

Inferring Viral Dynamics from Sequence Data

Neke Ibeh

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of Master of Science in
Biology with specialization in Bioinformatics

Biology
Faculty of Science
University of Ottawa

Abstract

One of the primary objectives of infectious disease research is uncovering the direct link that exists between viral population dynamics and molecular evolution. For RNA viruses in particular, evolution occurs at such a rapid pace that epidemiological processes become ingrained into gene sequences. Conceptually, this link is easy to make: as RNA viruses spread throughout a population, they evolve with each new host infection. However, developing a quantitative understanding of this connection is difficult. Thus, the emerging discipline of phylodynamics is centered on reconciling epidemiology and phylogenetics using genetic analysis. Here, we present two research studies that draw on phylodynamic principles in order to characterize the progression and evolution of the Ebola virus and the human immunodeficiency virus (HIV). In the first study, the interplay between selection and epistasis in the Ebola virus genome is elucidated through the ancestral reconstruction of a critical region in the Ebola virus glycoprotein. Hence, we provide a novel mechanistic account of the structural changes that led up to the 2014 Ebola virus outbreak. The second study applies an approximate Bayesian computation (ABC) approach to the inference of epidemiological parameters. First, we demonstrate the accuracy of this approach with simulated data. Then, we infer the dynamics of the Swiss HIV-1 epidemic, illustrating the applicability of this statistical method to the public health sector. Altogether, this thesis unravels some of the complex dynamics that shape epidemic progression, and provides potential avenues for facilitating viral surveillance efforts.

Résumé

Un objectif principal de la recherche sur les maladies infectieuses est de trouver un lien direct entre la dynamique des populations virales et leur évolution moléculaire. En particulier, les virus à ARN évoluent si rapidement que les processus épidémiologiques sont directement enregistrés dans leur séquences géniques. Si, conceptuellement, ce lien est facile à faire car ces virus évoluent avec chaque nouvelle infection de l'hôte, il est difficile à modéliser de façon quantitative. Ainsi, la nouvelle discipline de phylodynamique se place à l'intersection entre l'épidémiologie et la phylogénétique en utilisant une analyse génétique. Ici, nous présentons deux études qui font appel aux principes de la phylodynamique afin de caractériser la progression et l'évolution du virus Ebola et du virus de l'immunodéficience humaine (VIH). Dans la première étude, l'interaction entre sélection et épistasie dans le génome du virus Ebola est élucidée par la reconstruction ancestrale d'une région critique dans la glycoprotéine virale. Nous fournissons ainsi un nouveau modèle mécaniste des changements structuraux qui ont conduit à l'épidémie de 2014. La deuxième étude applique une approche approximative de calcul Bayésien (ABC) pour l'inférence des paramètres épidémiologiques. Tout d'abord, nous démontrons l'exactitude de cette approche grâce à des simulations. Nous déduisons ensuite la dynamique de l'épidémie du VIH en Suisse, illustrant l'applicabilité de cette méthode statistique pour le secteur de la santé publique. Au total, cette thèse dénoue la dynamique complexe qui donne forme à deux épidémies, et fournit des avenues possibles pour faciliter les efforts de surveillance virale.

Acknowledgements

None of the work presented here would have been possible without the unwavering guidance of my supervisor, Stéphane Aris-Brosou. There was never a question that was too complex, or a task that was too difficult. His caring nature, patience, and peculiar attention to detail, have all contributed to an amazing two years. Thank you, Stéphane, for all that you are: my biggest supporter and my fiercest adversary. It has been a great privilege working with you.

I would also like to sincerely thank my committee members, Drs. Nicolas Rodrigue and Frithjof Lutscher, for their time and valuable comments. In particular, I would like to thank Dr. Lutscher for taking a chance and delving into new territory. I thank my lab mates, past and present, for their support and insightful conversation.

I will forever be thankful to my dearest friends, who have supported me throughout the ups and downs of my Master's degree. I am constantly reminded of how lucky I am to be surrounded by such loving people.

My deepest thanks goes to my family. To my siblings, thank you for your compassion and understanding. I am blessed to have you both by my side. Finally, I would like to thank my wonderful mother. You are my strength and my fortress, and I thank God everyday for your unconditional love. I would not have made it this far without you, so thank you.

The research presented in this thesis was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the University of Ottawa.

Contents

List of Figures	viii
List of Tables	ix
1 Introduction	1
1.1 RNA viruses	1
1.1.1 Genomic structure and mutation	1
1.1.2 Genetic variation and viral evolution	2
1.1.3 The Ebola virus	3
1.1.4 The Human Immunodeficiency Virus (HIV)	7
1.2 Molecular epidemiology and viral evolution	11
1.2.1 Linking epidemiology and evolution	11
1.2.2 Viral phylodynamics	12
1.2.3 Approximate Bayesian Computation for parameter inference	19
1.3 Objectives of the thesis	20
2 Both Epistasis and Diversifying Selection Drive the Structural Evolution of the EBOV Glycoprotein Mucin-like Domain	23
2.1 Abstract	23
2.2 Contributions	24
2.3 Introduction	24

2.4	Materials and Methods	26
2.4.1	Sequence data	26
2.4.2	Phylogenetic analyses	27
2.4.3	Selection analysis	27
2.4.4	Timeline of amino acid substitutions	27
2.4.5	Detection of epistasis	28
2.4.6	Epistasis and selection as interacting agents in outbreak onsets	29
2.4.7	Ancestral MLD structure predictions	29
2.4.8	MLD structure evolution	30
2.5	Results and Discussion	31
2.5.1	Evidence for selection is weak	31
2.5.2	Evidence for epistasis links to selection	33
2.5.3	Selection and epistasis are linked to conformation and func- tion in the MLD	35
2.5.4	Adaptive substitutions as endpoints of epistatic changes . .	37
2.5.5	Insight into the GP MLD structure	40
2.5.6	Linking selection, epistasis, and outbreak onsets	42
3	Inferring viral epidemiological features from HIV genomic data using Approximate Bayesian Computation with Sequential Monte Carlo	43
3.1	Abstract	43
3.2	Contributions	44
3.3	Introduction	44
3.4	Materials and Methods	49
3.4.1	Birth-death tree simulations	49
3.4.2	Simulation study	51
3.4.3	Inferring parameters from Swiss HIV-1 sequence data	51

3.5	Results and Discussion	52
3.5.1	Simulated data	53
3.5.2	Swiss HIV-1 parameter estimates	58
3.5.3	Conclusions	61
4	Conclusions and Future Directions	62
4.1	Summary of research objectives	62
4.2	Main conclusions	63
4.2.1	Evolution of the Ebola virus	63
4.2.2	HIV parameter inference using ABC-SMC	65
4.3	Future directions	68
4.3.1	Linking epistasis and selection: DNA versus RNA viruses . .	68
4.3.2	Expanding the application of ABC-SMC for parameter in- ference	69
	Bibliography	92

List of Figures

1.1 Ebola virus outbreaks from 1976-2014	4
1.2 Linear schematic of EBOV glycoprotein (GP)	6
1.3 Transmission tree versus reconstructed phylogenetic tree	17
1.4 Host contact structure and phylogenetic tree balance	20
2.1 Lineage through time (LTT) plot derived from dated BEAST tree .	33
2.2 Timeline of adaptive amino acid substitutions using a relaxed molec- ular clock phylogenetic tree	34
2.3 Detected epistatic interactions within the seven EBOV genes	35
2.4 Molecular representation of EBOV GP trimer crystal structure . . .	36
2.5 All 6 MLD ancestral reconstructions with labeled sites of interest . .	38
2.6 Root Mean Squared Deviation (RMSD) plot of MLD structural alignments	39
3.1 ABC-SMC estimates of the posterior distribution for four taxa settings	53
3.2 Marginal MSEs for λ	55
3.3 Marginal MSEs for μ	56
3.4 Joint MSE heatmaps	57
3.5 Three-factorial ANOVA for marginal MSEs	58
3.6 Swiss HIV-1 regional trees and parameter estimates	59

List of Tables

2.1	MEME analysis results for codon-specific episodic diversifying selection in the Ebola virus genes	32
3.1	Inferred λ and μ values with 95% HPD intervals for the two Swiss HIV clusters	60
3.2	Inferred λ and μ values using the upper and lower 95% HPD bounds for age	61

Chapter 1

Introduction

1.1 RNA viruses

Viruses are microscopic agents that, by definition, contain either an RNA or DNA genome surrounded by a capsid shell. Structurally, RNA and DNA viruses are essentially identical: a core of genetic material enclosed within a protective protein coat, usually surrounded by a lipid envelope [74]. The general life cycles of these two classes of viruses are also very similar. These cycles consist of attachment to the host cell, penetration, genetic replication, capsid formation, and finally, emergence from the host cell [74]. The chief differences arise from the manner by which genetic information is stored. As their name suggests, RNA viruses store their genetic information in RNA, not DNA. This distinction has major consequences in the life cycle of the virus, both in viral replication and host cell interaction.

1.1.1 Genomic structure and mutation

RNA viruses are remarkably diverse in genome structure, and as a consequence, viral replication is specifically tailored for each genomic design. Viral RNA can be present in double stranded form (ds) or single stranded form (ss), and its genome may be

present on a single RNA segment or may occupy multiple individual segments [74]. Furthermore, a single-stranded genome's RNA strand can either be a sense strand (+ strand), which acts as a messenger RNA (mRNA), or an antisense strand (- strand) that is complementary to the sense strand [74]. Once inside a cell, sense viral RNA is capable of replicating directly, due to its ability to function as mRNA and initiate the translation process [74]. Conversely, antisense RNA has no translational capability, and thus cannot produce viral components [74].

The copying of RNA is carried out by a unique set of RNA polymerase enzymes that comprise the RNA transcriptase [66]. Unlike DNA polymerases, RNA polymerases do not proofread (correct mismatched base pairs), leading to a large number of errors and a high mutation rate. In general, RNA viruses have a mutation rate of 10^{-4} mutations per nucleotide during each cycle [74]. Humans on the other hand exhibit approximately 10^{-8} mutations per nucleotide per cycle [74]. The extraordinarily high mutation rates of RNA viruses relative to their host organisms lead to the continuous propagation of virus variants which possess immense potential for adaptability to new hosts.

1.1.2 Genetic variation and viral evolution

Just as cell-based life is subject to evolutionary pressures, so too are viruses. For evolution to act upon a viral population, genetic variation must be present. As previously mentioned, the high mutation rates of RNA viruses allow them to generate an abundance of variation quite rapidly. Moreover, their short generation times, high replication rates, and potential for recombination and re-assortment, all contribute to the production of genetic variation [34]. Thus, the genetic makeup of RNA populations can change due to selective pressures, where strains with favorable phenotypes increase in prevalence through time. Consequently, beneficial heritable traits, such as antibiotic resistance due to a gene variant, become increasingly common in the pop-

ulation [33]. An arms-race quickly ensues as the host and the virus undergo constant reciprocal adaptation [152]. However, the prolific replication rate of RNA viruses provides them with new opportunities for evading the host immune system, posing significant health concerns for their human hosts.

Evidence suggests that human RNA viruses are both diverse and highly dynamic, with existing strains frequently vanishing from human populations, and novel viruses emerging to replace them [38]. Nevertheless, the vast majority of these newly arising viruses remain rare [38]. Only a small subset of these viruses is capable of proliferating within human populations. The Ebola virus and the Human Immunodeficiency Virus (HIV) represent two of these very special cases in viral evolution, and they are the focal points of this thesis.

1.1.3 The Ebola virus

Background

The Ebola virus causes a severe hemorrhagic fever that is often fatal if left untreated. The case fatality rate is high, varying from 25% to 90% in recorded outbreaks [52]. To date, the five identified Ebola virus species are the Zaire ebolavirus (EBOV), Sudan ebolavirus, Bundibugyo ebolavirus, Reston ebolavirus, and Tai Forest ebolavirus [158]. Although four of the five Ebola species have been detected in animals native to Africa, there is no consensus on the natural reservoir of the Ebola virus [67]. Evidence provided by the host ranges of similar viruses suggests that fruit bats play a major role in Ebola virus transmission, and are most likely the main reservoir [67].

The first Ebola virus outbreak on record occurred in 1976 in what is now the Democratic Republic of Congo [67]. The remote location of this outbreak kept the disease relatively contained within an isolated area. Subsequent outbreaks (1995, 2000, 2007) also had a short duration of transmission and low death tolls [67]. However, the 2014 West African outbreak (first case recorded in March 2014) was the

largest and most severe outbreak on record, totalling more deaths than all previous outbreaks combined (Figure 1.1).

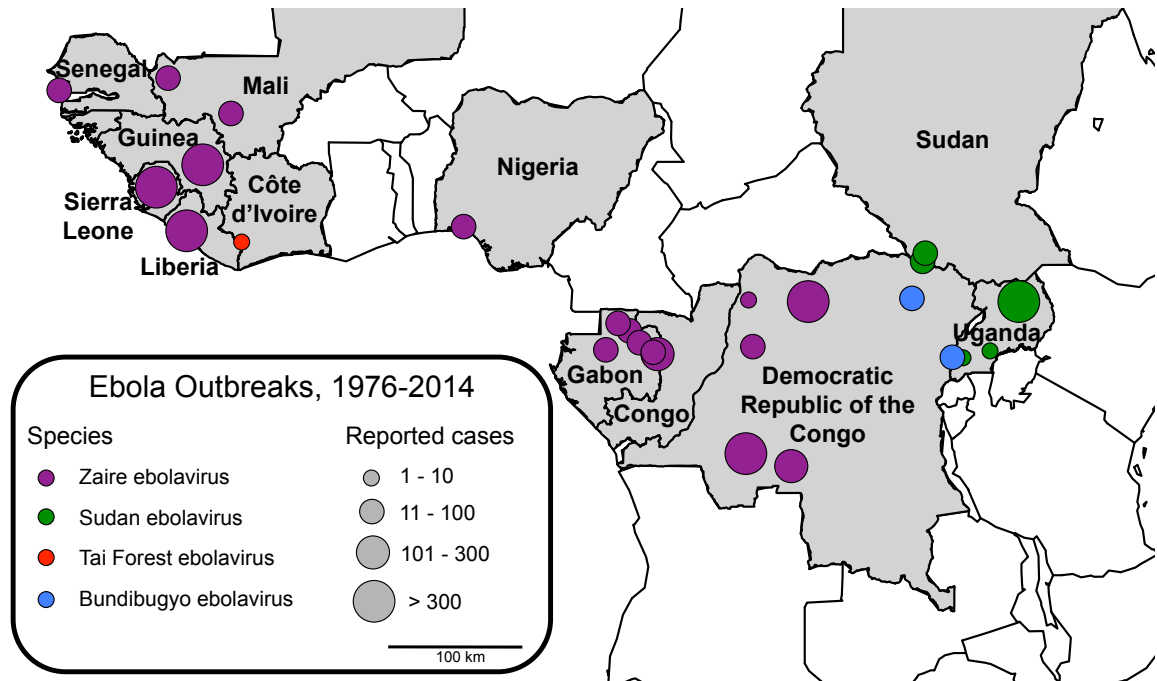


Figure 1.1: Ebola virus outbreaks from 1976-2014. Map shows regional distribution of the total number of reported cases. Map and labels produced with R (v.3.2.1). Data obtained from [158].

A total of 10 different countries reported cases of the Ebola virus from 2014 to 2015 [117]. As of April 2016, the worldwide case count reached over 28,000, and the death toll has surpassed 11,000 individuals [123]. The rapid spread and unprecedented virulence of the most recent strain heightened international efforts towards the development of feasible preventative treatments.

Functional roles of the EBOV proteins

To accelerate future containment efforts, it is essential that the function and structure of the Ebola virus proteins are well understood. The Ebola virus genome consists of seven genes that encode nine proteins, namely: VP40, VP24, L polymerase, VP35, VP30, nucleoprotein (NP), glycoprotein (GP), soluble GP (sGP), and the

small soluble GP (ssGP) [42]. The proteins GP, sGP, and ssGP are a product of the co-transcriptional editing and post-translational processing of the GP protein [42]. The VP40 matrix protein, the most abundant virion protein, shapes the virus and facilitates the budding process [42]. VP40 is capable of associating with lipid bilayers and forms homo-oligomers through various self-interacting domains [42]. As such, VP40 drives the processes of viral assembly and release. The secondary Ebola virus matrix protein is VP24, and it is posited to be a minor component of developed virions [72]. While the function of VP24 is yet to be elucidated, it is hypothesized that this protein possesses similar structural and biochemical characteristics as the VP40 matrix protein [72]. The Ebola genome also encodes four proteins that make up the ribonucleocapsid complex: the L polymerase protein, VP35, VP30, and the nucleoprotein (NP) [72]. The large L polymerase is an RNA-dependent RNA polymerase, and its role is to create new copies of the RNA genome once the virus has entered the host cell [42]. It possesses the enzymatic properties that are necessary for transcription and replication [42]. The proteins VP30 and VP35 are implicated in regulating the transcription and replication processes [72]. At the core of the virus, the nucleoprotein envelops the RNA, creating a helical nucleocapsid complex [72].

While little is known about the functional roles of the secreted proteins sGP and ssGP, the Ebola virus GP is well characterized due to its integral role in host cell entry. As the only transmembrane surface protein, GP is responsible for catalyzing membrane fusion and initiating host cell attachment [86]. In addition, the Ebola glycoprotein shields the viral surface from immune detection and contributes to viral stability outside of the host cell [85]. Protruding from the surface of the virus, GP is a highly dynamic protein consisting of two main subunits. GP1 is responsible for receptor binding, while subunit GP2 controls viral fusion to host cell membranes [86]. Prior to host cell fusion, these two domains form a metastable, pre-fusion conformation, with GP2 later releasing itself and triggering fusion to the cell membrane [85]. Within the GP1 subunit, there exists a highly glycosylated region termed the mucin-

like domain (MLD), which also facilitates viral attachment [85] (Figure 1.2). The MLD has been implicated in masking GP from antibody recognition and enhancing viral adhesion to the target cells [96]. As such, studying this region of GP has become of paramount importance in elucidating the Ebola virus mode of action.

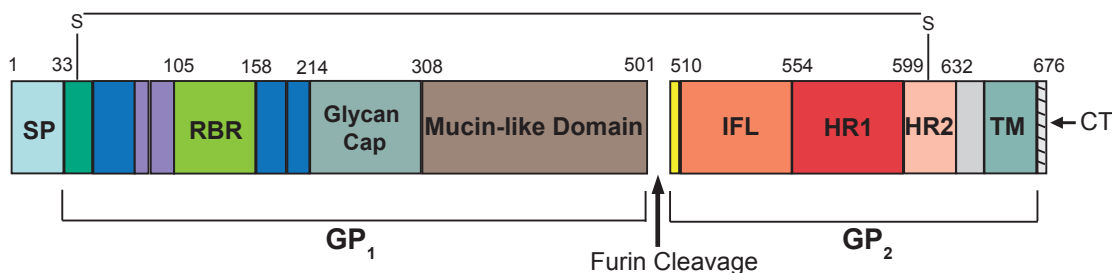


Figure 1.2: Linear schematic of EBOV glycoprotein (GP). SP, signal peptide; RBR, receptor binding region; IFL, internal fusion loop; HR1, heptad repeat 1; HR2, heptad repeat 2; TM, transmembrane domain; CT, cytoplasmic tail. Disulfide bonds are marked with an “S”.

Vaccine design

The MLD and surrounding GP1 domains are targeted by antibodies that bind to the underside of the glycoprotein [26]. As a result, the glycoprotein is the primary target for Ebola virus vaccines [27]. Currently, there is no licensed vaccine available for Ebola treatment. Nonetheless, several candidate vaccines for the prevention of Ebola virus infection are currently in Phase I of clinical trials [61]. The most promising candidate vaccine, rVSV-ZEBOV (recombinant Vesicular Stomatitis Virus Zaire EBOV vaccine), developed by the Public Health Agency of Canada, has been in Phase III of clinical trials since April 2015 [61]. In March 2016, the vaccine was used as a preliminary control measure for a small Ebola outbreak in Guinea [112]. The vaccine consists of a live vesicular stomatitis virus that expresses the Ebola virus GP instead of its own envelope protein [61]. Once injected into the host, this genetically engineered virus elicits the humoral immune response against the complete Ebola virus. Investigational treatments of this vaccine are currently ongoing in Guinea, Liberia,

Canada, Gabon, Germany, Kenya, Switzerland, and the United States [61]. While preliminary data suggests that this vaccine may prevent the spread of Ebola [61], the extent of protection provided by this treatment is unknown.

1.1.4 The Human Immunodeficiency Virus (HIV)

Background

For over three decades, the human immunodeficiency virus (HIV) has been one of the world's most urgent health pandemics [110]. Unlike other RNA viruses such as Ebola and Influenza, the human body is incapable of completely fighting off this virus. Thus, once contracted, the virus remains in the human host for life. HIV targets the host immune system, depleting the system of CD4 cells (T cells), which are critical for the immune response [44]. The host's weakened immune system is now highly susceptible to opportunistic infections or cancers, and this stage of infection signals that the individual has developed Acquired Immune Deficiency Syndrome (AIDS) [44].

Genomic studies suggest that a species of chimpanzee native to Central Africa was the source of HIV infection in humans [69]. These chimpanzees were commonly hunted for meat, and it is hypothesized that the simian immunodeficiency virus was transmitted to humans and subsequently mutated into HIV [69]. The virus then progressively spread across Africa and into other regions of the world. Records show that the virus has existed in North America since at least the mid-1970s [69]. The first isolation of HIV occurred in 1983, and subsequent studies revealed that two types of HIV were circulating in human populations [69]. HIV-1 consists of several recombinants and subtypes, with subtype B representing the predominant strain in North America and Western Europe [110]. HIV-2 is most commonly found in West African populations and is comprised of seven subtypes [110]. At the end of 2014, the World Health Organization declared that approximately 36.9 million people were

living with HIV/AIDS worldwide, 2.6 million of these being children [73].

Preventative methods and treatments

In light of the staggering number of individuals living with this virus, new global efforts have been mounted within the last decade to address this epidemic. Novel interventions in the form of laws and policies have facilitated HIV prevention on a large scale. Priority intervention methods proposed by the World Health Organization have effectively reduced HIV transmission in numerous high-risk groups [73]. Examples of such approaches include the provision of commodities and skilled birth attendants for pregnant women with HIV, access to needle and syringe programs for intravenous drug users, and the distribution of male and female condoms at a subsidized cost [111]. Advances in medicine have also aided in the plight against the global HIV pandemic. Antiretroviral therapy, or ART, is the primary treatment administered to HIV patients, and it significantly increases life expectancy while decreasing the chances of viral transmission to a new host [111]. However, no efficient vaccination for HIV presently exists. Vaccine development is continuously hindered by the rapid evolutionary rate of this virus.

HIV molecular evolution

The extensive genetic diversity of HIV can be attributed to a few key characteristics of the virus. As previously noted, the high mutation rate of this RNA virus provides various pathways by which genetic variation can be introduced into a viral population. One of the key factors that influences this mutation rate is HIV's retroviral lifecycle. Once the virus enters the host cell, its positive-sense RNA is reverse transcribed into a single strand of DNA, which is subsequently copied to form double-stranded DNA [66]. The enzyme responsible for these key steps (reverse transcriptase) lacks exonucleolytic proofreading activity, and the risk of polymerization errors is significantly increased

during either RNA-dependent or DNA-dependent DNA synthesis [66].

However, the high mutation rate alone does not account for the immense genetic variation that is observed in HIV relative to other viruses. It is the astounding rate of replication that this virus possesses that is the key factor. In the course of infection of a single host, the virus is capable of producing billions of copies of itself each day [44]. Another key factor in HIV variability is its ability to undergo recombination, creating new variants within one host individual. This phenomenon occurs when one host cell has been infected by two distinct HIV variants. During replication, reverse transcriptase can then alternate between the two genomic templates [44]. HIV has one of the highest recombination rates of any genetic system (minimum rate of 2.8 crossovers per genome per cycle), and the presence of recombination provides the virus with an efficient means of alleviating the accumulation of deleterious mutations [160, 168]. This ability to potentially eliminate deleterious mutations is especially important when the compact nature of the HIV genome is considered. While the presence of gene overlapping (one genomic region encoding multiple proteins) allows for the vast compression of information seen in many RNA viruses, it elevates the burden of deleterious mutations by making the virus more sensitive to negative mutational effects [137].

As a consequence of these conditions, a viral population can become sensitive to even the slightest selective pressures inflicted by the host immune system. HIV is continuously adapting as it passes through various host populations, and it inevitably evolves novel strategies for immune evasion. Studies suggest that mutations which are associated with immune escape are present at the population level [160]. Such population-wide genomic patterns are due to cytotoxic T-cell responses which detect infected cells through the display of small peptides by the human leukocyte antigen (HLA) [103]. The virus is capable of evading these responses due to the accumulation of mutations that inhibit epitope recognition and binding [103]. It has been reported that high frequency HLAs are promoting the fixation of a growing group of CTL

escape mutations in HIV populations [160]. In a similar manner, drug resistance can arise in a viral population as drug escape mutations become fixed due to excessive antiviral therapy [160]. In this context, recombination allows for the exchange of drug resistant mutations, leading to the rapid evolution of drug-resistant strains. Thus, recombination, coupled with a high rate of mutation, provides new pathways for evading the neutralizing effects of therapeutic drug cocktails and vaccine treatments.

HIV/AIDS clinical databases and long-term cohort studies

HIV databases represent a significant source of genetic data. Using these databases, valuable information can be inferred, including the clinical significance of HIV drug resistance or the transmission risks associated with gender and ethnicity. These databases are usually populated by sequences obtained from long-term observational cohort studies [119]. In an epidemiological context, observational cohort studies track a group of infected individuals (the cohort) over a period of time while collecting pertinent transmission data. Patients regularly visit study centers (i.e. every 6 months) in their respective regions, and the clinical progression of the virus is examined [119]. Treatment is regularly administered to these patients either at their own treatment center or at the study site. The clinical data obtained from each patient (CD4 counts and RNA levels) as well as the relevant personal data (i.e. gender, age, and mode of transmission) are recorded and stored [119]. While the value of such observational studies is immense, limitations naturally exist. Specifically, there may be potential bias involved when attempting to draw conclusions from the data. Misclassification bias may occur if the exposure or disease status of patients is wrongly recorded. Furthermore, information bias can be introduced if a certain group of patients is more difficult to follow-up with than another. Nonetheless, information from such studies is vital to the surveillance and prevention of HIV.

Numerous HIV cohort studies exist worldwide [111], and many of these have

been following patients from the early years of the HIV pandemic. The Swiss HIV Cohort Study (SHCS) is just one example of many such studies. Established in 1988 through the Swiss Federal Office of Public Health, the SHCS is an observational study of HIV-infected citizens of Switzerland that are older than 16 years of age [145]. The seven centers involved in this collaborative study are based in Basel, Bern, Geneva, Lausanne, Zurich, St.Gallen, and Lugano [145]. Approximately 69% of Swiss natives living with AIDS are involved in this study, and 45% of the total HIV infections recorded by Swiss authorities are included [145]. Relative to other western European countries, Switzerland has a high prevalence of HIV/AIDS, with 0.4% of the population being infected at the end of 2014 [109]. In contrast, 0.2% of Canadians were living with HIV/AIDS in 2014 [149]. The enrolment of infected individuals into the SHCS during the Swiss HIV epidemic (early 1980s to the early 2000s) facilitated the efforts made by Swiss health organizations to combat this growing health crisis [145]. Since then, the SHCS has made possible the publication of numerous scientific articles that have addressed questions on HIV infection and progression in an epidemiological and evolutionary context.

1.2 Molecular epidemiology and viral evolution

1.2.1 Linking epidemiology and evolution

The processes of describing and quantifying infectious disease dynamics necessitate an epidemiological framework. Indeed, the field of viral epidemiology is concerned with elucidating the transmission and incidence of infectious diseases within host populations through time. Evolution, on the other hand, aims at describing the transmission of genetic polymorphisms instead of viruses. Fundamentally, both of these disciplines provide a quantitative and predictive means of explaining basic biological processes. The link between them exists due to the fact that epidemiological

mechanisms inevitably influence the direction and extent of selective forces acting on viral populations. As a virus spreads from one host to the next, its rapid rate of mutation provides the necessary genetic variation for evolution to act upon. The rate at which these mutations are fixed is strongly influenced by epidemiological quantities such as the susceptible population size, the transmission rate, and the death rate [49]. As such, research at the interface of epidemiology and evolution seeks to uncover the dynamics and behaviour of infectious diseases using viral sequence data.

The recent surge of sequencing approaches in molecular biology has greatly facilitated the analysis of essentially any pathogen genome. Unfortunately, the vast majority of the collected pathogen data is under-utilized [49]. To alleviate this issue, interdisciplinary approaches are being integrated into viral genomic data analysis. Such disciplines include population genetics (mutation rates, recombination, and selection), ecology (spatial and temporal distributions of the virus and host), and phylogenetics (evolutionary relationship amongst viral strains). At the forefront of such studies is an integrated approach termed phylodynamics [55], which combines viral population dynamics with molecular phylogenetics.

1.2.2 Viral phylodynamics

Phylogenetic reconstruction

Encompassing a wide range of disciplines, the field of phylodynamics attempts to uncover the evolution and transmission of infectious diseases. As previously noted, epidemiological factors (such as transmission rate) along with selection have a significant bearing on the observed patterns of viral genetic variation. As a consequence, sequence phylogenies become a means by which viral population dynamics can be investigated [49]. In order to make such inferences, phylogenetic trees must be reconstructed. These trees reflect the inferred evolutionary relationship amongst a group of viral sequences. The reconstruction of viral phylogenetic trees is accomplished using

various methods, as there is no perfect approach. Methods for tree inference differ in the manner by which multiple sequence alignments are transformed into numerical data. The three main classes are distance, parsimony, and likelihood based methods.

Distance-based methods Distance-based methods employ a distance matrix, which is a table listing the evolutionary distances between all pairs of sequences included in the dataset [19]. These distances are based on the number of nucleotide differences per nucleotide site for each pair of sequences [48]. The bifurcation patterns and branch lengths within the generated tree reflect the values contained in the additive distance matrix. While conventional distance-based methods (such as the least-squares difference [48]) require parsing of the tree space to obtain the optimal distance tree, algorithmic approaches to the distance-based method are much more efficient. Distance-based algorithms such as the unweighted pair group method (UPGMA) [139] and the neighbour-joining method [134] do not require the scanning of a tree space; thus, they overcome the limitations faced with more classical distance methods. The UPGMA approach constructs a tree by first identifying the smallest distance value in the matrix, pairing the identified taxa into one taxonomic unit, and recalculating the distance matrix. This process is iterated for each pair of taxa in order to produce an ultrametric tree. The neighbour-joining algorithm is very similar to the UPGMA clustering approach, but it permits varying evolutionary rates on different tree branches [134]. This algorithm begins with all taxa forming a star-like pattern from a single central node. At the first clustering stage, the pair of sequences (taxa) that gives the smallest sum of branch lengths is chosen. These taxa are transferred to a second internal node, which is then connected to the initial central node. The algorithm proceeds by identifying pairs of taxa (neighbours) that minimize the total branch length at each stage of clustering [134]. Since distance-based methods allocate one value to every pair of sequences, they strongly reduce the phylogenetic information extracted from each sequence. The maximum parsimony method, a character-based

method, allows for the analysis of sequence characters throughout the reconstruction process.

Maximum parsimony method The maximum parsimony method operates under the assumption that evolution takes the shortest available route; thus, the correct phylogenetic tree is the one which requires the least number of nucleotide substitutions to reflect the divergence amongst sequences [47]. Each phylogenetic tree is scored based on the number of evolutionary changes necessary to produce the observed data. The optimal tree (or group of trees) is the one with the smallest number of required changes (the most parsimonious tree). One major limitation of parsimony analysis is that in cases where there are too many possible trees to construct, scoring becomes increasingly difficult. Parsimony algorithms that can search through a set of potential trees in a tree space without scoring all options have therefore been proposed to overcome this issue [47]. A common source of error in phylogenetic analyses (particularly those employing maximum parsimony) is a phenomenon called long-branch attraction (LBA). LBA is the tendency of highly divergent taxa to group together in a tree, irrespective of their true relatedness. Another major shortcoming of both distance and parsimony methods is their inability to determine how much more probable one tree is over another. Likelihood methods, as well as Bayesian methods, naturally provide a statistical approach to the tree reconstruction process, alleviating the obstacles noted above.

Maximum likelihood and Bayesian methods Maximum likelihood approaches determine the probability that the hypothesis (branching pattern) explains the observed data (sequence alignment) [46]. Similar to the above mentioned methods, maximum likelihood approaches compare potential trees and assign scores to each one. However, this score is based on the likelihood (or probability) that a given group of sequences evolved under a particular evolutionary history given the substi-

tution model [19]. The tree with the highest likelihood score is the optimal tree. As previously noted with parsimony and distance-based methods, maximum likelihood approaches are computationally intensive for large tree spaces. A much more flexible variant of maximum likelihood is the Bayesian approach to tree reconstruction. Bayesian methods allow the implementation of complex models of evolution while taking into account the prior probabilities of different hypotheses. Instead of calculating the probability that a certain set of model parameters gave rise to the observed data, Bayesian methods attempt to calculate the posterior probability of the model parameters, given the observed data [19]. In other words, if the data is denoted by x and model parameters are denoted by θ , the quantity calculated by maximum likelihood methods is $p(x|\theta)$, while that calculated by Bayesian methods is $p(\theta|x)$. Bayes' theorem reflects the relationship between these two quantities [71]:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \quad (1.2.1)$$

where $p(\theta)$ is the prior distribution and $p(x)$ is the marginal probability distribution of the observed data.

A major challenge for Bayesian approaches is that the analytical calculation of the posterior probability is hampered by the necessary calculation of the marginal probability of the data (the denominator in Eqn. 1.2.1), which requires integration over the entire parameter space [19]. When the hypothesis space is large, computing this denominator can be computationally difficult [130]. To facilitate this process, Markov chain Monte Carlo (MCMC) simulations [60, 97] are used to parse through the parameter and tree space. The MCMC approach generates a random walk in parameter space based on an initial prior density and uses a rejection method for proposed moves [60]. This approach enables the analysis of complex evolutionary hypotheses using a statistical framework.

Divergence dating While inferring the evolutionary patterns amongst a group of viral sequences is the primary goal of phylogenetic analysis, a secondary objective is determining the divergence times of the ancestral sequences. In order to assign a date to each branching event in a tree, molecular clocks are used. The strict molecular clock hypothesis posits that the rate at which nucleotide substitutions occur over time is relatively constant [169]. Consequently, the magnitude of genetic difference between two taxa can be used to reveal when their ancestral sequence diverged. The original hypothesis of a constant substitution rate through time (strict molecular clock) has since been modified into a more realistic approach, where molecular rates are allowed to vary among lineages (relaxed molecular clock) [135, 166]. Molecular clock analysis thus serves as an immensely valuable tool for reconstructing the evolutionary history of organisms that lack an extensive fossil record, such as viruses.

Mathematical modeling

Analytical methods currently play a significant role in addressing questions within the field of phylodynamics. Methods based on principles from coalescent theory [78] and birth-death processes [142] are widely applied to phylogenetic analysis and epidemiological parameter inference.

The Coalescent The coalescent is a population genetics model that determines the distribution of gene divergence in a genealogy of n sampled members of a population [78]. In a phylogenetic framework, coalescent theory models the ancestry of a group of sampled lineages backwards in time until the most recent common ancestor of the samples is reached [78]. In this way, the underlying population dynamics of a viral population can be inferred. While classical coalescent theory was based on constant population sizes with discrete generations, more recent implementations permit varying population growth and continuous (overlapping) generations [28, 126]. Nonetheless, the assumptions made by the coalescent are not suitable for modeling

epidemics. In particular, the coalescent requires that the number of infected hosts within the effective population is continuously large [132], but this requirement cannot be met in the early stages of an outbreak. Forward-in-time birth-death models have therefore been used as an alternative avenue for reconstructing viral epidemics, as they allow a more realistic description of early outbreaks [144].

The Birth-Death approach and compartmental models Unlike the coalescent, the birth-death diversification model can represent the exponential growth phase of an epidemic, while allowing random extinctions to occur [142]. In the birth-death model, a birth event (transmission) corresponds to a bifurcation in the tree, while a death event (becoming noninfectious) is reflected by the truncation of a lineage [142]. In this manner, the birth-death model generates a transmission tree.

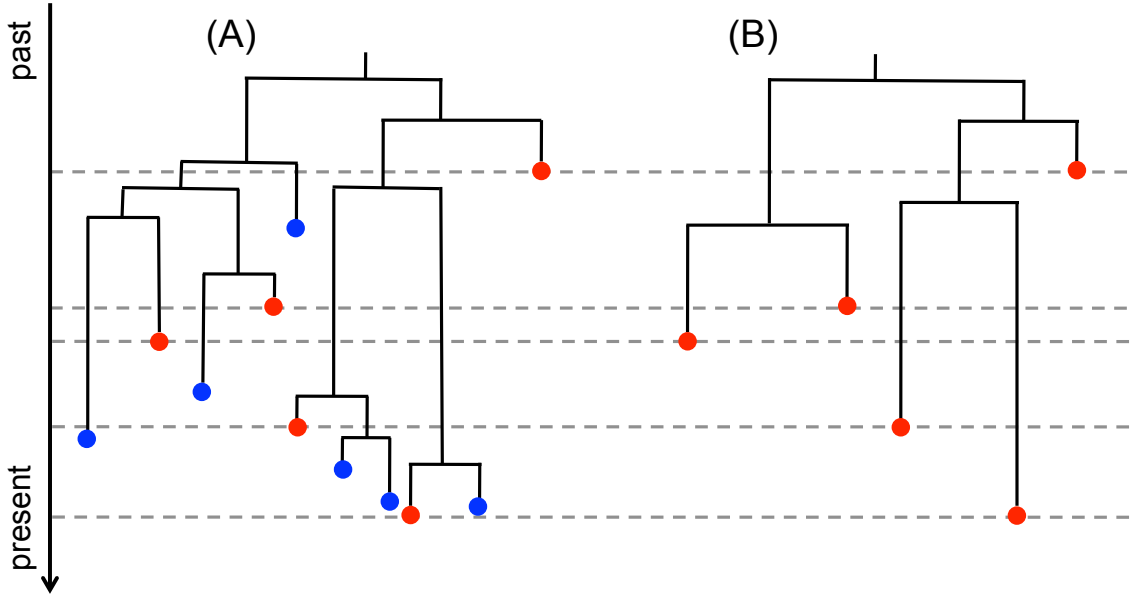


Figure 1.3: Transmission tree versus reconstructed phylogenetic tree. (A): A transmission tree produced by the birth-death process. Sampling events are denoted by the horizontal dashed lines. Sampled tips are marked in red, and removal (death) is marked in blue. Once sampled, an individual is considered non-infectious. (B): Phylogenetic tree obtained by suppressing all unsampled tips from the transmission tree. (Adapted with changes from [51, 80]).

The sampling of infected individuals from this transmission tree is captured by

a sampling rate (or sampling probability) [142]. To obtain the reconstructed phylogenetic tree, all edges without sampled descendants are suppressed from the transmission tree (Figure 1.3). Therefore, the transmission tree describes the sequence of transmission amongst infected individuals, while the phylogenetic tree describes the evolutionary relationships between the viral samples obtained from each individual.

Furthermore, birth-death models have recently been used in conjunction with epidemiological compartmental models in studies of infectious dynamics [80]. Compartmental models describe viral dynamics by categorizing the host population into multiple classes. The most notable example is the Susceptible-Infected-Recovered (SIR) system of differential equations, which models fractions of the population as a function of time [75]:

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dI}{dt} = \beta SI - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases} \quad (1.2.2)$$

In this set of equations, β represents the transmission rate to susceptible individuals, and γ is the recovery (non-infectious) rate [75]. When coupled with the birth-death model, compartmental models provide a very detailed depiction of outbreak dynamics. This approach, termed BDSIR, provides an approximation to the stochastic SIR model [80]. Using the infection trajectories provided by the compartmental model described above, the average piecewise transmission rate is determined. The likelihood of the phylogenetic tree is obtained using these transmission time series in conjunction with the given epidemiological parameters. While the BDSIR is a biologically realistic model, the computation of the exact likelihood is inefficient due to the complexity of the epidemiological dynamics involved [120]. This limitation has motivated the application of statistical techniques such as Approximate Bayesian Computation to phylodynamic inference.

1.2.3 Approximate Bayesian Computation for parameter inference

A key challenge in the field of viral evolution is the inference of epidemiological parameters from phylogenetic tree shapes. The development of methods to address such problems has been facilitated by the continuous growth in computer processing power that is available to researchers. Computational approaches such as the software package BEAST (Bayesian Evolutionary Analysis by Sampling Trees) [36] enable the calculation of the posterior probability of a phylogenetic tree within a coalescent or a birth-death framework. While such Bayesian approaches can account for phylogenetic uncertainty in the tree reconstruction process, these approaches often require the calculation of a complicated likelihood function, as previously mentioned. Hence, Approximate Bayesian Computation (ABC) has become a central framework for parameter inference.

ABC methods circumvent the issue of a costly likelihood function calculation by employing simulations, the outcomes of which are compared with observed data [12]. The goal is to accurately infer the parameters of the empirical model by minimizing the discrepancy between the simulated dataset (phylogenetic tree) and the target empirical dataset. The closeness between the simulated and target datasets is quantified by a distance function. Complex information such as phylogenetic tree shape or tree balance (Figure 1.4) is difficult to evaluate; hence, summary statistics are used to effectively capture the information represented by each tree. Various summary statistics have been proposed for characterizing phylogenetic trees, such as Sackin's index [133] and the total cophenetic index [100] for tree balance, and the Robinson-Foulds metric [129] for topological distance. ABC approaches range from the most simple rejection sampler [12], to the more efficient sequential Monte Carlo sampler [29]. Through the application of ABC simulation methods, inferring the underlying characteristics of viral epidemics has become a much more accurate and informative

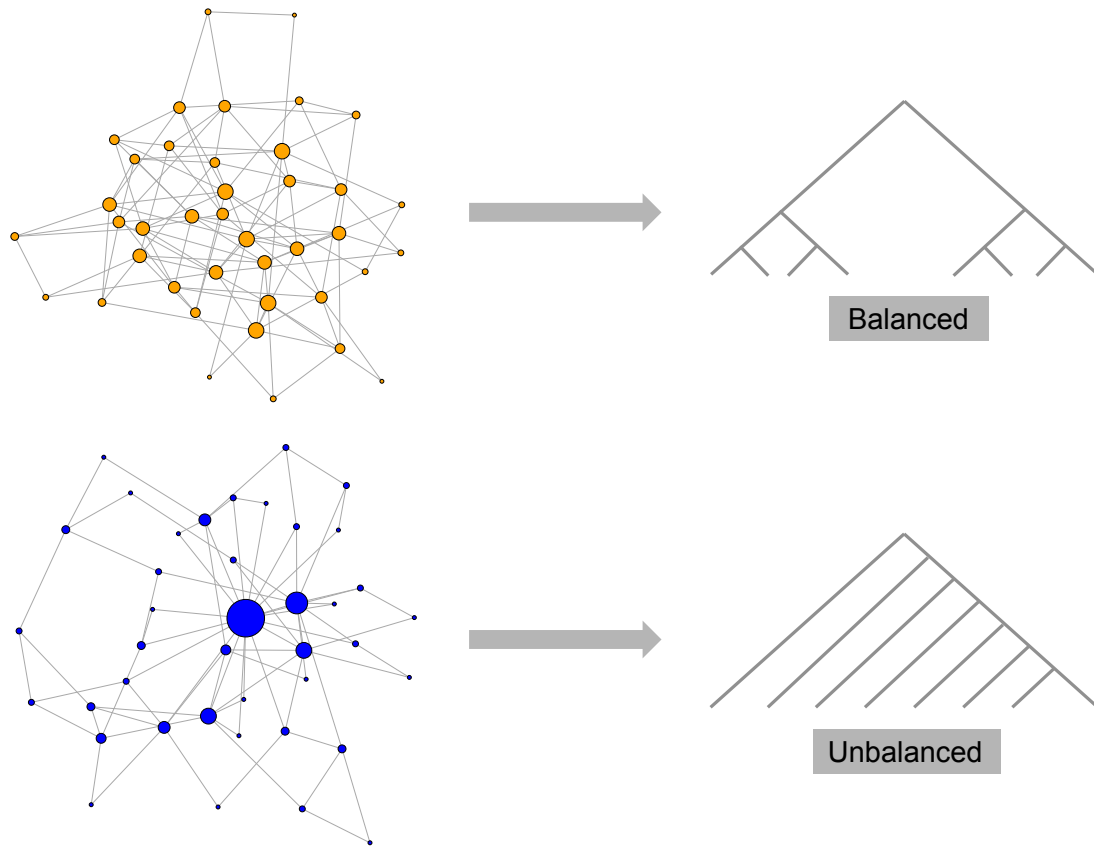


Figure 1.4: Host contact structure and phylogenetic tree balance. Node sizes in the contact networks reflect the degree (number of transmission events) of each node (host). The orange network depicts a balanced contact network, which yields a balanced viral phylogenetic tree. The blue contact network contains “superspreaders,” giving rise to an unbalanced viral tree.

process.

1.3 Objectives of the thesis

This thesis explores the intricate evolutionary dynamics of two of the deadliest RNA viruses: the Ebola virus and HIV. Using tools stemming from the fields of viral phylodynamics and molecular epidemiology, I address key questions pertaining to protein function and structure, virulence, diversifying selection, and viral transmission dynamics. In the case of the Ebola virus, I attempt to provide a functional and

mechanistic explanation for the evolutionary trajectory that shaped the 2014 outbreak. I then shift my focus to statistical inference and assess the scope and accuracy of an ABC approach when applied to HIV epidemic data.

The role that selection plays in the emergence of viral epidemics remains debated, particularly in the context of the 2014 Ebola virus (EBOV) outbreak. Most critically, should such evidence exist, it is generally unclear how this role relates to function and increased virulence. In Chapter 2, I show that the viral lineage leading up to the 2014 outbreak underwent a complex interplay between selection and correlated evolution (epistasis) in a protein region that is critical for immune evasion. The three-dimensional structure of this domain was reconstructed, and I show that the initial mutations along this lineage deformed the structure, while subsequent mutations restored part of the structure. Along this mutational path, the first and last mutations were adaptive, while the intervening ones were epistatic. Altogether, a novel mechanistic model is provided that explains how selection and epistasis acted on the structural constraints that materialized during the 2014 EBOV outbreak.

In Chapter 3, I demonstrate the application of an ABC-sequential Monte Carlo sampler to the inference of epidemiological parameters. The sequential Monte Carlo approach provides an efficient means of approximating the posterior by sampling from a sequence of predefined probability distributions. Using a wide range of epidemiological scenarios, viral epidemics are simulated and each phylogeny is summarized using a normalized count of the number of lineages present through time. The results of this implementation suggest that the ABC-SMC approach (coupled with the normalized lineages through time as a summary statistic) provides a reliable framework for inferring the transmission and removal rates of an epidemic. I then apply this algorithm to real data from the Swiss HIV-1 epidemic and demonstrate its ability to distinguish between topologically similar phylogenies. As such, this method will allow for real-time monitoring of viral epidemics, facilitating the validation of novel public health policies.

Finally, in Chapter 4, I leverage all of the above-mentioned results and conclusions, and discuss their relevance to the current state of viral phylodynamics. Links are made between the work presented here and current work addressing similar biological questions. Future directions for both projects are presented, with an emphasis on potential improvements to the experimental design.

Chapter 2

Both Epistasis and Diversifying Selection Drive the Structural Evolution of the EBOV Glycoprotein Mucin-like Domain

2.1 Abstract

Throughout the last three decades, Ebola virus (EBOV) outbreaks have been confined to isolated areas within Central Africa; however, the 2014 variant reached unprecedented transmission and mortality rates. While the outbreak was still underway, it was hypothesized that the variant leading up to this outbreak evolved faster than previous EBOV variants, but evidence for diversifying selection was undetermined. Here, we test this selection hypothesis and show that while previous EBOV outbreaks were preceded by bursts of diversification, evidence for site-specific diversifying selection during the emergence of the 2014 EBOV clade is weak. However, we show strong evidence supporting an interplay between positive selection and correlated evolution

(epistasis), particularly in the mucin-like domain (MLD) of the EBOV glycoprotein. By reconstructing ancestral structures of the MLD, we further propose a structural mechanism explaining how the substitutions that accumulated between 1918-1969 distorted the MLD one way, while more recent epistatic substitutions restored part of the structure, with the most recent substitution being adaptive. We suggest that it is this complex interplay between weak positive selection, epistasis, and structural constraints that has shaped the evolution of the 2014 EBOV variant.

2.2 Contributions

This work has been published in the Journal of Virology [68]. All data analysis and protein reconstruction was conducted by me. Jean-Claude Nshogozabahizi and Stéphane Aris-Brosou provided the code for the epistasis analysis. The manuscript was written by me and edited by Stéphane Aris-Brosou.

2.3 Introduction

The five viruses that constitute the genus *Ebolavirus* (*Zaire ebolavirus* (EBOV), *Sudan ebolavirus*, *Bundibugyo ebolavirus*, *Reston ebolavirus*, and *Tai Forest ebolavirus*) have been the cause of a major public health concern in sub-Saharan Africa for over three decades [9]. Historically, outbreaks have been confined to isolated areas within Central Africa; however, the 2014 outbreak reached an unprecedented level, making this the largest outbreak since the discovery of the virus in 1976 [67].

The primary analysis of EBOV isolates conducted by Gire and colleagues found a large number of nonsynonymous mutations between the 2014 EBOV sequences and all previously published EBOV sequences [52]. In particular, 50 fixed nonsynonymous changes were observed, all of which were unique to the 2014 EBOV variant. While these fixed mutations spanned all seven viral proteins (NP, VP35, VP40, GP, VP30,

VP24, L Polymerase), both epidemiological [158] and molecular [63] evidence suggest that these mutations did not affect the virulence or the transmissibility of the 2014 EBOV strain. Nonetheless, the authors of the primary study fell short of testing for the role of selection in the emergence of this outbreak.

A subsequent study further investigated the rate of nonsynonymous substitutions in the 2014 EBOV strain, focusing on the glycoprotein (GP) [117]. They discovered a strikingly high ratio of nonsynonymous-to-synonymous mutations concentrated in the heavily glycosylated mucin-like domain (MLD) of GP, a region that spans residues 308-501 [96, 151]. While it has been proposed that the MLD plays a key role in immune evasion [95], the flexible structure and dispensability of this domain [85] might explain its propensity to maintain high substitution rates throughout the outbreak. Alternatively, as GP plays a critical role in host cell attachment and membrane fusion [117], selective pressure exerted by the host's immune system may contribute to the high incidence of nonsynonymous mutations in this protein. Park and colleagues found a significantly higher rate of nonsynonymous mutations in experimentally validated GP epitopes than expected by chance, supporting the selection hypothesis [117].

While positive selection favors the strains (or phenotypes) best adapted to the changing host environment, epistatic interactions among these mutations are known, at least from a theoretical point of view, to increase the number and magnitude of diversification events [56]. Epistasis within and among genes is defined as the effect of one mutation being dependent on (*i.e.*, correlated to) the presence or absence of one or more different mutations [54]. The existence of such non-additive interactions within (and among) genes can be experimentally validated in a variety of viral genomes [41]. However, experimental validation is time intensive, cumbersome, and does not always capture the complex patterns of epistatic interactions [54]. Some of these complex patterns are manifested structurally through the stabilizing interactions among mutations. The presence of permissive (that expand the margin of

stability) and compensatory (that restore the stability of a compromised structure) mutations dictates the physical conformation of proteins, and may therefore offer critical insight into the evolutionary trajectory of the Ebola virus [16].

For the very first time here, we aim to disentangle how episodic diversifying selection and epistasis have shaped the evolution of the 2014 EBOV strain. Using the Mixed Effects Model of Evolution (MEME) [102] as a tool to test for site-specific evidence for positive selection, we identify in the EBOV genome codons that experienced diversifying selection in restricted lineages; using AEGIS (Analysis of Epistasis & Genomic Interacting Sites) [107], we identify sites that evolved under epistasis. We show that, in the MLD, these two categories of sites overlap along the lineage leading up to the 2014 sequences. By means of ancestral reconstructions of the MLD, we further develop a mechanistic model explaining how selection and epistasis brought about conformational changes that led to the MLD found in the 2014 sequences.

2.4 Materials and Methods

2.4.1 Sequence data

We retrieved from GenBank a total of 124 complete Ebola virus genome sequences previously published [52]. Of these, 99 EBOV genomes were from 78 confirmed 2014 Ebola virus disease patients, and two EBOV genomes (Guinea) were from the start of the outbreak in February 2014. Twenty-three additional genomes from earlier outbreaks were also retrieved in order to allow for a complete phylogenetic reconstruction of the *Ebolavirus* genus. These 23 genomes consisted of ten Sudan sequences, seven Reston sequences, five Bundibugyo sequences, and one Tai Forest sequence, all spanning 1976-2012.

2.4.2 Phylogenetic analyses

We aligned the nucleotide sequences (124) for each gene independently using MUSCLE [40]. Poorly aligned positions were eliminated from our alignment using default parameters in GBlocks [146]. Next, model selection was conducted based on the AIC as implemented in jModelTest [25]. The nucleotide alignments for VP40, VP24, and VP30 were analyzed with PhyML [59] using a (GTR+ Γ) model of nucleotide substitution, while a (TVM+ Γ) model was used for the proteins VP35, GP, NP, and L Polymerase. The resultant phylogenetic trees for each gene were visualized using FigTree [125].

2.4.3 Selection analysis

We tested for evidence of episodic diversifying selection within the Ebola virus genome using the Mixed Effects Model of Evolution (MEME) [102]. Available methods for identifying sites which undergo pervasive selection (constant through time) are not always able to detect episodic (sporadic) selection [57]. However, MEME, a tool for detecting site-specific diversifying selection, circumvents this issue of low sensitivity by identifying episodes of diversifying selection affecting a subset of branches at individual sites [5]. In particular, this analysis allowed us to identify codons under diversifying selection along the branch leading up to the 2014 EBOV strain. This data analysis was conducted using Datamonkey, a web-server of the HyPhy v2.12 package [31].

2.4.4 Timeline of amino acid substitutions

In order to map the timeline of accumulation of substitutions under selection in the Ebola virus genome, we analyzed genome data using a relaxed molecular clock. To obtain a tree of concatenated coding sequences, we first examined our full alignment for evidence of recombination using SBP (Single Breakpoint Recombination) of the

HyPhy v2.12 package [31]. Finding no recombinant sequences, an alignment of concatenated coding sequences was derived from our original seven protein alignments using Catfasta2phym [108]. This new alignment was analyzed in BEAST using a GTR+ Γ_4 model of nucleotide substitution, assuming an uncorrelated lognormal relaxed clock and a constant-population coalescent prior. We ran two separate Markov chain Monte Carlo samplers to check for convergence, visually assessed these with Tracer (tree.bio.ed.ac.uk/software/tracer), and removed a conservative 10% of the 1×10^{10} states as burn-in. The detected adaptive substitutions were then mapped onto the dated phylogenetic tree.

2.4.5 Detection of epistasis

To test for a link between sites under diversifying selection and those under epistasis, we ran an algorithm designed to detect correlated substitutions, as implemented in AEGIS (Analysis of Epistasis & Genomic Interacting Sites) [107]. Taking inspiration from a statistical approach originally developed to assess phenotypic correlation at pairs of traits [114], our approach determines the degree of correlation between pairs of amino acid positions. As the original model was developed to analyze binary traits, amino acid positions were recoded as ancestral/derived characters based on a majority rule consensus. More specifically, using the 2014 EBOV sequences as a reference, amino acids (at each site) which are present in at least 90% of the EBOV 2014 clade are denoted as “derived”, while those that differ from this reference are denoted as “ancestral”. The appropriate majority rule percentage was chosen based on the consensus amino acids observed across all 124 sequences.

2.4.6 Epistasis and selection as interacting agents in outbreak onsets

To elucidate how the interplay between epistasis and selection shaped the epidemiology of the virus, we reconstructed a lineage through time (LTT) plot derived from the BEAST tree. We used the `ltt.plot()` function from the R package `ape` [116] to do this.

2.4.7 Ancestral MLD structure predictions

To gain intuition into the biological significance of our results, we mapped the substitutions that are both under selection and involved in epistatic interactions on the 3D structure of the EBOV glycoprotein (PDB-ID: 3CSY) using Swiss PDB Viewer v4.1.0 [58]. The corresponding chains (A-H) of the Fab KZ52 human antibody complex were removed to facilitate visualization of the pre-fusion conformation of the GP trimer. However, all of the sites of interest (identified as under selection and/or epistasis) are within a region, the GP mucin-like domain (MLD), that was excised from protein 3CSY to aid its crystallization [83]. As such, homology modeling [15] could not be used to predict the structure of this region, and we had to resort to a more sophisticated approach. For this, we resorted to Robetta [76], a full-chain protein structure prediction server that uses the Rosetta software package [138]. Domain parsing and *ab initio* structure prediction of the GP MLD were carried out using the Ginzu [76] protocol method implemented in Robetta.

In order to understand the structural changes within this MLD at different time points along the lineage leading up to the 2014 sequences, we performed marginal ancestral reconstructions. Using our GP sequence alignment, we first mapped the branches along which each mutation of interest (identified as under selection and/or epistasis) occurred. This led us to selecting all internal nodes between the first and the last occurrences. At each of these internal nodes, we inferred the GP MLD from

the GP gene alignment and its corresponding phylogenetic tree using FastML [7], a web server for maximum-likelihood reconstruction of amino-acid sequences. To reconstruct a representative 2014 GP MLD, a sixth sequence was chosen at random from the 2014 clade (GenBank accession number: KM233078). We were able to select a random representative sequence without compromising our analysis due to the high global similarity and low pairwise distance among the GP MLD region of the 2014 EBOV sequences. Using the BLOSUM62 scoring matrix [62] implemented in CLUSTAL W [82], we obtained an average global similarity score of 0.996 (min: 0.89, max: 1.00) comparing the MLD region among the 81 2014 sequences. To further test the level of similarity within the MLD region of our 81 2014 sequences, we obtained an average pairwise distance using ConSurf 2010 [6], a server for calculating evolutionary conservation in protein sequences. With an average pairwise distance of 1.02×10^{-7} (compared to 0.67 for the entire alignment of 124 sequences), our 2014 sequences exhibit a high level of conservation within the MLD region of the glycoprotein. Given these findings, any one of our 2014 sequences would yield accurate MLD reconstructions representing the 2014 clade.

2.4.8 MLD structure evolution

To understand how selection and epistasis affected the structure of the MLD, we compared each predicted structure with the oldest reconstructed structure. To do this, we first performed structural alignments with Swiss PDB Viewer v4.1.0 [58]. From these, we derived root mean squared deviations (RMSDs), and plotted these as a function of time (as derived from the BEAST analysis).

Because of the complexity of the MLD structural changes, we further computed difference distance matrices among pairs of residues for each structural alignment with SuperPose, a server specializing in structure alignments and distance statistics [92]. Difference distance matrices were obtained by first generating a distance matrix for

each molecule (pairwise distances between all pairs of aligned backbone alpha carbon atoms), and then subtracting these matrices from one another. Corresponding heat maps were then plotted to reveal where and when structural divergence/convergence occurred and hence explain the interplay between epistasis and selection on the evolving shape of the MLD. A brief illustration on how these maps were obtained is available here https://en.wikipedia.org/wiki/Distance_matrix.

2.5 Results and Discussion

Major Ebola outbreaks are preceded by diversification events

First, using a lineage-through-time (LTT) analysis, we reconstructed the diversification of the Ebola virus through time and showed that all known major outbreaks are immediately preceded by a distinct diversification event (Figure 2.1: A). Such a pattern prompts the question of whether these diversification events are the result of diversifying selection acting episodically just prior to each outbreak.

2.5.1 Evidence for selection is weak

To test this hypothesis, we ran a MEME analysis that aims at detecting evidence for diversifying selection acting episodically, *i.e.* along particular lineages of the tree, without specifying these lineages *a priori* – in order not to bias our analysis towards ‘outbreak lineages’. This analysis revealed that of the seven proteins, three (GP, L Polymerase, and NP) were evolving under episodic diversifying selection, with $p \leq 0.01$ (Table 2.1). GP, L Polymerase, and NP play a major role in viral replication and host cell attachment [27], which suggests that these three proteins are the most vulnerable to selective pressures exerted by the host cell.

However, these adaptive substitutions do not show the phylogenetic clustering (Figure 2.2) that we expected if they were, on their own, responsible for viral diver-

2. Both Epistasis and Diversifying Selection Drive the Structural Evolution of the EBOV Glycoprotein Mucin-like Domain 32

Table 2.1: MEME analysis results for codon-specific episodic diversifying selection in the Ebola virus genes. Sites which are also involved in epistatic interactions are in bold type.

	L Polymerase	GP	VP40	NP	VP35	VP30	VP24
Number of Codons	2212	676	326	739	341	288	251
Model of Evolution ^a	GTR	GTR	GTR	GTR	GTR	GTR	GTR
Detected Codon ^b (p-value; q-value)	226 (0.0066; 1)	44* (0.0041; 1)	N/A	35* (0.0059; 1)	N/A	N/A	N/A
	790 (0.0039; 1)	215 (0.0062; 1)		111* (0.0004; 0.3176)			
	1060 (0.0099; 1)	308 (0.0061; 1)		398 (0.0026; 0.9617)			
	1364 (0.0038; 1)	428* (0.0044; 1)		721 (0.0036; 0.8912)			
	1694 (0.0096; 1)						

^aThe GTR (general time-reversible) model was selected as the best fitting model for all genes.

^bN/A denotes genes which had no signal of episodic diversifying selection.

*These codons were detected along the 2014 EBOV lineage.

sification bursts identified in the LTT analysis.

This disagreement is not caused by improper model selection; this MEME analysis was based on AIC model selection, which showed that the best model to use was GTR. To ensure that the MEME results were robust to this model choice, we ran the analysis again under a range of models of evolution and found that the location of inferred sites undergoing diversifying selection did not change depending on the model used: these substitutions still do not cluster on lineages leading up to outbreaks.

On the other hand, because the MEME analysis involves multiple testing of sites and branches, it may be critical to correct the test procedure by computing the false discovery rate of sites undergoing episodic diversifying selection [102]. Yet, after correcting for FDR in our analysis, no statistically significant sites remained.

Just like ours, a previous study also found strong evidence for episodic diversifying selection within the GP and L Polymerase proteins [153], at a 5% significance level and without correcting for FDR. Increasing our significance level from 1 to 5%, we found that only one residue, amino acid position 1492 of the L Polymerase protein, was detected in both our study and the previous analysis, with no other common sites. One possible explanation for this discrepancy may be that the previous study only used 78 EBOV sequences (and 13 additional unnamed sequences) instead of the 124 full genome sequences used here.

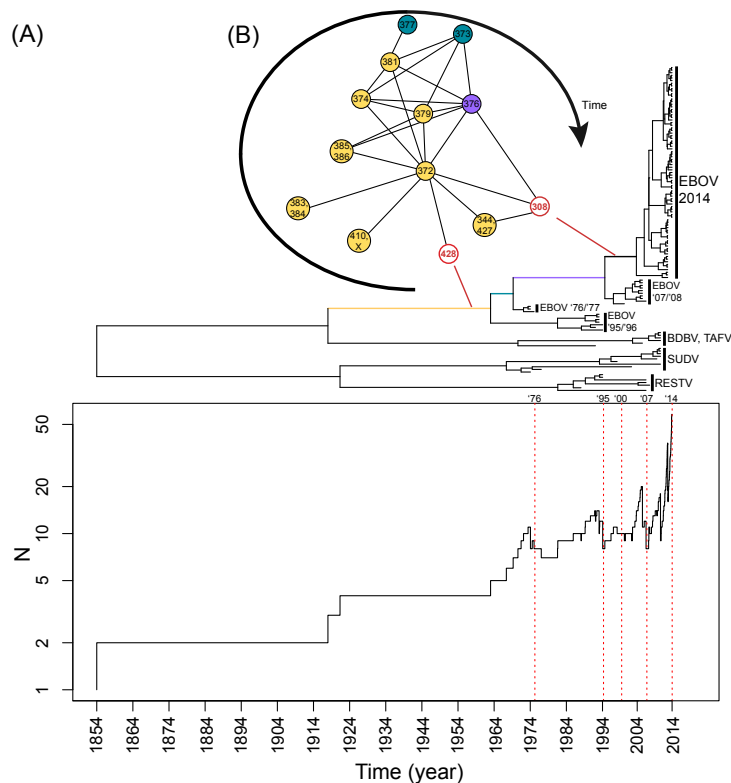


Figure 2.1: Lineage through time (LTT) plot derived from dated BEAST tree. (A): Number of lineages (log-transformed) as a function of time (years). The five major Ebola virus outbreaks are marked in red dashed lines (1976, 1995, 2000, 2007, 2014). (B): The glycoprotein (GP) network of epistatic sites is mapped onto the BEAST tree, with the emergence of the adaptive substitutions (red circles) indicated along the respective branches (solid red line). The “X” represents a group of sites with identical site patterns (i.e. identical recoding pattern based on majority rule consensus, as described in the text). This group of sites is listed in full in Table S1. The branches along which each mutation emerged are color coded (yellow, turquoise, purple). The elliptic arrow represents the timing of emergence for substitutions, starting from sites that mutated along the yellow branch to substitution at position 308; a full chronology for each site could not be inferred from the alignment.

2.5.2 Evidence for epistasis links to selection

In the absence of strong statistical evidence supporting selection at these positions, we assessed their potential biological significance. Based on the Bayesian relaxed molecular clock analysis conducted for the LTT analysis, we reconstructed a timeline of the emergence of these substitutions putatively under selection (Figure 2.2). The sequential and sustained mode of emergence of these adaptive substitutions suggests that epistatic interactions may have played a role in the evolution of the 2014 EBOV

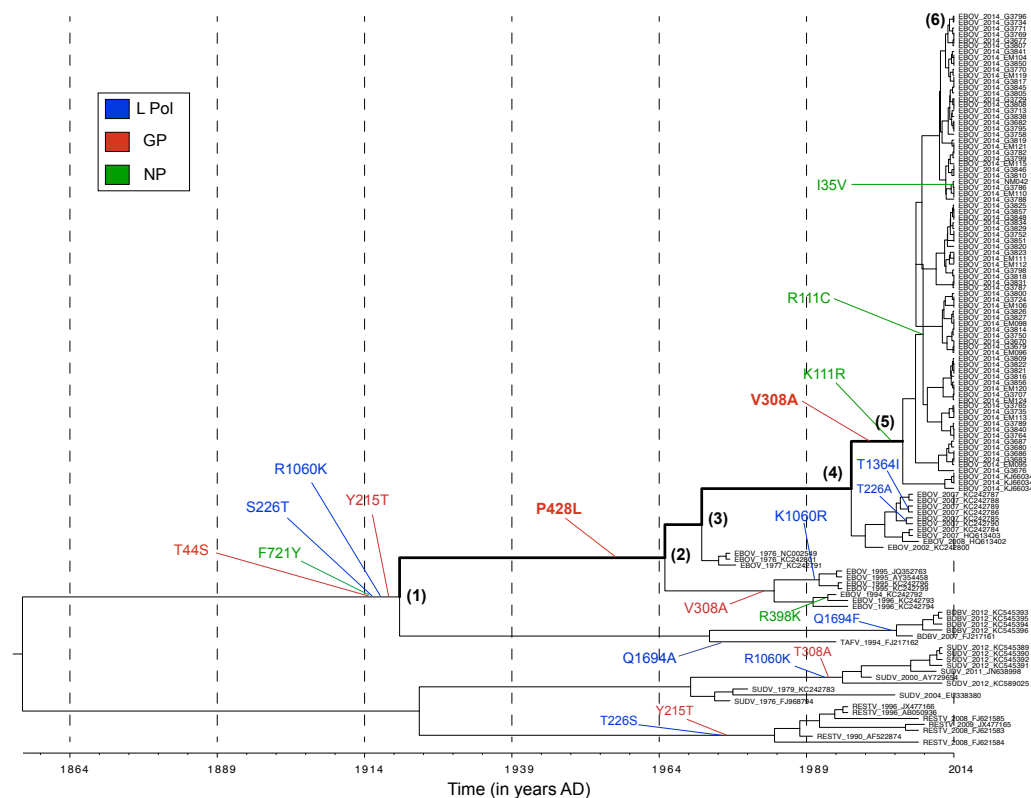


Figure 2.2: Timeline of adaptive amino acid substitutions using a relaxed molecular clock phylogenetic tree. Sites detected as evolving under episodic diversifying selection (MEME analysis) for three out of seven Ebola virus genes (L Polymerase, Glycoprotein, and Nucleoprotein) are mapped onto the dated BEAST phylogenetic tree. Numbers in brackets denote the five internal nodes and one 2014 EBOV sequence (GenBank accession number: KM233078) used for the ancestral reconstruction of the GP mucin-like domain. The GP sites (308 and 428) around which these nodes were chosen are in red bold-face type.

clade.

To test this hypothesis, we analyzed each protein alignment with AEGIS [107]. At a significance level of 4×10^{-4} (after FDR), we detected epistatic interactions in all seven Ebola virus genes (Tables S1-S7).¹ As previously uncovered [107], we found some overlap among pair members of the epistatic interactions, so that long chains or networks of epistatic sites could be reconstructed by transitivity (Figure 2.3).

Critically, three of these networks of epistatic sites also contained sites potentially under selection (Table 2.1): GP, L Polymerase, and NP. This kind of overlap suggests

¹Supplementary files: <http://jvi.asm.org/content/90/11/5475/suppl/DCSupplemental>

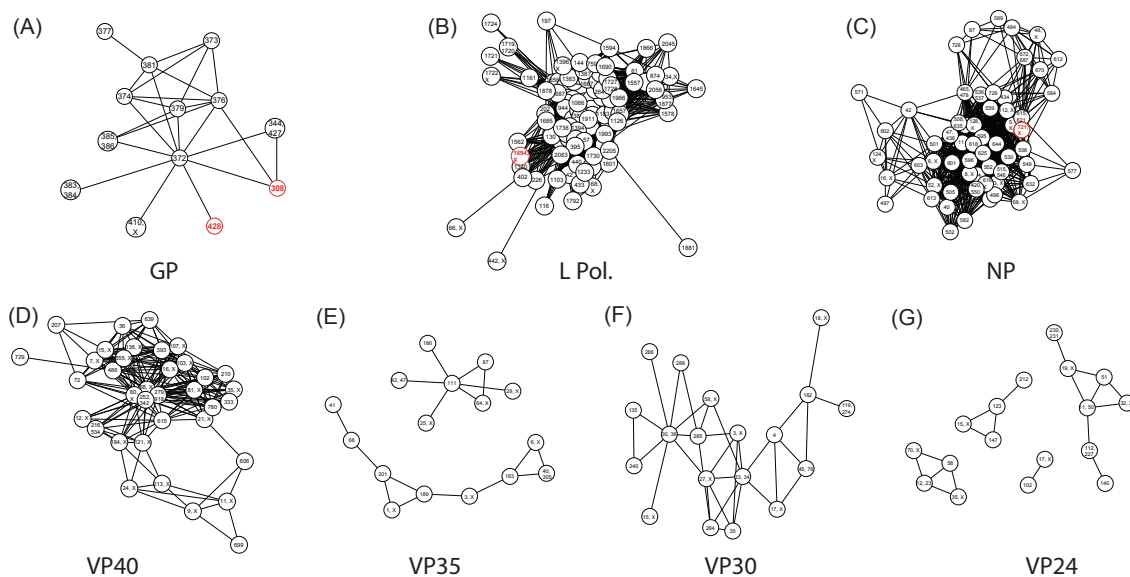


Figure 2.3: Detected epistatic interactions within the seven EBOV genes. (A-G): Epistatic interactions detected in all seven genes at a significance level of 4×10^{-4} (after FDR). Each circle represents an amino acid site, and connecting lines denote epistatic interactions detected by AEGIS. Sites in red font are those that were also detected as adaptive by the MEME analysis for episodic diversification. Overlapping site labels were omitted for ease of visualization; all interacting pairs are listed in Tables S1-S7.

a link between epistasis and selection. Such a link was theoretically shown to be possible in prokaryotes [56], but to the best of our knowledge, it has never been shown to exist in actual data, let alone in viruses such as Ebola.

2.5.3 Selection and epistasis are linked to conformation and function in the MLD

To investigate the nature of this link between selection and epistasis, we focused on the GP gene for three reasons. First, GP has become the primary target of drug design and human treatments. The newly developed recombinant, replication-competent, vesicular stomatitis virus-based vaccine (rVSV-ZEBOV) expresses EBOV glycoproteins in order to elicit an immune response against the complete virus [61]. Second, only in this gene are the detected epistatic sites clustered in a small region

2. Both Epistasis and Diversifying Selection Drive the Structural Evolution of the EBOV Glycoprotein Mucin-like Domain 36

of the protein. Indeed, all positions are in the tight [308, 428] amino acid range (Figure 2.3) and are physically close to each other on a 3D model of the protein structure (Figure 2.4). Possibly, this clustering of epistatic sites could be the product of a relatively high mutation rate in this region, reflecting its dispensability for cell entry [85], but this possibility has, to date, not been fully investigated. Third and most critically, this small cluster is right in the mucin-like domain (MLD), a region that spans residues 308-501 [96, 151] and that has been hypothesized to shield the glycoprotein from antibodies [95, 136].

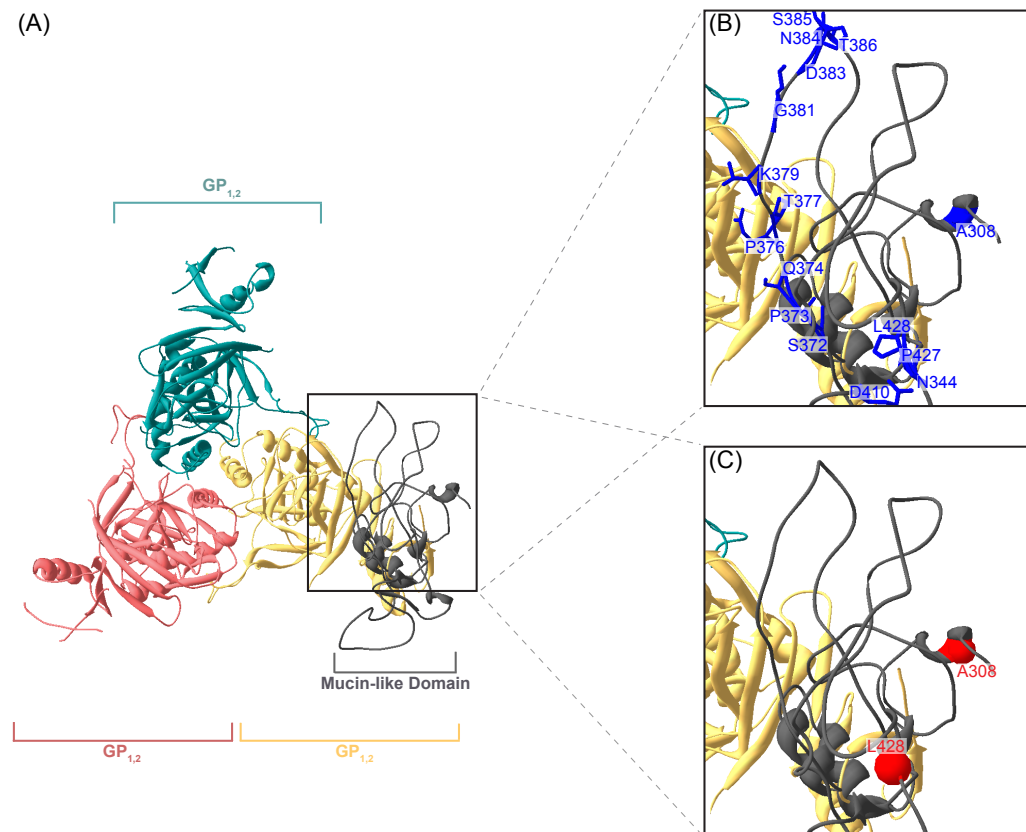


Figure 2.4: Molecular representation of EBOV GP trimer crystal structure. (A): Top view of the three GP_{1,2} monomers in pre-fusion conformation (PDB-ID: 3CSY). The Robetta structure prediction of the mucin-like domain (gray) is superimposed onto one of the GP_{1,2} monomers. The putative location of the MLD was determined using the MagicFit function in Swiss PDB viewer. (B): A close-up view of the GP MLD showing distribution of epistatic sites, each labeled in blue. (C): A close-up view of the GP MLD showing the position of the epistatic sites 308 and 428 that were also detected as adaptive by the MEME analysis.

Several lines of evidence support why episodic diversifying selection coupled with epistasis is operating on the GP gene. Most notably, the MLD is critical for evading the humoral immune response through the inhibition of surface protein recognition [95]. This region also contains multiple binding sites of early cocktail antibodies that are, however, non-neutralizing because the MLD is excised by host cathepsins in the endosome [26, 101]. It has also been speculated that in the absence of cleavage by furin, the host protein necessary for GP activation, it is the flexible MLD that provides the necessary conformational freedom within the internal fusion loop to activate GP [85]. As such, substitutions within this domain may be necessary to facilitate a metastable conformation of GP prior to fusion. The possibility of such compensatory interactions is made evident by the tight physical clustering of our detected sites, which suggests an interaction network driven by steric hinderance [113]. In such a network, any conformational change in one of the amino acids may necessitate a compensatory change in another neighboring amino acid. As such, this would lead to an intriguing hypothesis: a series of epistatic changes might have been required at some point in time to compensate for a previous substitution. Such an epistatic constraint, present for functional reasons, might have favored the emergence of an initial adaptive substitution (at position 428) (Figure 2.5: B). While conferring a selective advantage, this substitution might have distorted the 3D structure of the MLD, thereby requiring some epistatic changes for conformational reasons. These changes, in turn, permitted the emergence of a second adaptive substitution (at position 308), which appeared in the 2014 Ebola strain (Figure 2.2).

2.5.4 Adaptive substitutions as endpoints of epistatic changes

In order to test the plausibility of this scenario, we reconstructed the 3D structure of the MLD at key time points: from just before the emergence of the first adaptive substitution (at position 428) all the way to the emergence of the second adaptive

2. Both Epistasis and Diversifying Selection Drive the Structural Evolution of the EBOV Glycoprotein Mucin-like Domain 38

substitutions (at position 308). This led us to five reconstructions at nodes placed on the lineage leading up to the 2014 sequences (Figure 2.2: nodes 1-5), and one reconstruction representing a 2014 EBOV sequence (Figure 2.5); all six models are available as PDB files in the supplementary information. The pairwise superimpositions of the five latest MLD structures to the earliest structure reveal an approximate trajectory of structural divergence through time (Figure 2.6).

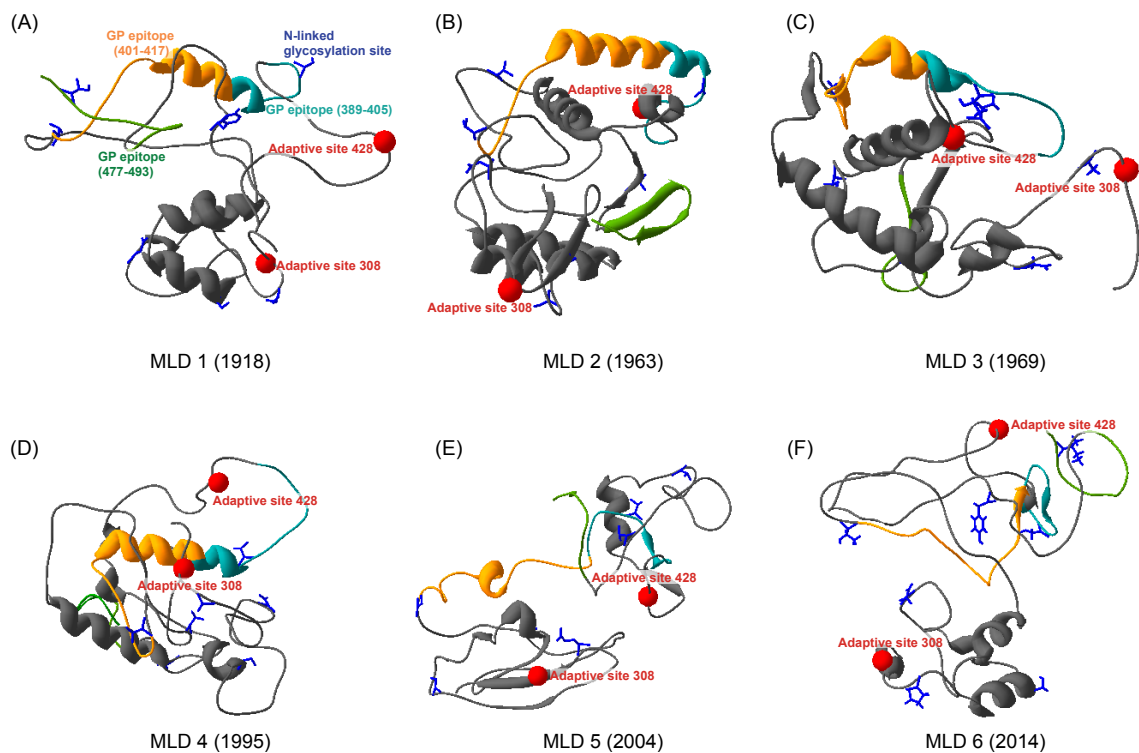


Figure 2.5: All 6 MLD ancestral reconstructions with labeled sites of interest. (A-F): MLD reconstructions for ancestral nodes at time points (years) 1918, 1963, 1969, 1995, 2004, 2014, respectively. The locations of GP linear epitopes within the MLD are labeled for each reconstruction, with amino acid numbers in brackets (panel (A) only). GP epitopes 401-417 (orange) and 389-405 (teal) overlap at sequence ‘ATQVE’. Known N-linked glycosylation sites throughout the MLD are marked in blue (stick representation is used). The locations of the detected adaptive sites (from the MEME analysis) are shown in red.

Notably, the difference distance plot of MLD1 vs MLD5 reveals a distinct square area of low amino acid distance bounded by the two adaptive substitutions at sites 308 and 428 (Figure 2.6: 1 *vs.* 5). These sites, therefore, seem to act as beginning

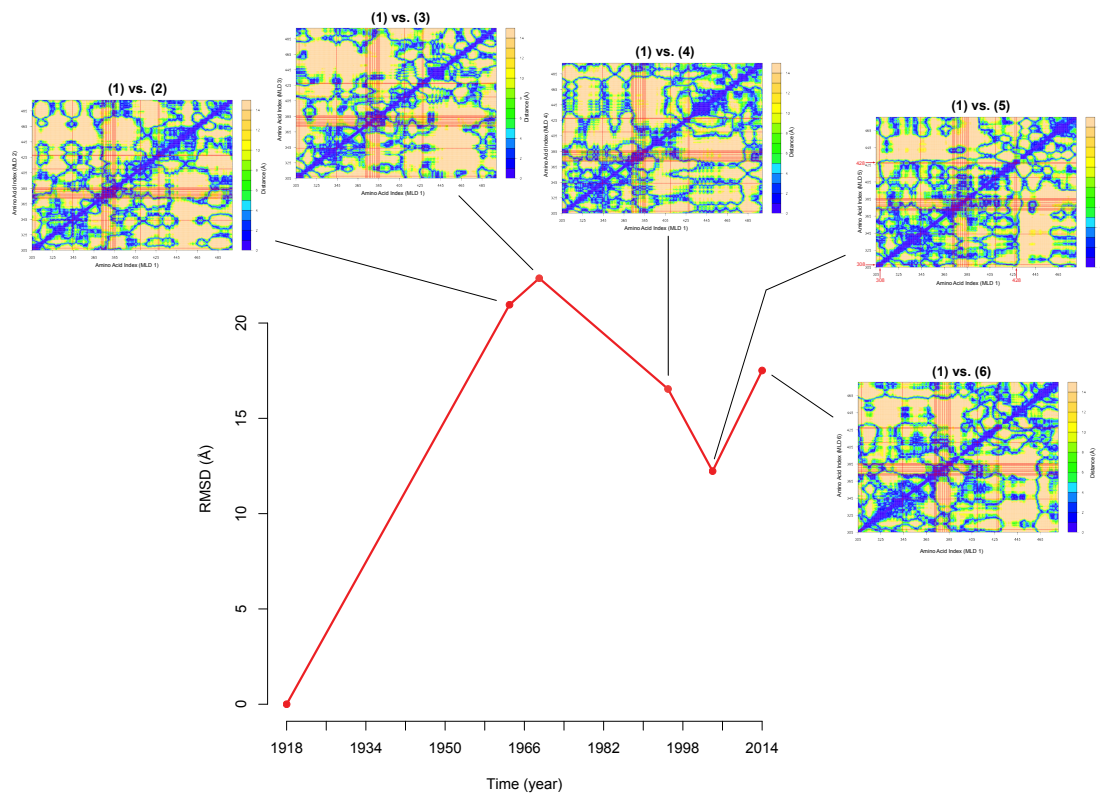


Figure 2.6: Root Mean Squared Deviation (RMSD) plot of MLD structural alignments. (Main): RMSD as a function of time for six reconstructed 3D structures of the GP MLD at key time points (before and after adaptive substitutions). The oldest reconstructed structure (year 1918) was used as a reference upon which each of the remaining structures were superimposed. The six plotted points (1918, 1963, 1969, 1995, 2004, 2014) correspond to the following structure comparisons: MLD1 vs MLD1, MLD1 vs MLD2, MLD1 vs MLD3, MLD1 vs MLD4, MLD1 vs MLD5, and MLD1 vs MLD6. (Insets): Corresponding heat maps of amino acid distance are plotted for each pairwise comparison (except for MLD1 vs itself). 2D locations of epistatic GP amino acid sites are represented with vertical and horizontal red lines. Adaptive sites 308 and 428 are annotated in red on the MLD1 vs MLD5 heat map.

and endpoints for the structural reestablishment of the mucin-like domain – both in time and 3D space. We suggest that adaptive mutation P428L arose as a permissive substitution that first allowed for a wide array of substitutions to accrue, changing the MLD structure considerably up to year 1966. The new conformation of the MLD then might have required compensatory changes, that (i) drastically reduced the extent of divergence from the original (1918) MLD structure, potentially creating a much more adapted glycoprotein secondary structure, and (ii) facilitated the emergence of the adaptive mutation at site 308, which appeared in the 2014 strain.

It can be noted that the structural similarities between the six reconstructions (Figure 2.5) do not seem to coincide with the global RMSD values obtained for the pairwise comparisons (Figure 2.6). For instance, MLD 1 and MLD 5 appear more diverged than MLD 1 *vs.* MLD 4 (Figure 2.5), while our global RMSD values show that the MLD 1 reference structure is the most similar to the MLD 5 structure. In cases where the aligned structures do not have identical or very similar sequences (as seen with the MLD), RMSD calculations only consider backbone alpha carbons and disregard side chain residues [53]. Hence, RMSDs do not reflect the rotameric states of all side chain residues shown in (Figure 2.5).

2.5.5 Insight into the GP MLD structure

The computational nature of our ancestral reconstructions of the MLD may make them somehow approximative, yet they are based on one of the best performing algorithms for *ab initio* structure prediction [76, 124]. As such, these reconstructions provide some structural insight that has not been previously available in the literature. In particular, we can better understand the spatial configuration of the GP MLD, in the absence of a crystallized structure to work from. Our structural prediction and superimposition (Figure 2.4) onto the glycoprotein (PDB-ID: 3CSY) show that the MLD is situated at the top-most vertex (spike) of each GP1 monomer, and protrudes at the sides of the GP structure. These results agree with previous predictions and findings about the MLD [151]. Each GP1 monomer contains a moderately glycosylated glycan cap, which is located atop the trimeric GP spike [83]. The spatial orientation of the predicted MLD (Figure 2.4) suggests that there is no physical contact among all three MLDs on the GP structure. Our reconstruction therefore supports the proposition that the MLD extends outwards from this glycan cap, forming a heavily glycosylated blanket across the top of the GP [37, 83].

Although the MLD has been predicted to be a generally unstructured region

[151], it contains three polypeptide sequences that have been experimentally identified as linear epitopes [159]. For the Zaire species in particular, antibodies 13F6-1-2, 6D8-1-2, and 12B5-1-1 target residues 389-405, 401-417, and 477-493, respectively [84, 86]. We can now highlight these epitopes on our ancestral reconstructions to gain functional insight through time. Our models suggest that both relative position and structure of these three epitopes vary widely from the first reference time point (1918) to the last (2014; Figure 2.5). Notably, the first four MLD reconstructions (Figure 2.5: A-D) predict the GP epitopes 389-405 and 477-493 as alpha helices, which then become highly unstructured regions in the last two reconstructions (Figure 2.5: E-F). The loss of the alpha helix secondary structure at these epitopes between 1995 and 2004 is intriguing and may warrant experimental data to understand its implications. Nonetheless, our reconstructions support previous hypotheses that regions involved in critical interactions with the host proteins are often these unique alpha helices [8, 21]. Note that one of the experimentally validated GP-bound linear epitopes (401-417) for 1976 and 1995 Ebola Zaire isolates [159] contains two of our detected epistatic sites: amino acids 410 and 414 (Table S1). Although this overlap on its own does not substantiate that epistasis is a major factor in epitope-antibody interaction, it opens up potential avenues for future epitope investigation. It is likely that these apical surface epitopes are more exposed and more flexible for antibody interaction than other GP epitopes situated deeper inside the GP spike, allowing the MLD to tolerate a greater amount of substitutions (and consequent epistatic interactions).

The MLD is also populated with N- and O- linked glycosylation sites [85, 87], some of which border the above-mentioned MLD linear epitopes. While our reconstructions (Figure 2.5) provide the structural positions of the validated N-linked glycans in the MLD, very little can be gathered about the potential structural consequences of these glycans on MLD function. It has been previously demonstrated that N-linked glycans are not critical for GP folding and that their removal from the MLD does not affect antiserum sensitivity [87]. It is therefore possible that the N-linked

glycans, which flank the GP epitope residues 389-405 and 401-417, influence the stability of this region, but are unlikely to bear any effect on antibody sensitivity, as previously investigated [87].

2.5.6 Linking selection, epistasis, and outbreak onsets

While our initial MEME analyses failed to show evidence for selection in most other major outbreaks, our epistasis analyses show widespread evidence for correlated evolution throughout the Ebola genome (Tables S1-S7). This begs the question as to whether epistasis can act alone to accelerate divergence. Based on the above findings, we argue that selection requires epistatic interactions, which are in turn driven by a tight physical linkage among amino acids.

Although experimental evidence of our argument would lead to the usual dual use conundrum [13] and would hence most likely be unfeasible, we presented here tantalizing evidence supporting a key role for epistasis, in conjunction with selection, in the diversification process leading up to the 2014 Ebola outbreak. We further support our claim with structural evidence, linking all processes together.

Chapter 3

Inferring viral epidemiological features from HIV genomic data using Approximate Bayesian Computation with Sequential Monte Carlo

3.1 Abstract

The use of phylogenetic methods for the reconstruction of viral ancestral relationships has garnered increasing interest, particularly in the characterization of HIV epidemics. The Swiss HIV Cohort Study (SHCS) represents a wealth of genomic data to which such methods have been applied. However, current approaches for quantifying the epidemiological dynamics of diseases are computationally intensive and fail to scale well with this magnitude of data. Here, we utilize an Approximate Bayesian Computation (ABC) approach based on sequential Monte Carlo (SMC). By

means of simulations, we show that our implementation is capable of inferring key epidemiological parameters of the Swiss HIV epidemic with great confidence. This ABC-based method circumvents the issue of a costly likelihood function calculation by using simulations, the outcomes of which are compared with observed data. Our results reflect the unique transmission dynamics of two differing subsamples from the SHCS. Moreover, sampling intensity has no significant effect on the accuracy of our estimates. Given these findings, we propose that this method will allow us to evaluate the impact of new public health policies (e.g., implementation of a needle exchange program in the case of HIV) based on genetic data sampled before and after the implementation of a new policy.

3.2 Contributions

All data analysis (simulations and parameter inference) was conducted by me. Stéphane Aris-Brosou provided guidance throughout the project. This manuscript was written by me and edited by Stéphane Aris-Brosou.

3.3 Introduction

Understanding the transmission patterns of infectious diseases is critical in order to monitor both their spread and the efficacy of public health policies. Uncovering these patterns requires quantifying epidemiological dynamics to obtain key insights into an outbreak. The most informative of these key quantities are the reproductive ratio (R_0) and the growth rate (r) of the epidemic. R_0 represents the number of secondary infections caused by a single infected individual (within a susceptible population) during their entire infectious period [18]. Therefore, it informs health officials of the transmission potential in a host population. Generally, $R_0 < 1$ reflects an epidemic that is under control, while $R_0 > 1$ reflects the opposite [90]. While the utility of

this threshold parameter as a measure of eradication has been questioned [90], it still provides a critical insight into an epidemic. The parameter r represents the speed at which the total number of infected individuals grows in a susceptible population [154]. So, together, the magnitudes of R_0 and r are used to predict the risk and the extent of spread of an emerging infectious disease.

Traditionally, both the estimation of these parameters and the reconstruction of transmission events have relied heavily on epidemiological surveys that are valuable, yet time-intensive [23]. However, recent developments in the field of phylodynamics have facilitated the use of phylogenetic methods for reconstructing ancestral relationships solely from viral sequence data, thus accelerating the inference of key quantities [23, 154]. These methods allow us to create pathogen phylogenies that are used to infer critical parameters relating to transmission dynamics within host populations. Two processes that are most commonly used for such inferences are the coalescent and the birth-death model. In essence, the coalescent takes n sampled individuals and traces their ancestral lineages backwards in time [148]. Historically, the coalescent has been the standard genealogical model in phylogenetic studies at the population level [78]. However, its inability to account for dense sampling of infected individuals, and its lack of a separate estimation for transmission and death rates, are crucial limitations that make it unsuitable for modeling epidemiological transmission [18].

Conversely, the birth-death model (BDM) provides a stochastic framework for the tree-generating process, thereby overcoming the limitations presented by the coalescent. Unlike the classic coalescent model, the BDM assumes stochastic changes in population size, where an individual can be born or can die at any point in time [141]. In an epidemiological context, a birth event corresponds to an individual becoming infected, while a death event corresponds to the removal of an infected individual from the population (death, treatment, or a change in habits) [141]. The BDM thus makes it possible to estimate parameters such as the transmission rate (λ) and the removal rate (μ) of an epidemic, which are in turn used to determine the reproductive

ratio($R_0 = \lambda/\mu$) and the growth rate ($r = \lambda - \mu$).

The aforementioned tree-generating processes are subject to both stochastic and systematic errors. In order to account for such errors, Bayesian inference has extensively been applied to phylogenetic reconstruction [2, 4, 130, 131]. These approaches, however, often require the calculation of a complicated likelihood function. Let us define the parameter vector as θ and the data as x . Given a prior distribution $p(\theta)$ and a likelihood function $p(x|\theta)$, the goal is to estimate the posterior distribution $p(\theta|x) \propto p(x|\theta)p(\theta)$, which is the probability distribution of the parameter vector θ having observed the data x . In cases where the likelihood function is too complex to derive [127, 165], an alternative approach is required. Hence, Approximate Bayesian Computation (ABC) has become a central framework for parameter inference [150]. ABC methods circumvent the issue of a costly likelihood function calculation by employing simulations, the outcomes of which are compared with observed data [12, 121]. In rejection sampling, the simplest ABC algorithm, closeness is determined using a distance function (d) and a tolerance value (ϵ). If the difference between the observed (x_0) and simulated (x^*) datasets $d(x_0, x^*)$ falls below this threshold, then the candidate parameter vector at the n^{th} step of the algorithm (θ_n) should be accepted. Otherwise, one rejects the proposed parameter and samples a new candidate parameter. These steps are iterated for a fixed number of samples, N . The ABC rejection sampler takes on the following form [121]:

Algorithm 1 Rejection ABC

```

1: procedure
2:    $n \leftarrow 1$ 
3:   while  $n < N + 1$  do
4:     repeat
5:        $\theta_n \sim p(\theta)$ 
6:        $x^* \stackrel{\text{sim}}{\sim} p(x|\theta_n)$ 
7:       until  $d(x_0, x^*) \leq \epsilon$ 
8:        $n \leftarrow n + 1$ 
9:   end while
10:  return  $\theta_1, \dots, \theta_N$ 
11: end procedure

```

One of the major limitations of the rejection sampler is that the acceptance rate is low when the prior distribution strongly differs from the posterior distribution. Thus, an ABC method based on Markov chain Monte Carlo was introduced to overcome this problem [93]. Although ABC-MCMC is guaranteed to converge to the target approximate posterior distribution, it also possesses some potential disadvantages. Notably, the efficiency of the algorithm is relatively low due to the correlated nature of samples coupled with a potentially low acceptance rate [150]. In order to alleviate the aforementioned issues, a sequential Monte Carlo (SMC)-based simulation approach was designed. By propagating sampled parameter values (particles) through a sequence of intermediate distributions defined by a tolerance schedule, parameters that represent a sample from the target posterior distribution will eventually be obtained [150].

The key to an efficient ABC sampler is the choice of an appropriate summary statistic for the data. To compute the distance between a simulated and an observed

phylogeny, each data set must be summarized in a manner that captures the information contained therein. Various summary statistics have been proposed for characterizing phylogenetic trees, such as Sackin’s index [17, 133] and the total cophenetic index [100], but very few of these can effectively capture the underlying epidemiological dynamics of a viral phylogeny. Specifically, most phylogenetic summary statistics are not informative enough to be used as a replacement for the likelihood in the parameter inference problem. Often, however, the comparison of two phylogenies has been made possible through the use of their Lineages-Through-Time (LTT) curves [104, 115]. Since the LTT curve of a phylogenetic tree depicts the number of lineages in the phylogeny through time, it reveals the distinct patterns of speciation and extinction that are specific to each tree. To avoid issues that arise when comparing trees of different sizes, Janzen and colleagues [70] proposed the normalized LTT (nLTT), which makes it possible to compare trees with differing height and number of extant tips.

Although attempts of applying ABC methods to the parameter inference problem have been made, these approaches are computationally intensive and fail to scale well with large data sets [65, 144]. Moreover, previous attempts at estimating contact structure as a means of making epidemiological inferences proved to be highly complex [89]. Here, to avoid these obstacles, we test the application of an ABC-SMC sampler to the inference of epidemiological parameters. By means of simulations, we show that ABC-SMC yields reliable and accurate estimates of the transmission (λ) and removal (μ) rates of an epidemic. Due to its superior performance against commonly used tree statistics [70], we hypothesized that the nLTT would be an appropriate summary statistic here. We validate our approach using data from the Swiss HIV-1 epidemic as a case study [145], where two subsamples obtained from different regions of the country are expected to exhibit unique transmission patterns. We report distinct parameter estimates for each region, highlighting the sensitivity of our approach.

3.4 Materials and Methods

3.4.1 Birth-death tree simulations

Taking inspiration from Stadler’s implementation of the birth-death model [142, 144], our epidemiological model follows a constant-rate birth-death process. The birth (transmission) rate with which one infected individual infects another uninfected individual is λ , and the removal rate in our model (μ) represents the rate at which an infected individual becomes non-infectious. Traditionally, the tree-generating process under the birth-death model has been executed under complete sampling. Recently, however, the model was extended to account for incomplete sampling by combining the birth-death process with a parameter called the sampling probability (ρ) [104, 141, 164]. Through this forward-in-time description of the epidemiological process, the BDM specifies sampling intensity and thus enables explicit modeling of the serially sampled infected class.

The birth-death model employed here produces transmission trees in the following manner. A transmission event is represented by a split (bifurcation) of one lineage (infected individual) into two lineages. Conversely, becoming non-infectious is represented by a truncated lineage. After the complete transmission tree is generated, a subset of the infected individuals is randomly selected to be included in the observed phylogeny. Each tip is sampled with probability ρ . This method of sampling reflects the fact that in empirical data, only a subset of the infected individuals are sampled and included into the epidemiological study [141]. It is on this reconstructed phylogeny that analyses were conducted. All edges without sampled descendants were suppressed from the tree.

While Stadler’s BDM implementation in R (package TreeSim) enables the inference of epidemiological dynamics under constant or episodic processes [143], its utility for such inference is limited to relatively small phylogenetic trees. A more recently

developed BDM R-package called TESS [64] provides the same features offered in TreeSim, with a more flexible framework for specifying diversification models. As such, TESS was used here as a means of simulating phylogenetic trees.

Next, to infer the epidemiological parameters imprinted on each phylogenetic tree, we employed an ABC-SMC method [150]. In ABC-SMC, a series of parameter values (often termed particles) are first sampled from the prior distribution. These particles are then propagated through a sequence of intermediate distributions defined by a tolerance schedule, where the threshold value (ϵ) for acceptance is sequentially lowered. This threshold value is based on a summary statistic that captures the information represented by the data. As the acceptance threshold approaches zero for each sequential population of particles, the intermediate distributions gradually morph towards the target posterior distribution. This population approach alleviates the major issue of getting stuck in areas of low probability, usually faced with the more standard ABC-MCMC approach [150].

Although the performance of ABC-SMC samplers has historically been superior to that of the basic rejection sampler and ABC-MCMC [88, 150], the efficiency of the ABC-SMC algorithm hinges on the choice of an appropriate summary statistic for the data. Given that the data observed here are in the form of phylogenetic trees, we hypothesized that the normalized LTT (nLTT) curves of each tree would serve as appropriate summary statistics. Using these curves, two phylogenies can effectively be summarized and compared against one another. We therefore implemented this summary statistic into our ABC-SMC algorithm as a means of comparing our simulated and observed phylogenies. We calculated the nLTT statistic using the function nLTTstat from the R-package nLTT in R 3.2.1 [70].

3.4.2 Simulation study

We simulated birth-death trees under several epidemiological scenarios in order to assess the ability of our method to infer both removal (μ) and transmission (λ) rates accurately. The parameters used in this model were λ , μ , ρ (sampling probability), T (total number of extant tips), and epidemic age. For λ and μ , the prior distribution was a uniform prior on $[0,5]$. Such a prior reflects the most realistic values for these parameters [70] and ensures neither parameter becomes too large. Possible empirical values for λ and μ were based on a geometric sequence from 0 to 5 with common ratio 1.5, yielding values $\{0.5, 0.75, 1.125, 1.6875, 2.53125, 3.796875\}$ for λ and $\{0.15, 0.225, 0.3375, 0.50625, 0.759375, 1.1390625\}$ for μ . The parameter μ was restricted as to be smaller than λ in order to avoid simulating uninformative phylogenies. Trees were simulated with sampling probabilities (ρ) of 1, 0.5, and 0.25. The number of taxa (T) was set to 500, 1000, 10,000, and 20,000. Age was also varied, with values set to 10, 20, and 30. These chosen values for number of taxa and epidemic age represent a wide range of potential real-world epidemic scenarios. The number of particles per iteration of the ABC-SMC algorithm was set to 1000. The ϵ schedule was set following the procedure described in Janzen *et al.* [30, 70]. The schedule was decreased based on an exponential progression, such that the ϵ value was $\epsilon_t = \epsilon_0 \exp(-0.5t)$, where t represents the current step in the sequence. For each parameter setting of λ , μ , ρ , T , and age, we simulated 25 trees. The accuracy of our algorithm was measured using the mean squared error (MSE). The adapted and modified script from [70] has been deposited at github.com/sarisbro.

3.4.3 Inferring parameters from Swiss HIV-1 sequence data

We used HIV-1 polymerase sequence data from the Swiss HIV Cohort Study [145]. The available Swiss HIV pol sequences (570) [79] were aligned using MUSCLE version 3.8.31 [40] and manually inspected using Jalview [157]. The sequences used

were divided based on the presence of a sequencing artifact (indel) in one group of sequences. This pattern reflects the phylogeographical signal resulting from two different sequencing methods in two regions. Therefore, we had two groups of sequences representing the two culturally unique regions within Switzerland that were sampled. Since the identity of each group could not be inferred from GenBank annotations, we labeled them as Region 1 (252 sequences) and Region 2 (318 sequences). In order to determine the age of each tree (length of the epidemic), the alignments were analyzed in BEAST v1.8.0 [36] using a GTR+ Γ_4 model of nucleotide substitution [81, 147, 163], assuming an uncorrelated lognormal relaxed clock and a constant-population coalescent prior. We ran two separate Markov chain Monte Carlo samplers for each alignment (10^{10} states) to check for convergence, visually assessed these with Tracer (tree.bio.ed.ac.uk/software/tracer), and removed 10% of the 10^{10} states as burn-in.

Using the two trees obtained through the BEAST analysis, we applied our algorithm to the Swiss HIV data. The empirical estimates for parameters μ and λ were obtained for Regions 1 and 2.

3.5 Results and Discussion

In order to circumvent the computational burden incurred by rigorous modeling and Markov chain Monte Carlo sampling, we implemented an epidemiological model employing an ABC-SMC sampler. By inferring the transmission and removal rates of an epidemic, critical parameters such as the basic reproductive number ($R_0 = \lambda/\mu$) and the growth rate ($r = \lambda - \mu$) were determined. Thus, our approach offers a potentially powerful tool for inferring key epidemiological parameters from sequence data alone. To validate our approach, we undertook an extensive simulation study.

3.5.1 Simulated data

We simulated 25 replicate trees for each parameter setting, and we assessed the ability of our method to recover the values for each combination of λ and μ . Simulation results show that the SMC algorithm converges upon the true posterior within 11 populations (Figure 3.1), using solely the nLTT as a summary statistic.

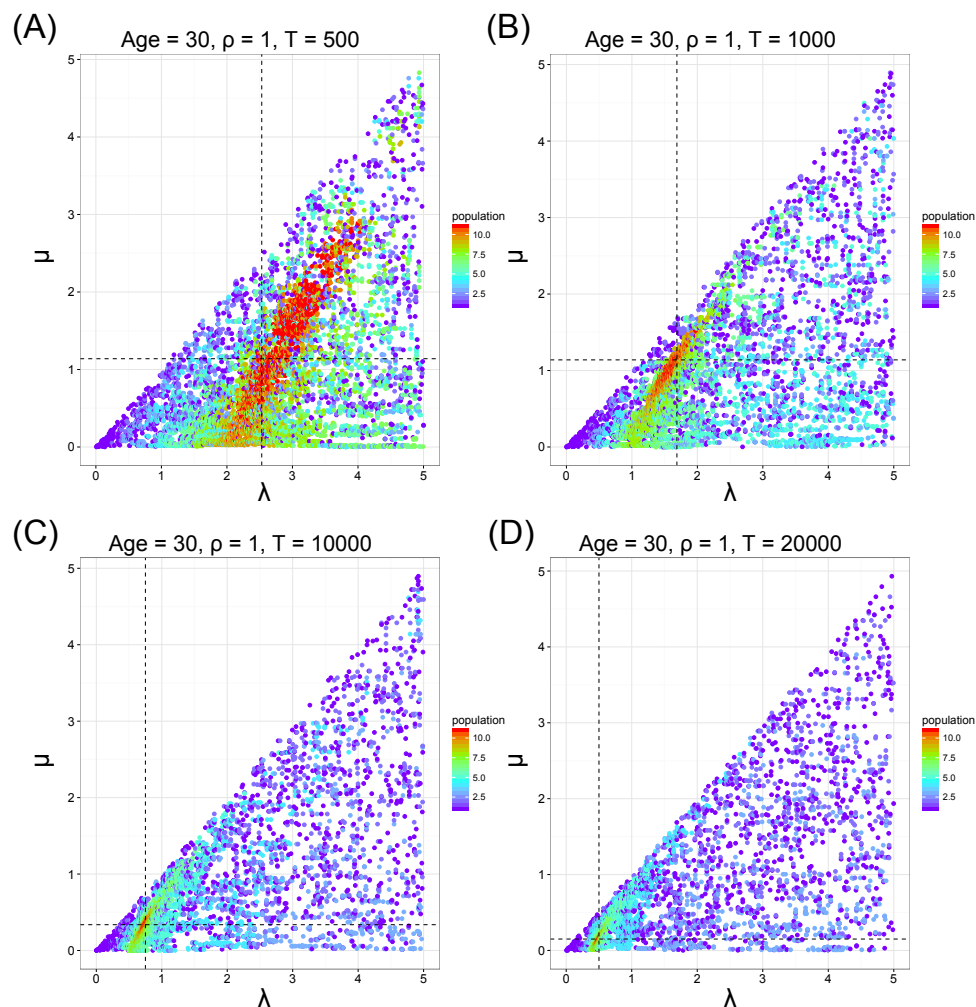


Figure 3.1: ABC-SMC estimates of the posterior distribution for four taxa settings. Number of populations required for convergence (max = 11) is indicated using color spectrum. Estimated parameter values for each setting are indicated with dashed lines. (A): ($\lambda=2.531$, $\mu=1.139$). (B): ($\lambda=1.125$, $\mu=1.139$). (C): ($\lambda=0.750$, $\mu=0.338$). (D): ($\lambda=0.500$, $\mu=0.150$). Particles per iteration = 1000. $N = 25$ replicate trees.

Such quick convergence supports our hypothesis that the nLTT is an appropriate and informative summary statistic, given the nature of our data. While previous authors [70] compared the performance of the nLTT to other phylogenetic summary statistics within an ABC-SMC framework, the parameter space they explored was limited. In particular, birth-death trees were only simulated under age 10 and no more than 178 tips (extant taxa). Here, we cover a broader parameter space, presenting a greater number of epidemiological scenarios. Still, the quick convergence onto the true posterior distribution further demonstrates that the nLTT is a suitable summary statistic for parameter inference [70].

The marginal MSEs for λ (Figure 3.2) and μ (Figure 3.3) illustrate that our SMC approach was successful in recovering the real values of these parameters using only the nLTT of each tree as a summary statistic. Marginal MSEs for λ were surprisingly low, with an upper bound of 0.2 for all combinations of age, number of taxa, and sampling intensity. While the marginal MSEs obtained for μ were also reasonably low, our method was not as accurate in recovering these estimates as it was for λ . This is evident when comparing the upper bound of marginal MSEs for μ (0.5) versus λ (0.2).

Simulations therefore show that the transmission rate is more accurately estimated than the removal rate. Inferring the removal rate seems to be complicated by the non-independence between λ and μ (Figure 3.1). In our approach, we limited μ to values less than λ to avoid the high computational cost of trees simulated with a low birth rate and high death rate. This constraint is potentially limiting our ability to accurately infer μ , as λ might leave a more detectable imprint than μ on each reconstructed tree, thereby increasing the estimated growth rate (r). Consequently, certain regions of the parameter space correspond to very high estimation accuracy (low MSEs), while other regions give rise to low accuracy (high MSEs) (Figure 3.4), as previously reported [20]. This issue is further compounded by the inevitable epidemiological correlation between the birth rate and death rate. Previous work [20] has

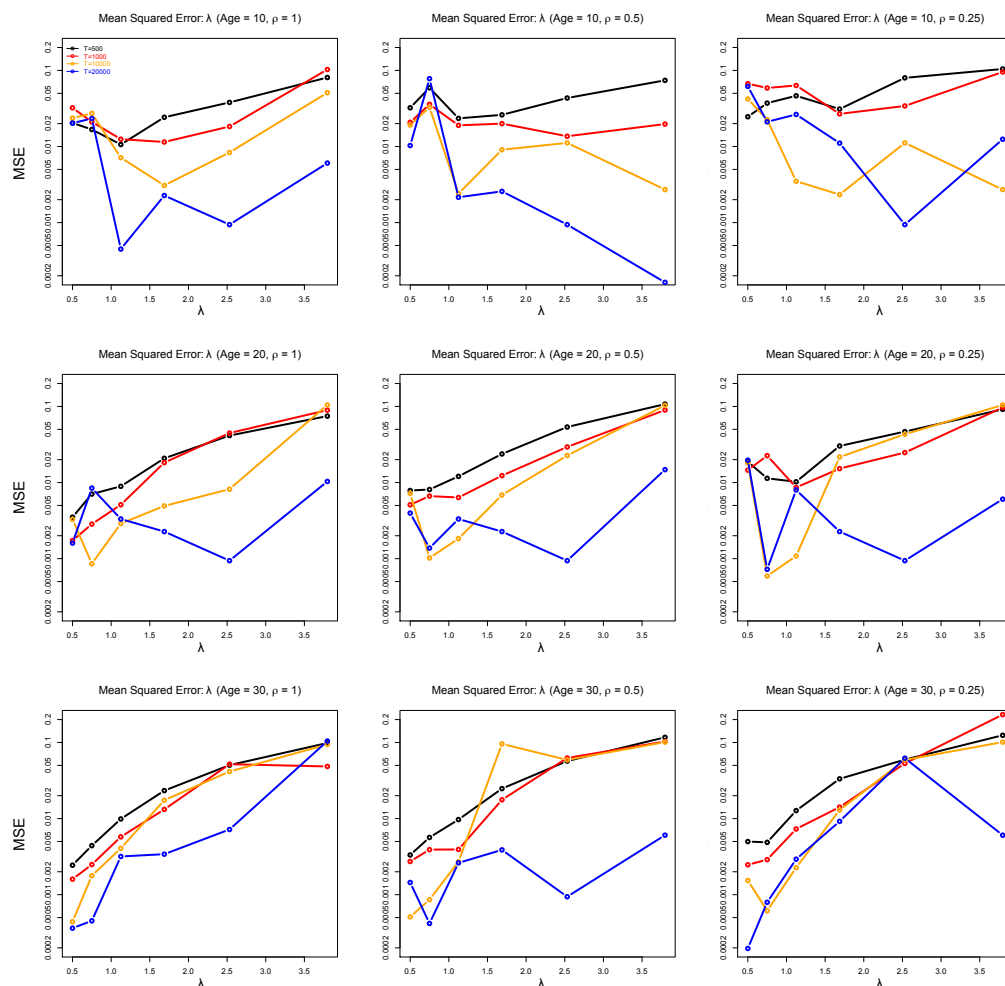


Figure 3.2: Marginal MSEs for λ . Mean squared errors for all epidemiological settings are shown. Y-axis is log-transformed (base 10). Each point represents an MSE value over 25 replicate trees.

shown that the transmission and removal rates in an epidemiological model are highly correlated, with the magnitude of the correlation efficient being strongly dependent on the value of R_0 . As R_0 approaches 1, the correlation coefficient peaks, and standard errors for point estimates increase significantly. Such a pattern can be observed in Figure 3.4. This correlation thus contributes to the identifiability constraints we observed here.

Although the correlation between these two parameters has been examined in

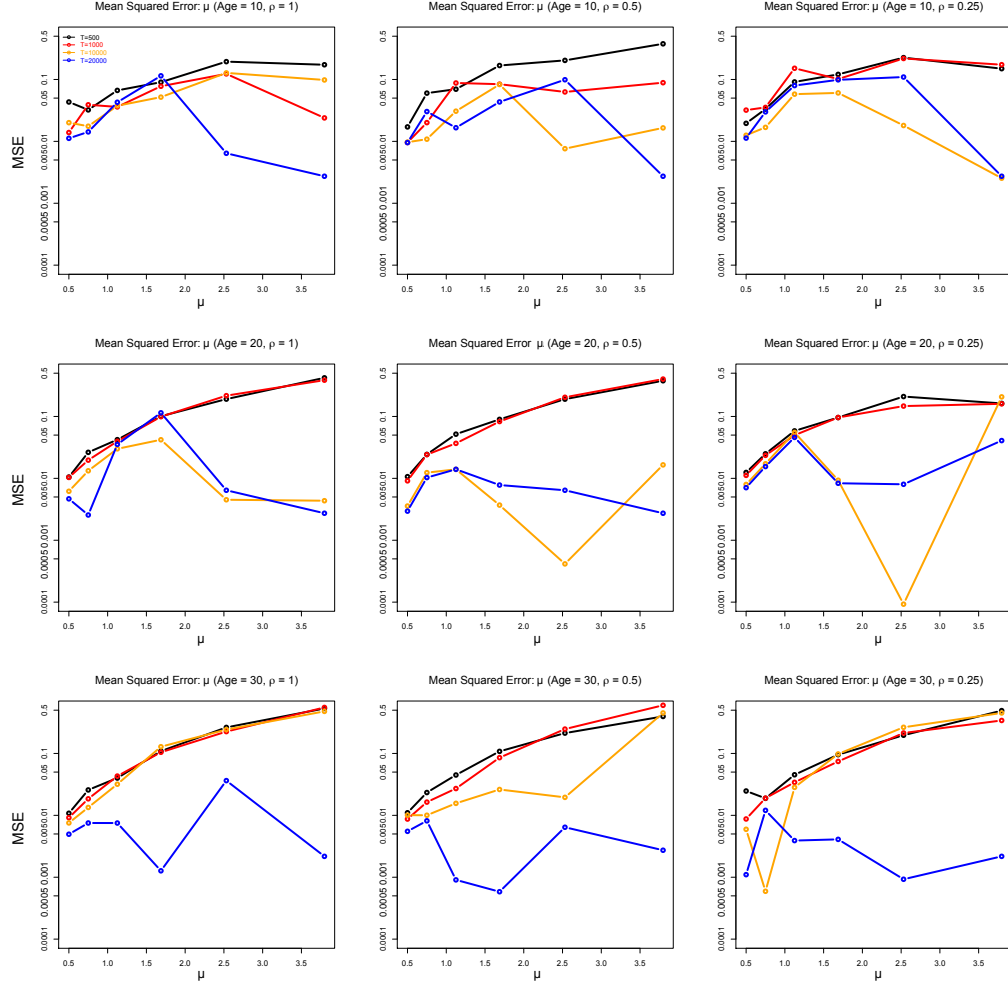


Figure 3.3: Marginal MSEs for μ . Mean squared errors for all epidemiological settings are shown. Y-axis is log-transformed (base 10). Each point represents an MSE value over 25 replicate trees.

previous studies [20, 91, 162], there is still no consensus on how identifiability constraints can be overcome. It has been noted that the time at which observations are made during an epidemic has a significant bearing on the precision of the estimates [20]. Sensitivity functions have been proposed as a means of determining points in time when the epidemiological system is most sensitive to parameter fluctuations [10, 11, 20]. Analysis based on such functions illustrates that the correlation between estimates of λ and μ is highest during the exponential growth phase, early in the

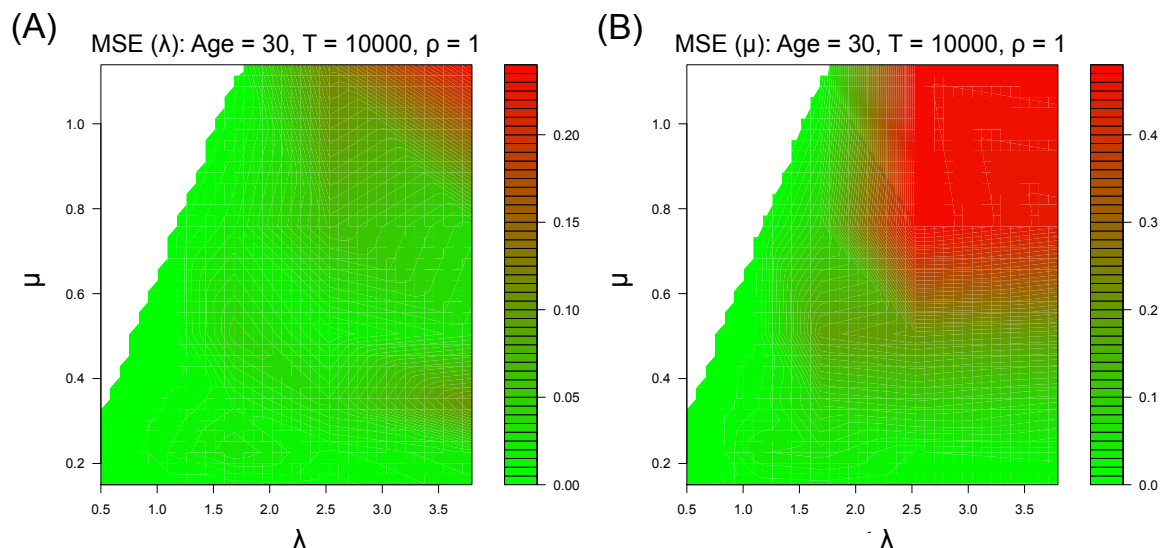


Figure 3.4: Joint MSE heatmaps. Mean squared error for a set age, taxa size, and sampling probability for λ (A) and μ (B). Axes show range of simulated values for all trees ($\lambda \leq \mu$).

epidemic [20]. As the constant-rate BDM employed here is limited to the exponential phase of an epidemic, such a strong correlation between λ and μ is expected. Therefore, this BDM technique could be further developed in future studies so that it also encompasses the saturation phase of an epidemic.

Our results also suggest that accuracy is solely affected by the number of taxa on the simulated tree. Essentially, as the number of taxa increases, our ability to accurately infer these epidemiological parameters increases as well (Figure 3.2 and Figure 3.3). Accuracy is particularly high for trees simulated with 1000 or more sequences (Figure 3.1). While our simulation method does not take into account tree length, it is evident that it is in fact tree depth (more than the number of taxa) that dictates the epidemiological signal we are trying to detect. Nonetheless, due to the model parameterization in TESS, we use the number of taxa as a proxy for tree depth.

Recent studies have also noted that parameter inference using the BDM is biased towards larger population sizes [18, 144]. This suggests that larger phylogenies provide

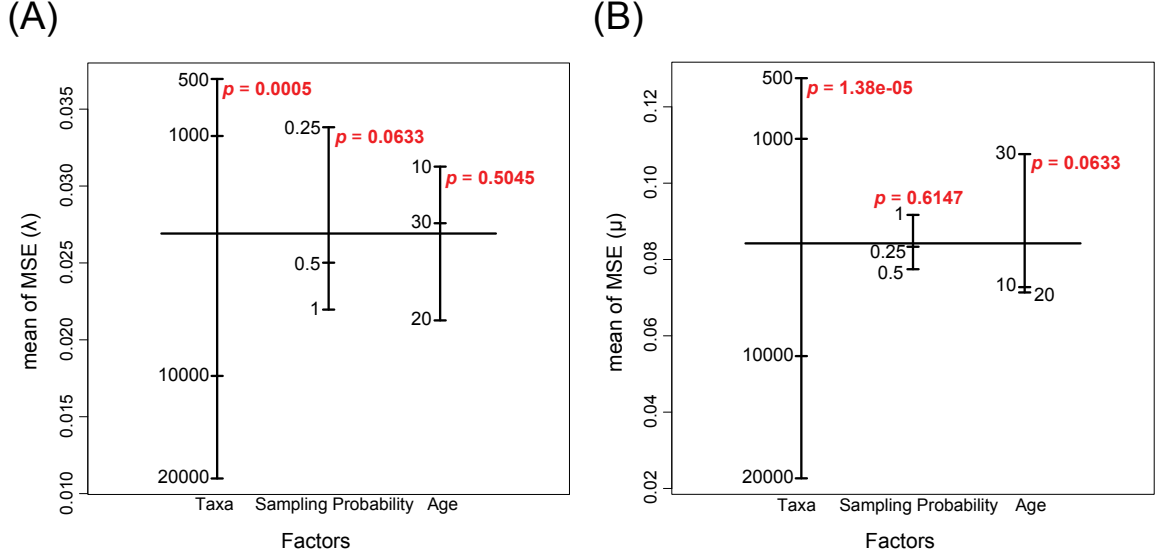


Figure 3.5: Three-factorial ANOVA for marginal MSEs. ANOVA results for λ (A) and μ (B). Independent factors listed are number of taxa, sampling probability, and age. P-value for each factor is denoted in red.

much more information about the dynamics of an epidemic, thereby facilitating the process of parameter inference. Uniquely here, we also find that sampling intensity (ρ) and age bear no statistically significant impact on accuracy (Figure 3.5). These results suggest that our approach is robust to most sampling designs (sampling effort and/or genetic diversity) and epidemic ages.

3.5.2 Swiss HIV-1 parameter estimates

We applied our method to HIV-1 sequence data from the Swiss HIV Cohort Study [145]. Using the two dated phylogenetic trees obtained from the BEAST analysis, we inferred the transmission and removal rates for Regions 1 and 2. We tested whether our method was capable of distinguishing between the transmission dynamics of these two geographically similar but culturally distinct samples.

First, the topological symmetry of both trees was compared using the normalized Sackin index [133], and we found that Region 1 and Region 2 had Sackin indices of

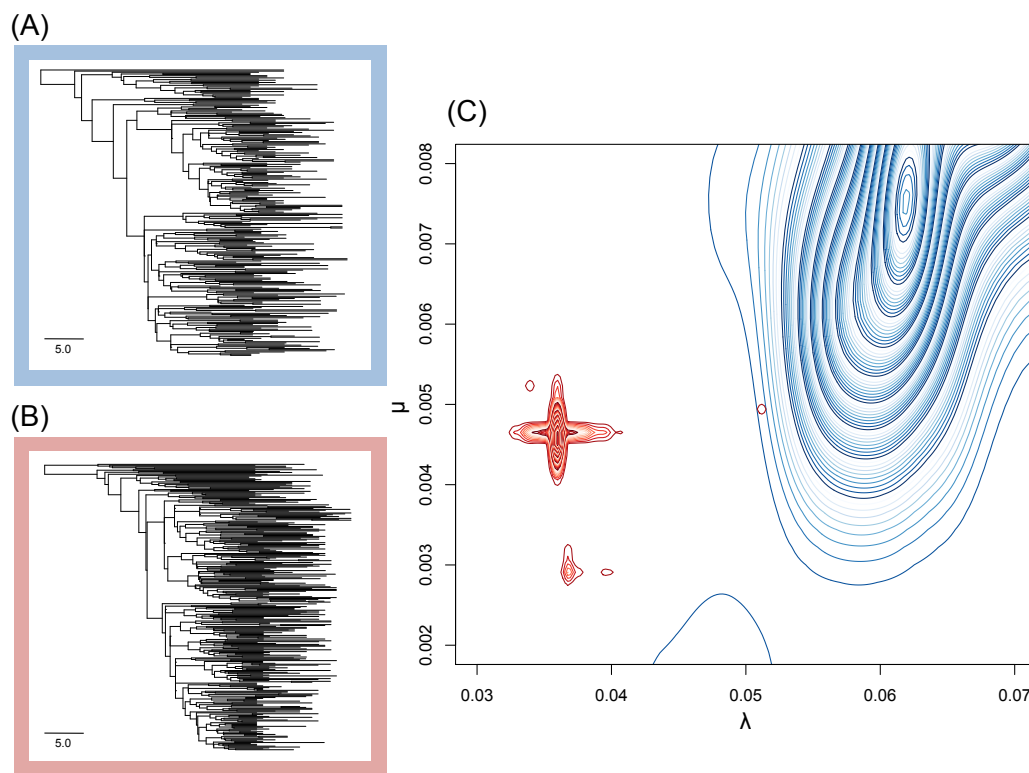


Figure 3.6: Swiss HIV-1 regional trees and parameter estimates. (A-B): Dated phylogenetic BEAST trees for Swiss HIV sub-clusters under a relaxed molecular clock. Scale is in years. (C): Joint posterior distribution of λ and μ approximated with ABC-SMC algorithm. Distribution for Region 1 is shown in blue, and Region 2 is in red.

0.918 and 0.939, respectively. While the trees corresponding to each region (Figure 3.6: A-B) suggest that the two phylogenies were very similar with respect to topology, our algorithm recovered two distinct epidemiological patterns (Figure 3.6: C). The contour plots represent the joint posterior distribution of λ and μ approximated by the ABC-SMC algorithm for the two regions. The parameter estimates and 95% HPD intervals for each region are listed in Table 3.1. Estimates using the upper and lower bounds of the 95% HPD intervals for age are listed in Table 3.2 for reference.

Table 3.1: Inferred λ and μ values with 95% HPD intervals for the two Swiss HIV clusters. R_0 and r values were obtained using λ and μ estimates.

	Tree Age ^a [95% HPD] ^b	λ [95% HPD]	μ [95% HPD]	R_0	r
Region 1	35 [23.7193, 51.3431]	0.0694 [0.0536, 0.0790]	0.0077 [0.0049, 0.0212]	9.0130	0.0617
Region 2	45 [32.9688, 60.1382]	0.0359 [0.0325, 0.0524]	0.0046 [0.0029, 0.0132]	7.8043	0.0313

^aAges (in years) were obtained from the BEAST analysis and were used to condition the ABC-SMC sampler.

^bThe 95% highest probability density (HPD) interval is defined as the shortest interval that contains 95% of the posterior probability.

Our results indicate that during the Swiss HIV epidemic, Region 1 had a higher growth rate than Region 2, suggesting that the pathogen was spreading at a faster pace throughout the population. Moreover, Region 1 had a higher basic reproductive ratio than Region 2 (Table 3.1). The R_0 estimates for Regions 1 and 2 are significantly greater than 1, indicating that the probability of infection and the extent of spread of the disease within these regions in Switzerland were high. While this is in line with previous predictions based on samples from the Swiss HIV Cohort Study [144], no study to date has conducted any analysis on the specific sequences and epidemic time-frame used here, so comparisons cannot be made. However, the R_0 estimates for other HIV outbreaks in Western Europe ($R_0 = 3.5-4.1$) [105], as well as well-known diseases such as Measles ($R_0 = 12-18$) [3], Influenza ($R_0 = 2-3$) [99], and SARS ($R_0 = 2-5$) [155] provide a point of reference. While disease dynamics cannot be compared based on R_0 values alone, our estimates seem to suggest that this outbreak was growing at a faster pace than previous Western European HIV outbreaks.

With no available data connecting the sequences to their exact origins, no conclusions can be made about the cultural effects of each region on the above estimates. Nonetheless, one factor that may have influenced the different epidemiological patterns in each region is their respective composition according to transmission group. The subsamples studied here may be dominated by different risk groups (i.e. men having sex with men, heterosexuals, and intravenous drug users) [79, 144], which may have an impact on the underlying dynamics of the subepidemics.

Table 3.2: Inferred λ and μ values using the upper and lower 95% HPD bounds for age. R_0 and r values were obtained using λ and μ estimates.

	Tree Age	λ [95% HPD]	μ [95% HPD]	R_0	r
Region 1	23	0.0322 [0.0266, 0.0570]	0.0157 [0.0091, 0.0316]	2.0510	0.0165
	51	0.0046 [0.0027, 0.0075]	0.0011 [0.0002, 0.0029]	4.1818	0.0035
Region 2	32	0.0058 [0.0031, 0.0069]	0.0007 [0.0006, 0.0009]	8.2857	0.0051
	60	0.0754 [0.0636, 0.0855]	0.0099 [0.0081, 0.0296]	7.6162	0.0655

3.5.3 Conclusions

Various studies have attempted to estimate epidemiological parameters directly from viral sequence data. Past approaches based on the coalescent as a transmission model were unable to directly infer transmission and death rates, and they had to resort to a sophisticated independent estimate for the duration of infectiousness [50, 154]. More recently, Stadler *et al.* [144] applied a Bayesian BDM to the parameter estimation problem. While they accurately inferred R_0 values, their approach was incapable of accurately estimating the transmission and death rates separately. Here, we show that our method can accurately recover the transmission and removal rates of simulated empirical trees. Most importantly, our reanalysis of data from the Swiss HIV Cohort Study shows that we are able to distinguish transmission patterns in two culturally different parts of Switzerland.

Our ability to infer the epidemiological parameters of two subsamples with very similar phylogenetic structures, highlights the utility and power of our method. Our approach provides useful information about an outbreak even when epidemiological data is incomplete, and it is thus informative within the earliest points of an outbreak. In light of these results, we propose that our method will make it possible for future studies to evaluate the impact of new public health policies (e.g., implementation of a needle exchange program in the case of HIV) based on genetic data sampled before and after the implementation of a new policy.

Chapter 4

Conclusions and Future Directions

4.1 Summary of research objectives

The strong growth within the field of viral evolution over the last decade can be attributed to the rising availability of genomic sequence data. Viral sequence data can be highly informative about the evolutionary trajectory of a strain, and the potential selective pressures acting upon it. Most importantly, sequence data can provide critical insight into epidemiological patterns and parameters that cannot consistently be obtained from standard surveillance data. Sequence data, therefore, not only provide a glimpse into the past, but also enable the prediction of future dynamics. As the sequencing of viral genomes continues to grow at an exponential pace, novel and interdisciplinary approaches are required in order to address key questions in viral evolution and epidemiology. Thus, drawing on principles from the fields of phylodynamics and Bayesian statistics, this thesis sheds light on the evolutionary processes dictating the spread of the Ebola virus and HIV.

The main research objectives for this thesis were as follows:

- i) Given the presence of a series of amino acid substitutions unique to the 2014

EBOV variant, what are the implications of these changes on the viral biology of the Ebola virus? Specifically, is there evidence for diversifying selection in this genome, and if it exists, to what extent was epistasis required for the accumulation of these mutations? How does the structural evolution of the EBOV glycoprotein reflect the interplay between selection and epistasis?

ii) Can an ABC-SMC sampler coupled with the normalized lineages through time as a summary statistic effectively infer the transmission and removal rates of an epidemic? For which epidemiological scenarios/parameter combinations does the algorithm provide the most reliable estimates? Is the algorithm capable of distinguishing between the unique epidemiological dynamics of two phylogenetically similar datasets?

4.2 Main conclusions

4.2.1 Evolution of the Ebola virus

The first ever analysis of Ebola virus isolates, conducted by Gire and colleagues [52], represents a monumental body of work in the plight against Ebola. However, while providing an immense amount of sequence data to facilitate future viral analysis, this previous study failed to present any evidence of adaptation within the EBOV genome. In Chapter 2, we reconstructed the history of this virus and proposed a mechanistic explanation for the structural evolution of the GP MLD through time.

First, using MEME as a means of detecting site specific episodic selection, we found tentative evidence for selection occurring in 3 out of the 7 EBOV genes. However, given that this evidence was not statistically significant, we attempted to assess the biological significance of our detected sites. A lineage-through-time analysis revealed that each EBOV outbreak was preceded by a marked burst of diversification, the most significant of which was right before the 2014 outbreak. This finding

strengthened the hypothesis that selection was in fact shaping the evolution of this viral strain. The sequential pattern of fixation of the adaptive substitutions led us to believe that epistasis may have played a role in the accumulation of these changes. Evidence for epistatic interactions was not only statistically significant, but also, a handful of genes displayed a pattern of epistatic networks. Moreover, the overlap between the detected adaptive and epistatic sites for these genes suggested a potential link between selection and epistasis.

The EBOV glycoprotein proved to be the ideal protein for exploring this potential link, since it is the sole EBOV surface protein and is significantly more susceptible to immune selective pressure than the other EBOV proteins [26]. Our results clearly corroborated this observation, as all of the detected interacting sites were found within a region (the MLD) that is critical for immune evasion. The ancestral reconstructions of the MLD revealed the sequence of conformational changes that occurred throughout the history of this virus, highlighting the effects that adaptive and epistatic substitutions 428 and 308 had on the configuration of this region. In particular, the first substitution (P428L) represented a very drastic change in terms of the physico-chemical state of the protein, given the exceptionally high conformational rigidity of proline (P) [14]. Such a drastic change might have been maintained in the population by genetic drift, as the effective population size in the years preceding the 2014 outbreak was quite small (Figure 2.1). These six structural models illustrated the spatial configuration of the GP epitopes and glycosylation sites, a perspective that has been unavailable in the literature.

Notwithstanding, there exists one particular area of this study that warrants closer inspection. One of the most pressing questions concerning the 2014 EBOV outbreak was whether selective pressures may have contributed to the emergence of this new strain; however, we were unable to detect selection in 4 out of the 7 genes. Our inability to find evidence for selection within these genes may be due to a few key reasons. Firstly, selective pressures targeting these genes may not be

strong enough to enhance the rate of non-synonymous substitutions to a detectable level. This may be due to the fact that the proteins encoded by these genes may have limited interaction with the host immune system. Two of the proteins in question, VP35 and VP30, are minor role players in the ribonucleocapsid complex [72]. VP24 is a very small matrix protein that plays a secondary role behind the VP40 matrix protein. Surprisingly, VP40, the most abundant EBOV protein, showed no evidence of evolving under positive selection, even though it plays a critical role in viral assembly and budding. Therefore, a second potential reason for our inability to detect selection may be the presence of strong functional constraints. In the case of VP40, if there exists an intolerance for mutations due to such constraints, the accumulation of substitutions (escape mutations) will be significantly reduced, irrespective of the strength of selection. Finally, our failure to detect selection in the aforementioned genes may simply be due to the detection method itself. While we used the most reliable and robust method currently available [102], the limited number of sequences permitted for the analysis (due to a sequence similarity threshold) may have reduced the overall strength of the dN/dS signal for each gene. It remains unclear which factor contributed to the weak selection signal, but our study provides a good starting point for functional analysis.

4.2.2 HIV parameter inference using ABC-SMC

Retracing the host contact patterns of a viral outbreak is arduous, especially when surveillance data is incomplete. The application of ABC approaches to epidemiological inference has greatly alleviated the burden of incomplete surveillance data in real-world epidemics [141]. Reconstructing transmission trees in an ABC framework provides a more reliable route for gaining insight into the rate of epidemic growth. Two key constituents of epidemic growth are the transmission rate and the removal (death) rate. The work presented in Chapter 3 focused on the application of the

ABC-SMC approach to the inference of these two quantities.

In this study, we assessed the performance of an ABC-SMC sampler using the normalized lineages-through-time (nLTT) as a summary statistic. While a previous simulation study showed that the nLTT outperformed three other commonly used tree distance measures [70], that study was highly limited in its epidemiological breadth. Moreover, the ABC-SMC method was not applied to real epidemic data. Here, not only was the algorithm expanded to account for a wider range of epidemiological scenarios, but its performance was further tested using HIV-1 epidemic data. Our results indicated the specific regions in parameter space for which the algorithm was most accurate. This information has crucial implications for the characterization of viral epidemics. Knowing which epidemiological scenarios the algorithm is most robust to will allow the user to determine the reliability of the estimates obtained. The simulation results also suggested that incomplete sampling does not have a significant bearing on the accuracy of our estimates. Incomplete sampling of cases is inevitable in real-world epidemics [141]. As such, a favorable feature of any epidemiological inference algorithm would be that the fraction of unreported or unsampled cases does not directly affect the model.

Although the simulations highlighted the wide range of applicability of this approach, computational time was a limiting factor. The increments chosen for the simulated parameter values of μ and λ were relatively large, due to the fact that much smaller values (closer to 0) yielded computationally expensive trees with limited information to infer. Moreover, the parameter values for μ were constrained to avoid creating computationally expensive trees. As a consequence, our analysis might have been biased towards trees with a larger growth rate (r). Compounded by the epidemiological correlation between λ and μ [20], this constraint had a definite impact on the accuracy of our estimates for μ (Figure 3.4). While all of these adjustments were necessary in order to make the simulation study feasible, they highlight potential areas of improvement to be explored in extended studies.

The accuracy of our μ estimates might have also been influenced by our simulation design. Simulated trees were conditioned on both the time of the process (age of the epidemic) and the number of sampled species (extant taxa). This simulation design was chosen in order to reflect a wider scope of real-world epidemiological scenarios than those presented in previous work [70]. However, this approach may have potentially limited our ability to accurately infer μ . Since each simulation run was terminated once the set number of extant taxa had been reached, it is possible that the total effect of μ might not have completely materialized during certain runs. To avoid this, conditioning solely on epidemic age (and potentially using the number of taxa as an additional summary statistic) might alleviate the identifiability issues faced with μ . It is clear, however, that an even wider range of epidemiological scenarios need to be explored in order to determine the generality of the claim that identifiability issues are greater with μ than with λ .

In spite of the identifiability issues faced in Chapter 3, we were able to infer the reproductive ratio (R_0) of the two studied regions. While this ratio represents a fundamental concept in the field of epidemiology, its application as a threshold parameter is questionable. The (R_0) value is commonly used by health officials to determine whether the epidemic is under control ($R_0 < 1$) or not ($R_0 > 1$). However, in many instances, diseases with $R_0 < 1$ can persist within a population, and diseases with $R_0 > 1$ can be eradicated [90]. This is due to the fact that the invasion and persistence dynamics of a disease are not solely dictated by the transmission and removal rates (λ and μ), but also by the temporal and spatial contact structure of the host population. Moreover, by definition, R_0 represents the average number of secondary infections in a susceptible population; however, this setting is not representative of all stages of an epidemic. As such, estimated R_0 values need to be accompanied by the underlying model assumptions, as well as more sensitive measures of transmissibility.

The HIV sequences that were available for this study were very limited, since a full publication of all sequences in the SHCS might endanger the privacy of the

patients. The sequences used here represent a random subset of 10% of all sequenced samples used in the original publication [79]. The full SHCS dataset is only made available to researchers who have had a project proposal approved by the SHCS scientific board. One such research group with full access to the data provided R_0 estimates [144], but comparing our R_0 values was difficult, as they used a different data set (consisting of many more sequences) and partitioned their data into 10 subepidemics. As more studies attempting to apply similar methods to the parameter inference problem become available, improving and evaluating these approaches will be simplified.

4.3 Future directions

4.3.1 Linking epistasis and selection: DNA versus RNA viruses

Without the availability of direct functional studies, some of the conclusions made in Chapter 2 are speculative, but they are not without purpose. The exploration of such hypotheses is crucial for the characterization of viruses that continue to plague global communities. Currently, it remains unclear to what extent the conclusions made in Chapter 2 can be extrapolated to other viruses. A member of our group (Louis Parent) has therefore undertaken the task of determining the generality of the link made here between epistasis and selection. Louis is contrasting two kinds of viruses: double stranded DNA (dsDNA) and single stranded RNA (ssRNA) viruses.

Double stranded DNA viruses have large genomes (can be as large as 305,000 bp) and low mutation rates, due to the proofreading capabilities of DNA polymerases [94]. Additionally, DNA viruses do not replicate as quickly as RNA viruses; therefore, they tend to have much smaller population sizes. On the other hand, ssRNA viruses (such as Ebola) have compact genomes with overlapping genes, high mutation rates, and large population sizes [94]. We hypothesize that selective forces acting on dsDNA

viruses will be much stronger than those acting on ssRNA viruses, due to the presence of genomic redundancy and small population sizes. However, epistatic interactions may be more prevalent in ssRNA viruses. In these small genomes, the amplified fitness effects of mutations can be alleviated by compensatory mutations. We therefore predict that functional epistasis reflects stronger evolutionary constraints in ssRNA viruses than in organisms with larger and more redundant genomes.

To test this, whole genome data of well-characterized human viruses was collected from GenBank. Co-evolution and selection analysis was conducted on both dsDNA antigen/non-antigen genes and ssRNA antigen/non-antigen genes. Preliminary results suggest that the overlap between co-evolving sites and those under selection is not statistically different amongst subgroups. However, the analysis of the co-evolutionary networks for each gene revealed evidence for unique topological characteristics for ssRNA viral genes versus dsDNA viral genes. Further analyses incorporating a larger subset of viral genes are currently ongoing.

4.3.2 Expanding the application of ABC-SMC for parameter inference

In Chapter 3, the ABC-SMC approach was applied to sequence data from the Swiss HIV epidemic. In order to better evaluate the performance of this approach, it would need to be applied to other datasets, particularly ones for which the epidemiological context and parameters are well characterized. Examples of such HIV studies originate from China [156], Spain [24], and Botswana [106], to name a few. The estimates obtained from field studies can then be used as a benchmark for ABC inference.

The ABC-SMC approach can also be used to assess the impact of host contact structure on the overall epidemiological dynamics of an outbreak. In Switzerland, the relatively high prevalence of HIV/AIDS has been attributed to the growth of certain population groups, such as men who have sex with men, intravenous drug

users, and migrants from Sub-Saharan Africa. Differences in contact structure within host populations help explain the high variation in HIV/AIDS prevalence amongst developed countries. In addition to HIV, this ABC-SMC approach can be extended to other RNA viruses, such as the Zika virus, in order to facilitate control efforts for ongoing outbreaks.

Quantifying epidemiological dynamics in the midst of an outbreak can be difficult, especially when the outbreak is growing at an exponential pace (as seen in Chapter 2). Data collection and sharing is often hampered by uncertainties over intellectual property rights, as well as inconsistent data release policies [39, 167]. Parameter inference via viral sequences offers a less strenuous avenue for outbreak surveillance, limiting the risk of bias and human error that is often present in incidence data [1]. Nonetheless, in order to improve the accuracy and precision of the computational approaches discussed here, classical epidemiological measurements must be considered, once they are made available. Combining these two techniques will inevitably yield valuable insight which can be used to tailor control efforts.

Bibliography

- [1] ALIZON, S., LION, S., MURALL, C. L., AND ABBATE, J. L. Quantifying the epidemic spread of Ebola virus (EBOV) in Sierra Leone using phylodynamics. *Virulence* 5, 8 (2014), 825–827.
- [2] ALONSO, R., CRAWFORD, A. J., AND BERMINGHAM, E. Molecular phylogeny of an endemic radiation of Cuban toads (Bufonidae: Peltophryne) based on mitochondrial and nuclear genes. *Journal of Biogeography* 39, 3 (2012), 434–451.
- [3] ANDERSON, R. M., MAY, R. M., AND ANDERSON, B. *Infectious diseases of humans: dynamics and control*, vol. 28. Wiley Online Library, 1992.
- [4] ANTONELLI, A., AND SANMARTIN, I. Mass extinction, gradual cooling, or rapid radiation? Reconstructing the spatiotemporal evolution of the ancient angiosperm genus *Hedyosmum* (Chloranthaceae) using empirical and simulated approaches. *Systematic Biology* 60, 5 (2011), 596–615.
- [5] ARUNACHALAM, R. Detection of site-specific positive Darwinian selection on pandemic influenza A/H1N1 virus genome: integrative approaches. *Genetica* 141, 4-6 (Jun 2013), 143–55.
- [6] ASHKENAZY, H., EREZ, E., MARTZ, E., PUPKO, T., AND BEN-TAL, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure

- of proteins and nucleic acids. *Nucleic Acids Research* 38, Web Server issue (Jul 2010), W529–33.
- [7] ASHKENAZY, H., PENN, O., DORON-FAIGENBOIM, A., COHEN, O., CANNAROZZI, G., ZOMER, O., AND PUPKO, T. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research* 40, W1 (2012), W580–W584.
- [8] AZZARITO, V., LONG, K., MURPHY, N. S., AND WILSON, A. J. Inhibition of [alpha]-helix-mediated protein-protein interactions using designed molecules. *Nature Chemistry* 5, 3 (2013), 161–173.
- [9] BAIZE, S., PANNETIER, D., OESTEREICH, L., RIEGER, T., KOIVOGUI, L., MAGASSOUBA, N., SOROPOGUI, B., SOW, M. S., KEÏTA, S., DE CLERCK, H., TIFFANY, A., DOMINGUEZ, G., LOUA, M., TRAORÉ, A., KOLIÉ, M., MALANO, E. R., HELEZE, E., BOCQUIN, A., MÉLY, S., RAOUL, H., CARO, V., CADAR, D., GABRIEL, M., PAHLMANN, M., TAPPE, D., SCHMIDT-CHANASIT, J., IMPOUMA, B., DIALLO, A. K., FORMENTY, P., VAN HERP, M., AND GÜNTHER, S. Emergence of Zaire Ebola virus disease in Guinea. *The New England Journal of Medicine* 371, 15 (Oct 2014), 1418–25.
- [10] BANKS, H., DEDIU, S., AND ERNSTBERGER, S. L. Sensitivity functions and their uses in inverse problems. *Journal of Inverse and Ill-posed Problems JIIP* 15, 7 (2007), 683–708.
- [11] BANKS, H., ERNSTBERGER, S. L., AND GROVE, S. L. Standard errors and confidence intervals in inverse problems: sensitivity and associated pitfalls. *Journal of Inverse and Ill-posed Problems JIIP* 15, 1 (2007), 1–18.

- [12] BEAUMONT, M. A., ZHANG, W., AND BALDING, D. J. Approximate Bayesian computation in population genetics. *Genetics* 162, 4 (2002), 2025–2035.
- [13] BERG, P. The dual-use conundrum. *Science* 337, 6100 (Sep 2012), 1273.
- [14] BETTS, M. J., AND RUSSELL, R. B. Amino acid properties and consequences of substitutions. *Bioinformatics for Geneticists* 317 (2003), 289.
- [15] BIASINI, M., BIENERT, S., WATERHOUSE, A., ARNOLD, K., STUDER, G., SCHMIDT, T., KIEFER, F., CASSARINO, T. G., BERTONI, M., BORDOLI, L., AND SCHWEDE, T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research* 42, Web Server issue (Jul 2014), W252–8.
- [16] BLOOM, J. D., GONG, L. I., AND BALTIMORE, D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* 328, 5983 (2010), 1272–1275.
- [17] BLUM, M. G., AND FRANÇOIS, O. On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited. *Mathematical Biosciences* 195, 2 (2005), 141–153.
- [18] BOSKOVA, V., BONHOEFFER, S., AND STADLER, T. Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. *PLOS Computational Biology* 10, 11 (2014), e1003913.
- [19] BROWN, T. A. *Genomes*. Garland Science, 2006.
- [20] CAPALDI, A., BEHREND, S., BERMAN, B., SMITH, J., WRIGHT, J., AND LLOYD, A. L. Parameter estimation and uncertainty quantification for an epidemic model. *Mathematical Biosciences and Engineering* (2012), 553.

- [21] CHAKRABORTY, S., RAO, B. J., ASGEIRSSON, B., AND DANDEKAR, A. Characterizing alpha helical properties of Ebola viral proteins as potential targets for inhibition of alpha-helix mediated protein-protein interactions. *Faculty of 1000 Research* 3 (2014).
- [22] CLARKE, K., AND WARWICK, R. A further biodiversity index applicable to species lists: variation in taxonomic distinctness. *Marine Ecology Progress Series* 216 (2001), 265–278.
- [23] COLIJN, C., AND GARDY, J. Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine, and Public Health* 2014, 1 (2014), 96–108.
- [24] CUEVAS, M. T., MUÑOZ-NIETO, M., THOMSON, M. M., DELGADO, E., IRIBARREN, J. A., CILLA, G., FERNÁNDEZ-GARCÍA, A., SANTAMARÍA, J. M., LEZAUN, M. J., JIMÉNEZ, L., ET AL. HIV-1 transmission cluster with T215D revertant mutation among newly diagnosed patients from the Basque Country, Spain. *Journal of Acquired Immune Deficiency Syndromes* 51, 1 (2009), 99–103.
- [25] DARRIBA, D., TABOADA, G. L., DOALLO, R., AND POSADA, D. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9, 8 (Aug 2012), 772.
- [26] DAVIDSON, E., BRYAN, C., FONG, R. H., BARNES, T., PFAFF, J. M., MABILA, M., RUCKER, J. B., AND DORANZ, B. J. Mechanism of binding to Ebola virus glycoprotein by the ZMapp, ZMAb, and MB-003 cocktail antibodies. *Journal of Virology* 89, 21 (2015), 10982–10992.
- [27] DE WIT, E., FELDMANN, H., AND MUNSTER, V. J. Tackling Ebola: new insights into prophylactic and therapeutic intervention strategies. *Genome Medicine* 3, 1 (2011), 5.

- [28] DEARLOVE, B., AND WILSON, D. J. Coalescent inference for infectious disease: meta-analysis of hepatitis C. *Philosophical Transactions of the Royal Society B* 368, 1614 (2013), 20120314.
- [29] DEL MORAL, P., DOUCET, A., AND JASRA, A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68, 3 (2006), 411–436.
- [30] DEL MORAL, P., DOUCET, A., AND JASRA, A. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing* 22, 5 (2012), 1009–1020.
- [31] DELPORT, W., POON, A. F. Y., FROST, S. D. W., AND KOSAKOVSKY POND, S. L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 26, 19 (Oct 2010), 2455–7.
- [32] DOMINGO, E., ESCARMIS, C., SEVILLA, N., MOYA, A., ELENA, S., QUER, J., NOVELLA, I., AND HOLLAND, J. Basic concepts in RNA virus evolution. *The Federation of American Societies for Experimental Biology Journal* 10, 8 (1996), 859–864.
- [33] DOMINGO, E., AND HOLLAND, J. RNA virus mutations and fitness for survival. *Annual Reviews in Microbiology* 51, 1 (1997), 151–178.
- [34] DRAKE, J. W., AND HOLLAND, J. J. Mutation rates among RNA viruses. *Proceedings of the National Academy of Sciences* 96, 24 (1999), 13910–13913.
- [35] DRUMMOND, A. J., HO, S. Y., PHILLIPS, M. J., AND RAMBAUT, A. Relaxed phylogenetics and dating with confidence. *PLoS Biology* 4, 5 (2006), e88.
- [36] DRUMMOND, A. J., AND RAMBAUT, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7, 1 (2007), 214.

- [37] DUBE, D., BRECHER, M. B., DELOS, S. E., ROSE, S. C., PARK, E. W., SCHORNBERG, K. L., KUHN, J. H., AND WHITE, J. M. The primed ebolavirus glycoprotein (19-kilodalton GP1, 2): sequence and residues critical for host cell binding. *Journal of Virology* 83, 7 (2009), 2883–2891.
- [38] DUFFY, S., SHACKELTON, L. A., AND HOLMES, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* 9, 4 (2008), 267–276.
- [39] DYE, C., BARTOLOMEOS, K., MOORTHY, V., AND KIENY, M. P. Data sharing in public health emergencies: a call to researchers. *Bull World Health Organ* 94, 3 (2016), 158.
- [40] EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32, 5 (2004), 1792–1797.
- [41] ELENA, S. F., SOLÉ, R. V., AND SARDANYÉS, J. Simple genomes, complex interactions: epistasis in RNA virus. *Chaos* 20, 2 (Jun 2010), 026106.
- [42] ELLIOTT, L., KILEY, M., AND MCCORMICK, J. Descriptive analysis of Ebola virus proteins. *Virology* 147, 1 (1985), 169–176.
- [43] FAN, Y., WU, R., CHEN, M.-H., KUO, L., AND LEWIS, P. O. Choosing among partition models in Bayesian phylogenetics. *Molecular Biology and Evolution* 28, 1 (2011), 523–532.
- [44] FAUCI, A. S. The human immunodeficiency virus: infectivity and mechanisms of pathogenesis. *Science* 239, 3840 (1988), 617–22.
- [45] FELSENSTEIN, J. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology* 22, 3 (1973), 240–249.

- [46] FELSENSTEIN, J., AND FELENSTEIN, J. *Inferring phylogenies*, vol. 2. Sinauer Associates Sunderland, 2004.
- [47] FITCH, W. M. On the problem of discovering the most parsimonious tree. *American Naturalist* (1977), 223–257.
- [48] FITCH, W. M., MARGOLIASH, E., ET AL. Construction of phylogenetic trees. *Science* 155, 3760 (1967), 279–284.
- [49] FROST, S. D., PYBUS, O. G., GOG, J. R., VIBOUD, C., BONHOEFFER, S., AND BEDFORD, T. Eight challenges in phylodynamic inference. *Epidemics* 10 (2015), 88–92.
- [50] FROST, S. D., AND VOLZ, E. M. Viral phylodynamics and the search for an ‘effective number of infections’. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365, 1548 (2010), 1879–1890.
- [51] GAVRYUSHKINA, A., WELCH, D., STADLER, T., AND DRUMMOND, A. J. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology* 10, 12 (2014), e1003919.
- [52] GIRE, S. K., GOBA, A., ANDERSEN, K. G., SEALFON, R. S. G., PARK, D. J., KANNEH, L., JALLOH, S., MOMOH, M., FULLAH, M., DUDAS, G., WOHL, S., MOSES, L. M., YOZWIAK, N. L., WINNICKI, S., MATRANGA, C. B., MALBOEUF, C. M., QU, J., GLADDEN, A. D., SCHAFFNER, S. F., YANG, X., JIANG, P.-P., NEKOU, M., COLUBRI, A., COOMBER, M. R., FONNIE, M., MOIGBOI, A., GBKIE, M., KAMARA, F. K., TUCKER, V., KONUWA, E., SAFFA, S., SELLU, J., JALLOH, A. A., KOVOMA, A., KONINGA, J., MUSTAPHA, I., KARGBO, K., FODAY, M., YILLAH, M., KANNEH, F., ROBERT, W., MASSALLY, J. L. B., CHAPMAN, S. B., BOCHICCHIO, J., MURPHY, C., NUSBAUM, C., YOUNG, S., BIRREN, B. W.,

- GRANT, D. S., SCHEIFFELIN, J. S., LANDER, E. S., HAPPI, C., GEVAO, S. M., GNIRKE, A., RAMBAUT, A., GARRY, R. F., KHAN, S. H., AND SABETI, P. C. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 345, 6202 (Sep 2014), 1369–72.
- [53] GODZIK, A. The structural alignment between two proteins: is there a unique answer? *Protein Science* 5, 7 (1996), 1325–1338.
- [54] GONG, L. I., SUCHARD, M. A., AND BLOOM, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* 2 (2013), e00631.
- [55] GRENFELL, B. T., PYBUS, O. G., GOG, J. R., WOOD, J. L., DALY, J. M., MUMFORD, J. A., AND HOLMES, E. C. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303, 5656 (2004), 327–332.
- [56] GRISWOLD, C. K. Epistasis can accelerate adaptive diversification in haploid asexual populations. *Proceedings of the Royal Society of London B: Biological Sciences* 282, 1802 (Mar 2015).
- [57] GRUEBER, C. E., WALLIS, G. P., AND JAMIESON, I. G. Episodic positive selection in the evolution of avian toll-like receptor innate immunity genes. *PLoS One* 9, 3 (2014), e89632.
- [58] GUEx, N., AND PEITSCH, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* 18, 15 (Dec 1997), 2714–23.
- [59] GUINDON, S., AND GASCUEL, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52, 5 (Oct 2003), 696–704.
- [60] HASTINGS, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 1 (1970), 97–109.

- [61] HENAO-RESTREPO, A. M., LONGINI, I. M., EGGER, M., DEAN, N. E., EDMUNDS, W. J., CAMACHO, A., CARROLL, M. W., DOUMBIA, M., DRAGUEZ, B., DURAFFOUR, S., ET AL. Efficacy and effectiveness of an rVSV-vectored vaccine expressing Ebola surface glycoprotein: interim results from the Guinea ring vaccination cluster-randomised trial. *The Lancet* 386, 9996 (2015), 857–866.
- [62] HENIKOFF, S., AND HENIKOFF, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* 89, 22 (1992), 10915–10919.
- [63] HOENEN, T., SAFRONETZ, D., GROSETH, A., WOLLENBERG, K. R., KOITA, O. A., DIARRA, B., FALL, I. S., HAIDARA, F. C., DIALLO, F., SANOGO, M., SARRO, Y. S., KONE, A., TOGO, A. C. G., TRAORE, A., KODIO, M., DOSSEH, A., ROSENKE, K., DE WIT, E., FELDMANN, F., EBIHARA, H., MUNSTER, V. J., ZOON, K. C., FELDMANN, H., AND SOW, S. Mutation rate and genotype variation of Ebola virus from Mali case sequences. *Science* 348, 6230 (Apr 2015), 117–9.
- [64] HÖHNA, S., MAY, M. R., AND MOORE, B. R. TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. *Bioinformatics* (2015), btv651.
- [65] HÖHNA, S., STADLER, T., RONQUIST, F., AND BRITTON, T. Inferring speciation and extinction rates under different sampling schemes. *Molecular Biology and Evolution* 28, 9 (2011), 2577–2589.
- [66] HOLLAND, J., SPINDLER, K., HORODYSKI, F., GRABAU, E., NICHOL, S., AND VANDEPOL, S. Rapid evolution of RNA genomes. *Science* 215, 4540 (1982), 1577–1585.

- [67] HOUSE, T. Epidemiological dynamics of ebola outbreaks. *Elife* 3 (2014), e03908.
- [68] IBEH, N., NSHOGOZABAHIZI, J. C., AND ARIS-BROSOU, S. Both epistasis and diversifying selection drive the structural evolution of the Ebola virus glycoprotein mucin-like domain. *Journal of Virology* 90, 11 (2016), 5475–5484.
- [69] ILLIFE, J. The African AIDS epidemic: a history. *Athens: Ohio University Press, Oxford: James Currey, Cape Town: Double Storey Publishers* (2006).
- [70] JANZEN, T., HÖHNA, S., AND ETIENNE, R. S. Approximate Bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nLTT. *Methods in Ecology and Evolution* 6, 5 (2015), 566–575.
- [71] JEFFREYS, H. *Scientific inference*. Cambridge University Press, 1973.
- [72] JOHNSON, R. F., BELL, P., AND HARTY, R. N. Effect of Ebola virus proteins GP, NP and VP35 on VP40 VLP morphology. *Virology Journal* 3, 1 (2006), 1.
- [73] JOINT UNITED NATIONS PROGRAMME ON HIV/AIDS. How AIDS changed everything. *Geneva, Switzerland: UNAIDS* (2015).
- [74] KATZE, M. G., KORTH, M. J., LAW, G. L., AND NATHANSON, N. *Viral pathogenesis: from basics to systems biology*. Academic Press, 2015.
- [75] KERMACK, W. O., AND MCKENDRICK, A. G. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences* (1927), vol. 115, The Royal Society, pp. 700–721.

- [76] KIM, D. E., CHIVIAN, D., AND BAKER, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research* 32, Web Server issue (Jul 2004), W526–31.
- [77] KIMURA, M., ET AL. Evolutionary rate at the molecular level. *Nature* 217, 5129 (1968), 624–626.
- [78] KINGMAN, J. F. C. The coalescent. *Stochastic Processes and their Applications* 13, 3 (1982), 235–248.
- [79] KOUYOS, R. D., VON WYL, V., YERLY, S., BÖNI, J., TAFFÉ, P., SHAH, C., BÖRGISSE, P., KLIMKAIT, T., WEBER, R., HIRSCH, B., ET AL. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *Journal of Infectious Diseases* 201, 10 (2010), 1488–1497.
- [80] KÜHNERT, D., STADLER, T., VAUGHAN, T. G., AND DRUMMOND, A. J. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. *Journal of the Royal Society Interface* 11, 94 (2014), 20131106.
- [81] LANAVE, C., PREPARATA, G., SACONE, C., AND SERIO, G. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution* 20, 1 (1984), 86–93.
- [82] LARKIN, M. A., BLACKSHIELDS, G., BROWN, N., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., ET AL. Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 21 (2007), 2947–2948.

- [83] LEE, J. E., FUSCO, M. L., HESSELL, A. J., OSWALD, W. B., BURTON, D. R., AND SAPHIRE, E. O. Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. *Nature* 454, 7201 (Jul 2008), 177–82.
- [84] LEE, J. E., KUEHNE, A., ABELSON, D. M., FUSCO, M. L., HART, M. K., AND SAPHIRE, E. O. Complex of a protective antibody with its Ebola virus GP peptide epitope: unusual features of a V λ x light chain. *Journal of Molecular Biology* 375, 1 (2008), 202–216.
- [85] LEE, J. E., AND SAPHIRE, E. O. Ebola virus glycoprotein structure and mechanism of entry. *Future Virology* 4, 6 (2009), 621–635.
- [86] LEE, J. E., AND SAPHIRE, E. O. Neutralizing ebolavirus: structural insights into the envelope glycoprotein and antibodies targeted against it. *Current Opinion in Structural Biology* 19, 4 (2009), 408–417.
- [87] LENNEMANN, N. J., RHEIN, B. A., NDUNGO, E., CHANDRAN, K., QIU, X., AND MAURY, W. Comprehensive functional analysis of N-linked glycans on Ebola virus GP1. *mBio* 5, 1 (2014), e00862–13.
- [88] LENORMAND, M., JABOT, F., AND DEFFUANT, G. Adaptive approximate Bayesian computation for complex models. *Computational Statistics* 28, 6 (2013), 2777–2796.
- [89] LEVENTHAL, G. E., KOUYOS, R., STADLER, T., VON WYL, V., YERLY, S., BÖNI, J., CELLERAI, C., KLIMKAIT, T., GÜNTARD, H. F., AND BONHOEFFER, S. Inferring epidemic contact structure from phylogenetic trees. *PLoS Computational Biology* 8, 3 (2012), e1002413.
- [90] LI, J., BLAKELEY, D., ET AL. The failure of 0. *Computational and mathematical methods in medicine 2011* (2011).

- [91] LI, P., AND VU, Q. D. Identification of parameter correlations for parameter estimation in dynamic biological models. *BMC Systems Biology* 7, 1 (2013), 91.
- [92] MAITI, R., VAN DOMSELAAR, G. H., ZHANG, H., AND WISHART, D. S. SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Research* 32, suppl 2 (2004), W590–W594.
- [93] MARJORAM, P., MOLITOR, J., PLAGNOL, V., AND TAVARÉ, S. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* 100, 26 (2003), 15324–15328.
- [94] MARTÍNEZ, M. A., MARTRUS, G., CAPEL, E., PARERA, M., FRANCO, S., AND NEVOT, M. Quasispecies dynamics of RNA viruses. In *Viruses: Essential Agents of Life*. Springer, 2012, pp. 21–42.
- [95] MARTINEZ, O., NDUNGO, E., TANTRAL, L., MILLER, E. H., LEUNG, L. W., CHANDRAN, K., AND BASLER, C. F. A mutation in the Ebola virus envelope glycoprotein restricts viral entry in a host species-and cell-type-specific manner. *Journal of Virology* 87, 6 (2013), 3324–3334.
- [96] MARTINEZ, O., TANTRAL, L., MULHERKAR, N., CHANDRAN, K., AND BASLER, C. F. Impact of Ebola mucin-like domain on antiglycoprotein antibody responses induced by Ebola virus-like particles. *Journal of Infectious Diseases* 204, suppl 3 (2011), S825–S832.
- [97] METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H., AND TELLER, E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 21, 6 (1953), 1087–1092.
- [98] METZNER, K. J., RAUCH, P., VON WYL, V., LEEMANN, C., GRUBE, C., KUSTER, H., BÖNI, J., WEBER, R., AND GÜNTARD, H. F. Efficient suppression of minority drug-resistant HIV type 1 (HIV-1) variants present

- at primary HIV-1 infection by ritonavir-boosted protease inhibitor-containing antiretroviral therapy. *Journal of Infectious Diseases* 201, 7 (2010), 1063–1071.
- [99] MILLS, C. E., ROBINS, J. M., AND LIPSITCH, M. Transmissibility of 1918 pandemic influenza. *Nature* 432, 7019 (2004), 904–906.
- [100] MIR, A., ROSSELLÓ, F., ET AL. A new balance index for phylogenetic trees. *Mathematical Biosciences* 241, 1 (2013), 125–136.
- [101] MURIN, C. D., FUSCO, M. L., BORNHOLDT, Z. A., QIU, X., OLINGER, G. G., ZEITLIN, L., KOBINGER, G. P., WARD, A. B., AND SAPHIRE, E. O. Structures of protective antibodies reveal sites of vulnerability on Ebola virus. *Proceedings of the National Academy of Sciences* 111, 48 (2014), 17182–17187.
- [102] MURRELL, B., WERTHEIM, J. O., MOOLA, S., WEIGHILL, T., SCHEFFLER, K., AND KOSAKOVSKY POND, S. L. Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics* 8, 7 (2012), e1002764.
- [103] MUSEY, L., HUGHES, J., SCHACKER, T., SHEA, T., COREY, L., AND MCEL-RATH, M. J. Cytotoxic-T-cell responses, viral load, and disease progression in early human immunodeficiency virus type 1 infection. *New England Journal of Medicine* 337, 18 (1997), 1267–1274.
- [104] NEE, S., MAY, R. M., AND HARVEY, P. H. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 344, 1309 (1994), 305–311.
- [105] NISHIURA, H. Correcting the actual reproduction number: a simple method to estimate r_0 from early epidemic growth data. *International journal of environmental research and public health* 7, 1 (2010), 291–302.
- [106] NOVITSKY, V., BUSSMANN, H., LOGAN, A., MOYO, S., VAN WIDENFELT, E., OKUI, L., MMALANE, M., BACA, J., BUCK, L., PHILLIPS, E., ET AL.

- Phylogenetic relatedness of circulating HIV-1C variants in Mochudi, Botswana. *PloS One* 8, 12 (2013), e80589.
- [107] NSHOGOZABAHIZI, J. C., DENCH, J., AND ARIS-BROSOU, S. Historical contingency in a fast-evolving organism: long-range chained epistasis is prevalent in influenza A viruses. Submitted, 2016.
- [108] NYLANDER, J. Catfasta2phym. Program distributed by the author. *Evolutionary Biology Center, Uppsala University* (2013).
- [109] ON HIV/AIDS (UNAIDS), J. U. N. P., ET AL. The gap report. *Geneva: UNAIDS* (2014).
- [110] ORGANIZATION, W. H., ET AL. The World Health report: 2004: Changing history. *WHO Press* (2004).
- [111] ORGANIZATION, W. H., ET AL. Priority interventions: HIV/AIDS prevention, treatment and care in the health sector. *WHO Press* (2010).
- [112] ORGANIZATION, W. H., ET AL. Hundreds of contacts identified and monitored in new Ebola flare-up in Guinea. Update from the field: 22 March 2016. Geneva, Switzerland: World Health Organization; 2016, 2016.
- [113] ORTLUND, E. A., BRIDGHAM, J. T., REDINBO, M. R., AND THORNTON, J. W. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* 317, 5844 (2007), 1544–1548.
- [114] PAGEL, M., AND MEADE, A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *American Naturalist* 167, 6 (Jun 2006), 808–25.
- [115] PARADIS, E. Time-dependent speciation and extinction from phylogenies: A least squares approach. *Evolution* 65, 3 (2011), 661–672.

- [116] PARADIS, E., CLAUDE, J., AND STRIMMER, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 2 (2004), 289–290.
- [117] PARK, D. J., DUDAS, G., WOHL, S., GOBA, A., WHITMER, S. L., ANDERSEN, K. G., SEALFON, R. S., LADNER, J. T., KUGELMAN, J. R., MATRANGA, C. B., ET AL. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* 161, 7 (2015), 1516–1526.
- [118] PENNING, P. S., HOLMES, S. P., AND SHAFER, R. W. HIV-1 transmission networks in a small world. *Journal of Infectious Diseases* 209, 2 (2014), 180–182.
- [119] PHILLIPS, A. N., GRABAR, S., TASSIE, J.-M., COSTAGLIOLA, D., LUNDGREN, J. D., EGGER, M., ET AL. Use of observational databases to evaluate the effectiveness of antiretroviral therapy for HIV infection: comparison of cohort studies with randomized trials. *Aids* 13, 15 (1999), 2075–2082.
- [120] POPINGA, A., VAUGHAN, T., STADLER, T., AND DRUMMOND, A. Bayesian coalescent epidemic inference: comparison of stochastic and deterministic SIR population dynamics. *arXiv preprint arXiv:1407.1792* (2014).
- [121] PRITCHARD, J. K., SEIELSTAD, M. T., PEREZ-LEZAUN, A., AND FELDMAN, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* 16, 12 (1999), 1791–1798.
- [122] PYBUS, O. G., AND HARVEY, P. H. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London B: Biological Sciences* 267, 1459 (2000), 2267–2272.
- [123] QUICK, J., LOMAN, N. J., DURAFFOUR, S., SIMPSON, J. T., SEVERI, E., COWLEY, L., BORE, J. A., KOUNDOUNO, R., DUDAS, G., MIKHAIL, A., ET AL. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 7589 (2016), 228–232.

- [124] RAMAN, S., VERNON, R., THOMPSON, J., TYKA, M., SADREYEV, R., PEI, J., KIM, D., KELLOGG, E., DIMAIO, F., LANGE, O., KINCH, L., SHEFFLER, W., KIM, B.-H., DAS, R., GRISHIN, N. V., AND BAKER, D. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 77 Suppl 9 (2009), 89–99.
- [125] RAMBAUT, A. Figtree, a graphical viewer of phylogenetic trees. See <http://tree.bio.ed.ac.uk/software/figtree> (2007).
- [126] RASMUSSEN, D. A., VOLZ, E. M., AND KOELLE, K. Phylodynamic inference for structured epidemiological models. *PLoS Computational Biology* 10, 4 (2014), e1003570.
- [127] RAYNER, G. D., AND MACGILLIVRAY, H. L. Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing* 12, 1 (2002), 57–75.
- [128] RIEDER, P., JOOS, B., VON WYL, V., KUSTER, H., GRUBE, C., LEEMANN, C., BÖNI, J., YERLY, S., KLIMKAIT, T., BÜRGISSER, P., ET AL. HIV-1 transmission after cessation of early antiretroviral therapy among men having sex with men. *Aids* 24, 8 (2010), 1177–1183.
- [129] ROBINSON, D. F., AND FOULDS, L. R. Comparison of phylogenetic trees. *Mathematical Biosciences* 53, 1 (1981), 131–147.
- [130] RODRIGUE, N., AND ARIS-BROSOU, S. Fast Bayesian choice of phylogenetic models: Prospecting data augmentation-based thermodynamic integration. *Systematic Biology* 60, 6 (2011), 881–887.
- [131] RONQUIST, F. Bayesian inference of character evolution. *Trends in Ecology and Evolution* 19, 9 (2004), 475–481.

- [132] ROUZINE, I. M., RODRIGO, A., AND COFFIN, J. Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. *Microbiology and Molecular Biology Reviews* 65, 1 (2001), 151–185.
- [133] SACKIN, M. “good” and “bad” phenograms. *Systematic Biology* 21, 2 (1972), 225–226.
- [134] SAITOU, N., AND NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4, 4 (1987), 406–425.
- [135] SANDERSON, M. J. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* 14, 12 (1997), 1218–1231.
- [136] SIMON-LORIERE, E., FAYE, O., FAYE, O., KOIVOGUI, L., MAGASSOUBA, N., KEITA, S., THIBERGE, J.-M., DIANCOURT, L., BOUCHIER, C., VANDENBOGAERT, M., CARO, V., FALL, G., BUCHMANN, J. P., MATRANGA, C. B., SABETI, P. C., MANUGUERRA, J.-C., HOLMES, E. C., AND SALL, A. A. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature* 524, 7563 (Aug 2015), 102–4.
- [137] SIMON-LORIERE, E., HOLMES, E. C., AND PAGÁN, I. The effect of gene overlapping on the rate of RNA virus evolution. *Molecular Biology and Evolution* 30, 8 (2013), 1916–1928.
- [138] SIMONS, K. T., BONNEAU, R., RUCZINSKI, I., AND BAKER, D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins: Structure, Function, and Bioinformatics* 37, S3 (1999), 171–176.

- [139] SOKAL, R. R. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38 (1958), 1409–1438.
- [140] SONG, Y., DIMAIO, F., WANG, R. Y.-R., KIM, D., MILES, C., BRUNETTE, T., THOMPSON, J., AND BAKER, D. High-resolution comparative modeling with RosettaCM. *Structure* 21, 10 (Oct 2013), 1735–42.
- [141] STADLER, T. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology* 261, 1 (2009), 58–66.
- [142] STADLER, T. Sampling-through-time in birth–death trees. *Journal of Theoretical Biology* 267, 3 (2010), 396–404.
- [143] STADLER, T. Simulating trees with a fixed number of extant species. *Systematic Biology* (2011), syr029.
- [144] STADLER, T., KOUYOS, R., VON WYL, V., YERLY, S., BÖNI, J., BÜRGISSER, P., KLIMKAIT, T., JOOS, B., RIEDER, P., XIE, D., ET AL. Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution* 29, 1 (2012), 347–357.
- [145] SWISS HIV COHORT STUDY. Cohort profile: the Swiss HIV Cohort study. *International Journal of Epidemiology* 39 (2010), 1179–1189.
- [146] TALAVERA, G., AND CASTRESANA, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* 56, 4 (Aug 2007), 564–77.
- [147] TAVARÉ, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* 17 (1986), 57–86.

- [148] TAVARÉ, S., BALDING, D. J., GRIFFITHS, R. C., AND DONNELLY, P. Inferring coalescence times from DNA sequence data. *Genetics* 145, 2 (1997), 505–518.
- [149] TOMAS, K., DHAMI, P., HOUSTON, C., OGUNNAIKE-COOKE, S., AND RANK, C. HIV in Canada: 2009 to 2014. *Canada Communicable Disease Report* 41, 12 (2015), 292.
- [150] TONI, T., WELCH, D., STRELKOWA, N., IPSEN, A., AND STUMPF, M. P. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface* 6, 31 (2009), 187–202.
- [151] TRAN, E. E., SIMMONS, J. A., BARTESAGHI, A., SHOEMAKER, C. J., NELSON, E., WHITE, J. M., AND SUBRAMANIAM, S. Spatial localization of the Ebola virus glycoprotein mucin-like domain determined by cryo-electron tomography. *Journal of Virology* 88, 18 (2014), 10958–10962.
- [152] VAN VALEN, L. A new evolutionary law. *Evolutionary Theory* 1 (1973), 1–30.
- [153] VOLZ, E., AND POND, S. Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. *PLoS Current Outbreaks* 6 (2014).
- [154] VOLZ, E. M., POND, S. L. K., WARD, M. J., BROWN, A. J. L., AND FROST, S. D. Phylodynamics of infectious disease epidemics. *Genetics* 183, 4 (2009), 1421–1430.
- [155] WALLINGA, J., AND TEUNIS, P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology* 160, 6 (2004), 509–516.
- [156] WANG, X., WU, Y., MAO, L., XIA, W., ZHANG, W., DAI, L., MEHTA, S. R., WERTHEIM, J. O., DONG, X., ZHANG, T., ET AL. Targeting HIV

- prevention based on molecular epidemiology among deeply sampled subnetworks of men who have sex with men. *Clinical Infectious Diseases* 61, 9 (2015), 1462–1468.
- [157] WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M., CLAMP, M., AND BARTON, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 9 (2009), 1189–1191.
- [158] WHO EBOLA RESPONSE TEAM. Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *The New England Journal of Medicine* 371, 16 (Oct 2014), 1481–95.
- [159] WILSON, J. A., HEVEY, M., BAKKEN, R., GUEST, S., BRAY, M., SCHMALJOHN, A. L., AND HART, M. K. Epitopes involved in antibody-mediated protection from Ebola virus. *Science* 287, 5458 (2000), 1664–1666.
- [160] WOODMAN, Z., AND WILLIAMSON, C. HIV molecular epidemiology: transmission and adaptation to human populations. *Current Opinion in HIV and AIDS* 4, 4 (2009), 247.
- [161] WORLD HEALTH ORGANIZATION. Global Health Observatory Data Repository 2013, 2013.
- [162] WU, H., ZHU, H., MIAO, H., AND PERELSON, A. S. Parameter identifiability and estimation of HIV/AIDS dynamic models. *Bulletin of Mathematical Biology* 70, 3 (2008), 785–799.
- [163] YANG, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution* 11, 9 (1996), 367–372.
- [164] YANG, Z., AND RANNALA, B. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Molecular Biology and Evolution* 14, 7 (1997), 717–724.

-
- [165] YILDIRIM, S., SINGH, S. S., DEAN, T., AND JASRA, A. Parameter estimation in hidden Markov models with intractable likelihoods using sequential Monte Carlo. *Journal of Computational and Graphical Statistics* 24, 3 (2015), 846–865.
- [166] YODER, A. D., AND YANG, Z. Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution* 17, 7 (2000), 1081–1090.
- [167] YOZWIAK, N. L., SCHAFFNER, S. F., AND SABETI, P. C. Data sharing: Make outbreak research open access. *Nature* 518, 7540 (2015), 477.
- [168] ZHUANG, J., JETZT, A. E., SUN, G., YU, H., KLARMANN, G., RON, Y., PRESTON, B. D., AND DOUGHERTY, J. P. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *Journal of Virology* 76, 22 (2002), 11273–11282.
- [169] ZUCKERKANDL, E., AND PAULING, L. Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins* 97 (1965), 97–166.