, SIV*cpz* itself was also generated by lentiviral host switching and

recombination. Based on the SIV sequences available at the time, early studies identified SIV*mon*/-*mus*/-*gsn* (which infect mona, mustached, and greater spot-nosed monkeys, respectively) and SIV*rcm* (which infects red-capped mangabeys) as probable donors [9]. Functional analysis of accessory genes from these putative parental lineages indicate that the specific donors and genomic locations of these recombination event(s) were crucial for enabling what became SIV*cpz* to cross the high species barrier and establish an endemic lineage in hominids [28].

The complex evolutionary history resulting in HIV illustrates the importance of natural history to modern day viral diversity, and although the history leading to HIV is well detailed, broader questions regarding cross-species transmission of primate lentiviruses remain [37]. There are over 45 known extant primate lentiviruses, each of which is endemic to a specific host species [2, 7, 93], but the history of viral transmission between these host species has not been characterized.

Using phylogenetic inference, I reconstructed the evolutionary history of primate lentiviral recombination and cross-species transmission to the degree possible given a limited sample of modern-day viruses. I assembled datasets from publicly available lentiviral genome sequences and conducted discrete trait analyses to infer rates of transmission between primate hosts. I find evidence for extensive interlineage recombination and identify many novel host switches that occurred during the evolutionary history of lentiviruses. I also find that specific lentiviral lineages exhibit a broad range of abilities to cross the species barrier. Finally, I also examined the origins of each region of the SIV*cpz* genome in greater detail than previous studies to yield a more nuanced understanding of its origins.

2.2 Results

2.2.1 SIV interlineage recombination

In order to reconstruct the lentiviral phylogeny, we must first address the issue of recombination, which is frequent among lentiviruses [18]. In the context of studying cross-species transmission, this is both a challenge and a valuable tool. Evidence of recombination between viral lineages endemic to different hosts is also evidence that at one point in time, viruses from those two lineages were in the same animal (*i.e.*, a cross-species transmission event must have occurred in order to generate the observed recombinant virus). However, this process also results in portions of the viral genome having independent evolutionary—and phylogenetic-histories.

To address the reticulate evolutionary history of SIVs, I endeavored to identify the extent and nature of recombination between lentiviral lineages. Extensive sequence divergence between lineages masks site-based methods for linkage estimation (Figure A.1). However, topology-based measures of recombination allow for borrowing of information across nearby sites, and are effective for this dataset. I thus utilized a phylogenetic model to group segments of shared ancestry separated by recombination breakpoints instantiated in the HyPhy package GARD [60]. For this analysis, I used a version of the SIV compendium alignment from the Los Alamos National Lab (LANL), modified slightly to reduce the overrepresentation of HIV sequences (N=64, see Methods) [62]. Importantly, because each virus lineage has only a few sequences present in this alignment, these inferences refer to inter-lineage recombination, and not the rampant intra-lineage recombination common among lentiviruses.

GARD identified 13 locations along the genome that had strong evidence of inter-lineage recombination (Figure 2.1). Here, evidence for a particular model is assessed via Akaike Information Criterion (AIC) and differences in AIC between models indicate log probabilities, so that a ΔAIC of 10 between two models would indicate that one model is $e^{(10/2)} \approx 148$ times as likely as the other [4]. In this case, ΔAIC values ranged from 154 to 436 for each included breakpoint, indicating that these breakpoints are strongly supported by the

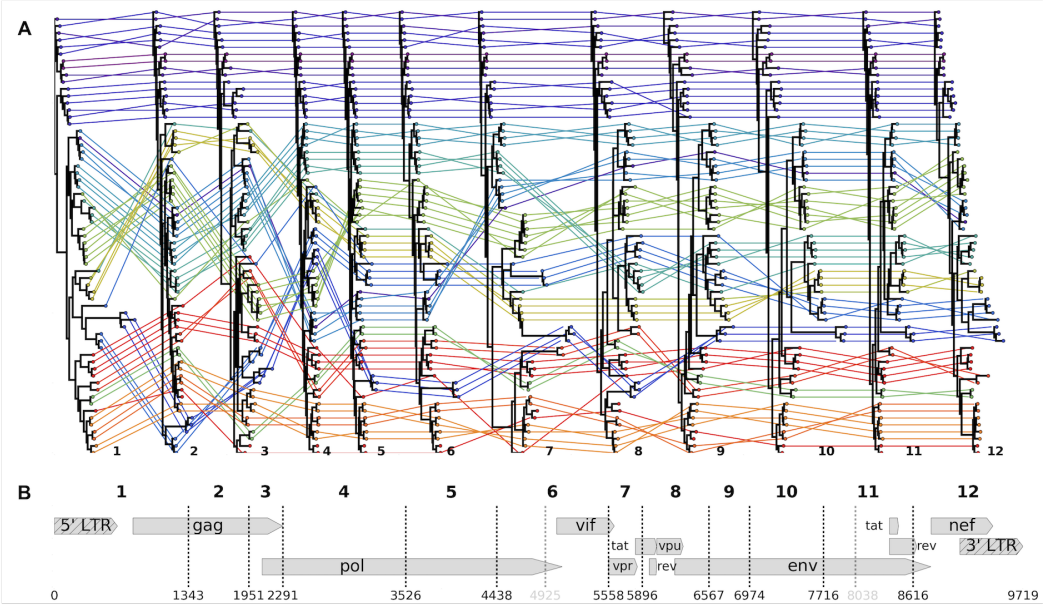

Figure 2.1. Inferred interlineage recombination breakpoints and supporting tree topologies The SIV LANL compendium, slightly modified to reduce overrepresentation of HIV, was analyzed with GARD to identify 13 recombination breakpoints across the genome (dashed lines in **B**; numbering according to the accepted HXB2 reference genome—accession K03455, illustrated). Two of these breakpoints were omitted from further analyses because they created extremely short fragments (< 500 bases; gray dashes in **B**). For each of the 11 remaining breakpoints used in further analyses, the compendium alignment was split along these breakpoints and used to build a maximum likelihood tree, displayed in **A**. Each viral sequence is color-coded by host species, and its phylogenetic position is traced between trees. Heuristically, straight, horizontal colored lines indicate congruent topological positions between trees (more likely not a recombinant sequence); criss-crossing colored lines indicate incongruent topological positions between trees (likely a recombinant sequence).

underlying tree likelihoods. The 14 resulting segments ranged in length from 351 to 2316 bases; in order to build reliable phylogenies, I omitted two of the less supported breakpoints from downstream analyses, yielding 12 segments ranging in length from 606 to 2316 bases. I found no evidence to suggest linkage between non-neighboring segments (Figure A.2). While it has been previously appreciated that several lineages of SIV are recombinant products

(*e.g.*, [9, 51]), the 13 breakpoints identified here provide evidence that there have been at least 13 inter-lineage recombination events during the evolution of SIVs. Identifying these recombination breakpoints allowed us to construct a putatively valid phylogeny for each segment of the genome that shares an internally cohesive evolutionary history.

2.2.2 Phylogenetic evidence of SIV cross-species transmission

Phylogenetic evidence of cross-species transmission may be found in the tree topologies of each of the 12 genomic segments. For this and all further analyses, I constructed a dataset from all publicly available primate lentivirus sequences, curated and subsampled by host and virus lineage to ensure an equitable distribution of data (see Methods). This primary dataset consists of virus sequences from the 24 primate hosts with sufficient data available (5 - 25 sequences per viral lineage, N=423, Figure A.3). Alignments used the fixed compendium alignment as a template (see Methods).

In phylogenetic trees of viral sequences, cross-species transmission appears as a mismatch between the host species of a sampled versus ancestral virus. To identify this pattern and estimate the frequency of CST, I modeled the host of each viral sample as a discrete state that can be inferred for each ancestral virus (internal node) in the phylogeny. This is analogous to treating the host state of each viral sample as an extra column in an alignment, and inferring the ancestral states across the phylogeny along with the rate of transition between them. This approach is similar to common phylogeographic approaches that model movement of viruses across discrete spatial regions [98] and has previously been applied to modeling discrete host state in the case of rabies virus [30].

Here, I took a fully Bayesian approach and sought the posterior distribution across phylogenetic trees, host transition rates and ancestral host states. The model integrates over parameter values using Markov chain Monte Carlo (MCMC) to yield phylogenetic trees for each segment annotated with ancestral host states alongside inferred host transition rates.

Figure 2.2 shows reconstructed phylogenies for 3 segments along with inferred ancestral host states. Trees are color coded by known host state at the tips, and inferred host state at

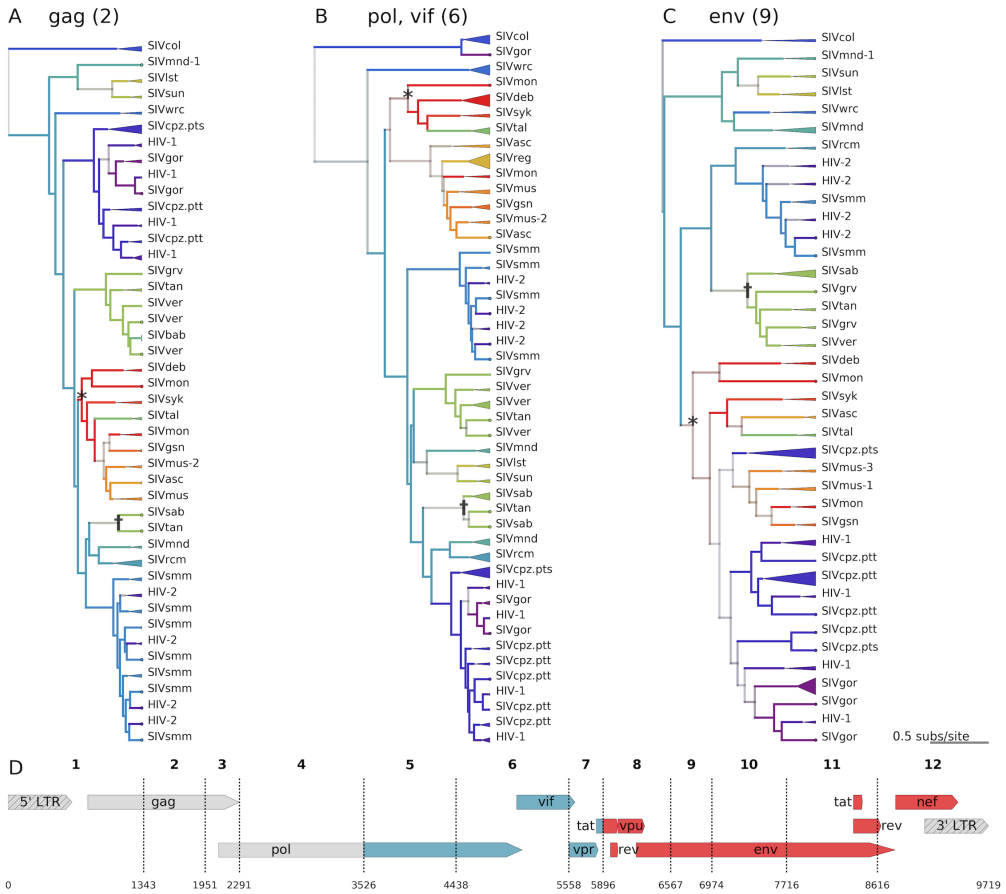

Figure 2.2. Lentiviral phylogenies highlighting the mosaic origins of SIV*cpz* and examples of how CST is inferred from the phylogenies **A,B,C** Bayesian maximum clade credibility (MCC) trees are displayed for segments 2 (*gag* - **A**), 6 (*int* and *vif* - **B**), and 9 (*env* **C**) of the main dataset (N=423). Tips are color coded by known host species; internal nodes and branches are colored by inferred host species, with saturation indicating the confidence of these assignments. Monophyletic clades of viruses from the same lineage are collapsed, with the triangle width proportional to the number of represented sequences. An example of likely cross-species transmission is starred in each tree, where the host state at the internal node (red / mona monkeys) is incongruent with the descendent tips' known host state (green / talapoin monkeys), providing evidence for a transmission from mona monkeys to talapoin monkeys. Another example of cross-species transmission of a recombinant virus among African green monkeys is marked with †. **D** The genome map of SIV*cpz*, with breakpoints used for the discrete trait analysis, is colored by the most likely ancestral host for each segment of the genome.

internal nodes/branches; color saturation indicates the level of certainty for each ancestral host assignment. A visual example of how the model identifies cross-species transmissions can

be seen in the SIV*mon*/SIV*tal* clade, which infect mona- and talapoin monkeys, respectively (starred in Figure 2.2A-C). Due to the phylogenetic placement of the SIV*mon* tips, the internal node at the base of this clade is red, indicating that the host of the ancestral virus was most likely a mona monkey. This contrasts with the host state of the samples isolated from talapoin monkeys (tips in green). These changes in the host state across the tree are what inform the model's estimates of the rate of transmission between host pairs. In total, the support for each possible transmission is derived from both A) whether the transmission is supported across the posterior distribution of phylogenies for a particular segment, and B) whether this is true for multiple genomic segments.

Notably, the tree topologies are substantially different between segments, which emphasizes both the extent of recombination and the different evolutionary forces that have shaped the phylogenies of individual portions of the genome. In all segments' trees, we also see frequent changes in the host state between internal nodes (illustrated as changes in color going up the tree), suggestive of frequent ancient cross-species transmissions. On average, primate lentiviruses switch hosts once every 6.25 substitutions per site per lineage across the SIV phylogeny.

The cross-species transmission events inferred by the model are illustrated in Figure 2.3, with raw rates and Bayes factors (BF) in Figure A.4. As shown, the model correctly infers nearly all pairs of hosts with previously identified (to my knowledge) CST events [9, 25, 33, 32, 52, 65], with the exception of the putative CST from sooty mangabeys to sabaeus monkeys reported by [51] (see discussion). Importantly, I also identify 14 novel cross-species transmission events with strong statistical support (cutoff of BF $\geq$ 10.0). Each of these transmissions is clearly and robustly supported by the tree topologies (all 12 trees are illustrated in Figure A.5).

To control for sampling effects, I repeated the analysis with a supplemental dataset built with fewer hosts, and more sequences per host (15 host species, subsampled to 16-40 sequences per viral lineage, N=510), and see consistent results. As illustrated in Figures A.6– A.8, we see qualitatively similar results. When directly comparing the average

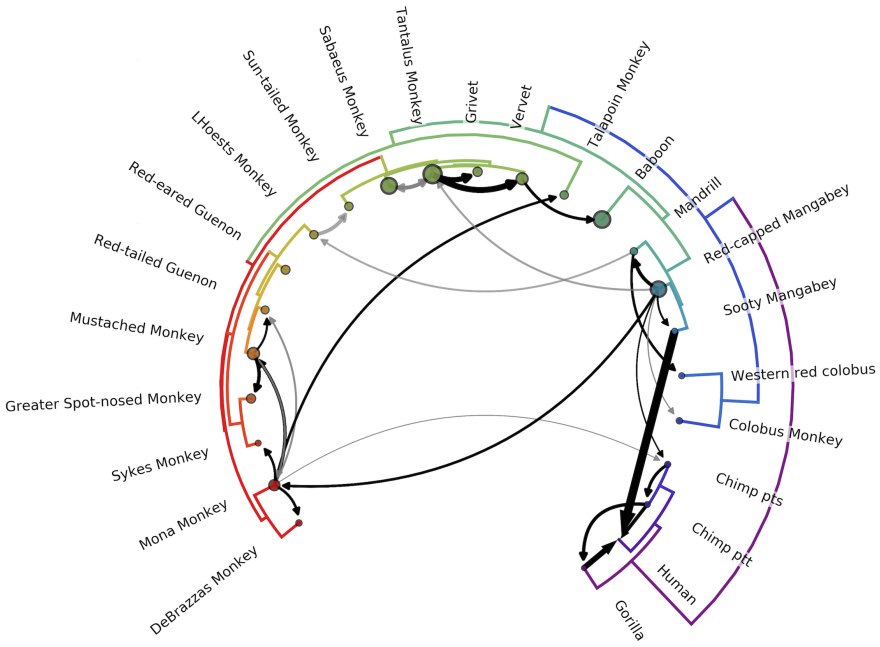

Figure 2.3. Network of inferred CSTs of primate lentiviruses The phylogeny of the host species' mitochondrial genomes forms the outer circle. Arrows represent transmission events inferred by the model with Bayes' factor (BF) ≥ 3.0; black arrows have BF ≥ 10, with opacity of gray arrows scaled for BF between 3.0 and 10.0. Width of the arrow indicates the rate of transmission. Circle sizes represent network centrality scores for each host. Transmissions from chimps to humans; chimps to gorillas; gorillas to humans; sooty mangabeys to humans; sabaeus to tantalus; and vervets to baboons have been previously documented. To my knowledge, all other transmissions illustrated are novel identifications.

indicator values, host state transition rates, and BF values between analyses, the results from the main and supplemental datasets are strongly correlated, indicating robust quantitative agreement between the two analyses (Figure A.9). There is similar agreement between replicates whenthe main dataset is regenerated via independent sampling draws (see online repository).

Taken together, these results represent a far more extensive pattern of CST among primate lentiviruses than previously described [7, 93]; nearly every primate clade has at least one inbound, robustly supported viral transmission from another clade. I thus conclude that the majority of lentiviruses have arisen from a process of host switching, followed by a

combination of intraclade host switches and host-virus coevolution.

2.2.3 Variable host specificity of SIVs

While most SIVs are the product of ancient recombination and host switching, the distribution of these host switches is not uniform; there is a broad range of network centrality across hosts (Figures 2.3, A.7, as node size), indicating some hosts act as sources in the SIV transmission network and other hosts act as sinks. From this, I infer that some viruses have either had greater opportunity or have a greater ability to cross the species barrier than others.

In particular, the SIVs from the four closely related species of African green monkeys (SIV*sab*, SIV*tan*, SIV*ver*, SIV*grv*; collectively, SIV*agm*) appear to exchange viruses with other host species frequently (Figure 2.3, 12 o-clock). An example of SIV*agm* CST events can be seen in the tree topologies from *gag*, *prot*, reverse transcriptase (*RT*), and *vif* (Figures 2.2 and A.5, segments 2, 3, 4, 6). Here, SIV*tan* isolates reported by Ayouba *et al.* [8] (denoted with † in Figure 2.2) clearly cluster with SIV*sab*, in a distant part of the phylogeny from the rest of the SIV*agm* viruses (including the majority of SIV*tan* isolates). For all other segments, however, the SIV*agm*s cluster together. I thus concur with the conclusion of Ayouba *et al.* that these samples represent a recent spillover of SIV*sab* from sabaeus monkeys to tantalus monkeys, and the model appropriately identifies this transmission.

Contrastingly, previous studies of the lentiviral phylogeny have noted that SIV*col* is typically the outgroup to other viral lineages, and have hypothesized that this may implicate SIV*col* as the original primate lentivirus [22]. I find this hypothesis plausible, but the evidence remains inconclusive. For the majority of genomic segments, we also observe SIV*col* as the clear outgroup (Figures 2.2 and A.5). In contrast, for portions of *gag*/*pol* (segments 3 and 4) and some of the accessory genes (segment 7), I find that there is not a clear outgroup. For these segments, many other lineages of SIV are just as closely related to SIV*col* as they are to each other. However, with the occasional exception of single heterologous taxa with poorly supported placement, SIV*col* remains a monophyletic clade (N=16), and does

not intercalate within the genetic diversity of any other lineage in this dataset. Based on these collective tree topologies, this model does not identify strong evidence for any specific transmissions out of colobus monkeys, and identifies only a single, weakly supported inbound transmission (likely noise in the model caused by the fact that red-capped mangabeys are the marginally supported root host state; see below). This is consistent with previous findings that the colobinae — in a different genus than most of the *Cercopithecus* primates in the datase — have a unique variant of the *APOBEC3G* gene, which is known to restrict lentiviral infection and speculated to be a barrier to cross-species transmission [22]. These observations generally support the idea of SIV*col* as having maintained a specific relationship with its host over evolutionary time.

Additionally, while most host species carry only one lineage of SIV, mandrils and mustached monkeys carry 2 and 3 lineages of SIV, respectively [2, 23, 67, 94]. In agreement with these previous studies, SIV*mnd*-1 and SIV*mnd*-2 do not always cluster together in the phylogeny; the same is true for SIV*mus*-1, SIV*mus*-2, and SIV*mus*-3, indicating that each of these viral lineages likely has a unique origin. This stands in stark contrast to baboons, which have only been infected by an SIV via a single documented spillover event [52].

Collectively, these examples demonstrate that the nature of the host-virus relationship is highly variable for primate lentiviruses, with some viruses switching hosts often while others putatively maintain strict host specificity. Likewise, while some hosts have acquired multiple SIV lineages, most are infected by only one SIV, or do not have an endemic SIV.

2.2.4 The evolutionary history of SIVcpz, the precursor to HIV-1

Unlike SIV*col*, SIV*cpz* appears to be the product of multiple CSTs and recombination events. SIV*cpz* actually encompasses two viral lineages: SIV*cpzPtt* infects chimpanzees of the subspecies *Pan troglodytes troglodytes*, and SIV*cpzPts* infects chimpanzees of the subspecies *Pan troglodytes schweinfurthii* [92]. There are two additional subspecies of chimpanzees that have not been found harbor an SIV despite extensive surveys, suggesting that SIV*cpzPtt* was acquired after chimpanzee sub-speciation [65]. Both this previous work and my own results

support the hypothesis that SIV*cpz* was later transmitted from one chimpanzee subspecies to the other, and SIV*cpzPtt* is the only SIV*cpz* lineage that has crossed into humans. Given the shared ancestry of the two lineages of SIV*cpz*, I use SIV*cpz* to refer specifically to SIV*cpzPtt*.

Based on the lentiviral sequences available in 2003, Bailes *et al* [9] suggested that the SIV*cpz* genome is a recombinant of just two parental lineages. SIV*rcm* (which infects red-capped mangabeys) was identified as the 5' donor, and an SIV from the SIV*mon*/*-mus*/*-gsn* clade (which infect primates in the *Cercopithecus* genus) was identified as the 3' donor. Since the time of this previous investigation many new lentiviruses have been discovered and sequenced. In incorporating these new data, I find clear evidence that the previous two-donor hypothesis may be incomplete.

The tree topologies from *env* in the 3' end of the genome (segments 8-11) support the previous hypothesis [9] that this region came from a virus in the SIV*mon*/*-mus*/*-gsn* clade. These viruses form a clear sister clade to SIV*cpz* with high posterior support (Figure 2.2C,D). I find strong evidence for transmissions from mona monkeys (SIV*mon*) to mustached monkeys (SIV*mus*), and from mustached monkeys to greater spot-nosed monkeys (SIV*gsn*) (see discussion of potential coevolution below). I also find more evidence in support of a transmission from mona monkeys to chimpanzees than from the other two potential donors, but additional sampling is required to firmly resolve which of these viruses was the original donor of the 3' end of SIV*cpz*.

I find phylogenetic evidence to support the previous hypothesis [9, 28] that the *int* and *vif* genes of SIV*cpz* (segments 4-6) originated from SIV*rcm*; however, I find equally strong evidence to support the competing hypothesis that *pol* came from SIV*mnd*-2, which infects mandrils (Figure 2.2B,D). In these portions of the genome, SIV*mnd*-2 and SIV*rcm* together form a clear sister clade to SIV*cpz*. The *vpr* gene, in segment 7, is also closely related to both SIV*rcm* and SIV*mnd*-2, but this sister clade also contains SIV*smm* from sooty mangabeys. Notably, the model infers a transmission from red-capped mangabeys to mandrils, but cannot determine whether this portion of the SIV*cpz* genome was acquired directly from SIV*rcm* or from SIV*mnd*-2.

Interestingly, I do not find evidence to support either SIV*rcm*/*mnd*-2 or SIV*mon*/-*mus*/-*gsn* as the donor for the 5' most end of the genome (segments 1-5), including the 5' *LTR*, *gag*, and *RT* genes. This is also true for the 3' *LTR* (segment 12). SIV*cpz* lacks a clear sister clade or ancestor in this region, and SIV*rcm* groups in a distant clade; I therefore find no evidence to suggest that an ancestor of an extant SIV*rcm* was the parental lineage of SIV*cpz* in the 5' most end of the viral genome as previously believed (Figure 2.2A,D). This may support the possibility of a third parental lineage, or a number of other plausible scenarios (discussed below).

2.3 Discussion

2.3.1 Limitations and strengths of the model

Additional sampling is required to fully resolve the history of CST among lentiviruses

In addition to the 14 strongly supported novel transmissions (BF $\geq$ 10) described above, I also find substantial evidence for an additional 8 possible novel transitions, but with lower support (BF $\geq$ 3) (Figures 2.3, A.4). These transmissions are more difficult to assess, because many of them are inferred on the basis of just a few outlier tips of the tree that group apart from the majority of viral samples from the same lineage. In each case, the tips' phylogenetic position is strongly supported, and the primary literature associated with the collection of each of these outlier samples clearly specifies the host metadata. However, due to the limited number of lentiviral sequences available for some hosts, I am unable to control for sampling effects for some of these lower-certainty transmissions. I report them here because it is unclear whether these outliers are the result of unidentified separate endemic lineages, one-time spillovers from other hosts, or species misidentification during sample collection. It is also important to note that while some of these less-supported transmissions are potentially sampling artifacts, many of them may be real, and may be less supported simply because they lack the requisite available data for some genome segments.

Ultimately, far more extensive sampling -- specifically, obtaining more full-length se-

quences from undersampled lineages -- of primate lentiviruses is required in order to resolve these instances. The dataset included only sequences at least 500 bases long; each taxon may contribute more informative sites to some segments than to others. When splitting the master alignment along breakpoints, any taxon that had no informative bases was removed from each segment. However, for each segment, there were between 0 and 13 (mean: 3.6) taxa that had some informative sites, but were very short (< 100 informative sites). Statistically, these short taxa contribute little information, and their placement in the topology for each segment has high uncertainty. This phylogenetic uncertainty is then propagated forward to the discrete traits model, meaning that these short taxa should not statistically influence these results in any meaningful way. Notably, though, their removal does result in extensive technical challenges (this pruned dataset results in poor mixing and rather divergent results, seen in Figures A.10 – A.12). Given the high congruence between results from independent sampling replicates of the main dataset and from an alternative sampling scheme, I believe this to be a technical issue, rather than reflecting true differences. However, this issue does further emphasize the importance of additional sampling in fully resolving the natural history of SIVs.

Most lentiviruses were originally acquired by CST and have since coevolved with their hosts

Some of these noisier transmission inferences, particularly within the same primate clade, may be the result of coevolution, *i.e.* lineage tracking of viral lineages alongside host speciation. Within the model, viral jumps into the common ancestor of two extant primate species appear as a jump into one of the extant species, with a secondary jump between the two descendants. For example, the model infers a jump from mona monkeys into mustached monkeys, with a secondary jump from mustached monkeys into their sister species, red-tailed guenons (Figure 2.3). Comparing the virus and host phylogenies, we observe that this host tree bifurcation between mustached monkeys and red-tailed guenons is mirrored in the virus tree bifurcation between SIV*mus* and SIV*asc* for most segments of the viral genome (Figures 2.2, 2.3). This heuristically suggests that the true natural history may be

an ancient viral transmission from mona monkeys into the common ancestor of mustached monkeys and red-tailed guenons, followed by host/virus coevolution during primate speciation to yield SIV*mus* and SIV*asc*. The possibility of coevolution means that while we also observe extensive host switching between primate clades, many of the observed jumps within a primate clade may be the result of host/virus coevolution. However, the species barrier is likely lower between closely related primates, making it challenging to rigorously disentangle coevolution vs. true host switches within a primate clade [17].

The model propagates and accounts for residual uncertainty from the recombination analyses

The deep phylogenies and extensive sequence divergence between SIV lineages makes any assessment of recombination imperfect. This analysis aims to 1) place a lower bound on the number of interlineage recombination events that must have occurred to explain the observed extant viruses, and 2) use this understanding to construct a model of CST among these viruses. As the most well developed package currently available for topology-based recombination analysis, GARD was an appropriate choice to identify in broad strokes the extent and nature of recombination among SIVs. Some (though not all) of the remaining uncertainty as to the exact location of breakpoints is represented within the posterior distribution of trees for each segment, and is thus propagated and accounted for in the discrete traits model (inferences of CST). Most other studies of SIV evolutionary history simply split the phylogeny along gene boundaries or ignore recombination (*e.g.*, [9, 17, 93]). Thus, while there is still some remaining uncertainty in the recombination analyses, these results still represent a major step forward in attempting to systematically assess recombination among all extant SIV lineages and to incorporate it into the phylogenetic reconstruction.

2.3.2 Constraints and drivers of cross-species transmission

Paleovirology, biogeography, and statistical models of lentiviral evolution estimate that primate lentiviruses share a common ancestor approximately 5-10 million years ago [3, 22, 74, 102]. This, along with the putative viral coevolution during primate speciation, suggests

that many of these transmissions were ancient, and have been acted on by selection for millions of years. Thus, given that the observed transmissions almost exclusively represent evolutionarily successful host switches, it is remarkable that lentiviruses have been able to repeatedly adapt to so many new host species. In the context of this vast evolutionary timescale, however, I conclude that while lentiviruses have a far more extensive history of host switching than previously understood, these events remain relatively rare overall. As noted above, these results illustrate that some SIVs cross the species barrier more readily than others, and some primate host species become infected with new viral lineages more commonly than others. This is likely governed by both ecological and biological factors. Ecologically, frequency and form of exposure are likely key determinants of transmission [69], but these relationships can be difficult to describe statistically.

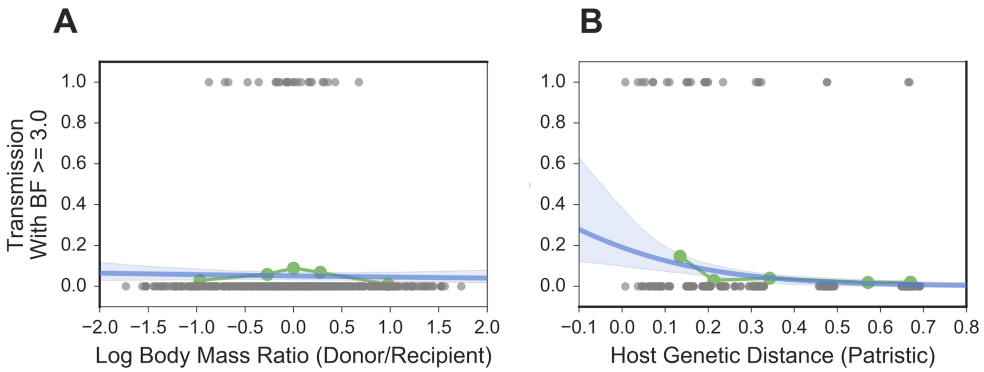

Figure 2.4. Logistic regressions of body mass ratios (a proxy for predation) and host genetic distance on the probability of CST (A) Log ratio of average body masses for each pair of hosts and **(B)** patristic genetic distance (from a maximum-likelihood tree of mtDNA) between each pair of hosts, versus indicator value for transmission between them (indicator = 1 if the BF for that transmission was ≥ 3.0, otherwise indicator = 0). Each plot shows raw predictor data in gray; the quintiles of the predictor data in green; and the logistic regression and 95% CI in blue.

For example, many primates are chronically exposed to many exogenous lentiviruses through predation [1, 38, 93]. Using log body mass ratios [50] as a proxy for predation, we do not see a statistically significant association between body mass ratio and non-zero transmission rate (Figure 2.4A, blue; $p = 0.678$, coef. 95% CI (-0.311, 0.202)). I believe the

lack of signal is likely due to the imperfect proxy, although it is also possible that predator-prey relationships do not strongly structure the CST network. It is also likely that geographic overlap and habitat similarity are ecological determinants of SIV CST, but modern primate habitats are likely very different since the time that these transmissions actually occurred.

Biologically, increasing host genetic distance has a clear negative association with the probability of cross-species transmission (Figure 2.4B, blue, $p < 0.001$, coef. 95% CI (-7.633, -2.213)). Importantly, as already discussed, the strength of this association may be inflated by instances of lineage tracking (virus/host cospeciation). However, it is well established that increasing host genetic distance is associated with a higher species barrier [17]. As previously documented in the literature, I expect this is due to several factors, such as the divergence of host restriction factor genes, which are key components of the innate immune system (reviewed in [45]) and differences in host cell receptor phenotypes (*e.g.*, [19, 80, 87]). Functional assays of these host phenotypes against panels of SIVs, while outside the scope of this study, will be important for further identifying the molecular bases of the species barriers that have led to the transmission patterns identified here.

2.3.3 Origins of HIV-1 and HIV-2

Epidemiological factors were key to the early spread of HIV

Understanding the underlying dynamics of lentiviral CST provides important ecological context to the transmissions that generated the HIV pandemic. As discussed above, these results support a view of lentiviral cross-species transmission as a rare event. Notably, only two lentiviruses have crossed the high species barrier from Old World monkeys into hominids: SIV*smm* and the recombinant SIV*cpz*. Both HIV-1 and HIV-2 have arisen in human populations in the last century [29, 93]. While it is possible that this has occurred by chance, even without increased primate exposure or other risk factors, it is striking that humans would acquire two exogenous viruses within such a short evolutionary timespan. Examining this phenomenon more closely, the history of HIV-2 is enlightening. HIV-2 has been

acquired through at least 8 independent spillover events from sooty mangabeys [93]. Notably, 6 of these transmissions have resulted in only a single observed infection (spillovers) [21, 20, 33]; only 2 of these events have established sustained transmission chains and successfully switched hosts to become endemic human pathogens [24, 49, 82]. This pattern, as well as serology-based reports of other limited spillovers of SIVs into humans [54, 94], suggest that there have been many isolated introductions of lentiviruses into humans over the past 200,000 years. However, these other viral exposures did not result in new endemic human pathogens either because of species-specific immune barriers, non-conducive epidemiological conditions, or a combination thereof. The rapid and repeated emergence of HIV-1 and HIV-2 is on a timescale more congruent with changes in epidemiological conditions than mammalian evolution, perhaps emphasizing the importance of the concurrent changes in human population structure and urbanization in facilitating the early spread of the epidemic [29]. Significantly, though, this also highlights the importance of careful public health surveillance and interventions to prevent future epidemics of zoonotic viruses.

*The exact origins of SIV*cpz *may not be identifiable*

The ancestry of SIV*cpz* appears to be tripartite: the unknown ancestry of the 5' end; the putatively SIV*rcm* or SIV*mnd*-2 derived *int* and *vif* genes; and the putatively SIV*gsn*/-*mon*/-*mus* derived 3' end of the genome. For any of these three portions of the genome, there are multiple evolutionary histories supported by available sequence data. It is possible that further sampling of lentivirus lineages (both known and currently undiscovered) will be able to definitively resolve the ancestry of SIV*cpz*. Alternatively, it may be that the ancestral virus(es) that gave rise to any of these three portions is extinct [6]. In the case of the last two portions of the SIV*cpz* genome, it may be that the common ancestor of these putative genetic donors (SIV*rcm*/-*mnd*-2 and SIV*mon*/-*mus*/-*gsn*, respectively) was the true source. However, it is also a distinct possibility that SIV*cpz* has sufficiently diverged since its acquisition by chimpanzees such that its history is obscured.

Evolutionary time obscures the identity of the original primate lentivirus

These results clearly illustrate that the vast majority of primate lentiviral lineages were originally acquired by cross-species transmission. It is intriguing to speculate as to which virus was the original source of all of these lineages. Because of its consistent position as the outgroup of primate lentiviral trees, it has been hypothesized [22] that SIV*col* was this original lentivirus among primates. While SIV*col* is certainly the most evolutionary isolated extant lentivirus that has been sampled to date, this does not definitively place it as the ancestral lentivirus. Alternative scenarios (also noted by [6, 22]) include an extinct original lentiviral lineage (and/or primate host species) or an unsampled ancestral lentivirus. It is also plausible that another known extant lentivirus was the original lineage, but has diverged and/or recombined to such an extant that its origins are obscured.

2.3.4 Conclusions

Here, I have shown that lentiviruses have a far more extensive history of host switching than previously described. These results also demonstrate that the propensity of each lentiviral lineage to switch between distant hosts, or to spillover among related hosts, is highly variable. In examining specific lineages, these findings are consistent with the prevalent hypothesis that SIV*col* has evolved in isolation from other SIVs. Contrastingly, I have also demonstrated that the mosaic origins of SIV*cpz* are far more complex than previously recognized; the currently available sequence data is unable to resolve the ancestry of nearly half of the SIV*cpz* genome. Together, these analyses move closer to a full understanding of the pattern of cross-species transmission among primate lentiviruses, but additional efforts to obtain high quality viral genome sequences from under sampled lineages will be necessary to fully resolve the natural history of these viruses.

2.4 Methods

Datasets antibody alignments

Lentiviral genomes are translated in multiple reading frames; I therefore utilized nucleotide sequence data for all analyses.

Recombination analysis dataset

For all recombination analyses, I used the 2015 Los Alamos National Lab (LANL) HIV/SIV Compendium [62]. The compendium is a carefully curated alignment of high-quality, representative sequences from each known SIV lineage and each group of HIV-1 and HIV-2. I reduced the overrepresentation of HIV in this dataset, but maintained at least one high quality sample from each group of HIV. In total, this dataset contains 64 sequences from 24 hosts (1-10 sequences per host). Maximum likelihood trees for each segment – used for the rSPR analysis and displayed in Figure 2.1 — were built with RAxML v.8.2.9 [95] with the rapid bootstrapping algorithm under a GTR model with 25 discrete bins of site to site rate variation, and were rooted to SIV*col* for display.

Main and supplemental datasets

Primate lentiviral sequences were downloaded from the comprehensive LANL HIV sequence database [62]. Sequences from lineages known to be the result of artificial cross-species transmissions (SIV*mac*, *-stm*, *-mne*, and *wcm*) were excluded. I also excluded any sequences shorter than 500 bases or that were flagged as problematic by standard LANL criteria. I grouped host subspecies together except for cases where there is a known specific relationship between host subspecies and virus lineage (chimpanzees and African green monkeys). To construct datasets with a more equitable distribution of sequences per host, I preferentially subsampled sequences from the LANL Compendium, followed by samples isolated from Africa (more likely to be primary sequences), and finally supplementing with samples isolated elsewhere (excluding experimentally generated sequences). For humans, mandrils

and mustached monkeys, which are host to 2, 2 and 3 distinct viral lineages, respectively, I allowed a few additional sequences (if available) to represent the full breadth of documented lentiviral diversity. The main dataset consists of 24 host species, with 5-31 sequences per host (total N=422). As an alternative dataset to control for sampling bias and data availability, I also constructed a supplemental dataset with just 15 hosts but with more viral sequences per host (16 - 40 sequences per lineage, N=510).

Alignments

Alignments were done with the L-INSI algorithm implemented in mafft v7.294b [55]; the compendium alignment was held fixed, with other sequences aligned to this template. Insertions relative to the fixed compendium were discarded. This alignment was then split along the breakpoints identified by GARD to yield the segment-specific alignments. Sequences that contained no bases for a given segment were removed; the 'pruned' dataset is identical to the 'main' dataset, but here I also removed sequences with fewer than 100 ungapped bases for a given segment.

Recombination analyses

Topology-based analysis

Each portion of the genome was analyzed with the Genetic Algorithm for Recombination Detection (GARD), implemented in HyPhy v.2.2.0 [60], with a nucleotide model selected by HyPhys NucModelCompare package (#012234) and general discrete distribution (3 bins) of site variation. Computational intensity was eased by analyzing the recombination dataset in 3kb long portions, with 1kb overlaps on either end (*e.g.*, bases 1:3000, 2000:5000, 4000:7000, etc.). To control for sites' proximity to the ends of the genomic portion being analyzed, this was repeated with the windows offset (*e.g.*, bases 1:2500, 1500:4500, 3500:6500, etc.). In total, this resulted in every site of the alignment being analyzed at least two-fold, with at least one of these replicates providing 500—1500 bases of genomic context on either side

(other than at the very ends of the total alignment). Disagreement between window-offset replicates for a given breakpoint were minimal and almost always agreed within a few bases, with two exceptions: offset replicates for the breakpoint between segments 7 and 8 disagreed by 529 bases, and offset replicates for the breakpoint between 8 and 9 disagreed by 263 bases. In these instances, I used the average site placement.

Sitewise linkage analysis

Biallelic sites were identified across the genome (ignoring gap characters and polymorphisms present at $< 5\%$). These biallelic sites were compared pairwise to generate an observed and expected distribution of haplotypes (combinations of alleles between the two sites), and assessed with the statistic $R^2 = \frac{\chi^2}{d*n}$, where χ^2 is chi-squared, d is the degrees of freedom, and N is the number of haplotypes for this pair of sites [46]. This statistic follows the canonical interpretation of R^2, *i.e.*, if the allele at site 1 is known, how well does it predict the allele at site 2 ($R^2 = 0$ indicating no linkage between sites, and $R^2 = 1$ indicating perfect linkage).

Segment linkage analysis

Segment linkage was assessed by comparing the similarity of tree topologies between segments. This was done with the pairwise method in the Rooted Subtree-Prune-and-Regraft (rSPR) package [101], which measures the number of steps required to transform one topology into another. Segment pairs with similar topologies have lower scores than segments with less similar topologies.

Phylogenies and discrete trait analysis

Empirical tree distributions

For each segment alignment, a posterior distribution of trees was generated with BEAST v. 1.8.2 under a general time reversible substitution model with γ- distributed rate variation and a strict molecular clock [26]. I used a Yule birth-death speciation model tree prior

[35]; all other priors were defaults. Trees were estimated using a Markov chain Monte Carlo (MCMC), and convergence was determined by visually inspecting the trace in Tracer v. 1.6.0 [84].

MCMC chain parameters and ESS values were as follows.

'Main' dataset: 25 million steps, sampled every 15,000 states, 10% burnin removed. Posterior distribution of ≈1600 trees per segment. All but two effective sample size (ESS) values were well over 200.

'Pruned' dataset: 65 million steps, sampled every 15,000 states, 10% burnin removed. Posterior distribution of ≈3900 trees per segment.

'Resampled' dataset: 65 million steps, sampled every 15,000 states, 10% burnin removed. Posterior distribution of ≈3900 trees per segment. All but one ESS value were well above 200.

Discrete trait analysis

These empirical tree distributions were used to estimate the transmission network. As in Faria *et al.* [30], I model hosts as states of a discrete trait, and model cross-species transmission as a stochastic diffusion process among n host states. I use a non-reversible state transition matrix with $n * (n1)$ individual transition parameters [27]. I also utilize standard methodology in using a Bayesian stochastic search variable selection process to encourage a sparse network, implemented as in [66]. An exponential distribution with mean=1.0 was used as a prior on each of the pairwise rate parameters (552 in total for the main and pruned datasets with 24 hosts; 210 parameters in total for the supplemental dataset with 15 hosts). The prior placed on the sum of the binary indicator variables corresponding to each pairwise rate parameter (*i.e.*, the number of transitions that are non-zero) was an exponential distribution with mean=20—23. Convergence was again assessed by visually inspecting the log files in Tracer.

MCMC chain parameters and ESS values were as follows.

'Main' dataset: 25 million steps, sampled every 5,000 states, 5 million states discarded as

burnin.

'Pruned' dataset: decreased phylogenetic noise in the empirical tree distributions for this dataset led to narrow tree distributions and poor mixing (but clear convergence). I thus ran 40 parallel chains of 25—35 million steps each, sampled every 20,000 states (trees sampled every 50,000 states). I discarded 10 million steps of each chain as burnin and concatenated the remaining states.

'Resampled' dataset: 45 million steps, sampled every 5,000 steps, 20 million steps discarded as burnin.

For the main and supplemental datasets, more than 97% of the rate, indicator, and actualRate (step-wise $rate * indicator$ values) parameters, and all other parameters, had an ESS well over 200 (most > 1000). The pruned dataset had a posterior ESS of 677, but the component tree likelihood ESS values ranged from 57 - 287; 91% of the rate, indicator, and actualRate parameters had an average ESS value > 200.

Statistical support for each transmission is summarized as a Bayes factor (BF), calculated by comparing the posterior and prior odds that a given rate is non-zero (*i.e.*, that there has been any transmission between a given pair of hosts) [66]. The ancestral tree likelihoods of each of the 12 tree distributions contribute equally to the inference of a shared transmission rate matrix. However, not every lineage has recombined along every breakpoint, which means that the tree likelihoods from each segment are not fully statistically independent. Thus, I report conservative estimates of BF by dividing all BF values by 12.

Chapter 3

DENGUE VIRUS ANTIGENIC EVOLUTION

This chapter is adapted from a manuscript, coauthored with Drs. Leah Katzelnick and Trevor Bedford, which is currently under review for publication [14].

3.1 Introduction to dengue virus

Dengue virus (DENV) is a mosquito-borne flavivirus which consists of four genetically distinct clades, canonically thought of as serotypes [63]. DENV circulates primarily in South America and Southeast Asia, infecting 400 million people annually. While most of these infections are asymptomatic or mild, ∼1–3% of cases progress to severe dengue hemorrhagic fever, causing approximately 10—15,000 deaths each year [15]. Unlike most infectious diseases, DENV secondary infections are more likely than primary infections to cause severe disease, with estimates of relative risk of severe dengue of ∼24 [75]. Primary DENV infection is more often mild and is thought to generate lifelong homotypic immunity and temporary heterotypic immunity, which typically wanes over six months to two years [58, 89, 85]. Subsequent heterotypic secondary infection induces broad cross-protection, and symptomatic tertiary and quaternary cases are rare [36, 79]. However, a small subset of secondary infections are enhanced by nonneutralizing, cross-reactive antibodies, resulting in severe disease [44, 57, 91]. Thus, the antigenic relationships between dengue viruses — describing whether the immune response generated after primary infection results in protection or enhancement of secondary infection — are key drivers of DENV case outcomes and epidemic patterns.

While each serotype is clearly genetically and antigenically distinct, it is not clear how sub-serotype clades of DENV interact antigenically. Each DENV serotype consists of broad genetic diversity, including canonical clades termed 'genotypes' (Figure 3.1) [86, 96]. Specific

genotypes have been associated with characteristically mild or severe disease, and heterogeneous neutralization titers suggest that the immune response to some genotypes is more cross-protective than others [34, 88]. Until recently, it has been assumed that these intraserotype differences are minimally important compared to interserotype differences. However, empirical evidence has demonstrated that these genotype-specific differences can drive case outcomes and epidemic severity (reviewed in Holmes and Twiddy, 2003). For example, analysis of a longitudinal cohort study demonstrated that specific combinations of primary infection serotype and secondary infection genotype can mediate individual case outcomes [78]. On a population scale, the DENV1-immune population of Iquitos, Peru, experienced either entirely asymptomatic or very severe epidemic seasons in response to two different genotypes of DENV2 [59].

One explanation for these and similar observations is that overlooked intraserotype antigenic variation contributes to these genotype-specific case outcomes and epidemic patterns. Recent efforts to antigenically characterize diverse dengue viruses suggests that each serotype may contain antigenic heterogeneity, but the source and impact of this heterogeneity is not clear [56]. Here, I take a phylogenetic approach to characterize the evolutionary basis for observed antigenic heterogeneity among DENV clades. I also quantify the impact of within- and between-serotype antigenic variation on real-world DENV population dynamics.

3.2 Results

3.2.1 Measuring antigenic relationships between dengue viruses

Antigenic distance between a pair of viruses i and j is experimentally quantified using neutralization titers, which measures how well serum drawn after infection with virus i is able to neutralize virus j *in vitro* [88]. To measure the pairwise antigenic distances for a panel of diverse dengue viruses (Figure B.1), Katzelnick *et al.* infected naive non-human primates (NHP) with each virus, drew sera at three months post-infection, and then titered this sera against a panel of test viruses [56]. To compare patterns of cross-protection in NHP and

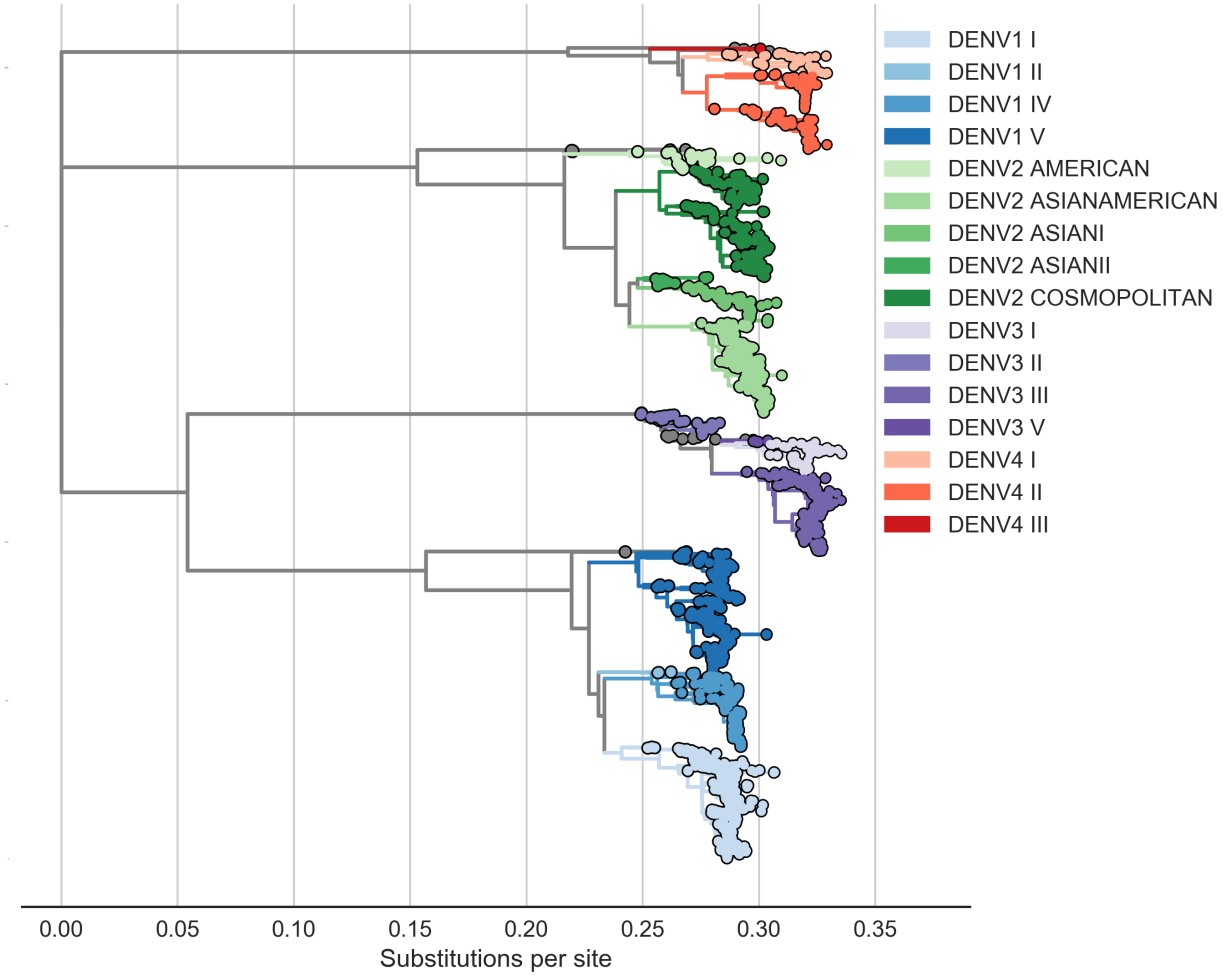

Figure 3.1. Phylogeny of dengue viral sequences Maximum likelihood phylogeny of dengue virus genomes, colored by canonical genotype assignment.

humans, they also drew sera from 31 study participants six weeks after inoculation with a monovalent component of the NIH dengue vaccine candidate. This sera was also titered against a broad panel of dengue viruses. As originally reported, I find generally consistent patterns of neutralization between the NHP and human sera data; see [56] for a detailed comparison. In total, this dataset consists of 454 NHP sera titrations spanning the breadth of DENV diversity, and 728 human sera titrations providing deep coverage of a small subset of viruses.

To standardize these measurements, I first take the $\log_2$ of each value, such that one titer unit corresponds to one, two-fold drop in neutralization. I then define antigenic distance between autologous virus-sera pairs (*i.e.*, virus i and serum i) as zero. Normalized antigenic distance between i and j are thus calculated as $D_{ij} = \log_2(T_{ii}) - \log_2(T_{ij})$, such that a higher value of D_{ij} indicates that serum i is less effective at neutralizing virus j, implying greater antigenic distance between viruses i and j.

The full dataset of standardized titer values is shown in Figure 3.2. Here, we see that homotypic virus-serum pairs are more closely related antigenically than heterotypic pairs. However, we also observe large variance around this trend, both within and between serotypes. This suggests that treating each serotype as antigenically uniform overlooks potentially important antigenic heterogeneity across viruses within each serotype.

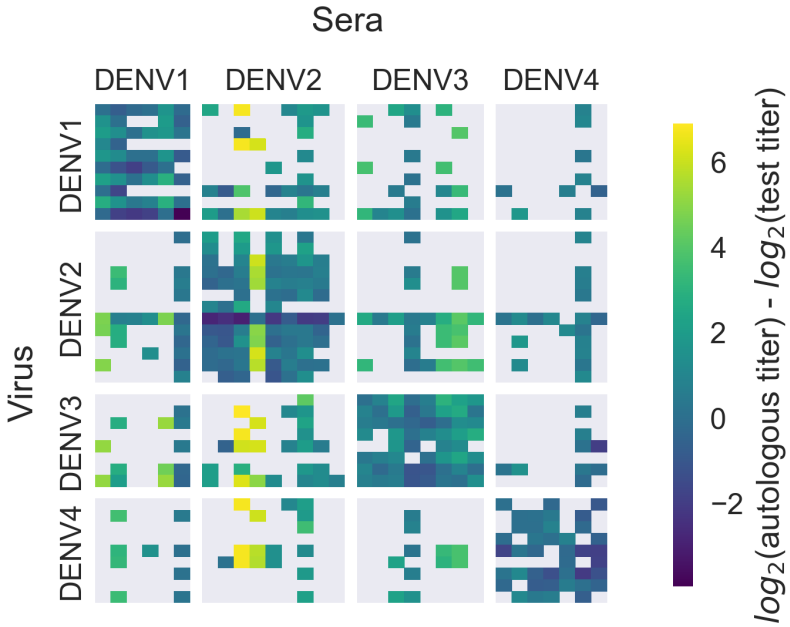

Figure 3.2. Normalized antigenic distance between pairs of dengue viruses and sera Aggregated plaque reduction neutralization titers from Katzelnick *et al.* are standardized such that the distance between autologous virus-serum pairs is 0, and each titer unit corresponds to one, two-fold change in PRNT50 value (see Methods). Light gray areas represent missing data. Higher values correspond to greater antigenic distance.

3.2.2 Mapping dengue antigenic evolution to phylogenetic divergence

Titer measurements are prone to noise, and there is a limited amount of available titer data. If the antigenic heterogeneity observed in the raw data is truly the result of an underlying evolutionary process, I expect that changes in antigenic phenotype correspond to underlying changes in viral genotype.

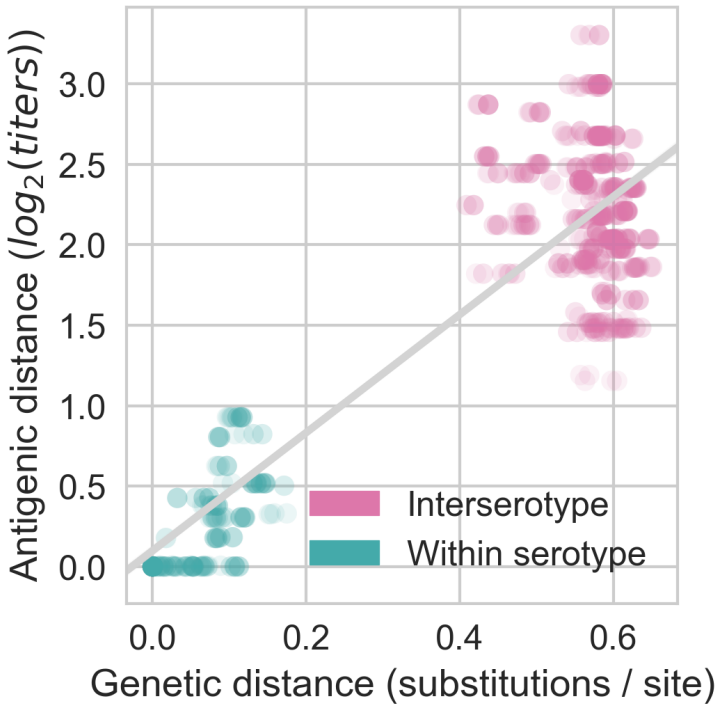

Figure 3.3. Normalized antigenic distance vs. genetic distance between pairs of dengue viruses Antigenic distances are the same as from Figure 3.2; genetic distances are the patristic distances between viral genomes on a maximum likelihood phylogeny.

Figure 3.3 shows the relationship between genetic and antigenic distance between each pair of viruses in this dataset. There are two groups of comparisons — between serotype and within serotype — however, even within serotypes there is significant genetic diversity and a correlation between increased genetic distance and increased antigenic distance. The relationship between genetic distance and antigenic distance is consistent within and between

serotypes, where increasing genetic divergence generally corresponds to increased antigenic distance.

3.2.3 Within-serotype antigenic evolution

To fully map the relationship between DENV genetic and antigenic evolution, I adapt a phylogeny-based model originally developed for influenza [77]. Conceptually, this model predicts titer values through three steps. First, I build a phylogeny of dengue virus sequences to establish the genetic relationships between viruses (Figures 3.1, B.2). Next, I infer how much antigenic change has occurred along each branch of the phylogeny by mapping titer changes to individual branches. This assigns each branch b an antigenic distance d_b. With this in hand, I estimate the antigenic distance between all pairs of viruses by summing branch-specific distances d_b for each branch along the path between them in the phylogeny (Methods, Eq. 3.2).

To learn these values of d_b, I first split the dataset into training (random 90% of measurements) and test data (the remaining 10% of values). I take the training data and fit d_b for each branch in the tree, subject to regularization as follows (also detailed in Methods, Eq. 3.3). Parsimoniously, I expect that antigenic change is more likely to occur through larger changes on a few branches than through small changes on many branches; correspondingly, the prior expectation of values of d_b is exponentially distributed such that most values of $d_b = 0$. This is analogous to lasso regression to identify a few parameters with positive weights and set other parameters to 0. Additionally, some viruses have greater binding avidity, and some sera are more potent than others (Figure B.3); these 'row' and 'column' effects, respectively, are normally distributed and are taken into account when estimating titers. The model uses convex optimization to learn the values of d_b that minimize the sum of squared errors (SSE) between observed and predicted titers in the training data.

This model formulation is an effective tool for estimating antigenic relationships between viruses based on their relative positions in the phylogeny. Variations of this model can be used to explicitly test whether the observed antigenic phenotypes are better explained by the

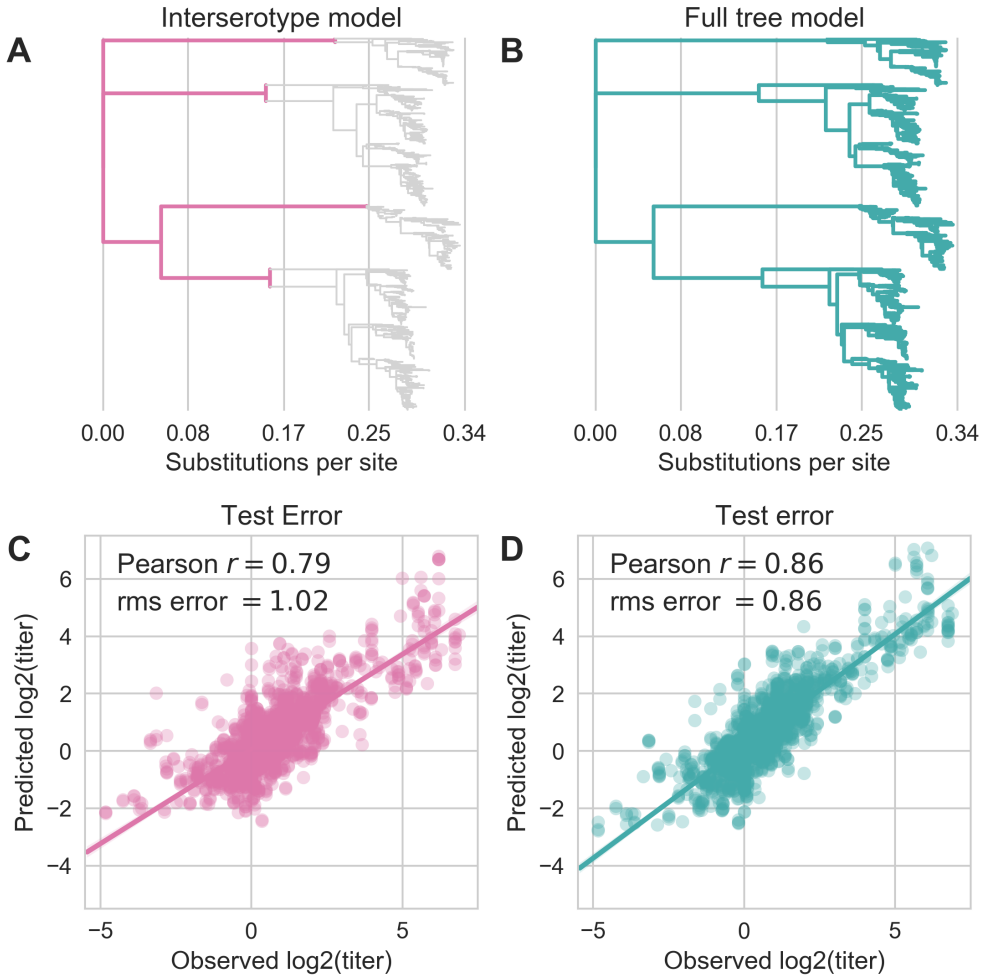

Figure 3.4. Titer model formulations and performance **A** The 'interserotype model' only allows branches that lie between serotypes to contribute to antigenic evolution. All other branches are assigned $d_b = 0$. **B** The 'full tree model' allows any branch in the phylogeny to contribute to antigenic evolution ($d_b \geq 0$). **C,D** Predictive performance of each model on the test dataset (aggregated from 10-fold cross-validation).

hypothesis that dengue serotypes are antigenically uniform ('interserotype model'), or by the hypothesis that serotypes are antigenically diverse ('full tree model'). In the interserotype model, $d_b = 0$ for all branches in the tree that do not lie between serotypes (Figure 3.4A). Alternatively, the 'full tree model' allows any branch in the phylogeny to contribute to antigenic evolution (Figure 3.4B). For each model, parameters are learned from the training

data, and then used to predict test data values. Model performance is assessed by comparing the predicted test titer values to the actual values, and indicates how well the hypothesis embedded in the model explains the observed data.

Serotype-level characterization alone explains observed antigenic phenotypes to a reasonable degree. On average, this interserotype model predicts titers within 1.02 $\log_2$ titer units of the true value (root mean squared error, RMSE), and explains 62% of the observed variation in neutralization titers overall (Figure 3.4).

However, accounting for within-serotype antigenic evolution substantially improves the model's ability to explain dengue antigenic phenotypes, as estimated by cross-neutralization titers. The full tree model is able to predict test titers within 0.86 $\log_2$ titer units of the true value (RMSE approaching the level of error intrinsic to the assay), and explains 74% of the observed variation in neutralization titers overall (Figure 3.4). Importantly, all reported error metrics refer to performance on test data, so this difference in model performance is not due to the number of free parameters. The full tree model performance is comparable to the model error from a cartography-based characterization of the same dataset (RMSE 0.65—0.8 $\log_2$ titer units), and to the error observed when this model was used to characterize an influenza dataset (RMSE of 0.5 $\log_2$ titer units) [56, 77]. From this, I conclude that there is antigenic evolution within each serotype of DENV, and that this is driven by underlying genetic divergence.

Collapsing putatively antigenically uniform clades, we observe at least 12 distinct antigenic phenotypes of DENV (Figure 3.5), rejecting the null hypothesis that all dengue viruses within a serotype are antigenically uniform. The titer dataset spans the breadth of canonical DENV genotypes, but in most cases lacks the resolution to detect within-genotype antigenic diversity. I thus expect that these results represent a lower-bound on the true extent of DENV intraserotype antigenic diversity.

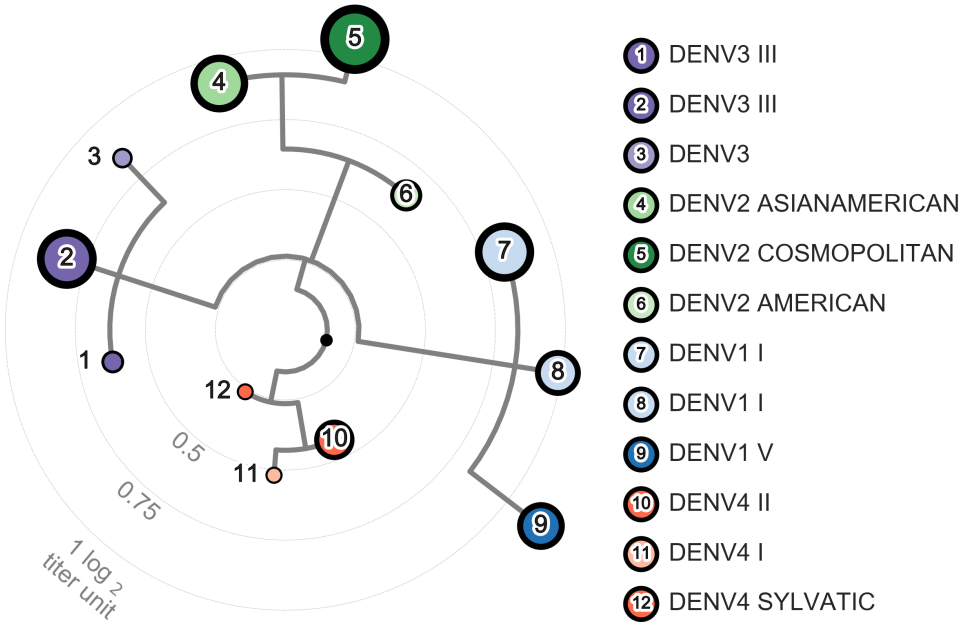

Figure 3.5. Tree of dengue antigenic phenotypes The topology is inferred from a maximum likelihood phylogeny of DENV genomes. Branch lengths are scaled to reflect d_b, the relative contributions of each branch to antigenic divergence, as inferred by the 'full tree' model of DENV evolution. Antigenically uniform clades are collapsed, and node diameter reflects the size of the collapsed clade. Figure B.4 offers an additional view of the same topology.

3.3 Discussion

3.3.1 Breadth of dengue antigenic diversity

These results show that mapping antigenic change to individual branches and interpolating across the DENV phylogeny is able to explain a large majority of the observed variation in antigenic phenotypes, as measured by neutralization titers. This demonstrates that DENV antigenic divergence is closely coupled to genetic divergence. Additionally, within-serotype antigenic evolution better explains the observed variation in antigenic phenotypes. This supports and expands upon previous reports [56, 99, 31] that the null hypothesis of antigenically uniform serotypes is inconsistent with observed patterns of cross-protection and susceptibility. We observe at least 12 distinct antigenic phenotypes that are both genetically

and antigenically distinct. Analysis of the recent CYD-TDV vaccine trial shows different vaccine efficacy against genotypes I and II of DENV4 [53]; the antigenic distance between these two genotypes is comparable to the antigenic distance separating each of the 12 antigenic phenotypes reported here (Figures 3.5, B.5). This suggests that these phenotypes may be sufficiently distinct to have important impacts on secondary case outcomes and vaccine efficacy.

3.3.2 Titer model strengths and limitations

Overall, I expect that these antigenic phenotypes represent a lower-bound on the extent, magnitude, and nature of antigenic heterogeneity with DENV. The current titer dataset spans the breadth of DENV diversity, but due to small sample size, it lacks the resolution to detect most sub-genotype antigenic variation or to identify which specific mutations correspond to changes in antigenic phenotype. The appearance of the deep antigenic divergence of the four serotypes, and the more recent antigenic divergences within each serotype, suggest that DENV antigenic evolution is likely an ongoing, though gradual, process. I therefore expect that future studies with richer datasets will find additional antigenic variation within each genotype. This dataset also contains many left-censored titer values, where we know two viruses are at least T titer units apart, but do not know exactly how far apart. If we knew the true value of these censored titers, many of them would indicate larger antigenic distances than the reported values, T, which are used to train the model. Thus, it is likely that this model systematically underestimates the magnitude of titer distances. Finally, antibody neutralization and escape (as measured by PRNT titers) is only one component of the immune response to DENV. Although analysis of a longitudinal cohort study shows that these neutralization titers correlate with protection from severe secondary infection, it is unclear how PRNT titers correspond to antibody-dependent enhancement [58]. It is also important to note that DENV case outcomes are partially mediated by interactions with innate and T-cell immunity, the effects of which are not captured in neutralization titers [39]. Overall, while richer datasets and the development of more holistic assays will be required

in order to fully characterize the extent of DENV antigenic diversity, it is clear that the four-serotype model is insufficient to explain DENV antigenic evolution.

3.4 Methods

Sequence Data

I downloaded all dengue virus sequences available from the Los Alamos National Lab Hemorrhagic Fever Virus Database as of March 7, 2018, that contained the full coding sequence of *E (envelope)* (total N=12,645) [61]. I discarded sequences which were putative recombinants, duplicates, lab strains, or which lacked an annotated sampling location and/or sampling date. I then randomly subsampled up to 8 viruses per region, per month, preferentially including records with available titer data and longer sequences. The final dataset consists of 2,563 viral sequences (Figure B.2). The annotated reference dataset from [83] was used to assign sequences to canonical genotypes.

Titer Data

Antigenic distance between pairs of viruses i and j is experimentally measured using a neutralization titer, which measures how well serum drawn after infection with virus i is able to neutralize virus j *in vitro* [88]. Briefly, two-fold serial dilutions of serum i are incubated with a fixed concentration of virus j. Titers represent the lowest serum concentration able to neutralize 50% of virus, and are reported as the inverse dilution. I used two publicly available plaque reduction neutralization titer (PRNT50) datasets generated by Katzelnick *et al.* in [56]. The primary dataset was generated by infecting each of 36 non-human primates with a unique strain of DENV. NHP sera was drawn after 12 weeks and titered against the panel of dengue viruses. The secondary dataset was generated by vaccinating 31 human trial participants with a monovalent component of the NIH DENV vaccine. Sera was drawn after 6 weeks and titered against the same panel of dengue viruses. As discussed in Katzelnick *et al.*, these two datasets show similar patterns of antigenic relationships between dengue viruses. In total, my dataset includes 47 virus strains, 36 serum strains, and 1182 measurements.

Titer Model

I compute standardized antigenic distance between virus i and serum j (denoted D_{ij}) from measured titers relative to autologous titers (denoted T_{ii} and T_{ij}, respectively), such that

$$D_{ij} = \log_2(T_{ii}) - \log_2(T_{ij}) \tag{3.1}$$

To predict unmeasured titers, I employ the 'tree model' from Neher *et al.* and implemented in Nextstrain, which assumes that antigenic evolution is driven by underlying genetic evolution [43, 77]. This pipeline uses RAxML version 8.2.11 and a GTRCAT nucleotide substitution model to build a maximum likelihood phylogeny of dengue viral genome sequences [95]. Observed titer drops are mapped to branches in the viral phylogeny after correcting for overall virus avidity, v_i, and serum potency, p_j ('row' and 'column' effects, respectively):

$$\hat{D}_{ij} \approx D_{ij} = \sum_{b \in path(i,j)} d_b + v_i + p_j \tag{3.2}$$

where d_b is the titer drop assigned to each branch, b, in the phylogeny. A random 10% of titer measurements are withheld as a test set. The remaining 90% of titer measurements are used as a training set to learn values for virus avidity, serum potency, and branch effects. As in Neher *et al.*, I formulate this as a convex optimization problem and solve for these parameter values to minimize the cost function:

$$C = \sum_{i,j} (\hat{D}_{ij} - D_{ij})^2 + \lambda \sum_b d_b + \gamma \sum_i v_i^2 + \delta \sum_j p_j^2 \tag{3.3}$$

Respectively, these terms represent the squared training error; an L1 regularization term on branch effects, such that most values of $d_b = 0$; and L2 regularization terms on virus avidities and serum potencies, such that they are normally distributed. These parameter values are then used to predict the antigenic distance between all pairs of viruses, i and j, in the phylogeny. I assess performance by comparing predicted to known titer values in the test data set, and present test error (aggregated from 10-fold cross-validation) throughout the manuscript.

Chapter 4

HUMAN POPULATION IMMUNITY AND DENGUE VIRUS CLADE DYNAMICS

This chapter is adapted from a manuscript, coauthored with Drs. Leah Katzelnick and Trevor Bedford, which is currently under review for publication [14]. *It extends work described in Chapter 3.*

4.1 Introduction to antigenic fitness

Based on the results from Chapter 3, we observe strong evidence that homotypic genotypes of dengue virus vary in their ability to escape antibody neutralization (Figures 3.4, B.5). However, antibody neutralization is only one of many factors that shape epidemic patterns. Thus, I next investigated whether this observed antigenic diversity influences dengue population dynamics in the real world.

The size of the viral population (*i.e.*, prevalence, commonly analyzed using SIR models as reviewed in [72]) is determined by many complex factors, and reliable values for population prevalence are largely unavailable. Contrastingly, the composition of the viral population (*i.e.*, the relative frequency of each viral clade currently circulating) can be estimated over time by examining historical sequence data [64, 77], and is primarily driven by viral fitness [10]. In meaningfully antigenically diverse viral populations, antigenic novelty (relative to standing population immunity) contributes to viral fitness: as a given virus i circulates in a population, the proportion of the population that is susceptible to infection with i—and other viruses antigenically similar to i—decreases over time as more people acquire immunity [11, 73, 42]. Antigenically novel viruses that are able to escape this population immunity are better able to infect hosts and sustain transmission chains, making them fitter than

the previously circulating viruses [103, 11]. Thus, if antigenic novelty constitutes a fitness advantage for DENV (e.g., as suggested by [71, 100]), then I would expect greater antigenic distance from recently circulating viruses to correlate with higher growth rates.

4.2 Results

4.2.1 Antigenic novelty and dengue serotype turnover

To test this hypothesis, I examine the composition of the dengue virus population in Southeast Asia from 1970 to 2015. I estimate the relative population frequency of each DENV serotype at three month intervals, $x_i(t)$ (Figure 4.1A), based on available sequence data (Methods, Eq. 4.1).

Fitter virus clades increase in frequency over time, such that $x_i(t+dt) > x_i(t)$. It follows that these clades have a growth rate—defined as the fold-change in frequency over time—greater than one: $\frac{x_i(t+dt)}{x_i(t)} > 1$. To isolate the extent to which antigenic fitness contributes to clade success and decline, I extend work by Łuksza and Lässig [73] to build a simple model that attempts to predict clade growth rates based on two variables: the antigenic fitness of the clade at time t, and a time-invariant free parameter representing the intrinsic fitness of the serotype the clade belongs to. I estimate the antigenic fitness of clade i at time t as a function of its antigenic distance from each viral clade j that has circulated in the same population over the previous two years, weighted by the relative frequency of j and adjusted for waning population immunity (Figure 4.1B; Methods, Eq. 4.4). Growth rates are estimated based on a five year sliding window (Figure 4.1C).

This simple model explains 62% of the observed variation in serotype growth rates, and predicts serotype growth vs. decline correctly for 80% of predictions (Figure 4.1D). This strongly suggests that antigenic fitness is a major driver of serotype population dynamics. This also demonstrates that this model captures key components of dengue population dynamics; examining the formulation of this model in more detail can yield insights into how antigenic relationships influence DENV population composition. The fitness model includes

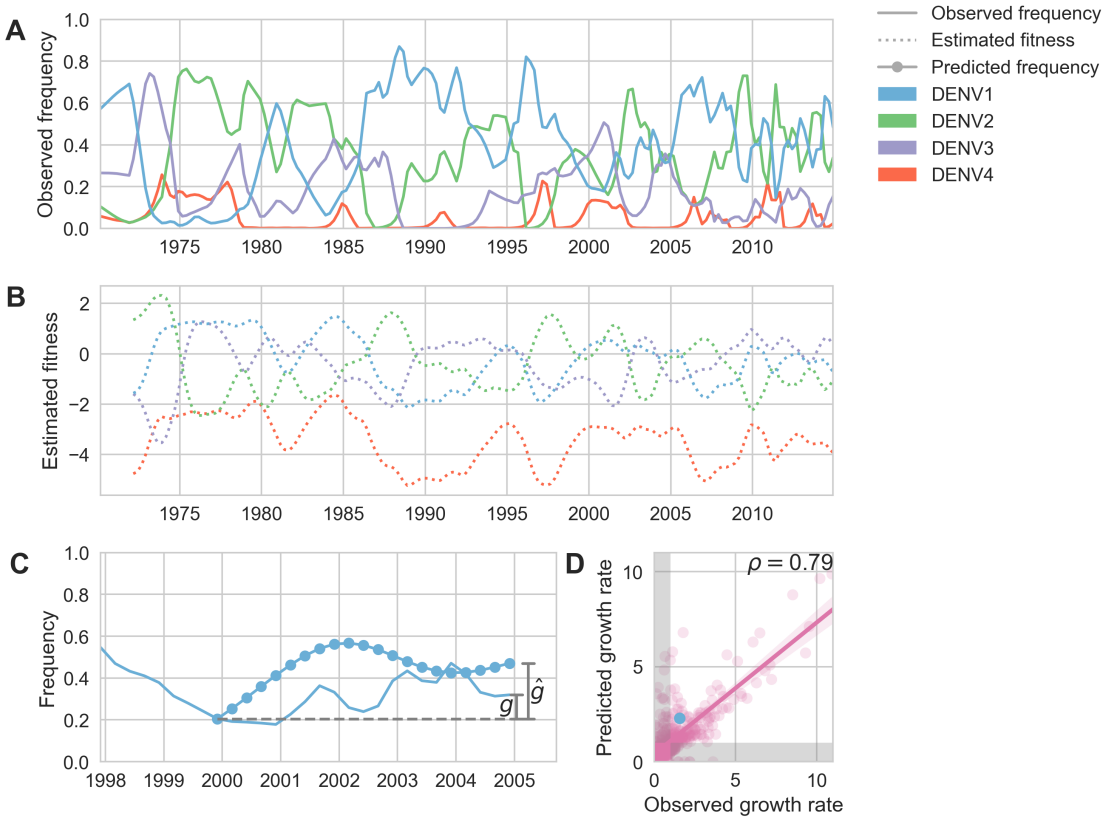

Figure 4.1. Antigenic novelty predicts serotype success **A** The relative frequency of each serotype, x_i, in Southeast Asia is estimated every three months based on available sequence data. **B** Antigenic fitness for each serotype over time is estimated as its frequency-weighted antigenic distance from recently circulating viruses. This is added to a time-invariant intrinsic fitness value to calculate total fitness (shown here, arbitrary units). **C** illustrates how the model predicts clade growth rates, with DENV1 from 1998 — 2005 as an example. At each timepoint t, the model is blinded to all empirical data from timepoints later than t and predict each serotype's future trajectory based on its initial frequency, time-invariant intrinsic fitness, and antigenic fitness at time t (Methods, Eq. 4.7). I predict forward in three-month increments for a total prediction period of $dt = 5$ years. At each increment, the model uses the predicted stepwise frequency change to adjust estimates of antigenic fitness on a rolling basis (Methods, Eq. 4.11). Predicted growth rates are calculated as $\hat{g} = \frac{\hat{x}_i(t+dt)}{x_i(t)}$ and compared to empirically observed growth rates, $g = \frac{x_i(t+dt)}{x_i(t)}$ in **D**. The example illustrated in **C** is also shown in **D** as the blue point. Serotype growth versus decline is accurate (*i.e.*, the predicted and actual growth rates are both > 1 or both < 1, all points outside the gray area) for 80% of predictions.

eight free parameters that are optimized such that the model most accurately reproduces the observed fluctuations in DENV population composition (reduction in prediction error relative to a null model, see Methods). I find that serotype fluctuations are most consistent with a model wherein population immunity wanes linearly over time, with the probability of protection dropping by about 48% per year for the first two years after primary infection. I also find that these dynamics are best explained by intrinsic fitness that moderately varies by serotype (Table B.1).

4.2.2 Antigenic fitness and genotype dynamics

To estimate how well antigenic fitness predicts genotype dynamics, I used the same model to predict genotype success and decline. As before, fitness of genotype i is based on the intrinsic fitness of the serotype i belongs to, and the antigenic distance between i and each other genotype, j, that has recently circulated (Figure 4.2B). Importantly, I can calculate antigenic distance between i and j at the serotype level (*i.e.*, the antigenic distances computed from the 'interserotype model' as illustrated in Figure 3.4A) or at the genotype level (*i.e.*, the antigenic distances computed by the 'full tree model' as illustrated in Figure 3.4B, which incorporates the observed within-serotype heterogeneity). If within-serotype antigenic heterogeneity contributes to genotype fitness, then I would expect estimates of antigenic fitness based on the 'full tree model' to better predict genotype growth rates.

I find that antigenic fitness contributes to genotype turnover, although it explains less of the observed variation than for serotypes. When antigenic distance is estimated from the 'interserotype model', I find that this model of antigenic fitness explains approximately 31% of the observed variation in genotype growth rates, and correctly predicts genotype growth vs. decline 68% of the time (Figure 4.2C). Perhaps surprisingly, more precise estimates of antigenic distance between genotypes from the 'full tree model' do not improve model predictions of genotype success ($R^2 = 0.28$, 70% accuracy; Figure 4.2D). This suggests that although we see strong evidence that genotypes vary in their ability to escape neutralizing antibodies (Figures 3.4, B.5, Chapter 3), these differences are subtle enough that they do

not impact broad-scale regional dynamics over time.

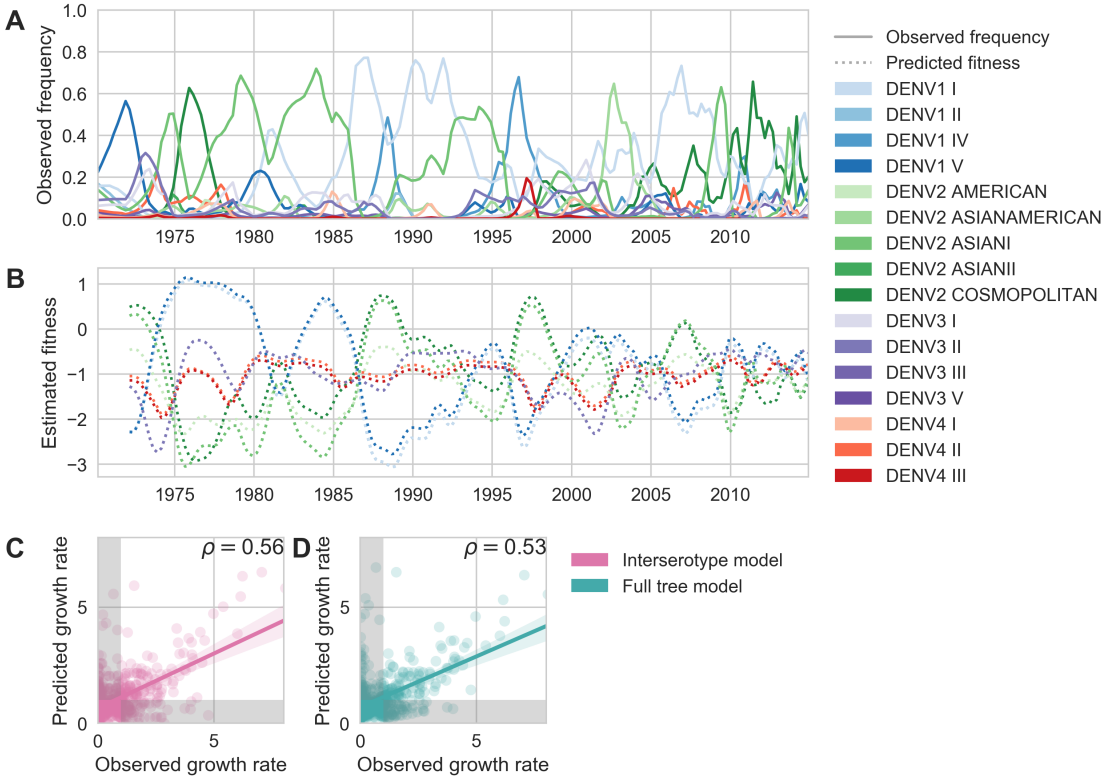

Figure 4.2. Antigenic novelty partially predicts genotype success **A** Relative frequencies of each canonical dengue genotype across Southeast Asia, estimated from available sequence data. **B** Antigenic fitness is calculated for each genotype as its frequency-weighted antigenic distance from recently circulating genotypes. This is added to a time-invariant, serotype-specific intrinsic fitness value to calculate total fitness (shown here, arbitrary units). Antigenic distance is estimated using either the 'full tree model' or the 'interserotype model' of antigenic relationships (Figure 3.4). This panel shows total fitness over time, incorporating estimates of antigenic fitness derived from the 'full tree' model. **C, D** Fitness estimates were used to predict clade growth rates over 5 years, compounding immunity every three months based on predicted frequency changes (Methods Eq. 4.11). Here, I compare observed vs. predicted growth rates for both formulations of the fitness model (using fitness derived from either the 'interserotype' or 'full tree' model estimates of antigenic distance). Growth versus decline was accurate (predicted and actual growth rates both > 1 or both < 1, points outside the gray shaded area) for 68% – 70% of predictions, respectively.

4.3 Discussion

4.3.1 Fitness model strengths and limitations

I use viral antigenic relationships inferred in Chapter 3 to directly quantify the impact of antigenic fitness on DENV population composition. To do so, I measure serotype frequencies across Southeast Asia over time and construct a model to estimate how they will fluctuate (Methods, Eq. 4.3–4.16). This model places a fitness value on each serotype that derives from a constant intrinsic component alongside a time-dependent antigenic component. Antigenic fitness declines with population immunity, which is accumulated via the recent circulation of antigenically similar viruses. I find that antigenic fitness is able to explain most of the observed variation in serotype growth and decline (Figure 4.1). This demonstrates that antigenic fitness is a strong determinant of DENV serotype dynamics in a real-world, hyper-endemic population.

I similarly use this model to quantify the effect of within-serotype antigenic variation on the success and decline of canonical DENV genotypes (Figure 4.2). As above, genotype antigenic fitness declines with population immunity. Here, I estimate population immunity based on antigenic distance from recently circulating genotypes, using distances derived from the 'interserotype model' or the 'full tree model' of DENV antigenic evolution. I then directly compare how strongly these coarser serotype-level versus specific genotype-level antigenic relationships impact DENV population dynamics. Overall, I find that antigenic fitness explains a moderate portion of the observed variation in genotype growth and decline. Surprisingly, however, I find that incorporating within-serotype antigenic differences does not improve these predictions (Figure 4.2C,D). This suggests that although genotypes are antigenically diverse, these differences do not appear to influence large-scale regional dynamics over time.

However, this observation is subject to caveats imposed by the available data and model assumptions. My estimates of antigenic fitness are informed by the antigenic distances inferred by the titer tree model; thus, as discussed in Section 3.3, I am unable to account for nuanced antigenic differences between sub-genotype clades of DENV due to limited titer

data. I estimate DENV population composition over time based on available sequence data, pooled across all of Southeast Asia (Methods, Eq. 4.1). As the vast majority of cases of DENV are asymptomatic, sequenced viruses likely represent a biased sample of more severe cases from urban centers where patients are more likely to seek and access care. I also assume that Southeast Asia represents a closed viral population with homogeneous mixing. However, increasing globalization likely results in some amount of viral importation that is not accounted for in this model [5]. Finally, although Southeast Asia experiences hyperendemic DENV circulation, the majority of DENV transmissions are hyper-local [90], and viral populations across this broad region may not mix homogeneously each season. Thus, it is possible that these sub-serotype antigenic differences impact finer-scale population dynamics, but I lack the requisite data to examine this hypothesis.

4.3.2 Conclusions

As presented in Chapters 3 and 4, I find that within-serotype antigenic evolution explains the observed patterns of cross-neutralization among dengue genotypes. These within-serotype differences are large enough that I believe they likely impact secondary case outcomes and small scale population dynamics. I also find that population immunity is a strong determinant of the composition of the DENV population across Southeast Asia, although this is putatively driven by coarser, serotype-level antigenic differences. As richer datasets become available, future studies that similarly combine viral genomics, functional antigenic characterization, and population modeling have great potential to improve our understanding of how DENV evolves antigenically and moves through populations.

4.4 Methods

Datasets

See section 3.4 for a description of the titer and sequence datasets.

Empirical Clade Frequencies

As discussed in Neher *et al* and Lee *et al*, empirical clade frequencies are estimated from 1970 to 2015 based on observed relative abundance of each clade in the 'slice' of the phylogeny corresponding to each quarterly timepoint [64, 77].

Briefly, the frequency trajectory of each clade in the phylogeny is modeled according to a Brownian motion diffusion process discretized to three-month intervals. Relative to a simple Brownian motion, the expectation includes an 'inertia' term that adds velocity to the diffusion and the variance includes a term $x(1-x)$ to scale variance according to frequency following a Wright-Fisher population genetic process. This results in the following diffusion process:

$$x(t+dt) = \mathcal{N}\left(x(t) + \epsilon dx,\ dt\sigma^2 x(t)(1-x(t))\right) \tag{4.1}$$

with 'volatility' parameter σ^2. The term dx is the increment in the previous timestep, so that $dx = x(t) - x(t-dt)$. I used $\epsilon = 0.7$ and $\sigma = 2.0$ to maximize fit to empirical trajectory behavior.

These estimates also include an Bernoulli observation model for clade presence / absence among sampled viruses at timestep t. This observation model follows

$$f(x,t) = \Pi_{v \in V} x(t)\ \Pi_{v \notin V}(1-x(t)) \tag{4.2}$$

where $v \in V$ represents the set of viruses that belong to the clade and $v \notin V$ represents the set of viruses that do not belong to the clade. Each frequency trajectory is estimated by simultaneously maximizing the likelihood of the process model and the likelihood of the observation model via adjusting frequency trajectory $\vec{x} = (x_1, ...x_n)$.

Population Immunity

For antigenically diverse pathogens, antigenic novelty represents a fitness advantage [68]. This means that viruses that are antigenically distinct from previously-circulating viruses are able to access more susceptible hosts, allowing the antigenically novel lineage to expand. I adapt a simple deterministic model from Łuksza and Lässig to directly quantify dengue antigenic novelty and its impact on viral fitness [73]. I quantify population immunity to virus i at time t, $P_i(t)$, as a function of which clades have recently circulated in the past N years, and how antigenically similar each of these clades is to virus i:

$$P_i(t) = \sum_{n=1}^{n=N} \left(w(n) \sum_j \left(x_j(t-n)\, C(D_{ij}) \right) \right) \tag{4.3}$$

Where D_{ij} is the antigenic distance between i and each non-overlapping clade j, n is the number of years since exposure, and $x_j(t-n)$ is the relative frequency of j at year $t-n$. Waning immunity is modeled as a non-negative linear function of time:

$$w(n) = max(-\gamma n + 1, 0) \tag{4.4}$$

The relationship between antigenic distance and the probability of protection, C, is also assumed to be non-negative and linear with slope $-\sigma$, such that:

$$C(D_{ij}) = max(-\sigma D_{ij} + 1, 0) \tag{4.5}$$

I model the effects of population immunity, $P_i(t)$, on viral antigenic fitness, $f_i(t)$, as:

$$f_i(t) = f_0 - \beta P_i(t) \tag{4.6}$$

where β and f_0 are fit parameters representing the slope of the linear relationship between immunity and fitness, and the intrinsic relative fitness of each serotype, respectively.

Frequency Predictions

Similar to the model implemented in Łuksza and Lässig, I estimate predicted clade frequencies at time $t + dt$ as

$$\hat{x_i}(t+dt) = \frac{x_i(t) e^{f_i(t)dt}}{\sum_i x_i(t) e^{f_i(t)dt}} \tag{4.7}$$

for short-term predictions (where $dt < 1$ year).

For long-term predictions, we must account for immunity accrued at each intermediate timepoint between t and dt. I divide the interval between t and dt into a total of U 3 month timepoints, $[t+u, t+2u, ...t+U]$, such that $t+U = dt$. I then compound immunity based on predicted clade frequencies at each intermediate timepoint:

$$\hat{x_i}(t+u) = x_i(t)e^{f_i(t)u} \tag{4.8}$$

$$\hat{x_i}(t+2u) = \hat{x_i}(t+u)e^{f_i(t+u)u} \tag{4.9}$$

...

$$\hat{x_i}(t+U) = x_i(t)e^{f_i(t)u}e^{f_i(t+u)u}e^{f_i(t+2u)u}...e^{f_i(t+U)u} \tag{4.10}$$

$$\hat{x_i}(t+dt) = \hat{x_i}(t+U) = x_i(t)e^{\sum_u f_i(t+u)u} \tag{4.11}$$

I then calculate clade growth rates, defined as the fold-change in relative clade frequency between time t and time $t+dt$:

$$\frac{\hat{x_i}(t+dt)}{x_i(t)} \tag{4.12}$$

Null model and model performance

To quantify the impact of antigenic fitness on DENV clade success, we can compare the antigenically-informed model to a null model wherein all viruses have equal antigenic fitness at all timepoints:

$$f_i^{null}(t) = 0 \tag{4.13}$$

$$\hat{x_i}^{null}(t+dt) = x_i(t)e^0 = x_i(t) \tag{4.14}$$

For both the null model and the antigenically-informed model, we can assess predictive power as the sum of squared error between predicted and empirical clade frequencies:

$$SSE = \sum_{i,t} \left(\hat{x_i}(t+dt) - x_i(t+dt)\right)^2 \tag{4.15}$$

We can then estimate how much more error is present in the null model than the antigenically-informed model:

$$\Delta SSE = SSE^{null} - SSE^{model} \tag{4.16}$$

The frequency prediction model has a total of 8 free parameters:

Table 4.1. Fitness model parameter definitions

β	Slope of linear relationship between population immunity and viral fitness
γ	Slope of linear relationship between titers and probability of protection
σ	Proportion of titers waning each year since primary infection
f_{s0}	Relative intrinsic fitness of each serotype ($f_0 = 0$ for DENV4)
N	Number of years of previous immunity that contribute to antigenic fitness
dt	Number of years in the future to predict clade frequencies

For each dataset, these parameters are jointly fit to maximize ΔSSE (Table B.1) via a sweep through parameter space. Optimization was confirmed by inspecting the profile likelihoods of each parameter.

Chapter 5

CONCLUSIONS

The previous chapters provide examples of phylodynamic methods which can help us understand how viruses enter populations, adapt to their hosts, and move through populations. Phylodynamics was initially developed for descriptive studies of genomic data. In Chapter 2, I present a retrospective analysis that contextualizes the HIV pandemic within the broader dynamics of lentiviral cross-species transmission (CST). While the direct origins of HIV are well-established, general patterns of lentiviral CST were not previously understood. I apply discrete-trait analysis to quantify the pairwise rate of lentiviral transmission between 24 primate host species. In doing so, I identify 14 novel, ancient CST events, placing the HIV pandemic into context alongside many other CSTs. The observed transmission patterns also highlight that lentiviruses vary widely in their ability and propensity to cross the species barrier. From this, I conclude that lentiviral CST is far more prevalent among primates than previously believed; however, given the vast evolutionary timescale, these events remain rare overall. This supports previous suggestions that the early spread of HIV was primarily driven by shifting epidemiological conditions [29].

More recent methodological advances have enabled explanatory studies that combine genomic and functional data. In preceding chapters, I present two complementary analyses of dengue antigenic diversity and population dynamics; here, I summarize this work and outline several promising areas for future investigation. In Chapter 3, I quantify the extent and nature of dengue antigenic evolution and diversity. Although it has been well-established that antigenic interactions determine dengue case outcomes and epidemic severity, it has long been assumed that each serotype of dengue is antigenically uniform. Here, I adapt a phylogeny-based model of antigenic change to map dengue antigenic variation to genetic

divergence. I identify a minimum of 12 distinct antigenic phenotypes of dengue, rejecting the null hypothesis that only four antigenic variants exist. This suggests that within-serotype antigenic diversity may mediate severity of heterotypic secondary infection, and has major implications for the selection of vaccine strains.

However, as extensively discussed in Chapter 3, the 12 antigenic phenotypes identified here are likely a significant underestimate of dengue antigenic diversity. The titer dataset used in this study spans the breadth of dengue genotypes, but does not capture within-genotype variation due to small sample size. Additionally, these titers are from non-human primates; although we see generally comparable patterns between NHP and human immune responses, there are likely important differences that are not captured in this dataset [56]. Future work is needed to generate a more comprehensive account of dengue antigenic phenotypes. To capture the full breadth of circulating antigenic diversity, I would suggest a rarefaction based approach, wherein batches of dengue strains are randomly sampled until the number of new, distinct antigenic phenotypes observed per batch levels off. Such a richer catalog of dengue antigenic diversity would enable two major advances.

First, a more nuanced understanding of dengue antigenic diversity would greatly facilitate vaccine strain selection. Similar analysis of a more comprehensive dataset (as proposed above) would yield a thorough accounting of antigenic distance (in titer units) between all pairs of dengue strains. Katzelnick *et al.* quantitatively define the relationship between titers and the probability of protection from severe secondary infection [58] and the risk of antibody-dependent enhancement [57]. Together, this could enable the selection of vaccine strains that collectively maximize coverage of antigenic space while minimizing the probability of severe subsequent infection.

Secondly, additional titer data would provide sufficient statistical power to identify which specific mutations in the dengue genome are associated with antigenic change; Neher *et al.* provide a methodological starting point in their 'substitution model' detailed in [77]. This would greatly improve our understanding of how dengue antigenic change arises and enable the identification of antigenically drifted strains based on genomic surveillance.

Building upon the work presented in Chapter 3, in Chapter 4 I investigate the impact of antigenic novelty on dengue population turnover. For many antigenically variable pathogens, such as influenza, antigenic novelty — relative to standing population immunity — is a strong determinant of viral fitness, thereby driving population turnover [11, 103]. While the results from Chapter 3 demonstrate that dengue contains far more antigenic diversity than previously believed, the breadth and magnitude of this variation is far more modest than typically seen for pathogens like influenza. Dengue evolution and population turnover is also much slower than influenza evolutionary dynamics; thus, it was unclear whether the antigenic diversity identified in Chapter 3 impacted real-world dengue population dynamics. I implement a simple, deterministic model to directly quantify the extent to which antigenic novelty contributes to dengue viral fitness.

Looking at quarterly timepoints spanning thirty years of dengue circulation in Southeast Asia, this model quantifies antigenic fitness of each viral clade as its frequency-weighted antigenic distance from recently circulating clades. The model then combines this time-variable antigenic fitness value with a time-invariant, serotype-specific intrinsic fitness value to predict the growth rate of each clade over the subsequent five year period. I demonstrate that although dengue antigenic dynamics are far more subtle than seen in influenza, serotype-level antigenic differences still drive dengue poulation turnover (Pearson $R^2 = 0.62$, accuracy = 80% for serotype growth and decline).

This has significant implications for dengue surveillance and epidemic preparedness. However, these predictions must be more accurate before they are actionable. I expect that error is primarily due to limited temporal, geographic, and antigenic resolution.

First, although dengue is a seasonal pathogen, all timepoints are treated equally in the current model. Previous work has shown that this seasonal fluctuation in viral population size results in stochastic population bottlenecking [72]. While it would be possible to incorporate this via stochastic simulations or other complex approaches, this is very difficult to capture accurately. However, past modeling work provides a useful heuristic: Lourencco *et al.* show that because of this seasonal bottlenecking, highly fit variants are most likely to rise to high

frequency if they are introduced when the viral population size is lower [70]. Thus, one way to incorporate seasonality into the current model would be to simply increase the value of β (which scales how strongly antigenic fitness influences predicted frequency change) for timepoints when the dengue population is expected to be lower.

Additionally, the current model formulation assumes that the entire region of Southeast Asia is a homogeneously mixed, closed population with no viral importation. However, the vast majority of dengue transmissions occur hyperlocally [90]; I expect that the influence of antigenic fitness on regional dynamics reflects an aggregation of these hyperlocal dynamics. As such, I hypothesize that hyperlocal transmission dynamics are driven by a combination of antigenic distance from local population immunity and occasional viral importation from neighboring areas. This could be incorporated by modeling the viral population on a finer geographic scale (e.g., each state / province across the region) and allowing for occasional spillover between demes (*e.g.*, the highest frequency variant of a given deme is introduced to neighboring demes at a set rate). To do so requires significantly increased genomic surveillance with a much finer spatial resolution (*e.g.*, monthly sampling and sequencing from each state / province across Southeast Asia).

Finally, dengue forecasting will also benefit from the richer catalog of antigenic diversity proposed above. Forecasts are based on antigenic fitness, estimated as the frequency-weighted antigenic distance from recently circulating strains. Thus, more accurate assessments of antigenic distance will also contribute to more accurate forecasts. Importantly, while the predictive power demonstrated in Chapter 4 suggests that these genomic surveillance-based estimates are surprisingly effective, these estimates should also be benchmarked against cross-sectional, representative human titers for each population of interest (and model parameters re-fit accordingly).

The advances proposed above would yield geographically specific predictions of which clades will dominate an upcoming epidemic season, alongside more precise estimates of standing population immunity against these incoming strains. Together, this could provide realistic estimates of the proportion of cases expected to be severe in an upcoming

dengue season. Currently, dengue public health interventions are limited to vector control and symptom management, and public health agencies typically have a geographically narrow remit and a small time window for actionable inference to be made. Accurate, actionable predictions of epidemic severity would facilitate efficient resource allocation in low-resource settings.

Collectively, the studies presented here use phylodynamics to augment our understanding of how viruses switch hosts, evolve to evade human immunity, and spread through populations. Clearly, phylodynamics has already become a powerful tool for investigating questions in basic virology and evolutionary biology; as the field progresses, it is also poised to become a vital tool for infectious disease epidemiology and public health.

BIBLIOGRAPHY

1. Avelin F Aghokeng, Ahidjo Ayouba, Eitel Mpoudi-Ngole, Severin Loul, Florian Liegeois, Eric Delaporte, and Martine Peeters. Extensive survey on the prevalence and genetic diversity of sivs in primate bushmeat provides insights into risks for potential new cross-species transmissions. *Infection, Genetics and Evolution*, 10(3):386–396, 2010.

2. Avelin F Aghokeng, Elizabeth Bailes, Severin Loul, Valerie Courgnaud, Eitel Mpoudi-Ngolle, Paul M Sharp, Eric Delaporte, and Martine Peeters. Full-length sequence analysis of sivmus in wild populations of mustached monkeys (cercopithecus cephus) from cameroon provides evidence for two co-circulating sivmus lineages. *Virology*, 360(2):407–418, 2007.

3. Pakorn Aiewsakun and Aris Katzourakis. Time-dependent rate phenomenon in viruses. *Journal of virology*, pages JVI–00593, 2016.

4. Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics*, pages 610–624. Springer, 1992.

5. Orchid M Allicock, Philippe Lemey, Andrew J Tatem, Oliver G Pybus, Shannon N Bennett, Brandi A Mueller, Marc A Suchard, Jerome E Foster, Andrew Rambaut, and Christine VF Carrington. Phylogeography and population dynamics of dengue viruses in the americas. *Molecular biology and evolution*, 29(6):1533–1543, 2012.

6. Cristian Apetrei, Thaidra Gaufin, Rajeev Gautam, Carol Vinton, Vanessa Hirsch, Mark Lewis, Jason Brenchley, and Ivona Pandrea. Pattern of sivagm infection in patas monkeys suggests that host adaptation to simian immunodeficiency virus infection

may result in resistance to infection and virus extinction. *The Journal of infectious diseases*, 202(Supplement_3):S371–S376, 2010.

7. Cristian Apetrei, David L Robertson, and Preston A Marx. The history of sivs and aids: epidemiology, phylogeny and biology of isolates from naturally siv infected non-human primates (nhp) in africa. *Front Biosci*, 9:225–254, 2004.

8. Ahidjo Ayouba, Richard Njouom, Julius Ebua Chia, Steve Ahuka-Mundeke, Anfumbom Kfutwah, Eitel Mpoudi Ngole, Eric Nerrienet, Eric Delaporte, and Martine Peeters. Molecular characterization of a new mosaic simian immunodeficiency virus in a naturally infected tantalus monkey (chlorocebus tantalus) from cameroon: A challenge to the virus–host co-evolution of sivagm in african green monkeys. *Infection, Genetics and Evolution*, 30:65–73, 2015.

9. Elizabeth Bailes, Feng Gao, Frederic Bibollet-Ruche, Valerie Courgnaud, Martine Peeters, Preston A Marx, Beatrice H Hahn, and Paul M Sharp. Hybrid origin of siv in chimpanzees. *Science*, 300(5626):1713–1713, 2003.

10. Trevor Bedford, Sarah Cobey, and Mercedes Pascual. Strength and tempo of selection revealed in viral gene genealogies. *BMC Evolutionary Biology*, 11(1):220, 2011.

11. Trevor Bedford, Andrew Rambaut, and Mercedes Pascual. Canalization of the evolutionary trajectory of the human influenza virus. *BMC Biol*, 10(1):38, 2012.

12. Sidney M Bell. Modern-day siv viral diversity generated by extensive recombination and cross-species transmission. Master's thesis, University of Washington, 2017.

13. Sidney M Bell and Trevor Bedford. Modern-day siv viral diversity generated by extensive recombination and cross-species transmission. *PLoS pathogens*, 13(7):e1006466, 2017.

14. Sidney M Bell, Leah Katzelnick, and Trevor Bedford. Dengue antigenic relationships predict evolutionary dynamics. *bioRxiv*, page 432054, 2018.

15. Samir Bhatt, Peter W Gething, Oliver J Brady, Jane P Messina, Andrew W Farlow, Catherine L Moyes, John M Drake, John S Brownstein, Anne G Hoen, Osman Sankoh, et al. The global distribution and burden of dengue. *Nature*, 496(7446):504, 2013.

16. Robin M Bush, Catherine A Bender, Kanta Subbarao, Nancy J Cox, and Walter M Fitch. Predicting the evolution of human influenza a. *Science*, 286(5446):1921–1925, 1999.

17. MA Charleston and DL Robertson. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Systematic biology*, 51(3):528–535, 2002.

18. Jianbo Chen, Douglas Powell, and Wei-Shau Hu. High frequency of genetic recombination is a common feature of primate lentivirus replication. *Journal of virology*, 80(19):9651–9658, 2006.

19. Zhiwei Chen, Douglas Kwon, Zhanqun Jin, Simon Monard, Paul Telfer, Morris S Jones, Chang Y Lu, Roberto F Aguilar, David D Ho, and Preston A Marx. Natural infection of a homozygous δ24 ccr5 red-capped mangabey with an r2b-tropic simian immunodeficiency virus. *Journal of experimental medicine*, 188(11):2057–2065, 1998.

20. Zhiwei Chen, Amara Luckay, Donald L Sodora, Paul Telfer, Patricia Reed, Agegnehu Gettie, James M Kanu, Ramses F Sadek, JoAnn Yee, David D Ho, et al. Human immunodeficiency virus type 2 (hiv-2) seroprevalence and characterization of a distinct hiv-2 genetic subtype from the natural range of simian immunodeficiency virus-infected sooty mangabeys. *Journal of virology*, 71(5):3953–3960, 1997.

21. Zhiwei Chen, P Telfier, Agegnehu Gettie, Patricia Reed, Linqi Zhang, David D Ho, and Preston A Marx. Genetic characterization of new west african simian immunodeficiency virus sivsm: geographic clustering of household-derived siv strains with

human immunodeficiency virus type 2 subtypes and genetically diverse viruses from a single feral sooty mangabey troop. *Journal of virology*, 70(6):3617–3627, 1996.

22. Alex A Compton and Michael Emerman. Convergence and divergence in the evolution of the apobec3g-vif interaction reveal ancient origins of simian immunodeficiency viruses. *PLoS pathogens*, 9(1):e1003135, 2013.

23. Valérie Courgnaud, Bernadette Abela, Xavier Pourrut, Eitel Mpoudi-Ngole, Severin Loul, Eric Delaporte, and Martine Peeters. Identification of a new simian immunodeficiency virus lineage with a vpu gene present among different cercopithecus monkeys (c. mona, c. cephus, and c. nictitans) from cameroon. *Journal of virology*, 77(23):12523–12534, 2003.

24. Florence Damond, Diane Descamps, Isabelle Farfara, Jean Noël Telles, Sophie Puyeo, Pauline Campa, Annie Leprêtre, Sophie Matheron, Françoise Brun-Vezinet, and François Simon. Quantification of proviral load of human immunodeficiency virus type 2 subtypes a and b using real-time pcr. *Journal of clinical microbiology*, 39(12):4264–4268, 2001.

25. Mirela D'arc, Ahidjo Ayouba, Amandine Esteban, Gerald H Learn, Vanina Boué, Florian Liegeois, Lucie Etienne, Nikki Tagg, Fabian H Leendertz, Christophe Boesch, et al. Origin of the hiv-1 group o epidemic in western lowland gorillas. *Proceedings of the National Academy of Sciences*, page 201502022, 2015.

26. Alexei J Drummond, Marc A Suchard, Dong Xie, and Andrew Rambaut. Bayesian phylogenetics with beauti and the beast 1.7. *Molecular biology and evolution*, 29(8):1969–1973, 2012.

27. Ceiridwen J Edwards, Marc A Suchard, Philippe Lemey, John J Welch, Ian Barnes, Tara L Fulton, Ross Barnett, Tamsin C O'Connell, Peter Coxon, Nigel Monaghan,

et al. Ancient hybridization and an irish origin for the modern polar bear matriline. *Current Biology*, 21(15):1251–1258, 2011.

28. Lucie Etienne, Beatrice H Hahn, Paul M Sharp, Frederick A Matsen, and Michael Emerman. Gene loss and adaptation to hominids underlie the ancient origin of hiv-1. *Cell host & microbe*, 14(1):85–92, 2013.

29. Nuno R Faria, Andrew Rambaut, Marc A Suchard, Guy Baele, Trevor Bedford, Melissa J Ward, Andrew J Tatem, João D Sousa, Nimalan Arinaminpathy, Jacques Pépin, et al. The early spread and epidemic ignition of hiv-1 in human populations. *science*, 346(6205):56–61, 2014.

30. Nuno Rodrigues Faria, Marc A Suchard, Andrew Rambaut, Daniel G Streicker, and Philippe Lemey. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 368(1614):20120196, 2013.

31. Brett M Forshey, Robert C Reiner, Sandra Olkowski, Amy C Morrison, Angelica Espinoza, Kanya C Long, Stalin Vilcarromero, Wilma Casanova, Helen J Wearing, Eric S Halsey, et al. Incomplete protection against dengue virus type 2 re-infection in peru. *PLoS neglected tropical diseases*, 10(2):e0004398, 2016.

32. Feng Gao, Elizabeth Bailes, David L Robertson, Yalu Chen, Cynthia M Rodenburg, Scott F Michael, Larry B Cummins, Larry O Arthur, Martine Peeters, George M Shaw, et al. Origin of hiv-1 in the chimpanzee pan troglodytes troglodytes. *Nature*, 397(6718):436, 1999.

33. Feng Gao, Ling Yue, Albert T White, Peter G Pappas, Joseph Barchue, Aloysius P Hanson, Bruce M Greene, Paul M Sharp, George M Shaw, and Beatrice H Hahn. Human infection by genetically diverse sivsm-related hiv-2 in west africa. *Nature*, 358(6386):495, 1992.

34. MK Gentry, EA Henchal, JM McCown, WE Brandt, and JM Dalrymple. Identification of distinct antigenic determinants on dengue-2 virus using monoclonal antibodies. *The American journal of tropical medicine and hygiene*, 31(3):548–555, 1982.

35. T Gernhard. Using birth-death model on trees. *J Theor Biol*, 253:769–778, 2008.

36. Robert V Gibbons, Siripen Kalanarooj, Richard G Jarman, Ananda Nisalak, David W Vaughn, Timothy P Endy, Mammen P Mammen Jr, and Anon Srikiatkhachorn. Analysis of repeat hospital admissions for dengue to estimate the frequency of third or fourth dengue infections resulting in admissions and dengue hemorrhagic fever, and serotype sequences. *The American journal of tropical medicine and hygiene*, 77(5):910–913, 2007.

37. Robert J Gifford. Viral evolution in deep time: lentiviruses and mammals. *Trends in Genetics*, 28(2):89–100, 2012.

38. Jane Goodall. *The chimpanzees of Gombe: Patterns of behavior*. Belknap Press of Harvard University Press, 1986.

39. Angela M Green, P Robert Beatty, Alexandros Hadjilaou, and Eva Harris. Innate immunity to dengue virus infection and subversion of antiviral responses. *Journal of molecular biology*, 426(6):1148–1160, 2014.

40. Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James LN Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *science*, 303(5656):327–332, 2004.

41. Nathan D Grubaugh, James Weger-Lucarelli, Reyes A Murrieta, Joseph R Fauver, Selene M Garcia-Luna, Abhishek N Prasad, William C Black IV, and Gregory D Ebel. Genetic drift during systemic arbovirus infection of mosquito vectors leads to decreased relative fitness during host switching. *Cell host & microbe*, 19(4):481–492, 2016.

42. Sunetra Gupta, Neil Ferguson, and Roy Anderson. Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science*, 280(5365):912–915, 1998.

43. James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, 2018.

44. Scott B Halstead. In vivo enhancement of dengue virus infection in rhesus monkeys by passively transferred antibody. *Journal of Infectious Diseases*, 140(4):527–533, 1979.

45. Reuben S Harris, Judd F Hultquist, and David T Evans. The restriction factors of human immunodeficiency virus. *Journal of Biological Chemistry*, 287(49):40875–40883, 2012.

46. WG Hill and Alan Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6):226–231, 1968.

47. Vanessa M Hirsch, Robert A Olmsted, Michael Murphey-Corb, Robert H Purcell, and Philip R Johnson. An african primate lentivirus (sivsmclosely related to hiv-2. *Nature*, 339(6223):389, 1989.

48. Edward C Holmes. *The evolution and emergence of RNA viruses*. Oxford University Press, 2009.

49. Koichi Ishikawa, Wouter Janssens, Jacob S Banor, Teiichiro Shinno, João Piedade, Tetsutaro Sata, William K Ampofo, James AM Brandful, Yoshio Koyanagi, Naoki Yamamoto, et al. Genetic analysis of hiv type 2 from ghana and guinea-bissau, west africa. *AIDS research and human retroviruses*, 17(17):1661–1663, 2001.

50. IUCN. International union for conservation of nature, red list of threatened species, 2016.

51. MJ Jin, HUXIONG Hui, DL Robertson, MC Müller, F Barré-Sinoussi, VM Hirsch, JS Allan, GM Shaw, PM Sharp, and BH Hahn. Mosaic genome structure of simian immunodeficiency virus from west african green monkeys. *The EMBO journal*, 13(12):2935–2947, 1994.

52. Mojun J Jin, Jeffrey Rogers, Jane E Phillips-Conroy, Jonathan S Allan, Ronald C Desrosiers, George M Shaw, Paul M Sharp, and Beatrice H Hahn. Infection of a yellow baboon with simian immunodeficiency virus from african green monkeys: evidence for cross-species transmission in the wild. *Journal of virology*, 68(12):8454–8460, 1994.

53. Michal Juraska, Craig A Magaret, Jason Shao, Lindsay N Carpp, Andrew J Fiore-Gartland, David Benkeser, Yves Girerd-Chambaz, Edith Langevin, Carina Frago, Bruno Guy, et al. Viral genetic diversity and protective efficacy of a tetravalent dengue vaccine in two phase 3 trials. *Proceedings of the National Academy of Sciences*, page 201714250, 2018.

54. Marcia L Kalish, Nathan D Wolfe, Clement B Ndongmo, Janet McNicholl, Kenneth E Robbins, Michael Aidoo, Peter N Fonjungo, George Alemnji, Clement Zeh, Cyrille F Djoko, et al. Central african hunters exposed to simian immunodeficiency virus. *Emerging infectious diseases*, 11(12):1928, 2005.

55. Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

56. Leah C Katzelnick, Judith M Fonville, Gregory D Gromowski, Jose Bustos Arriaga, Angela Green, Sarah L James, Louis Lau, Magelda Montoya, Chunling Wang, Laura A VanBlargan, et al. Dengue viruses cluster antigenically but not as discrete serotypes. *Science*, 349(6254):1338–1343, 2015.

57. Leah C Katzelnick, Lionel Gresh, M Elizabeth Halloran, Juan Carlos Mercado,

Guillermina Kuan, Aubree Gordon, Angel Balmaseda, and Eva Harris. Antibody-dependent enhancement of severe dengue disease in humans. *Science*, 358(6365):929–932, 2017.

58. Leah C Katzelnick, Magelda Montoya, Lionel Gresh, Angel Balmaseda, and Eva Harris. Neutralizing antibody titers against dengue virus correlate with protection from symptomatic infection in a longitudinal cohort. *Proceedings of the National Academy of Sciences*, 113(3):728–733, 2016.

59. Tadeusz J Kochel, Douglas M Watts, Scott B Halstead, Curtis G Hayes, Angelica Espinoza, Vidal Felices, Roxana Caceda, Christian T Bautista, Ysabel Montoya, Susan Douglas, et al. Effect of dengue-1 antibodies on american dengue-2 viral infection and dengue haemorrhagic fever. *The Lancet*, 360(9329):310–312, 2002.

60. Sergei L Kosakovsky Pond, David Posada, Michael B Gravenor, Christopher H Woelk, and Simon DW Frost. Gard: a genetic algorithm for recombination detection. *Bioinformatics*, 22(24):3096–3098, 2006.

61. Carla Kuiken, Jim Thurmond, Mira Dimitrijevic, and Hyejin Yoon. The lanl hemorrhagic fever virus database, a new platform for analyzing biothreat viruses. *Nucleic acids research*, 40(D1):D587–D592, 2011.

62. Los Alamos National Labs. Hiv sequence database, 2015.

63. Robert S Lanciotti, Duane J Gubler, and Dennis W Trent. Molecular evolution and phylogeny of dengue-4 viruses. *Journal of General Virology*, 78(9):2279–2284, 1997.

64. Juhye M Lee, John Huddleston, Michael B Doud, Kathryn A Hooper, Nicholas C Wu, Trevor Bedford, and Jesse D Bloom. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. *Proceedings of the National Academy of Sciences*, 2018.

65. Thomas Leitner, Marie-Christine Dazza, Michel Ekwalanga, Cristian Apetrei, and Sentob Saragosti. Sequence diversity among chimpanzee simian immunodeficiency viruses (sivcpz) suggests that sivcpz pts was derived from sivcpz ptt through additional recombination events. *AIDS research and human retroviruses*, 23(9):1114–1118, 2007.

66. Philippe Lemey, Andrew Rambaut, Alexei J Drummond, and Marc A Suchard. Bayesian phylogeography finds its roots. *PLoS computational biology*, 5(9):e1000520, 2009.

67. Florian Liégeois, Vanina Boué, Fatima Mouacha, Christelle Butel, Bertrand Mve Ondo, Xavier Pourrut, Eric Leroy, Martine Peeters, and François Rouet. New stlv-3 strains and a divergent sivmus strain identified in non-human primate bushmeat in gabon. *Retrovirology*, 9(1):28, 2012.

68. Marc Lipsitch and Justin J O'Hagan. Patterns of antigenic diversity and the mechanisms that maintain them. *Journal of the Royal Society Interface*, 4(16):787–802, 2007.

69. Sabrina Locatelli and Martine Peeters. Cross-species transmission of simian retroviruses: how and why they could lead to the emergence of new diseases in the human population. *Aids*, 26(6):659–673, 2012.

70. Jose Lourenço and Mario Recker. Viral and epidemiological determinants of the invasion dynamics of novel dengue genotypes. *PLoS neglected tropical diseases*, 4(11):e894, 2010.

71. Jose Lourenço and Mario Recker. Natural, persistent oscillations in a spatial multi-strain disease system with application to dengue. *PLoS computational biology*, 9(10):e1003308, 2013.

72. José Lourenço, Warren Tennant, Nuno R Faria, Andrew Walker, Sunetra Gupta, and Mario Recker. Challenges in dengue research: A computational perspective. *Evolutionary applications*, 11(4):516–533, 2018.

73. Marta Łuksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507(7490):57, 2014.

74. Kevin R McCarthy, Andrea Kirmaier, Patrick Autissier, and Welkin E Johnson. Evolutionary and functional analysis of old world primate trim5 reveals the ancient emergence of primate lentiviruses and convergent evolution targeting a conserved capsid interface. *PLoS pathogens*, 11(8):e1005085, 2015.

75. Kenji Mizumoto, Keisuke Ejima, Taro Yamamoto, and Hiroshi Nishiura. On the risk of severe dengue during secondary infection: a systematic review coupled with mathematical modeling. *Journal of vector borne diseases*, 51(3):153, 2014.

76. David M Morens, Gregory K Folkers, and Anthony S Fauci. Emerging infections: a perpetual challenge. *The Lancet infectious diseases*, 8(11):710–719, 2008.

77. Richard A Neher, Trevor Bedford, Rodney S Daniels, Colin A Russell, and Boris I Shraiman. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proceedings of the National Academy of Sciences*, 113(12):E1701–E1709, 2016.

78. Molly OhAinle, Angel Balmaseda, Alexander R Macalalad, Yolanda Tellez, Michael C Zody, Saira Saborío, Andrea Nuñez, Niall J Lennon, Bruce W Birren, Aubree Gordon, et al. Dynamics of dengue disease severity determined by the interplay between viral genetics and serotype-specific immunity. *Science translational medicine*, 3(114):114ra128–114ra128, 2011.

79. Sandra Olkowski, Brett M Forshey, Amy C Morrison, Claudio Rocha, Stalin Vilcarromero, Eric S Halsey, Tadeusz J Kochel, Thomas W Scott, and Steven T Stoddard.

Reduced risk of disease during postsecondary dengue virus infections. *The Journal of infectious diseases*, 208(6):1026–1033, 2013.

80. Ivona Pandrea, Cristian Apetrei, Shari Gordon, Joseph Barbercheck, Jason Dufour, Rudolf Bohm, Beth Sumpter, Pierre Roques, Preston A Marx, Vanessa M Hirsch, et al. Paucity of cd4+ ccr5+ t cells is a typical feature of natural siv hosts. *Blood*, 109(3):1069–1076, 2007.

81. Colin R Parrish, Edward C Holmes, David M Morens, Eun-Chung Park, Donald S Burke, Charles H Calisher, Catherine A Laughlin, Linda J Saif, and Peter Daszak. Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiology and Molecular Biology Reviews*, 72(3):457–470, 2008.

82. Danuta Pieniazek, Dennis Ellenberger, Luiz M Janini, Artur C Ramos, John Nkengasong, Madeleine Sassan-Morokro, Dale J Hu, Issa-Malick Coulibally, Ehounou Ekpini, Claudiu Bandea, et al. Predominance of human immunodeficiency virus type 2 subtype b in abidjan, ivory coast. *AIDS research and human retroviruses*, 15(6):603–608, 1999.

83. Alyssa T Pyke, Peter R Moore, Carmel T Taylor, Sonja Hall-Mendelin, Jane N Cameron, Glen R Hewitson, Dennis S Pukallus, Bixing Huang, David Warrilow, and Andrew F Van Den Hurk. Highly divergent dengue virus type 1 genotype sets a new distance record. *Scientific reports*, 6:22356, 2016.

84. A Rambaut, AJ Drummond, and M Suchard. Tracer v1. 6, 2014.

85. Nicholas G Reich, Sourya Shrestha, Aaron A King, Pejman Rohani, Justin Lessler, Siripen Kalayanarooj, In-Kyu Yoon, Robert V Gibbons, Donald S Burke, and Derek AT Cummings. Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. *Journal of The Royal Society Interface*, 10(86):20130414, 2013.

86. Rebeca Rico-Hesse. Molecular evolution and distribution of dengue viruses type 1 and 2 in nature. *Virology*, 174(2):479–493, 1990.

87. Nadeene E Riddick, Emilia A Hermann, Lamorris M Loftin, Sarah T Elliott, Winston C Wey, Barbara Cervasi, Jessica Taaffe, Jessica C Engram, Bing Li, James G Else, et al. A novel ccr5 mutation common in sooty mangabeys reveals sivsmm infection of ccr5-null natural hosts and efficient alternative coreceptor use in vivo. *PLoS pathogens*, 6(8):e1001064, 2010.

88. Philip K Russell and Ananda Nisalak. Dengue virus identification by the plaque reduction neutralization test. *The Journal of Immunology*, 99(2):291–296, 1967.

89. Albert B Sabin. Research on dengue during world war ii1. *The American journal of tropical medicine and hygiene*, 1(1):30–50, 1952.

90. Henrik Salje, Justin Lessler, Irina Maljkovic Berry, Melanie C Melendrez, Timothy Endy, Siripen Kalayanarooj, A Atchareeya, Sumalee Chanama, Somchai Sangkijporn, Chonticha Klungthong, et al. Dengue diversity across spatial and temporal scales: local structure and the effect of host population size. *Science*, 355(6331):1302–1306, 2017.

91. Nadhirat Sangkawibha, Suntharee Rojanasuphot, Sompop Ahandrik, Sukho Viriyapongse, Sujarti Jatanasen, Viraj Salitul, Boonluan Phanthumachinda, and Scott B Halstead. Risk factors in dengue shock syndrome: a prospective epidemiologic study in rayong, thailand. *American journal of epidemiology*, 120(5):653–669, 1984.

92. Mario L Santiago, Cynthia M Rodenburg, Shadrack Kamenya, Frederic Bibollet-Ruche, Feng Gao, Elizabeth Bailes, Sreelatha Meleth, Seng-Jaw Soong, J Michael Kilby, Zina Moldoveanu, et al. Sivcpz in wild chimpanzees. *Science*, 295(5554):465–465, 2002.

93. Paul M Sharp and Beatrice H Hahn. Origins of hiv and the aids pandemic. *Cold Spring Harbor perspectives in medicine*, 1(1):a006841, 2011.

94. Sandrine Souquière, Frédéric Bibollet-Ruche, David L Robertson, Maria Makuwa, Cristian Apetrei, Richard Onanga, Christopher Kornfeld, Jean-Christophe Plantier, Feng Gao, Katharine Abernethy, et al. Wild mandrillus sphinx are carriers of two types of lentivirus. *Journal of Virology*, 75(15):7086–7096, 2001.

95. Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

96. S Susanna Twiddy, Edward C Holmes, and Andrew Rambaut. Inferring the rate and time-scale of dengue virus evolution. *Molecular biology and evolution*, 20(1):122–129, 2003.

97. Fran Van Heuverswyn, Yingying Li, Cecile Neel, Elizabeth Bailes, Brandon F Keele, Weimin Liu, Severin Loul, Christelle Butel, Florian Liegeois, Yanga Bienvenue, et al. Human immunodeficiency viruses: Siv infection in wild gorillas. *Nature*, 444(7116):164, 2006.

98. Erik M Volz, Katia Koelle, and Trevor Bedford. Viral phylodynamics. *PLoS computational biology*, 9(3):e1002947, 2013.

99. Jesse J Waggoner, Angel Balmaseda, Lionel Gresh, Malaya K Sahoo, Magelda Montoya, Chunling Wang, Janaki Abeynayake, Guillermina Kuan, Benjamin A Pinsky, and Eva Harris. Homotypic dengue virus reinfections in nicaraguan children. *The Journal of infectious diseases*, 214(7):986–993, 2016.

100. Helen J Wearing and Pejman Rohani. Ecological and immunological determinants of dengue epidemics. *Proceedings of the National Academy of Sciences*, 103(31):11802–11807, 2006.

101. Chris Whidden and Frederick A Matsen IV. Ricci–ollivier curvature of the rooted phylogenetic subtree–prune–regraft graph. *Theoretical Computer Science*, 699:1–20, 2017.

102. Michael Worobey, Paul Telfer, Sandrine Souquière, Meredith Hunter, Clint A Coleman, Michael J Metzger, Patricia Reed, Maria Makuwa, Gail Hearn, Shaya Honarvar, et al. Island biogeography reveals the deep history of siv. *Science*, 329(5998):1487–1487, 2010.

103. Chunlin Zhang, Mammen P Mammen, Piyawan Chinnawirotpisan, Chonticha Klungthong, Prinyada Rodpradit, Patama Monkongdee, Suchitra Nimmannitya, Siripen Kalayanarooj, and Edward C Holmes. Clade replacements in dengue virus serotypes 1 and 3 are associated with changing serotype prevalence. *Journal of virology*, 79(24):15123–15130, 2005.

Appendix A

CHAPTER 1 SUPPLEMENTAL FIGURES

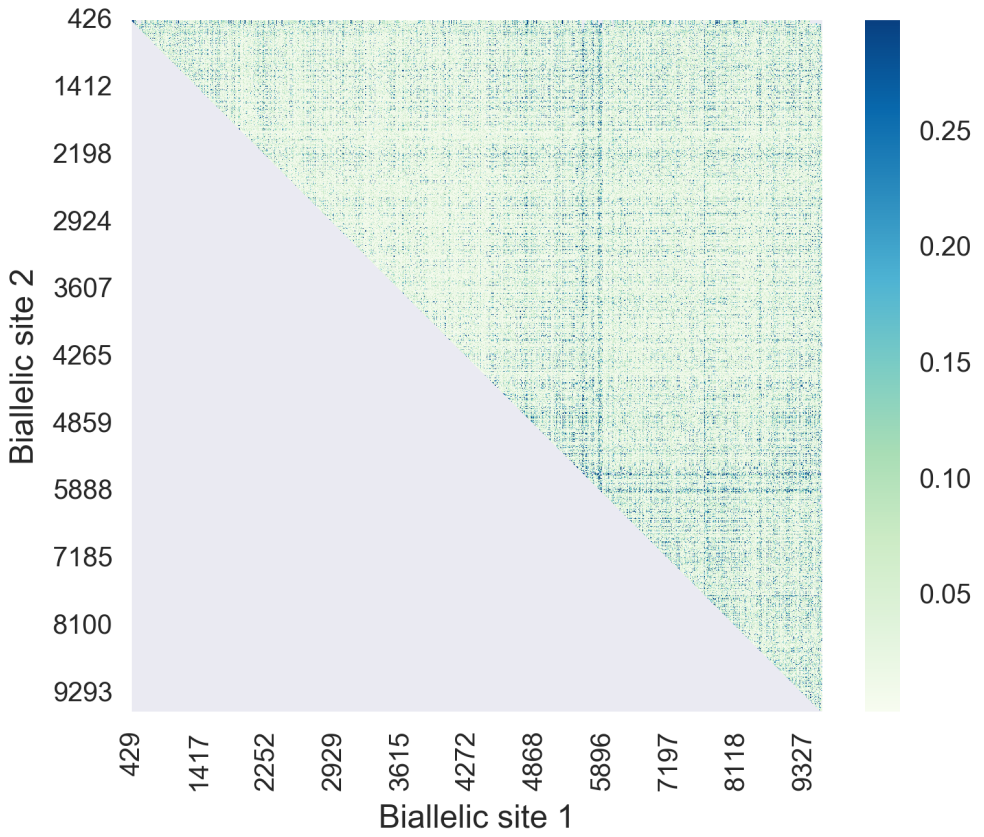

Figure A.1. Heatmap of R^2, a pairwise measure of genetic linkage between sites For pairs of biallelic sites (ignoring rare variants), R^2 was used to estimate how strongly the allele in one site predicts the allele in the second site, with values of 0 indicating no linkage and 1 indicating perfect linkage. The mean value of R^2 was 0.044, indicating very low levels of linkage overall.

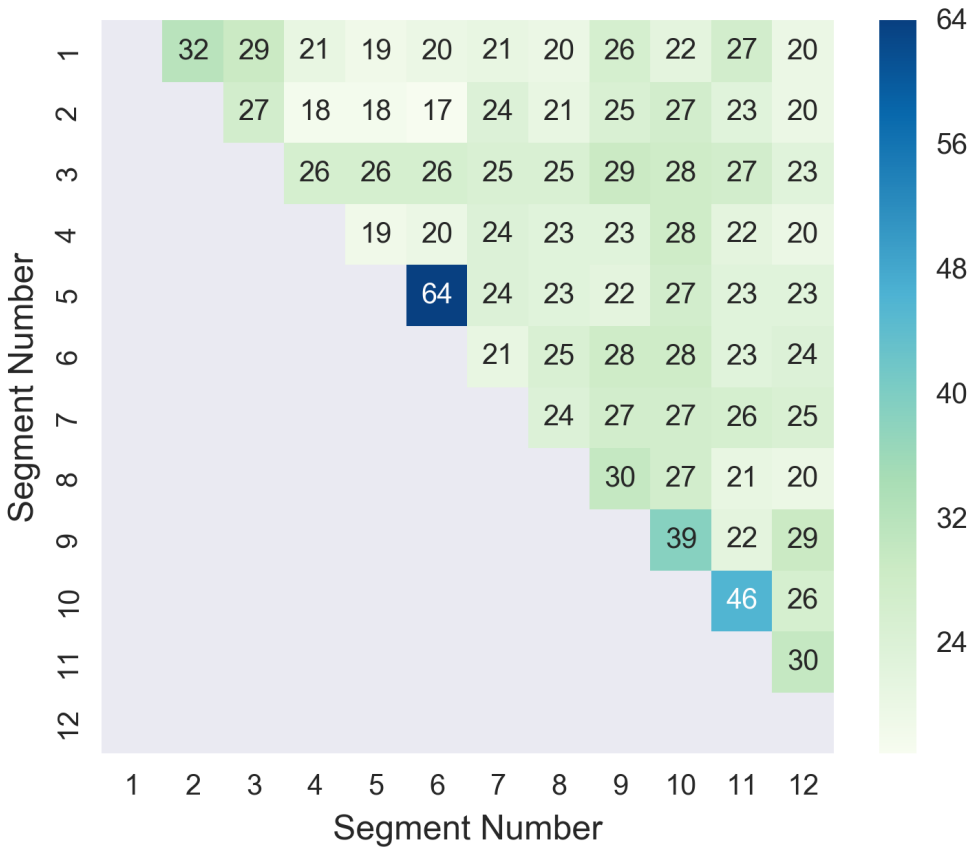

Figure A.2. Heatmap of rSPR scores estimating topological similarity between segments of the SIV genome The alignment used for GARD analyses (LANL compendium with HIV overrepresentation reduced) was split along the breakpoints identified by GARD to yield the 12 genomic segments, and a maximum likelihood tree was constructed for each. The number of steps required to turn one tree topology into another was assessed for each pair of trees with the Rooted Subtree-Prune-and-Regraft (rSPR) package. Segment pairs with similar topologies have lower scores than segments with less similar topologies.

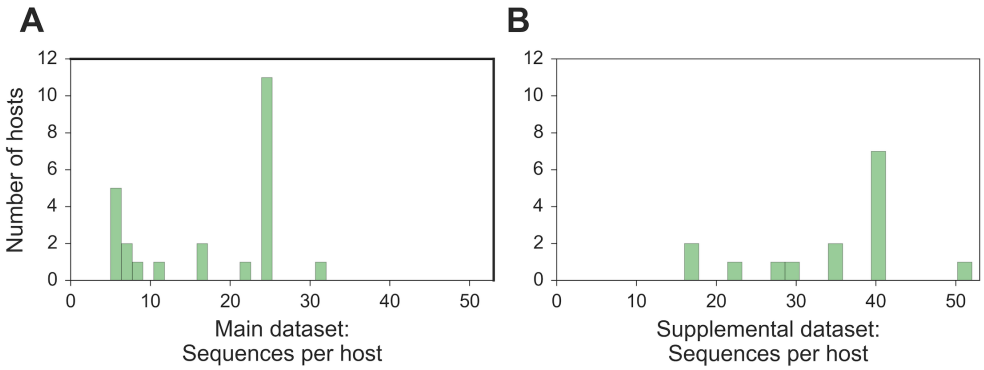

Figure A.3. Distribution of the number of sequences per host included in analyses **A** All available high-quality lentivirus sequences were randomly subsampled up to 25 sequences per host for the main dataset. I included the 24 hosts with at least 5 sequences available in this dataset. **B** For the supplemental dataset, I randomly subsampled up to 40 sequences per host, and included the 15 hosts with at least 16 sequences available in this dataset. For both datasets, a small number of additional sequences were permitted for the few hosts that are infected by multiple viral lineages in order to represent the full breadth of known genetic diversity of lentiviruses in each host population.

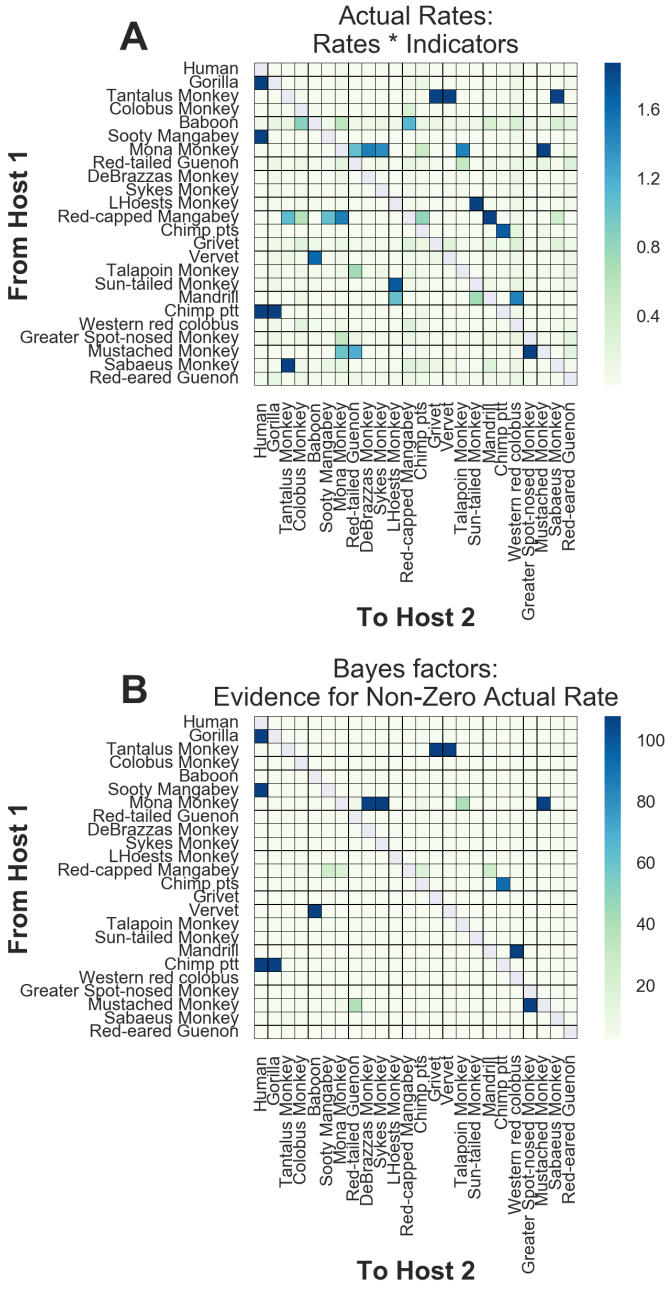

Figure A.4. Actual rates and Bayes factors for main dataset discrete trait analyses Values for the asymmetric transition rates between hosts, as estimated by the CTMC, were calculated as $rate * indicator$ (element-wise for each state logged). The average posterior values are reported above. Bayes factors represent a ratio of the posterior odds / prior odds that a given actual rate is non-zero. Because each of the 12 segments contributes to the likelihood, but they have not evolved independently, I divide all Bayes factors by 12 and report the adjusted values.

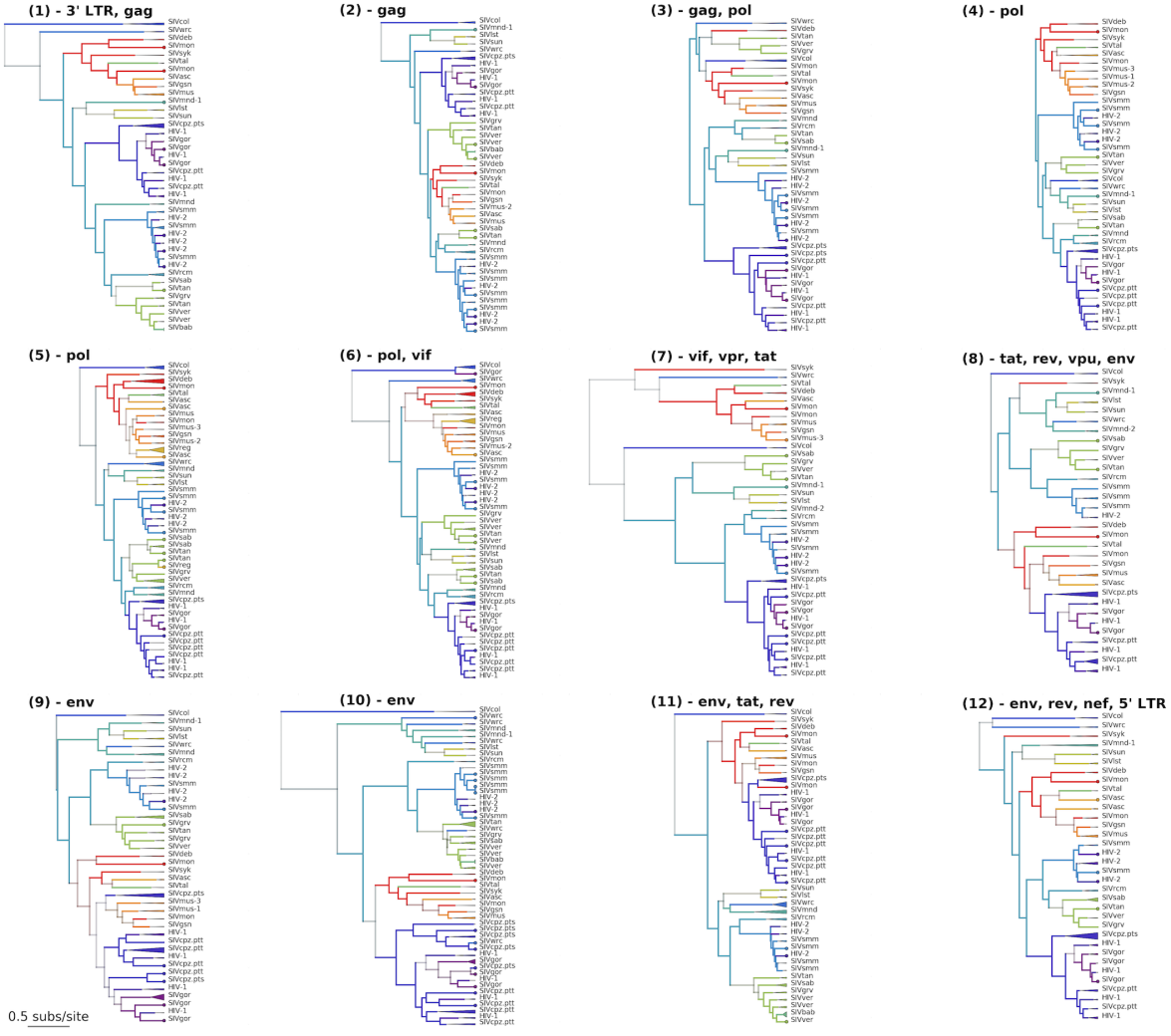

Figure A.5. Maximum clade credibility trees for each of the 12 GARD-identified genomic segments of the lentiviral genome (main dataset) Tips are color coded by known host state; branches and internal nodes are color coded by inferred host state, with color saturation indicating the confidence of these assignments. Monophyletic clades of viruses from the same lineage are collapsed, with the triangle width proportional to the number of represented sequences.

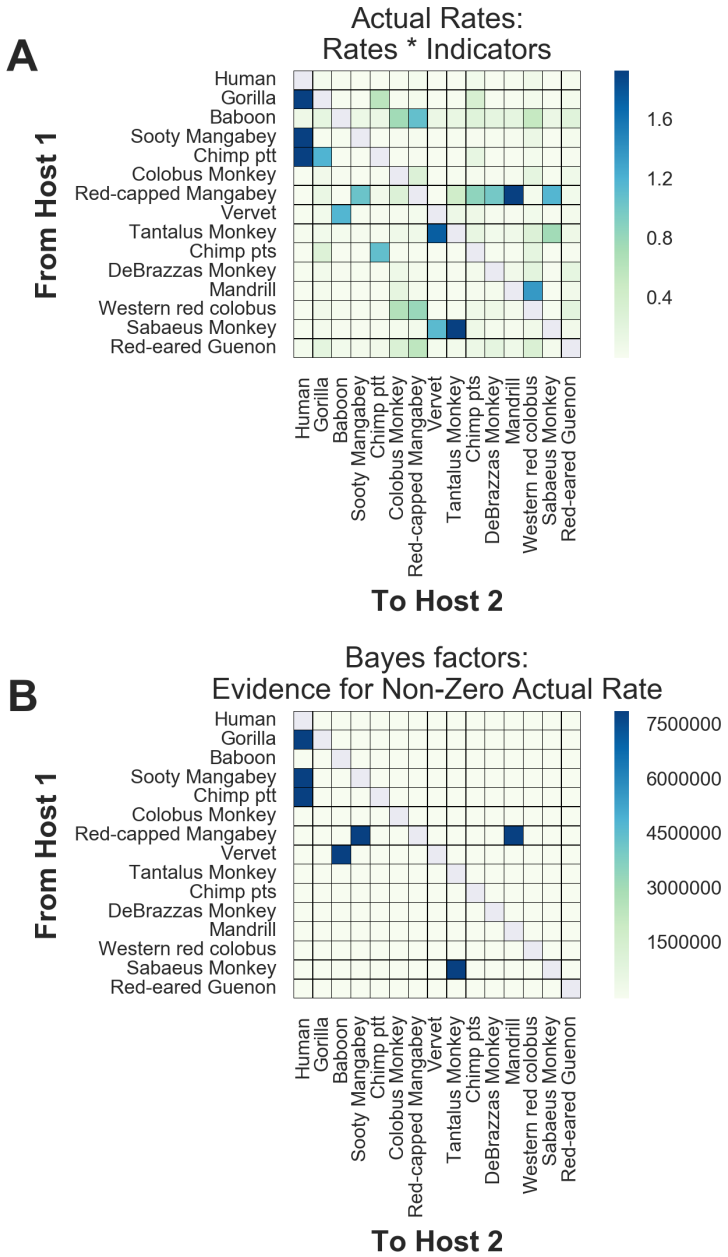

Figure A.6. Actual rates and Bayes factors for 'supplemental' dataset discrete trait analyses Values for the asymmetric transition rates between hosts, as estimated by the CTMC, were calculated as $rate * indicator$ (element-wise for each state logged). The average posterior values are reported above. Bayes factors represent a ratio of the posterior odds / prior odds that a given actual rate is non-zero. Because each of the 12 segments contributes to the likelihood, but they have not evolved independently, I divide all Bayes factors by 12 and report the adjusted values.

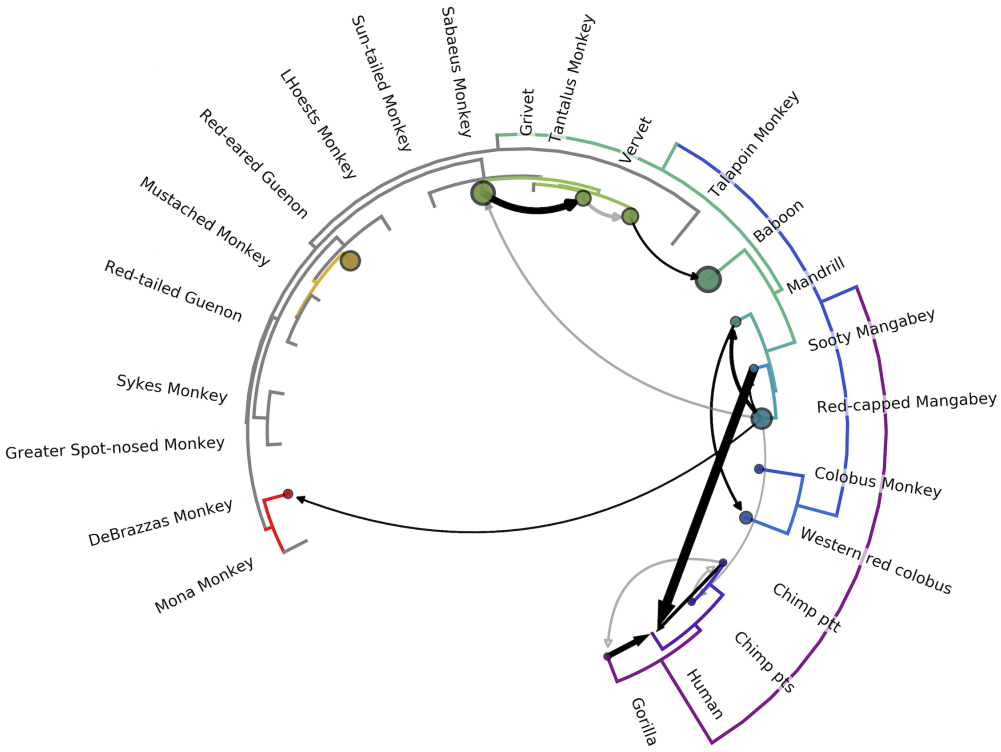

Figure A.7. Network of inferred CSTs of primate lentiviruses ('supplemental' dataset) The phylogeny of the host species' mitochondrial genomes forms the outer circle. Arrows represent transmission events inferred by the model with Bayes' factor (BF) ≥ 3.0; black arrows have BF ≥ 10, with opacity of gray arrows scaled for BF between 3.0 and 10.0. Transmissions with $2.0 \leq BF \leq 3.0$ have open arrowheads (see discussion). Width of the arrow indicates the rate of transmission. Circle sizes represent network centrality scores for each host.

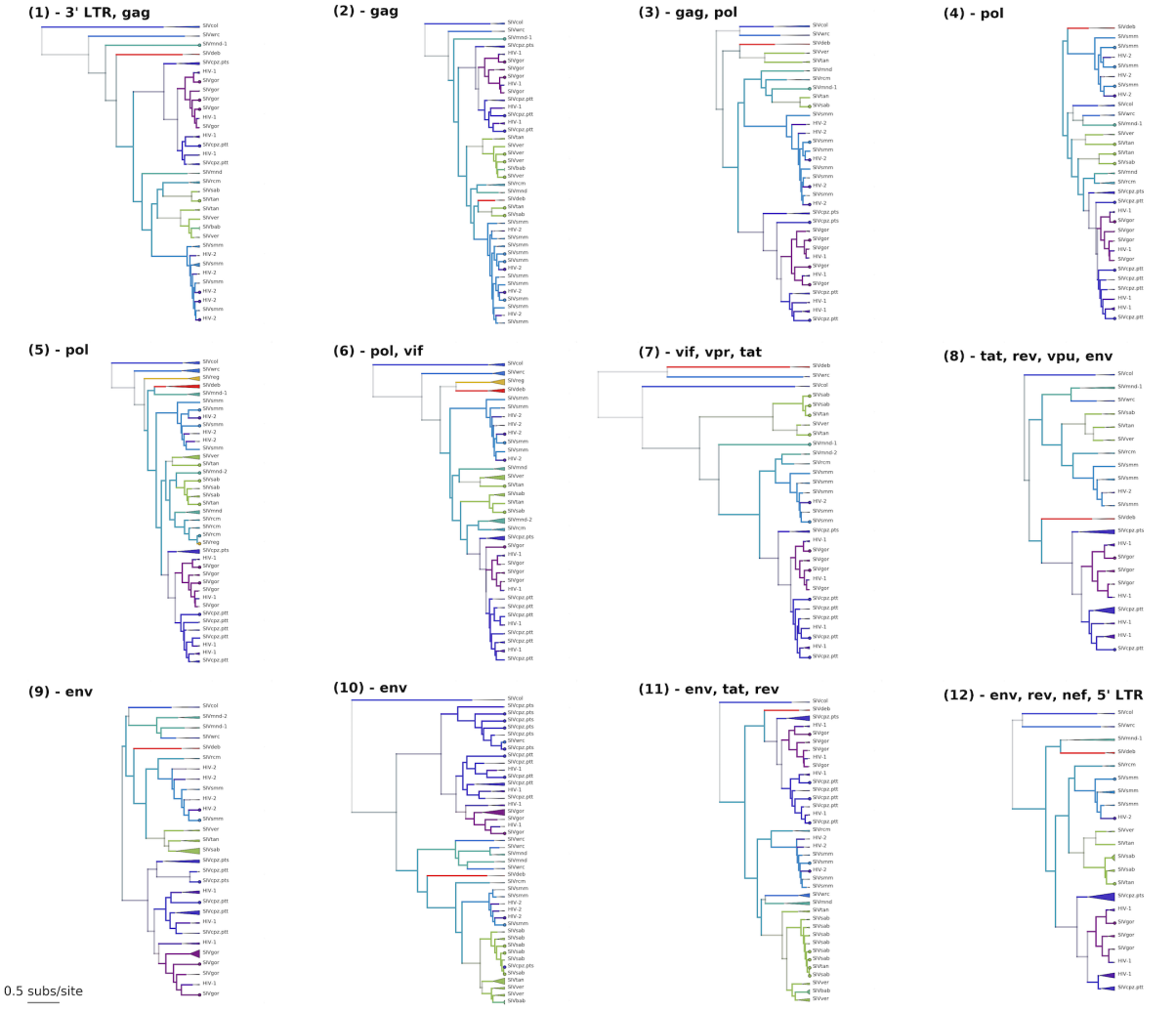

Figure A.8. Maximum clade credibility trees for each of the 12 GARD-identified genomic segments of the lentiviral genome ('supplemental' dataset) Tips are color coded by known host state; branches and internal nodes are color coded by inferred host state, with color saturation indicating the confidence of these assignments. Monophyletic clades of viruses from the same lineage are collapsed, with the triangle width proportional to the number of represented sequences.

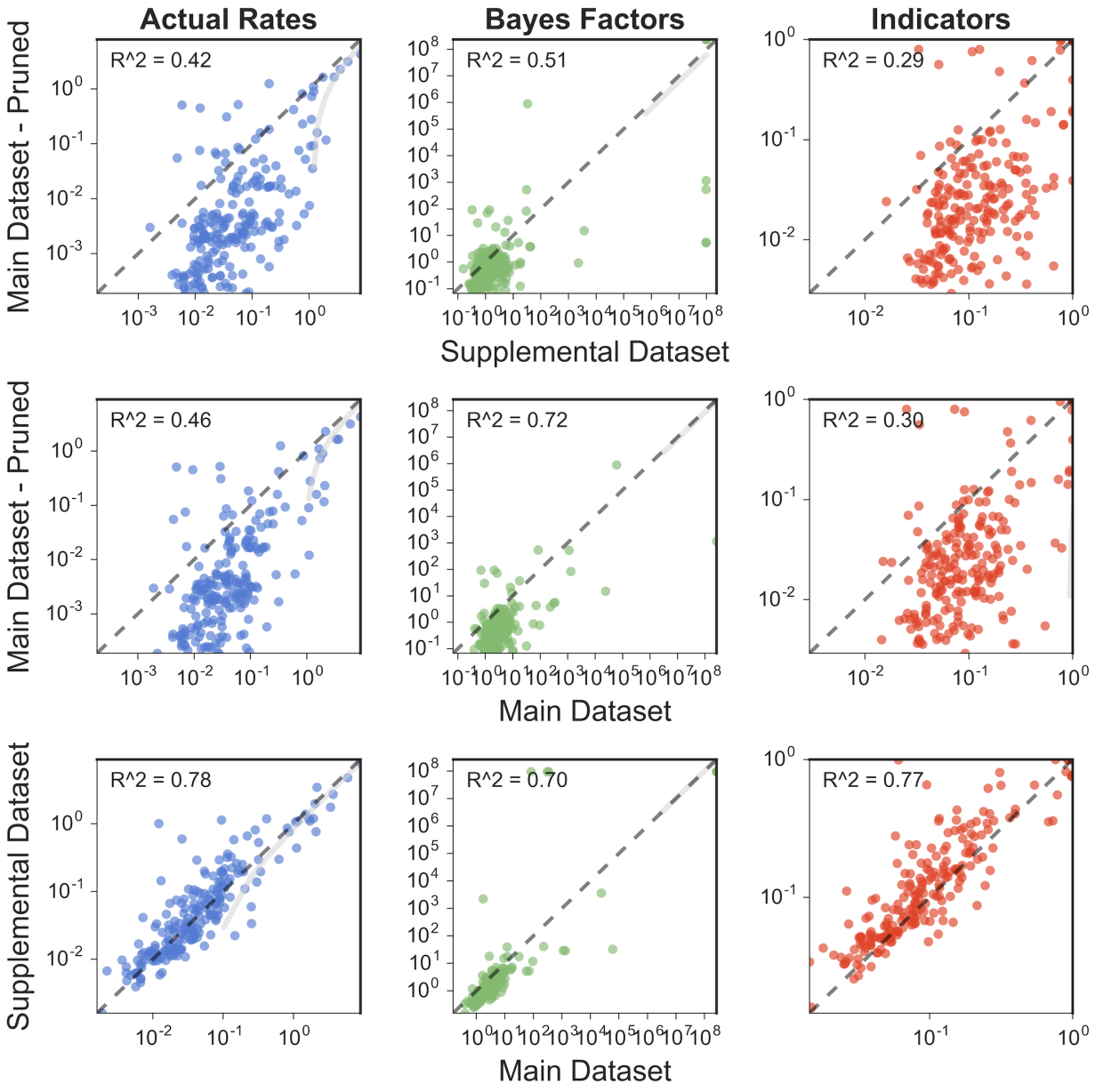

Figure A.9. Comparison of discrete trait analysis results Each datapoint represents one of the 210 possible transmissions between each pair of the 15 hosts present in all three datasets. The 'main dataset' consists of 5-25 sequences from each of 24 host species; the 'pruned dataset' is identical, but with short taxa removed from each segment. The 'supplemental' dataset consists of 16-40 sequences from each of 15 host species. The black dashed line shows y=x; the linear regression is shown in gray.

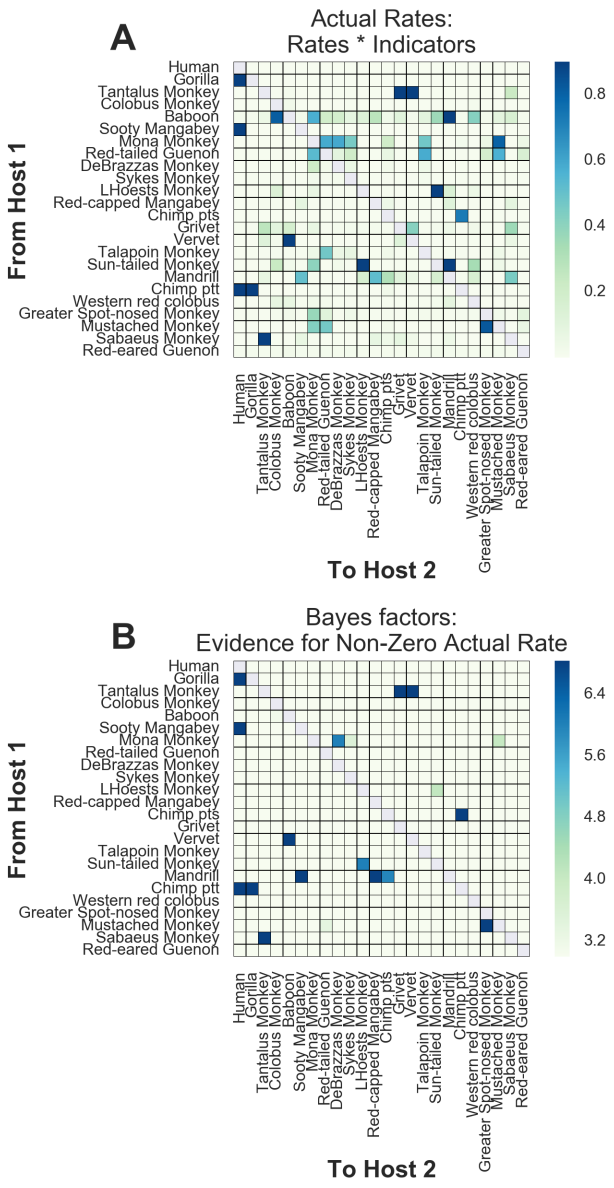

Figure A.10. Actual rates and Bayes factors for 'pruned' main dataset discrete trait analyses Values for the asymmetric transition rates between hosts, as estimated by the CTMC, were calculated as $rate * indicator$ (element-wise for each state logged). The average posterior values are reported above. Bayes factors represent a ratio of the posterior odds / prior odds that a given actual rate is non-zero. Because each of the 12 segments contributes to the likelihood, but they have not evolved independently, I divide all Bayes factors by 12 and report the adjusted values.

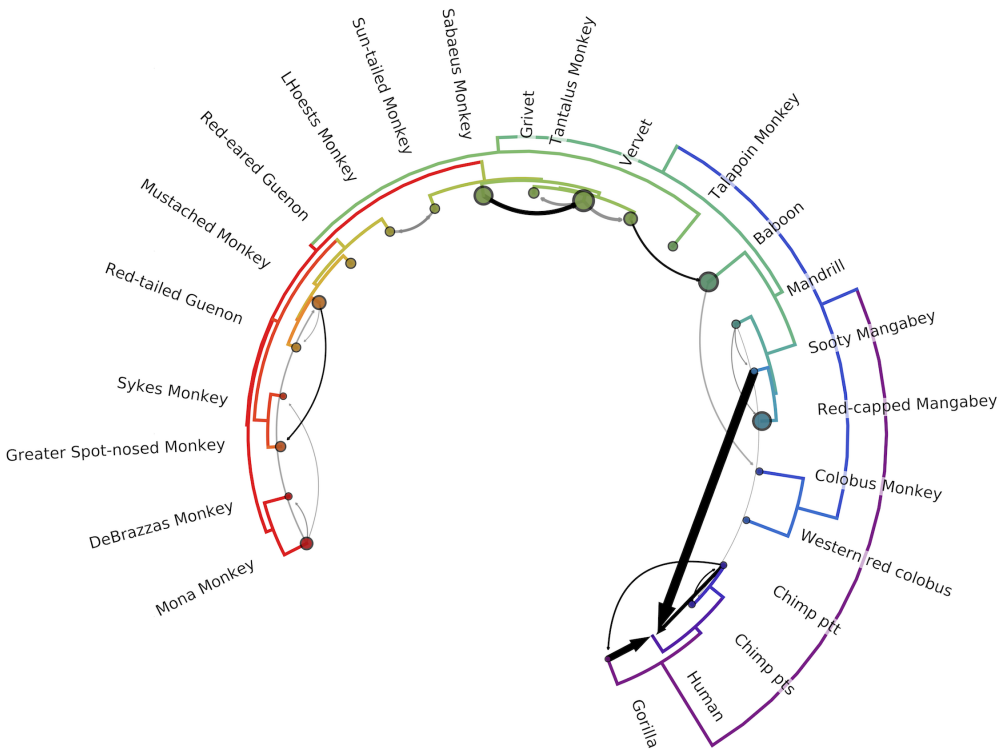

Figure A.11. Network of inferred CSTs of primate lentiviruses ('pruned' dataset) The phylogeny of the host species' mitochondrial genomes forms the outer circle. Arrows represent transmission events inferred by the model with Bayes' factor (BF) ≥ 3.0; black arrows have BF ≥ 10, with opacity of gray arrows scaled for BF between 3.0 and 10.0. Width of the arrow indicates the rate of transmission. Circle sizes represent network centrality scores for each host. As illustrated and discussed, pruning short taxa from the topology can introduce unusual model behavior.

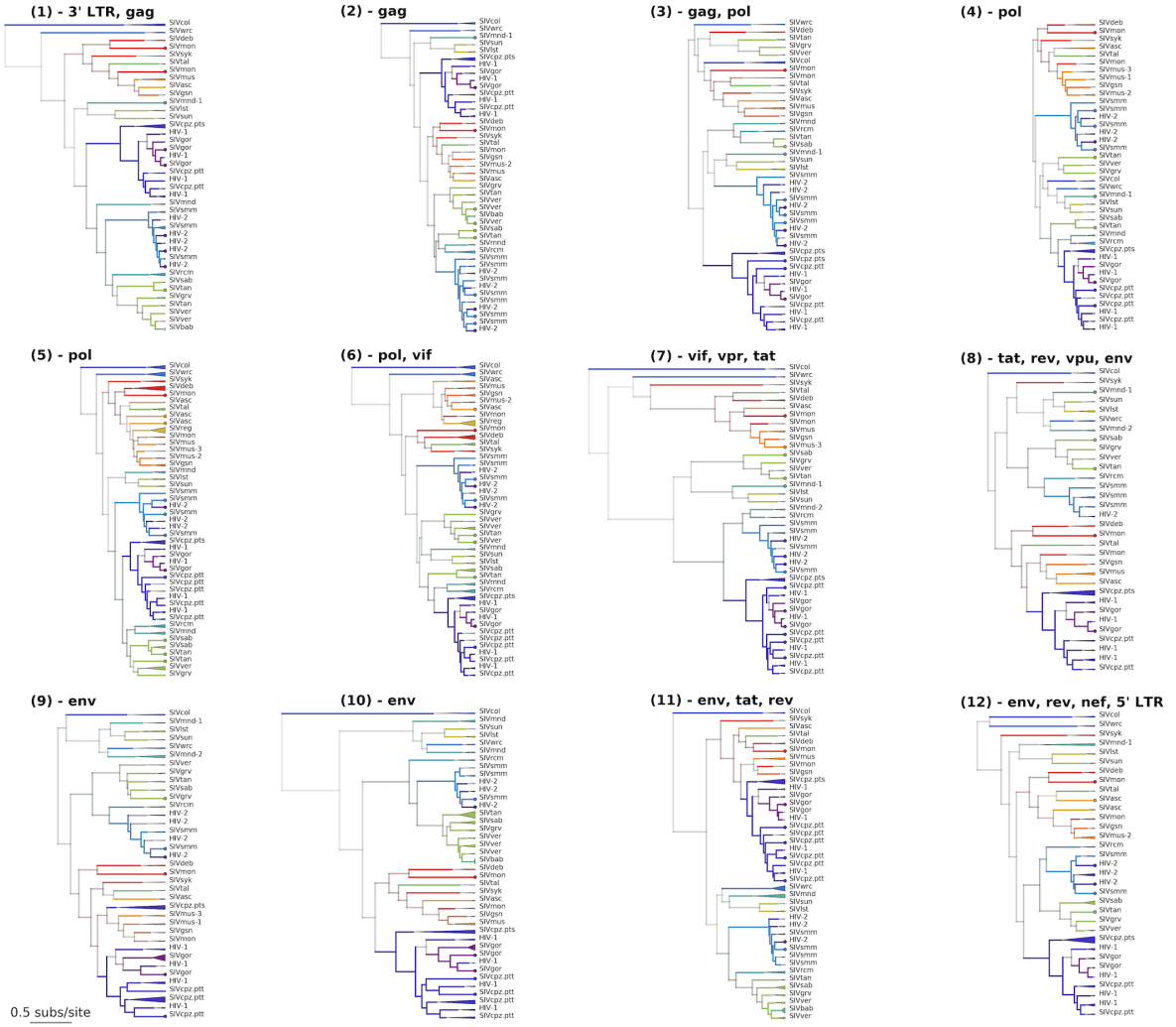

Figure A.12. Maximum clade credibility trees for each of the 12 GARD-identified genomic segments of the lentiviral genome ('pruned' dataset) Tips are color coded by known host state; branches and internal nodes are color coded by inferred host state, with color saturation indicating the confidence of these assignments. Monophyletic clades of viruses from the same lineage are collapsed, with the triangle width proportional to the number of represented sequences.

Appendix B

CHAPTERS 2 AND 3 SUPPLEMENTAL FIGURES

Table B.1. Optimized parameter values for fitness model

Genetic resolution	Antigenic resolution	Metric	Metric value	β	γ	σ	DENV1 f_0	DENV2 f_0	DENV3 f_0	DENV4 f_0
Serotype	Interserotype	Δ SSE	15.02	2.57	0.57	0.86	4.57	3.43	2.14	0.00
Serotype	Interserotype	Pearson R^2	0.63	2.57	0.57	0.86	3.43	2.29	0.71	0.00
Genotype	Interserotype	Δ SSE	14.83	2.57	0.57	0.86	5.71	4.57	3.57	0.00
Genotype	Interserotype	Pearson R^2	0.36	2.57	0.57	0.86	5.71	5.71	2.86	0.00
Genotype	Full tree	Δ SSE	14.22	1.71	0.57	0.43	1.40	0.80	0.40	0.00
Genotype	Full tree	Pearson R^2	0.33	1.29	0.57	0.43	1.40	1.60	0.40	0.00

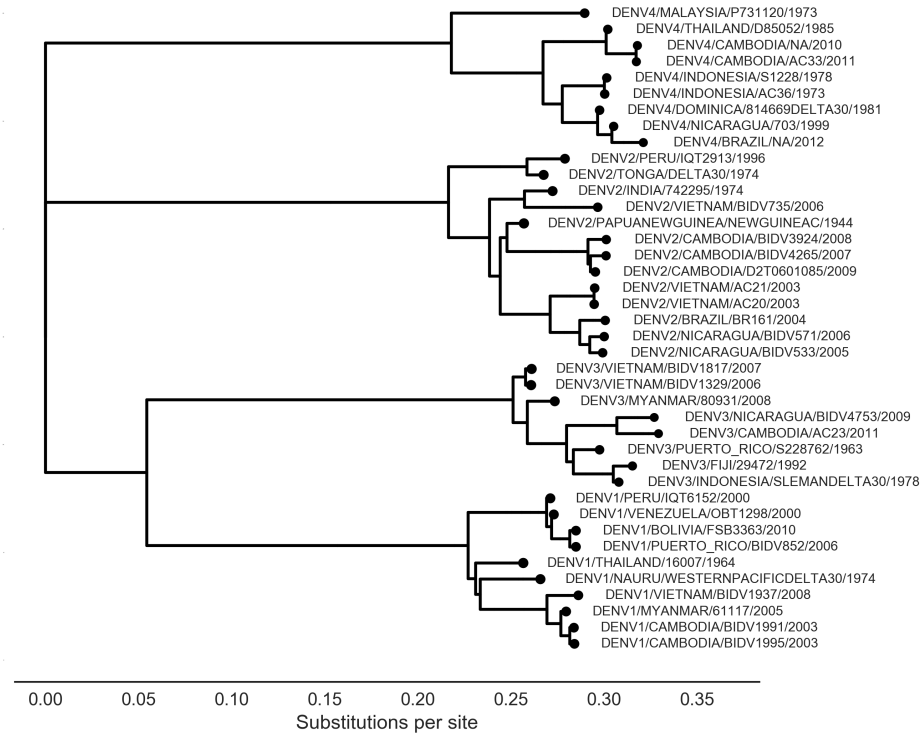

Figure B.1. Tree of dengue viruses in titer dataset Maximum likelihood phylogeny of all dengue viruses that were included in the titer dataset.

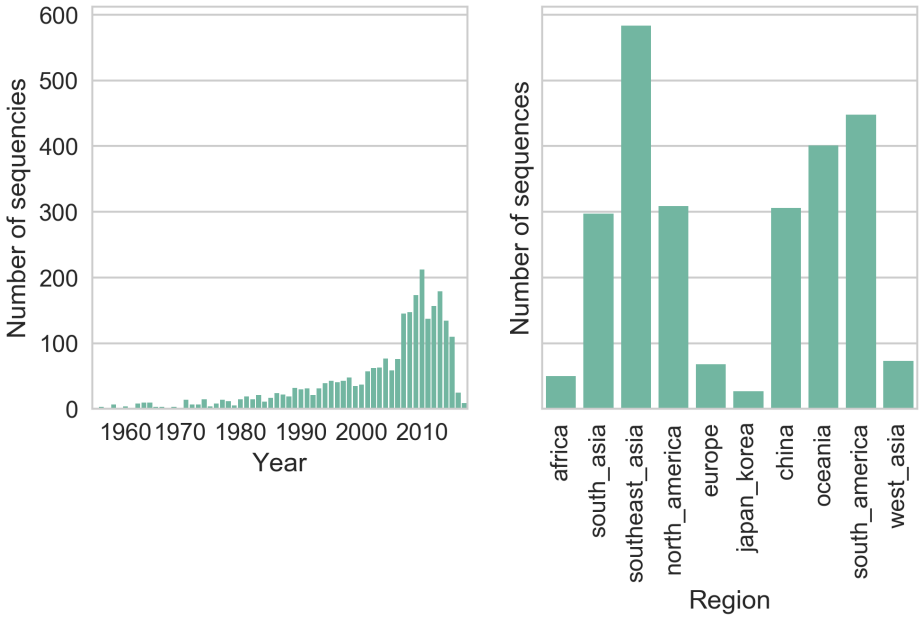

Figure B.2. Sequence dataset distribution Temporal and geographic distribution of sequences included in the dataset after subsampling.

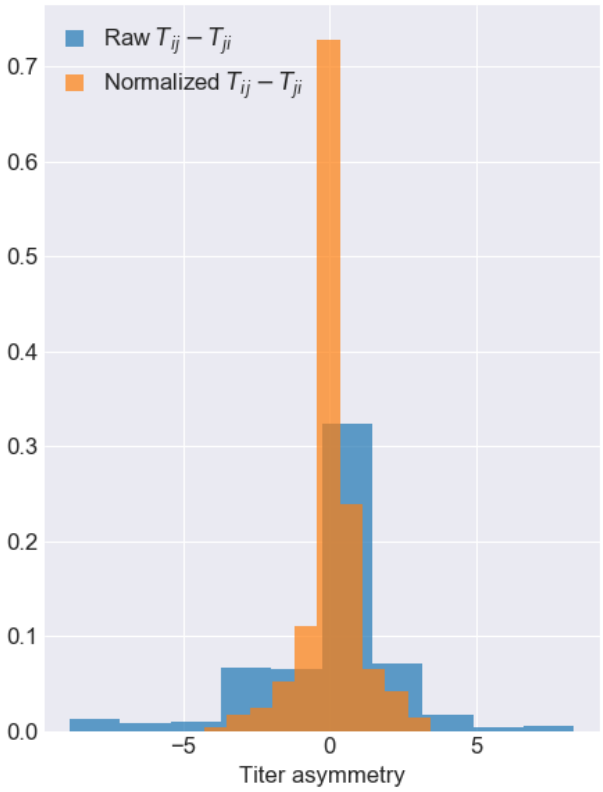

Figure B.3. Titer value symmetry Some viruses have greater avidity overall, and some sera are more potent overall. I normalize for these row and column effects (v_a and p_b, respectively) in the titer model. Once overall virus avidity and serum potency are accounted for, titers are roughly symmetric (*i.e.*, $D_{ij} \approx D_{ji}$).

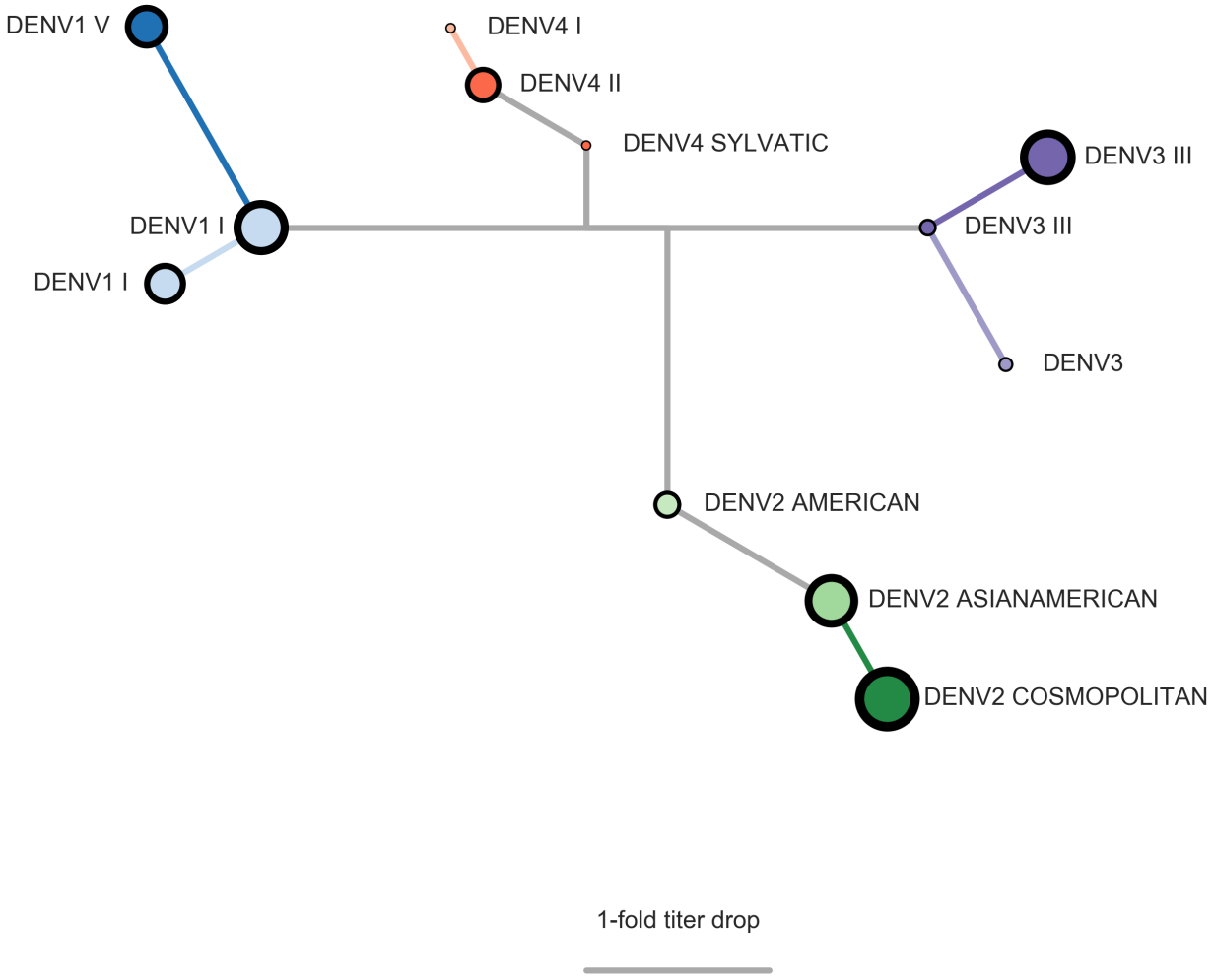

Figure B.4. Tree of dengue antigenic phenotypes (alternate view) As in Figure 3.5, the topology is inferred from a maximum likelihood phylogeny of DENV sequences. Branch lengths are scaled to represent the d_b assigned to each branch by the 'full tree' model of antigenic evolution.

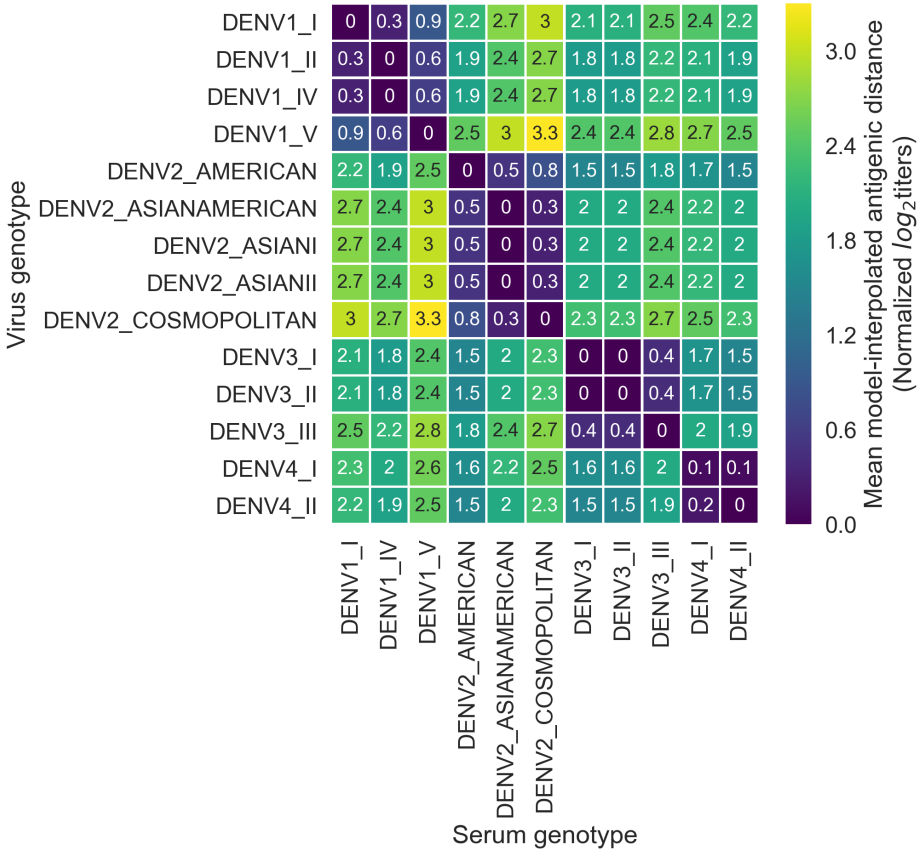

Figure B.5. Titer distance by genotype Values represent the mean interpolated antigenic distance between canonical dengue genotypes (in standardized $\log_2$ titer units).