THE UNIVERSITY OF CHICAGO

EVOLUTION OF ADAPTABILITY AND THE IMMUNE RESPONSE TO INFLUENZA AND HIV

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY
MARCOS COSTA VIEIRA

CHICAGO, ILLINOIS
JUNE 2020

For my parents, Vevé and Walmick.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am immensely grateful for the support of many colleagues, collaborators and mentors during my time in graduate school. My advisor, Sarah Cobey, has been a constant source of encouragement, inspiration, exciting ideas and constructive criticism. I am especially thankful for the people Sarah brought together in the lab and for their friendship and support: Frank Wen, Sylvia Ranjeva, Rahul Subramanian, Ed Baskerville, Daniel Zinder, Phil Arevalo, Kangchon Kim, Katie Gostic, Spencer Carran and Lauren McGough. As if such an amazing group of people wasn't enough, we have been fortunate to share a conference room turned lunch space and philosophical café with our friends from the Pascual Lab: Mauricio Santos-Vega, Pamela Martinez, Shai Pilosof, Qixin He, Victoria Aznar and Sergio Alcala-Corona.

The members of my thesis committee, Mercedes Pascual, Patrick Wilson, Joe Thornton and Chung-I Wu, helped me make sense of the diversity of scientific fields this dissertation attempts to bring together, from protein evolution to immunology and disease ecology. I was aided in this challenge by the excellent courses taught by the faculty at the University of Chicago. Stefano Allesina, Greg Dwyer, Joe Thornton and David Alonso were particularly inspiring teachers.

I am thankful to the many collaborators who contributed directly to the research presented here. I have acknowledged them at the end of each chapter.

My PhD would not have progressed nearly as sanely without the support of many friends in the Ecology and Evolution department and the Darwinian Cluster. Special thanks to Ayse, Arvind, Iuri, Darli, Débora, Rebecca, Natalia and Matt. Audrey, Bonnie, Connie, Jeff and Mary always made sure things went smoothly on the administrative front.

Writing a thesis while hunkering down to help mitigate a global pandemic was only possible due to the company and loving support of Rachel, my partner and friend of many years. I am as thankful for the times she brought my mind away from work as I am for the times she encouraged me to work on my research.

Finally, I would like to thank my family for their love and support. My parents, Vevé and Walmick, have taught me the love of science I share with my brothers. Arlete has supported me more than she can imagine. Obrigado a todos.

ABSTRACT

The ability of populations to adapt depends on a complex suite of traits including mutation and recombination rates and functional constraints on genes subject to selection. How much selection can act on these traits to promote adaptability remains an open question. This dissertation investigates the ultimate example of evolved adaptability, the adaptive immune system of jawed vertebrates. Adaptive immunity has been selected in jawed vertebrates to recognize pathogens through recombination, mutation and selection in populations of B cells rapidly evolving during the immune response within an individual. This short-term adaptability is enabled by the diversity of immunoglobulin genes that recombine to produce B cell receptors and by the receptors' mutation rate during the immune response, features that have been shaped by selection in vertebrate populations over hundreds of millions of years. We explore how these features contribute to the adaptability of the immune response through computational and statistical analyses of the response to influenza and HIV. We ask if despite possible epistatic interactions among recombining immunoglobulin genes, specificity for particular pathogens is associated with individual genes. We find that influenza infection in mice selects for B cell receptors using specific immunoglobulin genes during the immune response, suggesting that selection of immunoglobulin genes in the long term could lead to their specialization for particular pathogens. Because the adaptability of B cells depends on mutations that change affinity for the antigen, we also investigate the short-term evolution of mutational hotspots in B cell receptor sequences. While the long-term evolution of immunoglobulin genes led to an abundance of hotspots in the antingen-binding loops of the receptor, we find that selection and neutral mutations disrupt those hotspots over years of coevolution between B cells and HIV, a loss of mutability that might limit the adaptability of B cell responses to chronic or repeated infections. Finally, we investigate how immunity arises from infection history and how the resulting protection shapes the ecology of influenza virus lineages. We conclude by proposing ways to integrate life-history and ecology into the study of the adaptability and specialization of immunity.

CHAPTER 1

INTRODUCTION

The rate at which populations adapt depends both on the selective pressures imposed by the environment and on traits that affect the ability of the population to respond [99, 173, 174]. This complex suite of traits, which includes the rates at which variation is introduced by mutation and recombination and the functional constraints on the proteins subject to selection, defines the population's adaptability [152, 178, 16, 123].

The degree to which selection can promote adaptability by acting on traits that contribute to it depends on the rate at which environmental changes reinforce or modify selective pressures. Because mutations are more likely to be deleterious than beneficial [48], individuals with a lower mutation rate tend to be fitter on average because the reduction in deleterious mutations in the offspring more than compensates for the reduction in advantageous ones. Selection therefore tends to decrease the mutation rate [83, 159, 108, 154], incidentally decreasing adaptability. In rapidly changing environments, however, alleles that increase the mutation rate may become fixed by hitchhiking with beneficial mutations, which themselves arise more frequently in association with "mutator" alleles. Such indirect selection for an increased mutation rate has been demonstrated in theoretical models [97, 160, 151, 136] and experimental evolution studies [152, 123, 135]. Highly variable or rapidly changing environments can therefore select for traits that increase adaptability.

In jawed vertebrates, the highly variable and rapidly changing environment represented by pathogens led to an extreme case of evolved adaptability. Recognition of thousands of potential antigens is achieved through the recombination of immunoglobulin gene segments that recombine to produce B cell receptors, the precursors of antibodies [71, 19, 165, 75]. During this process, B cells undergo somatic hypermutation of the B cell receptor gene by specialized mutagenic enzymes and are selected based on their receptor's ability to recognize the antigen [63, 106, 168]. Mutation and selection of immunoglobulin genes therefore happen on two different scales: over hundreds of millions of years in vertebrate populations, during

which selection shapes the germline sequences of immunoglobulin genes prior to their recombination in B cells, and within weeks of an infection in B cell populations, during which selection acts on somatic mutations introduced into the recombined B cell receptor gene. The diversity of immunoglobulin genes and the frequency and distribution of the somatic mutations introduced into them during the immune response bear the signature of selection in in the long term to increase the adaptability of B cells in the short term. Understanding how these features facilitate the rapid adaptation of B cells in response to infection is the goal of this dissertation.

1.1 The evolutionary dynamics of antibody responses

Antigen recognition by the adaptive immune system of jawed vertebrates is done by B cells and T cells, each carrying on its surface specific antigen receptors produced by the recombination of germline gene segments [71, 19, 165, 75]. This combinatorial origin results in an extremely diverse repertoire of immune receptors collectively capable of binding a vast number of antigens [46, 47]. Once a previously inexperienced ("naive") B cell or T cell encounters an antigen, it divides and expands into a clonal population of cells that help clear the pathogen and persist as memory against future infections [22, 161]. In the case of B cells, the clonal expansion following the initial activation of the cell is accompanied by selection of mutations that improve antigen recognition by the B cell receptor [63, 106, 168]. B cells exiting this evolutionary process secrete their increasingly potent B cell receptors as antibodies.

The germline genes that recombine to produce the B cell receptor are known as variable (V), diversity (D) and joining (J) genes [71, 19, 165, 75]. Each B cell displays multiple identical copies of a single B cell receptor on its surface. Each receptor consists of two identical heavy chains and two identical light chains, and each chain has a variable region and a constant region. The variable region encodes the parts of the protein directly involved in antigen binding (loops known as "complementarity-determining regions", CDRs), inter-

spersed with structurally important "framework regions" (FRs) that are less often involved in antigen binding. The variable region of the heavy chain is formed by the recombination of one gene each from the V, D and J sets, while the variable region of the light chain lacks a D segment and is formed by the recombination of V and J sets that are different from the heavy chain V and J. Different combinations of V, D and J genes, and different combinations of the resulting heavy and light chains, result in B cell receptors with different molecular structures capable of recognizing different antigens.

While the long-term evolution of germline genes plays out in populations of organisms over hundreds of millions of years, its short-term counterpart plays out in populations of B cells in the days and weeks following the initial recognition of an antigen by a naive B cell [63, 106]. Activated B cells enter microanatomical structures known as germinal centers, where they bind and take up antigen at a rate that depends on the affinity of the B cell receptor [168, 163]. B cells then break down the antigens into small peptides and present them to helper T cells, which give B cells signals that allow them to survive and replicate. As B cells divide in germinal centers, they undergo "somatic hypermutation" of the variable region of the B cell receptor gene [164]. Somatic hypermutation introduces variation in affinity in the B cell population, and competition for antigen selects for B cells with high-affinity receptors. The ability of B cell populations to evolve rapidly during the course of the response is therefore key to the development of potent antibodies.

1.2 Adaptability and the diversity of germline immunoglobulin genes

Immunoglobulin V, D and J genes were present at the ancestor of jawed vertebrates around 500 million years ago and have since diversified via gene duplication and point mutations, with most of the diversity concentrated in the V gene set [110, 50, 34, 51, 75, 28]. Remarkably, jawless vertebrates have independently evolved a similar combinatorial mechanism for

generating immune receptors using a different family of gene segments [124].

Beyond the pressure to increase the number of potential antigens bound by the B cell repertoire, the selective pressures driving the long-term evolution of immunoglobulin germline genes are poorly understood. Theoretical models suggest the optimal distribution of specificities in the repertoire is shaped by a trade-off between binding commonly encountered pathogens quickly and having sufficient diversity to recognize rare pathogens [111]. How these demands affect the diversification of germline genes, however, is unclear. While the structure and specificity of the B cell receptor depend on the pair of recombined genes encoding the heavy and light chains and their specific V, D and J combinations, it is possible that some germline genes are more likely than others to recombine into a receptor that binds a particular pathogen well. These differences could arise by chance during the diversification of germline genes but could also be shaped by selection. For instance, it has been hypothesized that selection might increase the naive frequency of germline genes particularly likely to bind common pathogens and "hardwire" mutations into the germline sequence that would otherwise be acquired during the short-term evolution of B cell lineages using those genes [95, 28]. This process would amplify initial differences among germline genes in their propensity to recognize specific pathogens, yet it is unclear to what extent such variation exists.

In addition to the selection of specific B cell receptors, variation in germline genes' propensity to bind particular antigens would result in selection at the level of germline genes during the immune response. B cells whose receptor uses particular germline genes would be more successful than others at entering the experienced repertoire and undergoing clonal expansion. Selection of specific germline genes might cause antibodies using those genes to dominate the response of different individuals responding to the same pathogen. Evidence for this hypothesis comes from the observation that antibodies targeting the same antigen but isolated from multiple people tend to use similar sets of V genes [44]. Much of this evidence comes from antibodies studied for their ability to target particular epitopes in antigens of

interest, for instance conserved sites on the surface proteins of HIV and influenza viruses [59, 11, 87]. However, antibodies against a particular pathogen can potentially target several epitopes in different antigens, and this diversity of potential targets might select for antibodies using many different V genes [92]. A systematic attempt to test for V gene selection across the B cell repertoire found limited similarity in the V genes used by people immunized with influenza [156]. Differences in the epitopes and antigens targeted by different people, for instance due to differences in infection history, might explain people's discordant V gene usage.

1.3 Adaptability and the mutability of B cell receptors

The mutation rate of the B cell receptor is partly controlled by the B cell receptor sequence itself, as the enzymes responsible for introducing mutations target different nucleotide motifs in the sequence with different probabilities [138, 137, 121, 185, 31]. The frequency and distribution of highly mutable motifs – mutational "hotspots" – suggests germline gene sequences have been selected in the long-term to increase the short-term mutability of regions directly involved in antigen binding (CDRs) while reducing the mutability of structurally important regions (FRs) [122]. Due to the distribution of nucleotide motifs in the germline sequence, mutations in the antigen binding regions are also more likely to result in changes between amino acids with different biochemical properties [175, 81, 67, 140].

The sequences of germline genes therefore appear to have been selected to improve the adaptability of B cell receptors during the immune response by focusing mutations where they are likely to create variation in affinity while keeping mutations away from regions where they are likely to be deleterious. How the finely tuned distribution of mutability in the naive B cell receptor changes during the response and affects the subsequent adaptability of B cell populations, however, is poorly understood. Loss of mutational hotspots has been proposed as an explanation for the observed decline in the substitution rates of B cell lineages coevolving with HIV [145, 68], but the contributions of neutral mutations, selection for

affinity and indirect selection for mutability to hotspot gains and losses are unknown.

1.4 Adaptability and immune memory

In addition to the diversity of germline genes and the ability of B cell receptors to evolve high affinity, B cell responses owe their adaptability to memory. Specialized memory B cells emigrate from germinal centers at various points during the evolution of B cell lineages and persist in the body after the infection has been cleared [93]. Upon a new infection with a previously encountered pathogen, memory cells can be quickly re-activated to differentiate into antibody-secreting plasma cells. Memory B cells can also reenter germinal centers to undergo additional affinity maturation, but the extent to which they do so is unclear [38, 115].

Although immune memory completely prevents reinfection with many pathogens, antigenically variable pathogens can often reinfect hosts by escaping immune memory. For instance, the build-up of immune memory in the human population drives the evolution of influenza viruses by selecting for antigenically novel strains capable of evading existing memory [24, 49, 150, 14, 13]. As a result, humans are infected multiple times since early childhood with many different influenza strains belonging to different types, subtypes, lineages or clades [82, 139, 94].

The evolution of immune escape by influenza viruses has traditionally been studied under the assumption that host immune memory can remember each strain independently from the others [150, 12, 190]. Thus, hosts are assumed to be protected against any strain that is antigenically close to one of the strains the host has encountered in the past, regardless of the order in which past strains were encountered. This approach led to the development of "antigenic maps", where antigenic distances between strains are estimated by infecting a naive animal model with a strain and measuring the ability of the resulting antibodies to bind another strain [150, 85, 52].

While much has been learned by studying the evolution of immune escape under these assumptions, decades of research in immunology have established that the memory developed

against new strains is not in fact independent of the memory developed against previous ones [27]. Antibody titers appear to be highest against strains encountered early in life, a phenomenon termed "original antigenic sin" [35, 36, 53]. This effect suggests preexisting B cell lineages in the memory population interfere with the dynamics and evolution of new lineages recruited from the naive repertoire upon a new infection. For instance, preexisting antibodies might clear the antigen or mask epitopes on its surface, effectively limiting the amount of antigen available to stimulate the evolution of new B cell lineages in germinal centers [187, 188].

While the effect of immune history on the development of antibodies to new infections is clear, how this effect translates into differences in protection among hosts with different exposure histories is less clear. Early infections can lead to permanently increased protection against the variants first encountered and explain the distinct age distributions of medically attended infections with different groups and subtypes of influenza type A, a phenomenon termed "immune imprinting" [60, 61, 10]. However, for influenza variants closer to each other than are the influenza A subtypes, cross-protection across lineages is likely to play a role in addition to, or independently of, imprinting protection from the earliest exposures. The two lineages of influenza type B, for instance, diverged much more recently than the influenza A subtypes [139, 43, 94], and differences in the age distribution of medically attended infections with each lineage suggest possible immune history effects [162, 153, 171, 147, 128, 172]. Yet it is unclear how memory of past infections with each lineage leads to protection and how protection affects the lineages' distinct age distributions.

1.5 Overview of dissertation

The next three chapters investigate the adaptability of B cell responses in the context of the diversity and selection of germline genes, their subsequent evolution as B cell receptors under selection for increased affinity, and the development of protection from immune memory.

In Chapter 2, we investigated variation in V genes' propensity to recognize specific

pathogens, which might affect the predictability of short-term antibody evolution and shape the long-term evolution of immunoglobulin genes. To control for differences in infection history that might affect variation in the antigens and epitopes targeted by different individuals, we studied the B cell response of mice infected once or twice with the same influenza strain. We developed a simple mathematical framework for defining and estimating selection of germline genes, as opposed to selection of individual B cell receptors that use a specific combination of genes. Using a statistical test designed to detect differentially abundant taxa in microbiome studies, we found several V genes that consistently increased in frequency between the naive and experienced repertoires of different mice infected with influenza, suggesting those genes were positively selected by influenza antigens. We also found genes with evidence of negative selection. These results suggest influenza infections select for similar germline genes in naive individuals or individuals with identical infection histories. This similarity in V gene usage might help explain why antibodies targeting particular antigens or epitopes evolve with similar probability in different individuals. Systematic characterization of functional differences between antibodies using different V genes and of differences in V usage between people with different infection histories might help vaccine design efforts by identifying groups more or less likely to develop antibodies targeting particular epitopes.

Chapter 3 asked how the finely tuned mutability of immunoglobulin genes changes during the short term evolution of B cell receptors. During the response to chronic HIV infection, loss of mutational hotspots and changes in their distribution across CDRs and FRs are predicted to compromise the adaptability of B cell receptors, yet the contributions of different mechanisms to gains and losses of hotspots remain unclear. We reconstructed changes in anti-HIV B-cell receptor sequences and show that mutability losses were more frequent than gains in both CDRs and FRs, with the higher relative mutability of CDRs maintained throughout the response. Nonsynonymous substitutions caused most of the mutability loss in CDRs. Because CDRs also show strong positive selection, this result suggests that selection for mutations that increase binding affinity contributed to loss of mutability in antigen-binding

regions. Although recurrent adaptation to evolving viruses could indirectly select for high mutation rates, we found no evidence of indirect selection to increase or retain hotspots. Our results suggest mutability losses are intrinsic to both the neutral and adaptive evolution of B-cell populations and might constrain their adaptation to rapidly evolving pathogens such as HIV and influenza[1].

Chapter 4 investigated the development of immune memory and protection from infection history. Differences in cohorts' infection histories might explain the unusual age distributions of medically attended infections with the two lineages of influenza type B, but how a person's infection history translates into protection is not fully understood. Fitting a statistical model to case data, we found that differences in the lineages' age distributions could emerge from historical changes in lineage frequencies combined with strong cross-protection between strains of the same lineage and asymmetric cross-protection between lineages. Consistent with previous serological observations, B/Victoria infections reduce the risk of B/Yamagata medically attended infections, whereas B/Yamagata infections protect much less against B/Victoria infections. We hypothesize these results arise from asymmetric patterns of epitope immunodominance between the lineages. Characterizing antibody specificity after infection and vaccination could help explain variation in cohorts' susceptibility to influenza arising from differences in infection history.

Finally, Chapter 5 takes stock of these findings and proposes avenues for investigating the potentially conflicting demands for adaptability and specialization in the immune system in light of species' ecology and life-histories. Species' longevity and body size, in particular, might impose constraints on the adaptability and specialization of immune receptors.

1. The results from Chapter 3 have been published previously [170], and much of the text in this paragraph and in the chapter appears in the published manuscript.

CHAPTER 2

SELECTION OF GERMLINE IMMUNOGLOBULIN V GENES BY PRIMARY AND SECONDARY INFLUENZA INFECTIONS IN MICE

2.1 Introduction

The remarkable adaptability of the vertebrate immune system arises from the recombination of germline genes encoding immune receptors [71, 19, 165]. B cell receptors, the precursors of secreted antibodies, are formed by the recombination of variable (V), diversity (D) and joining (J) germline immunoglobulin genes [75]. In each maturing B cell, one gene from each set is recombined into a mature gene encoding the receptor's heavy chain, while one gene each from separate V and J sets are recombined into a gene encoding the light chain. The resulting combinatorial diversity, with additional variation from insertions and deletions at the segments' junctions, produces a repertoire of "naive" (antigen-inexperienced) B cell receptors collectively capable of binding a vast number of potential antigens [47]. Antigens encountered upon infection or vaccination select for naive B cells capable of binding the antigen with sufficiently high affinity [189, 167], and activated B cells can then undergo clonal expansion and selection for somatic mutations that further improve antigen binding [63, 106, 169]. Immunoglobulin genes therefore evolve on two different timescales. Their germline sequences evolve in populations of vertebrates over hundreds of millions of years, and their recombined sequences forming the B cell receptor evolve in populations of B cells in a single individual within weeks of infection.

The long-term evolution of the adaptive immune system has been shaped by selection to improve the short-term adaptability of the B cell response, but how these selective pressures shaped the diversity of immunoglobulin genes is unclear. Immunoglobulin genes originated in the ancestor of jawed vertebrates around 500 million years ago and have since diversified

by gene duplication and point mutations [110, 50, 34, 51]. Was this diversification purely driven by selection to maximize the spectrum of potential specificities in the B cell repertoire? Or have some B genes evolved in the long term under selection to recognize particular pathogens? Although the structure and specificity of a B cell receptor depend on the pairing of the heavy and light chains with their specific combinations of V, D and J genes, some immunoglobulin genes might be more likely than others to recombine into a receptor that binds well to a particular pathogen. These differences could arise by chance during diversification of germline immunoglobulin genes and be subsequently amplified by selection. For instance, it has been hypothesized that the germline sequences of immunoglobulin genes with a high propensity to bind commonly encountered pathogens might be selected to hard-code mutations that improve recognition of those pathogens [28]. Those mutations could also be somatically generated and selected during the short-term evolution of B cell lineages using those genes. However, when the pathogens driving this selection are common, selection in vertebrate populations might favor the same mutations in the germline sequence itself.

Whatever its origin, variation across germline genes in their propensity to recognize a specific pathogen might affect the predictability of the antibody response. Whether they arose by chance or were the product of selection in the long-term, these differences would result in short-term selection not only of individual B cell receptors with specific V, D and J combinations, but also selection of germline immunoglobulin genes themselves. B cells using germline genes more likely to recombine into successful receptors would be more likely to become activated by the antigen and subsequently expand into a clonal population. As a result, the frequency of those genes would increase between the naive and experienced B cell repertoires. Strongly selected genes might dominate the response to the same antigen in different individuals, potentially leading to functionally similar responses. Understanding the predictability of germline gene usage and its functional consequences might thus inform efforts to induce antibodies targeting specific epitopes, such as conserved sites in HIV and influenza virus antigens [104, 181, 40, 23, 78, 169].

How much individual germline immunoglobulin genes vary in their ability to recognize specific pathogens, however, is unclear. Studies of individually isolated monoclonal antibodies (mAbs) suggest mAbs targeting specific antigens isolated from different people use similar V genes (e.g. [74, 59, 11, 87], reviewed in [44]). However, most of these studies focused on mAbs chosen for their ability to bind specific epitopes on antigens of interest, such as conserved sites in HIV and influenza proteins that might become the targets of new vaccines [59, 11, 87]. The total B cell response against a pathogen, however, may consist of thousands of B cell lineages potentially targeting multiple epitopes on different antigens. Given this diversity of potential targets, it is unclear if different individuals responding to the same pathogen tend to use similar V genes. Most studies of monoclonal antibodies also have not tested similarity in V gene usage statistically.

Many studies have attempted to identify signatures of infection with particular pathogens in the B cell repertoire without focusing on specific epitopes, but few have systematically tested for selection of individual immunoglobulin genes. Most repertoire-level studies focus instead on identifying specific amino acid sequences in an antigen binding region spanning the V, D and J genes, sequences that could arise from many different combinations of genes and from selection of somatic mutations during the response [126, 76, 143]. However, two studies tested for selection of V genes by looking for V genes increasing in frequency over time across multiple people given the same influenza vaccine [30, 156]. Both studies found only 2-3 genes that appeared to be particularly successful at binding influenza. As one of those studies demonstrated, parallel increases in the frequency of some V genes across people can be an artifact of correlated V gene frequencies across people before vaccination [156].

Thus, there is limited evidence of V-gene level selection during the immune response to particular pathogens. Stochasticity might lead to different genes dominating the response in different individuals, for instance if the earliest responding naive B cells tend to dominate the response and they happen to use different genes in different people, or if highly beneficial mutations happen to appear in clones using different V genes. Alternatively, differences in the

sets of V genes selected by a particular pathogen might be due to differences in the antigens and epitopes targeted by different people, for instance due to differences in infection history. In influenza, infection history strongly affect antibody responses against strains encountered later [35, 36, 53, 27]. Differences in the antigens and epitopes targeted by people with different infection histories might explain why different sets of V genes increase in frequency in different people exposed to the same influenza strain. However, detailed infection histories are rarely known as humans are repeatedly infected with influenza [149, 54, 37, 70, 133], and studying the response in previously naive individuals is usually not possible since the primary infection occur in the first years of life [142, 17, 141].

To investigate selection at the level of immunoglobulin V genes during the immune response, we sequenced and analyzed the naive and experienced B cell repertoires of mice after primary and secondary infection with an H1N1 virus. We found that statistical methods designed to identify microbial species that vary in abundance across samples can be used to identify V genes under negative and positive selection, and we developed a simple mathematical framework to quantify selection at the level of V genes. Using these methods, we found 9 positively selected genes that increased in frequency consistently between the naive and experienced repertoires of individual mice infected with influenza, and 11 negatively selected genes that consistently decreased in frequency. These results suggest that despite the diversity of potential antigens and epitopes targeted and despite epistatic interactions among V, D and J genes and between heavy and light chains, some immunoglobulin genes are particularly likely to recombine into a receptor that binds a specific pathogen well. Selection of immunoglobulin genes by specific pathogens during the immune response might lead to selection on those genes' germline sequences in the long term to improve recognition of specific pathogens. Systematic characterization of the specificity and affinity of antibodies using different V genes might reveal how much a bias toward particular V genes constraints the antigens and epitopes targeted during the response to influenza and thus inform efforts to induce antibodies targeting specific epitopes.

2.2 Results

To study selection of V genes during the the primary and secondary response to influenza infection, we sorted B cells from ten infected C57BL/6 mice and two controls on each of 8, 16 and 24 days after infection with a mouse-adapted H1N1 strain (Materials and Methods: "Infection experiments"). Additional groups were given a secondary infection 32 days after the first infection and sacrificed on days 40 and 56 after the primary infection. Infected mice showed a marked increase in antibody titers against the infecting virus between days 8 and 16 (Supplementary Information, Fig. 2.3). We sorted naive, germinal center, memory and plasma cells from the spleen, bone marrow and lymph nodes and bulk sequenced their B cell receptor heavy chains using cDNA sequencing (Materials and Methods: "B cell receptor sequencing"). We did not sort cells based on influenza specificity and therefore cannot distinguish between receptors that are able to bind influenza and receptors that are not. We used partis [132, 130, 131] to identify the most likely V, D and J genes used by each B cell receptor sequence while simultaneously inferring new alleles and clustering sequences into clones (sets of sequences likely descended from the same naive B cell). After estimating and accounting for sequencing error (Materials and Methods: "Estimating amplification and sequencing error"), we computed the frequency of each V gene in the naive repertoire and in each compartment of the experienced repertoire (germinal center, memory and plasma cells) and the combined frequency in the experienced repertoire.

The median number of unique sequences across mice was 23,327 (range 1,327-55,190) for naive cells, 4,902 (1,139-13,900) for germinal center cells, 14,444 (1,397-38,359) for plasma cells and 6,142 (1,050-18,284) for memory cells. To limit error in the estimates of naive and experienced frequencies, these numbers exclude mouse-cell type combinations with fewer than 1,000 unique sequences, and we excluded mice with fewer than 1,000 unique naive sequences altogether. Mice used a median of 77 (range 53-91) V genes in the naive repertoire and 78 (range 43-92) V genes in the experienced repertoire, with no significant differences in the number of V genes used between infected mice and controls (medians for control,

primary and secondary groups 77, 79.5 and 77 V genes in the naive repertoire, 77, 78.5 and 77.5 V genes in the experienced repertoire). V gene frequencies in the naive and experienced repertoire varied by four orders of magnitude, with the 25 most common genes each typically used by 1-10% of the B cell receptor sequences (Supplementary Information, Figs. 2.4 and 2.5).

2.2.1 Defining and quantifying V gene selection

Positively (negatively) selected genes are expected to increase (decrease) in frequency between the naive and experienced repertoires because they are more (less) likely than non-selected genes to recombine into B cell receptors capable of recognizing a particular antigen or because they do so with higher (lower) affinity on average. To formalize this idea, consider a set of V genes such that the number of naive B cells whose receptor uses gene i is N_i. Given a background antigenic environment (e.g., the non-pathogenic microbes found in a mouse facility), assume α is the probability of activation across B cells regardless of which V gene they use and β is the baseline number of B cells produced by an activated B cell regardless of its V gene. The number of B cells using gene i in the experienced repertoire is then given by $E_i = N_i\alpha\beta$. Thus, this neutral model assumes that in general experienced and naive V gene frequencies are correlated, an assumption supported by the sequence data from our experiment (Supplementary Information, Fig. 2.6).

Positive and negative selection at the level of individual V genes during the immune response can be described by a selection factor S_i^α modifying the activation probability of a B cell using gene i and by a selection factor S_i^β modifying the average number of B cells produced by the activated cell, with values greater than 1 indicating positive selection and values between 0 and 1 indicating negative selection. After accounting for germline gene selection, the number of cells using gene i in the experienced repertoire is thus $E_i = N_i\alpha S_i^\alpha\beta S_i^\beta = N_i S_i\alpha\beta$, where the "total" selection factor of gene i, S_i, is the product of the activation and replication selection factors. Selection of a V gene in this context is therefore

separate from selection of individual B cell receptors using a specific combination of V, D and J, since B cell receptors using a non-selected V gene can still be selected (i.e., activated by antigen) with probability α. Selection of a V gene in this context is also different from the long-term selection of germline genes in vertebrate populations and from the short-term selection of maturing B cells before they enter the naive repertoire based on their ability to express a functional receptor that does not recognize self-antigens. Selection might occur without a specific pathogen challenge. In the case of our experiment, for instance, some genes might have a higher propensity than others to recognize non-pathogenic antigens present in the mouse facility. However, genes selected by influenza antigens are expected to increase in frequency between the naive and experienced repertoires in infected mice but not in controls.

The relationship between selection factors as defined above and the classical selection coefficient expressed as the difference between instantaneous per-capita growth rates depends on the precise dynamics of the B cell populations. We illustrate this relationship for the case of exponential growth (Materials and Methods: "Details of the mathematical"), but we note that selection factors can be used to estimate the strength of immunoglobulin gene selection when the precise dynamics of B cell populations are unknown and instantaneous growth rates cannot be easily estimated, as is the case for our data.

Detecting selected genes based on their change in frequency is challenging because a single positively or negatively selected gene changing in frequency will cause all other genes to decrease or increase in frequency, respectively, due to the constraint that frequencies sum to one. Under the model of selection above, the ratio of experienced-to-naive frequencies for a gene i, ρ_i, is given by (Materials and Methods: "Details of the mathematical model"):

$$\rho_i = \frac{S_i}{\Lambda} \tag{2.1}$$

where Λ is the mean selection factor in the naive repertoire. Thus, even a gene under no

selection ($S_i = 1$) can decrease or increase in frequency, if the average gene in the naive repertoire has a selection factor greater or smaller than one, respectively.

However, if selected and non-selected genes can be identified, Eq. (2.1) provides a way of estimating the selection factors of selected genes. Because $\rho_i = \Lambda^{-1}$ for non-selected genes, their experienced-to-naive frequency ratios can be used to estimate Λ, which can then be used to estimate the selection factors of selected genes from their observed experienced-to-naive ratios. An analogous approach has been used to estimate a compositional scale factor for differentially expressed genes or differentially abundant taxa without explicitly linking it to selection factoris [91].

Differential abundance analysis reveals positively and negatively selected genes after influenza infection

While the experienced-to-naive frequency ratio alone cannot be used to tell if a gene is under selection, non-selected genes are distinguished from selected genes by the fact that the ratio of frequencies of two non-selected genes is the same in the naive and experienced repertoires, whereas the ratio of frequencies between a selected gene and a non-selected gene changes between the naive and experienced repertoires (Materials and Methods: "Details of the mathematical model"). This property has been used to detect differentially abundant taxa in microbiome data or differentially expressed genes in gene expression studies [91, 20], with genes or taxa whose relative proportions are constant between samples inferred to each have the same abundance in different groups of samples.

We simulated experienced B cell repertoires by sampling V genes from the observed naive repertoires of the experimental mice while randomly choosing some genes to be under positive or negative selection, and we found that the method proposed by Brill *et al.* [20] correctly identifies positively and negatively selected genes for a range of simulated selection factors while having low false discovery rates (Materials and Methods: "Differential abundance test and power analysis"; Supplementary Information, Figs. 2.7 and 2.8). Briefly, the method

Table 2.1: Positively and negatively selected genes after primary and secondary influenza infection in mice.

	Infection group	V gene	Median selection factor (1st and 3rd quartiles)	n mice frequency increase (%)	n mice frequency decrease (%)
Positive selection	secondary	IGHV1-53*01	3.01 (1.85-4.27)	15 (100%)	0 (0%)
	secondary	IGHV7-1*03	3.21 (2.04-6.77)	15 (100%)	0 (0%)
	secondary	IGHV4-1*01	3.75 (1.85-8.74)	14 (93.33%)	1 (6.67%)
	secondary	IGHV6-3*01	2.38 (1.51-2.95)	14 (93.33%)	1 (6.67%)
	primary	IGHV12-3*01	2.50 (1.36-5.34)	16 (88.89%)	2 (11.11%)
	primary	IGHV14-4*01	1.59 (1.1-2.24)	15 (83.33%)	3 (16.67%)
	primary	IGHV2-9*01	1.91 (1.47-2.99)	15 (83.33%)	3 (16.67%)
	primary	IGHV4-1*01	1.92 (1.27-3.51)	15 (83.33%)	3 (16.67%)
	primary	IGHV7-1*03	1.53 (1.01-4.76)	14 (77.78%)	4 (22.22%)
	secondary	IGHV11-2*01	9.53 (0.95-16.25)	11 (73.33%)	4 (26.67%)
	primary	IGHV1-55*01	1.22 (0.95-1.62)	13 (72.22%)	5 (27.78%)
	primary	IGHV11-2*01	3.56 (1.44-15.31)	13 (72.22%)	5 (27.78%)
Negative selection	primary	IGHV1-5*01	0.43 (0.32-0.61)	3 (16.67%)	15 (83.33%)
	primary	IGHV1-54*01	0.57 (0.37-0.78)	3 (16.67%)	15 (83.33%)
	primary	IGHV5-4*01	0.51 (0.35-0.72)	3 (16.67%)	15 (83.33%)
	primary	IGHV5-9*01	0.47 (0.38-0.67)	3 (16.67%)	15 (83.33%)
	primary	IGHV1-15*01	0.51 (0.38-0.7)	2 (11.11%)	16 (88.89%)
	primary	IGHV5-2*02	0.37 (0.18-0.5)	2 (11.11%)	16 (88.89%)
	primary	IGHV5-6*01	0.42 (0.27-0.56)	2 (11.11%)	16 (88.89%)
	primary	IGHV1-59*01	0.45 (0.39-0.6)	1 (5.56%)	17 (94.44%)
	primary	IGHV5-17*03	0.50 (0.29-0.65)	1 (5.56%)	17 (94.44%)
	primary	IGHV5-9-1*02	0.51 (0.37-0.65)	1 (5.56%)	17 (94.44%)
	primary	IGHV1-81*01	0.54 (0.41-0.62)	0 (0%)	18 (100%)

assumes most genes are not under selection and heuristically identifies a reference set of non-selected genes by looking for a set of genes whose relative proportions remain the same across naive and experienced repertoires. Genes whose frequency relative to the combined frequency of reference genes is significantly different between the experienced and naive repertoires are inferred to be selected, but the test itself does not indicate the direction of the difference. We used Eq. (2.1) to estimate the selection factors of detected genes by taking the median estimate across individual mice, using the first and third quartiles as a measure of uncertainty.

We applied the method to the naive and experienced repertoires of infected and control mice and identified positively and negatively selected genes in the primary and secondary infection groups but not in the controls (Table 2.1). Seven genes were found to be under positive selection in mice subject to a primary infection alone, and five genes were found to be under positive selection in mice subject to both the primary and secondary infections, with

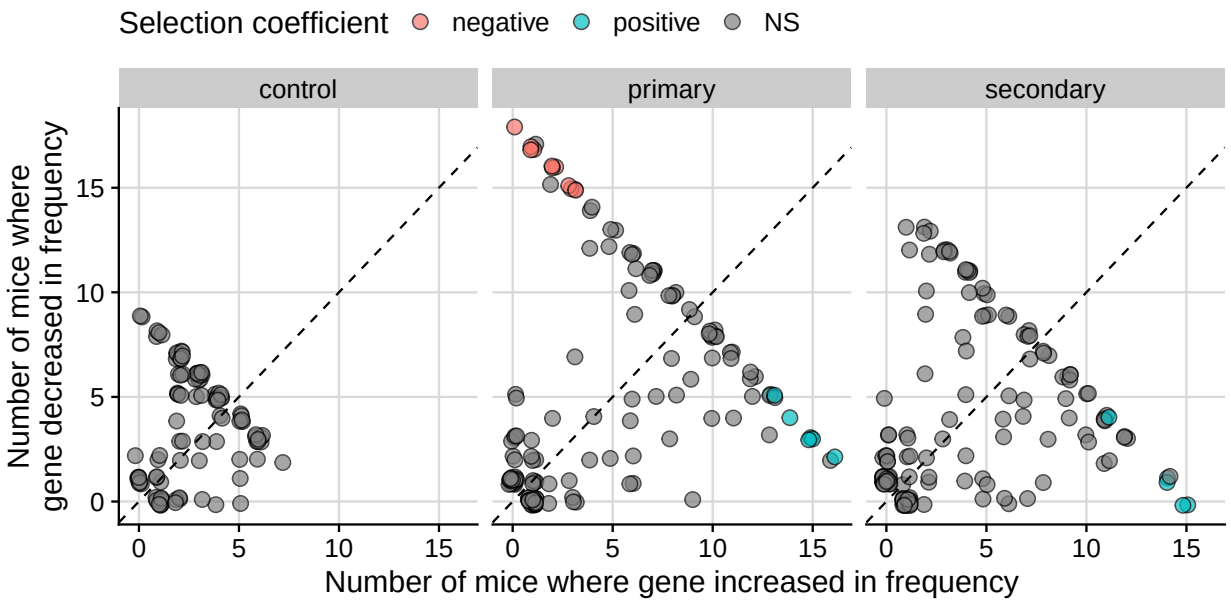

Figure 2.1: Number of mice in which each gene increased or decreased in frequency between the naive and experienced repertoires. Each circle represents a gene and is colored according to whether the gene is positively selected, negatively selected or non-significantly different in frequency between the experienced and naive repertoire. Circle positions have been slightly jittered to improve visualization. Not all genes are present in all mice.

three genes inferred to be under positive selection in both groups. Eleven genes were found to be under negative selection in the primary infection groups, but no genes were found to be under negative selection in the secondary infection groups, perhaps due to limited power to detect genes with a selection factor in the range of those detected in the primary infection group (Supplementary Information, Fig 2.7). Although there were fewer mice in the control group than in the infected groups, our power analysis suggests that if some genes were consistently increasing or decreasing in frequency in the controls, the differential abundance test would have been able to detect some of them (Supplementary Information, Fig 2.7). Thus, positive and negative selection of V genes in the infected mice appear to have been driven by influenza.

Naive B cells using the most consistently increasing genes (IGHV12-3*01 in 16/18 primary infection mice, IGHV1-53*01 and IGHV7-1*03 in 15/15 secondary infection mice) were 2-9 times more successful at entering the experienced repertoire of an infected mouse and

subsequently expanding into an experienced B cell population than B cells using non-selected genes (Table 2.1, Fig. 2.1). B cells using the most consistently decreasing gene (IGHV1-81*01 in 18/18 primary infection mice) were 40-60% as successful as B cells using non-selected genes.

2.2.2 Positively selected genes make up a limited fraction of the experienced repertoire but increase in frequency in late germinal centers

Next we asked if selection of V genes by the influenza infection differed between specific compartments of the experienced repertoire (germinal center, memory and plasma cells). Germinal center cells undergo affinity maturation via selection of mutations that improve antigen binding [168], and some germinal center cells exit germinal centers and differentiate into plasma cells, which actively secrete antibodies to help clear the infection, and memory cells, which can be reactivated upon future re-exposure. Because germinal center and plasma cell populations typically peak soon after an infection [163] whereas the memory compartment may reflect exposure to various previously encountered antigens, genes positively selected by influenza might be proportionally more common in the B cell receptors of germinal center cells or plasma cells than in memory cells.

Our power analysis suggests there is limited power to detect V genes specifically selected in individual compartments given the number of sequences in each compartment (Supplementary Information, Fig 2.9). Thus, instead of identifying different sets of selected genes specific for each compartment, we took the set of genes identified as positively selected at the level of the entire experienced repertoire of either primary or secondary groups and calculated the combined frequency of those same genes in each compartment, both in infected mice (in which those genes were inferred to be under positive selection) and in controls.

We found that positively selected genes made up a small proportion of the experienced repertoire of infected mice but appeared to increase in frequency in germinal centers late in the response. Despite selected genes increasing in frequency consistently across infected mice

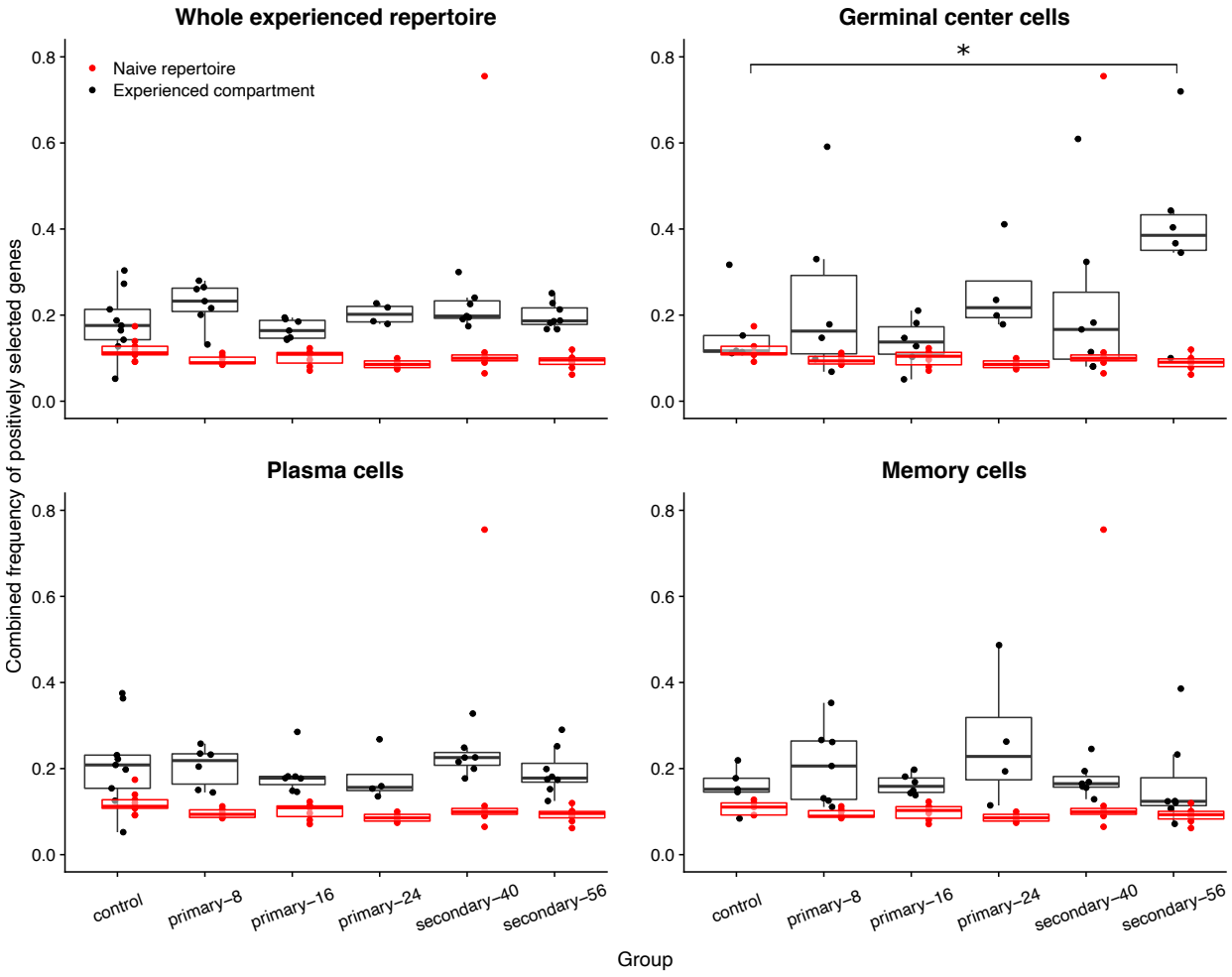

Figure 2.2: Combined frequency of genes identified as positively selected based on their frequencies in the entire experienced repertoire. The same genes are represented in each panel, with their combined frequency in the naive repertoire shown in read and the experienced repertoires in the compartment corresponding to each panel shown in black. Each point represents a mouse. A horizontal bar with an asterisk indicates a significant difference ($P < 0.05$).

but not in controls, the combined frequencies of selected genes in the experienced repertoire of infected mice was not significantly higher than the frequency of the same genes in controls (median 19%, range 5-30% across all mice) (Fig 2.2). This result might be due to influenza specific B cell receptors making up only a small proportion of the experienced repertoire of infected mice, leading to a lack of power to detect differences in the combined frequency of selected genes between infected mice and controls. The naive frequencies of selected genes spanned three orders of magnitude (Supplementary Information, Fig. 2.10), and despite increasing due to selection, the frequency of some selected genes in the experienced repertoire remained small (e.g., between 10^{-3} and 10^{-2}; Supplementary Information, Fig. 2.11).

However, 56 days after the primary infection (24 days after the secondary infection), positively selected genes made up about 40% (95% CI 20-59%) of the germinal center repertoire of mice infected twice, 1.2-3.6 times more than the combined frequency of the same genes in the controls (Fig 2.2). The combined frequency of positively selected genes in germinal centers was similar between controls and infected mice from earlier time points (Fig 2.2). Although non-significant, a trend toward higher combined frequency of positively selected genes in germinal centers was observed in the late primary response (24 days after primary infection). It is possible that the effects of positive selection become more apparent over time, with B cell lineages that use positively selected genes strongly outcompeting other lineages in germinal centers late into the response. Germinal centers formed during the primary response might have been re-stimulated by the secondary infection, providing more time for fitness differences to result in a high frequency of positively selected genes.

2.3 Discussion

In how many ways can the immune system solve the problem of recognizing a particular pathogen by recombining immunoglobulin genes? Our results suggest solutions using particular V genes are more successful than others in mice infected with influenza. Despite likely epistatic interactions among V, D and J genes and between the heavy and light chains,

some V genes seem more likely than others to recombine into receptors that recognize influenza well. B cells whose receptor uses those genes are more likely to enter the experienced repertoire and subsequently expand into a clonal population, and those genes consequently increase in frequency between the naive and experienced repertoire of different infected mice. We used a simple mathematical framework to quantify this selection at the V-gene level during the immune response. Although we do not know which B cell receptors in the data bound influenza and which did not, the lack of evidence for selected genes in the uninfected controls suggests selection in the infected mice was driven by influenza.

While positively selected genes are more likely than others to recognize influenza antigens, the contribution of selected genes to the influenza response is unclear. B cells using non-selected genes may have bound influenza antigens and subsequently divided, thus potentially contributing to the response. Our analysis simply identified V genes that made B cells more or less successful at entering the experienced repertoire of infected mice and subsequently dividing than the average B cell. Power analysis suggests we may have missed genes under weakly positive or weakly negative selection, including genes in the controls. The differential abundance test we used is also not designed to detect selected genes present in only a few individuals, as might be the case for alleles present in only one or a few mice (individual mice had a median of one allele not shared with other mice, range 0-6).

Short-term selection at the level of V genes during the immune response might lead to selection of germline V gene sequences in the long term to improve the recognition of specific pathogens. For instance, it has been hypothesized that in addition to being selected during the short-term evolution of B cell lineages, mutations that improve recognition of common pathogens could be hard-coded into the germline sequences of immunoglobulin genes particularly adept at recognizing those pathogens [28]. Because influenza viruses do not naturally infect mice in the wild, our results suggest such variation in V genes' ability to recognize particular pathogens can arise by chance from the diversification of V genes in vertebrates.

Selection at the level of V genes during the immune response might also affect the predictability of the immune response. While the response to simple antigens is often dominated by a single V gene [32, 6, 177], complex antigens such as influenza's surface protein hemagglutinin (HA) can allow B cells using many different V genes to coexist in mouse germinal centers [92], potentially reflecting different epitope specificities among B cell receptors using different V genes. Different antigens and epitopes are often targeted in similar proportions by antibodies made by different individuals [9]. This convergence in phenotype might correspond to a convergence in V gene usage in the B cell repertoire due to the same genes being positively selected with similar strengths in different individuals. However, whether B cell receptors using a particular V gene consistently target the same antigen or epitope is unclear. In humans, monoclonal antibodies (mAbs) targeting the HA stalk appear to use V genes in different proportions from mAbs targeting the HA head [44]. One gene in particular, IGHV1-69, is disproportionately more common in broadly-reactive antibodies targeting the HA stalk [180, 103, 8], with a specific allele of IGHV1-69 leading to higher antibody titers against the stalk in some people [125]. While the V gene usage of antibodies targeting the stalk has been well characterized because the stalk is as a potential target for a universal influenza vaccine [80], it is unclear if other epitopes or antigens are preferentially bound by antibodies using particular V genes and if such patterns affect the breadth and potency of the response.

Studies looking for selection of particular V genes in response to influenza at the repertoire level without focusing on particular antigens or epitopes found limited similarity in V gene usage among people [30, 156]. In contrast, we identified genes that consistently increased in frequency between the naive and experienced repertoires of 72-100% of the infected mice. In addition to differences in statistical approach and power between our study and these previous studies, higher similarity in V gene usage across mice than across people might be due to variation across people in the epitopes and antigens targeted in response to influenza. For instance, differences in infection history might cause different people to target different

antigens or epitopes.

We propose that characterizing the specificity of influenza antibodies using different V genes and variation in V gene usage across cohorts might help guide efforts to induce antibodies against specific epitopes by vaccination. Our results suggest that given an identical infection history –or the lack of one– influenza infection will select similar V genes in different individuals. If differences in V gene usage lead to functionally distinct responses, differences in V gene usage associated with differences in infection history might help explain variation in cohorts' protection against influenza [27]. For instance, people with low pre-existing titers to pandemic H1N1 viruses are more likely to produce HA stalk antibodies using IGHV1-69 following vaccination with a pandemic H1N1 virus than people with high pre-existing titers [8], suggesting infection history affects the use of this particular V gene and the ability to develop broadly neutralizing antibodies. However, differences in V gene usage arising from differences in infection history have yet to be characterized more broadly. This knowledge might inform vaccine design efforts by identifying variation across people in their ability to develop antibodies against desired epitopes.

2.4 Materials and Methods

2.4.1 Infection experiments

We infected 40 8-week-old female C57BL/6 mice weighing 20-22g (8 for each time point) intranasally with 0.5 LD_{50} of a mouse-adapted pandemic H1N1 strain (A/Netherlands/602/2009) in a total of 30 μL of PBS under full anesthesia. Two controls for each time point were given PBS only. All mouse experiments were approved by The University of Chicago Institutional Animal Care and Use Committee (IACUC protocol 71981). We sorted naive (IgD+B220+), plasma (IgD-Sca-1hiCD138hi), memory (IgD-B220+CD95-CD38hi) and germinal center (IgD-B220+CD95+ CD38loGL-7+) cells after excluding cells expressing CD4, CD8a, TER-199 or F4/80 (to exclude T cells and dendritic cells). After spinning down

cells and removing the PBS supernatant, we stored pellets at -20°C because we originally intended to sequence genomic DNA and not cDNA.

2.4.2 B cell receptor sequencing

We generated immunoglobulin heavy chain (IGH) DNA libraries from complementary DNA generated from 10-500 ng of total RNA using Superscript III (Invitrogen) reverse transcriptase and random hexamer primers. For PCR amplifications, we used multiplexed primers targeting the mouse framework region 1 (FR1) of IGHV in combination with isotype-specific primers targeting constant region exon 1 of IgA, IgD, IgE, IgG, or IgM (Table 2.2). We performed separate PCR reactions for each isotype to avoid formation of inter-isotype chimeric products. We barcoded each sample with 8-mer primer-encoded sequences on both ends of the amplicons and performed PCR amplification in two steps. First, we generated amplicons using primers with the partial Illumina adapter, the sample-specific barcode and the locus-specific sequence. In the second step, we performed another PCR to complete the Illumina adapter sequence and to ensure final products were not amplified to saturation. We purified pooled products by agarose gel electrophoresis and extraction. We used a 600 cycle v3 kit to sequence products using an Illumina MiSeq instrument.

2.4.3 Estimating amplification and sequencing error

We estimated the rate at which errors were introduced during amplification and sequencing by comparing the sequenced reads with the reference sequence for the corresponding isotype. Because the constant region does not undergo somatic hypermutation, we counted each mismatch between the end of the J gene and the beginning of the conserved region primer as an error introduced by sequencing and amplification. Based on 187,500 errors found out of 104,092,368 bases analyzed, we estimated the error rate to be 1.80 mutations per thousand bases (95% binomial CI 1.79-1.81).

To remove sequence variation attributable to sequencing and amplification error, we built

Table 2.2: Primers for mouse heavy chain B cell receptors.

Primer name	**Sequence**
P7-VH1-MsFR1-A	CCTGGGGCTTCAGTGA
P7-VH1-MsFR1-B	GCCTGGGACTTCAGTGA
P7-VH1-MsFR1-C	CCTGGGGCCTCAGTGA
P7-VH1-MsFR1-D	GCCTGGGGCTTCAGTAA
P7-VH2-MsFR1	CCCTCACAGAGCCTGT
P7-VH3-MsFR1	CTTCAGGAGTCAGGACCT
P7-VH5-MsFR1-A	GTCCCTGAAACTCTCCTGTG
P7-VH5-MsFR1-B	GCCTGGAAGGTCCGT
P7-VH5-MsFR1-C	GTCCCTGAAACTCTCCTG
P7-VH7-MsFR1	TTCTCTGAGACTCTCCTGTG
P7-VH9-MsFR1	TGGAGAGACAGTCAAGATCTCC
P7-VH10-MsFR1	GATTGGTGCAGCCTAAAGG
P7-VH11-MsFR1	GCTTGGTGCAACCTGG
P7-VH12-MsFR1	TGCTGTCATCAAGCCATCA
P7-VH14-MsFR1	AGTCAAGTTGTCCTGCA
Ms-Tim-IgM	GGGAAGACATTTGGGAAGGAC
Ms-Tim-IgD	TGAGAGGAGGAACATGTCAG
Ms-Inner-IgG1	GCTCAGGGAAATAGCCCTTGAC
Ms-Inner-IgG2	GCTCAGGGAAATAACCCTTGAC
Ms-Inner-IgG2b	ACTCAGGGAAGTAGCCCTTGAC
Ms-Inner-IgG3	GCTCAGGGAAGTAGCCTTTGAC
Ms-Tim-IgA	GTCAGTGGGTAGATGGTGG
Ms-Tim-IgEc	CCAGGCAGCCCAGGGTCATGG

Table 2.3: PCR conditions.

1st PCR	
usual initial mix	
	per rxn
10x buffer	3 μL
MgCl2	1.8
2mM dNTP	3
v primer	3
c primer	3
template	4μL (200ng total)
Taq	0.3
H20	11.9
total	30

1st PCR cycle	
94°C	7 min
94°C	30s ($\times$ 35 cycles)
56°C	45s
72°C	1.5 min
72°C	10 min

2nd PCR	
MM	10
Q mix	4
f primer	0.4
r primer	0.4
template	0.5
H2O	4.7

2nd PCR cycle	
usual illumina cycling	
95°C	15 min
95°C	30s ($\times$ 12 cycles)
60°C	45s
72°C	1.5 min
72°C	10 min

separate phylogenetic trees for each cell type (naive, germinal center, memory, and plasma cells) in each clonal lineage identified by partis. We calculated the nucleotide Hamming distance between each pair of sister sequences in the tree and conservatively counted sister sequences from the same cell type and tissue as a single unique sequence if they differed by two nucleotides or fewer.

2.4.4 Details of the mathematical model

Let $f_i^N = N_i / \sum_j N_j$ be the frequency of gene i in the naive repertoire and $f_i^E = E_i / \sum_j E_j$ the frequency of gene i in the experienced repertoire. The latter is is given by:

$$f_i^E = \frac{S_i N_i \alpha\beta}{\sum_j S_j N_j \alpha\beta} = \frac{S_i N_i}{\sum_j S_j N_j} \tag{2.2}$$

and the ratio ρ_i of the experienced and naive frequencies of gene i is therefore given by:

$$\rho_i = \frac{f_i^E}{f_i^N} = \frac{S_i N_i}{\sum_j S_j N_j} \times \frac{\sum_j N_j}{N_i} = \\ = \frac{S_i}{\sum_j S_j f_j^N} = \frac{S_i}{\Lambda} \tag{2.3}$$

where $\Lambda = \sum_j S_j f_j^N$ is the expected selection factor across genes in the naive repertoire.

From Eq. 2.2, notice that the ratio of the experienced frequencies of two genes i and j is $(S_i N_i)/(S_j N_j)$. Thus, if neither i nor j are under selection ($S_i = S_j = 1$), the ratio of experienced frequencies is simply N_i/N_j, which is equal to the ratio of the naive frequencies of i and j. In contrast, if i is selected, the ratio between its experienced frequency and the experienced frequency of each non-selected gene j is either greater or smaller than ratio of naive frequencies, depending on the selection factor of gene i. The stability of frequency ratios of non-selected genes between the naive and experienced repertoires can be used to heuristically identify genes unlikely to be under selection, which can be used to test for selection in the remaining genes [20].

The relationship between the selection factor of gene i, S_i, and the classical selection coefficient expressed as a difference between per capita instantaneous growth rates depends on the precise dynamics of B cell populations. We illustrate this relationship by considering the simple case of B cell populations growing exponentially upon activation. Let r be the instantaneous per capita growth rate of B cells using a non-selected gene (analogous to a wild-type growth rate in classical evolutionary genetics), and let r_i be the instantaneous per capita growth rate of B cells using selected gene i. Note that the size of the B cell population at the beginning of the exponential growth is the number of activated B cells using gene i, given by the product $N_i \alpha S_i^\alpha$, where S_i^α is the activation selection factor for gene i.

Combining the equation $E_i(t) = N_i S_i \alpha \beta$ with the equation for exponential growth, $E_i(t) = N_i S_i^\alpha \alpha e^{r_i t}$, we have:

$$\frac{S_i}{S_i^\alpha} = \frac{e^{r_i t}}{\beta} \tag{2.4}$$

By definition, β is the factor by which an activated B cell population using a non-selected gene j will have increased at time t:

$$E_j(t) = N_j \alpha \beta = N_j \alpha e^{rt} \implies \beta = e^{rt} \tag{2.5}$$

Therefore:

$$\frac{S_i}{S_i^\alpha} = e^{(r_i - r)t} = e^{s_i t} \tag{2.6}$$

where s_i is the selection coefficient of gene i. Taking the natural logarithm on both sides and solving for s_i:

$$s_i = \frac{1}{t} \ln \frac{S_i}{S_i^\alpha} \tag{2.7}$$

Note that if the selected gene only increases the probability of a B cell becoming activated but has no effect on the subsequent growth rate ($S_i = S_i^\alpha$ and $S_i^\beta = 1$), its selection coefficient will be zero. Since $S_i = S_i^\alpha S_i^\beta$, we can also express s_i in terms of the expansion selection factor S_i^β

$$s_i = \frac{1}{t} \ln S_i^\beta \tag{2.8}$$

2.4.5 Differential abundance test and power analysis

We used R package `dacomp` to perform the differential abundance test proposed by Brill *et al.* [20]. The test uses a heuristic method for identifying taxa genes (taxa) whose frequency ratios are constant across samples from the groups being tested. Genes that are not differentially abundant across groups should have similar frequency ratios across samples and can be used as a reference set to test if the remaining genes are differentially abundant. For each pair of genes (j, k), the following statistic is computed to summarize the variation in the ratio of j and k frequencies:

$$SD_{j,k} = \mathrm{sd}_{i=1}^{N} \left[\log_{10} \left(\frac{Z_{i,j} + 1}{Z_{i,k} + 1} \right) \right] \tag{2.9}$$

where i indexes a sample, in our case an experienced or naive repertoire from a single mouse, $Z_{i,j}$ is the number of unique sequences using gene j in sample i, $\mathrm{sd}_{i=1}^{N}$ indicates the standard deviation across samples and N is the total number of samples (across naive and experienced repertoires). For each gene j, the median value of $SD_{j,k}$ across all other genes k, S_j, gives a measure of how much the frequency of gene j varies relative to the frequency of other genes. Genes with the lowest value of S_j (below a critical value S_{crit}) are chosen as the reference set. We heuristically set S_{crit} to the 15th percentile of the distribution of S_j.

Given the set of reference genes, the test for each remaining gene j consists of testing if the frequency of j relative to the combined frequency of j plus the set of reference genes is different between naive and experienced repertoires. First, we find the smallest combined count of j plus genes in the reference set across all samples, λ_j. For each naive repertoire i, a random variable $X_{i,j}$ is sampled from a binomial distribution with λ_j trials and probability of success equal to the frequency of j in sample i relative to the combined frequency of j and the reference set in sample i. A random variable $Y_{k,j}$ is similarly drawn for each experienced repertoire k. If gene j is not under selection, the distributions of $X_{i,j}$ and $Y_{k,j}$ should be identical, a null hypothesis that can be tested with a Wilcoxon rank sum test. We performed

separate tests for uninfected controls, mice in the primary infection group, and mice in the secondary infection group.

To estimate the power and false discovery rate of this test when applied to detect selected V genes in mouse B cell receptor data, we generated synthetic experienced repertoires in which 10 genes were randomly chosen to be under selection with the same selection factor (0.1, 0.2, 0.25, 0.5, 1, 2, 4, 5, 10). Because the test is not designed to detect selection for genes that are only present in one or a few mice, only genes shared by at least 5 mice could be randomly chosen to be under selection. We generated 100 synthetic datasets for each selection factor. We sampled each synthetic dataset by taking a multinomial draw for each experienced repertoire compartment in each mouse. The size of the draw was equal to the number of unique sequences in that compartment in the real data, and the probability of sampling each V gene was proportional to the product of the V gene's naive frequency and its selection factor. We kept the number of mice in each group (controls, primary infection and secondary infection) fixed in the synthetic datasets. For each synthetic dataset, we calculated the test's sensitivity (the proportion of selected genes that were detected) and the false discovery rate (the proportion of detected genes that were not in fact under selection).

2.5 Acknowledgements

This work was only possible due to the collaboration of Anna-Karin Palm, Chris Stamper, Micah Tepora and Patrick Wilson, who were responsible for the infection experiments and the RNA extraction, and Ji-Yeun Lee, Tho D. Pham, Khoa Nguyen and Scott Boyd, who were responsible for library preparation, amplification and sequencing. This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under grant DP2 AI117921 and CEIRS Contract No. HHSN272201400005C and by a Complex Systems Scholar Award from the James S. McDonnell Foundation awarded to Sarah Cobey. The content is solely the responsibility of the authors and does not necessarily represent

the official views of the NIAID or the National Institutes of Health. Marcos Vieira was also supported by a William Rainey Harper Dissertation Fellowship by the University of Chicago. This work was completed in part with resources provided by the University of Chicago Research Computing Center.

2.6 Supplementary information

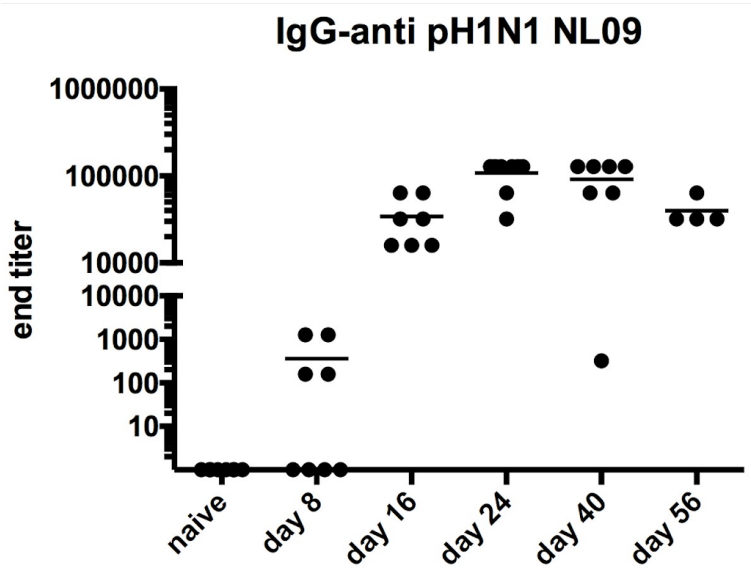

Figure 2.3: ELISA titers for IgG anti-hemagglutinin antibodies in mice infected with the mouse-adapted pandemic H1N1 strain A/Netherlands/602/2009.

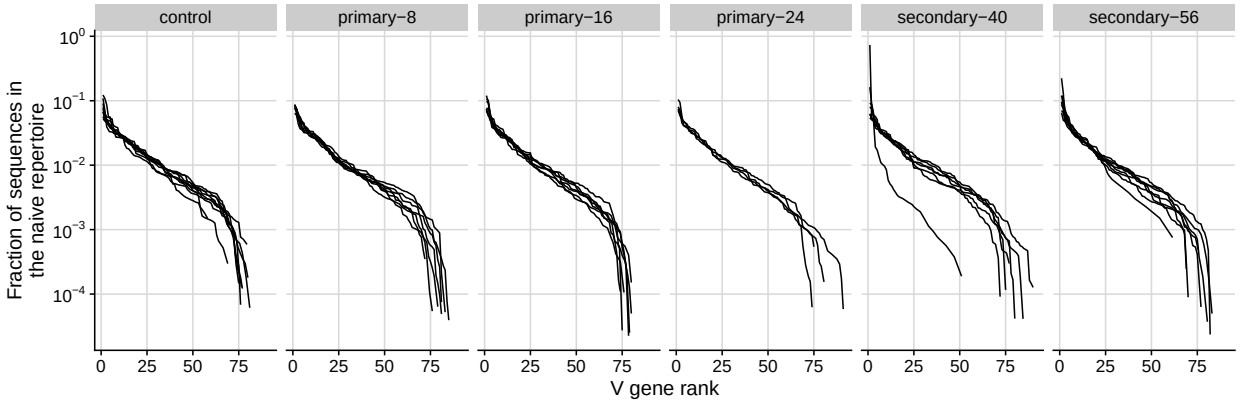

Figure 2.4: Rank-frequency plots of immunoglobulin heavy-chain V genes in the naive repertoire of C57BL/6 mice. Primary infection mice were infected with a mouse-adapted H1N1 influenza strain once and sacrificed 8, 16 or 24 days after the infection. Secondary infection mice were infected again 32 days after the primary infection and sacrificed 40 or 56 days after the initial infection.

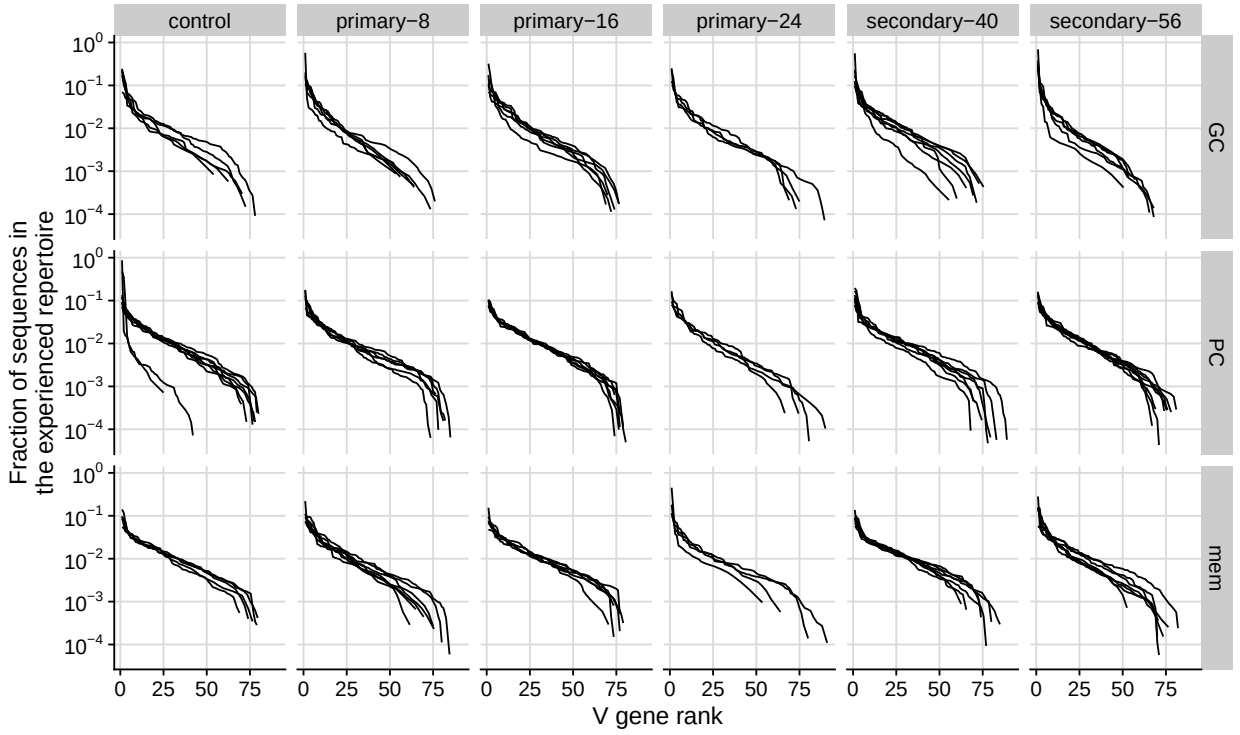

Figure 2.5: Rank-frequency plots of immunoglobulin heavy-chain V genes in the experienced repertoire of C57BL/6 mice. Primary infection mice were infected with a mouse-adapted H1N1 influenza strain once and sacrificed 8, 16 or 24 days after the infection. Secondary infection mice were infected again 32 days after the primary infection and sacrificed 40 or 56 days after the initial infection. GC: germinal center cells. PC: plasma cells. mem: memory cells.

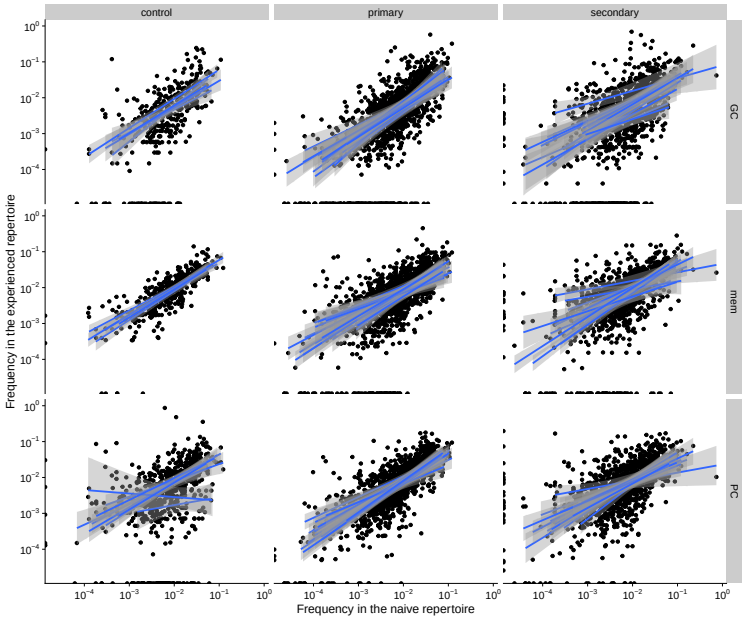

Figure 2.6: Correlation between naive and experienced V gene frequencies across mice infected once (primary) or twice (secondary) with a mouse-adapted H1N1 strain. Points for different mice in each group are shown together, while each line was fitted to the points of an individual mice. GC: germinal center cells. PC: plasma cells. mem: memory cells.

.

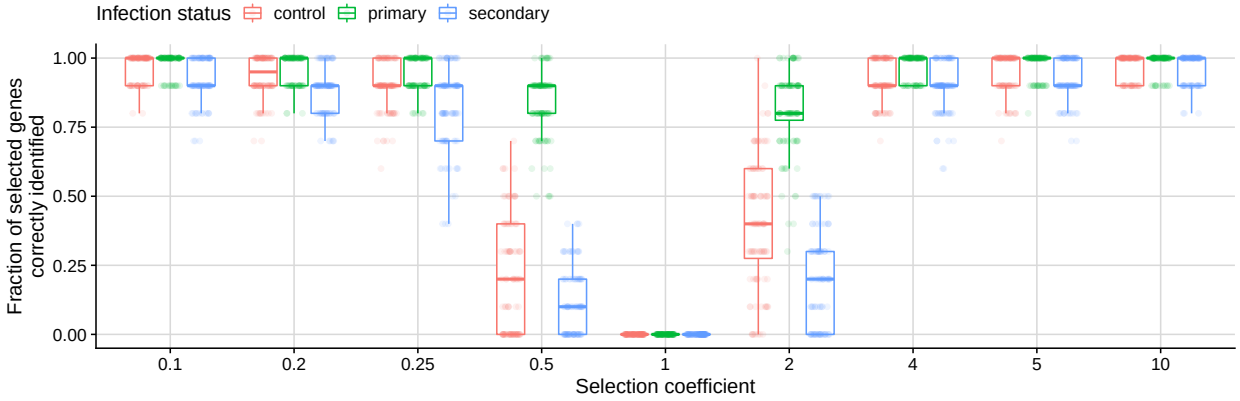

Figure 2.7: Sensitivity of the method developed by Brill et al. [20] applied to synthetic B cell receptor sequence data under different factors. For each synthetic dataset, we randomly chose ten genes to have the specified selection factor while setting all other genes to have a selection factor of one. We sampled the number of sequences using each gene in the experienced repertoire based on the genes' naive frequencies and selection factors. We generated 100 synthetic datasets for each selection factor.

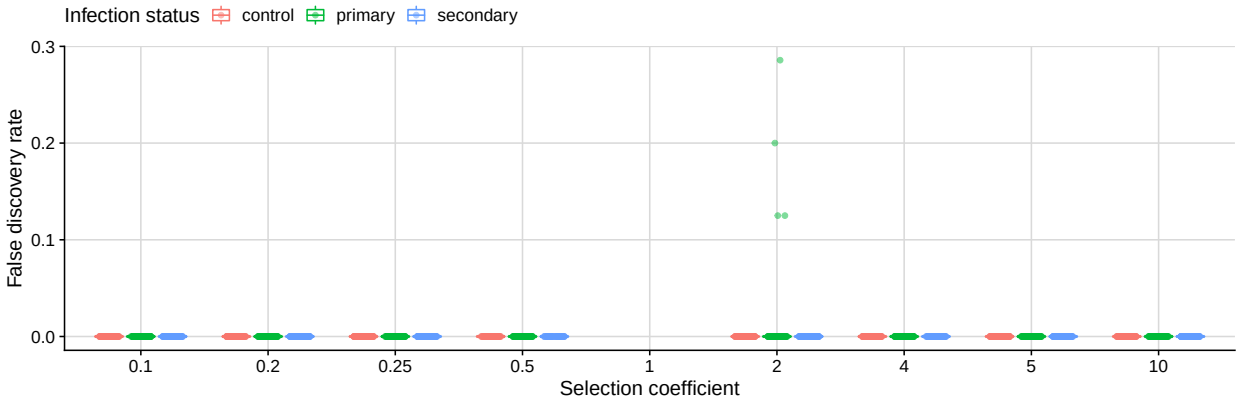

Figure 2.8: False discovery rate of the method developed by Brill *et al.* [20] applied to synthetic B cell receptor sequence data under different selection factors. For each synthetic dataset, we randomly chose ten genes to have the specified selection factor while setting all other genes to have a selection factor of one. We sampled the number of sequences using each gene in the experienced repertoire based on the genes' naive frequencies and selection factors. We generated 100 synthetic datasets for each selection factor. The false discovery rate is the proportion of genes identified by the test as differentially abundant between naive and experienced repertoires that were not in fact under selection.

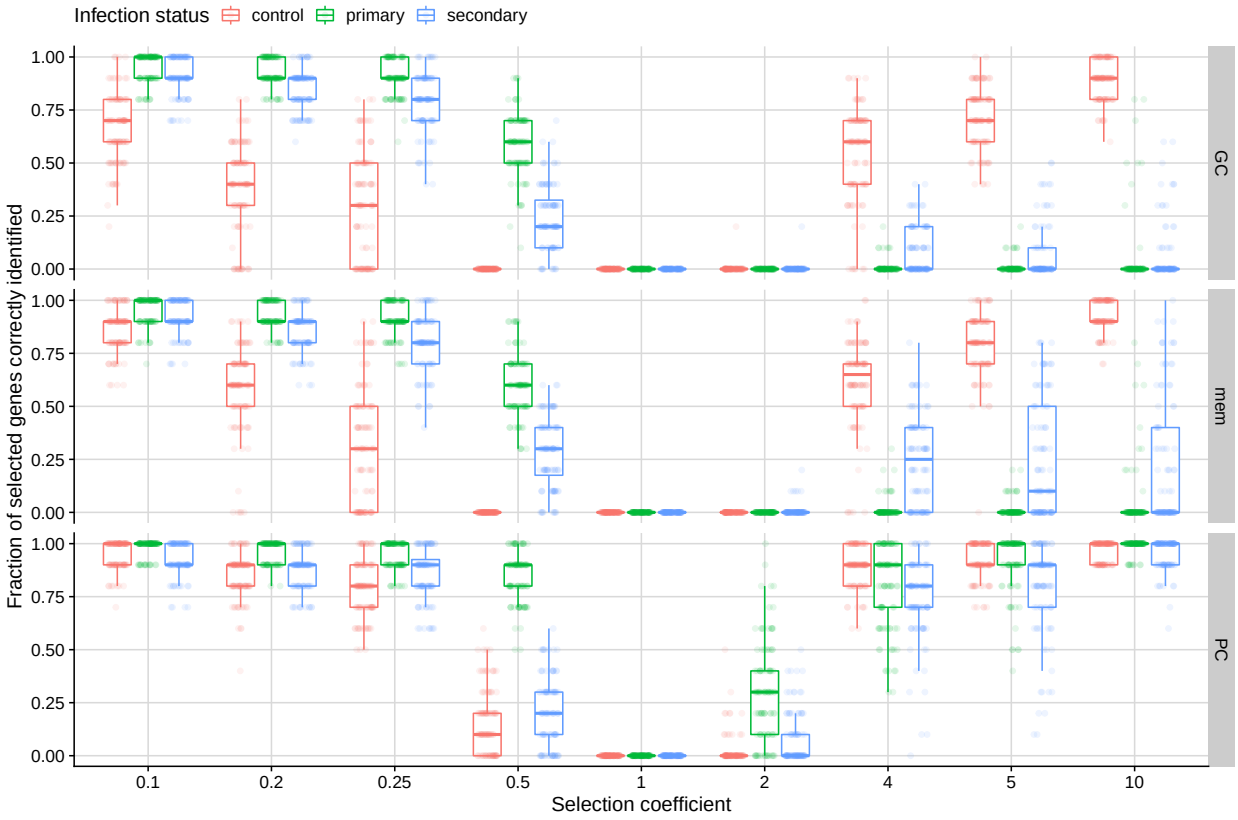

Figure 2.9: Sensitivity of the method developed by Brill *et al.* [20] applied to synthetic B cell receptor sequence data generated for different compartments of the experienced repertoire. For each synthetic dataset, we randomly chose ten genes to have the specified selection factor while setting all other genes to have a selection factor of one. Selected genes were the same across germinal center (GC), memory (mem) and plasma cells (PC), but a separate test was done for each compartment using each synthetic dataset. The number of sequences in each compartment and the number of mice in each group were identical to those in the observed data. We generated 100 synthetic datasets for each selection factor.

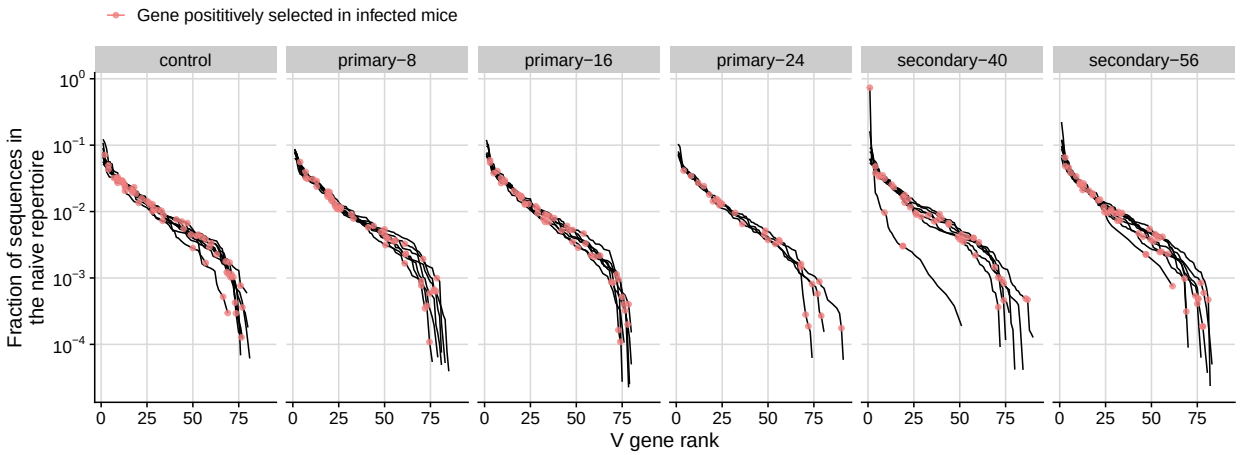

Figure 2.10: Rank-frequency plots of immunoglobulin heavy-chain V genes in the naive repertoire of C57BL/6 mice. Genes positively selected in mice infected with influenza are highlighted in red. Primary infection mice were infected with a mouse-adapted H1N1 influenza strain once and sacrificed 8, 16 or 24 days after the infection. Secondary infection mice were infected again 32 days after the primary infection and sacrificed 40 or 56 days after the initial infection.

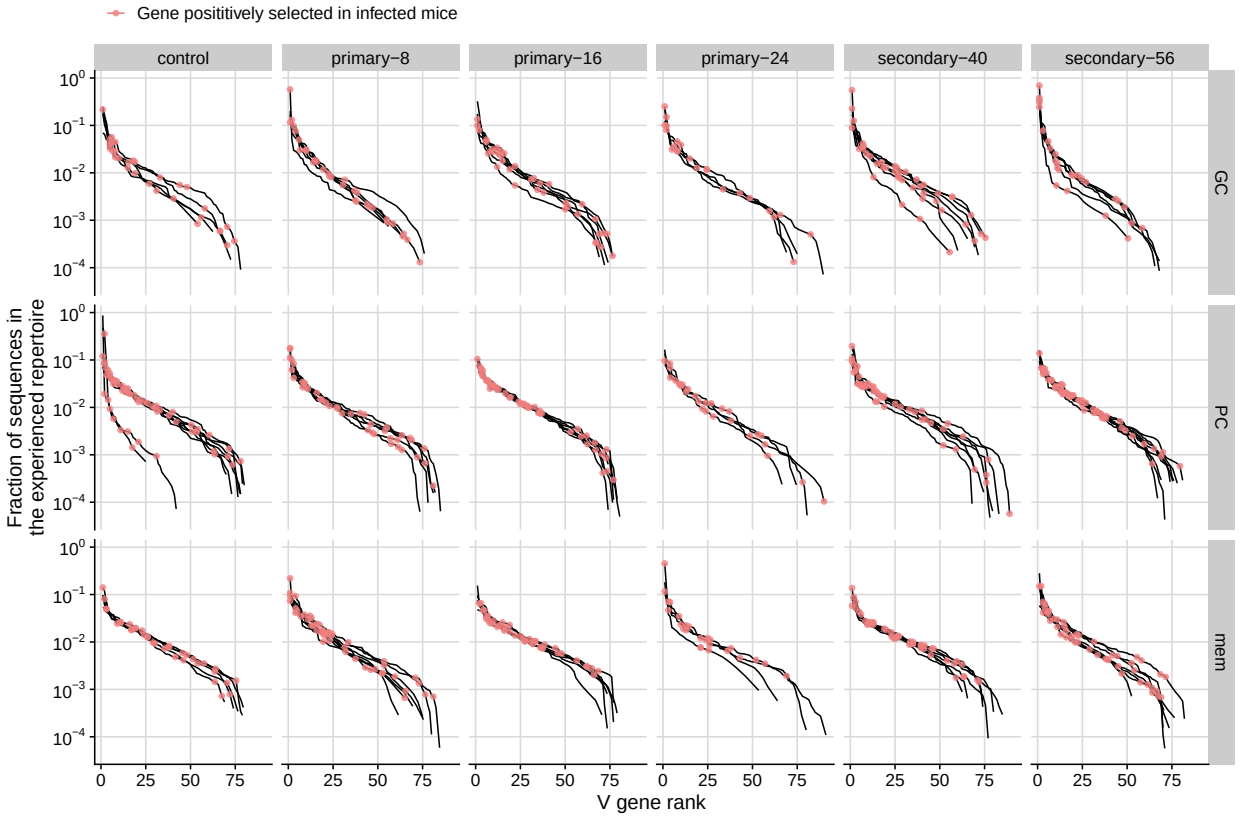

Figure 2.11: Rank-frequency plots of immunoglobulin heavy-chain V genes in the experienced repertoire of C57BL/6 mice. Genes positively selected in mice infected with influenza are highlighted in red. Primary infection mice were infected with a mouse-adapted H1N1 influenza strain once and sacrificed 8, 16 or 24 days after the infection. Secondary infection mice were infected again 32 days after the primary infection and sacrificed 40 or 56 days after the initial infection. GC: germinal center cells. PC: plasma cells. mem: memory cells.

CHAPTER 3

SELECTION AND NEUTRAL MUTATIONS DRIVE PERVASIVE MUTABILITY LOSSES IN LONG-LIVED ANTI-HIV B CELL LINEAGES

3.1 Introduction

High-affinity antibodies arise during the adaptive immune response from the very process that gave vertebrates an adaptive immune system in the first place: adaptation by natural selection. In response to infection or vaccination, mutagenic enzymes and error-prone polymerases cause somatic hypermutation of B cell receptors and thus create variation in their ability to bind antigen [164]. B cells with high-affinity receptors are more likely to receive survival and replication signals from helper T cells, and thus selection for improved antigen binding drives the development of high-affinity B cell receptors that are later secreted as antibodies [63, 109, 106, 168]. Understanding how the immune system evolved to facilitate the rapid adaptation of B cell receptors during infection may provide general insights into the evolution of adaptability.

Two features of B cell receptor genes suggest their long-term evolution has been shaped by selection for adaptability during infection. First, the germline genes that recombine to produce B cell receptors are enriched for nucleotide motifs that are targeted with high frequency by the mutagenic enzymes involved in somatic hypermutation [138, 137, 176]. High mutation rates provide the genetic variation required for B cell adaptation, and low B cell mutation rates have been linked to immunodeficiency disorders [102, 18] and to the decline in immune function with age [55]. Second, mutational "hotspots" occur where mutations are most likely to be beneficial. Hotspots are concentrated in loops of the B cell receptor protein that are directly involved in antigen recognition (complementarity determining regions, CDRs) [122]. In contrast, structurally important regions of the B cell receptor

(framework regions, FRs), which are usually less directly involved in antigen binding, are enriched with motifs that have low mutability [122]. In addition, mutations in the CDRs are more likely to be non-synonymous and involve amino acids with distinct biochemical properties [175, 81, 140, 67]. The differential mutability of CDRs and FRs appears to focus mutations to regions where they are likely to produce variation in antigen affinity without destabilizing the protein. The frequency and distribution of mutational hotspots in B cell receptor genes therefore seem to contribute to their adaptability during immune responses.

As B cells mutate during the immune response, however, changes in the frequency and distribution of mutational hotspots might affect the subsequent adaptability of B cell receptors. This change in adaptability may be especially important in B cell lineages that coevolve with pathogens like HIV and influenza. Experimental removal of hotspots decreases both the mutation rate of the altered site relative to others and the overall mutation rate of the sequence [176, 73]. Loss of highly mutable motifs has been hypothesized to occur during the immune response due to motifs' propensity to mutate [68, 155], and decreased mutation rates due to such "hotspot decay" might explain declines in the evolutionary rates of B cell lineages over several years of HIV infection [181, 145]. In addition, changes in the distribution of highly mutable motifs across FRs and CDRs might increase the frequency of deleterious mutations in the former and decrease the frequency of beneficial mutations in the latter.

Although hotspots have been shown to spontaneously decay through random mutations [68], a full picture of the factors influencing the evolution of B cell receptor mutability is missing. First, highly mutable motifs might be regained through mutation. Second, selection for affinity and protein stability might favor non-synonymous mutations that incidentally increase or decrease mutability. Finally, selection might act on somatic hypermutation rates themselves. Although selection in theory can favor low mutation rates due to the reduced frequency of deleterious mutations [83, 108], rapidly changing environments may indirectly select for a higher mutation rate through its association with beneficial mutations [97, 160, 58,

123]. Thus, although short-term differences in fitness among B cells arise from differences in the affinity of their receptors, B cells with more mutable CDRs might have a higher probability of producing high-affinity descendants able to keep up with evolving antigens in the long term. Indirect selection for mutability might therefore retain or increase the frequency of highly mutable motifs in CDRs, while mutability losses in FRs might be selected due to the reduced frequency of destabilizing mutations.

Understanding changes in mutability may reveal constraints on B cell adaptation, but the contributions of different mechanisms to changes in the frequency and distribution of mutational hotspots during B cell responses are largely unknown. We investigated the evolution of B cell receptor mutability by fitting phylogenetic models to sequences from long-lived anti-HIV B cell lineages. In characterizing mutability, we considered two sequence-based features that appear to have been strongly selected in the evolution of the adaptive immune system: overall mutability (the density of mutagenic nucleotide motifs) and also changes in the mutability of CDRs relative to FRs. First, we examined B cell mutability in the unmutated common ancestors of anti-HIV antibodies. Next, we investigated the effects of random mutations and positive selection for amino acid substitutions that increase affinity for antigen. Finally, we tested for selection to increase, retain or decrease the frequency of highly mutable motifs.

3.2 Results

3.2.1 Ancestral B cells have higher mutability in CDRs than FRs

To characterize changes in mutability during B cell evolution, we inferred the evolutionary histories of previously reported B cell lineages from three HIV-1 patients. The CH103 and VRC26 lineages comprise heavy and light chain B cell receptor sequences obtained from high-throughput sequencing over 144 and 206 weeks of infection in two patients, respectively [104, 39]. We also analyzed heavy chain sequences of three lineages from the VRC01 dataset,

which was sampled from a third patient over a 15-year period [181]. The lineages we analyzed were originally investigated for having evolved the ability to neutralize diverse HIV strains.

To infer changes in mutability over time, we used Bayesian phylogenetic analyses (Materials and Methods) to obtain a sample of time-resolved trees from the posterior distribution of each lineage's genealogy for the heavy and light chains separately, and to estimate the nucleotide sequences of all internal nodes. Mutabilities of the observed and the inferred internal sequences were estimated using the S5F model. This model assigns relative mutation rates to all five-nucleotide DNA motifs and is based on a large independent dataset of antigen-experienced B cells [185]. The mutability of each sequence was defined as the geometric mean of the S5F scores across all sites in the B cell receptor sequence. We estimated the number, magnitude and distribution of mutability changes on all branches by computing the difference in average log-S5F-score for all pairs of parent-descendant nodes. Fig. 3.1 illustrates mutability evolution in the heavy chain of lineage CH103.

To investigate the potential for mutability changes during the response, we first characterized the mutability of each lineage's ancestor. In the ancestors of all heavy and light chains, sites in CDRs had higher average mutability than sites in FRs. On average across the seven heavy and light chain ancestors, mutability was approximately 35% higher in CDRs than in FRs (range: 26-54%; Supplementary Information, Fig. 3.6). Previous analyses of germline V genes (which, together with D and J genes, recombine to produce mature B cell receptors) showed that mutability is lower in FRs and higher in CDRs than expected based on their amino acid sequences, suggesting selection to increase the frequency of highly mutable motifs in CDRs and decrease their frequency in FRs [122, 140]. Consistent with those analyses, we found that the FRs of lineage ancestors had lower S5F mutability than expected based on their amino acid sequences. On average across all heavy and light chains, the mean FR mutability of ancestral B cell receptors was lower than 99% of sequences obtained by randomizing their codons (according to usage frequencies in humans [119]) while keeping the amino acid sequences constant (range: 96-100%; Fig. 3.6). However, while previous studies

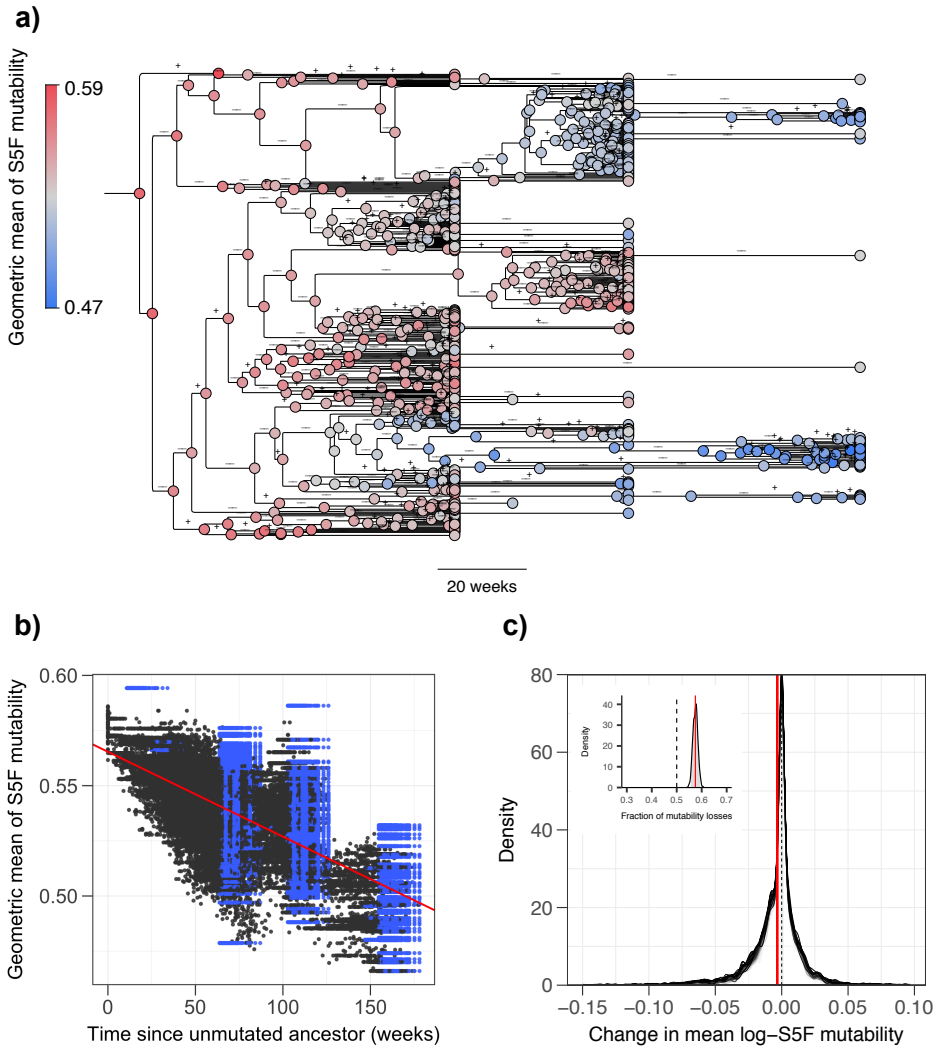

Figure 3.1: Evolution of S5F-mutability in the heavy-chain CH103 B cell lineage. **a**) Long-term declines in mutability across the maximum clade-credibility tree. Nodes are colored according to the geometric mean of the S5F mutability scores across the their sequences. Branches measured in weeks are annotated to indicate gains (+) or losses (–) of mutability. **b**) Mutability over time for a combined sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to inferred internal nodes. The red line represents an average of regression lines calculated for each tree in a sample of 1000 trees. **c**) For each tree in the posterior distribution of trees obtained for each lineage, we computed the magnitude of mutability changes on all branches, producing one distribution per tree. The distributions for a sample of 100 trees are shown in the main plot as overlaid densities. We also computed, for each tree, the fraction of changes in that tree that were losses, producing one value per tree. A distribution of such values for the full sample of 1000 trees is shown in the inset plot, with the 95% highest-posterior density interval shown in gray. Red lines indicate the means of the distributions.

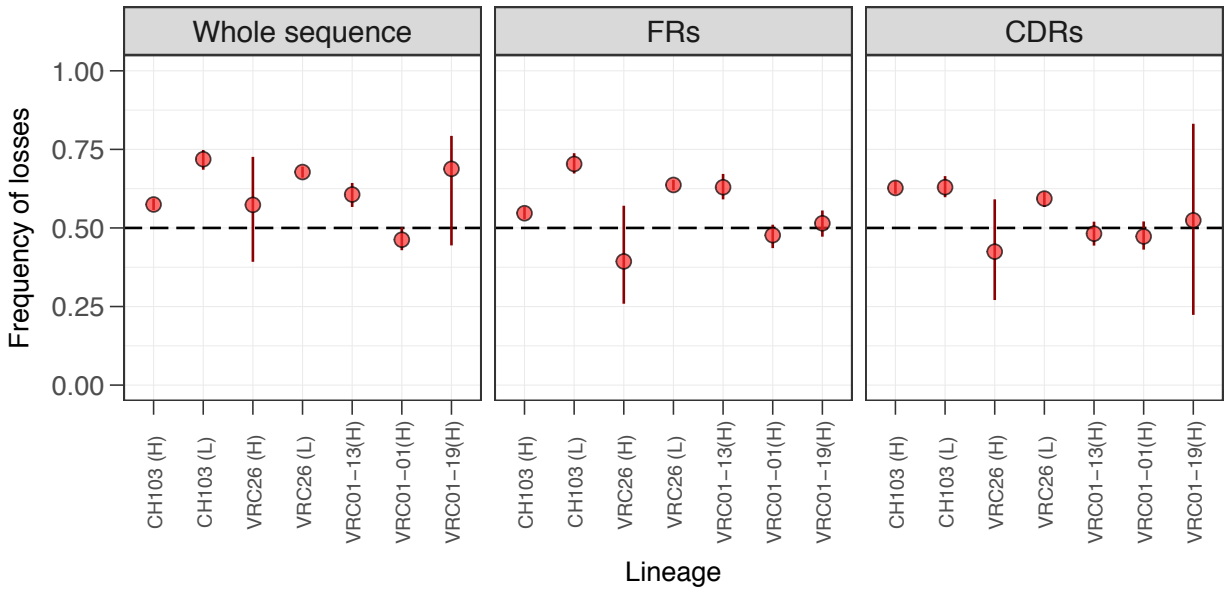

Figure 3.2: Frequency of losses relative to the total number of changes in mean log-S5F mutability during the evolution of anti-HIV B cell lineages. Results are shown for the entire analyzed region of the B cell receptor, and separately for framework regions (FRs) and complementarity determining regions (CDRs). Each point denotes the fraction of changes in mean log-S5F mutability that were losses, averaged across a sample of 1000 trees from the posterior distribution. Vertical red lines indicate the 95% highest-posterior density interval.

found that the CDRs of V genes are typically more mutable than expected based on their amino acid sequences [122, 140], we found the S5F mutability of CDRs in ancestral B cell receptors to be, on average, greater than only 60% of randomized sequences with the same amino acid sequence (range: 22-92%, Supplementary Information, Fig. 3.6). This difference in the results for CDRs may be due to the different mutability metrics used in those studies. Both studies defined mutability as the propensity for non-synonymous mutations. In contrast, we used the S5F model to quantify the propensity for all mutations. These results suggest that loss of S5F mutability in FRs is only possible if those regions undergo amino acid changes.

3.2.2 Mutability is more often lost than gained

Next we investigated long-term changes in mutability as B cell receptors evolved from their ancestors. Consistent with previous analyses [145, 68], average mutability decreased with time in four of the seven heavy and light chains (Fig. 3.1a,b; Supplementary Information, Figure 3.7). To investigate the factors contributing to those net long-term trends, we analyzed the frequency of mutability gains and losses across branches. Mutability losses should arise from hotspot decay, positive selection of the amino acid changes that incidentally decrease mutability, selection for lower mutability due to the reduction in the frequency of deleterious mutations, or a combination of those factors. Mutability gains should reflect spontaneous hotspot gains through mutation, positive selection of amino acid changes that incidentally increase mutability, indirect selection for higher mutation rates by association with beneficial mutations, or a combination of those factors. To summarize the net contributions of mutability-decreasing and mutability-increasing mechanisms, we computed the fraction of branches with mutability losses out of all branches with mutability changes. By computing the frequency of mutability losses for each tree in the posterior sample, we estimated the posterior distribution (Fig. 3.1c).

Across the entire BCR sequence, mutability losses occurred more frequently than gains in four of the seven heavy and light chains (Fig. 3.1a,c, Fig. 3.2). On average across the seven datasets, approximately 61% of changes in mutability were losses (range: 46-72%). The four chains where mutability losses were significantly more frequent than gains include three chains where mutability decreased in the long term (heavy and light chains of CH103 and light chain of VRC26; Supplementary Information, Figure 3.7) and one chain where mutability increased in the long term despite the high frequency of mutability losses (VRC01-13). This last result illustrates how mutability can increase in the long term despite the propensity for mutability to be lost on any given branch, potentially reflecting the effects of genetic drift, selection for mutations that incidentally increase mutability, or selection for mutability itself.

The frequency of mutability losses was not significantly different from the frequency of gains in the heavy chains of VRC26 (95% highest posterior interval 39-73%) and VRC01-19 (44-79%) and was slightly lower than the frequency of gains in lineage VRC01-01 (43-50%).

Mutability changes in FRs and CDRs were consistent with the changes observed across the entire B cell receptor sequence. Long-term declines in mutability occurred in the FRs of four of the seven heavy and light chains (Supplementary Information, Figure 3.8) and in the CDRs of five of the seven heavy and light chains (Supplementary Information, Figure 3.9), and the difference between CDR and FR mutabilities changed little (Supplementary Information, Fig. 3.10). Consistent with the net long-term trends, both CDRs and FRs had similar frequencies of mutability losses (FR average 56%, range: 39-70%; CDR average 54%, range: 42-63%; Fig. 3.2). Despite considerable amino acid divergence from the unmutated ancestors in FRs (16-67% for different heavy and light chains), sequences from the last sampling time in each dataset had lower FR mutability than 94% of randomizations with the same amino acid sequence (range: 61-99.9%; Supplementary Information, Figure 3.11). The lower mutability of ancestral FRs relative to their amino acid sequences was therefore retained throughout the evolution of the B cell lineages. CDRs, however, became less mutable relative to their amino acid sequences than the ancestors. On average across the seven heavy and light chains, CDRs of sequences from the last sampling times were more mutable than 35% of their corresponding randomizations (range: 7-64%; Supplementary Information, Figure 3.11), down from 60% in ancestral CDRs.

3.2.3 Hotspot decay and selection for amino acid substitutions contribute to mutability losses

Highly mutable motifs have been hypothesized to decay due to their propensity to mutate [68, 155], but motifs should also be influenced by positive selection. The first case is straightforward. For instance, the hypothetical sequence CAG<u>C</u>TT contains the highly mutable cytosine at the center of the AGCTT motif [137, 185, 176], and a C→T mutation

in the underlined position would decrease the mutability of the site approximately 5-fold [185]. Positive selection on amino acid substitutions should influence motifs in two ways. Mutational hotspots can be disrupted if selection favors amino acid sequences whose codons happen to be, on average, less mutable than codons in the ancestral sequence. Selection can also affect mutability in neighboring codons. For example, if selection led to the replacement of CAG (glutamine) for CGG (arginine) in the sequence CAG<u>C</u>TT, the mutability of the underlined C nucleotide (not involved in the substitution) would decrease approximately 13-fold [185].

To test if the observed losses of S5F mutability were consistent with the decay of highly mutable motifs expected under the S5F model, we simulated B cell receptor evolution under a model that allows for variation in mutation rates across motifs based on their S5F mutability scores [185] (Materials and Methods). We compared changes in mutability simulated under the S5F-based model to changes under models that do not allow for motif-driven mutation rate variation. Instead, these models assume that the mutation rate is identical across all sites ("uniform" model) or depends on the position within a codon ("codon-position" model), with the relative rate for each position estimated from the data (Materials and Methods). For each branch on the MCC tree of each lineage, we started from the inferred nucleotide sequence of the branch's parent node and simulated 100 replicates of a descendant sequence. We constrained the simulations to produce the same number of synonymous and non-synonymous changes as inferred from the data but allowed them to occur at different positions. Thus, the overall amount of selection is expected to be the same in the data and in the simulations, and the models differ in the locations of synonymous and non-synonymous mutations. We then partitioned observed and simulated changes in mean log-S5F mutability into changes caused by synonymous mutations and changes caused by non-synonymous mutations.

Observed synonymous changes in mutability were consistent with the decay of mutational hotspots simulated under the S5F model (Fig. 3.3a). In contrast, synonymous changes

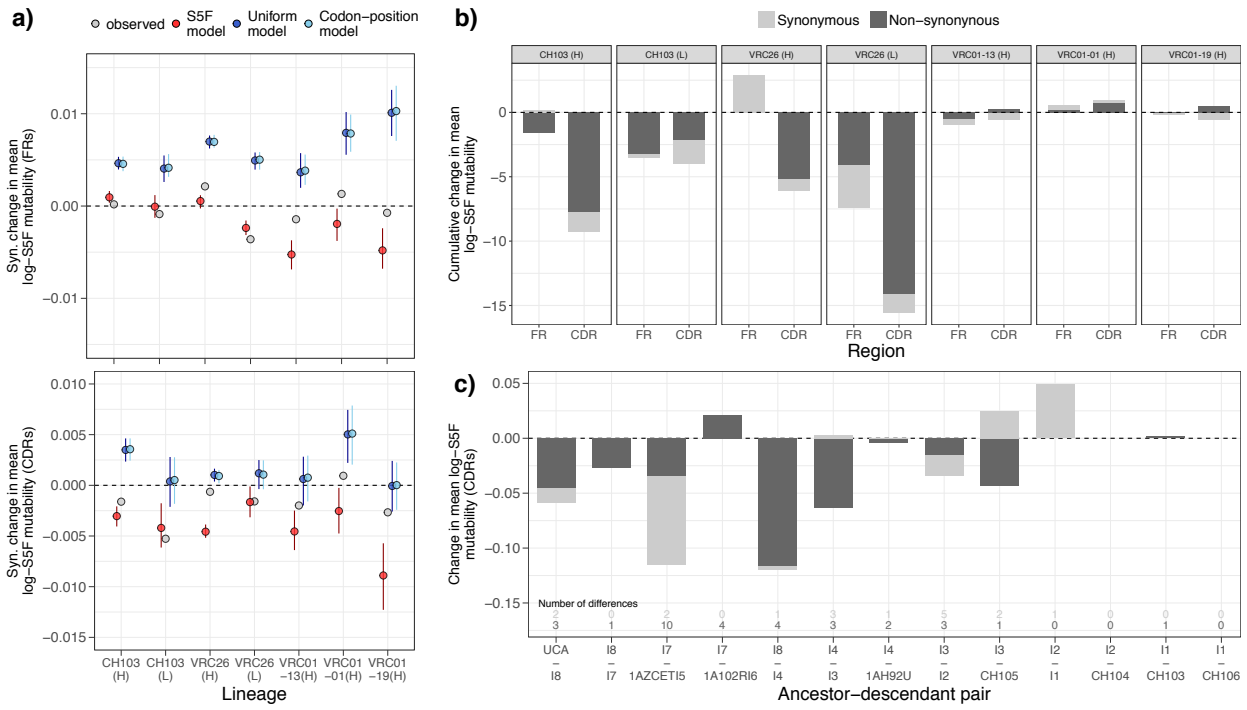

Figure 3.3: Changes in mean log-S5F mutability due to synonymous and non-synonymous changes in anti-HIV B cell lineages. **a**) Changes in mutability due to synonymous changes, averaged across all branches. Gray points indicate values inferred from the data, and colored points indicate values obtained by simulation under different models. Red indicates an S5F-based model where different nucleotide motifs mutate with different rates, dark blue indicates a model with no mutation rate variation across sites, and light blue indicates a model with different mutation rates for each position of a codon. Simulations were performed independently for each branch on the MCC tree of different anti-HIV B cell lineages, starting from the inferred sequence of the parent node. Each simulated sequence was constrained to have the same number of non-synonymous and synonymous changes as observed in the branch. Vertical bars indicate the 95% range obtained from 100 simulations per model. **b**) Changes in mean log-S5F mutability due to synonymous and non-synonymous changes summed across all branches of the B cell genealogies. **c**) Changes in mean-log S5F mutability in the CDRs of 13 ancestor-descendant B cell receptor sequence pairs where binding affinity to HIV-1 increased [104].

simulated with constant mutation rates across motifs (uniform and codon-position models) increased mutability in all lineages. Those results suggest that hotspot decay explains most of the mutability loss caused by neutral mutations, and that the S5F model, in particular, accurately describes neutral sequence evolution in B cell lineages.

However, the average non-synonymous loss simulated under the S5F model was greater than the observed loss in 13 of 14 regions/lineages (except in the CDRs of light-chain VRC26; Supplementary Information, Figure 3.12). While those differences are within the 95% range of values simulated under the S5F model, the consistent overestimation of non-synonymous mutability losses by the S5F model suggests the distribution of non-synonymous substitutions is affected by mechanisms other than across-motif variation in mutation rates, such as positive or purifying selection on specific amino acid mutations. Purifying selection may counter mutability losses if it eliminates non-synonymous changes that happen to decrease mutability. As previously reported by [145], we found that several non-synonymous changes occurred with high probability under the S5F mutability model but were scarce in the MCC trees inferred from the data (Supplementary Information, Figure 3.13), which might indicate they were under purifying selection during the evolution of the B cell lineages. Mutability losses caused by the ten most common transitions simulated under the S5F model were on average 75% greater than the losses caused by the ten most common transitions observed in the data (which in fact caused mutability gains on average in two datasets). Purifying selection against some of the most likely amino acid transitions under the S5F model would therefore contribute to smaller non-synonymous mutability losses than expected.

To evaluate the strength of positive selection in the B cell lineages, we used BASELINe [184] to quantify deviations in the frequency of non-synonymous substitutions from its expected value in the absence of selection (and under the mutational biases captured by the S5F model). In line with previous analyses of the same datasets [145] and of B cell lineages from healthy donors [183], we detected an enrichment of non-synonymous changes in the CDRs of four of the seven heavy and light chains (lineages CH103 and VRC26) and a lower frequency

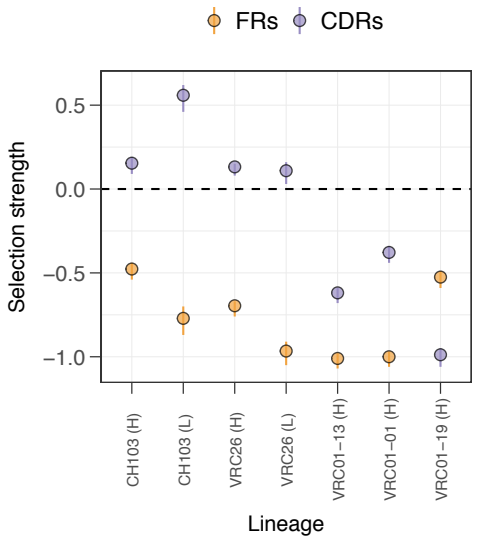

Figure 3.4: Selection in framework regions (FRs) and complementarity determining regions (CDRs) of B cell receptors from anti-HIV B cell lineages. Selection strength is measured as the log odds ratio between the observed ratio of non-synonymous to synonymous substitution frequencies, $\pi/(1-\pi)$, and the ratio expected under the S5F mutability model in the absence of selection, $\hat{\pi}/(1-\hat{\pi})$ [184]. Selection strength values greater than zero indicate positive selection, and values smaller than zero indicate purifying selection.

of non-synonymous mutations than expected in the FRs of all heavy and light chains (Fig. 3.4). This result contrasts with whole-repertoire analyses showing predominantly purifying selection across both types of regions [113], and suggests positive selection predominates in the CDRs. However, interpretation of such dN/dS ratios as selection strengths is complicated by the fact that most non-synonymous changes observed in the B cell sequences are polymorphisms, not fixations.

The enrichment of non-synonymous mutations in CDRs suggests that positive selection is concentrated in those regions. To estimate the potential contribution of positive selection to the mutability loss in the CDRs, we summed non-synonymous changes in the mean log-S5F mutability of CDRs across all branches of the MCC trees. Non-synonymous changes caused a net loss of CDR mutability in the same four heavy and light chains where an enrichment of non-synonymous changes was detected (Fig. 3.3b). Similarly, synonymous substitutions caused a net loss of CDR mutability in the same lineages. On average across those four

lineages, non-synonymous substitutions accounted for approximately 79% of the inferred mutability loss in CDRs (range: 54-91%). On average, selection for amino acid substitutions might therefore contribute to as much as 79% of the loss of mutability in the CDRs (in the extreme case where all non-synonymous changes in the CDRs were under positive selection).

To investigate changes in mutability due to non-synonymous changes for which there is evidence of positive selection, we analyzed 13 pairs of ancestor-descendant sequences experimentally characterized by [104]. In all of those pairs, binding affinity increased to at least one of the two tested HIV-1 envelope proteins. In ten of the pairs, the increase in affinity was associated with 1–10 non-synonymous changes in the CDRs. At least some of those non-synonymous changes were thus likely under positive selection for their effect on affinity. Our analysis showed that, in 8 of the 10 pairs where CDRs underwent non-synonymous changes, those non-synonymous changes caused CDR mutability to decrease (Fig. 3.3c). Positive selection of amino acid mutations that incidentally disrupted highly mutable motifs therefore likely contributed to the observed mutability loss in the CDRs.

3.2.4 No evidence of selection on mutability itself

Under persistent or recurrent selection, alleles that increase the mutation rate may increase in frequency by hitchhiking with beneficial mutations, which themselves arise more frequently due to the increased mutation rate [97, 160, 58, 123]. Such indirect selection for increased mutation rates might therefore lead to the conservation of highly mutable motifs in CDRs, where mutations that improve affinity are selected for during B cell evolution. In contrast, selection to reduce the frequency of deleterious mutations [83, 108] might directly favor mutability losses in FRs.

To test if mutability is subject to direct or indirect selection over the long term, we compared the frequency of synonymous changes in mutability between terminal and internal branches of the B cell trees. Terminal branches are expected to be enriched for deleterious mutations [179, 86, 45, 98]. If mutability losses are deleterious in the long term, branches

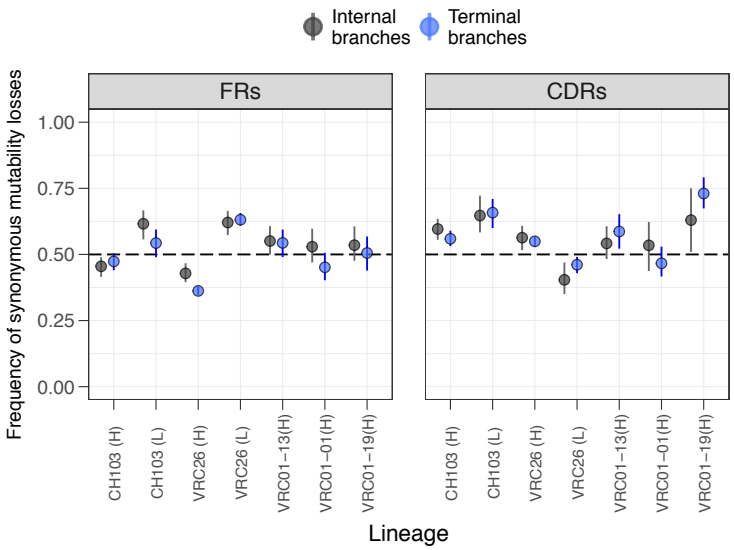

Figure 3.5: Frequency of losses relative to the total number of changes in mean log-S5F mutability caused by synonymous substitutions during the evolution of anti-HIV B cell lineages. Blue indicates changes on terminal branches, and black indicates changes on internal branches. Results are shown separately for framework regions (FRs) and complementarity determining regions (CDRs). Each point denotes the frequency of changes in mutability that were losses, averaged across a sample of 1000 trees from the posterior distribution. Vertical lines indicate the 95% highest-posterior density interval.

where mutability losses occurred should be less likely to contribute descendants to the B cell population, and would therefore be more likely terminal, rather than internal branches. In contrast, mutability losses should be more frequent on internal branches if mutability losses are beneficial. Because changes in mutability due to non-synonymous substitutions may be driven by selection for B cell receptor affinity and stability, we first restricted the analysis to mutability changes from synonymous substitutions.

We found no consistent evidence of selection to increase, retain, or decrease mutability (Fig. 3.5). In one instance (the CDRs of light chain VRC26), mutability losses were more frequent on terminal branches than on internal branches, possibly indicating selection against mutability losses. In another instance (FRs of heavy chains VRC26), mutability losses were more frequent on internal branches than on terminal branches, suggesting selection for mutability losses. In the remaining cases, the frequencies of synonymous mutability losses were similar in terminal and internal branches for both FRs and CDRs. No general trends were seen when we compared terminal branches, internal branches that belong to the main "backbone" of the tree, and other internal branches [98] (Supplementary Information, Figure 3.14), or when we included mutability changes due to non-synonymous substitutions (results not shown).

3.2.5 Results are consistent for three of four mutability metrics

In addition to the S5F model [185], we repeated the analyses using three other mutability metrics. The different mutability metrics were estimated from databases of somatic mutations not subject to the effect of selection, such as synonymous mutations, mutations in non-coding flanking regions of V genes, and mutations in unproductive B cell receptor genes that are not expressed by B cells but still undergo somatic hypermutation. Two of the alternative metrics are based on a discrete classification of DNA motifs into either hotspots or regular motifs: the number of WRCH and DGYW hotspots [137] or the number of "overlapping" (WGCW) hotspots [176], where $W = \{A/T\}$, $R = \{A/G\}$, $H = \{A/T/C\}$, $D = \{A/T/G\}$

and $Y = \{C/T\}$. We also quantified mutability using the 7-mer model, which, similar to the S5F model, assigns relative mutation rates to different motifs on a continuous scale, but does so for seven- instead of five-nucleotide motifs [47].

Consistent with the results for S5F mutability, hotspots and overlapping hotspots decreased in number over time (Supplementary Information, Figs. 3.15 and 3.16) and were lost more frequently than gained (Supplementary Information, Fig. 3.18). In contrast, average 7-mer mutability increased over time in four of five lineages, and losses of 7-mer mutability were approximately as frequent as gains across the entire B cell receptor sequence (Supplementary Information, Figs. 3.17 and 3.18).

3.3 Discussion

We used phylogenetic methods to reconstruct changes in B cell receptor sequences and found consistent loss of mutational hotspots over the course of the adaptive immune response. Our analyses shed light on previous studies of long-term trends in mutability [145, 68] by quantifying the contributions of different mechanisms to these losses. Selection for amino acid substitutions appears to have contributed to the mutability loss in the CDRs –precisely the regions where high mutation rates contribute the most to adaptation. However, mutability changes caused by synonymous substitutions throughout the sequence demonstrate mutability loss from the spontaneous decay of highly mutable motifs through neutral mutations. Disruption of highly mutable motifs through selection and hotspot decay therefore seems to be intrinsic to the evolution of B cells during affinity maturation. These factors counteract the high mutation rate selected in the evolution of immunoglobulin genes, and they suggest adaptability may inevitably become compromised in evolving B cell lineages.

Observed mutability losses due to non-synonymous changes were on average smaller than expected under the S5F model. Purifying selection against amino acid changes that occur with high probability under the S5F model may have contributed to this result. For instance, we found tyrosine-to-cysteine to be one one of the most common amino acid mutations under

the S5F model, and that mutation typically resulted in a loss of mutability in simulations (Supplementary Information, Figure 3.13). Yet Y-C mutations occurred with much lower frequency in the data, suggesting that purifying selection against Y-C mutations reduced the non-synonymous mutability loss relative to the S5F expectation.

While mutability losses occurred both in FRs and CDRs, the higher mutability of CDRs relative to FRs observed in ancestral receptors persisted over at least several years of B cell evolution, suggesting that some degree of adaptability may still be maintained. This observation may reflect survival bias: lineages in which mutability differences between CDRs and FRs decrease over time may be less likely to persist in the long term.

Theoretical simulations [160] and experimental evolution of bacteria [58, 123] suggest recurrent adaptation in asexual populations can select for increased mutation rates over the long term, but we found no consistent evidence of selection to increase or retain mutability during B cell evolution. Any long-term benefits of high mutation rates (in terms of increased chance of producing beneficial mutations) appear insufficient to overcome the rapid decay of highly mutable motifs through mutation and positive selection on amino acid substitutions. We also did not detect selection for reduced mutability in FRs, despite the potential for mutability losses to decrease the rate of destabilizing mutations. Because mutability is already low in the FRs of V genes [122, 140] and ancestral B cell receptors (this study), it is possible that the fitness effect of mutability losses in FRs is so small that such losses are nearly neutral given the effective size of the B cell population. Under this "drift-barrier" hypothesis [108], selection for decreased mutability would therefore not be able to fix mutability losses.

Losses of mutational hotspots in CDRs may reduce the adaptability of experienced B cells since lower mutation rates reduce the supply of genetic variation available for selection. Declines in mutation rates caused by hotspot losses may have contributed to reported declines in the evolutionary rates of broadly neutralizing B cell lineages during chronic HIV infection [145]. However, these losses may not be important for lineages that bind to conserved sites. As B cell receptors evolve high affinity for conserved sites, beneficial substitutions

become rare, and the corresponding transition from positive to purifying selection should cause substitution rates to fall [145]. In line with the previous analysis by [145], we found the oldest lineages to be under the strongest purifying selection in the CDRs. Those lineages also had minimal cumulative changes in mutability over the sampling period, suggesting most of the mutability loss occurred during early adaptation. However, as a result of hotspot loss, these lineages might adapt poorly if formerly conserved antigenic sites suddenly acquire mutations.

A direct link between motif-based mutability scores and rates of B cell evolution is still lacking. Hotspot loss is associated with reduced substitution rates [181, 145], which are influenced not only by the underlying mutation rate but also by changes in selective pressures and generation times [145]. Using robust counting [120] and a random local clock model [41], we found no evidence for consistent declines in the synonymous or non-synonymous substitution rates of these lineages (Supplementary Information, Figs. 3.19-3.22). Simulation-based power analysis (Supplementary Information) revealed that synonymous and non-synonymous substitution rates had to decline at least 50% to be detected with this method (Figs. 3.23-3.25). Observed long-term declines in S5F mutability were less than 50% (Supplementary Information, Figure 3.7) and suggest their effects on substitution rates would not have been detected, assuming a one-to-one relationship between the geometric mean of S5F scores and the actual mutation rate. Additionally, inference of B cell genealogies may be affected by non-equilibrium nucleotide frequencies, distorting inferences of absolute rates [68]. Standard phylogenetic methods such as the ones used in this study further assume that sites evolve independently, an assumption violated by the context-dependency of mutations in B cell receptor sequences. More elaborate substitution models that account for the particularities of B cell evolution [113, 68, 31] might overcome some of these limitations and help quantify the relationships between mutability metrics and absolute mutation and substitution rates.

Our results suggest a trade-off between the short-term and long-term adaptability within B cell lineages. Repeated infections by antigenically related pathogens, such as influenza

viruses, often recall memory B cell populations [180, 103, 77, 8]. Preferential recruitment of experienced, less mutable B cells at the expense of naive B cells may compromise the long-term adaptability of the immune repertoire to these pathogens. Protection against such pathogens may rely on a robust naive response that can complement preexisting B cell lineages once they fail to adapt to new antigens. Weaker naive responses, for instance in the elderly [55], may thus be problematic. Finally, some strategies for eliciting universal responses against HIV and influenza attempt to recapitulate the evolution of long-lived, highly mutated B cell lineages that were found to produce broadly neutralizing antibodies in infected patients [65]. Our results suggest those approaches may be hindered by a decrease in the adaptability of highly mutated antibodies.

3.4 Materials and Methods

3.4.1 Sequence data

We analyzed BCR sequences from three published studies of B cell lineages sampled longitudinally in individual HIV-1 patients not subject to antiretroviral therapy. The CH103 lineage comprises broadly neutralizing antibodies (bnAbs) isolated from individual B cells and clonally related sequences bioinformatically isolated from high-throughput sequencing over 144 weeks of HIV-1 infection in a single donor [104]. Likewise, the VRC26 lineage comprises both bnAbs isolated from individually sorted B cells and clonally related heavy-chain sequences obtained from high-throughput sequencing over 206 weeks of HIV-1 infection in a different patient [39]. In both cases, BCR sequences obtained from high-throughput sequencing were classified as clonally related to the isolated antibodies based on V and J gene usage. A third lineage, VRC01, was sampled longitudinally from a third HIV patient for 15 years starting from approximately five years after the date of infection. VRC01 is a large lineage that possibly consists of multiple independent B cell lineages [181, 145]. We therefore used PARTIS, a recently developed hidden-markov-model-based method [130] to partition

Table 3.1: BCR alignments analyzed. The number of sequences includes the V+J or V germline sequences

Lineage (chain)	V gene	J gene	N seqs	N sites	Sampling times (w.p.i)	Seqs. per point
CH103 (heavy)	IGHV4-59*01	IGHJ4*01	460	321	14, 53, 92, 66, 136, 140, 144	2-232
CH103 (light)	IGLV3-1*01	–	175	285	14, 53, 92, 136, 144	3-83
VRC26 (heavy)	IGHV3-30*18	IGHJ3*01	681	489	38, 48, 59, 119, 176, 206	4-273
VRC26 (light)	IGLV1-51*02	-	464	294	38, 48, 59, 119, 176, 206	6-212
VRC01-13 (heavy)	IGHV1-2*04	IGHJ1*01	157	351	240, 544, 580, 784, 808, 832, 856, 884, 924,948	2-43
VRC01-01 (heavy)	IGHV-ORF15-1*04	IGHJ4*03	124	372		0-35
VRC01-19 (heavy)	IGHV1-2*04	IGHJ1*01	110	438		1-28

w.p.i = weeks post-infection

the heavy-chain VRC01 dataset into sets of sequences likely to have descended from the same naive B cell, and to identify their most likely germline V and J genes. We analyzed the three largest lineages identified in this way, hereafter VRC01-13, VRC01-01 and VRC01-19. We aligned each set of sequences in MACSE v1.01b [134] along with their concatenated V and J genes (for the heavy chains) or along with their V genes only (light chains) (Table 3.1). Including the J genes in the light chain datasets produced bad alignments across the J region.

3.4.2 Phylogenetic inference

Using BEAST v.1.8.2 [42], we fit Bayesian phylogenetic models to the BCR sequence data in order to estimate time-resolved genealogies and internal node sequences for heavy and light chains separately. We used a GTR nucleotide substitution model, assumed a random local clock model to account for potential variation in substitution rates [41], and enabled robust counting [120] to estimate the numbers of synonymous and non-synonymous changes on each branch (Table 3.2). To reduce the number of parameters, we used empirical base frequencies and assumed a shared nucleotide transition matrix across codon positions for the robust counting inference. The inferred dynamics of mutability losses and gains were qualitatively robust to the choice of demographic model (constant population size, logistic or exponential growth) used to calculate coalescent prior probabilities. We set the V+J or V germline sequence of each alignment as the outgroup and assigned them a sampling time

of zero to represent the assumption that, at the start of the infection, the ancestral BCR sequence was close to its corresponding germline genes (except for insertions and deletions at the junctions). For each dataset, four independent MCMC chains were set to run for 500 million steps and sampled every 1,000 steps (for parameter values) and every 10,000 steps (for trees). We downsampled the set of trees recovered by the MCMC chains to obtain 1,000 trees per chain, sampled at regular intervals between the end of the the burn-in and the end of the entire chain.

Because of the long computation times for the larger datasets (lineages VRC26 and CH103), their MCMC chains were interrupted before 500 million steps had been reached. With the exception of heavy-chain VRC26, interrupted chains had ESSs close to or greater than 200 for most parameters and for the likelihood, prior and posterior probabilities. Although MCMC chains failed to converge for the heavy chain data of lineage VRC26, estimates of the parameters of interest (mean mutability changes and fraction of mutability losses across the tree) were numerically close across replicate MCMC chains, and we therefore present the results in the main text.

To investigate their ability to detect consistent changes in mutability, we repeated the phylogenetic analyses on alignments simulated using different population genetic models (Supplementary Methods). For alignments simulated under a model where all sites are equally likely to mutate, mutability losses were estimated to be approximately as frequent as gains. As expected, mutability losses were estimated to be more frequent than gains in datasets simulated under an S5F-based model, where motifs with high S5F score have a high probability of mutating.

3.4.3 Quantifying mutability

We quantified the mutability of different sites based on the S5F mutability model by [185], which assigns relative mutation probabilities to each of the 1024 five-nucleotide DNA motifs. Briefly, the scores were derived by counting the number of synonymous mutations associ-

ated with each motif in a curated database of mutations from high-throughput sequencing studies, followed by normalization of mutational counts based on the frequencies with which the motifs occurred in the data and statistical estimation of the scores for absent or under-represented motifs. The use of synonymous mutations presumably eliminates the influence of selection and thus captures the intrinsic propensity of the different motifs to mutate.

3.4.4 Mutability of randomized sequences

To estimate the expected mutability of B cell receptors given their amino acid sequence under random codon usage, we randomized each sequence by sampling, for each amino acid, a codon from the set of codons encoding that same amino acid. The probability of sampling each possible codon was equal to its relative usage frequency in humans [119]. We computed the geometric mean of S5F mutability scores for 1000 randomizations of each sequence.

3.4.5 Simulations of sequence evolution on MCC trees

For each branch on an MCC tree, we compared the nucleotide sequences of the parent and descendant nodes to identify the number of codon sites with non-synonymous and synonymous differences. Starting from the parent sequence, we simulated an alternative descendant sequence constrained to have the same number of non-synonymous and synonymous differences from the ancestor. To introduce each mutation, we sampled a nucleotide site according to different models for variation in mutation rates across sites (see below). We then mutated the nucleotide at the sampled site, and either kept the mutation if the number of mutations of the same type (synonymous or non-synonymous) was less than the corresponding number inferred from the data, and otherwise rejected it.

To simulate sequence evolution under an S5F-based model, we sampled mutated sites with sampling probabilities proportional to the S5F-scores of the five nucleotide motifs centered at that each site, as in [185]. In addition to the S5F-based model, we performed simulations under two models that assume mutation rates do not depend on the local nucleotide sequence

around each site. First, we simulated sequence evolution under a uniform model where all sites are equally likely to mutate. Second, we used a codon-position-based model, where codon positions 1, 2 and 3 have different mutation rates. Estimates of the relative mutation rates of each codon position were obtained from the robust counting inference performed in BEAST and used to parameterize the simulations.

To model nucleotide transitions, we made use of the fact that, in addition to estimating the relative mutation rate of different motifs, the S5F model includes the probability of transitions between nucleotides, given the motif where a mutation occurs. All simulations used the S5F-based nucleotide transition probabilities from [185], regardless of whether they assumed variable or constant mutation rates across motifs.

3.4.6 Quantifying dN/dS

Because of variation in relative mutation rates across sites, standard dN/dS tests are unreliable when applied to B cell receptor sequence data. For example, an excess of non-synonymous mutations relative to synonymous mutations may result from high mutability in sites where mutations tend to be non-synonymous, leading to incorrect inference of positive selection. To estimate the strength of selection in B cell receptor sequences, we followed [145] and used BASELINe [184] to compare the observed ratio of synonymous and non-synonymous mutations to the ratio expected under motif-specific variation in mutation rates. Briefly, selection strength is quantified as the log-odds ratio between $\pi(1-\pi)$ and $\hat{\pi}(1-\hat{\pi})$, where π and $\hat{\pi}$ are the observed and expected frequencies of non-synonymous mutations, respectively. The numbers of synonymous and non-synonymous mutations for each sequence are obtained by comparing the sequence to a reference sequence, which, in our case, was the ancestral sequence of the corresponding lineage.

3.5 Acknowledgements

Daniel Zinder contributed to this project and was a coauthor in the published paper. This work was supported in part by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health National (grant number DP2 AI117921) and by a Complex Systems Scholar Award from the James S. McDonnell Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIAID or the National Institutes of Health. This work was completed in part with resources provided by the University of Chicago Research Computing Center. Code implementing the analyses and figures is available at `https://github.com/cobeylab/evolution_of_mutability`. We thank three anonymous reviewers for comments and suggestions.

3.6 Supplementary Information

Simulation of B cell receptor alignments

To assess the power of a random local clock model [41] combined with robust counting [120] to detect changes in synonymous and non-synonymous substitution rates during B cell evolution, we analyzed B cell alignments simulated under different underlying mutation rates. In each simulation, the mutation rate was either constant or variable over time, and either constant or variable across sites. In simulations with variable mutation rates across sites, site-specific rates either depended on the site's S5F mutability [185] or on whether the site was a WRCH/DGYW hotspot [137]. In simulations with variable mutation rates across sites, mutation rates also changed over time as the S5F mutability of the sequences (or the number of WRCH/DGYW hotspots in them) changed. Simulations are summarized below.

Scenario 1: Mutation rate constant across time and across sites.

Scenario 2: Mutation rate constant across sites but decreasing over time.

2a: 10% decrease in rate over simulation period.

2b: 20% decrease in rate.

2c: 30% decrease in rate.

2d: 40% decrease in rate.

2e: 50% decrease in rate.

Scenario 3: Variable mutation rates over time and across sites.

3a: Site-specific mutation rates depend on S5F mutability.

3b: WRCH/DGYW hotspots are three times more likely to mutate.

3c: WRCH/DGYW hotspots are 30 times more likely to mutate.

For the simulations, we modified a simple forward-time Wright-Fisher model to impose a fitness cost on non-synonymous mutations. The status of a mutation (synonymous or non-synonymous) is defined relative to a fixed reference sequence. As a reference sequence, we used the heavy chain sequence at the root node of lineage CH103, inferred under a logistic growth prior (98-100% identical to sequences inferred in four replicate chains under a constant population size prior). An initial population is generated consisting of a single copy of the reference sequence. At each subsequent generation t, N_t sequences are produced by replicating sequences from the previous generation with the possibility of mutation. We assumed N grows logistically:

$$N_t = N_{t-1} + rN_{t-1}\left(1 - \frac{N_{t-1}}{K}\right) \tag{3.1}$$

Logistic growth assumes the B cell population initially expands exponentially with intrinsic growth rate r and then saturates at the carrying capacity K.

At each generation t, we generate a number of new sequences equal to the nearest integer to N_t. The probability that a newly generated sequence at generation t descends from sequence i in generation $t-1$ is equal to the fitness of sequence i, w_i, normalized by the sum of fitness values across the entire population at $t-1$. Any sequence i whose amino acid translation is the same as that of the reference sequence has fitness $w_i = 1$. Each

non-synonymous mutation relative to the reference sequence adds s to the value of w_i, with negative values of s representing a fitness cost (w_i cannot go below 0). Newly generated sequences undergo mutations at fixed or variable rates, depending on the scenario.

Modeling relative and absolute mutation rates

Different models have been proposed to describe variation in somatic hypermutation rates across different nucleotide motifs [137, 185, 47]. For site i in sequence j, those models can be used to assign a *relative* mutation rate $m_{i,j}$ to site i, based on its local sequence context in sequence j. However, the precise relationship between the *relative* mutability of a site and the site's *absolute* mutation rate is unclear. To model absolute mutation rates as a function of the relative mutability of a sequence's motifs, we assume that the average mutation rate per site per generation for sequence j, $\overline{\mu}_j$, is proportional to the sequence's average relative mutability, $\overline{m}_j$:

$$\overline{\mu}_j = k \times \overline{m}_j \tag{3.2}$$

Let $\overline{m}_0$ be the average relative mutability of the reference sequence. We choose k so that the reference sequence has an average mutation rate per site per generation, $\overline{\mu}_0$, equal to $1/4L$, where L is the sequence length (in number of nucleotides):

$$k = \frac{1}{\overline{m}_0 4L} \tag{3.3}$$

In preliminary simulations under default values for other parameters (see below), this choice of $\overline{\mu}_0 = 1/4L$ produced alignments visually similar to those observed for real B cell lineages in terms of overall nucleotide diversity. For any sequence j, the average mutation rate per site per generation is then given by:

$$\overline{\mu}_j = \frac{\overline{m}_j}{\overline{m}_0 4L} \tag{3.4}$$

Thus, a sequence with half the average relative mutability of the reference sequence has half the average mutation rate per site per generation. Finally, site-specific mutation rates per generation for sequence j are given by:

$$\mu_{i,j} = m_{i,j} \times k = \frac{m_{i,j}}{\overline{m}_0 4L} \tag{3.5}$$

Note that scaling site-specific mutation rates to $1/4L$ does not change the relative mutability of the sites. For example, for sites a and b in the same sequence j:

$$\frac{\mu_{a,j}}{\mu_{b,j}} = \frac{m_{a,j}}{m_{b,j}}$$

To models scenarios 1 and 2 (where mutation rates are independent of sequence context) we let $\mu_{i,j}(t) = c(t) \times 1/4L$ be the mutation rate per time per generation for all sites in all sequences at generation t, where $0 \leq c(t) \leq 1$. In scenario 1 we let $c(t) = 1$ for all t, whereas in scenario 2 we choose decreasing values of $c(t)$ for different time intervals

In the S5F-based parameterization (scenario 3a), we set $m_{i,j}$ to the relative mutability scores from [185], based on a five-nucleotide window centered on site i. The first two sites and the last two sites, for which S5F is indeterminate since the neighbors are unknown, are assigned mutability zero. We used motif-specific transition probabilities between nucleotides inferred by the S5F model along with motif-specific mutation rates.

In the "hotspot" parameterization (scenarios 3b-c), we let $m_{i,j}$ be either 1, if site i is not at the central position of a hotspot, or h, if it is. Site i is at the central position of a hotspot if it is occupied by the underlined nucleotide in a $WR\underline{C}H$ or a $D\underline{G}YW$ motif, where $W = \{A/T\}$, $R = \{A/G\}$, $H = \{A/T/C\}$, $D = \{A/T/G\}$ and $Y = \{C/T\}$ [137]. A site is assigned mutability 1 f it cannot be determined whether or not that site is at the center of a WR$\underline{\text{C}}$H or a D$\underline{\text{G}}$YW motif (for example, the left neighbors of the first site in a sequence and the right neighbors of the last site in a sequence are unknown). We assumed uniform transition probabilities between nucleotides.

We ran simulations for 2000 generations, sampling 25 sequences every 250 generations starting at generation 500. We decreased $c(t)$ by 0.2 every 250 generations starting at generation $t = 1000$ for scenario 2a, by 0.4 every 250 generations starting at generation $t = 1000$ for scenario 2b, by 0.5 every 250 generations starting at generation $t = 750$ for scenario 2c, by 0.8 every 250 generations starting at generation $t = 1000$ for scenario 2d, and by 0.1 every 250 generations starting at generation $t = 1000$ for scenario 2d.

We ran all simulations with $s = -0.01$, $r = 0.7$ and $K = 1000$. For scenario 2, we repeated the simulations with $s = -0.005$, $r = 0.01$, $K = 2000$. Under the first set of parameters, the simulated population reached the carrying capacity after approximately 20 generations, before the start of sampling at generation 500. Under the second set of parameters, carrying capacity was reached after approximately 1500 generations, well into the sampling period.

We analyzed the simulated alignments using BEAST v.1.8.2 to test if declines in synonymous and non-synonymous substitution rates are correctly detected in scenarios 2 and 3.

Changes in substitution rates over time

For each tree in the posterior distributions inferred for observed and simulated alignments, we computed the estimated synonymous and non-synonymous substitution rates for each branch by dividing estimated counts of synonymous and non-synonymous substitutions (obtained by robust counting) by the branch's length measured in time. For each tree in the posterior distributions of simulated alignments, we estimated pairwise linear regression coefficients between branch times (predictor variable, measured at each branch's parent node) and total, synonymous and non-synonymous substitution rates (response variables).

Supplementary figures and tables

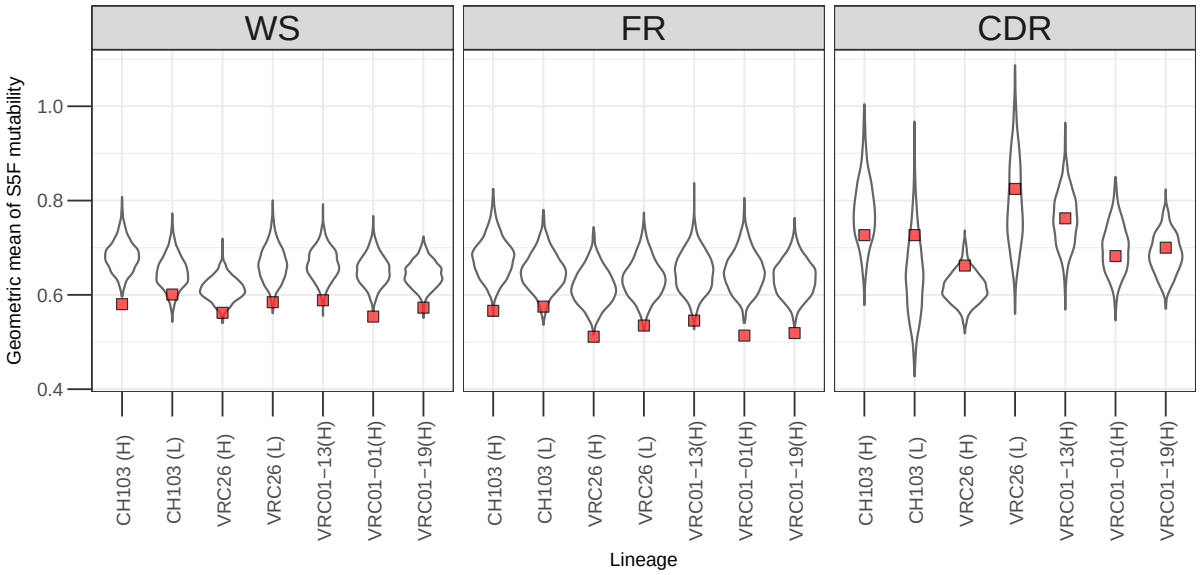

Figure 3.6: Mutability of the inferred ancestral sequences of long-lived B cell lineages (red squares) compared with the distribution of mutability values obtained by randomizing the ancestral codon sequence while keeping the amino acid sequence constant. Mutability was measured as the geometric mean of the S5F scores across sites. Results are shown for the whole sequences (WS) and separately for framework regions (FRs) and complementarity determining regions (CDRs).

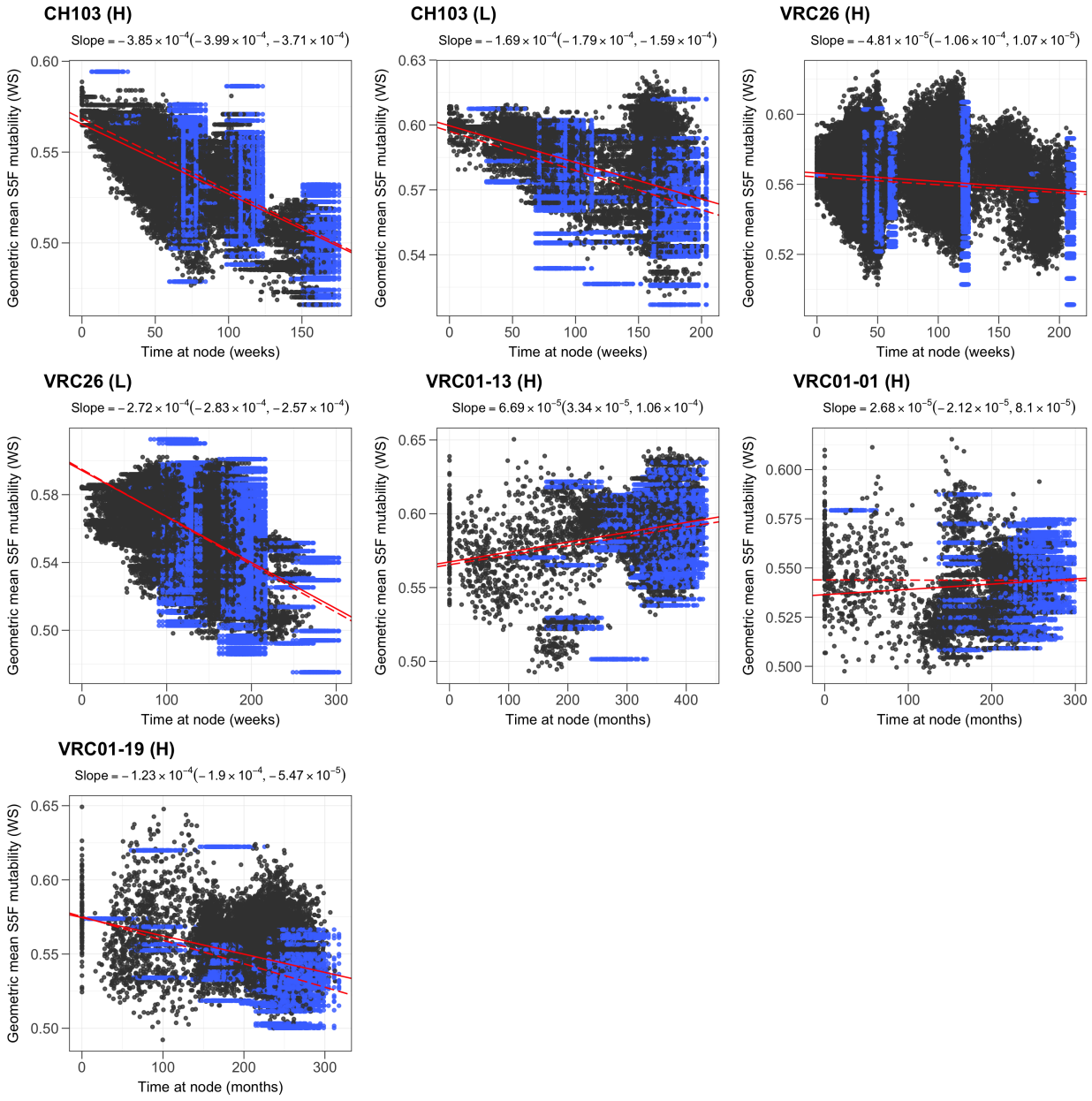

Figure 3.7: Evolution of mutability in long-lived B cell lineages. Mutability was measured as the geometric mean of S5F scores across all sites in the sequence. Scatterplots show mutability over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.

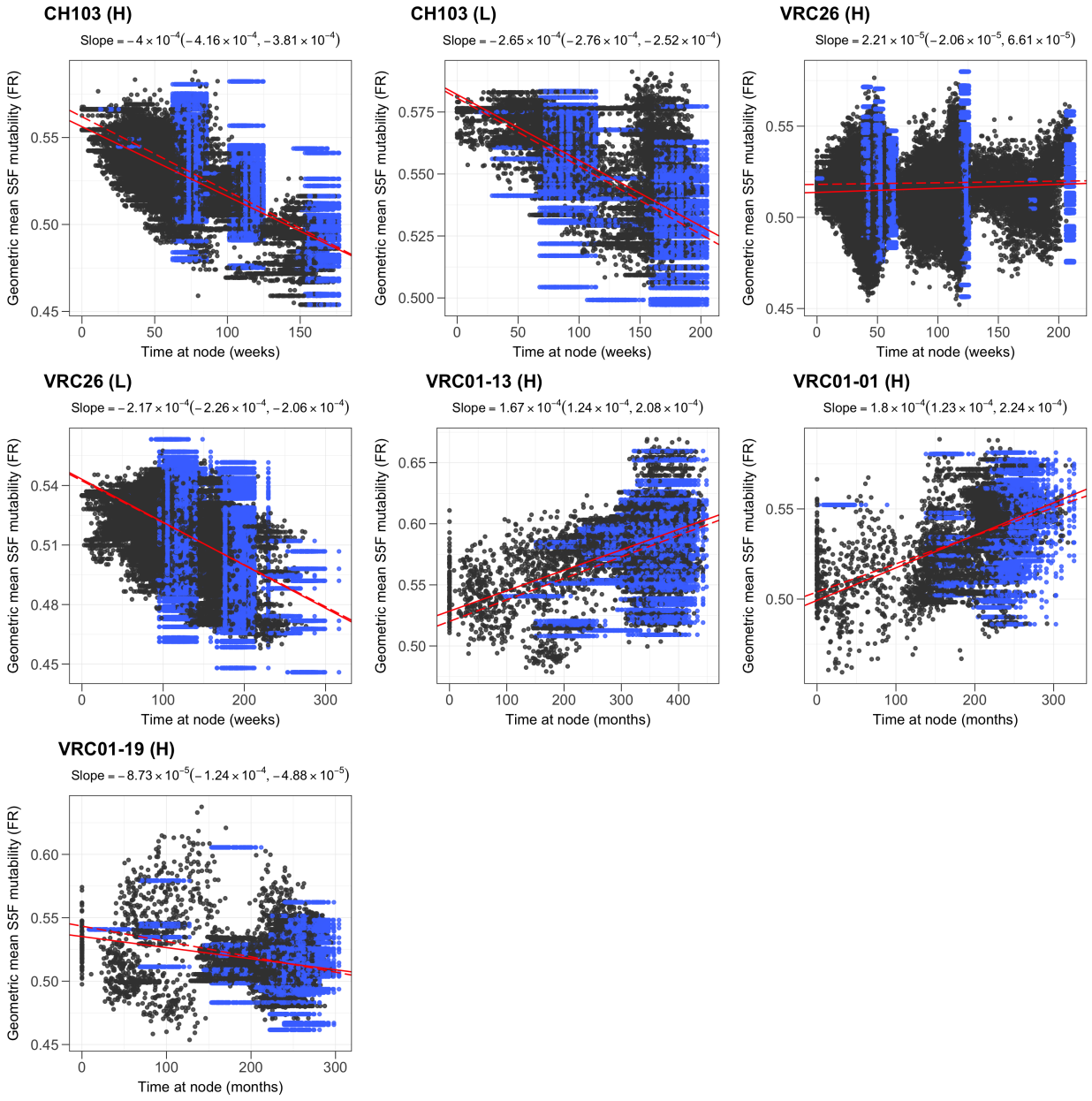

Figure 3.8: Evolution of mutability in the framework regions (FRs) of long-lived B cell lineages. Mutability was measured as the geometric mean of S5F scores across all sites in FRs. Scatterplots show mutability over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.

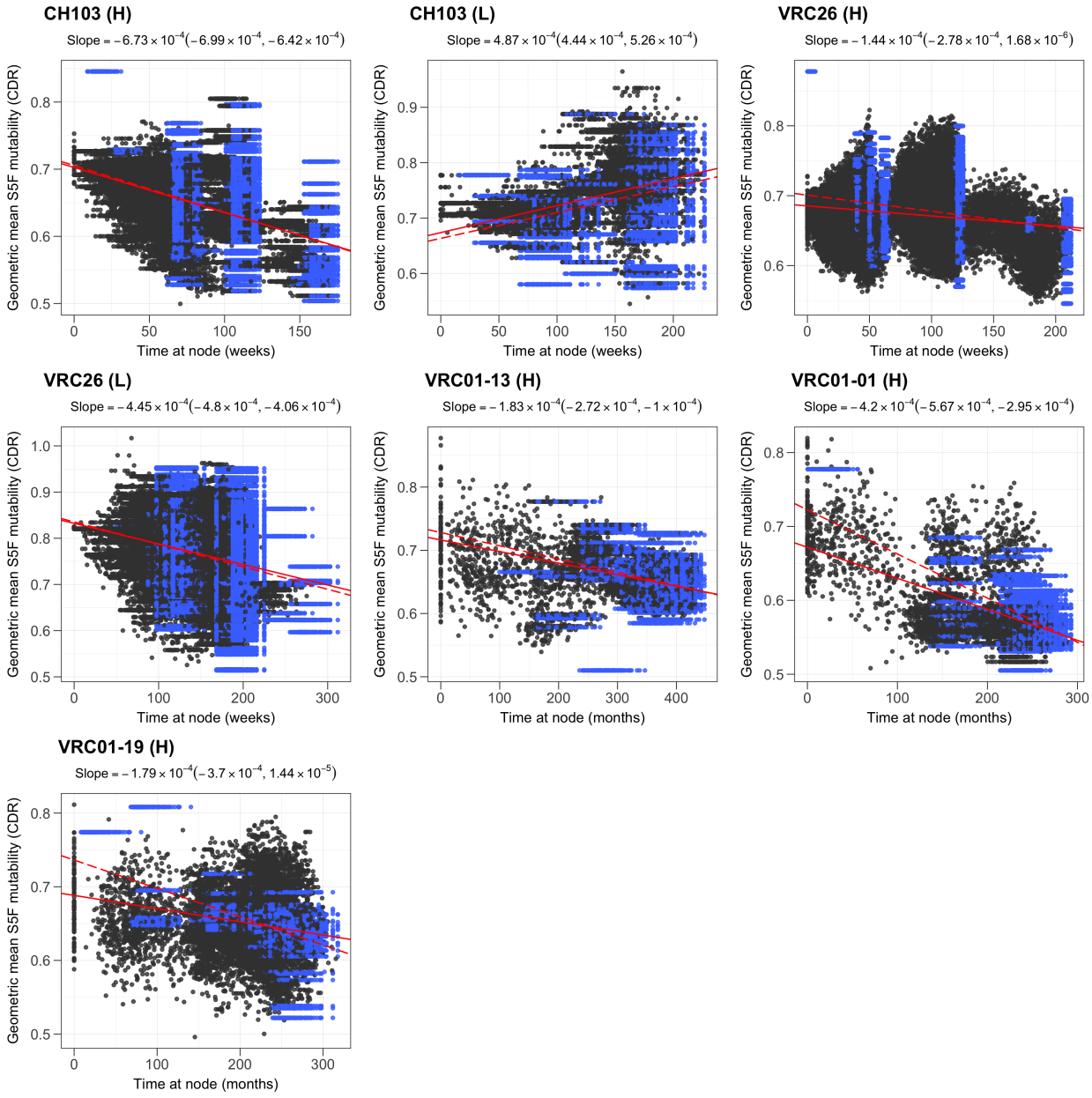

Figure 3.9: Evolution of mutability in the complementarity determining regions (CDRs) of long-lived B cell lineages. Mutability was measured as the geometric mean of S5F scores across all sites in CDRs. Scatterplots show mutability over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.

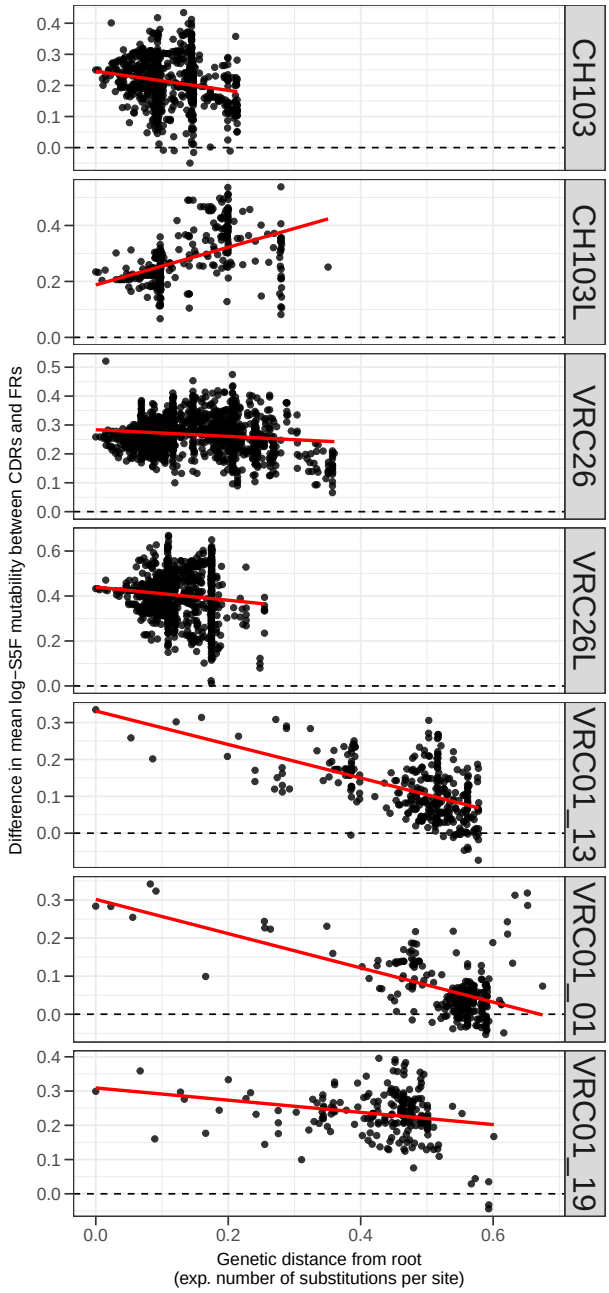

Figure 3.10: Evolution of the difference in mutability between complementarity determining regions (CDRs) and framework regions (FRs) in long-lived B cell lineages. The relative difference is calculated as the average log-S5F mutability of CDRs minus the average log-S5F mutability of FRs. Each point corresponds to a node in the maximum-clade-credibility tree of each lineage. Differences in CDR and FR mutability are plotted as a function of genetic distance from the root of the tree, measured as the expected number of substitutions per site since the root.

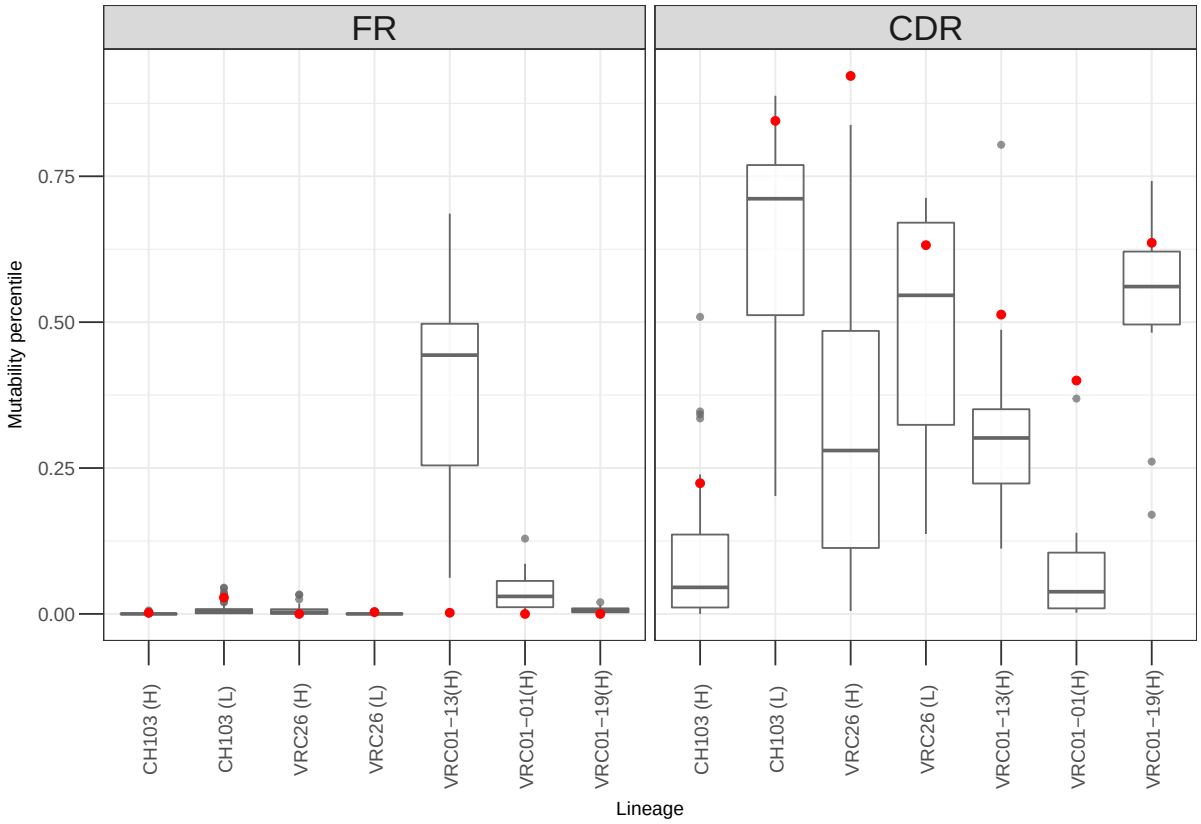

Figure 3.11: Mutability of B cell receptor sequences from different B cell lineages relative to the expected distribution of mutability values obtained by randomizing codon sequences while keeping the amino sequences constant. Mutability was measured as the geometric mean of S5F scores across sites. The distribution of mutability percentiles obtained for sequences sampled at the last sampling time point in each dataset is shown in gray. The mutability percentile of each lineage's ancestor is shown in red. Results are shown separately for framework regions (FR) and complementarity determining regions (CDRs).

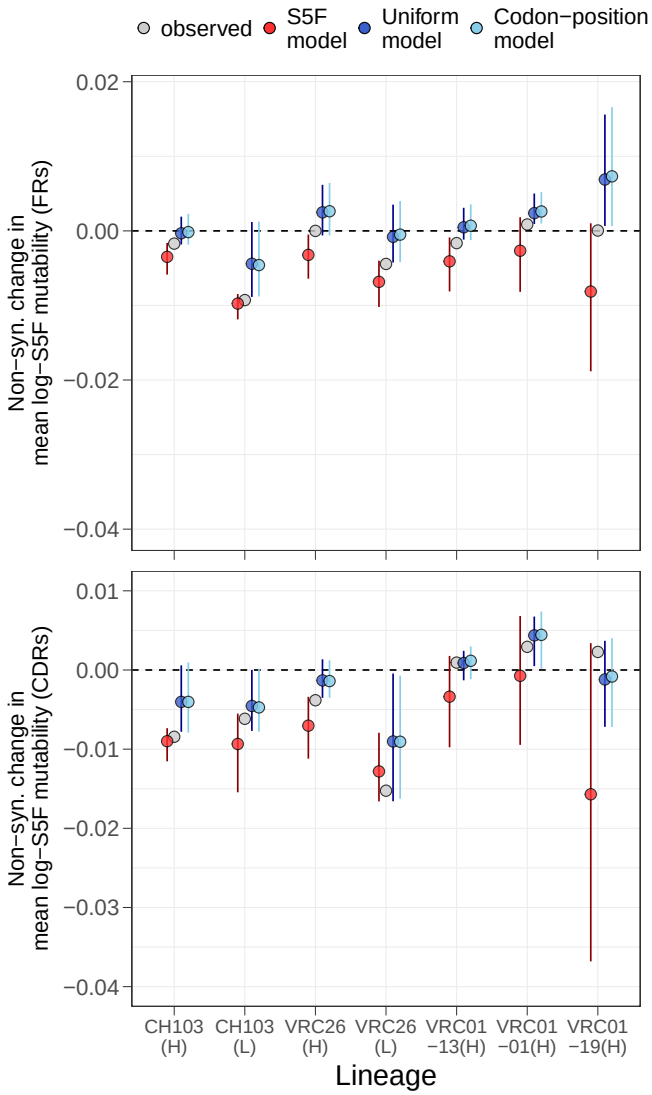

Figure 3.12: Changes in mean log-S5F mutability due to non-synonymous changes averaged across all branches of different anti-HIV B cell lineages. Gray points indicate values inferred from the data, and colored points indicate values obtained by simulation under different models. Red indicates an S5F-based model where different nucleotide motifs mutate with different rates, dark blue indicates a model with no mutation rate variation across sites, and light blue indicates a model with different mutation rates for each position of a codon. Simulations were performed independently for each branch on the MCC tree of different anti-HIV B cell lineages, starting from the inferred sequence of the parent node. Each simulated sequence was constrained to have the same number of non-synonymous and synonymous changes as observed in the branch. Vertical bars indicate the 95% range obtained from 100 simulations per model.

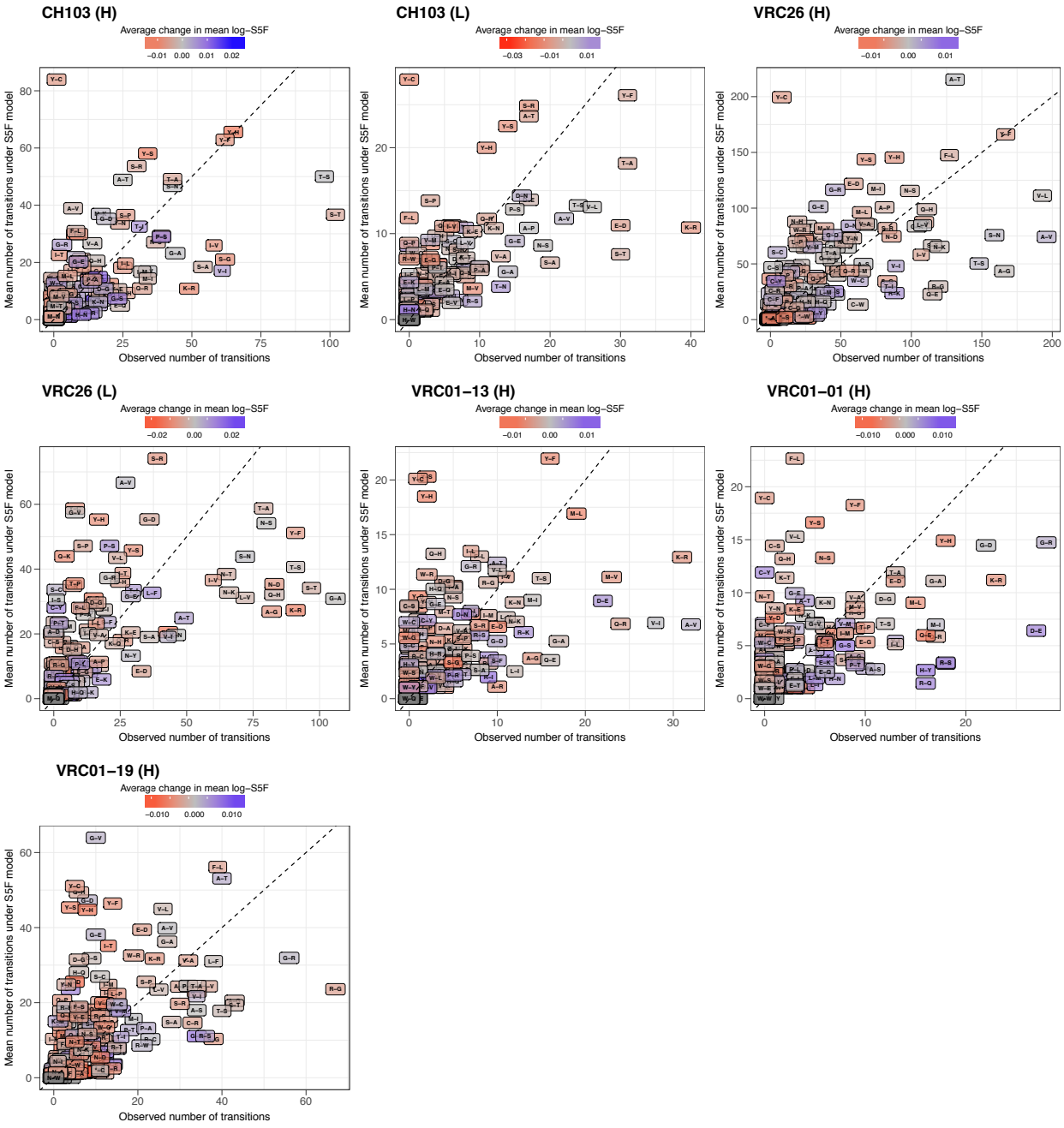

Figure 3.13: Frequency of amino acid transitions simulated under the S5F model compared to the frequencies in the MCC trees of B cell lineages inferred from the data. Amino acid transitions are colored according to their average effect on mean log-S5F mutability in 100 simulations.

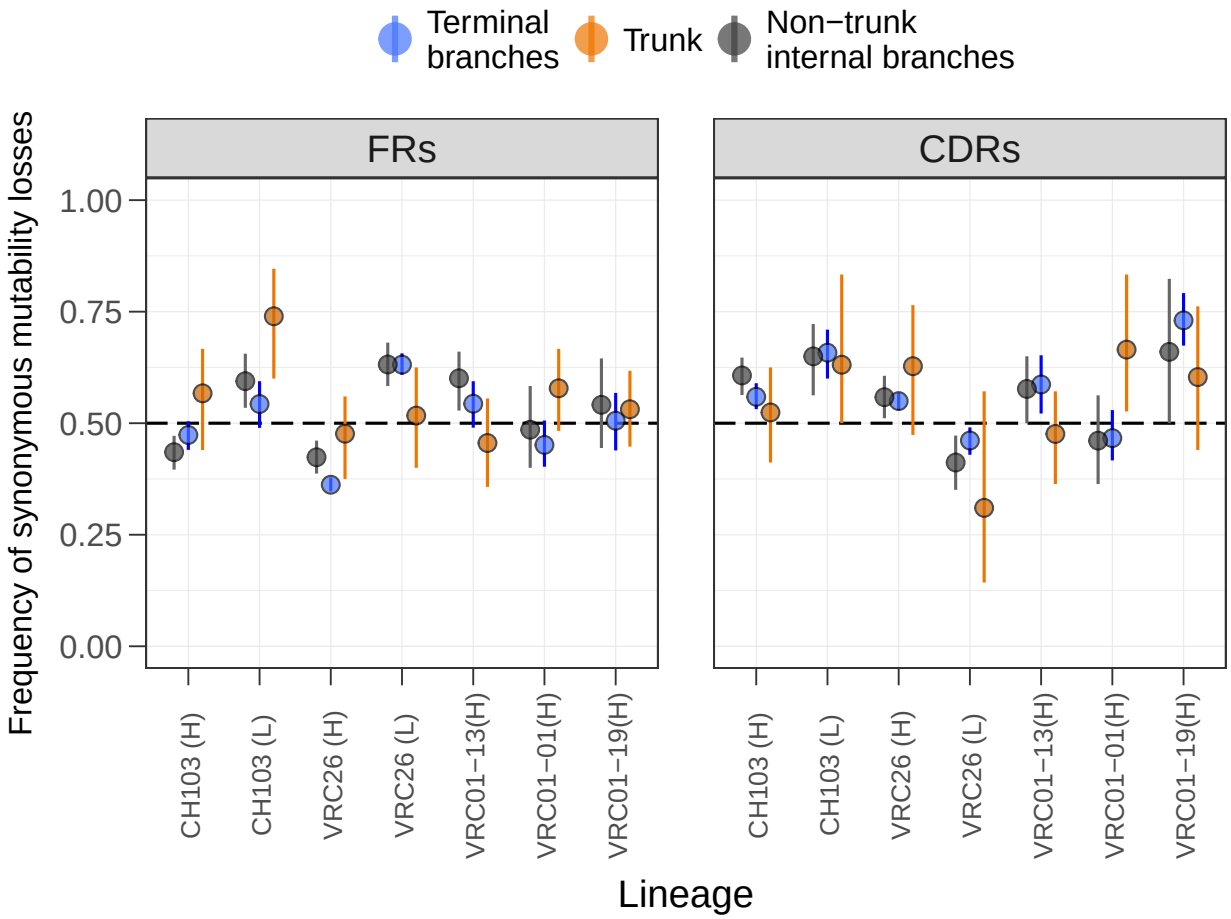

Figure 3.14: Frequency of losses relative to the total number of changes in mean log-S5F mutability caused by synonymous substitutions during the evolution of anti-HIV B cell lineages. Blue indicates changes that on terminal branches, orange indicates changes along the trunk of the tree, and black indicates changes on the remaining internal branches. Results are shown separately for framework regions (FRs) and complementarity determining regions (CDRs). Each point denotes the frequency of changes in mutability that were losses, averaged across a sample of 1000 trees from the posterior distribution. Vertical lines indicate the 95% highest-posterior density interval.

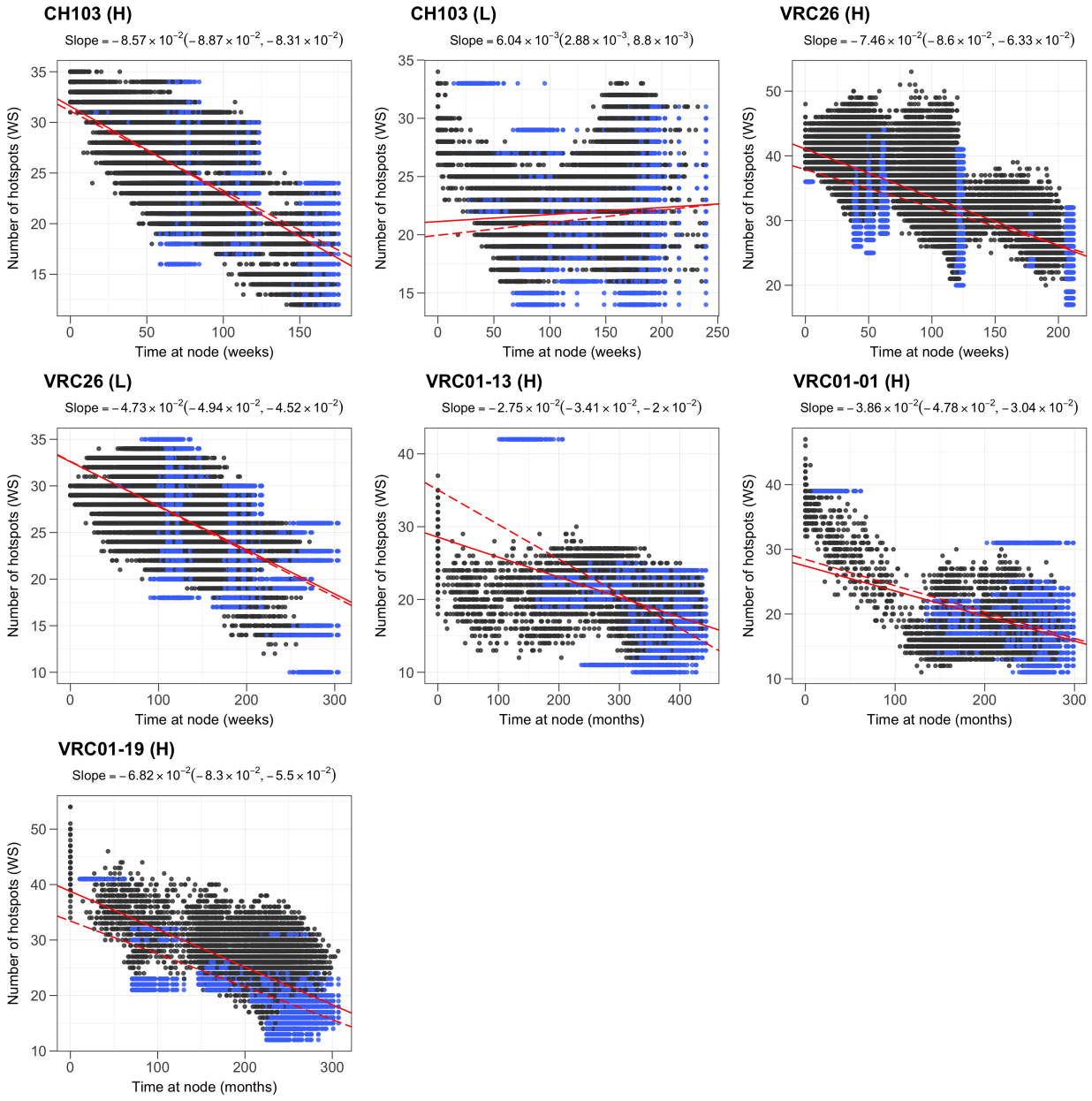

Figure 3.15: Evolution of the number of WRCH/DGYW hotspots in long-lived B cell lineages. Scatterplots show the number of hotspots over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.

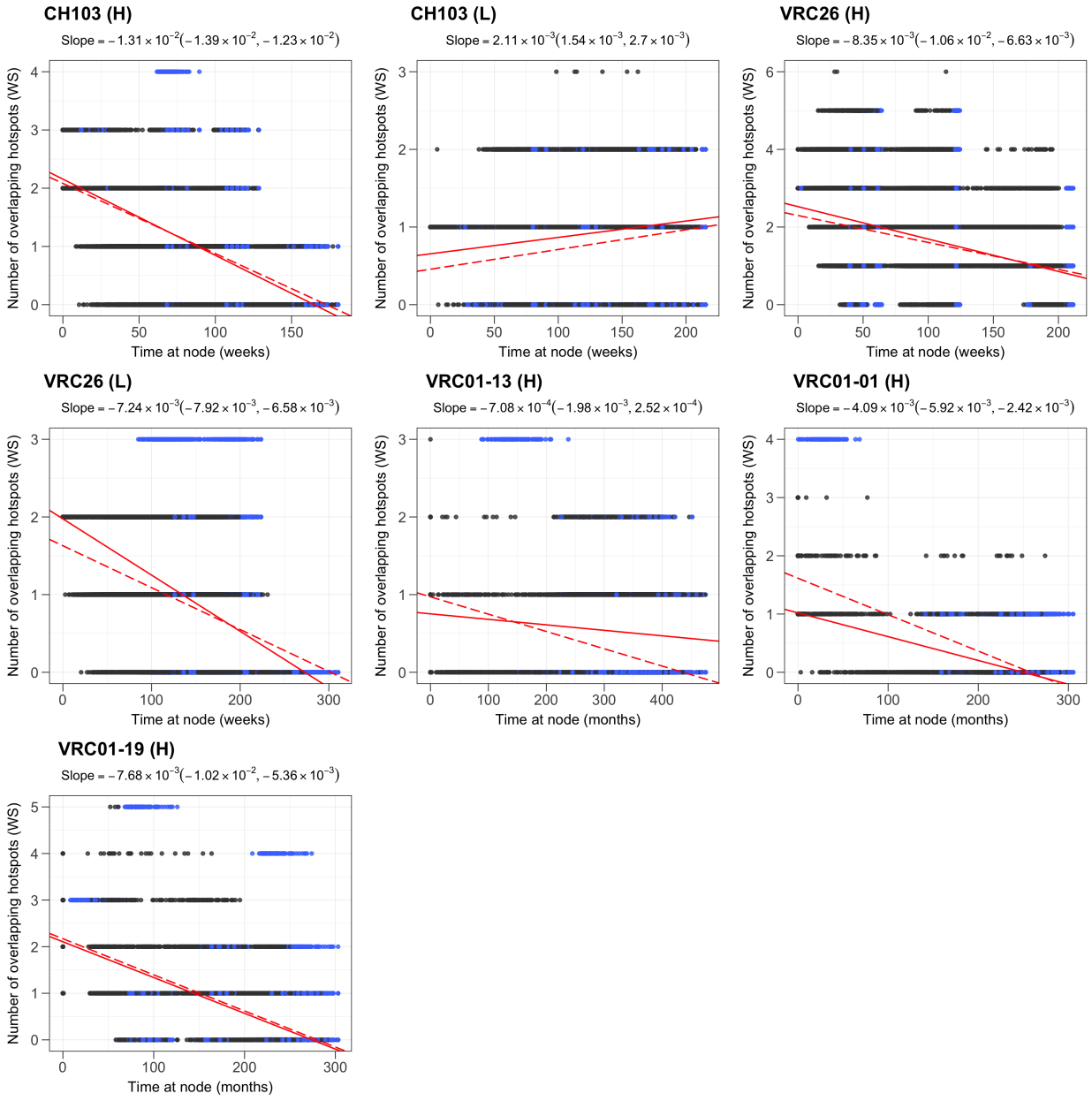

Figure 3.16: Evolution of the number of overlapping hotspots in long-lived B cell lineages. Scatterplots show the number of overlapping hotspots over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.

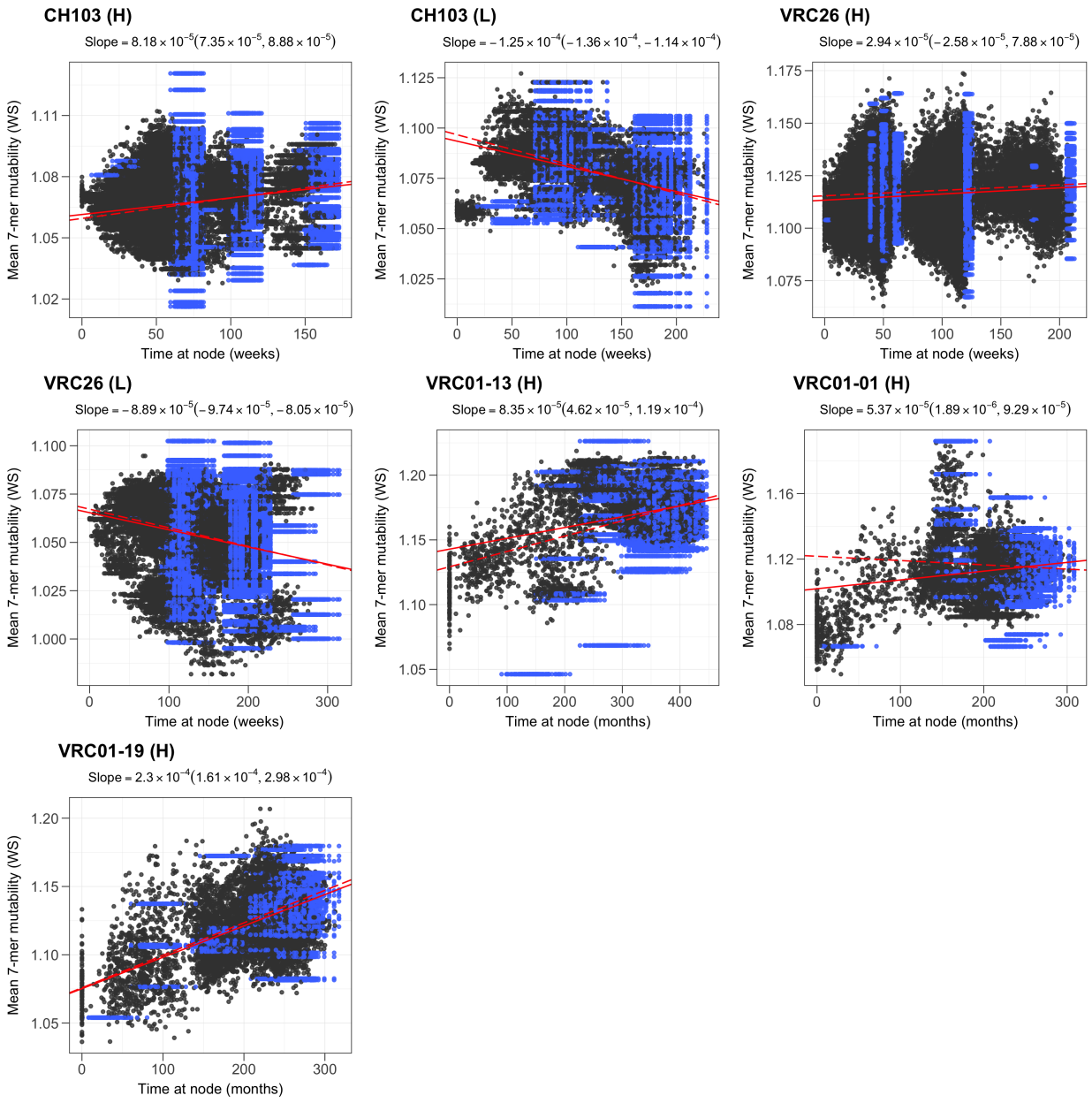

Figure 3.17: Evolution of 7-mer mutability in long-lived B cell lineages. Scatterplots show 7-mer mutability over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.

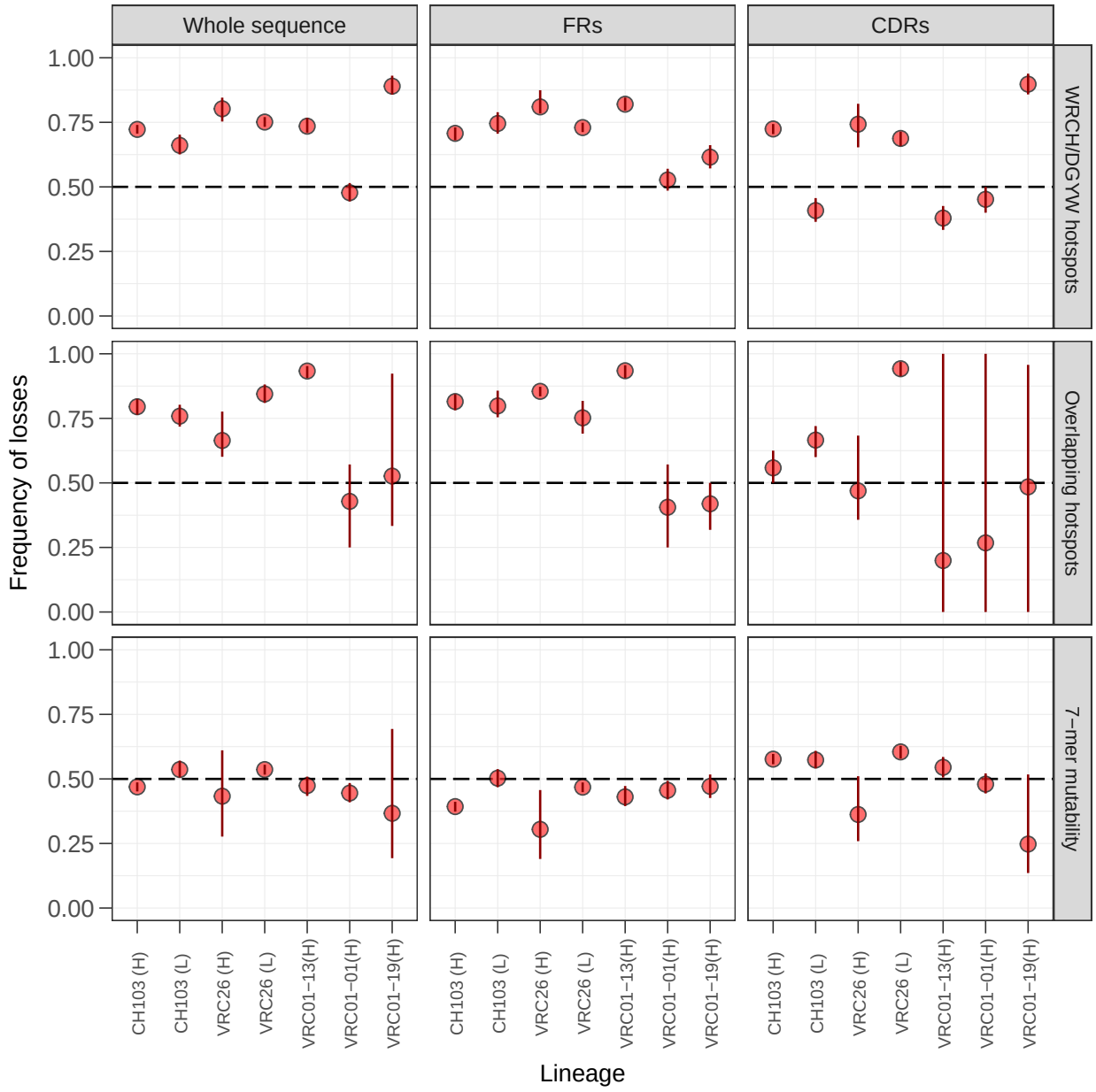

Figure 3.18: Frequency of mutability losses relative to the total number of changes in mutability during the evolution of anti-HIV B cell lineages. Rows correspond to different mutability metrics, and column contain results obtained for the whole analyzed region of the BCR sequence, and separately for framework regions (FRs) and complementarity determining regions (CDRs). Each point denotes the frequency of changes in mutability that were losses, averaged across a sample of 1000 trees from the posterior distribution. Vertical red lines indicate the 95% highest-posterior density interval.

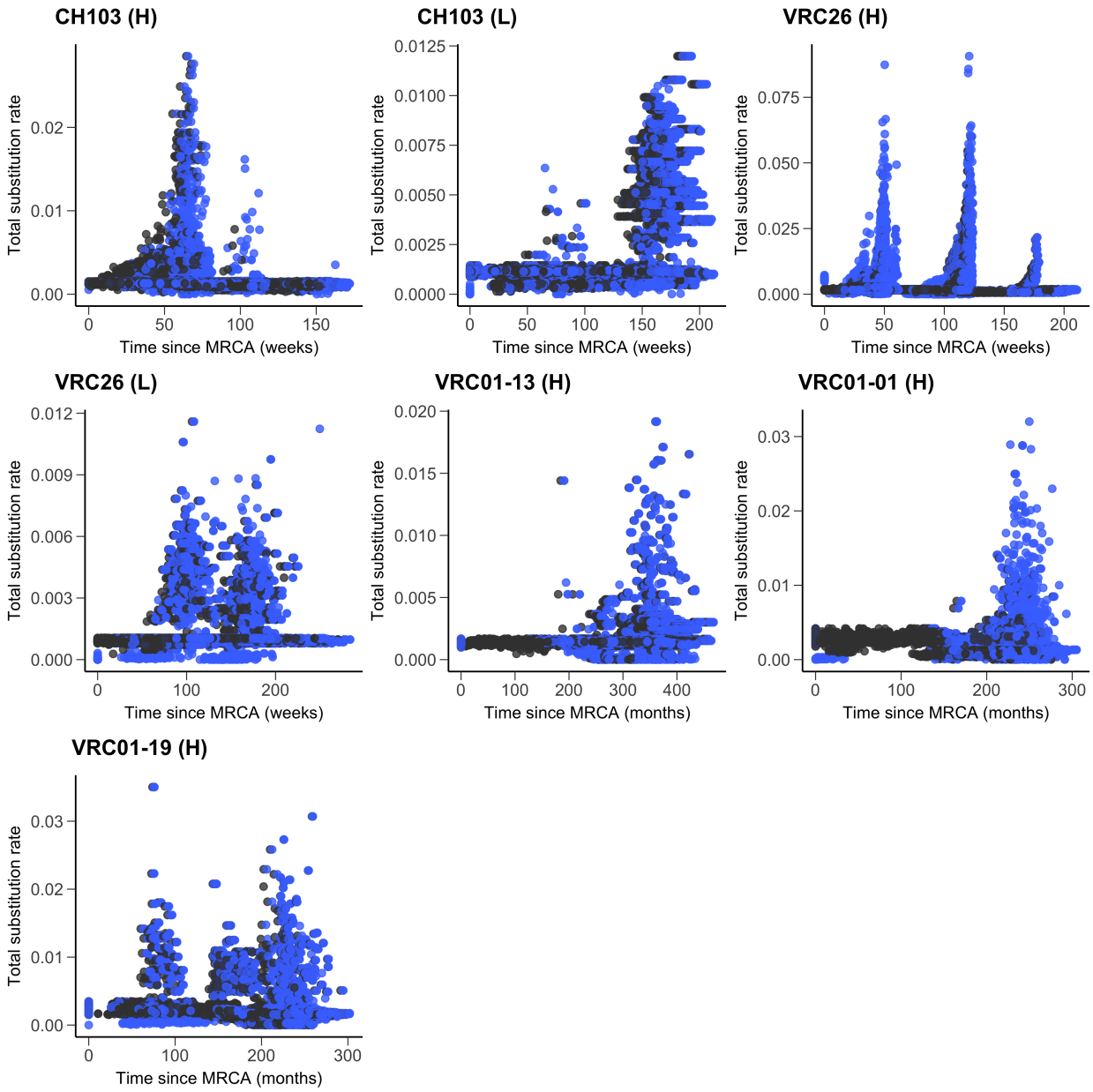

Figure 3.19: Total substitution rate inferred from the random local clock model as a function of time for the observed lineages. Each plot shows the points corresponding to a sample of 100 trees from the posterior distribution inferred by BEAST. Terminal branches are shown in blue, and internal branches are shown in black.

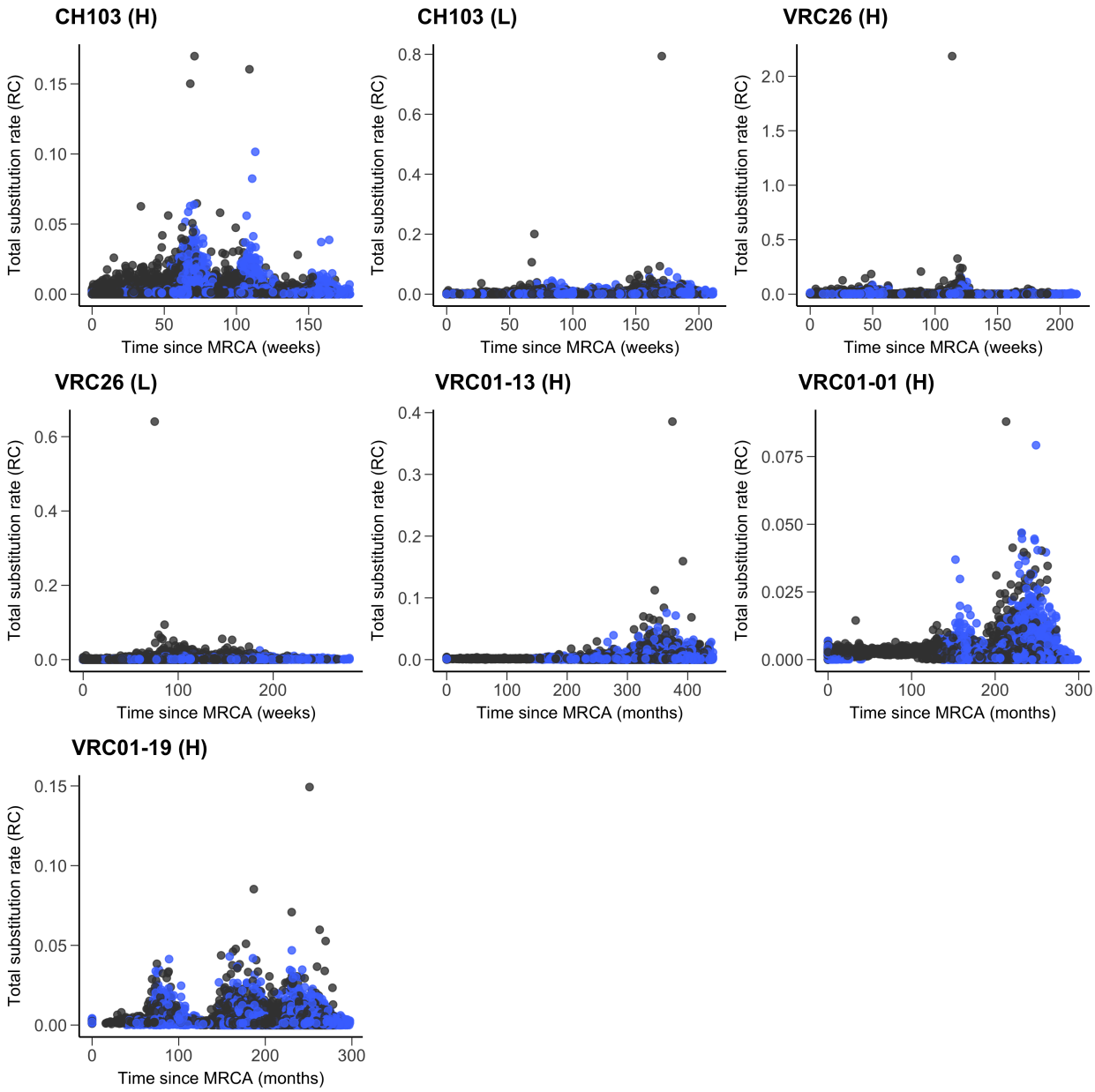

Figure 3.20: Total substitution rate inferred from robust counting (summing synonymous and non-synonymous rate estimates), as a function of time for the observed lineages. Each plot shows the points corresponding to a sample of 100 trees from the posterior distribution inferred by BEAST. Terminal branches are shown in blue, and internal branches are shown in black.

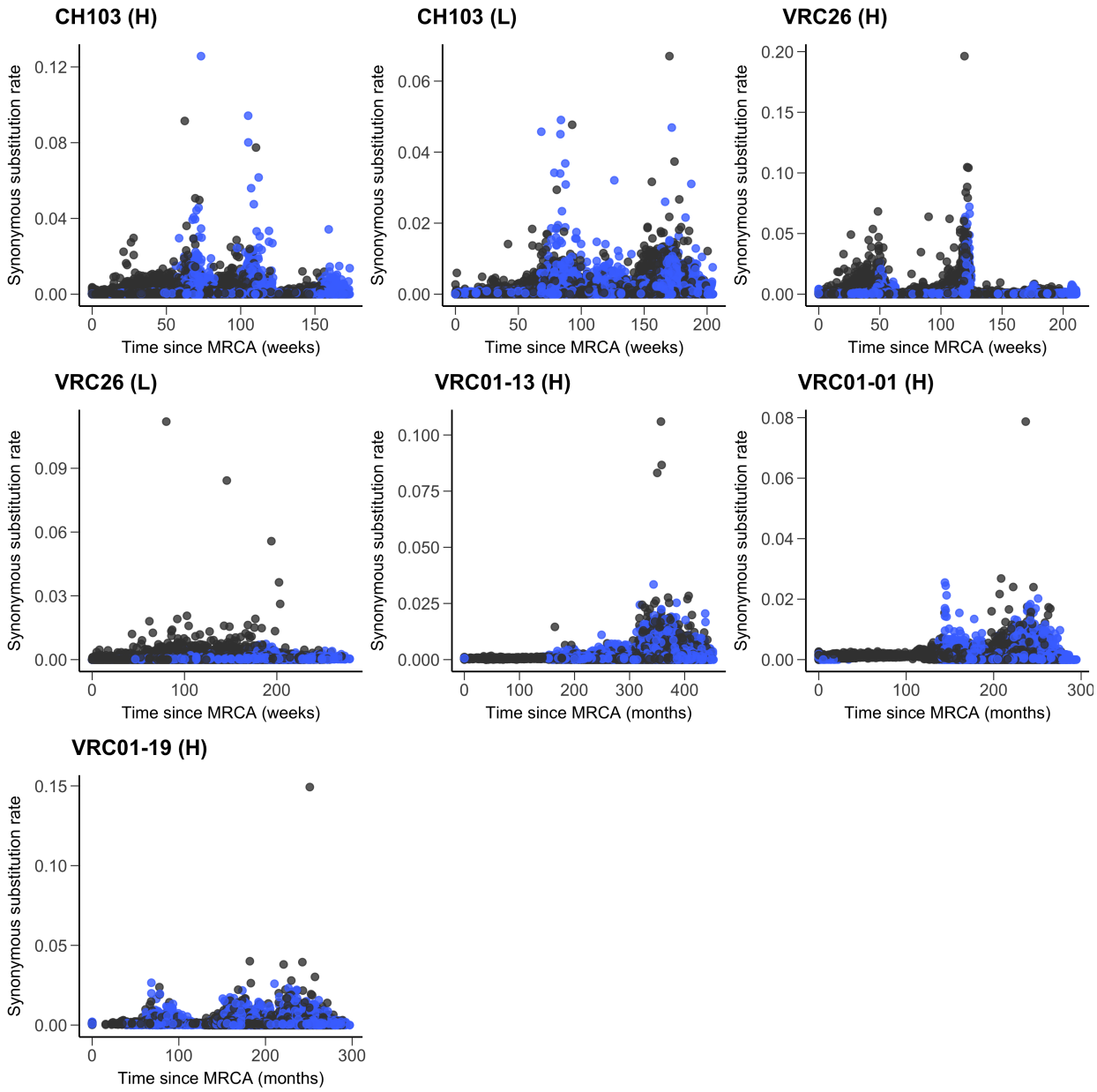

Figure 3.21: Robust counting synonymous substitution rate as a function of time for the observed lineages. Each plot shows the points corresponding to a sample of 100 trees from the posterior distribution inferred by BEAST. Terminal branches are shown in blue, and internal branches are shown in black.

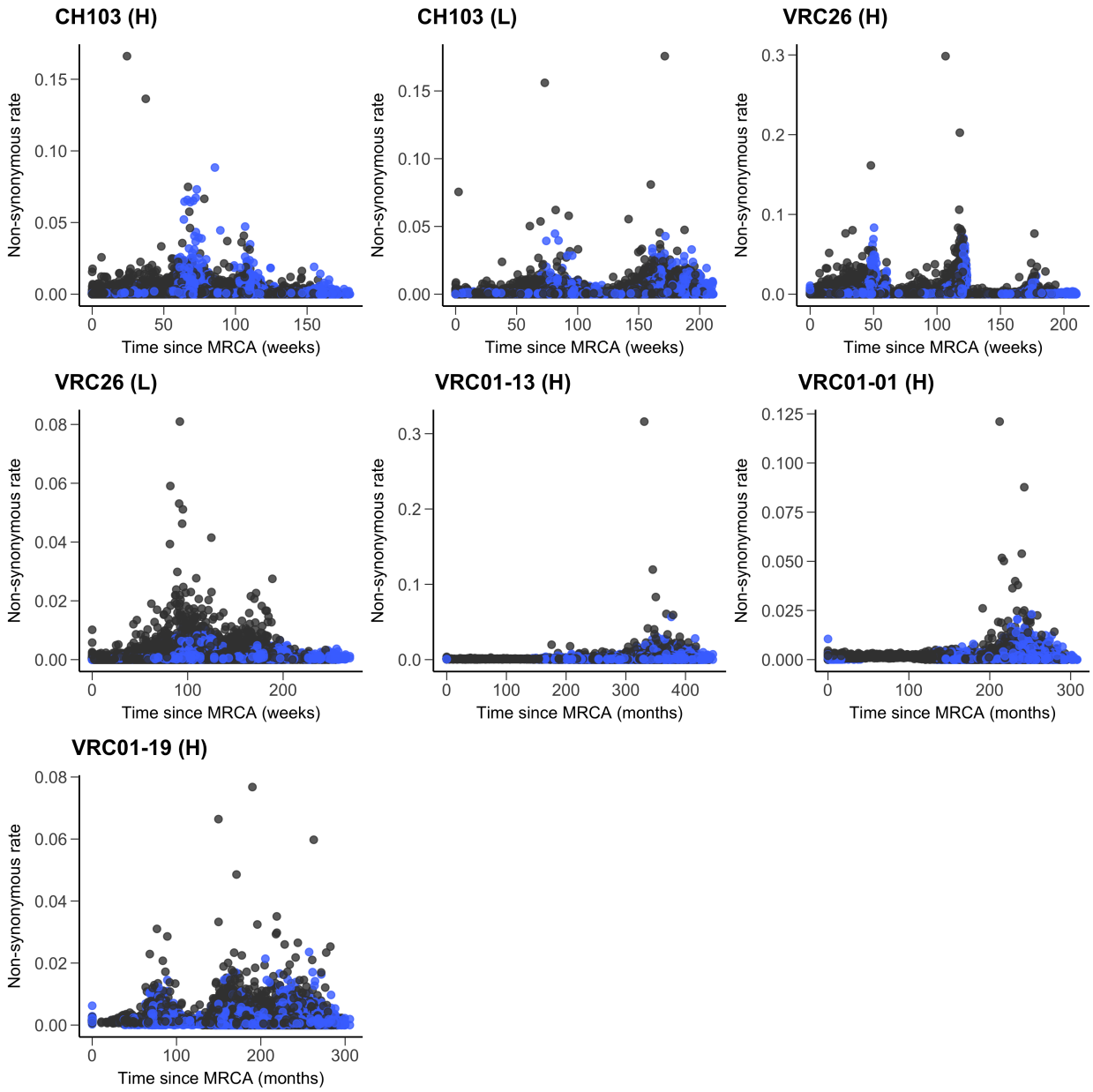

Figure 3.22: Robust counting non-synonymous substitution rate as a function of time for the observed lineages. Each plot shows the points corresponding to a sample of 100 trees from the posterior distribution inferred by BEAST. Terminal branches are shown in blue, and internal branches are shown in black.

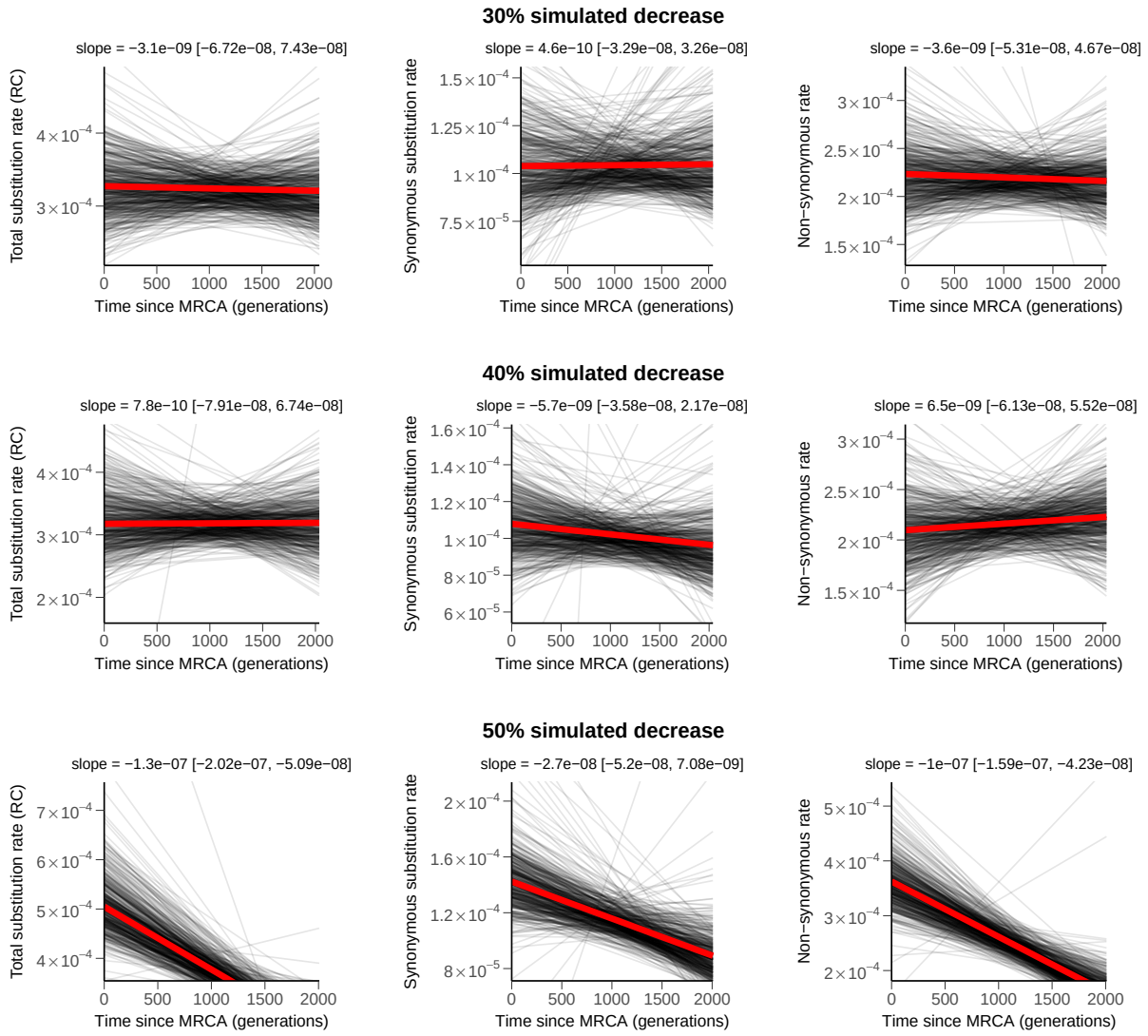

Figure 3.23: Relationship between robust counting substitution rates and time for simulations performed under different levels of decline in the overall mutation rate. Each black line is the linear regression line between branch-specific rates and times for a single tree from the posterior distribution inferred for a simulated alignment using BEAST. Each plot shows a sample of 500 lines. The red lines are the "average" regression lines, with the average intercept and the average slope calculated from a larger sample of 1000 trees from each distribution. Parameter values for the simulations were fitness cost $s = -0.01$, intrinsic growth rate $r = 0.7$ and carrying-capacity $K = 1000$.

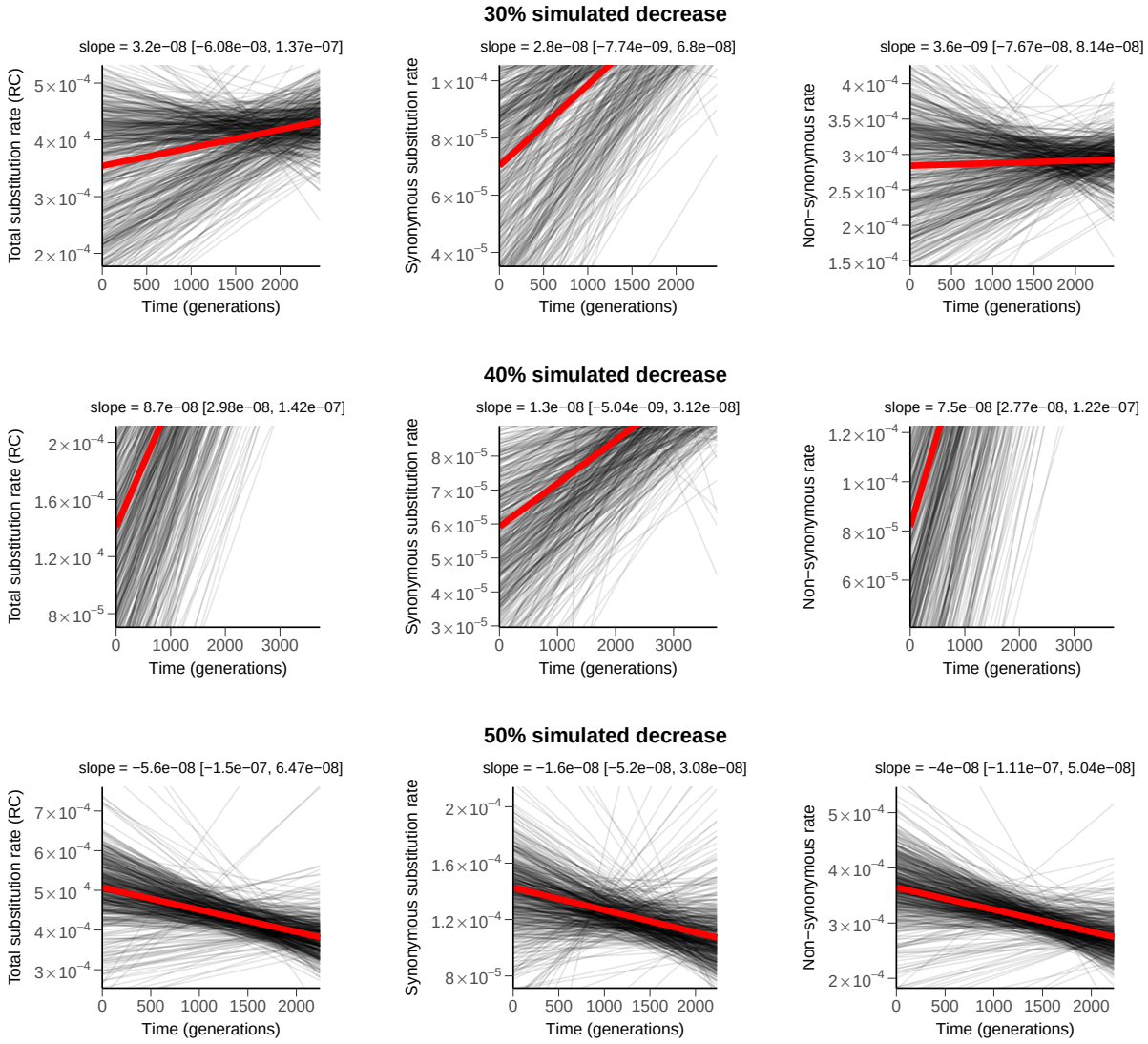

Figure 3.24: Relationship between robust counting substitution rates and time for simulations performed as in Figure 3.23, but using a different set of parameters (fitness cost $s = -0.01$, intrinsic growth rate $r = 0.7$ and carrying-capacity $K = 1000$). Each black line is the linear regression line between branch-specific rates and times for a single tree from the posterior distribution inferred for a simulated alignment using BEAST. Each plot shows a sample of 500 lines. The red lines are the "average" regression lines, with the average intercept and the average slope calculated from a larger sample of 1000 trees from each distribution.

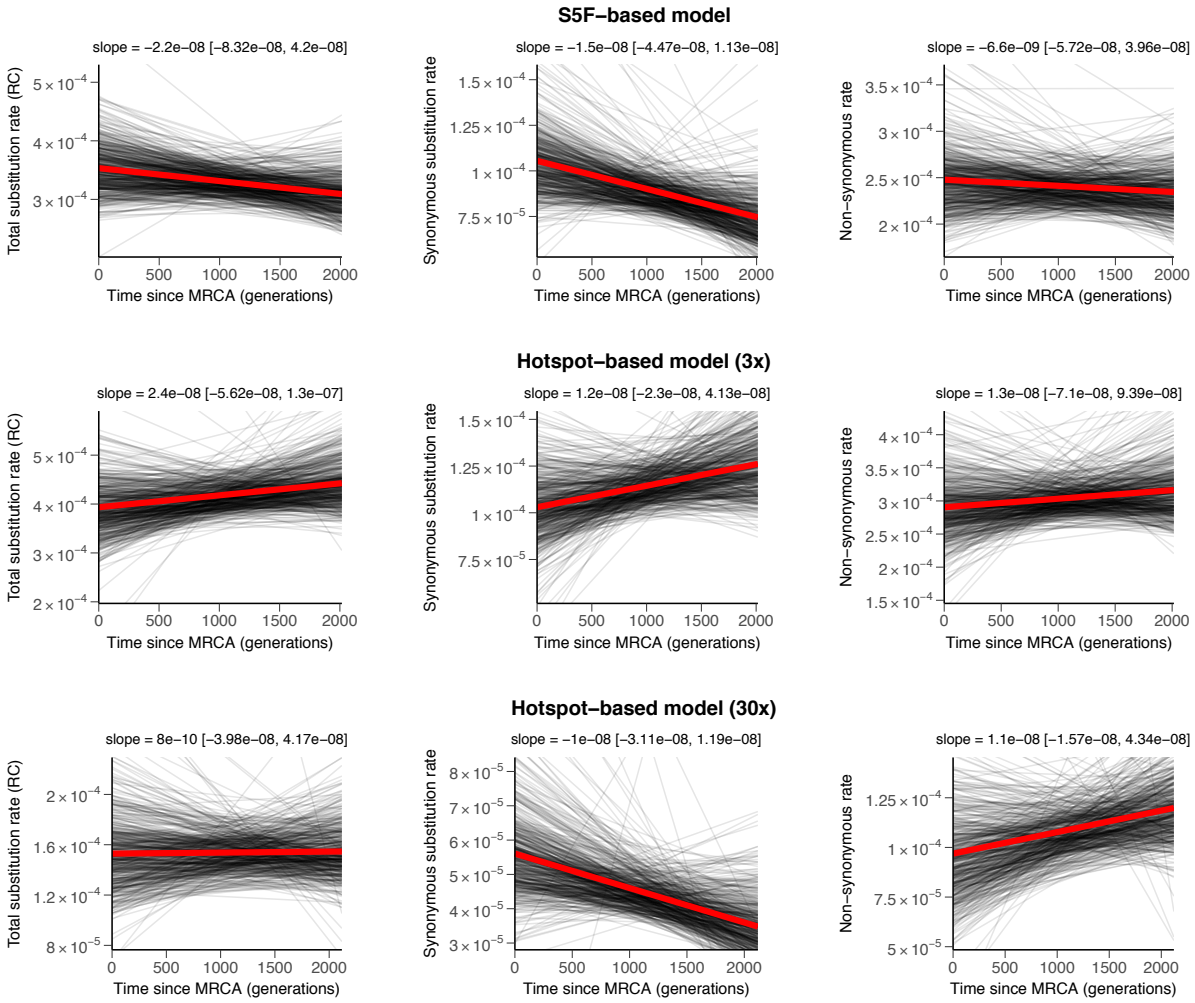

Figure 3.25: Relationship between robust counting substitution rates and time for simulations performed under models where the mutation rate at each site depends on its S5F mutability or on whether that site is at the center of a WRCH/DGYW hotspot (in which case it mutates either 3 or 30 times more frequently than non-hotspots sites). Each black line is the linear regression line between branch-specific rates and times for a single tree from the posterior distribution inferred for a simulated alignment using BEAST. Each plot shows a sample of 500 lines. The red lines are the "average" regression lines, with the average intercept and the average linear coefficient calculated from a larger sample of 1000 trees from each distribution.

Table 3.2: BEAST priors. Priors are left at the default choices expect where specified below.

Parameter	**Description**	**Prior**	**Initial value**
CP1.mu	Relative rate for codon position 1	Γ[0.05, 10]	1
CP2.mu	Relative rate for codon position 2	Γ [0.05, 10]	1
CP3.mu	Relative rate for codon position 3	Γ [0.05, 10]	1
clock.rate	Substitution rate	Uniform[0,10]	0.0001
logistic.popSize	Pop. size under logistic growth	Default	100
logistic.growRate	Logistic growth rate	Default	0.5

CHAPTER 4

ASYMMETRIC CROSS-LINEAGE PROTECTION SHAPES THE DISTINCT AGE DISTRIBUTIONS OF INFLUENZA B LINEAGES

4.1 Introduction

The incidence of pathogens that induce long-lasting immunity, such as measles or mumps, typically peaks at a young age and subsequently decreases as older hosts are more likely to have been infected [56, 7, 62]. However, many pathogens are able to infect the same host multiple times as antigenically distinct strains. Changes in the prevalences of strains over time can lead to different infection histories in hosts born at different times, and such differences in infection history might lead to more complicated distributions of infection [64, 88, 89]. Influenza viruses, for instance, circulate as multiple antigenically distinct variants, including the "types" A and B, "subtypes" of influenza A, "lineages" of influenza B, and clades within them. Changes in variant prevalence over time [82, 139, 144] generate different infection histories that are correlated across birth cohorts, but how differences in infection history translate into protection and shape the age distribution of influenza infections is not fully understood.

Early childhood infections with influenza A have long-lasting consequences for protection against influenza A subtypes, a phenomenon termed "immune imprinting" [60, 61, 10]. Subtypes are distinguished by their surface proteins, hemagglutinin (HA) and neuraminidase (NA). Different subtypes of influenza A have circulated in the 20th century [82], and protection against severe infection and death is higher to subtypes whose HAs are genetically similar to the HA of the subtype with which a person was likely first infected [60, 61, 10]. For instance, early infection with H1N1 or H2N2 is associated with lifelong protection against severe infections with H5N1 and H1N1, and early H3N2 infection protects against severe

infection with H7N9 and H3N2 (the H3 and H7 HAs are more related to each other than to the H1, H2, and H5 HAs). Subtypes H1N1 and H2N2 were the only subtypes circulating between 1918 and 1968, and H3N2 has been more common than H1N1 since 1968. Thus, the age distributions of clinical infections with H1N1 and H5N1 skew young, and the distributions of H7N9 and H3N2 infections skew old [60, 61, 10], because of the lasting impact of childhood immunity to the first HA encountered.

Despite its durability, imprinting protection does not completely prevent re-infection with the same subtype, and models based on imprinting protection alone cannot completely recapitulate the age distribution of cases [61, 10]. Longitudinal analyses of antibody titers, which reveal subclinical infections, suggest that protection after infection with influenza type A decreases significantly within 3.5-7 years [90, 133]. Repeated clinical infections of the same subtype have been observed in the same person [149, 54, 37, 70] and are likely enabled by antigenic evolution of HA and NA, which experience strong positive selection [24, 49, 150, 14]. Thus, protection against infection with a subtype appears to depend not only on imprinting protection (early infection with that subtype or another) but also cross-protection from recent infections with that subtype. This cross-protection can be sensitive to the precise strains with which a person was infected, apparent as birth cohort effects [105, 127].

Like the different subtypes of influenza A, the two lineages of influenza type B have distinct age distributions of medically attended infections. B/Victoria and B/Yamagata diverged in the 1970s to early 1980s [139, 43] and circulated with varying frequencies since (Fig. 4.1A), causing 25% of global influenza cases detected in 2000-2018 [25]. While the incidence of medically attended infections is highest in children for both lineages, B/Yamagata is less common than B/Victoria in teenagers and young adults but more common in the middle-aged (Fig. 4.1B). This difference has been observed in the 2000s and 2010s in Oceania [171], East Asia [162, 182], Europe [153, 128] and North America [147] and globally in sequence databases [172]. Changes in the expression of sialic acid receptors with age have

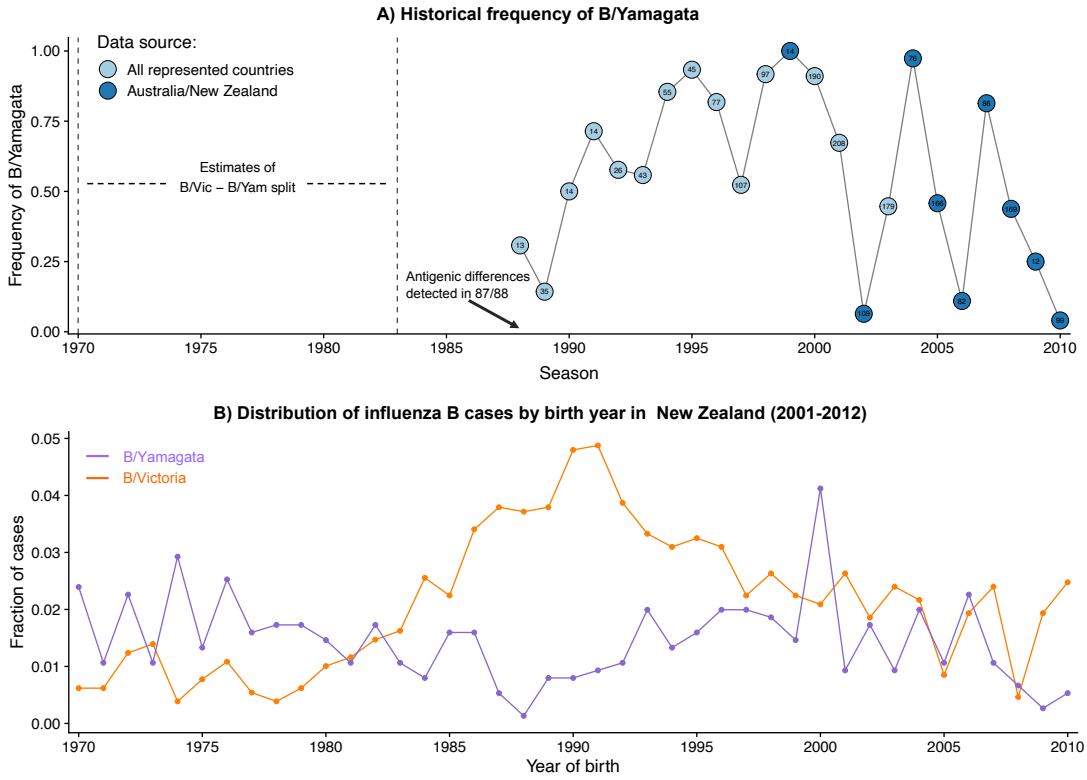

Figure 4.1: **Historical frequencies and age distributions of the influenza B lineages.** (A) Frequency of B/Yamagata estimated from sequences deposited on GISAID and the NCBI Influenza Virus Database. Frequencies were estimated for New Zealand and Australia combined, except when fewer than 10 isolates from those countries were available. Numbers inside the circles show the total number of isolates for each season. Only seasons with at least 10 isolates, either in New Zealand and Australia combined or across all countries represented in the data, are shown. (B) Distribution of medically attended influenza B cases in New Zealand in 2001-2012 by birth year of the infected person [171]. The fraction of cases in each birth year was calculated relative to all cases observed for each lineage (including birth years before 1970, not shown in the figure).

been proposed to explain the lower mean age of B/Victoria cases [171], but this explanation does not account for the higher frequency of B/Yamagata cases in middle-aged people.

Alternatively or in addition to physiological changes, differences in cohorts' susceptibility to influenza B lineages might arise from differences in cohorts' infection histories. It is unclear if people have increased protection against their first infecting influenza B lineage. One hypothesized mechanism for immune imprinting in influenza A is antibodies to conserved epitopes [60, 61]. Since B/Victoria and B/Yamagata diverged from each other more recently [139, 26, 43] and evolve antigenically much more slowly than the influenza A subtypes [14, 13, 171], conserved epitopes within a lineage might also be conserved between lineages, leading to strong protection by antibodies targeting those epitopes, regardless of which lineage was encountered first.

To investigate how protection might arise from infections with B/Victoria and B/Yamagata and contribute to differences in their age distributions, we fitted a statistical model to medically attended influenza B cases in New Zealand and Australia. The model used the estimated historical frequencies of the lineages to estimate the probabilities of different infection histories and the strength of within- and cross-lineage protection from previous infections. We found that differences in cohorts' susceptibility to each lineage are consistent with strong cross-protection between strains of the same lineage but asymmetric cross-protection between the lineages, with strong protection against B/Yamagata from B/Victoria but weak protection against B/Victoria from B/Yamagata. Similar to the immune imprinting reported for influenza A subtypes, we found additional protection against B/Yamagata in people first infected with it, but the strength of a similar effect for B/Victoria could not be estimated with confidence. The asymmetric cross-lineage protection is consistent with previous serological observations and suggests the immune response to primary B/Victoria infections focuses on epitopes shared between the lineages, whereas the response to primary infections with B/Yamagata focuses on epitopes that are not shared by B/Victoria. Characterizing individuals' antibody specificity after influenza A and B infections, and understanding how such

specificity arises from past exposure, could help explain variation in infection risk over time and the unexpected age distributions of influenza A and B cases.

4.2 Results

4.2.1 Changes in the age distributions of B/Victoria and B/Yamagata over time suggest cohort effects

To evaluate explanations for the distinct age distributions of B/Victoria and B/Yamagata, we analyzed medically attended influenza B infections detected by influenza surveillance from 2001 to 2012 in New Zealand and Australia. These data were previously analyzed by Vijaykrishna *et al.* [171] (Materials and Methods: "Case data"; Fig. 4.1B, Figs. 4.4-4.5). We focused on the New Zealand cases because they could be designated as coming from sentinel general practitioners (sentinel data, representing mild cases) or from hospital samples mostly from inpatients (non-sentinel data, representing more severe cases). Age distributions can differ between inpatients and outpatients [153]. Although Australian cases came mostly from hospital samples, the proportions coming from inpatients and from hospital clinic or emergency room outpatients were unknown.

If the age distribution of cases is strongly shaped by cohort-specific factors, such as childhood infections, the distribution of cases by birth year should change modestly over time, as each cohort retains approximately a similar risk of infection [60]. We tested this prediction using the case data from New Zealand and found that the mean birth year of cases did not change noticeably over time (95% CIs for regression coefficient: -0.14-0.39 and -0.15-0.38 for B/Victoria and B/Yamagata) (Fig. 4.6). Although some change in the mean birth year of cases may have occurred due to deaths in the elderly, births, and infection of naive children, the resulting change in the overall distribution of cases by birth year was too small to detect. If birth year better predicts infection risk than age, the age distribution of cases should be unstable over time [60], although complex multiannual dynamics can also cause

short-term fluctuations in mean age. From 2001 to 2012, the mean age of cases increased by 0.88 year each year for both B/Victoria and B/Yamagata (95% CI 0.61-1.14 for B/Victoria, 0.62-1.15 for B/Yamagata) (Fig. 4.6), more than the average annual increase in the mean age of the general population (0.16 year, Materials and Methods: "Demographic data"). These results suggest susceptibility to medically attended influenza B infections depends on cohort-related factors in addition to factors related to age itself.

4.2.2 Statistical model of influenza B infections by birth year

To test if differences in cohorts' susceptibility arise from different infection histories via immune imprinting and cross-protection within and across lineages, we fitted a statistical model to the observed cases by maximum likelihood (Materials and Methods: "Statistical model of influenza B susceptibility based on infection history"). To limit model complexity, we fitted to cases from people born since 1952, i.e., to people 60 years old or younger at the time the cases were observed. Including older people (Fig. 4.4) would have required additional parameters to describe age-related changes in susceptibility, vaccination, healthcare-seeking behavior, and contact rates, and potentially multiple ancestral strains. We found similar parameter estimates when moving the birth year cutoff five years in each direction (1947 and 1957).

The model describes the probability that a case of a particular lineage in a particular season occurs in a person born in a particular birth year, and it bases these probabilities on the following factors:

1. *Demography.* The fraction of the population born in a particular birth year gives a baseline probability that a case occurs in a person born in that year.

2. *Age-specific effects on infection risk.* Age can affect rates of contact with infectious people and susceptibility to infection [117, 114]. We assumed that preschoolers (0-5 years old), school-age children and teenagers (6-17) and people 18 and older had

different baseline probabilities of infection.

3. *Age-specific effects on reporting.* We assumed infections in children 0-4 years old had a different probability of receiving medical attention and being reported than infections in the rest of the population (estimates of protection were similar when we used an alternative interval of 0-2 years old).

4. *Infection history.* The probability of observing a case of a lineage in a birth cohort could be modified by the average susceptibility of people in that cohort to that lineage, based on the cohort's infection history. Susceptibility is defined as the probability of infection given exposure, relative to that of a naive person. We estimated the probabilities of different infection histories for each cohort using a discrete-time model in which the probability of becoming infected with either lineage depends on the lineages' historical frequencies (Materials and Methods: "Infection history probabilities"), which we estimated using sequence databases (Materials and Methods: "Historical frequencies of influenza B lineages"). Consistent with trends observed in other countries [144] (Fig. 4.7), we found that B/Yamagata was the dominant lineage in New Zealand in the 1990s. B/Victoria started circulating at high frequencies in the early 2000s, and the two lineages have circulated with alternating frequencies since (Fig. 4.1A).

   We examined three effects of past infection (Table 4.1):

   (a) *Within-lineage protection.* Any previous infection with B/Victoria or B/Yamagata decreases susceptibility to future infections with the same lineage by fractions χ_{VV} and χ_{YY}, respectively.

   (b) *Cross-lineage protection* Cross-lineage protection against a lineage was estimated as a fraction (γ) of the corresponding within-lineage protection against that lineage: $\chi_{VY} = \gamma_{VY} \times \chi_{YY}$ and $\chi_{YV} = \gamma_{YV} \times \chi_{VV}$, where χ_{VY} is the protection from B/Victoria against B/Yamagata and χ_{YV} is the protection from B/Yamagata against B/Victoria. We assumed that once a person was infected

with a lineage, within-lineage protection superseded protection from previous infection with the other lineage.

(c) *Lineage-specific imprinting.* To represent increased protection against the lineage of first infection, susceptibility to B/Victoria and B/Yamagata could be further reduced by R_V or R_Y if a person was first infected with the corresponding lineage.

(d) *Infection with influenza B strains before 1988.* Because sequence data were too scarce before 1988 to reliably estimate the frequencies of B/Victoria and B/Yamagata, we treated all infections before 1988 as infections with a separate "ancestral" lineage A and estimated protection from those infections against B/Victoria (χ_{AV}) and B/Yamagata (χ_{AY}). Those infections encompass the ancestral influenza B lineage before the split between B/Victoria and B/Yamagata and also strains circulating between the split and 1988. An infection with A, if it occurred, was necessarily a person's first infection. We modeled those infections to capture the middle part of the age distributions of cases, but because of the uncertain identity and antigenic phenotype of strains circulating in that period, we interpreted the associated parameters with caution.

We assumed that the protection conferred by an infection against future infections depends on lineage but not on the time between infections. Consistent with this assumption, average genetic divergence between lineages was greater than evolution within each lineage over the study period (Fig. 4.8; Materials and Methods: "Sequence divergence analysis"). By 2012, the last year in the case data, the amino acid divergence of the lineages' HAs was approximately 7% from their B/Victoria and B/Yamagata founders in the late 1980s (Fig. 4.8), whereas divergence between the lineages in 2012 was approximately 14%. Assuming cross-protection from past exposures does not decay appreciably over time allowed us to calculate infection history probabilities exactly without the need for dynamical simulations (Materials and Methods: "Infection history probabilities").

Table 4.1: Possible infection histories in terms of the lineage of first infection and lineages encountered since.

1st infection	Lineages encountered later	Symbol	Susceptibility to B/Vic	Susceptibility to B/Yam
None	None	P_0	1	1
V	None	$P_{V,0}$	$(1-\chi_{VV})(1-R_V)$	$(1-\chi_{VY})$
V	Y	$P_{V,Y}$	$(1-\chi_{VV})(1-R_V)$	$(1-\chi_{YY})$
Y	None	$P_{Y,0}$	$1-\chi_{YV}$	$(1-\chi_{YY})(1-R_Y)$
Y	V	$P_{Y,V}$	$1-\chi_{VV}$	$(1-\chi_{YY})(1-R_Y)$
A	None	$P_{A,0,0}$	$1-\chi_{AV}$	$1-\chi_{AY}$
A	V	$P_{A,V,0}$	$1-\chi_{VV}$	$1-\max(\chi_{AY},\chi_{VY})$
A	Y	$P_{A,Y,0}$	$1-\max(\chi_{AV},\chi_{YV})$	$1-\chi_{YY}$
A	V and Y (any order)	$P_{A,\{VY\}}$	$1-\chi_{VV}$	$1-\chi_{YY}$

V indicates B/Victoria, Y indicates B/Yamagata, and A indicates strains circulating before 1988. Susceptibility to each lineage depends on within-lineage (χ_{VV}, χ_{YY}) and cross-lineage (χ_{VY}, χ_{YV}, χ_{AV}, χ_{AY}) cross-protection from prior infections regardless of their order and on additional protection against the lineage first encountered (R_V and R_Y). Only the strongest cross-lineage protection term is assumed to affect susceptibility. Within-lineage protection is constrained to be stronger than cross-lineage protection against the same lineage.

Influenza vaccine coverage was low in New Zealand during the study period, except in the elderly (5% or lower for ages 0-49, 14 % for ages 50-64 and 64% for ages 65 and older in 2012 [107]), and we thus ignored protection by vaccination.

4.2.3 Infection probabilities estimated by the model are consistent with estimates from independent serological data

To test if our model produced realistic estimates of infection risk, we compared estimates from the model fitted to the New Zealand case data with estimates based on cross-sectional serology from children in the Netherlands [17] (Materials and Methods: "Model validation with independent serological data"). The annual probability of influenza B exposure (which, for naive people, is equal to the probability of infection) for preschoolers was very similar between our model (β_1 = 12%, 95% CI 9-15%, Table 4.2) and that inferred from Dutch seroprevalence data (12%, 95% CI 10-14%). These estimates are also within the range of infection probabilities estimated from longitudinal studies [116, 70, 66, 72]. However, the estimated exposure probability for school-age children (6-17 years old) was significantly lower in our model than in the serological study (β_2 = 14%, 95% CI 11-17% vs 22%, 95% CI 16-

Table 4.2: Parameter estimates for the model fitted to the distribution of mild (sentinel) and severe (non-sentinel) cases in New Zealand.

Parameter	MLE (95% CI)	Definition
β_1	0.12 (0.09-0.15)	Annual exposure probability for preschoolers (0-5 years old)
β_2	0.14 (0.11-0.17)	Annual exposure probability for people 6-17 years old
β_3	0.16 (0.11-0.22)	Annual exposure probability for people 18+ years old
χ_{VV}	0.93 (0.91-0.97)	Protection against B/Vic from any prior B/Vic infection
χ_{YY}	0.83 (0.64-0.92)	Protection against B/Yam from any prior B/Yam infection
γ_{VY}	1.00 (0.83-1.00)	Protection from B/Vic infection against B/Yam (as a fraction of χ_{YY})†
γ_{YV}	0.00 (0.00-0.19)	Protection from B/Yam infection against B/Vic (as a fraction of χ_{VV})†
R_V	0.00 (0.00-0.95)	Additional B/Vic protection if 1st infection was with B/Vic
R_Y	1.00 (0.62-100)	Additional B/Yam protection if 1st infection was with B/Yam
γ_{AV}	1.00 (0.97-1.00)	Protection against B/Vic from pre-1988 infections (as a fraction of χ_{VV})†
γ_{AY}	0.45 (0.00-1.00)	Protection against B/Yam from pre-1988 infections (as a fraction of χ_{YY})†
ρ	0.90 (0.69-1.18)	Reporting factor for children 0-4

†In the model, cross-lineage protection does not apply when within-lineage protection is present.

27%), perhaps because only children up to 7 years old were represented in the serological data. The fraction of children with detectable antibodies against B/Victoria was close to the prediction from our model, but more children had detectable B/Yamagata antibodies than predicted by the model (Fig. 4.9). This discrepancy is consistent with the presence of antibodies from B/Victoria infections that cross-react with B/Yamagata but not vice-versa [139, 100, 101, 148, 96].

The infection probabilities estimated from our model and from cross-sectional serology represent an average probability across influenza seasons. Differences between estimates from our model and those based on serology might reflect differences in the intensity of influenza B circulation or in lineage frequencies between New Zealand and the Netherlands in the periods represented in each dataset.

4.2.4 Evidence for asymmetric cross-lineage protection and immune imprinting

We first fitted the model to the complete New Zealand data to estimate protection against medically attended influenza B infections in general, including mild (general practice) and more severe (hospital-associated) cases. The distributions of B/Victoria and B/Yamagata

are most consistent with strong within-lineage protection, asymmetric cross-lineage protection, and imprinting protection against B/Yamagata (Table 4.2, Fig. 4.10). Any previous B/Victoria infection decreased the probability of medically attended infection with B/Victoria by 93% (χ_{VV} = 93%, 95% CI 91-97%), and any previous B/Yamagata infection decreased the probability of future medically attended B/Yamagata infection by 83% (χ_{YY} = 83%, 95% CI 64-92%). Protection against B/Yamagata after B/Victoria infections appeared as strong as protection after B/Yamagata infection itself (γ_{VY} = 100%, 95% CI 83-100%), corresponding to an 83% reduction in susceptibility to B/Yamagata in people infected only with B/Victoria (χ_{VY} = 83%, 95% CI 68-83%). However, B/Yamagata infections provided little to no protection against medically attended B/Victoria infections (γ_{YV} = 0, 95% CI 0-19%; χ_{VY} = 0, 95% CI 0-18%). People for whom B/Yamagata was the lineage of first infection had an additional 62-100% reduction in susceptibility against medically attended B/Yamagata infection compared with people infected with B/Yamagata after a primary B/Victoria infection (R_Y = 100%, 95% CI 62-100%). However, the strength of imprinting for B/Victoria was effectively non-identifiable (R_V = 0%, 95% CI 0-95%).

While the data support both asymmetric cross-lineage protection and imprinting protection against B/Yamagata, imprinting has a modest impact on the shape of the age distribution of B/Yamagata cases compared with the effect of asymmetric cross-lineage protection. Fitting the model while constraining B/Yamagata to be strongly protective against B/Victoria (setting γ_{YV} = 0.9), or constraining B/Victoria to give no protection against B/Yamagata (γ_{VY} = 0), underestimated cases of B/Victoria and overestimated cases of B/Yamagata in cohorts born around 1990 (Δ_{AIC} = 105 and 20, respectively; Figs. 4.11 and 4.12). In contrast, fitted without imprinting protection against B/Yamagata (R_Y = 0), the model could still capture the low incidence of B/Yamagata in cohorts born around 1990, albeit with a poorer fit (Δ_{AIC} = 9, Fig. 4.13).

The model therefore explains the lower incidence of B/Yamagata compared to B/Victoria among people born in the late 1980s and early 1990s primarily via strong within-lineage

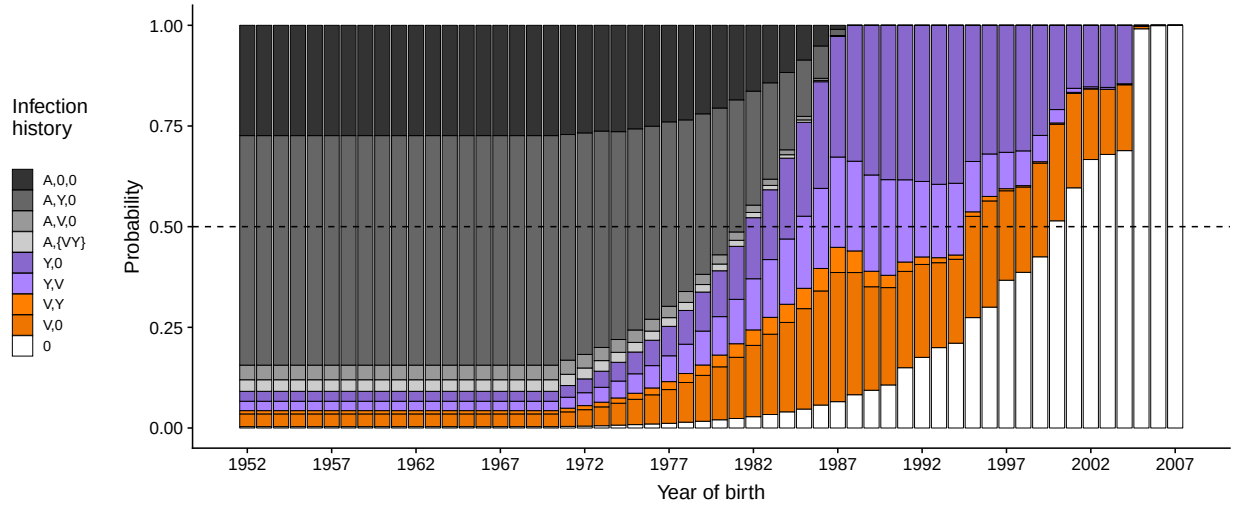

Figure 4.2: Probabilities of different infection histories with influenza B in New Zealand for people born between 1952 and 2007 and observed in 2007. Infection histories consist of the lineage of first infection and lineages encountered later regardless of their order. (A,0,0): First infection before 1988 and no subsequent infections with either B/Victoria or B/Yamagata. (A,Y,0) and (A,V,0): First infection before 1988 followed by B/Yamagata but not B/Victoria and by B/Victoria but not B/Yamagata, respectively. (A,{V,Y}): First infection before 1988 followed by infections with both B/Victoria and B/Yamagata in any order. (Y,V) and (Y,0): First infection with B/Yamagata, with and without a subsequent B/Victoria infection. (V,Y) and (V,0): First infection with B/Victoria, with and without a subsequent B/Yamagata infection. (0): fully naive to influenza B.

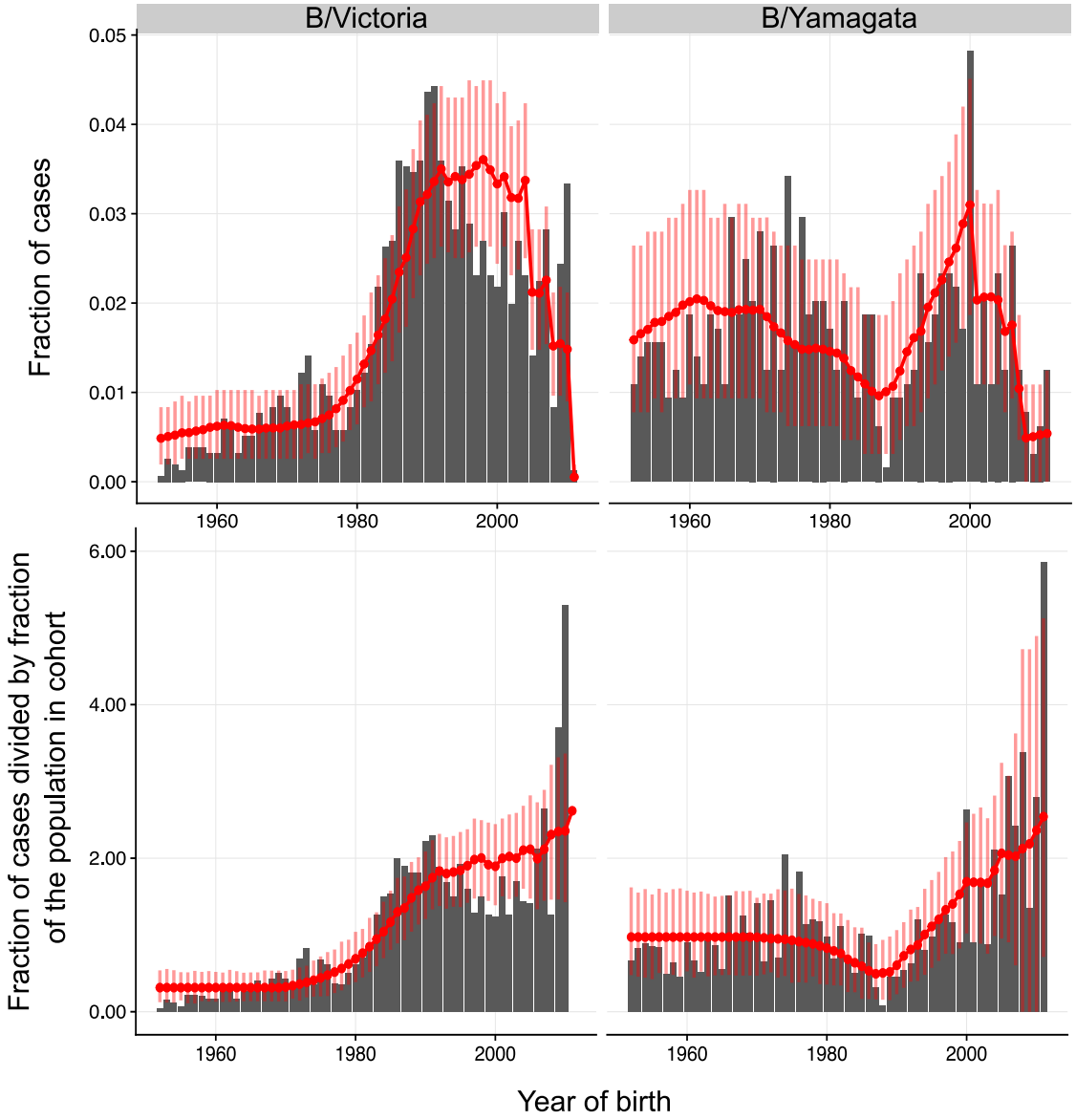

Figure 4.3: Observed and predicted distributions of mild and severe influenza B cases in New Zealand by birth year. The model was simultaneously fitted to the age distributions in each observation year from 2001 to 2012, accounting for uncertainty in the birth year of each reported case given the patient's age. For plotting, we pooled observed and predicted numbers of cases across observation years for each birth year, assuming the earliest possible birth year for each age (e.g., an age of 10 in 2000 was assumed to correspond to the birth year 1989). Red lines and dots show the predicted distribution under the model. Vertical bars are 95% bootstrap confidence intervals. In the bottom row, predicted and observed fractions of cases were normalized by dividing by the fraction of the population born in that birth year (i.e., the null expectation if all birth years were infected at the same rate). Cases in people born in 2011 (with wide bootstrap CIs) were omitted from the bottom row to improve visualization but are shown in Fig 4.14.

protection and asymmetric cross-protection between the lineages. By 2007 (the midpoint of the surveillance period), 62% of people born between 1987 and 1993 had been infected with B/Yamagata (of which 94% had B/Yamagata as their first influenza B infection, Fig. 4.2). Another 26% had never been infected with B/Yamagata but had been infected with B/Victoria and therefore had cross-lineage protection against B/Yamagata. Thus, the model estimates that 88% of people born in 1987-1993 were protected against medically attended infection with B/Yamagata in 2007. In contrast, only the 51% of people who had been previously infected with B/Victoria had any protection against B/Victoria, since people infected only with B/Yamagata had little to no protection against B/Victoria. As a result, average susceptibility to B/Victoria in cohorts born in 1987-1993 was three times higher than susceptibility to B/Yamagata (53% versus 18% that of a naive person) (Fig. 4.3).

The model estimates that infections with strains circulating before 1988 provided strong protection against medically attended B/Victoria infections in 2001-2012 (χ_{AV} = 93.3%, 95% CI 93.2-93.3%) and less protection against B/Yamagata (χ_{AY} = 38%, 95% CI 0-67%) (Table 4.2). This result suggests that the ancestral strains that preceded the lineage split were antigenically more similar to B/Victoria than to B/Yamagata. However, because the protection from the ancestral strains against B/Yamagata was weaker than the protection from B/Victoria against B/Yamagata, ancestral strains also seem to have been antigenically farther from B/Yamagata than is B/Victoria. Because we grouped infections before the lineages split with infections that occurred between the split and 1988, it is possible that the stronger protection estimated against B/Victoria derives from higher incidence of B/Victoria during that period (which we are unable to detect).

Together, protection against B/Victoria from early strains, cross-lineage protection against B/Yamagata from B/Victoria, and potentially stronger imprinting from B/Yamagata infections can explain why B/Yamagata cases in 2001-2012 were less common in people born around 1990 than in older people. We estimated that susceptibility to B/Yamagata was 47% higher on average in people born before 1986 than in people born between 1987 and

1993 (Fig. 4.3). Most people born before 1986 (78%) were first infected with strains circulating before 1988 (Fig. 4.2). Although most (64%) of those people had been infected with B/Yamagata by 2007, they lacked the imprinting protection of people born in the late 1980s and early 1990s. Most people born before 1986 also lacked the cross-protection from B/Victoria infections against B/Yamagata, since infection with strains circulating before 1988 prevented B/Victoria infections in most (80%) of them.

These results were qualitatively robust to underestimation of sentinel cases in New Zealand (Materials and Methods: "Estimating missing sentinel cases and cases with missing lineage information"; Fig. 4.15).

4.2.5 Strong within-lineage protection dwarfs the effects of imprinting on severe cases

We next fitted the model separately to general practice and hospital-associated cases to investigate protection against cases likely to be relatively mild or severe, respectively. General practice cases correspond to the sentinel branch of the New Zealand data, while hospital-associated cases include the non-sentinel branch of the New Zealand data and cases from Australia. Although the proportions of inpatients and outpatients in the Australian data are unknown, the distribution of cases by birth year in Australia was more strongly correlated with the birth year distribution of non-sentinel New Zealand cases, which came mostly from hospital inpatients (B/Victoria: $r = 0.76$, 95% CI 0.65-0.84; B/Yamagata: $r = 0.53$, 95% CI 0.36-0.66), than with the distribution of New Zealand sentinel cases, which were recorded in general practice outpatients (B/Victoria: $r = 0.36$, 95% CI 0.14-0.54; B/Yamagata: $r =$ -0.18, 95% CI -0.4--0.06) (Fig. 4.16). This result suggests cases observed in Australia were similar in severity to those observed in the New Zealand non-sentinel data.

When fitting the model to general practice cases alone, we found estimates of protection against B/Victoria similar to those obtained for the complete New Zealand data, with strong within-lineage protection and little to no protection from B/Yamagata (Figs. 4.17 and 4.18).

The few B/Yamagata cases did not allow protection against B/Yamagata to be estimated with confidence from general practice cases alone, but these data were still most consistent with strong cross-lineage protection from B/Victoria against B/Yamagata (Fig. 4.17).

The estimated strength of within-lineage protection against hospital-associated B/Yamagata cases was higher than when fitting to general practice and hospital-associated cases combined. Infection with B/Yamagata reduced susceptibility to hospital-associated B/Yamagata infections by 70-100% in New Zealand and by 89-100% in Australia (maximum-likelihood estimates 89% and 94%, Figs. 4.19-4.22) compared to 64-92% protection estimated against general practice and hospital-associated B/Yamagata cases together. The strong within-lineage protection limited the potential for immune imprinting to reduce susceptibility to B/Yamagata further, making the strength of imprinting protection unidentifiable (95% CI for R_Y = 0-100% for both New Zealand and Australia hospital-associated cases, Fig. 4.23).

4.2.6 Evidence of protection from B/Yamagata against hospital-associated B/Victoria cases

While most estimates of protection were similar when fitting the model to hospital-associated cases from New Zealand and Australia, protection from B/Yamagata against B/Victoria was stronger when estimated from Australia (χ_{YV} = 77%, 95% CI 67-82%) than from New Zealand (χ_{YV} = 0%, 95% CI 0-36%). This difference was associated with a change in the birth year distribution of hospital-associated B/Victoria cases over time in New Zealand that was not evident in cases from Australia (Figs. 4.24 and 4.25). Hospital-associated B/Victoria cases from New Zealand had a peak in people born around 1990 during the 2005 season which decreased in subsequent seasons and was never as pronounced in Australia. Removing hospital-associated B/Victoria cases observed in New Zealand in 2005 led to a higher estimate of protection from B/Yamagata against B/Victoria (43%, 95% CI 1-67%), suggesting the lack of protection from B/Yamagata against severe B/Victoria cases in New Zealand might be due to the model's inability to fit this change in the birth year distribution of hospital-

associated B/Victoria cases. Thus, while B/Yamagata infections appear to provide weak protection against medically attended B/Victoria infections overall, B/Yamagata infections might reduce susceptibility to severe B/Victoria cases. However, estimates from cases in Australia suggest protection from B/Yamagata against severe B/Victoria cases is still weaker than protection from B/Victoria against B/Yamagata (67-82% versus 85-93%).

While we do not know why the 2005 season in New Zealand had proportionally more hospital-associated B/Victoria cases in people born in the early 1990s than other seasons, we estimate that 2005 had the most intense influenza B epidemic in New Zealand since 1995, almost 3 times as intense as the average season (we defined intensity as the product of the incidence of medically attended influenza-like illness [ILI] and the fraction of ILI cases positive for influenza B; Methods: "Intensity Scores"). In contrast, Australia had an influenza B epidemic of approximately average intensity in 2005. B/Victoria cases attended by New Zealand general practitioners typically peaked in those same cohorts born in the early 1990s, suggesting that in 2005 some B/Victoria infections that in other seasons would have been treated by general practitioners were treated at a hospital. The 2005 epidemic was also the first major B/Victoria epidemic in New Zealand since B/Victoria acquired its neuraminidase gene from B/Yamagata 2000-2001 [94] (a smaller epidemic occurred in 2002). Because of this reassortment, children born in the late 1980s and early 1990s might have lost some of the within-lineage protection they had against B/Victoria, leading to an increase in severe cases in those cohorts in 2005.

4.2.7 Severe cases in young children can explain discrepancies from model predictions in the most recent birth years

The main discrepancy between the data and the model predictions was an excess of B/Victoria cases in the most recent birth years (2009-2012). This excess was driven by severe B/Victoria cases in children 0-2 years old. The model generally underestimated the number of cases in those children across observation years (Fig. 4.24), perhaps because we used a single

parameter to describe the reporting probability for children 0-4 years-old relative to the rest of the population (ρ = 0.90, 95% CI 0.69-1.18). Because 2011 had many B/Victoria cases, the excess of cases in young children in that season led to an excess of cases in recent birth years in the distribution combined across observation years. Using an alternative 2 year-old cutoff for differential reporting led to significantly increased reporting in young children (ρ = 2.25, 95% CI 1.92-2.77), similar estimates of protection and a better prediction of the number of cases in people born in 2009-2012 (Figs. 4.26 and 4.27).

4.3 Discussion

Differences in the age distributions of the influenza B lineages are consistent with strong within- and cross-lineage protection from previous infections. The data suggest that any previous infection with B/Victoria or B/Yamagata strongly protects against future medically attended infections with the same lineage. However, cross-protection between the lineages is asymmetric: B/Victoria infections strongly protect against B/Yamagata, but protection from B/Yamagata infections appears weaker and is potentially limited to severe cases. The data are consistent with some imprinting protection against B/Yamagata but are agnostic about the strength of imprinting for B/Victoria.

High protection within and across lineages is not unexpected. The major antigens HA and NA evolve more slowly within the influenza B lineages than within influenza A subtypes [14, 13, 171], and their amino acid sequence divergences are much lower between B/Victoria and B/Yamagata ($\approx$14%, 2% and 7% for the HA head, the HA stalk, and NA, respectively) than between H1N1 and H3N2 ($\approx$66%, 50% and 60%, respectively) (Fig. 4.28). HA and NA epitopes shared across lineages might provide the basis for cross-protection between them, while epitopes conserved within the lineages but variable across them might be the basis for imprinting protection against B/Yamagata and potentially B/Victoria, as has been hypothesized for influenza A subtypes [61].

Our analysis is limited by the scarcity of data on the antigenic phenotype of influenza

B strains circulating prior to the lineage split and on lineage frequencies shortly after. Our results suggest influenza B strains prior to the lineage split were antigenically more similar to B/Victoria than to B/Yamagata, although they appear to have been less protective against B/Yamagata strains than were later B/Victoria strains. Consistent with this result, the study that first identified the existence of two lineages in the late 1980s found that antibodies induced by B/Victoria strains from that period were more cross-reactive than those induced by B/Yamagata strains against an influenza B strain from the early 1980s [139]. However, B/Victoria appears to have since undergone faster antigenic evolution than B/Yamagata [14, 171], such that more recent B/Yamagata strains might be closer to the ancestor.

The asymmetric cross-lineage protection estimated between B/Victoria and B/Yamagata is consistent with serological studies of strains from the late 1980s, 2000s and 2010s. In mice, ferrets, sheep and children, primary exposure to B/Victoria induces antibodies that inhibit hemagglutination by B/Yamagata strains, whereas exposure to B/Yamagata often induces low or undetectable levels of antibodies cross-reactive to B/Victoria [139, 100, 101, 148, 96]. Together, these observations and our results suggest a model in which the immune response to primary infections with each lineage is focused on different epitopes with different degrees of conservation between the lineages. In response to a primary infection with B/Victoria, antibodies might predominantly target epitopes that are relatively conserved between the lineages, consistent with our estimate of strong protection from B/Victoria against B/Yamagata and with the ability of B/Victoria-induced antibodies to bind B/Yamagata. If the primary infection is with B/Yamagata, however, the response might focus on a different set of epitopes that are less conserved between the lineages, consistent with our estimate of weak protection from B/Yamagata against B/Victoria and the poor binding of B/Victoria strains by B/Yamagata-induced antibodies. This proposed model could be tested by identifying which epitopes are preferentially targeted upon primary infection with each lineage, for instance by measuring antibody binding to mutant viruses sharing a single epitope each with the infecting strain [9].

While most studies of the antibody response to influenza focus on HA, our estimates of within and cross-lineage protection may also reflect protection by antibodies targeting NA. Serological evidence suggests immune responses targeting NA instead of HA are more common against influenza B than against influenza A, especially in children [72]. In 2000-2001, B/Victoria viruses acquired an NA gene by reassortment from a B/Yamagata clade [94]. Although all B/Victoria cases in the data occurred after this reassortment, our results suggest infections with B/Yamagata did not induce significant cross-lineage protection based on the newly acquired NA. Another potential consequence of the reassortment might have been to increase the cross-lineage protection from B/Victoria infections against B/Yamagata. Since we did not include this effect in our model (due to the increased complexity of infection histories and the difficulty of calculating their probabilities), our estimate of protection from B/Victoria against B/Yamagata may represent an average of protection before and after the acquisition of a B/Yamagata NA by B/Victoria strains.

Although our results are consistent with weaker protection from B/Yamagata infections against B/Victoria than vice-versa, protection against B/Victoria by trivalent seasonal vaccines containing B/Yamagata has been reported. Meta-analyses suggest lineage-mismatched trivalent seasonal vaccines were effective on average [166], but few studies report lineage-specific estimates of vaccine effectiveness that could reveal asymmetric cross-lineage protection. A test-negative design study estimated that inactivated vaccines containing B/Yamagata had over 50% effectiveness against B/Victoria infections, and vice-versa, in a population of mostly adults [146]. However, a meta-analysis of randomized controlled trials of live attenuated vaccines in children estimated that the efficacy of vaccines containing B/Yamagata in seasons dominated by B/Victoria was low to nonexistent (B/Yamagata seasons in which a B/Victoria vaccine was used were not available) [15].

It remains unclear precisely how differences in people's antibody responses and past infections shape their susceptibility to influenza. Antibody targeting of NA and of particular HA epitopes changes with age [118, 129, 157], but few studies have linked antibody specificity

to the infection history of particular cohorts and their susceptibility to particular variants [105, 8, 127]. Characterizing antibody specificity after infection and vaccination might reveal the origins of differences in the antigens and epitopes targeted by different people and cohorts. This resolution could improve understanding of influenza epidemiology and the response to influenza vaccination.

4.4 Materials and methods

4.4.1 Case data

Medically attended influenza B cases detected by influenza surveillance in New Zealand and Australia were sent to the World Health Organization (WHO) Collaborating Centre for Reference and Research on Influenza in Melbourne, Australia, and were previously compiled and analyzed by Vijaykrishna *et al.* [171]. New Zealand cases were identified from samples taken from patients with influenza-like illness attended by a network of sentinel general practitioners and from non-sentinel hospital samples analyzed by regional diagnostic laboratories and by the WHO National Influenza Centre at the Institute for Environmental Science and Research [3]. Cases from Australia were mostly identified from hospital samples and may include both severe and mild infections, as the fees charged for some general practice consultations may encourage patients to seek emergency care at a hospital [69].

4.4.2 Statistical model of influenza B susceptibility based on infection history

For lineage V (B/Victoria), we modeled the number of cases in people born in birth year b observed in season y as a multinomial draw with probabilities given by:

$$\theta_V(b,y) = D(b,y)\beta(b,y)Z_V(b,y)\rho(b,y) \tag{4.1}$$

with an analogous equation defining the multinomial distribution $\theta_Y(b, y)$ for lineage Y (B/Yamagata). $D(b, y)$ is the country-specific fraction of the population that was born in year b as of observation season y. $Z(b, y)$ is the susceptibility to lineage V during season y of a person born in year b relative to that of an unexposed person. $\beta(b, y)$ is a baseline probability of infection with influenza B that captures differences in transmission associated with age (thus depending on b and y) and is equal to β_1 if people born in year b are in preschool during season y (0-5 years old), β_2 if they are school-age children or teenagers (6-17 years old), or β_3 if they are 18 or older. $\rho(b, y)$ is an age-specific reporting rate equal to a country-specific parameter ρ if people are less than 5 years old and 1 otherwise. While $\beta(b, y)$ represents true differences in infection probabilities between preschoolers and post-preschool individuals and thus affects the infection history probabilities in the calculation of $Z(b, y)$, $\rho(b, y)$ represents differential reporting in children and does not affect those probabilities. For each lineage and observation season, values of θ from Eq. (4.1) are normalized by their sum across birth years to make them proper multinomial probabilities.

We defined relative susceptibility to V, $Z_V(b, y)$, as an expectation over all possible immune histories in terms of the lineage of first infection and subsequent infection with the other lineage. We let susceptibility be 1 for a person never exposed to influenza B. Cross-immunity from any previously encountered strain of V or Y decreased susceptibility to the corresponding lineage by χ_{VV} and χ_{YY}, respectively. Susceptibility was further reduced by R_V or R_Y if the lineage of first infection was V or Y. In the absence of a previous homologous infection, previous infection with Y decreased susceptibility to V by χ_{YV}, and previous infection to V decreased susceptibility to Y by χ_{YV}. Protection due to homologous infections superseded cross-lineage protection. Protection against V from a previous Y infection was constrained to be a fraction of the within-lineage protection to V (and vice-versa): $\chi_{YV} = \chi_{VV} \cdot \gamma_{YV}$, $\chi_{VY} = \chi_{YY} \cdot \gamma_{VY}$, where $0 \leq \gamma_{YV}, \gamma_{VY} \leq 1$. Similarly, pre-1988 infections reduced susceptibility to V by $\chi_{AV} = \chi_{VV} \cdot \gamma_{AV}$ and to Y by $\chi_{AY} = \chi_{YY} \cdot \gamma_{AY}$. Finally, $Z_V(b, y)$ was calculated as the sum of susceptibilities in Table

1 weighted by the probabilities of the corresponding infection histories (below). Relative susceptibility to Y, $Z_Y(b, y)$ was defined analogously.

We estimated parameters by maximum likelihood using R (version 3.4.3) and package optimParallel. We calculated the total likelihood as the product of the likelihood for each lineage in each observation year, with the number of cases in each combination treated as an independent multinomial draw. For plotting, we summed the observed and predicted cases for each observation year.

Code implementing the analyses and figures is available at `https://github.com/cobeylab/influenza-B`.

4.4.3 Infection history probabilities

To calculate the probabilities in Table 4.1, we assumed infections occur in discrete time measured in units of annual influenza seasons. We considered possible infections in each season between birth and the last season before observation season y. For a person born in year b and previously unexposed to influenza B, we let $a_{b,i}$ be the probability of becoming infected in season i. This probability was then modified by protection from previous infections as in Table 4.1. Given that an infection occurred in season i, we assumed that the probability it was caused by A, V or Y was equal to their frequencies in that season, $f_{A,i}$, $f_{V,i}$ and $f_{Y,i}$, with $f_{A,i} = 1$ for all i before 1988, and $f_{V,i} + f_{Y,i} = 1$ since. For simplicity, we assumed that people could not be infected more than once in each season (including simultaneous infections by the two lineages.)

Let $\Phi^B_{i,j}$ be the probability that no infections with influenza B occurred for a naive person born in b from seasons i to j (inclusive). It is given by:

$$\Phi^B_{i,j} = \prod_{k=i}^{j}(1 - a_{b,k}) \tag{4.2}$$

where k indexes years (influenza seasons). Thus the first probability in Table (4.1), that of

being fully naive to influenza B, is given by:

$$P_0(b, y) = \Phi^B_{b,y-1} \tag{4.3}$$

To shorten the expressions for the remaining probabilities, we let $\Phi^V_{i,j}(h)$, $\Phi^Y_{i,j}(h)$ and $\Phi^{VY}_{i,j}(h)$ be the probability that no V infections, no Y infections, and neither V nor Y infections occurred between seasons i and j (inclusive), respectively. Unlike Φ^B, which applies to naive people, Φ^V, Φ^Y and Φ^{VY} depend on the person's infection history, h. To calculate the probability of an initial infection with A but no subsequent infections with V or Y, $P_{A,0,0}(b, y)$, we integrated, across all seasons i of first infection, the joint probability that the person's first influenza B infection occurred in season i with the ancestral lineage A and that no subsequent infections with V or Y occurred from $i + 1$ to $y - 1$:

$$P_{A,0,0}(b, y) = \sum_{i=b}^{y-1} \left[\Phi^B_{b,i-1} \cdot a_{b,i} \cdot f_{A,i} \cdot \Phi^{VY}_{i+1,y-1}(A) \right] \tag{4.4}$$

where the probability that the first infection to influenza B occurred in season i with the ancestor A is obtained by multiplying the probability of no previous infections ($\Phi^B_{b,i-1}$) by the probability of an infection in season i ($a_{b,i}$) and by the frequency of lineage A in season i ($f_{A,i}$). The probability of no infections with either V or Y after the initial infection with A in season i, $\Phi^{VY}_{i+1,y-1}(A)$, depends on protection from the A infection against V (χ_{AV}) and Y (χ_{AY}):

$$\Phi^{VY}_{i+1,y-1}(A) = \prod_{k=i+1}^{y-1} \{1 - a_{b,k}[f_{V,k}(1 - \chi_{AV}) + f_{Y,k}(1 - \chi_{AY})]\} \tag{4.5}$$

Similarly, the joint probability of first infection with the ancestor A and subsequent infection

with V, but not with Y, is given by:

$$P_{A,V,0}(b,y) = \sum_{i=b}^{y-1}\left[\Phi^{B}_{b,i-1}a_{b,i}f_{A,i}\cdot\sum_{j=i+1}^{y-1}\Phi^{VY}_{i+1,j-1}(A)a_{b,j}f_{V,j}(1-\chi_{AV})\Phi^{Y}_{j+1,y+1}(A,V)\right] \tag{4.6}$$

where we again integrated over all possible seasons i when the first infection with influenza B occurred. Given the first infection occurred in season i with the "ancestral" lineage A, we calculated the probability of subsequent infection with V, but not with Y, by integrating over all possible seasons j when the first infection with V may have occurred. Given the initial infection with A in season i, the joint probability that the first V infection occurred in j is given by the probability that neither V nor Y infections occurred from $i+1$ to $j-1$, $\Phi^{VY}_{i+1,j-1}(A)$, times the probability of a V infection in season j, given by $a_{b,j}f_{V,j}(1-\chi_{AV})$. The probability $\Phi^{Y}_{j+1,y-1}(A,V)$ of no subsequent Y infections after season j given the previous A and V infections is then given by:

$$\Phi^{Y}_{j+1,y-1}(A,V) = \prod_{k=j+1}^{y-1}\{1-a_{b,k}f_{Y,k}[1-\max(\chi_{AY},\chi_{VY})]\} \tag{4.7}$$

where we assumed that only the strongest cross-protection (from the previous A and V infections) applies. By analogy, for $P_{A,Y,0}(b,y)$ we have:

$$P_{A,Y,0}(b,y) = \sum_{i=b}^{y-1}\left[\Phi^{B}_{b,i-1}a_{b,i}f_{A,i}\cdot\sum_{j=i+1}^{y-1}\Phi^{VY}_{i+1,j-1}(A)a_{b,j}f_{Y,j}(1-\chi_{AY})\Phi^{V}_{j+1,y+1}(A,Y)\right] \tag{4.8}$$

where

$$\Phi^{V}_{j+1,y-1}(A,Y) = \prod_{k=j+1}^{y-1} \{1 - a_{b,k} f_{V,k}[1 - \max(\chi_{AV}, \chi_{YV})]\} \tag{4.9}$$

To calculate $P_{A,\{VY\}}(b,y)$, we first computed the probabilities for the particular cases where either V or Y were the second infection, $P_{A,V\to Y}(b,y)$ and $P_{A,Y\to V}(b,y)$, such that $P_{A,\{VY\}}(b,y)$ is the sum of the two. The first is given by:

$$P_{A,V\to Y}(b,y) = \sum_{i=b}^{y-1} \left[\Phi^{B}_{b,i-1} a_{b,i} f_{A,i} \cdot \sum_{j=i+1}^{y-1} \Phi^{VY}_{i+1,j-1}(A) a_{b,j} f_{V,j} (1 - \chi_{AV})[1 - \Phi^{Y}_{j+1,y+1}(A,V)] \right] \tag{4.10}$$

and the second is given by:

$$P_{A,Y\to V}(b,y) = \sum_{i=b}^{y-1} \left[\Phi^{B}_{b,i-1} a_{b,i} f_{A,i} \cdot \sum_{j=i+1}^{y-1} \Phi^{VY}_{i+1,j-1}(A) a_{b,j} f_{Y,j} (1 - \chi_{AY})[1 - \Phi^{V}_{j+1,y+1}(A,Y)] \right] \tag{4.11}$$

Next we write down the probabilities of infection histories with either V or Y as first infections and no subsequent infections. For $P_{V,0}(b,y)$:

$$P_{V,0}(b,y) = \sum_{i=b}^{y-1} \Phi^{B}_{b,i-1} \cdot a_{b,i} \cdot f_{V,i} \cdot \Phi^{Y}_{i+1,y-1}(V) \tag{4.12}$$

where

$$\Phi^{Y}_{i+1,y-1}(V) = \prod_{k=i+1}^{y-1} [1 - a_{b,k} f_{Y,k}(1 - \chi_{VY})] \tag{4.13}$$

By analogy, for $P_{Y,0}(b,y)$:

$$P_{Y,0}(b,y) = \sum_{i=b}^{y-1} \Phi^{B}_{b,i-1} \cdot a_{b,i} \cdot f_{Y,i} \cdot \Phi^{V}_{i+1,y-1}(Y) \tag{4.14}$$

where

$$\Phi^{V}_{i+1,y-1}(Y) = \prod_{k=i+1}^{y-1} [1 - a_{b,k} f_{V,k} (1 - \chi_{YV})] \tag{4.15}$$

Finally, $P_{V,Y}(b,y)$ and $P_{Y,V}(b,y)$ are given by:

$$P_{V,Y}(b,y) = \sum_{i=b}^{y-1} \Phi^{B}_{b,i-1} \cdot a_{b,i} \cdot f_{V,i} \cdot [1 - \Phi^{Y}_{i+1,y-1}(V)] \tag{4.16}$$

$$P_{Y,V}(b,y) = \sum_{i=b}^{y-1} \Phi^{B}_{b,i-1} \cdot a_{b,i} \cdot f_{Y,i} \cdot [1 - \Phi^{V}_{i+1,y-1}(Y)] \tag{4.17}$$

Because case data had information on age and not the exact birth year, we averaged each exposure history probability across the two possible birth years given the age and the observation year (for instance, a 10-year-old in 2000 may have been born in either 1989 or 1990). Because the probabilities of different infection histories become very similar for cohorts born long before the lineages split, we used the probabilities calculated for the birth year 1970 for all previous cohorts to decrease computation time.

4.4.4 Season-specific attack rates

Let $\mathrm{P}_{\mathrm{inf}}(b,i,t)$ be the probability that a previously unexposed person born in year b has been infected with influenza B after experiencing fraction $t \in [0,1]$ of season i. Assuming a

constant instantaneous attack rate $\alpha_{b,i}$ throughout the season, $\mathrm{P}_{\mathrm{inf}}(b,i,t)$ is given by:

$$\mathrm{P}_{\mathrm{inf}}(b,i,t) = 1 - e^{-\alpha_{b,i}t} \tag{4.18}$$

We let the instantaneous attack rate $\alpha_{b,i}$ be equal to an age-specific baseline multiplied by an intensity score S_i representing the strength of influenza B circulation in season i relative to other seasons:

$$\alpha_{b,i} = -\ln[1 - \beta(b,y)] \cdot S_i, \tag{4.19}$$

where $\beta(b,y)$ takes on value β_1, if birth year b corresponds to an age of less than 5 in year y, β_2, if the corresponding age is 6-17, and β_3, for ages 18 and older. The probability of infection for an unexposed person born in year b across the entire season, $a_{b,i}$, is obtained by substituting $\alpha_{b,i}$ in Eq. [4.18] and setting $t = 1$:

$$a_{b,i} = 1 - e^{-\alpha_{b,i}} = 1 - [1 - \beta(b,y)]^{S_i} \tag{4.20}$$

The definition of $\alpha_{b,i}$ in to Eq. (4.19) was chosen such that for a season with average influenza intensity ($S_i = 1$), the annual probability of infection for an unexposed person is equal to $\beta(b,y)$.

For the season corresponding to the first year of life ($i = b$), people are only susceptible to infections during a fraction of the season, depending on when they were born and how long they were protected by maternal antibodies. In those cases, we defined $a_{b,i}$ as the expected probability of infection across all possible weeks of birth:

$$a_{b,i} = \frac{1}{W_b} \sum_{w} 1 - e^{-\alpha_{b,i}\phi_i(b,w+M)}, \quad \text{for } i = b \tag{4.21}$$

where $\phi_i(b, w)$ is the fraction of cases in season i observed in or after week w of year b and W_b is the number of weeks in year b. Because people are assumed to be completely protected against infection by maternal antibodies for the first M weeks following birth, ϕ_i was computed for an effective birth week $w + M$. Based on the fraction of children under the age of 1 with detectable antibodies to influenza B [17], we set M to 26 weeks (approximately 6 months). Averaging $\phi_i(b, w + M)$ over all possible birth weeks w in year b gives the expected fraction of season i experienced by a person born in year b assuming births are distributed uniformly in time. We estimated $\phi_i(b, w)$ by fitting the incomplete beta function to the cumulative fraction of cases in the seasons for which we had case data, using R package FlexParamCurve. Following Gostic *et al.* [60], we truncated season-specific infection probabilities so that they never exceed 0.75 even in years of high estimated influenza B intensity.

4.4.5 Intensity scores

We defined the intensity score S_i in Eq. (4.19) as:

$$S_i = \frac{\%\text{ influenza B in ILI specimens} \times \text{ILI incidence, for season } i}{\text{mean [\% influenza B in ILI specimens} \times \text{ILI incidence] across seasons}} \tag{4.22}$$

where ILI stands for "influenza-like illness". Annual influenza surveillance reports from New Zealand's Institute of Environmental Science and Research (ESR) available from 2003 to 2016 [3] give the "isolation" or "detection" rate (the number of influenza-positive swabs divided by the number of swabs tested), the percentage of influenza A and B viruses among all influenza-positive isolates (both sentinel and non-sentinel), and the estimated number of ILI cases in New Zealand for each season. The reports do not directly give the fraction of ILI specimens that were influenza B-positive. Instead, we calculated the fraction of ILI isolates that were influenza B-positive in a season by multiplying the fraction of ILI isolates that were influenza-positive by the fraction of influenza-positive specimens that were influenza B. For

seasons without data on the fraction of influenza-positive specimens in ILI specimens (1988 to 2000), without data on the fraction of influenza B in influenza-positive specimens (1988-89), or without estimates of the total number of ILI cases (1988-2001), we used the average values of those quantities across the remaining seasons. Although reports are not available for 1990-2002, the 2003 annual report lists the frequency of influenza B in influenza-positive specimens for those seasons.

The World Health Organization's FluNet has weekly data on the fraction of ILI specimens that were influenza B-positive in Australia from 1997 to the present. However, in data from before 2003, the number of influenza-positive specimens was usually the same as the reported number of specimens processed for that week, suggesting strong case ascertainment or reporting bias. We thus used data on the percent of influenza-positive ILI cases from 2003 on. Annual influenza reports from 1994 to 2010 are available from the Australian government's Department of Health website [1]. They report numbers of influenza A- and B-positive isolates but not the total number of specimens tested. Thus, only the fraction of influenza B in influenza isolates (but not in ILI isolates) can be estimated from those reports. We therefore used data from the WHO to calculate the fraction of influenza B-positive specimens in ILI specimens for 2003-2017. For 1994-2002, we multiplied the fraction of influenza B in influenza-positive specimens for each season (from the Department of Health reports) by the average annual fraction of influenza-positive specimens in ILI specimens from 2003-2017 (from the WHO data) to arrive at the fraction of influenza B positive ILI specimens. Finally, for seasons where data were missing altogether (1988-1993), we used the average annual fraction of influenza B positive specimens in ILI specimens for subsequent seasons (1994-2017).

To estimate ILI incidence in Australia, we used the maximum weekly number of ILI cases per 1000 consultations for each season from Department of Health annual and weekly reports (weekly reports are available for years since the last annual report in 2010). Different ILI definitions were used from 1994-2003 and from 2004-2010, and starting in 2009 reported

weekly ILI rates were averaged from multiple branches of the Australian influenza surveillance system. We thus normalized values by the average value within each of those periods (1994-2003, 2004-2008 and 2009-18) to arrive at a normalized peak number of ILI cases per 1000 consultations in Australia.

4.4.6 Historical frequencies of influenza B lineages

To estimate historical frequencies of B/Victoria and B/Yamagata, we downloaded data on lineage and date and country of isolation for all influenza B isolates on the Global Initiative on Sharing All Influenza Data (GISAID) website collected until 09/30/2018. To complement these data, we searched the NCBI Influenza Virus Database for all protein-coding HA sequences of influenza B viruses isolated from humans and excluding laboratory strains (information on passage history for GISAID entries was scarce and non-standardized and so we did not filter out laboratory strains from the GISAID data). Because lineage information was missing for virtually all sequences retrieved from NCBI, we used BLAST to assign each sequence to either B/Victoria or B/Yamagata based on the highest bit score match with reference sequences B/Victoria/2/87 and B/Yamagata/16/1988.

We combined data from both databases to estimate the frequency of B/Yamagata and B/Victoria isolates in each season. Isolates collected in year y in Europe or North America were assigned to season $y-1/y$, if collected before October, and to season $y/y+1$, if collected in October-December. Because most European and North American isolates were collected before October of the respective year (median across years = 83% for Gisaid and 82% for NCBI), we assumed isolates with missing month of collection in those regions were collected before October and thus in season $y - 1/y$.

Isolates with the same name but reported for different countries or seasons were considered separately. We condensed multiple occurrences of the same isolate in the same country and season (within or across datasets) into one, disregarding isolates for which different lineages were assigned in different countries/seasons. Using isolates present in both databases,

we found that our BLAST lineage assignment matched the lineage reported on GISAID in 98% (3,159/3,217) of cases. We disregarded isolates for which our BLAST assignment and the reported GISAID assignment disagreed. The final dataset consisted of 35,158 isolates, 23 of which (0.07%) were represented more than once (in different countries or seasons). We estimated the frequency of a lineage as the number of isolates belonging to that lineage divided by the total number of influenza B isolates collected in a season.

Because the numbers of isolates collected in New Zealand were low for many years, we estimated pooled estimates from New Zealand and Australia to estimate lineage frequencies. For seasons with fewer than 10 isolates reported in Australia and New Zealand combined, we used frequencies estimated from isolates collected in all countries represented in the data. Frequencies estimated from all other countries combined were strongly correlated with estimates based on isolates from Australia and New Zealand only (Pearson's correlation coefficient = 0.91, 95% CI 0.81-0.96; Fig. 4.29). We also considered using frequencies estimated from isolates collected in the United States, which were also correlated with frequencies in Australia and New Zealand (Figs. 4.7 and 4.29), but the correlation was weaker (0.73, 95% CI 0.48-0.88). We hoped to use the sequence databases to get more reliable estimates of lineage frequencies in the 1980s than those provided by early antigenic characterization [139], but fewer than 10 isolates were available for each year before 1988. To accommodate uncertainty, we grouped infections with B/Victoria and B/Yamagata before 1988 with infections by the ancestral influenza B strains circulating before the lineages split.

We compared our estimates of lineage frequencies based on sequence data to estimates based on antigenic characterization of circulating strains from epidemiological surveillance reports (Fig. 4.7). Surveillance reports from Australia are available from the Australian Government's Department of Health website [1]. Surveillance reports from New Zealand are available from the website of New Zealand's Institute of Environmental Science and Research (ESR) [3]. Although reports from New Zealand are only available from 2003 on, Fig. 27 of the 2012 report shows B/Victoria and B/Yamagata frequencies from 1990 to 2002 (without

reporting the number of isolates used to estimate those frequencies). Annual summaries of influenza surveillance in the United States are published by the Centers for Disease Control and Prevention (e.g., [57]).

4.4.7 Model validation with independent serological data

We compared the fraction of children predicted by our model to have been previously infected with B/Victoria, B/Yamagata or either lineage with the fraction of children that had detectable antibodies against the corresponding lineage (or any influenza B strain) in the Netherlands [17]. Sera from children 0-7 years old collected between February 2006 and June 2007 were tested using the hemagglutination inhibition assay against a panel of reference B/Victoria and B/Yamagata strains as well of strains isolated in the Netherlands during the study period. Sera were considered positive if their hemagglutination inhibition titer was ≥ 10 against at least one strain from the corresponding set (all influenza B strains, B/Victoria strains, and B/Yamagata strains). We compared these data with predictions under the maximum likelihood-parameter estimates of our model fitted to the complete New Zealand data.

We also used the seroprevalence data to independently estimate the annual probability of infection for preschoolers and school-age children, equivalent to the β_1 and β_2 parameters in our model. Assuming a constant instantaneous attack rate α, an individual of age A years is still naive (and therefore seronegative) to influenza B with probability $P_N(A)$ given by:

$$P_N(A) = e^{-\alpha A} \tag{4.23}$$

The probability of observing X seronegative individuals in a sample of n individuals of age A can be calculated assuming $X \sim \text{Binomial}[n, P_N(A)]$, and α thus can be estimated by maximum likelihood. The annual attack rate can then be calculated from the instantaneous attack rate α as $\beta_{\text{Netherlands}} = 1 - e^{-\alpha}$.

We make two modifications to Eq. (4.23) to account for the presence of maternal antibodies early in life and for uncertainty in the age of individuals when their serum was collected. First, we assume individuals spend a time m (in units of years) fully protected against influenza B due to the presence of maternal antibodies. Consistent with the fraction of children under the age of 1 with detectable antibodies to influenza B [17], we assumed $m = 0.5$ year. Second, because ages were reported at the resolution of one year (e.g. an individual 2.6 years old is reported as being 2 years old), we assume individuals with recorded age A were sampled at a randomly distributed time $T \in [0, 1)$ during the interval between the ages of A and $A + 1$. Thus, we let $P_N(A)$ be given by the expectation over T:

$$P_N(A) = E[e^{-\alpha(A+T-m)}] = \int_0^1 e^{-\alpha(A+t-m)} f(t)dt \tag{4.24}$$

where $f(t)$ is the probability density function of T. Assuming T is uniformly distributed between 0 and 1 (i.e., $f(t) = 1$), we have:

$$\begin{aligned} P_N(A) &= \int_0^1 e^{-\alpha(A+t-m)} dt \\ &= e^{-\alpha(A-m)} \int_0^1 e^{-\alpha t} dt \\ &= e^{-\alpha(A-m)} \left[\frac{-e^{-\alpha t}}{\alpha} \right] \Bigg|_0^1 \\ &= e^{-\alpha(A-m)} \frac{(1 - e^{-\alpha})}{\alpha} \end{aligned} \tag{4.25}$$

valid for $A > m$ and $\alpha > 0$. Letting α_1 and α_2 be the instantaneous attack rates for preschoolers and school-age children:

$$P_N(A) = \begin{cases} e^{-\alpha_1(A-m)} \frac{(1-e^{-\alpha_1})}{\alpha_1}, & \text{if } A \leq A_s \\ e^{-\alpha_1(A_s-m)} e^{-\alpha_2(A-A_s)} \frac{(1-e^{-\alpha_2})}{\alpha_2}, & \text{if } A > A_s \end{cases} \tag{4.26}$$

where A_s is the age at which children start going to school (4 years old in the Netherlands

[4]). Note that for school-age children (the equation for $A > A_s$ on the bottom) the correction term for uncertainty in sampling is not necessary for the time spent in preschool (assumed to be exactly A_s years), only for the time after preschool ($A - A_s$).

4.4.8 Estimating missing sentinel cases and cases with missing lineage information

Estimating missing sentinel cases

Because a maximum of three samples from ILI patients were sent for testing by sentinel general practices each week, the relative proportions of sentinel and non-sentinel cases in the New Zealand are different from the true proportions of cases leading to general practice visits and hospitalizations in New Zealand, and milder cases are thus likely underrepresented compared to more severe cases.

We estimated the number of influenza B cases attended by general practitioners missing from the data by multiplying the total number of general practice visits due to ILI in New Zealand estimated by surveillance reports[3] by the fraction of influenza-like illness specimens that tested positive for influenza B (Materials and Methods: "Intensity scores"). For each season, we calculated a correction factor by dividing the estimated total number of general practice visits due to influenza B by the number of cases present in the data for that season (for 2001, we used the average of correction factors for other reasons because the estimated total number of ILI cases in New Zealand was missing). We then generated an adjusted dataset by multiplying the number of cases observed in each birth year by the correction factor for the corresponding seasons, and we fitted the model to this adjusted dataset. We assumed the correction factor was constant across birth years because we could not estimate the total number of general practice visits due to influenza B by individual birth years or age groups.

Handling cases with missing lineage information

We assumed cases with missing lineage information in 2002 (11 sentinel and 51 non-sentinel) and 2011 (90 sentinel and 222 non-sentinel) belonged to B/Victoria, since 99% of identified cases in those seasons were B/Victoria (86/87 cases in 2002, 276/280 cases in 2011) as were 94% and 92% of isolates from sequence databases (for Australia and New Zealand combined). Unidentified cases were reported in 2005 and 2012, but those seasons were not clearly dominated by a single lineage and we thus disregarded unidentified cases in those seasons. Removing unidentified cases altogether led to similar parameter estimates.

4.4.9 Sequence divergence analysis

To estimate the amount of evolution within and between lineages, we analyzed all complete HA and NA sequences from human influenza B isolates available on GISAID in July 2019. The set of isolates used in this analysis differs from the set used to estimate lineage frequencies because we required isolates to have complete sequences (although not all sequences listed as complete on GISAID were in fact complete). Two isolates collected in 2000 (B/Hong Kong/548/2000 and B/Victoria/504/2000) were deposited as B/Victoria but our BLAST assignment indicated they were in fact B/Yamagata (their low divergence from B/Yamagata strains was a clear outlier). NA sequences from isolates B/Kanagawa/73 and B/Ann Arbor/1994 were only small fragments (99 and 100 amino acids long) poorly aligned with other sequences and were thus excluded. We also excluded NA sequences from B/Yamagata isolates B/Catalonia/NSVH100773835/2018 and B/Catalonia/NSVH100750997/2018 because they were extremely diverged (60% and 38%) from the reference strain B/Yamagata/16/88 and aligned poorly with other sequences.

To compare sequence diversity within and between lineages over time, we aligned sequences using MAFFT v. 7.310 [79] and calculated percent amino acid differences in pairs of sequences from the same lineage and in pairs with one sequence from each. For each year, we sampled 100 sequences from each lineage (or used all sequences if 100 or fewer were

available) to limit the number of pairwise calculations. To estimate how much B/Yamagata and B/Victoria evolved since the late 1980s, we calculated percent amino acid differences between each B/Yamagata and B/Victoria sequence and the corresponding HA and NA sequences of reference strains B/Yamagata/16/88 and B/Victoria/2/87. Unlike in the analysis of pairwise divergence within each time point, we used all sequences from each lineage in each year. We excluded sites in which one or both sequences had gaps or ambiguous amino acids.

To compare HA and NA divergence between influenza B lineages with divergence between influenza A subtypes, we downloaded complete HA and NA sequences from H3N2 and H1N1 isolated since 1977 and available on GISAID in August 2019. Homologous sites in the HA of H3N2 and H1N1 are difficult to identify by conventional sequence alignment, and instead we used the algorithm by Burke & Smith [21] implemented on the Influenza Research Database website [2]. Both H3N2 and H1N1 sequences were aligned with the reference H3N2 sequence A/Aichi/2/68. We verified that this method matched sites on the stalk and head of the H1N1 HA with sites on the stalk and head of H3N2 HA by comparing the resulting alignment with the alignment in Fig. S2 of Kirkpatrick *et al.* [84]. To limit the total number of influenza A sequences analyzed we randomly selected 100 H3N2 and 100 H1N1 sequences for years in which more than 100 sequences were available and used all available sequences for the remaining years. Isolates A/Canterbury/58/2000, A/Canterbury/87/2000 and A/Canterbury/55/2000 were excluded because both H1N1-like and H3N2-like sequences were available under the same isolate name on GISAID.

4.4.10 Demographic data

We obtained annual age distributions for the general population in New Zealand from StatsNZ [5]. When calculating changes in the mean age of the population over time, we excluded people over 90 years old, who were pooled together in New Zealand's demographic tables.

4.5 Acknowledgments

Celeste M. Donato, Philip Arevalo, Guus F. Rimmelzwaan, Liza Lopez, Q. Sue Huang, Vijaykrishna Dhanasekaran and Katia Koelle collaborated on this project and will be coauthors in the paper. Frank Wen helped with data collection from GISAID. This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under grant DP2 AI117921 and CEIRS Contract No. HHSN272201400005C awarded to Sarah Cobey. Marcos Vieira was also supported by a William Rainey Harper Dissertation Fellowship by the University of Chicago. This work was completed in part with resources provided by the University of Chicago Research Computing Center.

4.6 Supplementary Information

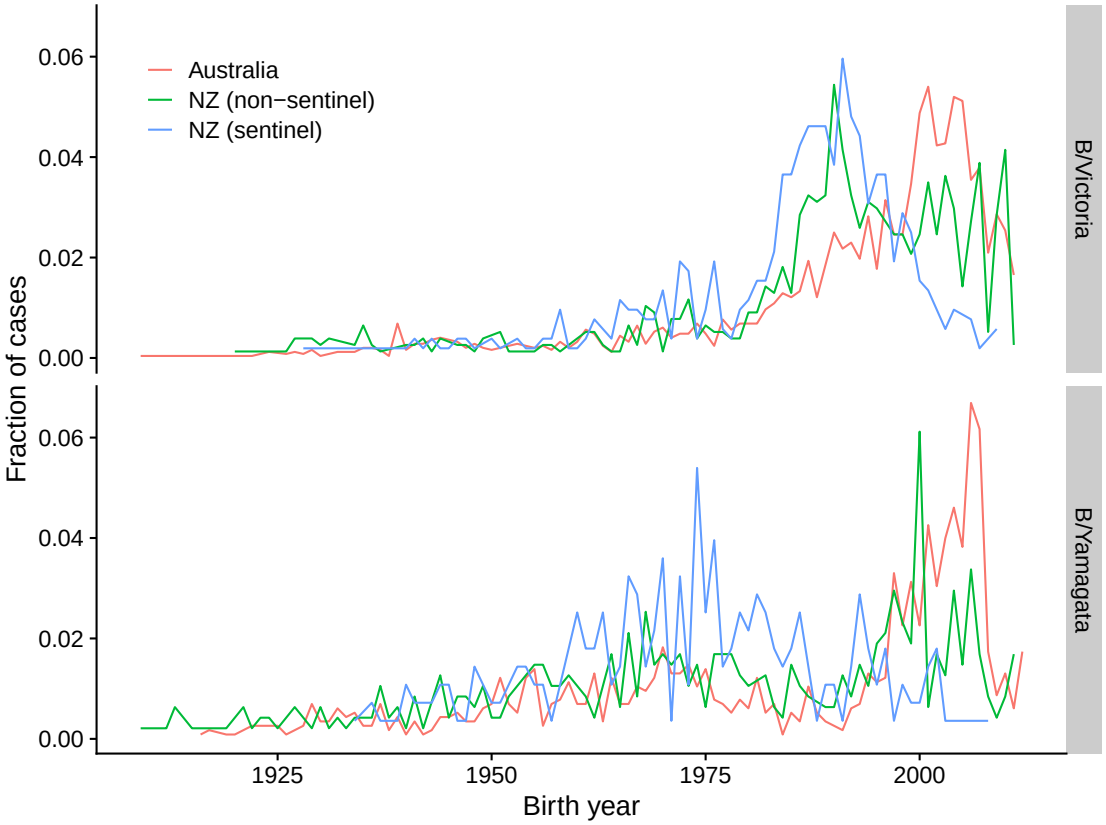

Figure 4.4: Distribution of medically attended influenza B cases in Australia (2002-2013) and New Zealand (2001-2012) by birth year of the infected person. The fraction of cases in each birth year was calculated relative to all cases observed for each lineage (separately by type of surveillance in New Zealand).

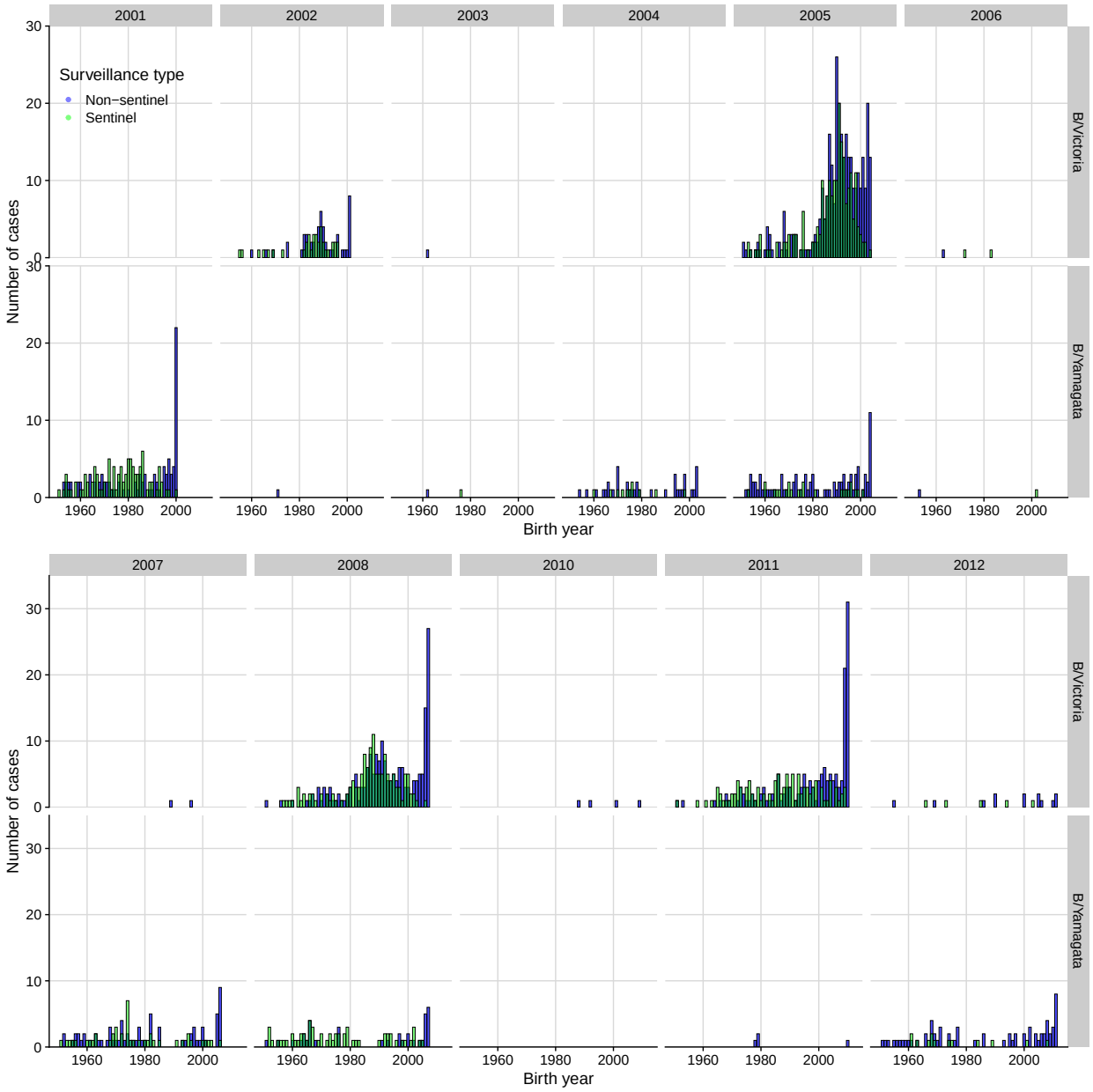

Figure 4.5: Distribution of medically attended influenza B cases in New Zealand by season, lineage and type of surveillance. Sentinel influenza B cases were identified by testing swabs taken from influenza-like illness pacients attended by general practices partitipating in New Zealand's sentinel surveillance system. Non-sentinel cases com primarily from hospital inpatients. Birth years prior to 1950 were omitted for clarity.

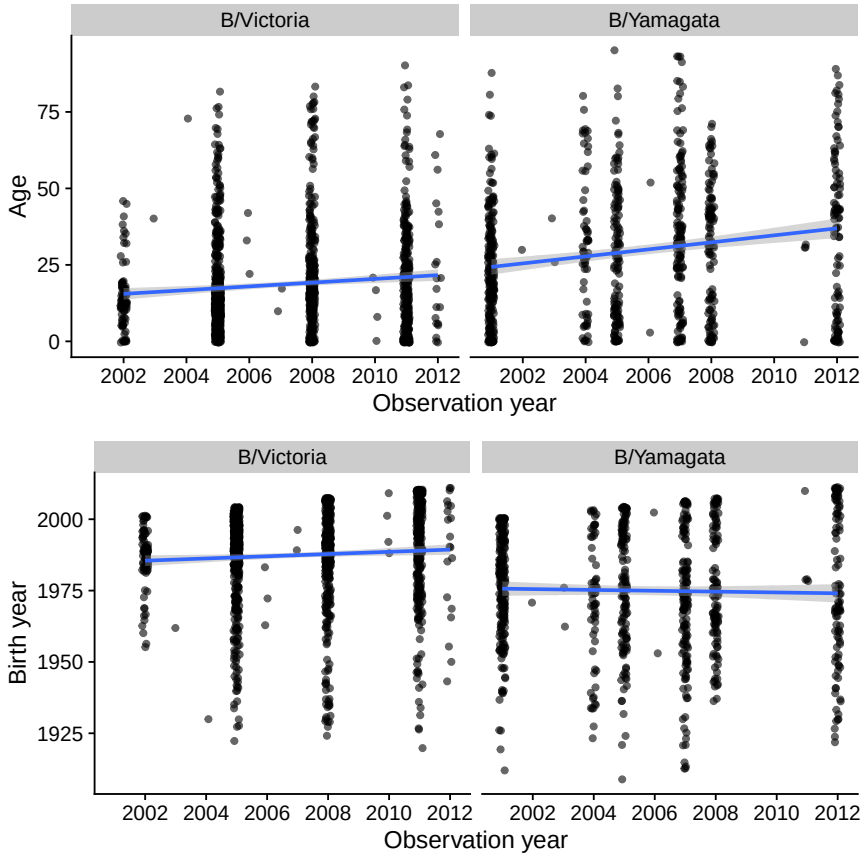

Figure 4.6: Changes in age and birth year distributions of influenza B cases over time. Age (top panels) and birth year (bottom panels) of people with medically attended influenza B infections in New Zealand in 2001-2013 as a function of the year when the infection occurred.

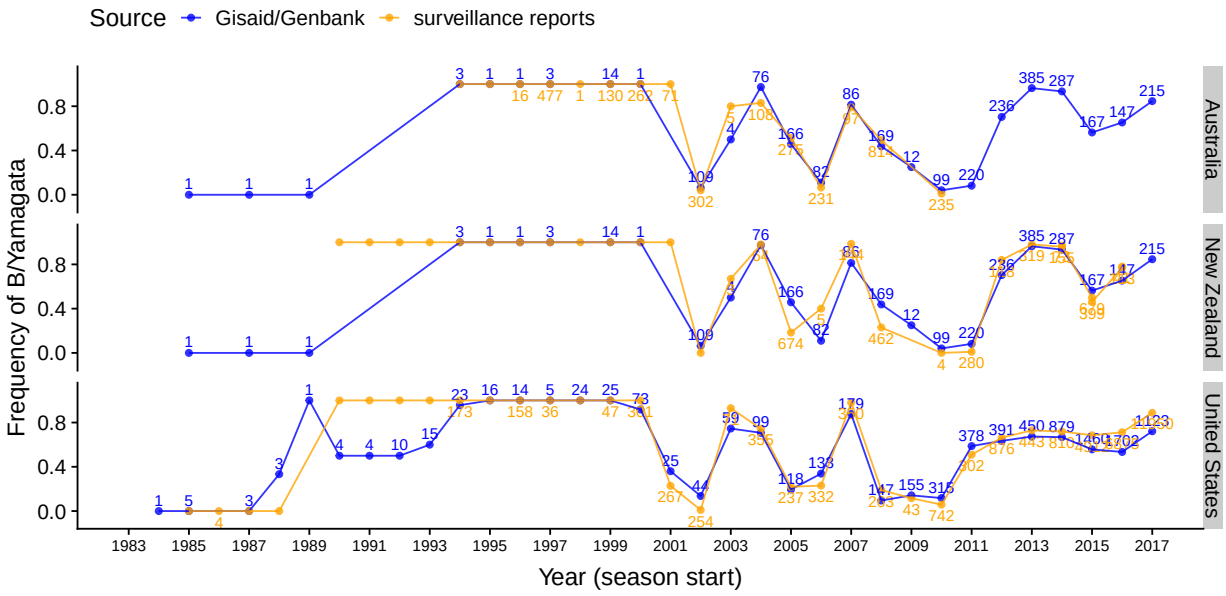

Figure 4.7: Comparison of lineage frequency estimates based on sequence data and surveillance reports. Frequencies are shown by annual influenza seasons, which span multiple calendar years in the United States but are contained in a single calendar year in Australia and New Zealand (e.g., 2006 refers to the 2006/2007 influenza season in the United States and to the 2006 season in Australia and New Zealand). Numbers indicate the number of isolates collected in that season and deposited on GISAID or the NCBI Influenza Virus Database (blue) or tested by surveillance (orange; the total number of isolates tested was not reported in some of the surveillance reports). Sequence database isolates from Australia and New Zealand were grouped together to estimate lineage frequencies.

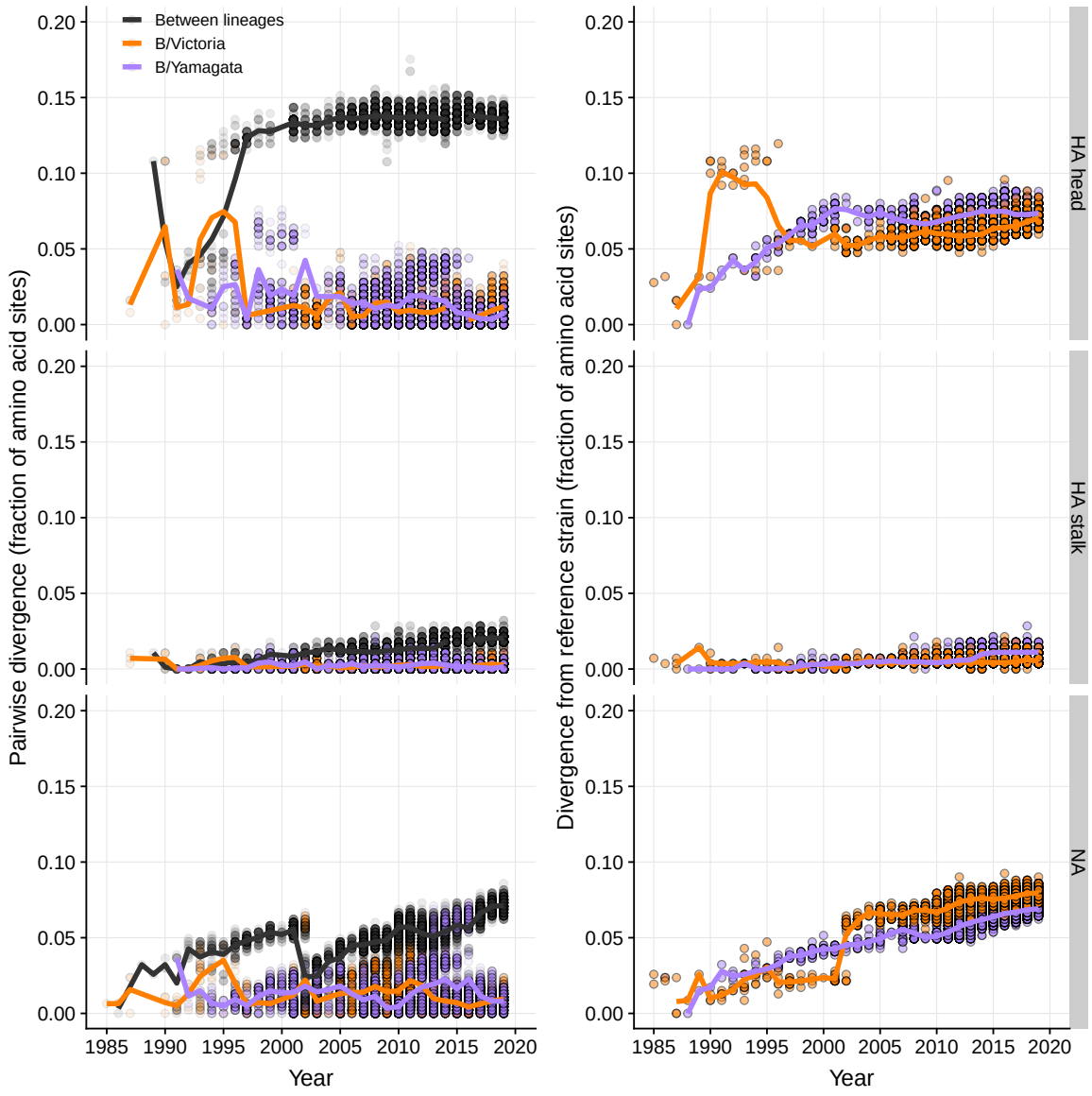

Figure 4.8: Amino acid divergence in the hemagglutinin (HA) and neuraminidase (NA) proteins within and between influenza B lineages. The left panel shows the fraction of amino acid sites that differ between pairs of sequences in each season (each point is a pair). Up to 500 randomly chosen pairs are shown, while the solid lines represent the annual average calculated from all pairs given two samples of up to 100 sequences from each lineage (i.e, up to 4,950 pairs). The right panel shows divergence of strains circulating at different times from reference strains B/Victoria/2/87 and B/Yamagata/16/88.

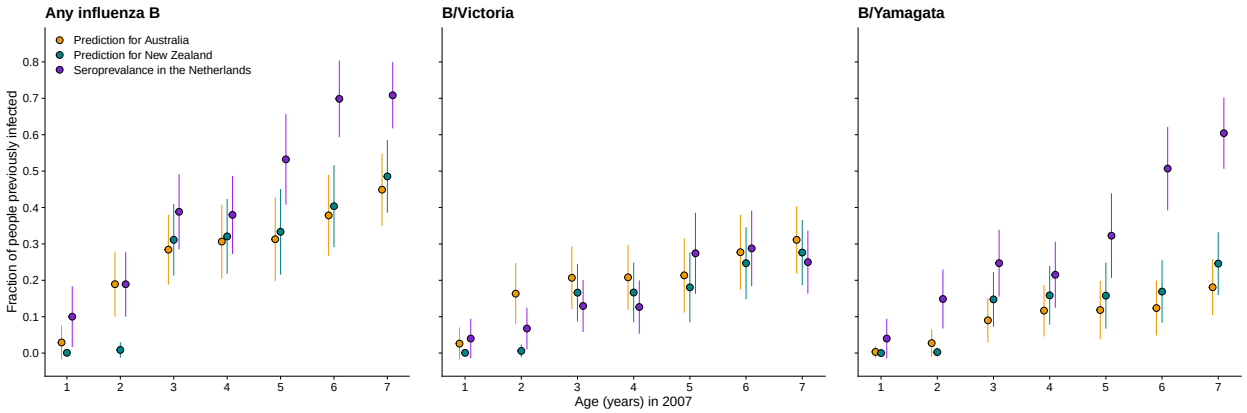

Figure 4.9: Predicted and observed frequency of past influenza B infections in children. The fraction of children previously infected with influenza B in the Netherlands in 2006-2007 was estimated by Bodewes *et al.* [17] as the fraction having serum antibodies against at least one influenza B strain from a panel of viruses (left panel), at least one B/Victoria strain (center panel) or at least one B/Yamagata strain (right panel). Bars show 95% binomial confidence intervals. Fractions for Australia and New Zealand were generated independently from the seroprevalence data by fitting a statistical model to medically attended influenza B infections. For the model predictions, binomial confidence intervals assume a sample size equal to the number of children with the corresponding age in the seroprelavence data.

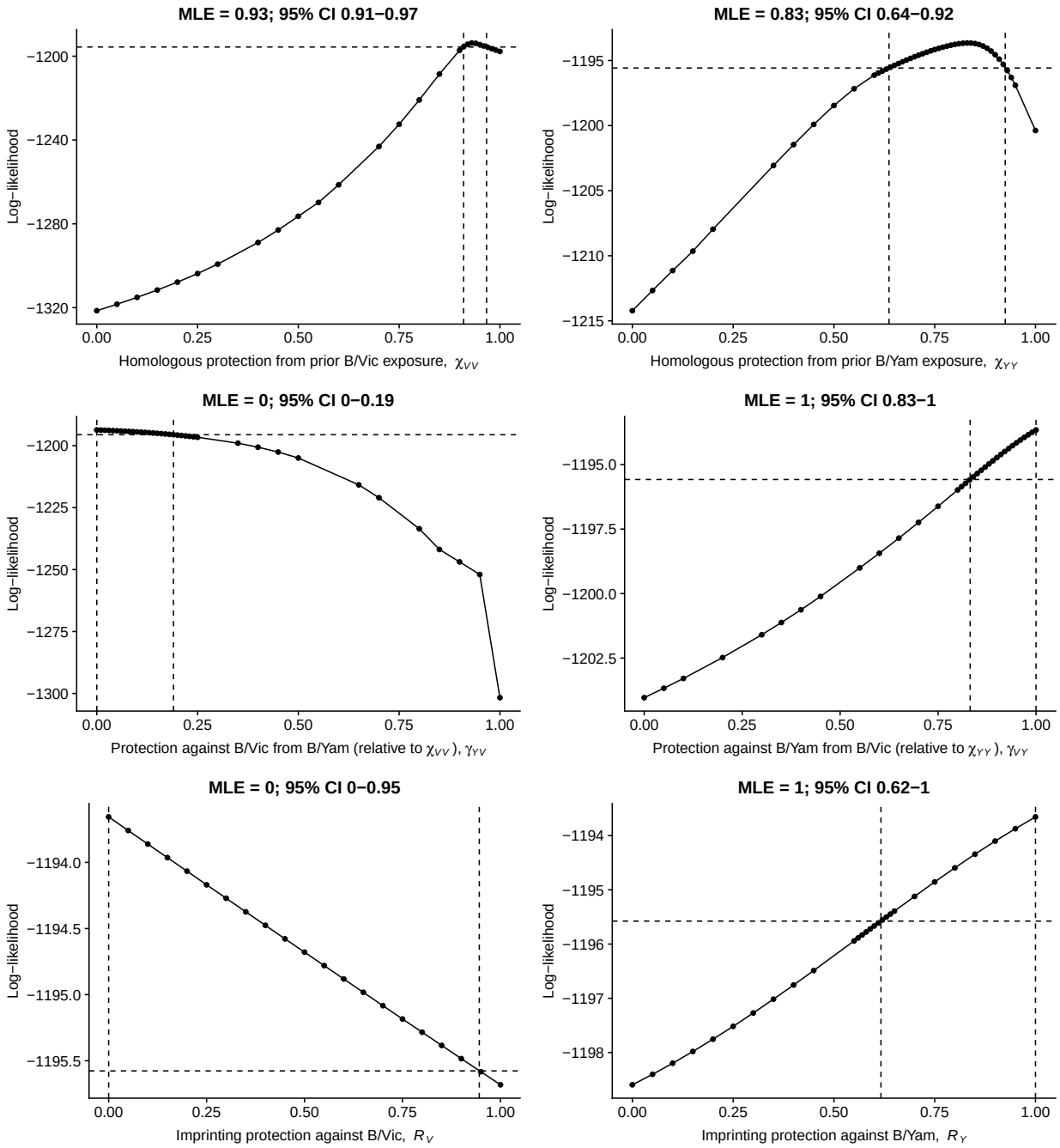

Figure 4.10: Likelihood profiles for protection parameters estimated from the complete New Zealand data, including general practice (sentinel) and hospital-associated (non-sentinel) cases. The 95% confidence interval based on a likelihood ratio test with one degree of freedom is indicated by the vertical dashed lines.

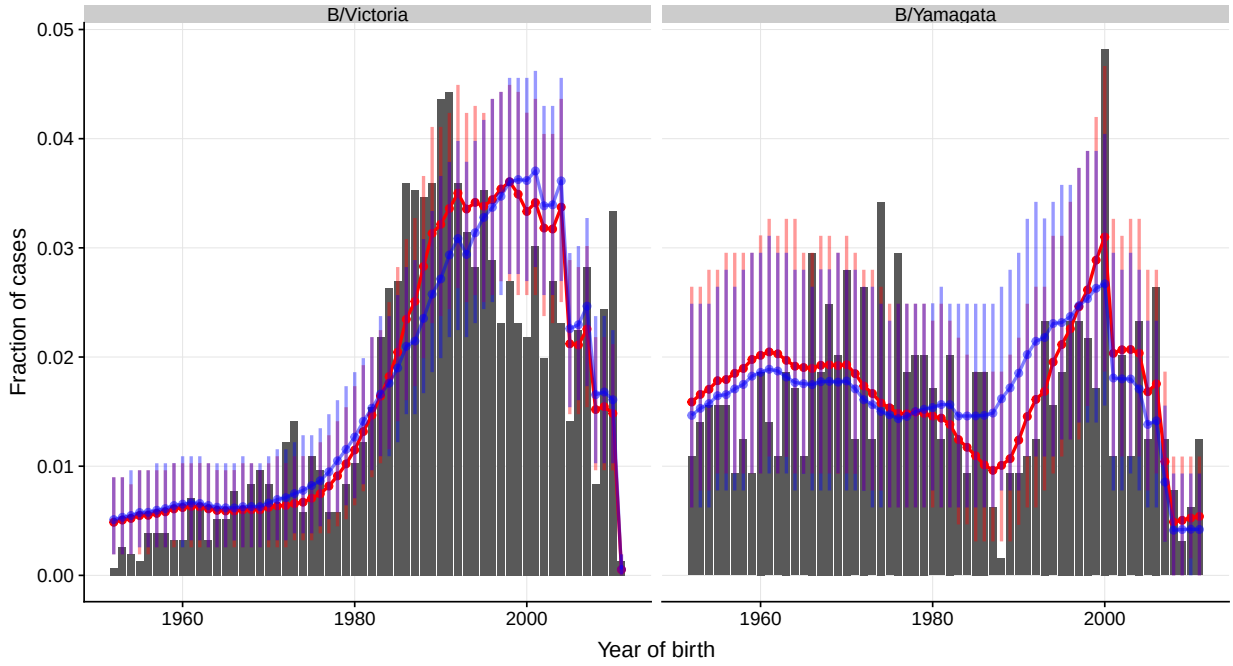

Figure 4.11: Predicted distributions of medically attended influenza B cases in New Zealand under strong protection against B/Victoria from B/Yamagata. Cases include general practice (sentinel) and hospital-associated (non-sentinel) cases. Red shows the predicted distribution under the maximum likelihood parameter estimates, whereas blue shows the best fit obtained while constraining protection against B/Victoria from B/Yamagata to 90% of B/Victoria's within-lineage protection ($\gamma_{YV} = 0.9$). Vertical bars are 95% bootstrap confidence intervals.

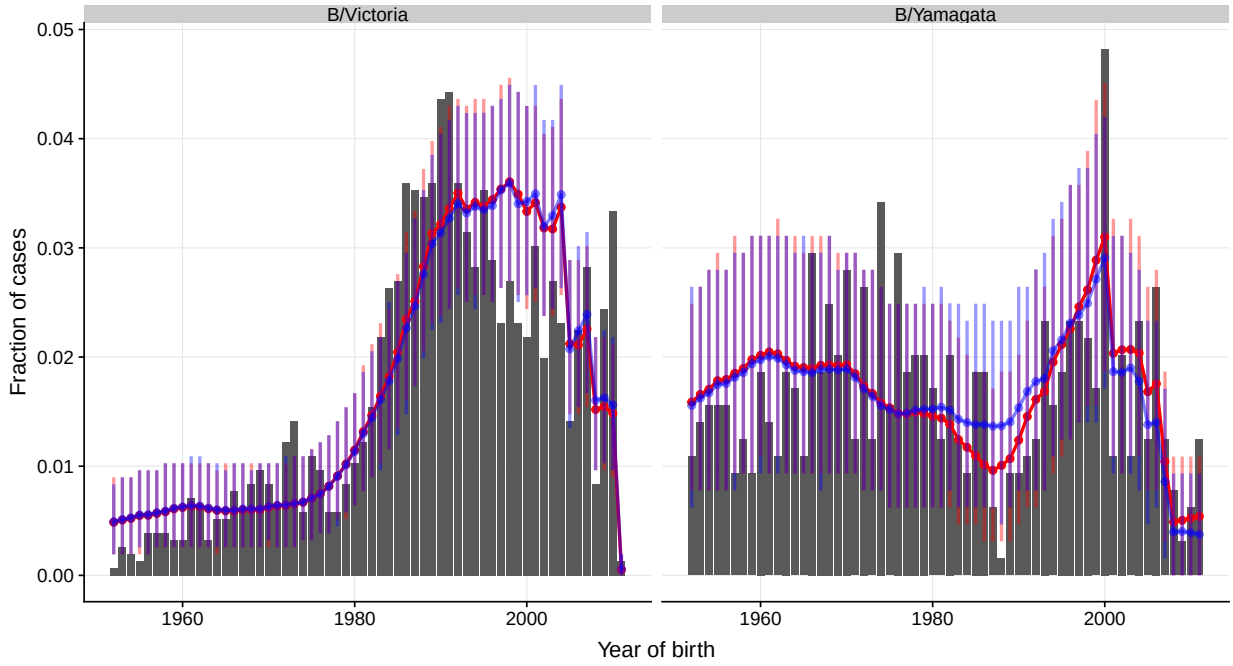

Figure 4.12: Predicted distributions of medically attended influenza B cases in New Zealand under no protection against B/Yamagata from B/Victoria. Cases include general practice (sentinel) and hospital-associated (non-sentinel) cases. Red shows the predicted distribution under the maximum likelihood parameter estimates, whereas blue shows the best fit obtained while constraining B/Victoria infections to give no protection against B/Yamagata ($\gamma_{VY} = 0$). Vertical bars are 95% bootstrap confidence intervals.

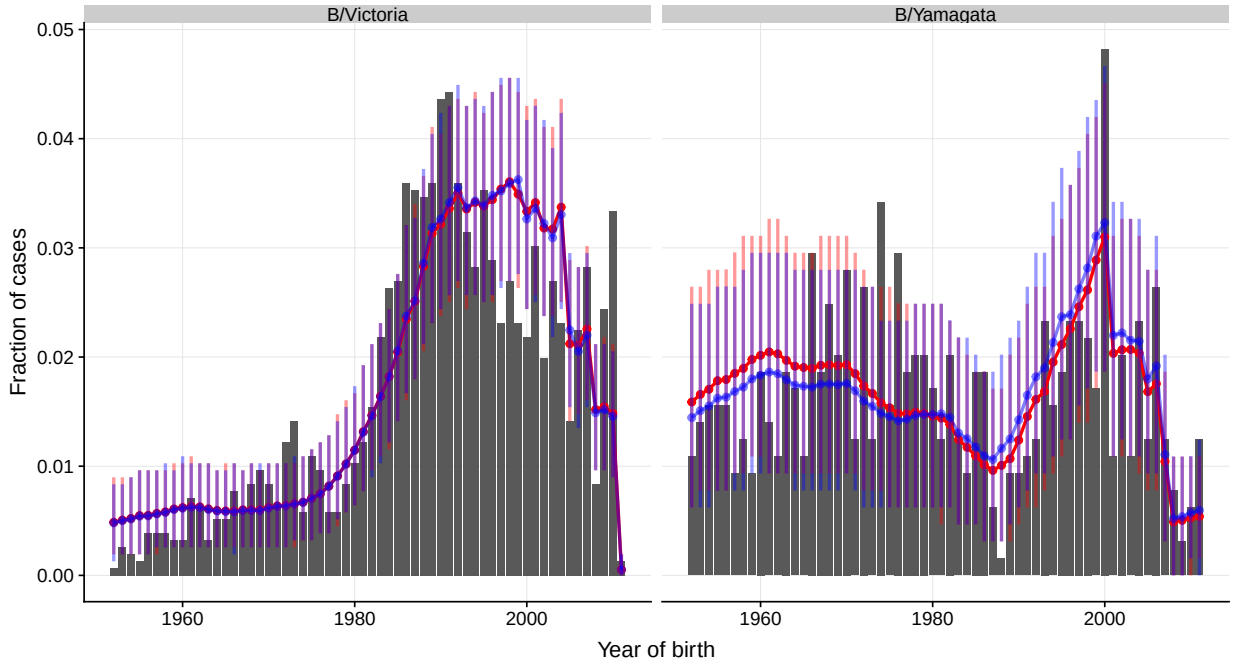

Figure 4.13: Predicted distributions of medically attended influenza B cases in New Zealand under no imprinting protection against B/Yamagata. Cases include general practice (sentinel) and hospital-associated (non-sentinel) cases. Red shows the predicted distribution under the maximum likelihood parameter estimates, whereas blue shows the best fit obtained while constraining the model to assign no additional protection against B/Yamagata in people for whom B/Yamagata was the lineages of first infection ($R_Y = 0$). Vertical bars are 95% bootstrap confidence intervals.

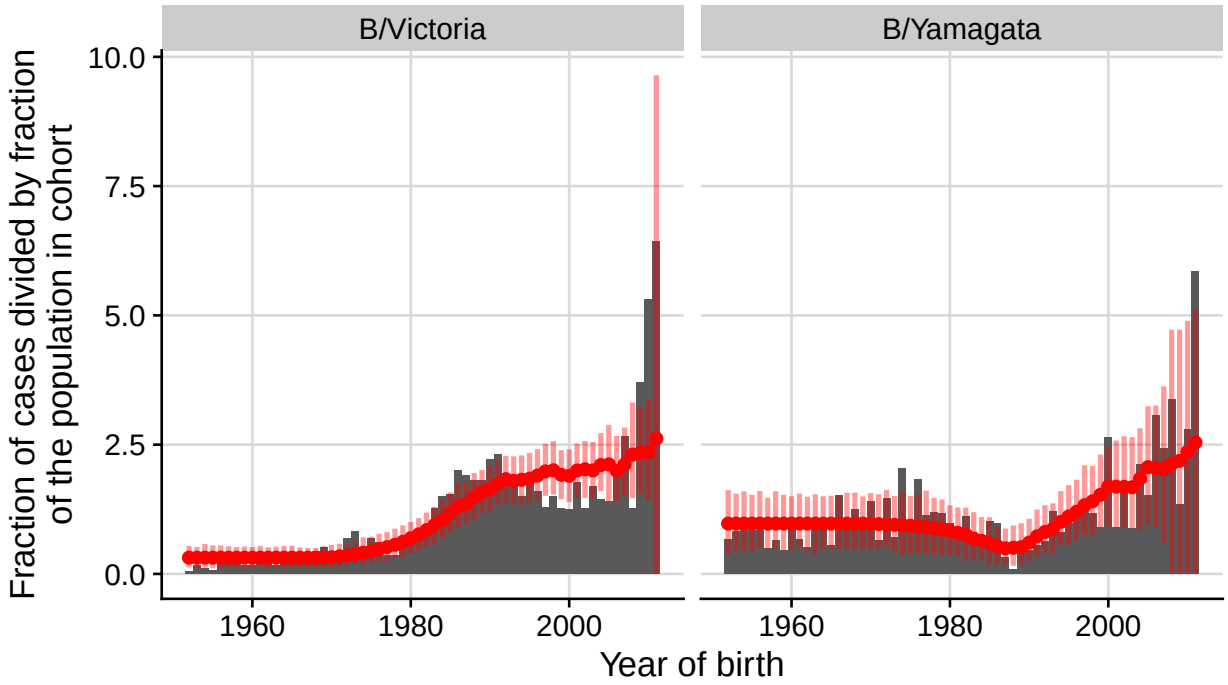

Figure 4.14: Observed and predicted distributions of medically attended influenza B cases in New Zealand by birth year normalized by demographic expectation. Cases include general practice (sentinel) and hospital-associated (non-sentinel) cases. Red lines and dots show the predicted distribution under the model. Vertical bars are 95% bootstrap confidence intervals. Predicted and observed fractions of cases were divided by the fraction of the population born in that birth year (i.e., the null expectation if all birth years were infected at the same rate). Cases in people born in 2011 (with wide bootstrap CIs) were omitted in Fig. 4.3 in the main text to improve visualization.

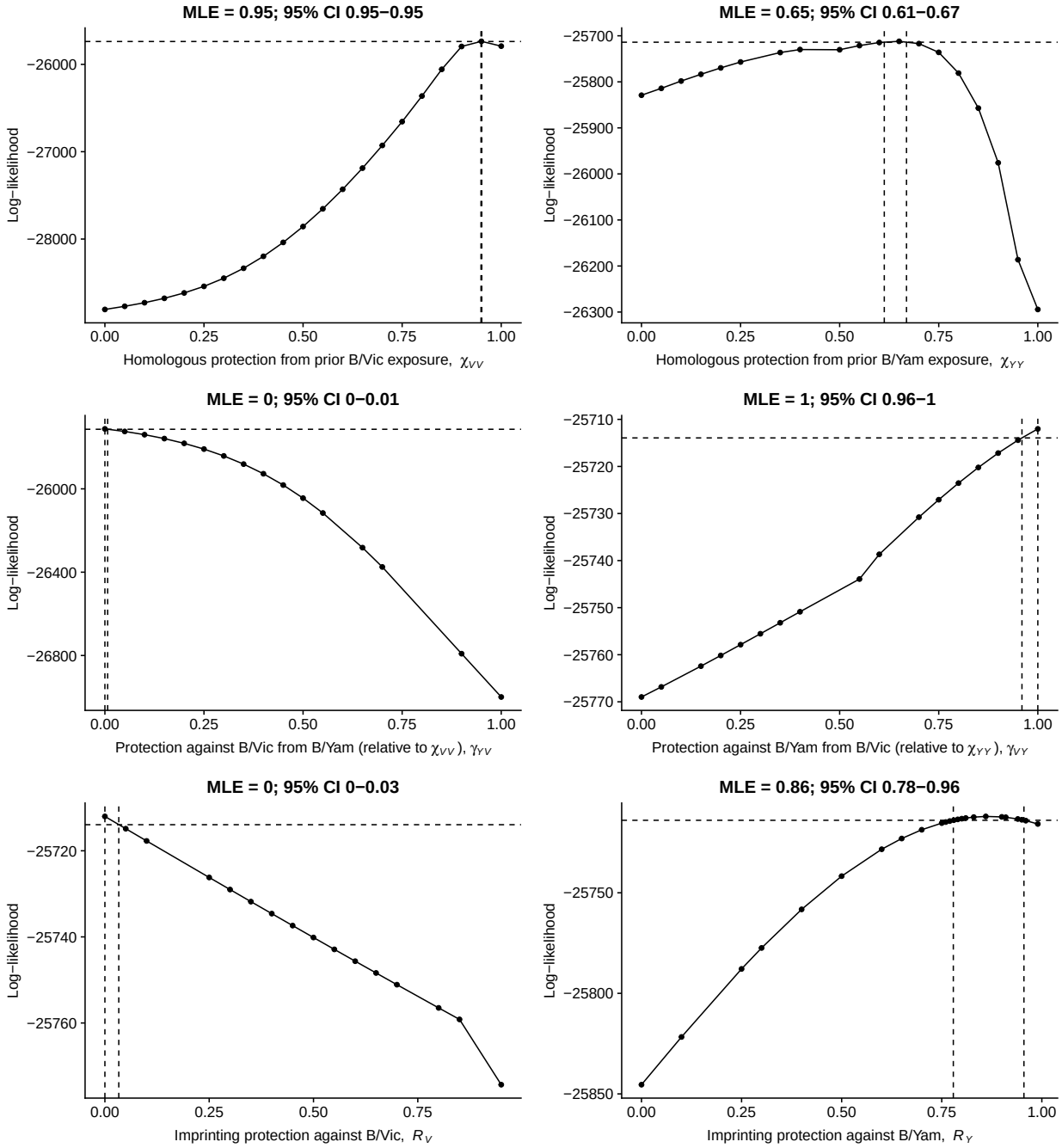

Figure 4.15: Likelihood profiles for protection parameters estimated from medically attended influenza B cases in New Zealand after including the estimated number of mild cases not caught by sentinel surveillance. The 95% confidence interval based on a likelihood ratio test with one degree of freedom is indicated by the vertical dashed lines.

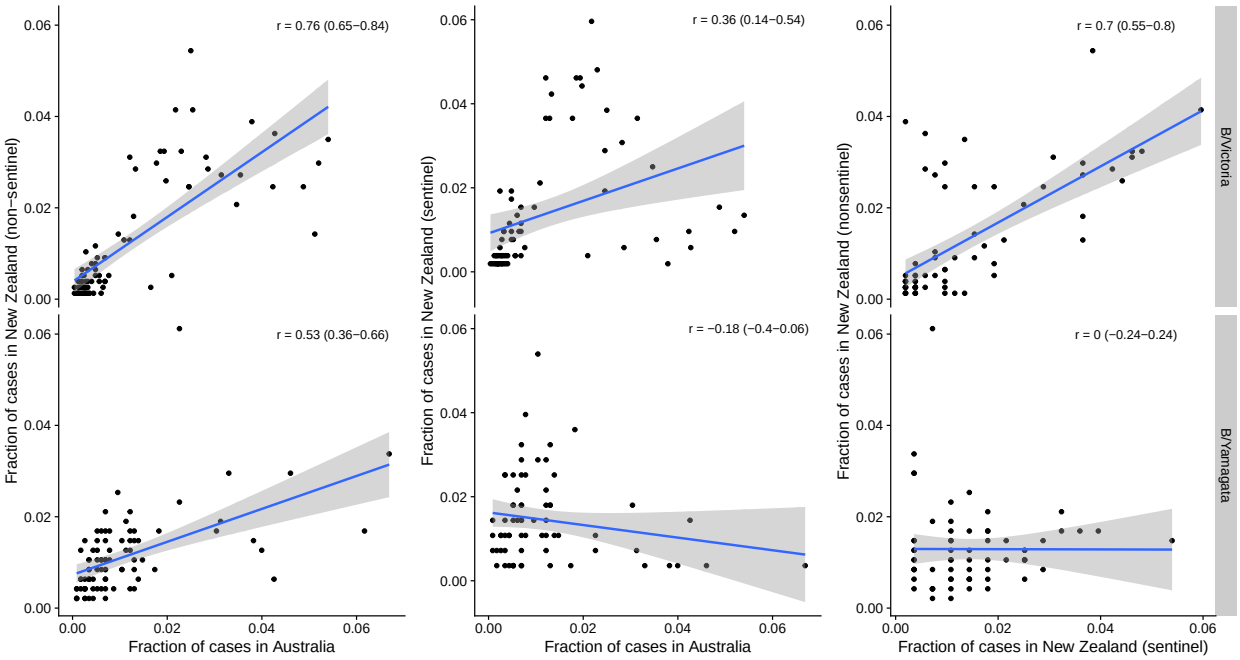

Figure 4.16: Correlations between the birth year distribution of influenza B cases in Australia and New Zealand. Each point represents a birth year. The fraction of cases in each birth year was calculated relative to all cases observed for each lineage in each country.

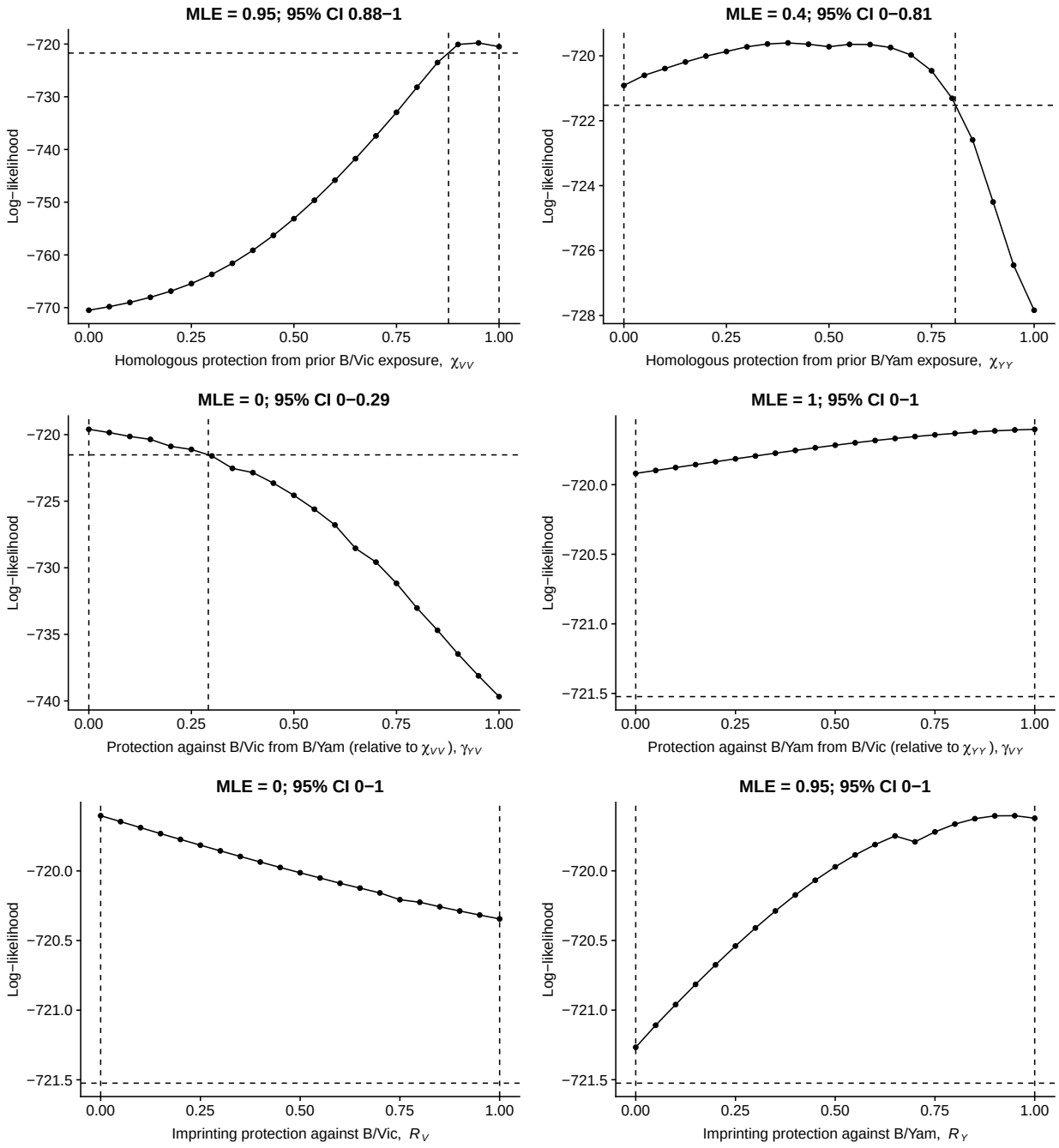

Figure 4.17: Likelihood profiles for protection parameters estimated from influenza B cases attended by general practitioners (sentinel cases) in New Zealand. The 95% confidence interval based on a likelihood ratio test with one degree of freedom is indicated by the vertical dashed lines.

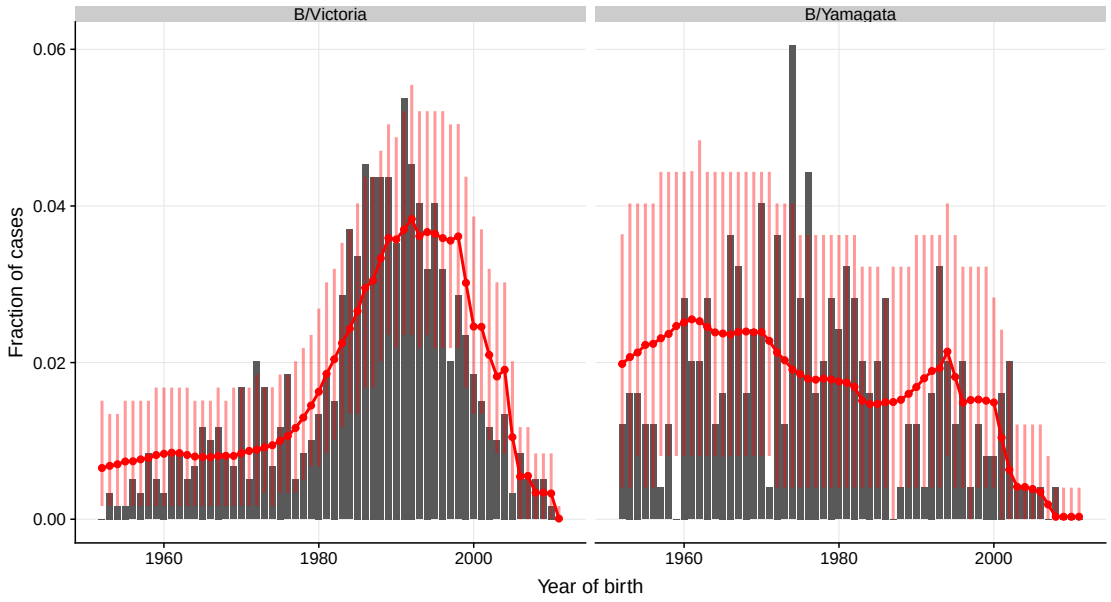

Figure 4.18: Observed and predicted birth year distributions of influenza B cases attended by general practitioners (sentinel cases) in New Zealand by observation year. Red lines and dots show the predicted distribution under the model. Vertical bars are 95% bootstrap confidence intervals.

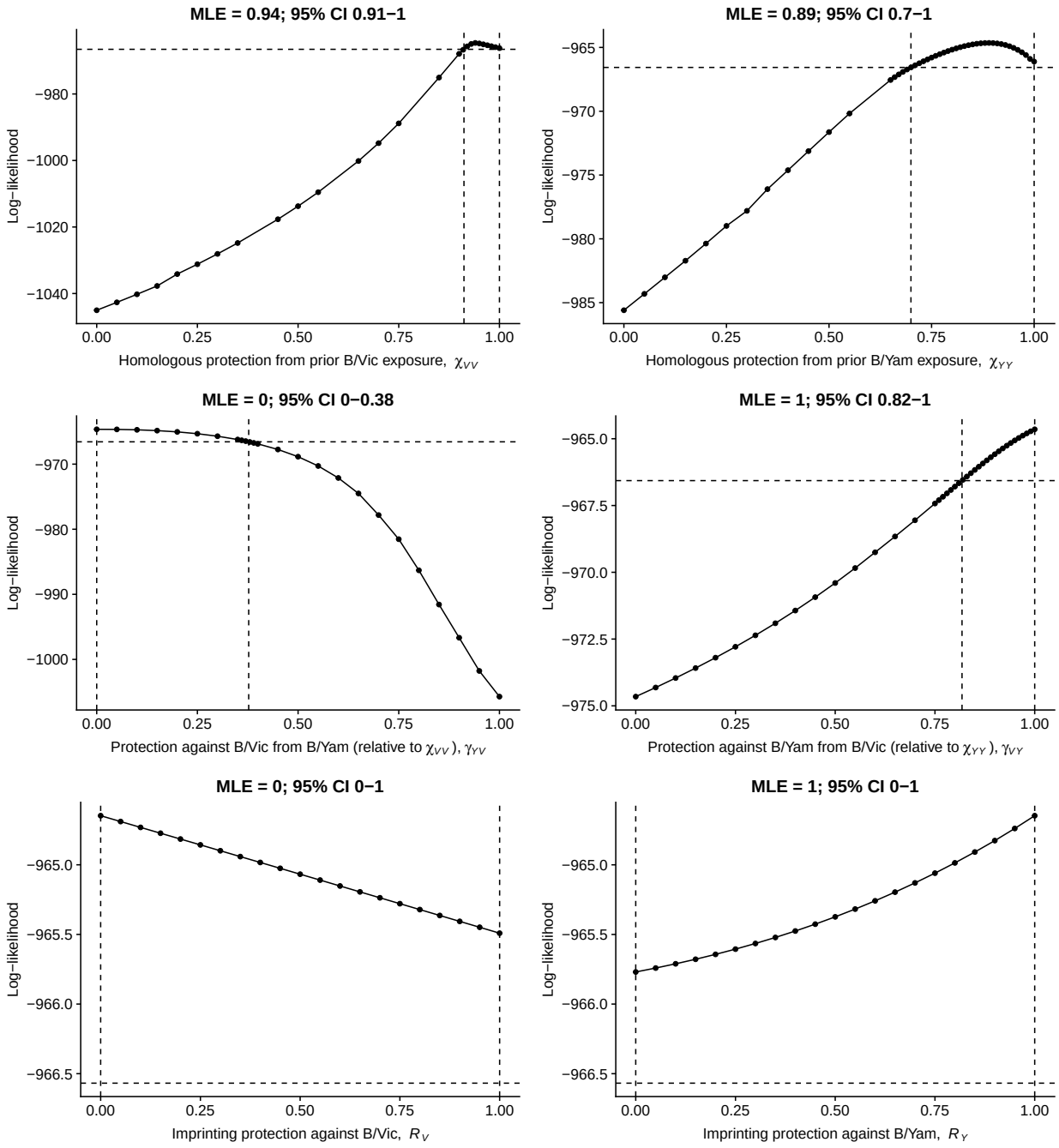

Figure 4.19: Likelihood profiles for protection parameters estimated from hospital-associated (non-sentinel) influenza B cases in New Zealand. The 95% confidence interval based on a likelihood ratio test with one degree of freedom is indicated by the vertical dashed lines.

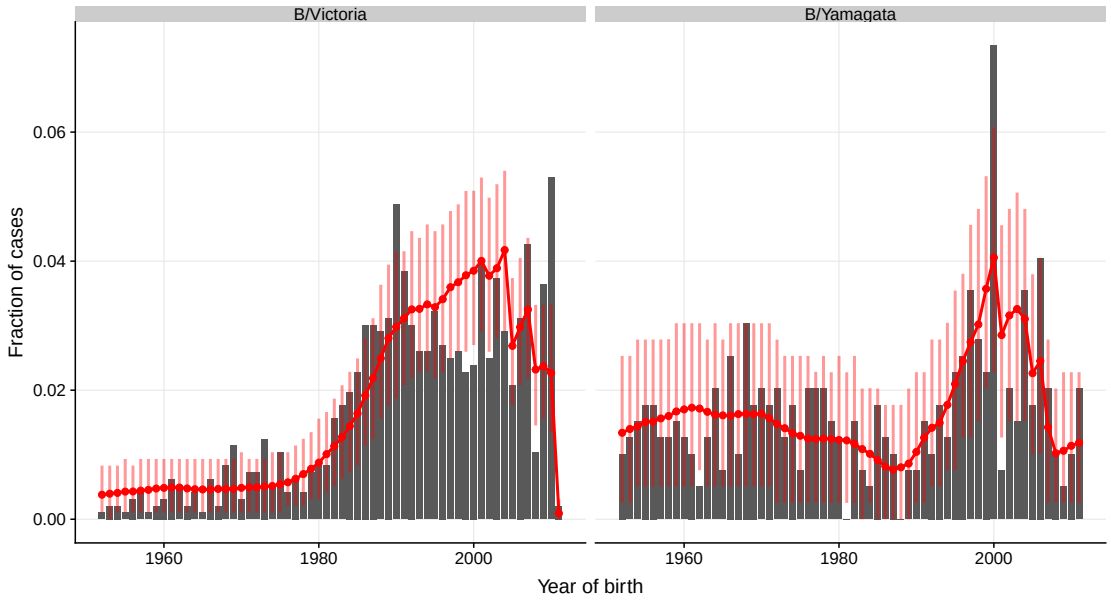

Figure 4.20: Observed and predicted distributions of hospital-associated (non-sentinel) influenza B cases in New Zealand by birth year. Red lines and dots show the predicted distribution under the model. Vertical bars are 95% bootstrap confidence intervals.

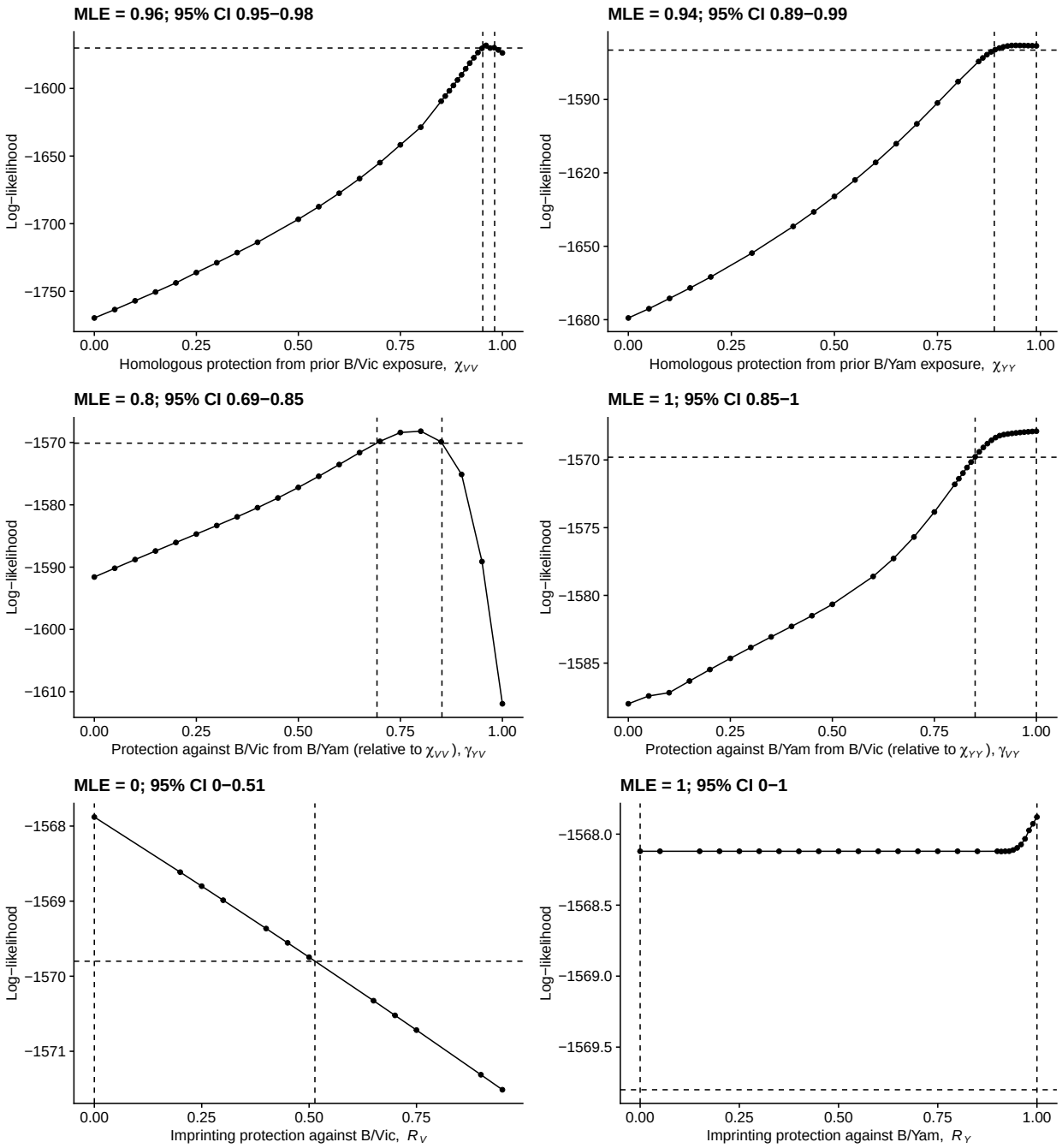

Figure 4.21: Likelihood profiles for protection parameters estimated from cases in Australia. The 95% confidence interval based on a likelihood ratio test with one degree of freedom is indicated by the vertical dashed lines.

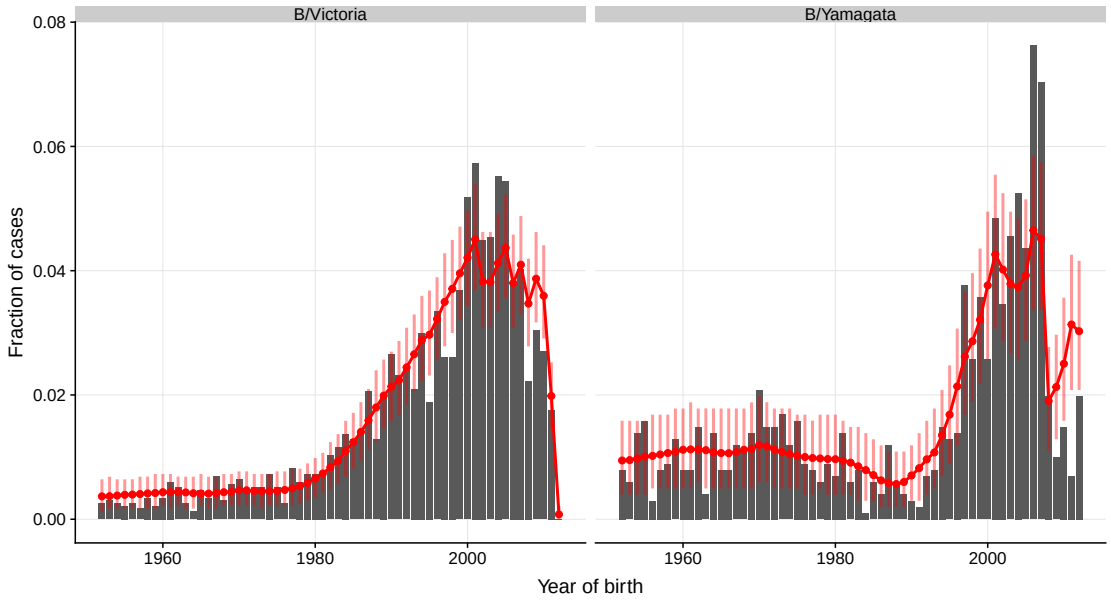

Figure 4.22: Observed and predicted distributions of influenza B cases in Australia by birth year. Red lines and dots show the predicted distribution under the model. Vertical bars are 95% bootstrap confidence intervals.

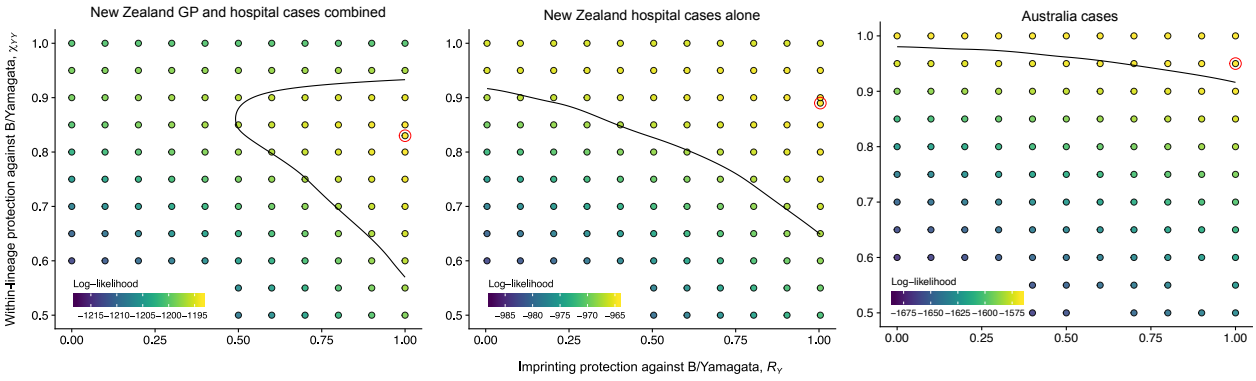

Figure 4.23: Bivariate likelihood profiles for within-lineage lineage protection against B/Yamagata and additional protection in people first infected with it (imprinting). The maximum-likelihood estimate is highlighted by the red circle, and the curve shows the 95% confidence region based on a likelihood-ratio test with 2 degrees of freedom under a LOESS fit.

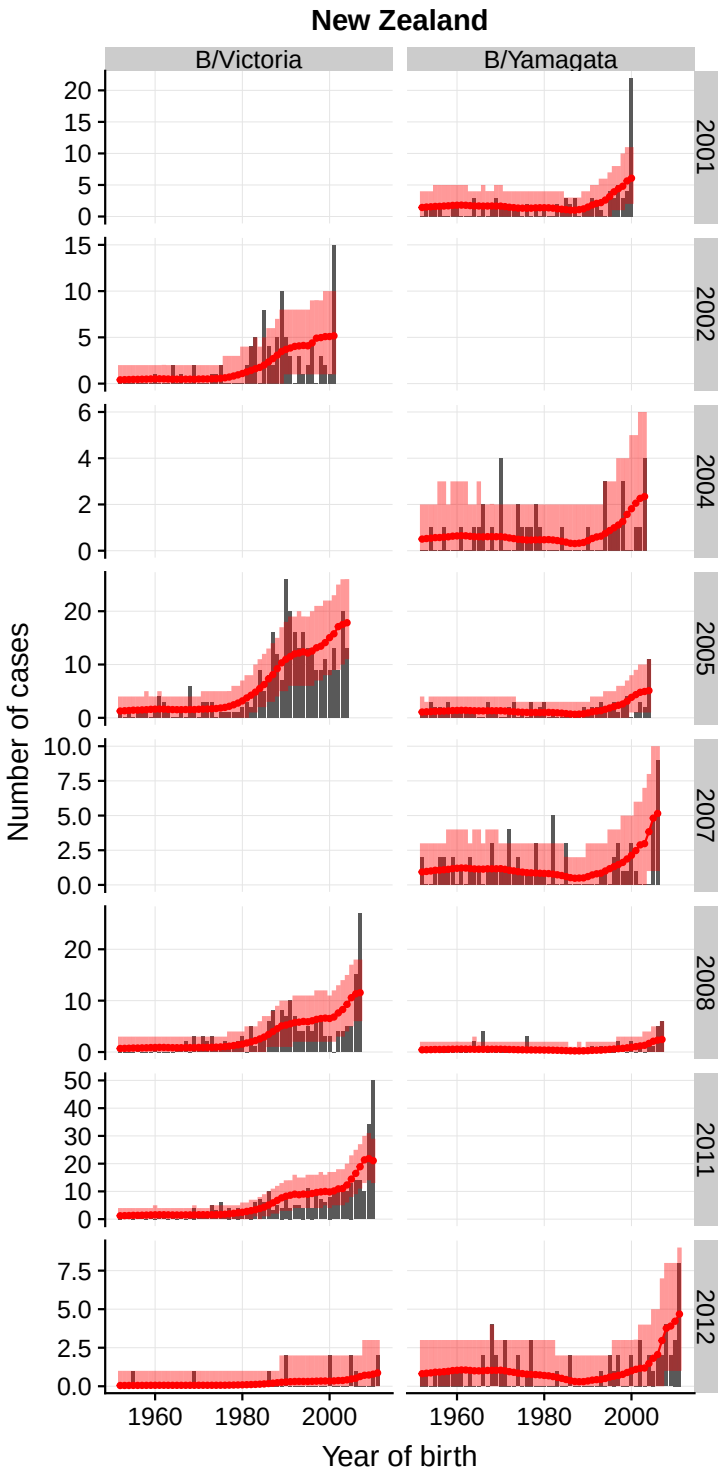

Figure 4.24: Observed and predicted distributions of hospital-associated (non-sentinel) influenza B cases in New Zealand by observation year. Vertical red bars are 95% bootstrap confidence intervals.

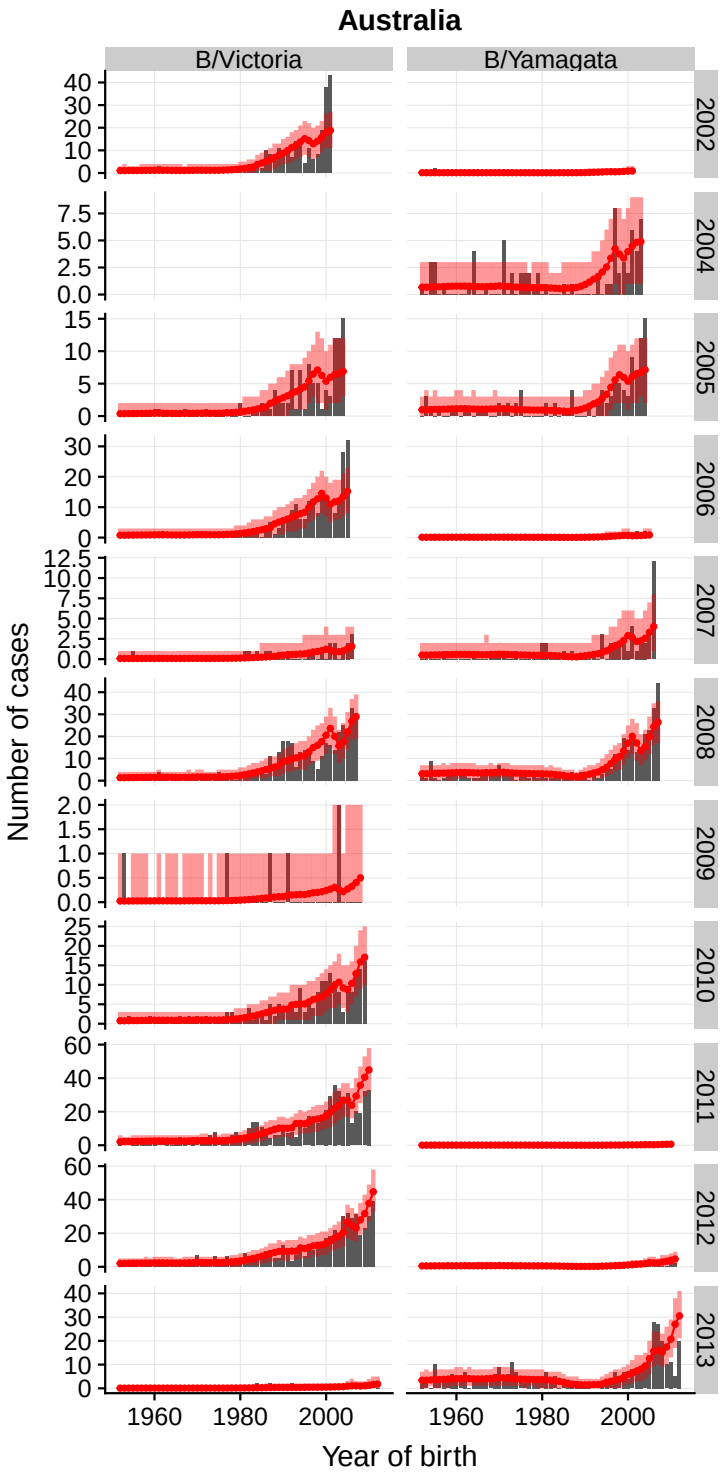

Figure 4.25: Observed and predicted distributions of influenza B cases in Australia by observation year. Vertical red bars are 95% bootstrap confidence intervals.

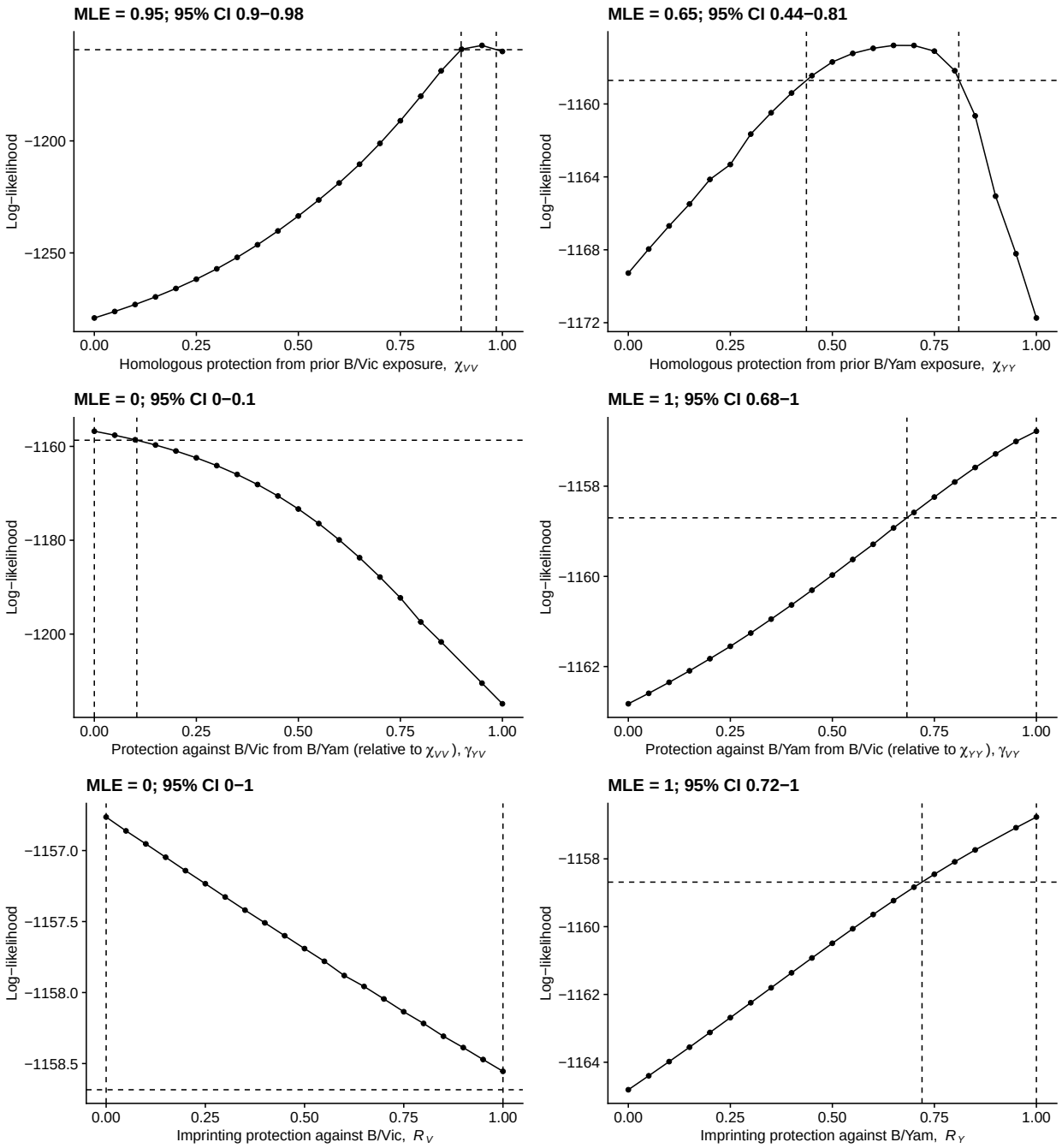

Figure 4.26: Likelihood profiles for protection parameters estimated from the complete New Zealand data assuming an alternative age cutoff for differential reporting in children (0-2 years old instead of 0-4). The 95% confidence interval based on a likelihood ratio test with one degree of freedom is indicated by the vertical dashed lines.

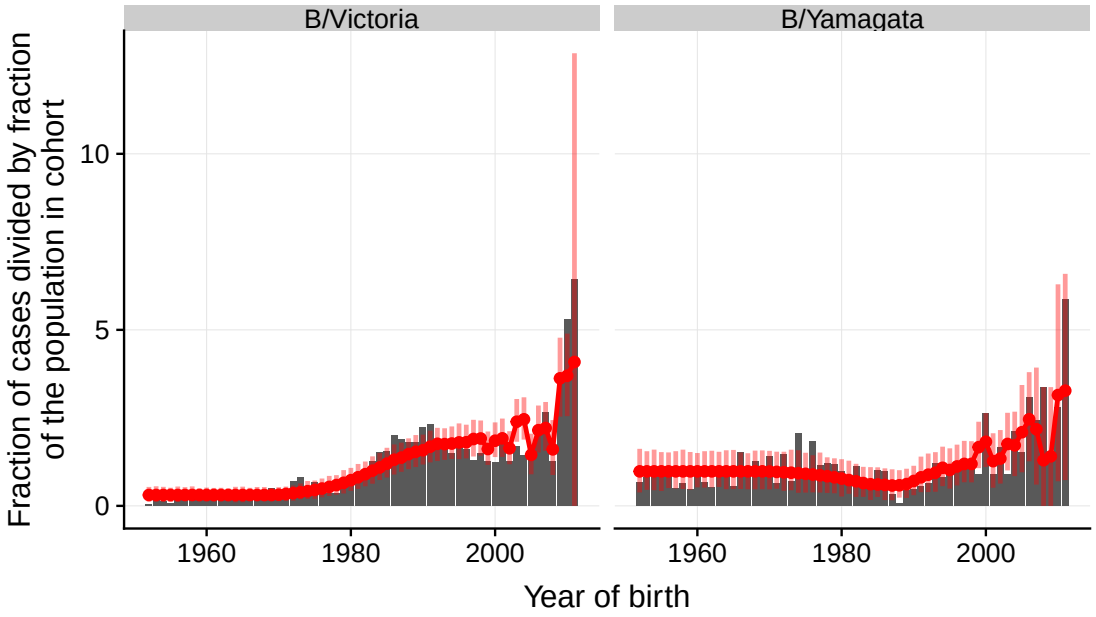

Figure 4.27: Model predictions for the complete New Zealand data assuming an alternative age cutoff for differential reporting in children (0-2 years old instead of 0-4). Predicted and observed fractions of cases were divided by the fraction of the population born in the corresponding birth year (i.e., the null expectation if all birth years were infected at the same rate). Red lines and dots show the predicted distribution under the model. Vertical bars are 95% bootstrap confidence intervals.

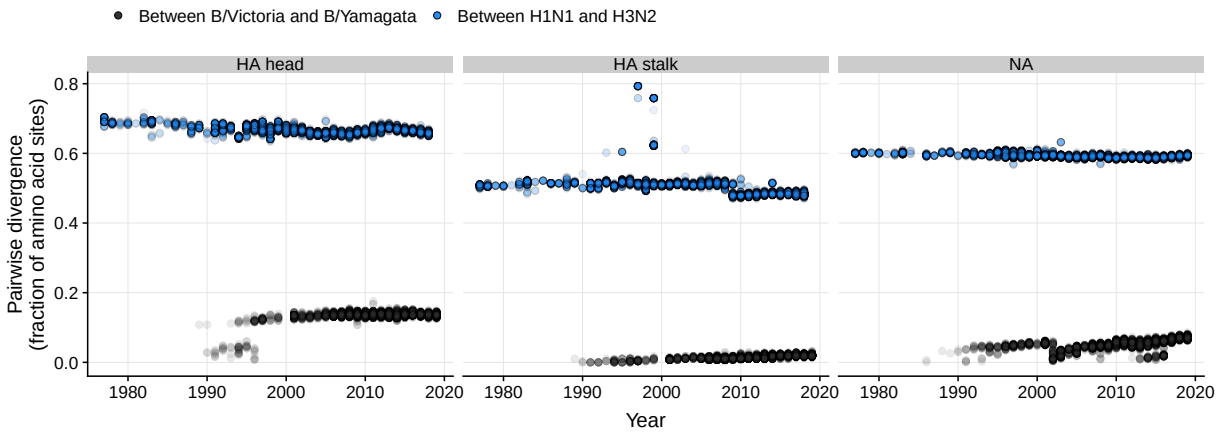

Figure 4.28: Amino acid divergence in the hemagglutinin and neuraminidase surface protein between influenza A subtypes and between influenza B lineages. Divergence in hemagglutinin is shown separately for the antigenically variable head region (left) and the more conserved stalk region (center). Each point represents a pair of strains isolated in the same year and deposited on GISAID. When more than 100 isolates from a subtype or lineage were isolated in a single year, we used a random sample of 100 isolates, resulting in up to 4,950 pairs of each kind in each year. Up to 500 randomly chosen pairs are shown for each year. Average divergence values reported in the main text are based on the full set of pairs given the randomly sampled sequences.

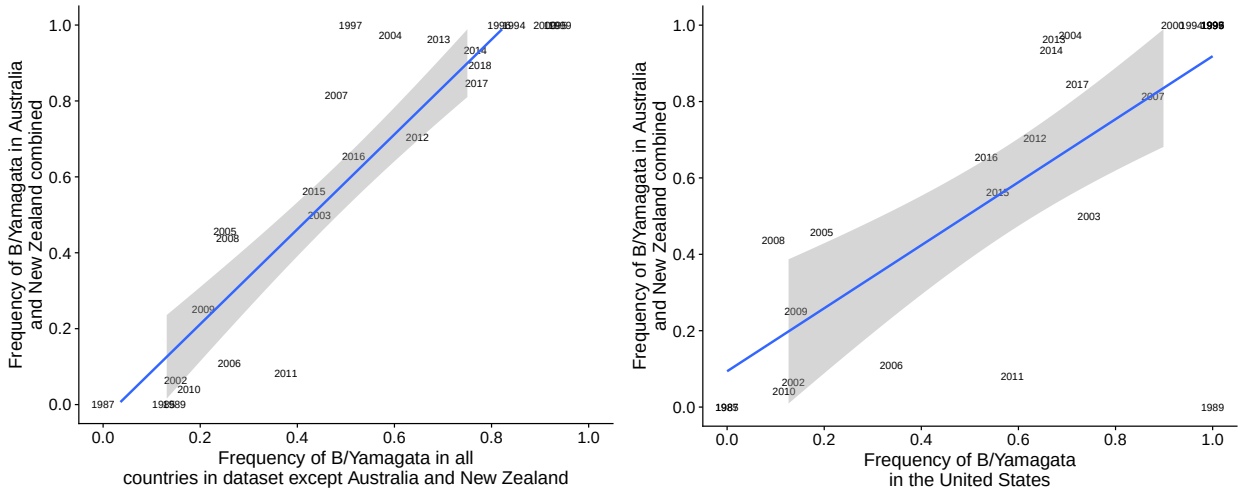

Figure 4.29: Proxies for B/Yamagata frequency in Australia and New Zealand for seasons with scarce data. We compared the frequency of B/Yamagata estimated from sequence isolates from Australia and New Zealand with the frequency estimated from isolates collected in all other countries represented in the dataset (left), or isolates collected in the United States (right). Because the former was a better proxy, we calculated frequencies using isolates from all countries represented in the data (including Australia and New Zealand) for seasons with fewer than 10 isolates in Australia and New Zealand.

CHAPTER 5

CONCLUSIONS

Faced with the problem of recognizing thousands of potential antigens, the evolution of jawed vertebrates produced not a fixed set of solutions but a system that uses recombination, mutation and selection to find solutions adaptively. The adaptability of B cell receptors in the short term is a direct result of selection acting on germline immunoglobulin genes in the long term. This dissertation investigated factors contributing to this short-term adaptability and how they change over the course of B cell evolution.

The results from Chapter 2 suggest that although the diversity of germline genes allows for many potential solutions to the problem of recognizing a particular pathogen, solutions based on some V genes are more successful than solutions based on others. Selection of V genes by specific pathogens during the immune response might help explain some of the phenotypic similarities observed in individual antibody responses. For instance, antibodies produced by different individuals often target the pathogen's antigens or the epitopes of a particular antigen in similar proportions. However, whether each V gene selected by a particular pathogen consistently targets a specific antigen or epitope is still unclear. Our results also suggest selected genes do not generally dominate the repertoire of experienced B cells in infected individuals, perhaps because the experienced repertoire encompasses responses against many antigens other than the ones in the infecting pathogen.

Chapter 3 found that the mutability of germline genes, selected over hundreds of millions of years to facilitate the short-term adaptation of B cell receptors, changes during the course of the immune response as neutral mutations and selection for affinity disrupt highly mutable "hotspot" motifs. While hotspots might maintain or increase their frequency by hitchhiking with beneficial mutations, which are more likely to appear in sequences containing many hotspots, we found no evidence of such indirect selection to maintain or increase mutability despite the sustained selective pressure imposed by a rapidly evolving HIV population. Since the experimental removal of hotspots has been shown to effectively decrease the mutation

rate [176, 73], the loss of mutational hotspots during the response might compromise the adaptability of B cell populations. While the decline in mutation rates due to the disruption of mutational hotspots might not impact the response to acute infections that lead to sterilizing immunity in a few weeks, loss of mutability might compromise the adaptability of B cell lineages evolving in response to chronic infections such as HIV. Loss of mutability might also compromise the adaptability of memory cells reentering germinal centers, for instance upon infection with a new influenza strain, although the extent to which memory cell reenter germinal centers is unclear [38, 115]. The ability to recruit new B cell lineages from the naive repertoire contributes to maintain the adaptability of B cell responses, since new lineages would have the initial distribution of mutational hotspots present in the germline genes. However, memory B cells can interfere with the generation of *de novo* responses [187, 188].

Chapter 4 found that protection against the two lineages of influenza type B arises from strong protection within lineages but asymmetric cross-protection between them. Differences in the memory from previous infections across human cohorts arose from historical changes in lineage frequencies. The resulting differences in cohorts' protection against each lineage can explain the differences in the age distributions of medically attended infections with B/Victoria and B/Yamagata. To explain the asymmetry in cross-lineage protection between B/Victoria and B/Yamagata, we proposed a model in which the response to primary infections with each lineage targets different epitopes with different degrees of conservation between the lineages. These differences in epitope targeting might be associated with selection of different V genes in people with different histories of infection with influenza B, a hypothesis that remains to be tested.

Implications and future directions

Based on our findings and previous work, we hypothesize that immunoglobulin genes might in the long term under potentially conflicting selective pressures favoring diversity and adaptability, on the one hand, and specialization for common pathogens, on the other. To rec-

ognize the vast number of antigens most species encounter, selection seems to favor the evolution of new immunoglobulin genes by gene duplication and their molecular diversification [110, 50, 34, 51]. To facilitate short-term evolution of higher affinity, selection in the long term also seems to favor mutations to germline genes that make antigen binding regions more mutable and structurally sensitive regions less mutable during short-term B cell evolution [122, 175, 81, 67, 140]. However, in addition to this selection for diversity and adaptability, selection might favor mutations that improve recognition of commonly encountered pathogens by specific immunoglobulin genes [28]. Instead of being repeatedly acquired by somatic mutation during the short-term evolution of B cell receptors, mutations that improve recognition of common pathogens would thus become hard-coded in the immunoglobulin genes themselves. This hard-coded protection would be similar to the one provided by innate immunity but might be easier to evolve, since functional constraints on germline immunoglobulin genes are likely weaker than those on innate receptors that do not arise from recombination and instead have specific structures and functions.

Specialization of particular genes for a specific pathogen requires variation in germline genes' propensity to recombine into receptors able to recognize the pathogen's antigens. Consistent with this hypothesis, we found V genes particularly good at binding influenza antigens in mice. Influenza does not appear to naturally infect mice in the wild and the two therefore have little shared evolutionary history, suggesting the variation in the ability to recognize specific pathogens can arise by chance as a byproduct of immunoglobulin genes' long-term diversification.

Specialization of germline genes to recognize common pathogens might come at the cost of reduced diversity and adaptability. For instance, selection to increase the number of copies of an immunoglobulin gene that recognizes a common pathogen might increase its frequency among B cell receptors and thus decrease repertoire diversity. Different immunoglobulin genes might independently evolve similar mutations that improve binding to a common antigen, decreasing their molecular diversity. Selection might also favor fewer mutational

hotspots in immunoglobulin genes that recognize common pathogens to prevent this binding from being disrupted by somatic mutations during short-term B cell evolution. Fewer mutational hotspots might then decrease adaptability in response to other pathogens.

We hypothesize that life-history differences could lead to stronger selection for diversity and adaptability in some species and stronger selection to recognize common pathogens in others. Lifespan differences, in particular, affect how species experience the "pathogenic environment" around them, defined as the set of pathogens individuals can potentially encounter during their lifespan, the relative frequencies of those pathogens and their virulence [111, 112]. Mathematical models suggest that minimizing the cost imposed by the pathogenic environment requires many receptors able to recognize common pathogens but also, counterintuitively, more receptors able to recognize rare pathogens than expected based on their frequency [111]. Absent from those models, however, is how life-history parameters affect the relative burden of rare pathogens. For instance, short-lived organisms might have limited opportunity to encounter rare pathogens and might thus be expected to allocate proportionally more receptors to common enemies than to rare ones. Long-lived species, in contrast, have more opportunity to encounter rare pathogens, and because long-lived species are also typically larger and thus have more B cells, they can afford to have more receptors targeting rare pathogens [28]. Thus, if the burden of rare pathogens is stronger in long-lived species, immunoglobulin genes from long-lived species should experience stronger selection for diversity and adaptability, while short-lived organisms should experience stronger selection for hard-coded recognition of common pathogens. Long-lived species should also have more opportunity to accumulate memory at the population level, leading to stronger selection for pathogens to escape existing antibodies. Faster antigenic evolution of pathogens would then further increase the benefit of highly adaptable germline genes in long-lived species. Broadly consistent with these hypotheses, mouse immunoglobulin genes undergo fewer nucleotide insertions and deletions when recombining to produce B cell receptors compared to human genes [29]. As a result, receptor specificities seem to be more strongly determined by the

germline sequences of immunoglobulin genes in mice than in humans [29, 28].

The pervasive mutability losses we found to be inherent to the evolution of B cell lineages might impact short- and long-lived vertebrates differently. Loss of adaptability due to a decline in the mutation rate might be more important in long-lived species than in short-lived ones because long-lived organisms can have longer chronic infections and have more opportunity to be reinfected with antigenically novel strains of the same pathogen. Long-term selection might therefore favor higher mutability in the germline immunoglobulin genes of long-lived species compared with short-lived ones, so that the B cell receptors of long-lived species could withstand extensive loss of mutability while maintaining some adaptability. This difference might be attenuated by the ability to recruit new B cell lineages from the naive repertoire.

In addition to traits that affect organisms' perception of their pathogenic environment, such as lifespan, the pathogenic environment itself can vary in time and space due to biotic and abiotic factors. Selection might favor increased immunoglobulin gene diversity in regions of higher pathogen diversity, and specialization for different pathogens common in different parts of species' geographic ranges might lead to sustained polymorphisms in immunoglobulin genes.

These hypotheses could be formalized using mathematical models and tested with comparative analyses of immunoglobulin genes across vertebrates. Incorporating lifespan and body size into models linking repertoire structure to the pathogenic environment [111] could generate predictions for how those parameters affect the number of receptors targeting rare and common pathogens in an optimal repertoire. We expect that long-lived species will be predicted to allocate proportionally more receptors targeting rare pathogens than will short-lived species. Those models can also be extended to capture the effect of longevity on the population-level strength of immune memory and thus the rate of antigenic evolution by pathogens.

Because the somatic hypermutation rate partly determines the the adaptability of B cell

receptors, stronger selection for adaptability in long-lived species might lead to more mutational hotspots in the germline genes of long-lived species than in those of short-lived ones. This prediction could be tested by constructing empirical models that quantify the mutability of immunoglobulin genes based on their sequence [185, 47]. By comparing the observed number of mutational hotspots with the expected number under null models that randomize codon usage, it is possible to quantify how much selection has optimized immunoglobulin gene sequences to increase mutability in antigen binding regions and reduce it in structurally important ones [122, 175, 81, 67, 140]. High mutability in antigen binding regions relative to the null expectation should increase with species' longevity. Because mutational hotspots can differ between species [31], different mutability models could be constructed for different species using receptor sequence datasets containing synonymous mutations or unproductive sequences which undergo mutations but not selection. One limitation to this approach is that knowledge of the germline immunoglobulin genes is scarce for many groups, which limits the power of comparative analyses of those genes [33]. However, mutability might still be estimated using certain motifs found to be highly mutable across different vertebrate groups (e.g., AGCT in humans [185], mice [31], horses [158] and fish [186]).

This dissertation contributed to this research program by investigating the short-term evolution of B cell receptors. We have shown there is significant variation in immunoglobulin genes' propensity to recognize specific pathogens, a necessary though not sufficient condition for the specialization of immunoglobulin genes via long-term selection on their germline sequences. Our results also shed light on the loss of mutability in B cell receptors during the immune response, which potentially leads to a short-term loss of adaptability that might affect species differently depending on their lifespan. Understanding the selective pressures for adaptability and specialization of immunoglobulin genes in light of these short-term evolutionary dynamics will help us understand the long-term evolution of adaptive immunity in vertebrates.

REFERENCES

[1] Australian National Influenza Surveillance Scheme - Annual reports. Available at https://www.health.gov.au/internet/main/publishing.nsf/content/cda-pubs-annlrpt-fluannrep.htm (accessed 6/21/2019 at 2:13 p.m.).

[2] HA Subtype Numbering Conversion (beta) - Influenza Research Database. Available at https://www.fludb.org/brc/hanumbering.spg? (accessed 8/27/2019 at 5:10 p.m.).

[3] New Zealand Institute of Environmental Science and Research Influenza Annual Reports. Available at `https://surv.esr.cri.nz/virology/influenza_annual_report.php` (accessed 6/21/2019 at 2:23 p.m.).

[4] Primary education in the Netherlands. Available at `https://www.government.nl/topics/primary-education` (accessed 1/11/2020 at 6:22 p.m.).

[5] StatsNZ Infoshare - Estimated Resident Population by Age and Sex (1991+) (Annual-Dec). Available at `http://archive.stats.govt.nz/infoshare/` (accessed 8/1/2019 at 4:54 p.m.).

[6] D. Allen, T. Simon, F. Sablitzky, K. Rajewsky, and A. Cumano. Antibody engineering for the analysis of affinity maturation of an anti-hapten response. *The EMBO Journal*, 7(7):1995–2001, July 1988.

[7] R. M. Anderson, J. A. Crombie, and B. T. Grenfell. The epidemiology of mumps in the UK: a preliminary study of virus transmission, herd immunity and the potential impact of immunization. *Epidemiology & Infection*, 99(1):65–84, August 1987.

[8] Sarah F Andrews, Yunping Huang, Kaval Kaur, Lyubov I Popova, Irvin Y Ho, Noel T Pauli, Carole J Henry Dunand, William M Taylor, Samuel Lim, Min Huang, Xinyan Qu, Jane-Hwei Lee, Marlene Salgado-Ferrer, Florian Krammer, Peter Palese, Jens

Wrammert, Rafi Ahmed, and Patrick C Wilson. Immune history profoundly affects broadly protective B cell responses to influenza. *Science Translational Medicine*, 7(316):316ra192, December 2015.

[9] Davide Angeletti, James S. Gibbs, Matthew Angel, Ivan Kosik, Heather D. Hickman, Gregory M. Frank, Suman R. Das, Adam K. Wheatley, Madhu Prabhakaran, David J. Leggat, Adrian B. McDermott, and Jonathan W. Yewdell. Defining B cell immunodominance to viruses. *Nature Immunology*, 18(4):456–463, April 2017.

[10] Philip Arevalo, Huong Q. McLean, Edward A. Belongia, and Sarah Cobey. Earliest infections predict the age distribution of seasonal influenza A cases. *medRxiv*, page 19001875, July 2019.

[11] Yuval Avnir, Aimee S. Tallarico, Quan Zhu, Andrew S. Bennett, Gene Connelly, Jared Sheehan, Jianhua Sui, Amr Fahmy, Chiung-yu Huang, Greg Cadwell, Laurie A. Bankston, Andrew T. McGuire, Leonidas Stamatatos, Gerhard Wagner, Robert C. Liddington, and Wayne A. Marasco. Molecular Signatures of Hemagglutinin Stem-Directed Heterosubtypic Human Neutralizing Antibodies against Influenza A Viruses. *PLOS Pathogens*, 10(5):e1004103, May 2014.

[12] Trevor Bedford, Andrew Rambaut, and Mercedes Pascual. Canalization of the evolutionary trajectory of the human influenza virus. *BMC Biology*, 10(1):38, January 2012.

[13] Trevor Bedford, Steven Riley, Ian G. Barr, Shobha Broor, Mandeep Chadha, Nancy J. Cox, Rodney S. Daniels, C. Palani Gunasekaran, Aeron C. Hurt, Anne Kelso, Alexander Klimov, Nicola S. Lewis, Xiyan Li, John W. McCauley, Takato Odagiri, Varsha Potdar, Andrew Rambaut, Yuelong Shu, Eugene Skepner, Derek J. Smith, Marc A. Suchard, Masato Tashiro, Dayan Wang, Xiyan Xu, Philippe Lemey, and Colin A. Rus-

sell. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523:217–220, June 2015.

[14] Trevor Bedford, Marc a Suchard, Philippe Lemey, Gytis Dudas, Victoria Gregory, Alan J Hay, John W McCauley, Colin a Russell, Derek J Smith, and Andrew Rambaut. Integrating influenza antigenic dynamics with molecular evolution. *eLife*, 3:e01914, January 2014.

[15] Robert B. Belshe, Kathleen Coelingh, Christopher S. Ambrose, Jennifer C. Woo, and Xionghua Wu. Efficacy of live attenuated influenza vaccine in children against influenza B viruses by lineage and antigenic similarity. *Vaccine*, 28(9):2149–2156, February 2010.

[16] Jesse D. Bloom, Sy T. Labthavikul, Christopher R. Otey, and Frances H. Arnold. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences*, 103(15):5869–5874, April 2006.

[17] R. Bodewes, G. de Mutsert, F. R. M. van der Klis, M. Ventresca, S. Wilks, D. J. Smith, M. Koopmans, R. a. M. Fouchier, A. D. M. E. Osterhaus, and G. F. Rimmelzwaan. Prevalence of Antibodies against Seasonal Influenza A and B Viruses in Children in Netherlands. *Clinical and Vaccine Immunology*, 18(3):469–476, March 2011.

[18] D. Bonhomme, L. Hammarstrm, D. Webster, H. Chapel, O. Hermine, F. Le Deist, E. Lepage, P. H. Romeo, and Yves Levy. Impaired Antibody Affinity Maturation Process Characterizes a Subset of Patients with Common Variable Immunodeficiency. *The Journal of Immunology*, 165(8):4725–4730, October 2000.

[19] Christine Brack, Minoru Hirama, Rita Lenhard-Schuller, and Susumu Tonegawa. A complete immunoglobulin gene is created by somatic recombination. *Cell*, 15(1):1–14, September 1978.

[20] Barak Brill, Amnon Amir, and Ruth Heller. Testing for differential abundance in

compositional counts data, with application to microbiome studies. *arXiv:1904.08937 [q-bio, stat]*, August 2019.

[21] David F. Burke and Derek J. Smith. A Recommended Numbering Scheme for Influenza A HA Subtypes. *PLOS ONE*, 9(11):e112302, November 2014.

[22] M Burnet. *The clonal selection theory of acquired immunity.* Cambridge University Press, Cambrige, UK, 1959.

[23] Dennis R Burton, Pascal Poignard, Robyn L Stanfield, and Ian A Wilson. Broadly neutralizing antibodies present new prospects to counter highly antigenically diverse viruses. *Science*, 337(6091):183–6, July 2012.

[24] R. M. Bush, W. M. Fitch, C. A. Bender, and N. J. Cox. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular Biology and Evolution*, 16(11):1457–1465, November 1999.

[25] Saverio Caini, Gabriela Kusznierz, Vernica Vera Garate, Sonam Wangchuk, Binay Thapa, Francisco Jos de Paula Jnior, Walquiria Aparecida Ferreira de Almeida, Richard Njouom, Rodrigo A. Fasce, Patricia Bustos, Luzhao Feng, Zhibin Peng, Jenny Lara Araya, Alfredo Bruno, Domnica de Mora, Mnica Jeannette Barahona de Gmez, Richard Pebody, Maria Zambon, Rocio Higueros, Rudevelinda Rivera, Herman Kosasih, Maria Rita Castrucci, Antonino Bella, Herv A. Kadjo, Coulibaly Daouda, Ainash Makusheva, Olga Bessonova, Sandra S. Chaves, Gideon O. Emukule, Jean-Michel Heraud, Norosoa H. Razanajatovo, Amal Barakat, Fatima El Falaki, Adam Meijer, G A. Donker, Q. Sue Huang, Tim Wood, Angel Balmaseda, Rakhee Palekar, Brechla Moreno Arvalo, Ana Paula Rodrigues, Raquel Guiomar, Vernon Jian Ming Lee, Li Wei Ang, Cheryl Cohen, Florette Treurnicht, Alla Mironenko, Olha Holubka, Joseph Bresee, Lynnette Brammer, Mai T. Q. Le, Phuong V. M. Hoang, Clotilde El Guerche-Sblain, John Paget, and the Global Influenza B. Study Team. The epidemio-

logical signature of influenza B virus and its B/Victoria and B/Yamagata lineages in the 21st century. *PLOS ONE*, 14(9):e0222381, September 2019.

[26] Rubing Chen and Edward C. Holmes. The Evolutionary Dynamics of Human Influenza B Virus. *Journal of Molecular Evolution*, 66(6):655, May 2008.

[27] Sarah Cobey and Scott E Hensley. Immune history and influenza virus susceptibility. *Current Opinion in Virology*, 22:105–111, February 2017.

[28] Andrew M. Collins and Katherine J. L. Jackson. On being the right size: antibody repertoire formation in the mouse and human. *Immunogenetics*, 70(3):143–158, March 2018.

[29] Andrew M. Collins, Yan Wang, Krishna M. Roskin, Christopher P. Marquis, and Katherine J. L. Jackson. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676):20140236, September 2015.

[30] Bernardo Cortina-Ceballos, Elizabeth Ernestina Godoy-Lozano, Juan Tllez-Sosa, Marbella Ovilla-Muoz, Hugo Smano-Snchez, Andrs Aguilar-Salgado, Rosa Elena Gmez-Barreto, Humberto Valdovinos-Torres, Irma Lpez-Martnez, Rodrigo Aparicio-Antonio, Mario H. Rodrguez, and Jess Martnez-Barnetche. Longitudinal analysis of the peripheral B cell repertoire reveals unique effects of immunization with a new influenza virus strain. *Genome Medicine*, 7(1):124, November 2015.

[31] Ang Cui, Roberto Di Niro, Jason A. Vander Heiden, Adrian W. Briggs, Kris Adams, Tamara Gilbert, Kevin C. OConnor, Francois Vigneault, Mark J. Shlomchik, and Steven H. Kleinstein. A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig Sequencing Data. *The Journal of Immunology*, 197(9):3566–3574, November 2016.

[32] Ana Cumano and Klaus Rajewsky. Structure of primary anti-(4-hydroxy-3-nitrophenyl)acetyl (NP) antibodies in normal and idiotypically suppressed C57BL/6 mice. *European Journal of Immunology*, 15(5):512–520, 1985.

[33] Sabyasachi Das, Masayuki Hirano, Chelsea McCallister, Rea Tako, and Nikolas Nikolaidis. Chapter 4 - Comparative Genomics and Evolution of Immunoglobulin-Encoding Loci in Tetrapods. In Frederick W. Alt, editor, *Advances in Immunology*, volume 111, pages 143–178. Academic Press, January 2011.

[34] Sabyasachi Das, Masafumi Nozawa, Jan Klein, and Masatoshi Nei. Evolutionary dynamics of the immunoglobulin heavy chain variable region genes in vertebrates. *Immunogenetics*, 60(1):47–55, January 2008.

[35] F. M. Davenport, A. V. Hennessy, C. H. Stuart-Harris, and T. Francis. Epidemiology of Influenza. Comparative Serological Observations in England and the United States. *Lancet*, pages 469–74, 1955.

[36] Fred M. Davenport and Albert V. Hennessy. Predetermination by Infection and by Vaccination of Antibody Response to Influenza Virus Vaccines. *Journal of Experimental Medicine*, 106(6):835–850, December 1957.

[37] J. R. Davies, E. A. Grilli, and A. J. Smith. Influenza A: infection and reinfection. *Epidemiology & Infection*, 92(1):125–127, February 1984.

[38] Ismail Dogan, Barbara Bertocci, Valrie Vilmont, Frdric Delbos, Jrome Mgret, Sbastien Storck, Claude-Agns Reynaud, and Jean-Claude Weill. Multiple layers of B cell memory with different effector functions. *Nature Immunology*, 10(12):1292–1299, December 2009.

[39] Nicole A Doria-Rose, Chaim A Schramm, Jason Gorman, Penny L Moore, Jinal N Bhiman, Brandon J DeKosky, Michael J Ernandes, Ivelin S Georgiev, Helen J Kim,

Marie Pancera, Ryan P Staupe, Han R Altae-Tran, Robert T Bailer, Ema T Crooks, Albert Cupo, Aliaksandr Druz, Nigel J Garrett, Kam H Hoi, Rui Kong, Mark K Louder, Nancy S Longo, Krisha McKee, Molati Nonyane, Sijy O'Dell, Ryan S Roark, Rebecca S Rudicell, Stephen D Schmidt, Daniel J Sheward, Cinque Soto, Constantinos Kurt Wibmer, Yongping Yang, Zhenhai Zhang, James C Mullikin, James M Binley, Rogier W Sanders, Ian A Wilson, John P Moore, Andrew B Ward, George Georgiou, Carolyn Williamson, Salim S Abdool Karim, Lynn Morris, Peter D Kwong, Lawrence Shapiro, and John R Mascola. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature*, 509(7498):55–62, 2014.

[40] Cyrille Dreyfus, Nick S. Laursen, Ted Kwaks, David Zuijdgeest, Reza Khayat, Damian C. Ekiert, Jeong Hyun Lee, Zoltan Metlagel, Miriam V. Bujny, Mandy Jongeneelen, Remko van der Vlugt, Mohammed Lamrani, Hans J. W. M. Korse, Eric Geelen, zcan Sahin, Martijn Sieuwerts, Just P. J. Brakenhoff, Ronald Vogels, Olive T. W. Li, Leo L. M. Poon, Malik Peiris, Wouter Koudstaal, Andrew B. Ward, Ian A. Wilson, Jaap Goudsmit, and Robert H. E. Friesen. Highly Conserved Protective Epitopes on Influenza B Viruses. *Science*, 337(6100):1343–1348, September 2012.

[41] A J Drummond and M A Suchard. Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, 8(114), 2010.

[42] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973, August 2012.

[43] Gytis Dudas, Trevor Bedford, Samantha Lycett, and Andrew Rambaut. Reassortment between Influenza B Lineages and the Emergence of a Coadapted PB1PB2HA Gene Complex. *Molecular Biology and Evolution*, 32(1):162–172, January 2015.

[44] Dunand Carole J. Henry and Wilson Patrick C. Restricted, canonical, stereotyped and

convergent immunoglobulin responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676):20140238, September 2015.

[45] C. T. T. Edwards, E. C. Holmes, O. G. Pybus, D. J. Wilson, R. P. Viscidi, E. J. Abrams, R. E. Phillips, and A. J. Drummond. Evolution of the Human Immunodeficiency Virus Envelope Gene Is Dominated by Purifying Selection. *Genetics*, 174(3), 2006.

[46] Yuval Elhanati, Anand Murugan, Curtis G. Callan, Thierry Mora, and Aleksandra M. Walczak. Quantifying selection in immune receptor repertoires. *Proceedings of the National Academy of Sciences*, 111(27):9875–9880, July 2014.

[47] Yuval Elhanati, Zachary Sethna, Quentin Marcou, Curtis G Callan, Thierry Mora, and Aleksandra M Walczak. Inferring processes underlying B-cell repertoire diversity. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1676):20140243, September 2015.

[48] Adam Eyre-Walker and Peter D. Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–618, August 2007.

[49] Neil M. Ferguson, Alison P. Galvani, and Robin M. Bush. Ecological and immunological determinants of influenza evolution. *Nature*, 422(6930):428–433, March 2003.

[50] Martin F. Flajnik. Comparative analyses of immunoglobulin genes: surprises and portents. *Nature Reviews Immunology*, 2(9):688, September 2002.

[51] Martin F. Flajnik and Masanori Kasahara. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics*, 11(1):47–59, January 2010.

[52] J. M. Fonville, S. H. Wilks, S. L. James, A. Fox, M. Ventresca, M. Aban, L. Xue, T. C. Jones, Y. Wong, A. Mosterin, L. C. Katzelnick, D. Labonte, G. van der Net,

E. Skepner, C. a. Russell, T. D. Kaplan, G. F. Rimmelzwaan, N. Masurel, J. C. de Jong, A. Palache, W. E. P. Beyer, H. F. L. Wertheim, a. C. Hurt, a. D. M. E. Osterhaus, I. G. Barr, R. a. M. Fouchier, P. W. Horby, and D. J. Smith. Antibody landscapes after influenza virus infection or vaccination. *Science*, 346(6212):996–1000, November 2014.

[53] Judith M. Fonville, Pieter L. A. Fraaij, Gerrie de Mutsert, Samuel H. Wilks, Ruud van Beek, Ron A. M. Fouchier, and Guus F. Rimmelzwaan. Antigenic Maps of Influenza A(H3N2) Produced With Human Antisera Obtained After Primary Infection. *The Journal of Infectious Diseases*, 213(1):31–38, January 2016.

[54] Arthur L. Frank and Larry H. Taber. Variation in frequency of natural reinfection with influenza A viruses. *Journal of Medical Virology*, 12(1):17–23, 1983.

[55] Daniela Frasca and Bonnie B Blomberg. Effects of aging on B cell function. *Current Opinion in Immunology*, 21(4):425–430, August 2009.

[56] N. S. Galbraith, JenniferJ. Pusey, SusanE. J. Young, D. L. Crombie, and J. P. Sparks. Mumps Sruveillance in England and Wales 1962-81. *The Lancet*, 323(8368):91–94, January 1984.

[57] Rebecca Garten, L Blanton, A. I. A. Elal, N Alabi, J Barnes, M Biggerstaff, L Brammer, Alicia P Budd, Erin Burns, C. N. Cummins, T Davis, Shikha Garg, L. Gubareva, Y Jang, K Kniss, N Kramer, S Lindstrom, D Mustaquim, A O'Halloran, W Sessions, C Taylor, Xu Xiyan, V. G Dugan, A. M. Fry, David E Wentworth, J. Katz, and D Jernigan. Update: Influenza Activity in the United States During the 201718 Season and Composition of the 201819 Influenza Vaccine. *MMWR. Morbidity and Mortality Weekly Report*, 67:634–642, 2018.

[58] Antoine Giraud, Ivan Matic, Olivier Tenaillon, Antonio Clara, Miroslav Radman, Michel Fons, and Franois Taddei. Costs and Benefits of High Mutation Rates: Adap-

tive Evolution of Bacteria in the Mouse Gut. *Science*, 291(5513):2606–2608, March 2001.

[59] Miroslaw K. Gorny, Xiao-Hong Wang, Constance Williams, Barbara Volsky, Kathy Revesz, Bradley Witover, Sherri Burda, Mateusz Urbanski, Phillipe Nyambi, Chavdar Krachmarov, Abraham Pinter, Susan Zolla-Pazner, and Arthur Nadas. Preferential use of the VH5-51 gene segment by the human immune response to code for antibodies against the V3 domain of HIV-1. *Molecular Immunology*, 46(5):917–926, February 2009.

[60] Katelyn M. Gostic, Monique Ambrose, Michael Worobey, and James O. Lloyd-Smith. Potent protection against H5N1 and H7N9 influenza via childhood hemagglutinin imprinting. *Science*, 354(6313):722–726, November 2016.

[61] Katelyn M. Gostic, Rebecca Bridge, Shane Brady, Ccile Viboud, Michael Worobey, and James O. Lloyd-Smith. Childhood immune imprinting to influenza A shapes birth year-specific risk during seasonal H1N1 and H3N2 epidemics. *PLOS Pathogens*, 15(12):e1008109, December 2019.

[62] Bryan T Grenfell and R. M. Anderson. The estimation of age-related rates of infection from case notifications and serological data. *Journal of Hygiene*, 95(2):419–436, 1985.

[63] Gillian M. Griffiths, Claudia Berek, Matti Kaartinen, and Cesar Milstein. Somatic mutation and the maturation of immune response to 2-phenyl oxazolone. *Nature*, 312(5991):271, November 1984.

[64] Sunetra Gupta, Neil Ferguson, and Roy Anderson. Chaos, Persistence, and Evolution of Strain Structure in Antigenically Diverse Infectious Agents. *Science*, 280(5365):912–915, May 1998.

[65] Barton F. Haynes, Garnett Kelsoe, Stephen C. Harrison, and Thomas B. Kepler. B-cell

lineage immunogen design in vaccine development with HIV-1 as a case study. *Nature Biotechnology*, 30(5):423–433, May 2012.

[66] Andrew C Hayward, Ellen B Fragaszy, Alison Bermingham, Lili Wang, Andrew Copas, W John Edmunds, Neil Ferguson, Nilu Goonetilleke, Gabrielle Harvey, Jana Kovar, Megan S C Lim, Andrew McMichael, Elizabeth R C Millett, Jonathan S Nguyen-Van-Tam, Irwin Nazareth, Richard Pebody, Faiza Tabassum, John M Watson, Fatima B Wurie, Anne M Johnson, and Maria Zambon. Comparative community burden and severity of seasonal and pandemic influenza: results of the Flu Watch cohort study. *The Lancet Respiratory Medicine*, 2(6):445–454, June 2014.

[67] Uri Hershberg and Mark J Shlomchik. Differences in potential for amino acid change after mutation reveals distinct strategies for kappa and lambda light-chain variation. *Proceedings of the National Academy of Sciences of the United States of America*, 103(43):15963–8, October 2006.

[68] Kenneth B Hoehn, Gerton Lunter, and Oliver Pybus. A Phylogenetic Codon Substitution Model for Antibody Lineages. *Genetics*, 206:417–427, 2017.

[69] Sandra Hopkins and Nathan Speed. The decline in free general practitioner care in Australia: reasons and repercussions. *Health Policy*, 73(3):316–329, September 2005.

[70] Peter Horby, Le Quynh Mai, Annette Fox, Pham Quang Thai, Nguyen Thi Thu Yen, Le Thi Thanh, Nguyen Le Khanh Hang, Tran Nhu Duong, Dang Dinh Thoang, Jeremy Farrar, Marcel Wolbers, and Nguyen Tran Hien. The Epidemiology of Interpandemic and Pandemic Influenza in Vietnam, 20072010The Ha Nam Household Cohort Study I. *American Journal of Epidemiology*, 175(10):1062–1074, May 2012.

[71] N. Hozumi and S. Tonegawa. Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proceedings of the National Academy of Sciences*, 73(10):3628–3632, October 1976.

[72] Q. Sue Huang, Don Bandaranayake, Tim Wood, E. Claire Newbern, Ruth Seeds, Jacqui Ralston, Ben Waite, Ange Bissielo, Namrata Prasad, Angela Todd, Lauren Jelley, Wendy Gunn, Anne McNicholas, Thomas Metz, Shirley Lawrence, Emma Collis, Amanda Retter, Sook-san Wong, Richard Webby, Judy Bocacao, Jennifer Haubrock, Graham Mackereth, Nikki Turner, Barbara McArdle, John Cameron, Edwin G. Reynolds, Michael G. Baker, Cameron C. Grant, Colin McArthur, Sally Roberts, Adrian Trenholme, Conroy Wong, Susan Taylor, Paul Thomas, Jazmin Duque, Diane Gross, Mark G. Thompson, Marc-Alain Widdowson, Kathryn Haven, Bhamita Chand, Pamela Muponisi, Debbie Aley, Claire Sherring, Miriam Rea, Judith Barry, Tracey Bushell, Julianne Brewer, Catherine McClymont, Shona Chamberlin, Reniza Ongcoy, Kirstin Davey, Emilina Jasmat, Maree Dickson, Annette Western, Olive Lai, Sheila Fowlie, Faasoa Aupaau, Louise Robertson, Pam Kawakami, Susan Walker, Robyn Madge, Amanda des Barres, Helen Qiao, Fifi Tse, Mahtab Zibaei, Tirzah Korrapadu, Louise Optland, and Cecilia Dela Cruz. Risk Factors and Attack Rates of Seasonal Influenza Infection: Results of the Southern Hemisphere Influenza and Vaccine Effectiveness Research and Surveillance (SHIVERS) Seroepidemiologic Cohort Study. *The Journal of Infectious Diseases*, 219(3):347–357, January 2019.

[73] Joyce K. Hwang, Chong Wang, Zhou Du, Robin M. Meyers, Thomas B. Kepler, Donna Neuberg, Peter D. Kwong, John R. Mascola, M. Gordon Joyce, Mattia Bonsignori, Barton F. Haynes, Leng-Siew Yeap, and Frederick W. Alt. Sequence intrinsic somatic mutation mechanisms contribute to affinity maturation of VRC01-class HIV-1 broadly neutralizing antibodies. *Proceedings of the National Academy of Sciences*, 114(32):8614–8619, August 2017.

[74] Martin Ivanovski, Federico Silvestri, Gabriele Pozzato, Shubha Anand, Cesare Mazzaro, Oscar R. Burrone, and D. G. Efremov. Somatic hypermutation, clonal diversity, and preferential expression of the VH 51p1/VL kv325 immunoglobulin gene combina-

tion in hepatitis C virus-associated immunocytomas. *Blood*, 91(7):2433–2442, 1998.

[75] Katherine J L Jackson, Marie J. Kidd, Yan Wang, and Andrew M. Collins. The shape of the lymphocyte receptor repertoire: Lessons from the B cell receptor. *Frontiers in Immunology*, 4(SEP):1–12, 2013.

[76] Katherine J. L. Jackson, Yi Liu, Krishna M. Roskin, Jacob Glanville, Ramona A. Hoh, Katie Seo, Eleanor L. Marshall, Thaddeus C. Gurley, M. Anthony Moody, Barton F. Haynes, Emmanuel B. Walter, Hua-Xin Liao, Randy A. Albrecht, Adolfo Garca-Sastre, Javier Chaparro-Riggers, Arvind Rajpal, Jaume Pons, Birgitte B. Simen, Bozena Hanczaruk, Cornelia L. Dekker, Jonathan Laserson, Daphne Koller, Mark M. Davis, Andrew Z. Fire, and Scott D. Boyd. Human Responses to Influenza Vaccination Show Seroconversion Signatures and Convergent Antibody Rearrangements. *Cell Host & Microbe*, 16(1):105–114, July 2014.

[77] N. Jiang, J. He, J. A. Weinstein, L. Penland, S. Sasaki, X.-S. He, C. L. Dekker, N.-Y. Zheng, M. Huang, M. Sullivan, P. C. Wilson, H. B. Greenberg, M. M. Davis, D. S. Fisher, and S. R. Quake. Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination. *Science Translational Medicine*, 5(171):171ra19, February 2013.

[78] Gunilla B Karlsson Hedestam, Ron A M Fouchier, Sanjay Phogat, Dennis R Burton, Joseph Sodroski, and Richard T Wyatt. The challenges of eliciting neutralizing antibodies to HIV-1 and to influenza virus. *Nature reviews. Microbiology*, 6(2):143–55, February 2008.

[79] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, July 2002.

[80] Kaval Kaur, Meghan Sullivan, and Patrick C Wilson. Targeting B cell responses in

universal influenza vaccine design. *Trends in immunology*, 32(11):524–31, November 2011.

[81] T. B. Kepler. Codon bias and plasticity in immunoglobulins. *Molecular Biology and Evolution*, 14(6):637–643, June 1997.

[82] Edwin D. Kilbourne. Influenza Pandemics of the 20th Century. *Emerging Infectious Diseases*, 12(1):9–14, January 2006.

[83] Motoo Kimura. On the evolutionary adjustment of spontaneous mutation rates*. *Genetics Research*, 9(1):23–34, February 1967.

[84] Ericka Kirkpatrick, Xueting Qiu, Patrick C. Wilson, Justin Bahl, and Florian Krammer. The influenza virus hemagglutinin head evolves faster than the stalk domain. *Scientific Reports*, 8(1):1–14, July 2018.

[85] Bjrn F Koel, David F Burke, Theo M Bestebroer, Stefan van der Vliet, Gerben C M Zondag, Gaby Vervaet, Eugene Skepner, Nicola S Lewis, Monique I J Spronken, Colin A Russell, Mikhail Y Eropkin, Aeron C Hurt, Ian G Barr, Jan C de Jong, Guus F Rimmelzwaan, Albert D M E Osterhaus, Ron A M Fouchier, and Derek J Smith. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*, 342(6161):976–9, November 2013.

[86] Sergei L. Kosakovsky Pond, Simon D. W. Frost, Zehava Grossman, Michael B. Gravenor, Douglas D. Richman, and Andrew J. Leigh Brown. Adaptation to Different Human Populations by HIV-1 Revealed by Codon-Based Analyses. *PLoS Computational Biology*, 2(6):e62, 2006.

[87] Jens C Krause, Tshidi Tsibane, Terrence M Tumpey, Chelsey J Huffman, Bryan S Briney, Scott A Smith, Christopher F Basler, and James E Crowe. Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence diver-

gence and interclonal convergence. *Journal of immunology (Baltimore, Md. : 1950)*, 187(7):3704–11, October 2011.

[88] Adam J. Kucharski and Julia R. Gog. Age profile of immunity to influenza: Effect of original antigenic sin. *Theoretical Population Biology*, 81(2):102–112, March 2012.

[89] Adam J. Kucharski and Julia R. Gog. The Role of Social Contacts and Original Antigenic Sin in Shaping the Age Pattern of Immunity to Seasonal Influenza. *PLOS Computational Biology*, 8(10):e1002741, October 2012.

[90] Adam J Kucharski, Justin Lessler, Jonathan M Read, Huachen Zhu, Chao Qiang Jiang, Yi Guan, Derek A T Cummings, and Steven Riley. Estimating the life course of influenza A(H3N2) antibody responses from cross-sectional data. *PLoS biology*, 13(3):e1002082, March 2015.

[91] M. Senthil Kumar, Eric V. Slud, Kwame Okrah, Stephanie C. Hicks, Sridhar Hannenhalli, and Hctor Corrada Bravo. Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics*, 19(1):799, November 2018.

[92] Masayuki Kuraoka, Aaron G. Schmidt, Takuya Nojima, Feng Feng, Akiko Watanabe, Daisuke Kitamura, Stephen C. Harrison, Thomas B. Kepler, and Garnett Kelsoe. Complex Antigens Drive Permissive Clonal Selection in Germinal Centers. *Immunity*, 44(3):542–552, March 2016.

[93] Tomohiro Kurosaki, Kohei Kometani, and Wataru Ise. Memory B cells. *Nature Reviews Immunology*, 15(3):149–159, March 2015.

[94] Pinky Langat, Jayna Raghwani, Gytis Dudas, Thomas A. Bowden, Stephanie Edwards, Astrid Gall, Trevor Bedford, Andrew Rambaut, Rodney S. Daniels, Colin A. Russell, Oliver G. Pybus, John McCauley, Paul Kellam, and Simon J. Watson. Genome-wide evolutionary dynamics of influenza B viruses on a global scale. *PLOS Pathogens*, 13(12):e1006749, December 2017.

[95] Uri Laserson, Francois Vigneault, Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, Jason A. Vander Heiden, William Kelton, Sang Taek Jung, Yi Liu, Jonathan Laserson, Raj Chari, Je-Hyuk Lee, Ido Bachelet, Brendan Hickey, Erez Lieberman-Aiden, Bozena Hanczaruk, Birgitte B. Simen, Michael Egholm, Daphne Koller, George Georgiou, Steven H. Kleinstein, and George M. Church. High-resolution antibody dynamics of vaccine-induced immune responses. *Proceedings of the National Academy of Sciences*, 111(13):4928–4933, April 2014.

[96] Karen L. Laurie, William Horman, Louise A. Carolan, Kok Fei Chan, Daniel Layton, Andrew Bean, Dhanasekaran Vijaykrishna, Patrick C. Reading, James M. McCaw, and Ian G. Barr. Evidence for Viral Interference and Cross-reactive Protective Immunity Between Influenza B Virus Lineages. *The Journal of Infectious Diseases*, 217(4):548–559, January 2018.

[97] Egbert Giles Leigh. Natural Selection and Mutability. *The American Naturalist*, 104(937):301–305, 1970.

[98] Philippe Lemey, Sergei L Kosakovsky Pond, Alexei J Drummond, Oliver G Pybus, Beth Shapiro, Helena Barroso, Nuno Taveira, and Andrew Rambaut. Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS computational biology*, 3(2):e29, February 2007.

[99] Richard E. Lenski, Jeffrey E. Barrick, and Charles Ofria. Balancing Robustness and Evolvability. *PLOS Biology*, 4(12):e428, December 2006.

[100] R A Levandowski, P A Gross, M Weksler, E Staton, M S Williams, and J Bonelli. Cross-reactive antibodies induced by a monovalent influenza B virus vaccine. *Journal of Clinical Microbiology*, 29(7):1530–1532, July 1991.

[101] R. A. Levandowski, H. L. Regnery, E. Staton, B. G. Burgess, M. S. Williams, and J. R.

Groothuis. Antibody responses to influenza B viruses in immunologically unprimed children. *Pediatrics*, 88(5):1031–1036, November 1991.

[102] Yves Levy, Neetu Gupta, Franoise Le Deist, Corinne Garcia, Alain Fischer, Jean-Claude Weill, and Claude-Agns Reynaud. Defect in IgV gene somatic hypermutation in Common Variable Immuno-Deficiency syndrome. *Proceedings of the National Academy of Sciences*, 95(22):13135–13140, October 1998.

[103] G.-M. Li, C. Chiu, J. Wrammert, M. McCausland, S. F. Andrews, N.-Y. Zheng, J.-H. Lee, M. Huang, X. Qu, S. Edupuganti, M. Mulligan, S. R. Das, J. W. Yewdell, a. K. Mehta, P. C. Wilson, and R. Ahmed. Pandemic H1N1 influenza vaccine induces a recall response in humans that favors broadly cross-reactive memory B cells. *Proceedings of the National Academy of Sciences*, 109(23):9047–9052, 2012.

[104] Hua-Xin Liao, Rebecca Lynch, Tongqing Zhou, Feng Gao, S. Munir Alam, Scott D. Boyd, Andrew Z. Fire, Krishna M. Roskin, Chaim A. Schramm, Zhenhai Zhang, Jiang Zhu, Lawrence Shapiro, Jesse Becker, Betty Benjamin, Robert Blakesley, Gerry Bouffard, Shelise Brooks, Holly Coleman, Mila Dekhtyar, Michael Gregory, Xiaobin Guan, Jyoti Gupta, Joel Han, April Hargrove, Shi-ling Ho, Taccara Johnson, Richelle Legaspi, Sean Lovett, Quino Maduro, Cathy Masiello, Baishali Maskeri, Jenny McDowell, Casandra Montemayor, James Mullikin, Morgan Park, Nancy Riebow, Karen Schandler, Brian Schmidt, Christina Sison, Mal Stantripop, James Thomas, Pam Thomas, Meg Vemulapalli, Alice Young, James C. Mullikin, S. Gnanakaran, Peter Hraber, Kevin Wiehe, Garnett Kelsoe, Guang Yang, Shi-Mao Xia, David C. Montefiori, Robert Parks, Krissey E. Lloyd, Richard M. Scearce, Kelly A. Soderberg, Myron Cohen, Gift Kamanga, Mark K. Louder, Lillian M. Tran, Yue Chen, Fangping Cai, Sheri Chen, Stephanie Moquin, Xiulian Du, M. Gordon Joyce, Sanjay Srivatsan, Baoshan Zhang, Anqi Zheng, George M. Shaw, Beatrice H. Hahn, Thomas B. Kepler, Bette T. M. Korber, Peter D. Kwong, John R. Mascola, and Barton F. Haynes.

Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature*, 496(7446):469–476, 2013.

[105] Susanne L Linderman, Benjamin S Chambers, Seth J Zost, Kaela Parkhouse, Yang Li, Christin Herrmann, Ali H Ellebedy, Donald M Carter, Sarah F Andrews, Nai-Ying Zheng, Min Huang, Yunping Huang, Donna Strauss, Beth H Shaz, Richard L Hodinka, Gustavo Reyes-Tern, Ted M Ross, Patrick C Wilson, Rafi Ahmed, Jesse D Bloom, and Scott E Hensley. Potential antigenic explanation for atypical H1N1 infections among middle-aged adults during the 2013-2014 influenza season. *Proceedings of the National Academy of Sciences of the United States of America*, 111(44):15798–15803, October 2014.

[106] Y.-J. Liu, D. E. Joshua, G. T. Williams, C. A. Smith, J. Gordon, and I. C. M. MacLennan. Mechanism of antigen-driven selection in germinal centres. *Nature*, 342(6252):929, December 1989.

[107] L Lopez and Q Sue Huang. Influenza surveillance in New Zealand - 2012, April 2013.

[108] Michael Lynch, Matthew S. Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W. Kelley Thomas, and Patricia L. Foster. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, 17(11):704–714, November 2016.

[109] Ian C. M. MaClennan and David Gray. Antigen-Driven Selection of Virgin and Memory B Cells. *Immunological Reviews*, 91(1):61–86, 1986.

[110] John J. Marchalonis, Samuel F. Schluter, Ralph M. Bernstein, Shanxiang Shen, and Allen B. Edmundson. Phylogenetic Emergence and Molecular Evolution of the Immunoglobulin Family. In Frank J. Dixon, editor, *Advances in Immunology*, volume 70, pages 417–506. Academic Press, January 1998.

[111] Andreas Mayer, Vijay Balasubramanian, Thierry Mora, and Aleksandra M. Walczak.

How a well-adapted immune system is organized. *Proceedings of the National Academy of Sciences*, 112(19):5950–5955, May 2015.

[112] Andreas Mayer, Thierry Mora, Olivier Rivoire, and Aleksandra M. Walczak. Diversity of immune strategies explained by adaptation to pathogen statistics. *Proceedings of the National Academy of Sciences*, 113(31):8630–8635, August 2016.

[113] Connor O McCoy, Trevor Bedford, Vladimir N Minin, Philip Bradley, Harlan Robins, and Frederick A Matsen. Quantifying evolutionary constraints on B-cell affinity maturation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1676):20140244–, September 2015.

[114] Alessia Melegaro, Mark Jit, Nigel Gay, Emilio Zagheni, and W. John Edmunds. What types of contacts are important for the spread of infections? Using contact survey data to explore European mixing patterns. *Epidemics*, 3(3):143–151, September 2011.

[115] Luka Mesin, Arin Schiepers, Jonatan Ersching, Alexandru Barbulescu, Ceclia B. Cavazzoni, Alessandro Angelini, Takaharu Okada, Tomohiro Kurosaki, and Gabriel D. Victora. Restricted Clonality and Limited Germinal Center Reentry Characterize Memory B Cell Reactivation by Boosting. *Cell*, 180(1):92–106.e11, January 2020.

[116] Arnold S. Monto, James S. Koopman, and Ira M. Longini. Tecumseh Study of Illness XIII. Influenza infection and disease, 19761981. *American Journal of Epidemiology*, 121(6):811–822, June 1985.

[117] Jol Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, Janneke Heijne, Malgorzata Sadkowska-Todys, Magdalena Rosinska, and W. John Edmunds. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLoS Medicine*, 5(3):e74, March 2008.

[118] Raffael Nachbagauer, Angela Choi, Ruvim Izikson, Manon M. Cox, Peter Palese, and Florian Krammer. Age Dependence and Isotype Specificity of Influenza Virus Hemagglutinin Stalk-Reactive Antibodies in Humans. *mBio*, 7(1):e01996–15, March 2016.

[119] Yasukazu Nakamura, Takashi Gojobori, and Toshimichi Ikemura. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research*, 28(1):292–292, January 2000.

[120] John D O'Brien, Vladimir N Minin, and Marc A Suchard. Learning to count: robust estimates for labeled distances between molecular sequences. *Molecular biology and evolution*, 26(4):801–14, April 2009.

[121] Line Ohm-laursen, Torben Barington, and Email Alerts. Analysis of 6912 Unselected Somatic Hypermutations in Human VDJ Rearrangements Reveals Lack of Strand Specificity and Correlation between Phase II Substitution Rates and Distance to the Nearest 3' Activation-Induced Cytidine Deaminase Target. *Journal of Immunology*, 178:4322–4334, 2007.

[122] Mihaela Oprea and Thomas B Kepler. Genetic Plasticity of V Genes Under Somatic Hypermutation : Statistical Analyses Using a New Resampling-Based Methodology. *Genome Research*, 919:1294–1304, 1999.

[123] Csaba Pal, Mara D. Maci, Antonio Oliver, Ira Schachar, and Angus Buckling. Coevolution with viruses drives the evolution of bacterial mutation rates. *Nature*, 450(7172):1079–1081, December 2007.

[124] Zeev Pancer, Chris T. Amemiya, Gtz R. A. Ehrhardt, Jill Ceitlin, G. Larry Gartland, and Max D. Cooper. Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature*, 430(6996):174–180, July 2004.

[125] Leontios Pappas, Mathilde Foglierini, Luca Piccoli, Nicole L. Kallewaard, Filippo Turrini, Chiara Silacci, Blanca Fernandez-Rodriguez, Gloria Agatic, Isabella Giacchetto-

Sasselli, Gabriele Pellicciotta, Federica Sallusto, Qing Zhu, Elisa Vicenzi, Davide Corti, and Antonio Lanzavecchia. Rapid development of broadly influenza neutralizing antibodies through redundant mutations. *Nature*, 516(7531):418–422, October 2014.

[126] Poornima Parameswaran, Yi Liu, Krishna M. Roskin, Katherine K. L. Jackson, Vaishali P. Dixit, Ji-Yeun Lee, Karen L. Artiles, Simona Zompi, Maria Jos Vargas, Birgitte B. Simen, Bozena Hanczaruk, Kim R. McGowan, Muhammad A. Tariq, Nader Pourmand, Daphne Koller, Angel Balmaseda, Scott D. Boyd, Eva Harris, and Andrew Z. Fire. Convergent Antibody Signatures in Human Dengue. *Cell Host & Microbe*, 13(6):691–700, June 2013.

[127] Joshua G. Petrie, Kaela Parkhouse, Suzanne E. Ohmit, Ryan E. Malosh, Arnold S. Monto, and Scott E. Hensley. Antibodies Against the Current Influenza A(H1N1) Vaccine Strain Do Not Protect Some Individuals From Infection With Contemporary Circulating Influenza A(H1N1) Virus Strains. *The Journal of Infectious Diseases*, 214(12):1947–1951, December 2016.

[128] Simona Puzelli, Angela Di Martino, Marzia Facchini, Concetta Fabiani, Laura Calzoletti, Giuseppina Di Mario, Annapina Palmieri, Paola Affanni, Barbara Camilloni, Maria Chironna, Pierlanfranco DAgaro, Simone Giannecchini, Elena Pariani, Caterina Serra, Caterina Rizzo, Antonino Bella, Isabella Donatelli, Maria Rita Castrucci, Filippo Ansaldi, Rosaria Arvia, Alberta Azzi, Patrizia Bagnarelli, Fausto Baldanti, Maria Rosaria Capobianchi, Silvana Castaldi, Maria Eugenia Colucci, Cristina Galli, Valeria Ghisetti, Andrea Orsi, Elisabetta Pagani, Giorgio Pal, Maurizio Sanguinetti, Riccardo Smeraglia, Fabio Tramuto, Francesco Vitale, and the Italian Influenza Laboratory Network. Co-circulation of the two influenza B lineages during 13 consecutive influenza surveillance seasons in Italy, 20042017. *BMC Infectious Diseases*, 19(1):990, November 2019.

[129] Madhusudan Rajendran, Raffael Nachbagauer, Megan E. Ermler, Paul Bunduc, Fatima

Amanat, Ruvim Izikson, Manon Cox, Peter Palese, Maryna Eichelberger, and Florian Krammer. Analysis of Anti-Influenza Virus Neuraminidase Antibodies in Children, Adults, and the Elderly by ELISA and Enzyme Inhibition: Evidence for Original Antigenic Sin. *mBio*, 8(2):e02281–16, May 2017.

[130] D. K. Ralph and F. A. IV Matsen. Likelihood-based inference of B cell clonal families. *arXiv*, 1603.08127, 2016.

[131] Duncan K. Ralph and Frederick A. Matsen Iv. Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data. *PLOS Computational Biology*, 15(7):e1007133, July 2019.

[132] Duncan K Ralph and Frederick A Matsen. Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. *PLoS computational biology*, 12(1):e1004409, January 2016.

[133] Sylvia Ranjeva, Rahul Subramanian, Vicky J. Fang, Gabriel M. Leung, Dennis K. M. Ip, Ranawaka A. P. M. Perera, J. S. Malik Peiris, Benjamin J. Cowling, and Sarah Cobey. Age-specific differences in the dynamics of protective immunity to influenza. *Nature Communications*, 10(1):1–11, April 2019.

[134] Vincent Ranwez, Sbastien Harispe, Frdric Delsuc, Emmanuel J. P. Douzery, and K Beal. MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLoS ONE*, 6(9):e22594, September 2011.

[135] Yevgeniy Raynes, Matthew R. Gazzara, and Paul D. Sniegowski. Mutator dynamics in sexual and asexual experimental populations of yeast. *BMC Evolutionary Biology*, 11(1):158, June 2011.

[136] Yevgeniy Raynes, C. Scott Wylie, Paul D. Sniegowski, and Daniel M. Weinreich. Sign of selection on mutation rate modifiers depends on population size. *Proceedings of the National Academy of Sciences*, 115(13):3422–3427, March 2018.

[137] I. B. Rogozin and M. Diaz. Cutting Edge: DGYW/WRCH Is a Better Predictor of Mutability at G:C Bases in Ig Hypermutation Than the Widely Accepted RGYW/WRCY Motif and Probably Reflects a Two-Step Activation-Induced Cytidine Deaminase-Triggered Process. *The Journal of Immunology*, 172(6):3382–3384, March 2004.

[138] Igor B. Rogozin and Nikolai A. Kolchanov. Somatic hypermutagenesis in immunoglobulin genes II. Influence of neighbouring base sequences on mutagenesis. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1171(1):11–18, November 1992.

[139] P. A. Rota, T. R. Wallis, M. W. Harmon, J. S. Rota, A. P. Kendal, and K. Nerome. Cocirculation of two distinct evolutionary lineages of influenza type B virus since 1983. *Virology*, 175(1):59–68, March 1990.

[140] Jasmine Saini and Uri Hershberg. B cell Variable genes have evolved their codon usage to focus the targeted patterns of somatic mutation on the complementarity determining regions. *Molecular Immunology*, 65(1):157–167, May 2015.

[141] A. Sauerbrei, T. Langenhan, A. Brandstdt, R. Schmidt-Ott, A. Krumbholz, H. Girschick, H. Huppertz, P. Kaiser, J. Liese, A. Streng, T. Niehues, J. Peters, A. Sauerbrey, H. Schroten, T. Tenenbaum, S. Wirth, and P. Wutzler. Prevalence of antibodies against influenza A and B viruses in children in Germany, 2008 to 2010. *Eurosurveillance*, 19(5):20687, February 2014.

[142] Andreas Sauerbrei, R. Schmidt-Ott, H. Hoyer, and P. Wutzler. Seroprevalence of influenza A and B in German infants and adolescents. *Medical Microbiology and Immunology*, 198(2):93, February 2009.

[143] Aaron G Schmidt, Matthew D Therkelsen, Shaun Stewart, Thomas B Kepler, Hua-Xin Liao, M Anthony Moody, Barton F Haynes, and Stephen C Harrison. Viral receptor-binding site antibodies with diverse germline origins. *Cell*, 161(5):1026–34, May 2015.

[144] Michael W. Shaw, Xiyan Xu, Yan Li, Susan Normand, Robert T. Ueki, Gail Y. Kunimoto, Henrietta Hall, Alexander Klimov, Nancy J. Cox, and Kanta Subbarao. Reappearance and Global Spread of Variants of Influenza B/Victoria/2/87 Lineage Viruses in the 20002001 and 20012002 Seasons. *Virology*, 303(1):1–8, November 2002.

[145] Zizhang Sheng, Chaim A. Schramm, M Connors, Lynn Morris, John R Mascola, Peter D Kwong, and Lawrence Shapiro. Effects of Darwinian Selection and Mutability on Rate of Broadly Neutralizing Antibody Evolution during HIV-1 Infection. *PLOS Computational Biology*, 12(5):e1004940, May 2016.

[146] Danuta M Skowronski, Catharine Chambers, Gaston De Serres, Suzana Sabaiduc, Anne-Luise Winter, James A Dickinson, Jonathan B Gubbay, Steven J Drews, Kevin Fonseca, Hugues Charest, Christine Martineau, Rebecca Hickman, Tracy Chan, Agatha Jassem, Martin Petric, Caren Rose, Nathalie Bastien, Yan Li, and Mel Krajden. Vaccine Effectiveness Against Lineage-matched and -mismatched Influenza B Viruses Across 8 Seasons in Canada, 20102011 to 20172018. *Clinical Infectious Diseases*, 68(10):1754–1757, May 2019.

[147] Danuta M Skowronski, Catharine Chambers, Gaston De Serres, Suzana Sabaiduc, Anne-Luise Winter, James A Dickinson, Jonathan B Gubbay, Kevin Fonseca, Steven J Drews, Hugues Charest, Christine Martineau, Mel Krajden, Martin Petric, Nathalie Bastien, and Yan Li. Age-Related Differences in Influenza B Infection by Lineage in a Community-Based Sentinel System, 20102011 to 20152016, Canada. *The Journal of Infectious Diseases*, 216(6):697–702, September 2017.

[148] Danuta M. Skowronski, Marie-Eve Hamelin, Naveed Z. Janjua, Gaston De Serres, Jennifer L. Gardy, Chantal Rhaume, Xavier Bouhy, and Guy Boivin. Cross-Lineage Influenza B and Heterologous Influenza A Antibody Responses in Vaccinated Mice: Immunologic Interactions and B/Yamagata Dominance. *PLoS ONE*, 7(6):e38929, June 2012.

[149] A. J. Smith and Joan R. Davies. Natural Infection with Influenza A (H3N2). The Development, Persistence and Effect of Antibodies to the Surface Antigens. *The Journal of Hygiene*, 77(2):271–282, 1976.

[150] Derek J. Smith, Alan S. Lapedes, Jan C. de Jong, Theo M. Bestebroer, Guus F. Rimmelzwaan, Albert D. M. E. Osterhaus, and Ron A. M. Fouchier. Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science*, 305(5682):371–376, July 2004.

[151] Paul D. Sniegowski, Philip J. Gerrish, Toby Johnson, and Aaron Shaver. The evolution of mutation rates: separating causes from consequences. *BioEssays*, 22(12):1057–1066, 2000.

[152] Paul D. Sniegowski, Philip J. Gerrish, and Richard E. Lenski. Evolution of high mutation rates in experimental populations of E. coli. *Nature*, 387(6634):703–705, June 1997.

[153] Maja Soan, Katarina Prosenc, Veronika Uakar, and Nataa Berginc. A comparison of the demographic and clinical characteristics of laboratory-confirmed influenza B Yamagata and Victoria lineage infection. *Journal of Clinical Virology*, 61(1):156–160, September 2014.

[154] Kathleen Sprouffske, Jos Aguilar-Rodrguez, Paul Sniegowski, and Andreas Wagner. High mutation rates limit evolutionary adaptation in Escherichia coli. *PLOS Genetics*, 14(4):e1007324, April 2018.

[155] Christopher T. Stamper and Patrick C. Wilson. What Are the Primary Limitations in B-Cell Affinity Maturation, and How Much Affinity Maturation Can We Drive with Vaccination? Is Affinity Maturation a Self-Defeating Process for Eliciting Broad Protection? *Cold Spring Harbor Perspectives in Biology*, 10(5):a028803, May 2018.

[156] Nicolas B. Strauli and Ryan D. Hernandez. Statistical inference of a convergent antibody repertoire response to influenza vaccine. *Genome Medicine*, 8(1):60, June 2016.

[157] Weina Sun, Davina S. Kang, Allen Zheng, Sean T. H. Liu, Felix Broecker, Viviana Simon, Florian Krammer, and Peter Palese. Antibody Responses toward the Major Antigenic Sites of Influenza B Virus Hemagglutinin in Mice, Ferrets, and Humans. *Journal of Virology*, 93(2), January 2019.

[158] Yi Sun, Chunyan Wang, Yating Wang, Tianyi Zhang, Liming Ren, Xiaoxiang Hu, Ran Zhang, Qingyong Meng, Ying Guo, Jing Fei, Ning Li, and Yaofeng Zhao. A comprehensive analysis of germline and expressed immunoglobulin repertoire in the horse. *Developmental & Comparative Immunology*, 34(9):1009–1020, September 2010.

[159] Way Sung, Matthew S. Ackerman, Samuel F. Miller, Thomas G. Doak, and Michael Lynch. Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences*, 109(45):18488–18492, 2012.

[160] F. Taddei, M. Radman, J. Maynard-Smith, B. Toupance, P. H. Gouyon, and B. Godelle. Role of mutator alleles in adaptive evolution. *Nature*, 387(6634):700–702, June 1997.

[161] David W. Talmage. Immunological Specificity: Unique combinations of selected natural globulins provide an alternative to the classical concept. *Science*, 129(3364):1643–1648, June 1959.

[162] Yi Tan, Wenda Guan, Tommy Tsan-Yuk Lam, Sihua Pan, Shiguan Wu, Yangqing Zhan, Cecile Viboud, Edward C. Holmes, and Zifeng Yang. Differing Epidemiological Dynamics of Influenza B Virus Lineages in Guangzhou, Southern China, 2009-2010. *Journal of Virology*, 87(22):12447–12456, November 2013.

[163] J. M. J. Tas, L. Mesin, G. Pasqual, S. Targ, J. T. Jacobsen, Y. M. Mano, C. S. Chen, J.-C. Weill, C.-A. Reynaud, E. P. Browne, M. Meyer-Hermann, and G. D. Victora.

Visualizing antibody affinity maturation in germinal centers. *Science*, 58(12):7250–7, 2016. arXiv: 1011.1669v3 ISBN: 1498224114983.

[164] Grace Teng and F Nina Papavasiliou. Immunoglobulin somatic hypermutation. *Annual Review of Genetics*, 41:107–120, 2007.

[165] Susumu Tonegawa. Somatic generation of antibody diversity. *Nature*, 302(5909):575, April 1983.

[166] Andrea C. Tricco, Ayman Chit, Charlene Soobiah, David Hallett, Genevieve Meier, Maggie H. Chen, Mariam Tashkandi, Chris T. Bauch, and Mark Loeb. Comparing influenza vaccine efficacy against mismatched and matched strains: a systematic review and meta-analysis. *BMC Medicine*, 11(1):153, June 2013.

[167] Menno C. van Zelm, Tomasz Szczepaski, Mirjam van der Burg, and Jacques J. M. van Dongen. Replication history of B lymphocytes reveals homeostatic proliferation and extensive antigen-induced B cell expansion. *Journal of Experimental Medicine*, 204(3):645–655, March 2007.

[168] Gabriel D. Victora and Michel C. Nussenzweig. Germinal Centers. *Annual Review of Immunology*, 30(1):429–457, 2012.

[169] Gabriel D. Victora and Patrick C. Wilson. Germinal Center Selection and the Antibody Response to Influenza. *Cell*, 163(3):545–548, October 2015.

[170] Marcos C. Vieira, Daniel Zinder, and Sarah Cobey. Selection and Neutral Mutations Drive Pervasive Mutability Losses in Long-Lived Anti-HIV B-Cell Lineages. *Molecular Biology and Evolution*, 35(5):1135–1146, May 2018.

[171] Dhanasekaran Vijaykrishna, E. C. Holmes, Udayan Joseph, Yvonne C. F. Su, Rebecca Halpin, Raphael T. C. Lee, Yi-Mo Deng, Vithiagaran Gunalan, Xudong Lin,

Timothy B. Stockwell, Nadia B. Fedorova, Bin Zhou, Natalie Spirason, Denise Khnert, Veronika Bokov, Tanja Stadler, Anna-Maria Costa, Dominic E. Dwyer, Q Sue Huang, Lance C Jennings, W. Rawlinson, Sheena G. Sullivan, Aeron C Hurt, Sebastian Maurer-Stroh, David E Wentworth, Gavin J. D. Smith, and Ian G Barr. The contrasting phylodynamics of human influenza B viruses | eLife. *eLife*, 4:e05055, 2015.

[172] Ramandeep K. Virk, Jayanthi Jayakumar, Ian H. Mendenhall, Mahesh Moorthy, Pauline Lam, Martin Linster, Julia Lim, Cui Lin, Lynette L. E. Oon, Hong Kai Lee, Evelyn S. C. Koay, Dhanasekaran Vijaykrishna, Gavin J. D. Smith, and Yvonne C. F. Su. Divergent evolutionary trajectories of influenza B viruses underlie their contemporaneous epidemic activity. *Proceedings of the National Academy of Sciences*, 117(1):619–628, January 2020.

[173] Andreas Wagner. Robustness, evolvability, and neutrality. *FEBS Letters*, 579(8):1772–1778, 2005.

[174] Andreas Wagner. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B: Biological Sciences*, 275(1630):91–100, January 2008.

[175] S. D. Wagner, C. Milstein, and M. S. Neuberger. Codon bias targets mutation. *Nature*, 376(6543):732, August 1995.

[176] Lirong Wei, Richard Chahwan, Shanzhi Wang, Xiaohua Wang, Phuong T. Pham, Myron F. Goodman, Aviv Bergman, Matthew D. Scharff, and Thomas MacCarthy. Overlapping hotspots in CDRs are critical sites for V region diversification. *Proceedings of the National Academy of Sciences*, pages E728–E737, 2015.

[177] U. Weiss and K. Rajewsky. The repertoire of somatic antibody mutants accumulating in the memory compartment after primary immunization is restricted through affinity maturation and mirrors that expressed in the secondary response. *Journal of Experimental Medicine*, 172(6):1681–1689, December 1990.

[178] Claus O. Wilke, Jia Lan Wang, Charles Ofria, Richard E. Lenski, and Christoph Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333, July 2001.

[179] S. Williamson and M. E. Orive. The Genealogy of a Sequence Subject to Purifying Selection at Multiple Sites. *Molecular Biology and Evolution*, 19(8):1376–1384, August 2002.

[180] Jens Wrammert, Dimitrios Koutsonanos, Gui-Mei Li, Srilatha Edupuganti, Jianhua Sui, Michael Morrissey, Megan McCausland, Ioanna Skountzou, Mady Hornig, W Ian Lipkin, Aneesh Mehta, Behzad Razavi, Carlos Del Rio, Nai-Ying Zheng, Jane-Hwei Lee, Min Huang, Zahida Ali, Kaval Kaur, Sarah Andrews, Rama Rao Amara, Youliang Wang, Suman Ranjan Das, Christopher David O'Donnell, Jon W Yewdell, Kanta Subbarao, Wayne a Marasco, Mark J Mulligan, Richard Compans, Rafi Ahmed, and Patrick C Wilson. Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *The Journal of Experimental Medicine*, 208(1):181–193, 2011.

[181] Xueling Wu, Zhenhai Zhang, Chaim A. Schramm, M. Gordon Joyce, Young Do Kwon, Tongqing Zhou, Zizhang Sheng, Baoshan Zhang, Sijy ODell, Krisha McKee, Ivelin S. Georgiev, Gwo-Yu Chuang, Nancy S. Longo, Rebecca M. Lynch, Kevin O. Saunders, Cinque Soto, Sanjay Srivatsan, Yongping Yang, Robert T. Bailer, Mark K. Louder, James C. Mullikin, Mark Connors, Peter D. Kwong, John R. Mascola, and Lawrence Shapiro. Maturation and Diversity of the VRC01-Antibody Lineage over 15 Years of Chronic HIV-1 Infection. *Cell*, 161(3):470–485, April 2015.

[182] Cuiling Xu, Kwok-Hung Chan, Tim K. Tsang, Vicky J. Fang, Rita O. P. Fung, Dennis K. M. Ip, Simon Cauchemez, Gabriel M. Leung, J. S. Malik Peiris, and Benjamin J. Cowling. Comparative Epidemiology of Influenza B Yamagata- and Victoria-Lineage

Viruses in Households. *American Journal of Epidemiology*, 182(8):705–713, October 2015.

[183] Gur Yaari, Jennifer I. C. Benichou, Jason A. Vander Heiden, Steven H. Kleinstein, and Yoram Louzoun. The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676):20140242, September 2015.

[184] Gur Yaari, Mohamed Uduman, and Steven Kleinstein. Quantifying selection in high-throughput Immunoglobulin sequencing datasets (58.4). *The Journal of Immunology*, 188(1 Supplement):58.4–58.4, May 2012.

[185] Gur Yaari, Jason A Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Namita Gupta, Joel N H Stern, Kevin C O'Connor, David A Hafler, Uri Laserson, Francois Vigneault, and Steven H Kleinstein. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Frontiers in Immunology*, 4:358, January 2013.

[186] Feixue Yang, Geoffrey C. Waldbieser, and Craig J. Lobb. The Nucleotide Targets of Somatic Mutation and the Role of Selection in Immunoglobulin Heavy Chains of a Teleost Fish. *The Journal of Immunology*, 176(3):1655–1667, February 2006.

[187] Veronika I Zarnitsyna, Ali H Ellebedy, Carl Davis, Joshy Jacob, Rafi Ahmed, and Rustom Antia. Masking of antigenic epitopes by antibodies shapes the humoral immune response to influenza. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1676):20140248–, September 2015.

[188] Veronika I. Zarnitsyna, Jennie Lavine, Ali Ellebedy, Rafi Ahmed, and Rustom Antia. Multi-epitope Models Explain How Pre-existing Antibodies Affect the Generation of Broadly Protective Responses to Influenza. *PLOS Pathogens*, 12(6):e1005692, June 2016.

[189] Menno C. van Zelm, Mirjam van der Burg, and Jacques J. M. van Dongen. Homeostatic and Maturation-associated Proliferation in the Peripheral B-Cell Compartment. *Cell Cycle*, 6(23):2890–2895, December 2007.

[190] Daniel Zinder, Trevor Bedford, Sunetra Gupta, and Mercedes Pascual. The Roles of Competition and Mutation in Shaping Antigenic and Genetic Diversity in Influenza. *PLoS Pathogens*, 9(1), 2013. ISBN: 1553-7374 (Electronic)\r1553-7366 (Linking).